Human Molecular Genetics, 2020, Vol. 29, No. 13 2275–2284

doi: 10.1093/hmg/ddaa083
Advance Access Publication Date: 3 June 2020
Association Studies Article

ASSOCIATION STUDIES ARTICLE

Population-specific reference panels are crucial for genetic analyses: an example of the *CREBRF* locus in Native Hawaiians

Meng Lin[1,†], Christian Caberto[2], Peggy Wan[1], Yuqing Li[3], Annette Lum-Jones[2], Maarit Tiirikainen[2,‡], Loreall Pooler[1], Brooke Nakamura[1], Xin Sheng[1,¶], Jacqueline Porcel[1], Unhee Lim[2], Veronica Wendy Setiawan[1], Loïc Le Marchand[2], Lynne R. Wilkens[2], Christopher A. Haiman[1], Iona Cheng[3] and Charleston W. K. Chiang[1,4,*,‖]

[1]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA, [2]Epidemiology Program, University of Hawai'i Cancer Center, University of Hawai'i at Mānoa, Honolulu, HI 96813, USA, [3]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA 94518, USA, and [4]Quantitative Computational Biology Section, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

*To whom correspondence should be addressed at: University of Southern California, Los Angeles, CA 90033, USA. Tel: +323-442-8052; Email: charleston.chiang@med.usc.edu

Abstract

Statistical imputation applied to genome-wide array data is the most cost-effective approach to complete the catalog of genetic variation in a study population. However, imputed genotypes in underrepresented populations incur greater inaccuracies due to ascertainment bias and a lack of representation among reference individuals, further contributing to the obstacles to study these populations. Here we examined the consequences due to the lack of representation by genotyping in a large number of self-reported Native Hawaiians ($N = 3693$) a functionally important, Polynesian-specific variant in the *CREBRF* gene, rs373863828. We found the derived allele was significantly associated with several adiposity traits with large effects (e.g. ~ 1.28 kg/m^2 per allele in body mass index as the most significant; $P = 7.5 \times 10^{-5}$), consistent with the original findings in Samoans. Due to the current absence of Polynesian representation in publicly accessible reference sequences, rs373863828 or its proxies could not be tested through imputation using these existing resources. Moreover, the association signals at the entire *CREBRF* locus could not be captured by alternative approaches, such as admixture mapping. In contrast, highly accurate imputation can be achieved even if a small number (<200) of internally constructed Polynesian reference individuals were available; this would increase sample size and improve the statistical evidence of associations. Taken together, our results suggest the alarming possibility that lack of representation in reference panels could inhibit discovery of functionally important loci such as *CREBRF*. Yet, they could be easily detected and prioritized with improved representation of diverse populations in sequencing studies.

[†]Meng Lin, http://orcid.org/0000-0003-4603-0718
[‡]Maarit Tiirikainen, http://orcid.org/0000-0002-8124-3818
[¶]Xin Sheng, http://orcid.org/0000-0002-6844-2740
[‖]Charleston W. K. Chiang, http://orcid.org/0000-0002-0668-7865
Received: November 18, 2019. **Revised:** March 16, 2020. **Accepted:** March 17, 2020

Introduction

Statistical imputation of untyped variants is a crucial step for large-scale genetic investigations of complex traits. By comparing to an appropriate reference panel, often based on whole-genome sequences of individuals, statistical imputation infers the genotype at variant sites not covered on genotyping arrays (1). Therefore, imputation benefits many study cohorts with its balance between budget and coverage of the genome. Nevertheless, inadequacy and inaccuracy of imputed markers can impede downstream genetic studies or clinical screening. This problem could arise when reference panels are not genetically close to the population of interest, further exacerbated by ascertainment bias of existing genotyping array. Today, it is increasingly recognized that there has been a severe bias toward studying individuals of European origin in genome-wide association studies (GWAS) (2–5); the same bias also exist in the largest available reference panel for imputation (e.g. the Haplotype Reference Consortium (6)). The inability to statistically impute diverse populations further hinders progress in studying these diverse populations that otherwise are already underserved (7,8).

One of the major challenges in studying diverse non-European populations, particularly for indigenous communities such as the Native Hawaiians, is the inability to accrue large sample sizes (9). For example, GWAS in European ancestry-based cohorts numbers in greater than 1 million (10). By contrast, there are only approximately 1.2 million individuals in total living in the United States that may derive some part of their ancestry to Native Hawaiians, according to the US 2010 census survey (https://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf). Therefore, the focus in studying diverse populations is often in (1) evaluating the transferability of findings from large-scale European ancestry-based studies and (2) identifying population-specific variants that may contribute to genetic risks in non-European populations. As they are often very rare or missing in Europeans, population-specific variants in non-Europeans are usually absent in most genotyping arrays. These variants would rely on high-quality imputation in order to be captured, unless the population of interest is whole-genome sequenced at large scale. One recent example of a variant that bears important consequences to the health and disease risk of a population is the Polynesian-specific missense variant rs373863828 in gene *CREBRF*. This locus was initially detected because of an association signal of a proxy variant, rs12513649, which was on the Affymetrix 6.0 array. The missense rs373863828 was then discovered through targeted resequencing of a small number of private Samoan sequences, followed by imputation into the entire Samoan cohort and validation of the imputed genotypes. Rs373863828 was found to have a large effect on body mass index (BMI) as well as on a number of other adiposity, metabolic and anthropometric traits in Samoans (11). Despite an estimated 26% allele frequency in Samoans, the derived allele is only segregating in the few Pacific Islands populations and not found elsewhere in the world (12–16). Because of a lack in Polynesian haplotypes in publicly accessible sequencing databases (e.g. 1000 Genomes Project or Haplotype Reference Consortium), this variant could not be directly imputed and studied by researchers with publicly available resources.

In this study, we use the *CREBRF* locus as an example to examine the potential limitations of post-imputation analyses in the absence of a proper representation in reference panels. We genotyped the variant rs373863828 in self-reported Native Hawaiians from the Multiethnic Cohort (MEC). The Hawai'i archipelago in northeast Polynesia was first settled between 1200 and 1800 *ya* (17–20). Historically, Native Hawaiians have remained relatively isolated on the northern Pacific islands, until their recent encounter with intercontinental migrants from Europe around the late eighteenth century and from East Asia (mainly China and Japan) during the nineteenth to twentieth century, followed by minor contributions from other populations around the world (17,18). Compared to other populations, contemporary Native Hawaiians have higher incidence rate of obesity-related medical conditions, such as diabetes and cardiovascular diseases (21). Therefore, it is of clinical importance to characterize the impact of variants with potentially large effects on adiposity, such as rs373863828 in *CREBRF*, in this population. We demonstrate that despite a strong impact on adiposity in Native Hawaiians, the *CREBRF* locus could not be discovered using conventional mapping methods with currently available resources, suggesting important challenges for discovering additional variants contributing to population-specific genetic risks. However, our findings also suggest that these challenges could be mitigated if the representation in reference sequences were improved, even if just marginally so.

