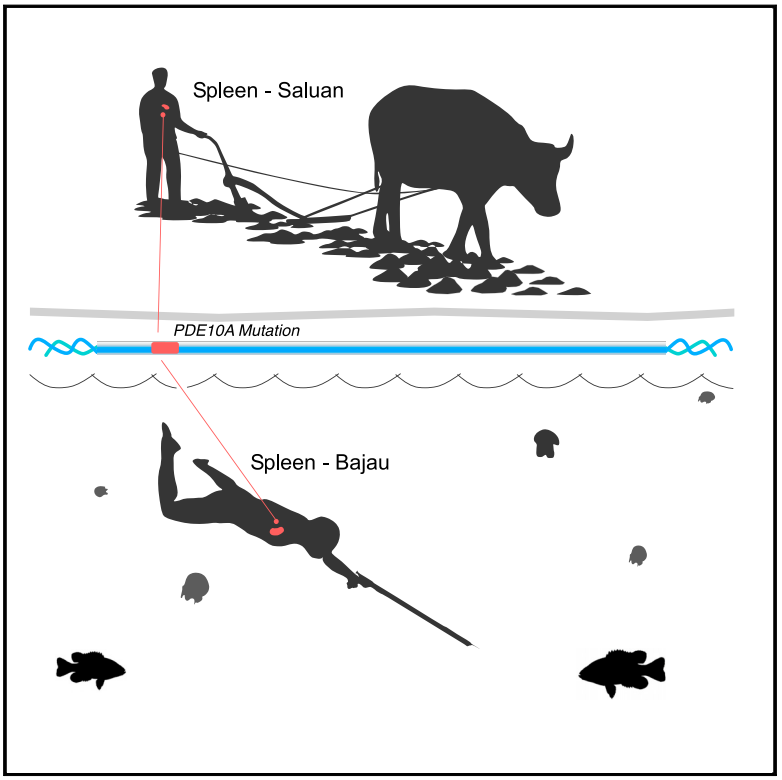
Article

Cell

Physiological and Genetic Adaptations to Diving in Sea Nomads

Graphical Abstract



Highlights

- The Bajau, or "Sea Nomads," have engaged in breath-hold diving for thousands of years
- Selection has increased Bajau spleen size, providing an oxygen reservoir for diving
- We find evidence of additional diving-related phenotypes under selection
- These findings have implications for hypoxia research, a pertinent medical issue

Authors

Melissa A. Ilardo, Ida Moltke, Thorfinn S. Korneliussen, ..., Suhartini Salingkat, Rasmus Nielsen, Eske Willerslev

Correspondence

rasmus_nielsen@berkeley.edu (R.N.), ewillerslev@snm.ku.dk (E.W.)

In Brief

Genetic and physiological adaptations enable the remarkable breath-holding ability of marine nomads.

Ilardo et al., 2018, Cell *173*, 569–580
April 19, 2018
https://doi.org/10.1016/j.cell.2018.03.054

CellPress

Cell

Physiological and Genetic Adaptations to Diving in Sea Nomads

Melissa A. Ilardo,[1] Ida Moltke,[2] Thorfinn S. Korneliussen,[1,3] Jade Cheng,[4] Aaron J. Stern,[4,5] Fernando Racimo,[1] Peter de Barros Damgaard,[1] Martin Sikora,[1] Andaine Seguin-Orlando,[1,6] Simon Rasmussen,[7] Inge C.L. van den Munckhof,[8] Rob ter Horst,[8] Leo A.B. Joosten,[8] Mihai G. Netea,[8,9] Suhartini Salingkat,[10] Rasmus Nielsen,[1,4,12,*] and Eske Willerslev[1,3,11,*]

[1]Centre for GeoGenetics, University of Copenhagen, Copenhagen 1350, Denmark
[2]Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark
[3]Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK
[4]Department of Integrative Biology, University of California at Berkeley, Berkeley, CA 94720, USA
[5]Department of Computational Biology, University of California at Berkeley, Berkeley, CA 94720, USA
[6]Danish National High-throughput DNA Sequencing Centre, University of Copenhagen 1353, Denmark
[7]Bioinformatics, Technical University of Denmark, Lyngby 2800, Denmark
[8]Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen 6525, the Netherlands
[9]Department for Genomics and Immunoregulation, Life and Medical Sciences Institute (LIMES), University of Bonn, Bonn 53115, Germany
[10]Tompotika Luwuk Banggai, Tompotika University, Luwuk 94711, Indonesia
[11]Wellcome Trust, Sanger Institute, Hinxton CB10 1SA, UK
[12]Lead Contact
*Correspondence: rasmus_nielsen@berkeley.edu (R.N.), ewillerslev@snm.ku.dk (E.W.)
https://doi.org/10.1016/j.cell.2018.03.054

SUMMARY

Understanding the physiology and genetics of human hypoxia tolerance has important medical implications, but this phenomenon has thus far only been investigated in high-altitude human populations. Another system, yet to be explored, is humans who engage in breath-hold diving. The indigenous Bajau people ("Sea Nomads") of Southeast Asia live a subsistence lifestyle based on breath-hold diving and are renowned for their extraordinary breath-holding abilities. However, it is unknown whether this has a genetic basis. Using a comparative genomic study, we show that natural selection on genetic variants in the PDE10A gene have increased spleen size in the Bajau, providing them with a larger reservoir of oxygenated red blood cells. We also find evidence of strong selection specific to the Bajau on BDKRB2, a gene affecting the human diving reflex. Thus, the Bajau, and possibly other diving populations, provide a new opportunity to study human adaptation to hypoxia tolerance.

INTRODUCTION

Humans are the only mammals to have colonized all of Earth's most extreme environments, from high altitude mountain chains to the remote islands of the Pacific. Human phenotypic adaptations to extreme environments have been the subject of much research (Beall, 2006; Yi et al., 2010), in part because locally adapted populations provide an opportunity to study the genetic and physiological consequences of environmental perturbations. For example, research on adaptations in the people of Tibet (Beall et al., 2010; Peng et al., 2011; Simonson et al., 2010; Wuren et al., 2014; Xiang et al., 2013; Xu et al., 2011; Yang et al., 2017; Yi et al., 2010) and other high altitude populations