Results

The missense variant in CREBRF is correlated with the proportion of Polynesian ancestry

To characterize the functional missense variant, rs373863828-A, in Native Hawaiians, we genotyped this single variant in 3693 self-reported Native Hawaiian individuals from the MEC and utilized the genotype data along with their existing genome-wide array data (Supplementary Material, Table S1). We also genotyped this variant in 1538 MEC individuals from other continental populations (Materials and Methods). Consistent with previous reports that this variant is found exclusively in Pacific Islanders (11,12), we estimated the derived allele frequency to be 5.9% in Native Hawaiians, but is monomorphic in all other ethnicities we genotyped. As Native Hawaiians derive a large proportion of their ancestry from Polynesians, we also found significant correlations between the derived allele frequency (DAF) at rs373863828 and bins of individuals' estimated Polynesian ancestry proportions ($r = 0.98$, $P = 6.2e^{-7}$, Supplementary Material, Fig. S1) and between the direct genotypes and individual ancestry proportions (GLM $\beta = 2.67$, $P < 2e^{-16}$). For brevity, we refer to the 152 individuals with estimated Polynesian ancestry >90% as unadmixed with respect to major continental ancestry (Europeans, East Asians, and Africans) while fully acknowledging that this practice is for convenience, rather than for precision or accuracy. We find that in these unadmixed individuals, the derived allele segregates at the frequency of 12.8% (Supplementary Material, Fig. S1).

The lower frequency (12.8%) among unadmixed Native Hawaiian individuals is unexpected, given that the allele has been reported to be under positive selection and is segregating at 26% in Samoans (11). We attempted to replicate the signal of selection at this locus in the Native Hawaiians using the nSL (22) among the 152 individuals with estimated Polynesian ancestry >90%. Compared to randomly drawn variants throughout the genome matched by derived allele frequency, we found no evidence of rs373863828-A being positively selected in these 152 individuals (nSL = 0.72; $P = 0.57$; Supplementary Material, Fig. S2), although our power may be limited due to the small sample size of unadmixed individuals.

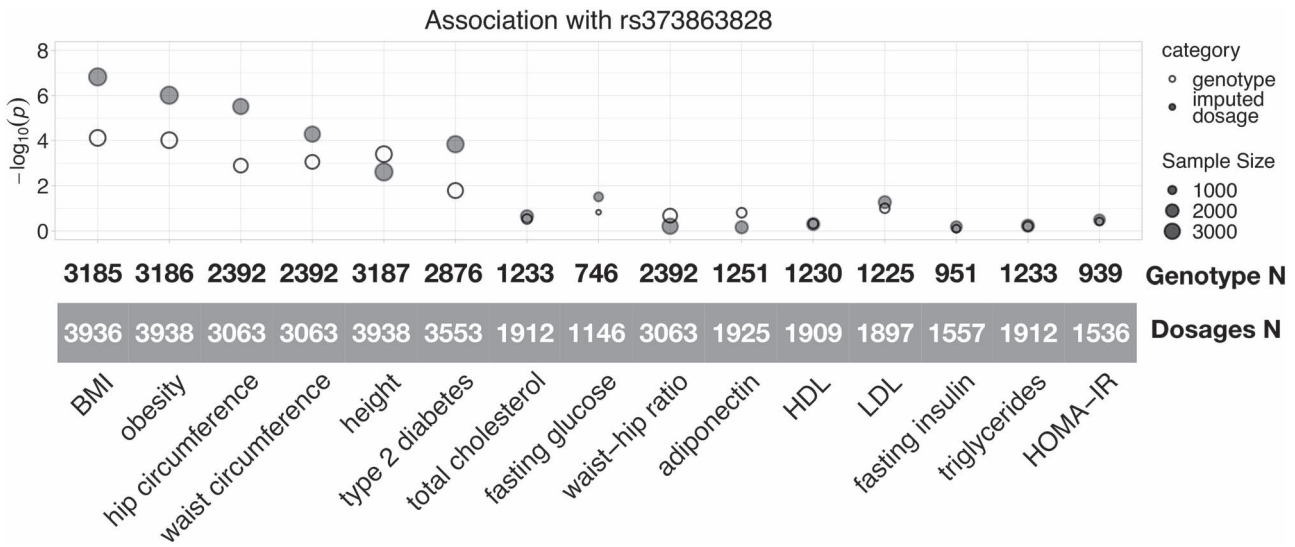

Figure 1. Association of rs373863828 with adiposity and lipid traits from PAGE cohort in Native Hawaiians. Log-transformed *P* values from association of rs373863828 with adiposity- and lipid-related traits. Associations with direct genotypes and imputed dosages are denoted in unfilled and filled circles, respectively. Point size is proportional to the sample size in each test and also shown below the trait.

The CREBRF variant is associated with adiposity traits in Native Hawaiians

The *CREBRF* variant was reported to have a large effect on body mass index (BMI) in several populations from the Pacific Islands and was significantly associated with height and other adiposity traits in Samoans (11,23). To explore its impact in Native Hawaiians, where the derived allele is present at a relatively lower frequency compared to the Samoan discovery cohorts (26%), we conducted linear mixed model association tests for the variant among genotyped individuals who also had available a collection of quantitative anthropometric, metabolic, and lipid phenotypes (Supplementary Material, Tables S2 and S3). Using a threshold of 0.0025 (Bonferroni correction for 15 phenotypes tested in (11,23) plus five cardiovascular traits below), we replicated the increasing effect of the derived allele for rs373863828 on BMI ($\beta = 0.214$ s.d. per allele, $P = 7.55e^{-5}$) and height ($\beta = 0.182$ s.d. per allele, $P = 3.96E^{-4}$). Based on these genotyped Native Hawaiians, this variant explains 0.52% and 0.37% of the phenotypic variance in BMI and height, respectively. These are much larger than the largest effect loci found in Europeans for these traits (in Europeans, rs11642015 in *FTO* and rs143384 in *GDF5* explain 0.25% and 0.2% of the variation for BMI and height, respectively, based on ~360 000 individuals studied in UK BioBank; http://www.nealelab.is/uk-biobank/). We also found the derived allele was associated with increases in waist and hip circumferences ($\beta = 0.215$ s.d. per allele and 0.206 s.d. per allele, $P = 8.7E^{-4}$ and 0.00127, respectively), but was not associated with waist–hip ratio (Supplementary Material, Table S4). Consistent with the Samoan study, we found no association between rs373863828-A and several metabolic and serum lipid traits, including fasting insulin, HOMA-IR, adiponectin, HDL, LDL, or triglycerides (Fig. 1, Supplementary Material, Table S4). In addition, we did not find an association of rs373863828 with total cholesterol or HDL, which was reported otherwise in Samoans (11).

We also examined the association for a number of dichotomous disease phenotypes. As expected for its large effect on BMI, we found the derived allele to increase the risk for obesity (OR = 1.096, $P = 9.52E^{-5}$). In addition, we nominally replicated the variant's protective effect to risk of type 2 diabetes (T2D) (OR = 0.935, $P = 0.0162$). As BMI is a major risk factor for cardiovascular diseases and the Native Hawaiians exhibit excess risk for cardiovascular diseases when compared to Europeans (24,25), we also tested the effect of the derived allele of rs373863828 on five traits among the Medicare FFS participants for the same cohort, namely, heart failure (HF), hyperlipidemia, hypertension, ischemic heart disease (IHD), and stroke/transient ischemic attack (TIA), but we found no significant associations regardless of whether or not we controlled for BMI in the analysis (Supplementary Material, Table S5).

Finally, to examine the effect of this allele on more refined measures of body fat distribution, we tested the association of the derived allele with 10 additional adiposity traits collected on a subset of 307 Native Hawaiian individuals in our study (26). Using a P value threshold of 0.0071 (Bonferroni correction for 7 effective traits among the 10 correlated traits; Materials and Methods), we found this variant was significantly associated with overall fat mass ($\beta = 0.69$ s.d. per allele, $P = 0.001$) and with whole-body fat percentage ($\beta = 0.58$ s.d. per allele, $P = 0.007$). However, the role it has on fat distribution is less clear to characterize, as the signals were nominally or near-nominally significant at best with lean mass in the arms and legs ($P = 0.056$ and 0.013, respectively) and subcutaneous fat mass ($P = 0.088$), and the variant is not associated with other adiposity traits (Table 1).

GWAS with current imputation resources or admixture mapping are unlikely to discover the CREBRF variant

While we have generally replicated the large effect rs373863828 exerts on BMI and other adiposity traits in the Native Hawaiians, approaches to discover variants like this also exemplifies one of the main goals in genetic studies of diverse populations. The derived allele of rs373863828 has large effects and is population specific, suggesting that genotype and/or ancestry at this locus is important for risk assessment in the Native Hawaiian and other Pacific Islanders. Using *CREBRF* as a test case, we thus examined whether a similar locus like this could have been discovered using currently available resources. If the entire cohort were

Table 1. Associations of rs373863828 with adiposity traits measured from DXA and abdominal MRI

Trait	Typed genotype			Imputed dosages		
	Sample size	Effect size[a]	*P* value	Sample size	Effect size[a]	*P* value
Total fat mass	294	0.69	0.001	298	0.68	0.001
Fat percentage	294	0.575	0.007	298	0.565	0.008
Trunk fat percentage	291	0.085	0.672	295	0.08	0.691
Abdominal fat	291	0.13	0.303	295	0.131	0.301
Liver fat percentage	276	0.148	0.497	280	0.159	0.465
Subcutaneous fat	278	0.349	0.088	282	0.335	0.1
Visceral fat	278	0.118	0.571	282	0.12	0.564
Lean mass in arm	283	-0.399	0.056	287	-0.396	0.057
Lean mass in leg	288	-0.507	0.013	292	-0.507	0.012
Total lean mass	292	0.233	0.233	296	0.229	0.242

[a]Effect size is based on inverse normalized transformation of original phenotypes (Materials and Methods) and are thus in units of standard deviations.

whole-genome sequenced, we estimated moderate power to identify this variant: at genome-wide significance threshold (i.e. 5×10^{-8}), we have 41% power with the current sample size with BMI ($N = 3940$). The power at the same significance threshold is much greater (75%) with the entire MEC Native Hawaiian cohort ($N \sim 5400$ individuals) (Materials and Methods).

However, it is not yet feasible to sequence the whole genome of all MEC Native Hawaiians. Therefore, statistical imputation is the most efficient strategy for gene discovery. Currently, 1000 Genomes Project (Phase 3; 1KGP) is the most diverse public sequencing database for imputation. Because rs373863828 is absent in 1KGP, it cannot be imputed directly. Yet there is the possibility to impute a proxy variant nearby that could tag the causal variant. We thus imputed the array genotype data in all Native Hawaiian samples using 1KGP data (Supplementary Material, Fig. S9) and conducted a scan for association across the *CREBRF* locus using the linear mixed model for BMI or type 2 diabetes as examples of quantitative or dichotomous trait. However, we found no association around the *CREBRF* region (+/− 100 kb) for either phenotype nearing the genome-wide significance threshold of 5×10^{-8} (lowest $P = 1.2 \times 10^{-4}$ and 2.5×10^{-3} for BMI and T2D, respectively; Supplementary Material, Fig. S3). Indeed, variants at this locus that are genotyped or well-imputed (Minimac $R^2 > 0.4$) using 1KGP are not in strong LD with rs373863828 (Supplementary Material, Fig. S4) to tag the association signal. Taken together, our results suggest that the current imputation resource is not sufficient for detecting this locus in Native Hawaiians.

There is a known proxy variant, rs12513649, with which the initial association in Samoans led the researchers to hone in on resequencing the entire *CREBRF* locus (11). This variant is in high LD with rs373863828 ($r^2 = 0.988$ in the Samoans (11); $r^2 = 0.99$ in 15 previously genotyped unadmixed Native Hawaiians that had rs12513649 data (50), for whom we genotyped rs373863828 here) and segregates in 1000 Genomes East Asians at ~6%. Thus, rs12513649 was potentially imputable, but it was poorly imputed (Minimac $R^2 = 0.35$) in our dataset and would be filtered out by standard GWAS quality control measures. Even if we were to analyze this variant, we found no genome-wide level of significance in associations between rs12513649 and BMI ($P = 0.28$) or T2D ($P = 0.04$). Upon further inspection of the imputation quality of rs12513649, we found that despite 80% of imputed genotypes at the locus have posterior genotype probability (GP) > 0.9, this proportion was driven by the homozygous ancestral genotype; the confidence of imputed genotype dropped sharply among carriers of derived allele at either the proxy (rs12513649) or the missense (rs373863828) variant (14.0% and 52.8%, respectively, Supplementary Material, Table S6).

An alternative approach to discover this locus in the Native Hawaiians would be to take advantage of the recent admixture and conduct admixture mapping. Given locally resolved assignment of ancestry across the genome in an admixed population, admixture mapping tests the association of local ancestry with a quantitative or dichotomous phenotype. There is reasonable *a priori* expectation that admixture mapping could successfully identify the *CREBRF* locus for its association with BMI or T2D, given that it is population specific, correlates strongly with Polynesian ancestry (Supplementary Material, Fig. S1), and exerts a large effect on BMI, which is differentially distributed between ancestral populations. However, we found no significant association via linear mixed models with local Polynesian ancestry across the gene ($P = 0.057$ for BMI, 0.452 for T2D, compared to a conventional statistical threshold of $5e^{-5}$ for admixture mapping). We estimated the discovery power for this locus, given the current sample size of 3940 and other assumptions of the allelic effect (Materials and Methods), to be as low as 0.2% for a $P \leq 5E^{-5}$ threshold (Fig. 2, Supplementary Material, Fig. S6). In fact, even at $P \leq 0.05$, a threshold typically used for replication, the power of replicating the *CREBRF* region via admixture mapping is only 18.2% (Supplementary Material, Fig. S7).

Taken together, without an appropriate imputation reference panel for Native Hawaiians, alternative approaches currently available could not have efficiently mapped this locus.

Imputation using internally constructed Polynesian reference boosts association signals

The obstacles described so far to conduct genetic analysis in diverse population are much attributed to the lack of representation from diverse populations in public whole genome sequences. We thus simulated a situation where a small number of individuals were available as a part of the reference panel to test if the key variant, rs373863828, could be imputed well and mapped in association studies (Supplementary Material, Fig. S8). Using as reference of Polynesian ancestry the 152 unrelated and unadmixed individuals who had rs373863828 successfully genotyped, we merged the genotype data with 1KGP to construct an imputation reference panel and imputed the genotype of rs373863828 among all remaining Native Hawaiian samples on the five different array platforms (Materials and Methods, Supplementary Material, Table S1). Regardless of the

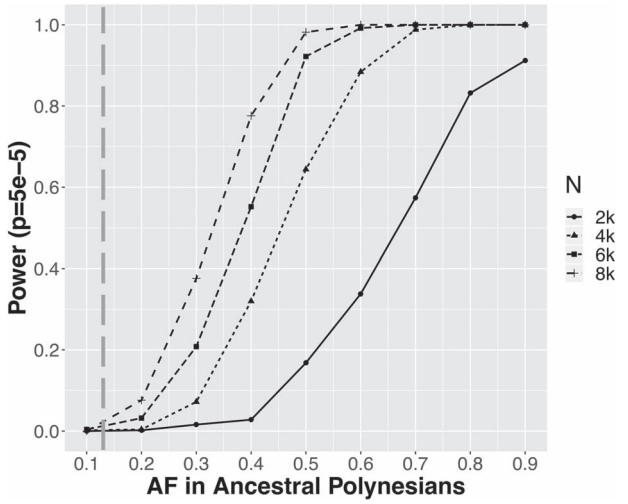

Figure 2. Statistical power of admixture mapping to discover *CREBRF* with BMI. Power was estimated through simulation given a range of allele frequencies in ancestral Polynesians and sample sizes. The effect sizes were assumed using rs373863828 as example, and the vertical dashed line denotes the empirical allele frequency of rs373863828 in unadmixed Native Hawaiians (Materials and Methods). The significance threshold for genome-wide discovery via admixture mapping is set at $5e^{-5}$.

platform a dataset was genotyped on, we found the imputed dosages to be highly correlated with the directly assayed genotypes at rs373863828: $r^2 > 0.9$ in all datasets tested, except for one dataset genotyped on the Oncoarray (0.76; Fig. 3, Supplementary Material, Table S7). Other measures of imputation quality, including concordance rate and allele frequency based internal R^2 from Minimac (27), also supported the high imputation accuracy (Supplementary Material, Table S7). While we included all relatively unadmixed individuals as the reference panel, we also found that imputation quality remained largely unchanged (Supplementary Material, Table S8) when a set of randomly selected individuals of the same size were used as reference panel, despite the allele segregating at lower frequency (Fig. 1).

We refined the associations with traits using imputed dosages of rs373863828. We found the statistical evidence for association generally improved, particularly for traits that previously showed significant or marginally significant association (Fig. 1), due to inclusion of larger sample sizes for analysis.. These were reflected mostly in adiposity phenotypes, such as BMI ($P = 1.5E^{-7}$), waist and hip circumference ($P = 5.15E^{-7}$ and $3.07E^{-6}$, respectively), and in dichotomous traits, obesity ($P = 9.7E^{-7}$). Type 2 diabetes, which previously did not survive multiple-trait testing, would now be significantly replicated ($P = 1.43e^{-4}$). Fasting glucose, which was previously insignificantly associated, would now surpass nominal significance level after imputation ($P = 0.031$, Fig. 1, Supplementary Material, Table S4). Given the statistical power of our study sample, our results suggest that rs373863828 is unlikely to exert an effect as large as it has on BMI to traits that we did not find an association (Supplementary Material, Table S9), though in some cases the upper bound of the effect size estimates are very large given the small sample sizes available (Supplemental Material, Table S10).

Discussion

We demonstrated the urgent need of having proper reference sequences in order to explore population-specific variants in diverse populations by using rs373863828 in *CREBRF* as an example in this study. We replicated the increasing effect of the derived allele of this variant on anthropometric and adiposity traits in Native Hawaiians and its protective effect on type 2 diabetes, consistent with reports in other populations from the Pacific Islands (11,15,16). When examining more refined measure of body fat distribution, we also found the derived allele to be associated with increasing total fat mass and whole-body fat percentage, even though we only have data available on ~300 Native Hawaiians. Most importantly, using the *CREBRF* locus in Native Hawaiian as an example, we have shown that even though this locus exhibits some of the largest effects on BMI observed in humans, its poor coverage in publicly available reference database due to the population specificity prevents the efficient mapping of this locus. Alternative mapping strategies such as admixture mapping also would not be powered enough to identify this locus. Then, without a specific staged study design to investigate diverse populations, we would not have been able to identify this variant that might contribute to health disparity between populations.

While our findings largely support those reported in the Samoans (11,23), we did not replicate the reported association with total cholesterol, even after imputing the variant in all individuals with phenotype available. For fasting glucose, the association was also only nominally significant after imputation. This may suggest that the allelic effect is potentially mediated by environmental factors that are found in the Samoans only, or potentially more likely, the sample sizes with these traits available in the Native Hawaiians are still insufficient.

We were also unable to detect the reported signature of natural selection at the *CREBRF* locus in Native Hawaiians. Even though we cannot rule out the lack of power since we only tested for signature of selection in 152 unadmixed individuals, this observation is consistent with the lowered derived allele frequency in Native Hawaiians (approximately 13% in relatively unadmixed Native Hawaiians vs. 26% in Samoans). While the settlement in Polynesia is believed to have occurred in a west-to-east direction across the Pacific, the derived allele is also found in lower frequency (~2 to 19%) in other populations in Pacific, including Tongans and New Zealand Maori living west of Samoa (12–15). It is unclear whether the difference of allele frequencies among the Pacific Islanders is mostly attributed to the differences in selection strength, the bottleneck and genetic drift in the founding Polynesians, different admixture histories, or some combinations of the above. A more detailed explanation will require a better construction of demographic history of the different groups.

The opposite effect on obesity and type 2 diabetes, two typically comorbid conditions, suggests that rs373863828-A could play a role in fat distribution among different body areas. One possible explanation is that the derived allele promotes accumulation of subcutaneous fat mass better than that of visceral fat, as the former can lead to obese phenotype, while excess of the latter is associated with insulin resistance and contributes to peripheral insulin sensitivity (28,29). To test this hypothesis, we examined the association with subcutaneous and visceral fat measures available in 278 Native Hawaiians, controlling for their total fat mass. We did not identify significant association with either trait, likely due to lack of power on this small subset, but the estimated effect size was larger for subcutaneous fat than for visceral fat. Further investigations are needed to explore the role of rs373863828 in general on fat deposition to better understand obesity-related metabolic disease.

Finally, of most immediate concerns is the need to construct population-specific reference panels to aid further genetic

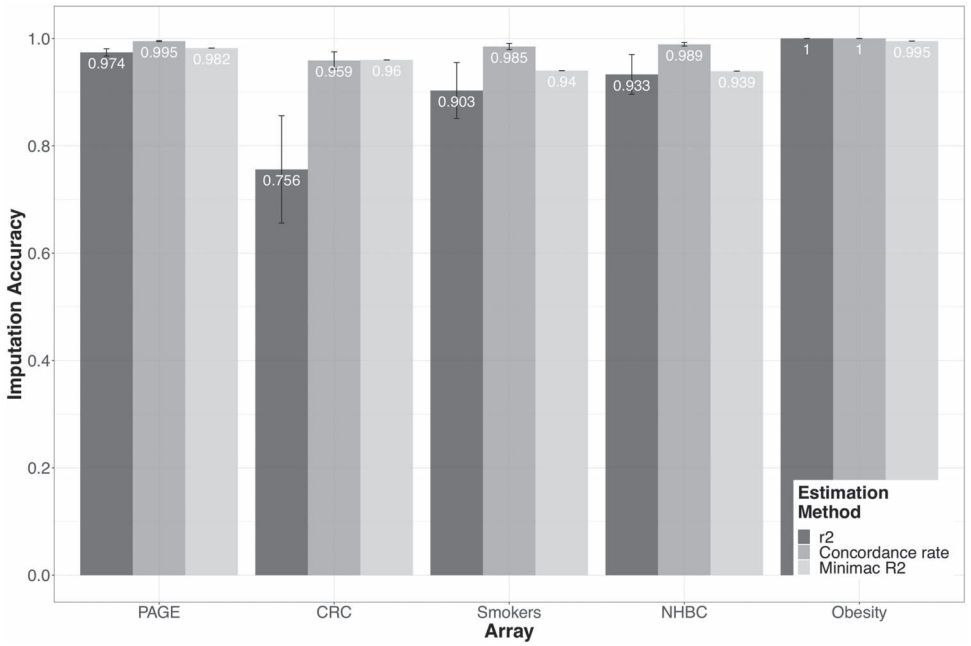

Figure 3. Assessment of imputation accuracy of rs373863828 using internally constructed references. The reference panel included 152 unadmixed Native Hawaiians, and other populations from 1000 Genomes Project (Materials and Methods). Imputation accuracy was estimated by comparing the true genotypes at rs373863828 and the imputed dosages, via different estimation methods. Standard errors were obtained through bootstrap.

analysis of these populations. Again taking the rs373863828 as an example, the variant is not found in either the 1000 Genomes Project or the Haplotype Reference Consortium, because the variant is exceedingly rare outside of Polynesia and some other Pacific Islands (the derived allele frequency = $3e^{-5}$ in gnomAD version 2.1), and neither of the reference datasets contained Polynesians. The recent release of the Human Genome Diversity Project (30,31) contained Pacific Islanders, but the variant was also not found probably due to small sample sizes ($N = 28$). A subset of the Samoan cohort was sequenced as part of the TOPMed consortium (32). Public release of this reference data for imputation will help if we can assume population continuity between Samoans, Native Hawaiians, and their most recent common ancestors, although the difference in frequency at rs373863828 between the two populations potentially suggests post-divergence drift between the two. Alternatively, another approach to overcome the issue of under-representation for investigators is to sequence a small number of individuals within the population of interest; we have shown that a sample size less than 200 is likely to be adequate for accurate imputation of this locus, echoing an earlier suggestion based on a coalescent framework that inclusion of a modestly sized internal reference panel improves imputation accuracy in the population of interest (32). Sequencing data, complemented by efforts to expand the cohort, would have the potential to detect other population-specific alleles of importance to health disparity.

In summary, there is an urgent need to increase diversity in public genome sequencing reference panels. The current deficiency could lead to adverse consequences such as failing to discover population-specific risk variants. This is particularly important for rare variants with potentially large effects, as rare variants tend to be geographically restricted and yet bear a significant role in understanding genetic architectures among populations (2,33,34). As sequencing cost continues to decline, it will be more feasible to establish diverse, population-specific references, empowering investigations for these functionally important variants like rs373863828 in the *CREBRF* gene.

Materials and Methods

Study subjects

The Multiethnic Cohort (MEC) is a population-based cohort study that examines lifestyle and genetic risk factors for cancer. It consists of 215 251 adult men and women from Hawai'i and California (primarily Los Angeles County), with ages ranging from 45 to 75 at recruitment (1993–1996). The cohort includes mainly five ethnicities: Native Hawaiians, African Americans, Japanese Americans, Latinos, and European Americans. Participants entered the cohort by completing a questionnaire on diet, height, weight, demographic information, and other risk factors. Participants are followed up with questionnaires every 5 years and linked with cancer registries annually. More details of the MEC can be found in Kolonel et al. (35) The institutional review boards of the University of Hawai'i and the University of Southern California approved the study protocol. All participants signed an informed consent form.

Genotyping and QC

In total, 4990 MEC Native Hawaiian participants were genotyped by genome-wide SNP arrays across different arrays platforms. Among them, 3940 individuals were genotyped on Illumina MEGA array as part of the PAGE consortium (2); 307 were genotyped on Illumina MEGA^{EX} as part of a collection of additional obesity-related anthropometric and imaging measurements of adiposity (26,50). An additional 266, 318, and 492 individuals were genotyped on Illumina Infinium Oncoarray, Illumina Human 1 M Duo BeadChip, and Illumina 660 W arrays, respectively, in the past for studies of colorectal cancer, nicotine metabolism, and breast cancer (36–38). In the main text and

Supplementary Material, Table S1, we refer to these past studies using MEC Native Hawaiian samples as PAGE, obesity, CRC, smokers, and NHBC substudies, respectively. In this report, we focused most of the analyses on Native Hawaiian participants in the PAGE consortium as this study had the largest sample size and SNP density, compared to other studies.

The genotyping calling process and quality control filtering for all of the genotyped datasets above are described in the corresponding references.

We applied additional uniform quality controls as follows: all variant names were updated to dbSNP v144; duplicated loci and indels were removed; triallelic variants or variants with non-matching alleles to 1000 Genomes Project phase 3 (1KGP) (39) were discarded; loci with unique positions not found in 1KGP were removed from the dataset; alleles were standardized to the positive strand by comparing to 1KGP. Finally, a genotype missingness filter of 5% was applied.

Additionally, we genotyped rs373863828 in *CREBRF* in a total of 4331 MEC Native Hawaiians using a TaqMan Assay; genotypes for 4214 of these individuals were called successfully, 3693 of which also have genome-wide array data and thus formed the dataset of this study. This variant was also genotyped in an additional 407 European Americans, 313 African Americans, 432 Japanese Americans, and 386 Latinos (both American and non-American-born Hispanics) from the MEC; the variant was monomorphic in all these other populations examined.

Phenotypes analyzed

We focused on a total of 30 quantitative and dichotomous traits related to obesity, type 2 diabetes, and cardiovascular diseases, chosen because Native Hawaiians have shown to have excess risk for these traits compared to other populations (21,24,25). These include 13 quantitative traits (BMI, hip circumference, waist circumference, adult standing height, waist–hip ratio, total cholesterol, fasting glucose, adiponectin, HDL, LDL, hypertension, fasting insulin, and HOMA-IR) and 2 dichotomous traits (obesity, type 2 diabetes). We also analyzed an additional 10 adiposity traits measured by whole-body dual-energy X-ray absorptiometry (DXA) and abdominal magnetic resonance imaging (MRI) that were available for a subset of 307 individuals (26): total fat mass, total lean mass, lean mass in leg, lean mass in arm, percentage of total fat, trunk fat mass, percentage of liver fat, visceral fat area, and abdominal fat area. Finally, we examined five additional dichotomous disease traits related to cardiovascular outcomes in MEC participants enrolled in the Medicare fee-for-service program including: heart failure (HF), hyperlipidemia (HYPERL), hypertension (HYPERT), ischemic heart disease (IHD), and stroke/transient ischemic attack (TIA). These phenotypes were identified by CMS Chronic Conditions Data Warehouse using algorithms that search the Medicare claims data for specific diagnosis or procedure codes (https://www2.ccwdata.org/web/guest/condition-categories). Please refer to Supplementary Material, Table S2 for details of the phenotype transformation.

Imputation of rs373863828

In order to impute rs373863828 using a population-specific reference panel, we identify MEC Native Hawaiian individuals with the highest amount of Polynesian ancestry to serve as the Polynesian reference panel. Specifically, we first estimated the global ancestry proportions of 3940 subjects from PAGE, using ADMIXTURE v1.3 (40), which adopts a block relaxation method to update ancestral allele frequency and ancestry fraction. We modeled the Native Hawaiian ancestry as four ancestral components: the majority Polynesian component, with recent admixtures from Europeans, East Asians, and Africans. We thus included all 1000 Genomes Project (39) (1KGP; Phase 3), together with other three MEC ethnicities, including 3465 Japanese, 30 Hispanic/Latinos, and 5325 African Americans, which were genotyped on the same MEGA array to better infer ancestry components. After pruning the dataset to exclude loci with genotype missingness $> 5\%$ and minor allele frequency $< 1\%$, we stratified individuals into a group of closely related individuals (first-degree or second-degree relatives) estimated from KING (41) and a separate group of relatively unrelated individuals. We then pruned variants such that no two variants have an LD above r^2 of 0.1, per recommendation of ADMIXTURE, and conducted an unsupervised run across the unrelated individuals at $k = 4$. We then projected the estimated ancestral allele frequency to the related samples to infer the genomic ancestries of these individuals. We performed five independent iterations with randomized seed numbers and found no minor mode at $k = 4$. The final result was integrated by averaging the estimated proportions after matching ancestry clusters across five runs. The European, East Asian, and African ancestral components could easily be defined with their corresponding representation from other reference populations in the panel. We treated the remaining component, found predominately in our cohort of Native Hawaiians, as the Hawaiian-specific indigenous component that we presumed to be 'Polynesian' in origin. From this analysis we identified 178 MEC Native Hawaiian individuals with $> 90\%$ Polynesian ancestry and have a refined kinship coefficient (through PC-relate (42), below) < 0.2 among reference samples. We further subset to 152 individuals who were genotyped successfully at rs373863828; we refer to this group of individual as 'unadmixed' throughout the manuscript for brevity. To construct the imputation reference panel, the genotype of the rs373863828 was merged to the array genotypes on chromosome 5 for these 152 individuals and merged with the 1KGP, with 1KGP sample carrying homozygous reference alleles at rs373863828. We statistically phased this constructed reference panel again with EAGLE2. We used Minimac3 (27) (version 2.0.1) and the constructed reference panel to impute the genotype of rs373863828 in the rest of Native Hawaiian individuals with genome-wide array data. After imputation, we had genotype dosages of rs373863828 available for all samples, covering those phenotyped but not particularly genotyped by TaqMan assay, which increased the sample size in the refined associations of this locus with all traits available.

Constructing the genetic relatedness matrix

Due to sample relatedness and population substructure within the MEC Native Hawaiians, standard approaches for constructing principal components and kinship estimates could each bias one another. Thus, we used PC-air and PC-relate from GENESIS v2.4.0 package (42,43), which performs a principal component analysis robust to family structures and infers genetic relatedness unbiased from unspecified population structures, respectively. Based on our initial kinship estimates from KING (41), we obtained the top 10 eigenvectors reflecting ancestry influences at the default unrelatedness cutoff of $2^{\wedge}(-11/2)$. The unbiased eigenvectors were in turn used to refine the kinship coefficients in the genetic relatedness matrix (GRM).

Linear mixed model

We used a linear or logistic mixed model (LMM) implemented in EMMAX (44) to perform all association tests in this study. We used the GRM generated from PC-relate (above) as a random effect in the model and the inverse normalized residuals and covariates as fixed effects for each trait (Supplementary Material, Table S2). In association tests of the *CREBRF* region on chromosome 5, only 15 334 markers that were genotyped or had Minimac internal R^2 score > 0.4 were included. In admixture mapping with BMI and T2D, genotyped positions with probabilities of being on Native Hawaiian haplotypes, as inferred from local ancestry inference (below), were used as dosages. Bonferroni correction was directly applied to associations conducted in PAGE subjects. For the additional 10 adiposity traits in the separate 307 individuals, due to the likely correlations among the traits, we determined the number of independent tests as following: we decomposed the phenotypic matrix by principle component analysis and calculated the accumulative variance explained by eigenvectors until it surpasses 95% of that of the phenotypic matrix. We found the corresponding number of PCs surpassed the threshold to be 7, suggesting a significance threshold of 0.05/7 = 0.0071.

Local ancestry inference

We first phased all MEC Native Hawaiians from the PAGE dataset against 1000 Genomes Project Phase 3, using Sanger Imputation Service (https://imputation.sanger.ac.uk/). We then merged these samples with the subset of the phased 1KGP that served as reference, namely, GBR, CEU, TSI, and IBS as European references; CHB, JPT, CHS, CDX, and KHV as East Asian references; and YRI and LWK as African references, and phased them together again, as rephasing was shown to improve imputation accuracy(6), and we found that rephasing improved genetic ancestry inference (Supplementary Material, Fig. S10). We then applied RFMix2 (v2.03-r0) (45)on the rephased data, which adopts a discriminative modeling approach using a conditional random field to call local ancestry based on genome-wide data. We used our constructed imputation reference panel as reference panel of the four components of ancestry. HapMap2 (46) pooled recombination rate (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20110106_recombination_hotspots/) was used as the genetic map. We adopted the default parameters of RFMix2 as we found no notable difference when enabling expectation–maximization for five iterations or when enabling re-analysis of reference individuals (data not shown). From the output of RFMix tsv file, we also computed global ancestry estimates for each person after we excluded tracts with any ancestry probability lower than 0.9. The global ancestry proportions estimated from RFMix is highly concordant with those from ADMIXTURE (Supplementary Material, Fig. S10); the main deviation is due to individuals detectably related to the 178 reference individuals (maximal kinship coefficients with the 178 internal individuals are significantly higher than the rest; Mann–Whitney (one way) $P = 5.14E^{-58}$).

Power estimate of single variant association

We estimated the power of discovering the locus rs373863828 via single variant association (at alpha level = $5E^{-8}$), as well as the effect size per allele that would be detectable with 80% power in this study (at alpha level = $2.5E^{-3}$ for traits tested in Fig. 1; alpha level = $7.1E^{-3}$ for the DXA and MRI adiposity traits). We followed a standard power estimate for quantitative traits: assuming a chi-square model with degree of freedom as 1 applies, the power equals to the left tail probability of a chi-square value of corresponding alpha probability (5×10^{-8} as the genome-wide significance threshold), but with the non-centrality parameter shifted as the product of sample size and heritability explained by the single locus. For rs373863828,

$$h_{SNP}{}^2 = 2p\ (1-p)\,\beta_{INT}{}^2$$

where p is the MAF (6%) and β_{INT} is the effect size in standard deviation unit from this study (e.g. 0.248 for BMI).

For dichotomous trait, we computed power using the GAS power calculator (47) (http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/). We used prevalence estimates in Native Hawaiians obtained from Hawai'i indicator-based information system (Hawaii-IBIS) (Retrieved Sat, 14 March 2020 from Hawai'i State Department of Health: http://ibis.hhdw.org/ibisph-view/), Hawai'i Health Data Warehouse (http://hhdw.org/wp-content/uploads/HHS_Hypertension_IND_00001_2010.pdf), and from literature (Juarez et al. (48) for hyperlipidemia).

Power simulation for admixture mapping

We assumed the following in our power simulation: (1) the derived allele frequency of rs373863828 in the ancestral, unadmixed, Native Hawaiians is 13%, as estimated from current individuals who have > 90% Polynesian ancestry; (2) the effect size of the derived allele and the phenotypic standard error are transformed to the same units as reported in the discovery cohort in Samoans by Minster et al.(11), i.e. 1.36 and 6.9 kg/m^2; (3) the percentage of local ancestry at *CREBRF* region among all samples is similar to the average of global ancestry proportions (i.e. no strong selection at the locus) (Supplementary Material, Fig. S5).

Given a target sample size, we assigned the genotype of each individual assuming Hardy–Weinberg equilibrium and the derived allele frequency of 5.9% (matching the frequency for rs373863828 in Native Hawaiians). Based on the genotype, we then assigned the ancestral origin of haplotype for each individual with the following rule:

1) Homozygous-derived genotype implies the individual derived both alleles from Polynesian haplotypes.
2) Heterozygous genotype implies the individual carries at least one copy of Polynesian haplotype. For the other, non-derived, allele, the probability that this allele derives from a Polynesian haplotype is:

$$P\left(Polynesian\ origin \mid non-derived\ allele\right) = \frac{f_P - f_P^* p_P}{1 - f_P^* p_P}$$

where $f_P$is the global Polynesian ancestry percentage and $p_P$is the derived allele frequency in non-admixed ancestral Polynesians (13%). Thus, the probability that this individual carries exactly one copy of Polynesian haplotype is$1 - P\left(Polynesian\ origin \mid non-derived\ allele\right)$.

3) Homozygous ancestral genotype (GG): the number of Polynesian tracts each individual carries corresponds to a binomial sampling at the probability of P (Polynesian origin | non-derived allele) as described, and the number of trials is 2 (diploid).

To simulate phenotypes of individuals given their genotypes, we sampled from a normal distribution with mean shifted by 2β,

β, and 0 standard deviations, where β is the reported effect size. Given simulated genotype, local ancestry, and phenotype, we tested the association between local ancestry and the phenotype to simulate admixture mapping. Power was calculated as the number of times a positive association at or below the specified statistical threshold was achieved in 500 iterations.

Selection test

We performed nS_L scan (22) implemented in Selscan (49), which identifies ongoing positive selection in genome based on phased haplotypes and is robust to recombination rate variation, on the 152 internal reference individuals at the *CREBRF* locus. To standardize the nSL score at the *CREBRF* locus, we constructed the null distribution based on genome-wide nSL scores from variants with derived allele frequency between 12 and 14%.

Supplementary Material

Supplementary Material is available at HMG online

Acknowledgements

We would like to thank all Native Hawaiian participants in the Multiethnic Cohort that are involved in this study. We would also like to thank Ryan L. Minster at University of Pittsburgh and Stephen T. McGarvey at Brown University for their comments on the manuscript, and University of Hawai'i Cancer Center's Native Hawaiian Community Advisory Board for reviewing the study proposal.

Conflict of Interest

Authors declare that they have no competing interest.

Funding

This study was funded by National Cancer Institute (U01CA164973, P01CA168530), National Human Genome Research Institute (U01HG007397), and University of Hawai'i Cancer Center Support Grant for Genomics and Bioinformatics Shared Resource (P30CA071789) Computation for this project was supported by the University of Southern California's Center for High-Performance Computing (https://hpcc.usc.edu).

References

1. Marchini, J. and Howie, B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, **11**, 499.
2. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L. *et al.* (2019) Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, **570**, 514–518.
3. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D. and Kenny, E.E. (2017) Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.*, **100**, 635–649.
4. Popejoy, A.B. and Fullerton, S.M. (2016) Genomics is failing on diversity. *Nat. News*, **538**, 161.
5. Bustamante, C.D., Vega, F.M.D.L. and Burchard, E.G. (2011) Genomics for the world. *Nature*, **475**, 163.
6. Consortium, H.R., McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P. *et al.* (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, ng.3643.
7. Hindorff, L.A., Bonham, V.L., Brody, L.C., Ginoza, M.E.C., Hutter, C.M., Manolio, T.A. and Green, E.D. (2017) Prioritizing diversity in human genomics research. *Nat. Rev. Genet.*, **19**, 175.
8. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M. and Daly, M.J. (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.*, **51**, 584–591.
9. Sirugo, G., Williams, S.M. and Tishkoff, S.A. (2019) The missing diversity in human genetic studies. *Cell*, **177**, 1080.
10. Evangelou, E., Warren, H.R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., Ntritsos, G., Dimou, N., Cabrera, C.P., Karaman, I. *et al.* (2018) Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.*, **50**, 1412–1425.
11. Minster, R.L., Hawley, N.L., Su, C.-T., Sun, G., Kershaw, E.E., Cheng, H., Buhule, O.D., Lin, J., Reupena, M.S., Viali, S. *et al.* (2016) A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nat. Genet.*, **48**, ng.3620.
12. Naka, I., Furusawa, T., Kimura, R., Natsuhara, K., Yamauchi, T., Nakazawa, M., Ataka, Y., Ishida, T., Inaoka, T., Matsumura, Y. *et al.* (2017) A missense variant, rs373863828-A (p.Arg457Gln), of CREBRF and body mass index in oceanic populations. *J. Hum. Genet.*, **62**, 847–849.
13. Ohashi, J., Naka, I., Furusawa, T., Kimura, R., Natsuhara, K., Yamauchi, T., Nakazawa, M., Ishida, T., Inaoka, T., Matsumura, Y. *et al.* (2018) Association study of CREBRF missense variant (rs373863828:G > A; p.Arg457Gln) with levels of serum lipid profile in the Pacific populations. *Ann. Hum. Biol.*, **45**, 215–219.
14. Berry, S.D., Walker, C.G., Ly, K., Snell, R.G., Carr, P.E.A., Bandara, D., Mohal, J., Castro, T.G., Marks, E.J., Morton, S.M.B. *et al.* (2017) Widespread prevalence of a CREBRF variant amongst Māori and Pacific children is associated with weight and height in early childhood. *Int. J. Obesity*, **42**, 603.
15. Krishnan, M., Major, T.J., Topless, R.K., Dewes, O., Yu, L., Thompson, J.M.D., McCowan, L., de Zoysa, J., Stamp, L.K., Dalbeth, N. *et al.* (2018) Discordant association of the CREBRF rs373863828 A allele with increased BMI and protection from type 2 diabetes in Māori and Pacific (Polynesian) people living in Aotearoa/New Zealand. *Diabetologia*, **61**, 1603–1613.
16. Hanson, R.L., Safabakhsh, S., Curtis, J.M., Hsueh, W.-C., Jones, L.I., Aflague, T.F., Sarmiento, J.D., Kumar, S., Blackburn, N.B., Curran, J.E. *et al.* (2019) Association of CREBRF variants with obesity and diabetes in Pacific islanders from Guam and Saipan. *Diabetologia*, **62**, 1647–1652.
17. Kim, S.K., Gignoux, C.R., Wall, J.D., Lum-Jones, A., Wang, H., Haiman, C.A., Chen, G.K., Henderson, B.E., Kolonel, L.N., Marchand, L.L. *et al.* (2012) Population genetic structure and origins of native Hawaiians in the multiethnic cohort study. *PLoS One*, **7**, e47881.
18. Nordyke, E. C. (1989) The peopling of Hawai'i. University of Hawai'i Press, Honolulu, HI.
19. Burney, D.A., James, H.F., Burney, L.P., Olson, S.L., Kikuchi, W., Wagner, W.L., Burney, M., McCloskey, D., Kikuchi, D., Grady, F.V. *et al.* (2001) Fossil evidence for a diverse biota from Kaua 'i and its transformation since human arrival. *Ecol. Monogr.*, **71**, 615–641.
20. Wilmshurst, J.M., Hunt, T.L., Lipo, C.P. and Anderson, A.J. (2011) High-precision radiocarbon dating shows recent and

rapid initial human colonization of east Polynesia. *Proc. Natl Acad. Sci.*, **108**, 1815–1820.
21. Maskarinec, G., Grandinetti, A., Matsuura, G., Sharma, S., Mau, M., Henderson, B.E. and Kolonel, L.N. (2009) Diabetes prevalence and body mass index differ by ethnicity: the multiethnic cohort. *Ethnic. Dis.*, **19**, 49–55.
22. Ferrer-Admetlla, A., Liang, M., Korneliussen, T. and Nielsen, R. (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.*, **31**, 1275–1291.
23. Carlson, J.C., Rosenthal, S.L., Russell, E.M., Hawley, N.L., Sun, G., Cheng, H., Naseri, T., Reupena, M.S., Tuitele, J., Deka, R. *et al.* (2020) A missense variant in CREBRF is associated with taller stature in Samoans. *Am J Hum Biol*, e23414.
24. Tung, W.-C. and Barnes, M. (2014) Heart diseases among native Hawaiians and Pacific islanders. *Home Heal Care Manage. Pract.*, **26**, 110–113.
25. Grandinetti, A., Chen, R., Kaholokula, J.K., Yano, K., Rodriguez, B.L., Chang, H.K. and Curb, J.D. (2002) Relationship of blood pressure with degree of Hawaiian ancestry. *Ethnic. Dis.*, **12**, 221–228.
26. Lim, U., Monroe, K.R., Buchthal, S., Fan, B., Cheng, I., Kristal, B.S., Lampe, J.W., Hullar, M.A., Franke, A.A., Stram, D.O. *et al.* (2018) Propensity for intra-abdominal and hepatic adiposity varies among ethnic groups. *Gastroenterology*, **156**, 966–975.e10.
27. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M. *et al.* (2016) Next-generation genotype imputation service and methods. *Nat. Genet.*, **48**, 1284–1287.
28. Gastaldelli, A., Miyazaki, Y., Pettiti, M., Matsuda, M., Mahankali, S., Santini, E., DeFronzo, R.A. and Ferrannini, E. (2002) Metabolic effects of visceral fat accumulation in type 2 diabetes. *J. Clin. Endocrinol. Metabolism*, **87**, 5098–5103.
29. Neeland, I.J., Turer, A.T., Ayers, C.R., Powell-Wiley, T.M., Vega, G.L., Farzaneh-Far, R., Grundy, S.M., Khera, A., McGuire, D.K. and de Lemos, J.A. (2012) Dysfunctional adiposity and the risk of Prediabetes and type 2 diabetes in obese adults. *JAMA*, **308**, 1150–1159.
30. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J. *et al.* (2019) Insights into human genetic variation and population history from 929 diverse genomes. *Science*, **367**, eaay5012.
31. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A. *et al.* (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.
32. Jewett, E.M., Zawistowski, M., Rosenberg, N.A. and Zöllner, S. (2012) A coalescent model for genotype imputation. *Genetics*, **191**, 1239–1255.
33. Mathieson, I. and McVean, G. (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.*, **44**, 243.
34. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D., Altshuler, D.L. *et al.* (2011) Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci.*, **108**, 11983–11988.
35. Kolonel, L.N., Henderson, B.E., Hankin, J.H., Nomura, A.M.Y., Wilkens, L.R., Pike, M.C., Stram, D.O., Monroe, K.R., Earle, M.E. and Nagamine, F.S. (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.*, **151**, 346–357.
36. Siddiq, A., Couch, F.J., Chen, G.K., Lindström, S., Eccles, D., Millikan, R.C., Michailidou, K., Stram, D.O., Beckmann, L., Rhie, S.K. *et al.* (2012) A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum. Mol. Genet.*, **21**, 5373–5384.
37. Wang, H., Burnett, T., Kono, S., Haiman, C.A., Iwasaki, M., Wilkens, L.R., Loo, L.W.M., Berg, D.V.D., Kolonel, L.N., Henderson, B.E. *et al.* (2014) Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat. Commun.*, **5**, 4613.
38. Patel, Y.M., Park, S.L., Han, Y., Wilkens, L.R., Bickeboller, H., Rosenberger, A., Caporaso, N. and L, i, M. T., Bruske, I., et al. (2016) Novel association of genetic markers affecting CYP2A6 activity and lung cancer risk. *Cancer Res.*, **76**, 5768–5776.
39. Consortium, 1000 Genomes Project, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68.
40. Alexander, D.H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–1664.
41. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. and Chen, W.-M. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867–2873.
42. Conomos, M.P., Reiner, A.P., Weir, B.S. and Thornton, T.A. (2016) Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.*, **98**, 127–148.
43. Conomos, M.P., Miller, M.B. and Thornton, T.A. (2015) Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.*, **39**, 276–293.
44. Kang, H. M., Sul, J. H., Service S.K., Zaitlen, N. A., Kong, S., Freimer, N. B., Sabatti, C. and Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348.
45. Maples, B.K., Gravel, S., Kenny, E.E. and Bustamante, C.D. (2013) RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.*, **93**, 278–288.
46. Consortium, T. I. H (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851.
47. Skol, A.D., Scott, L.J., Abecasis, G.R. and Boehnke, M. (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.*, **38**, 209–213.
48. Juarez, D.T., Davis, J.W., Brady, S.K. and Chung, R.S. (2012) Prevalence of heart disease and its risk factors related to age in Asians, Pacific islanders, and whites in Hawai'i. *J. Health Care Poor Underserved*, **23**, 1000–1010.
49. Szpiech, Z.A. and Hernandez, R.D. (2014) Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.*, **31**, 2824–2827.
50. Park, SL, Li, Y, Sheng, X, Xia, L, Zhao, K, Pooler, L, Setiawan, VW, Lim, U, Monroe, KR, Wilkens, LR, Kristal, BS, Lampe, JW, Hullar, MA, Shepherd, J, Loo, L, Ernst, T, Franke, AA, Tiirikainen, M, Haiman, CA, Stram, DO, Marchand, LL and Cheng, I () Genome-Wide Association Study of Liver Fat: The Multiethnic Cohort Adiposity Phenotype Study. *Hepatology Communications*, 2020. In Press.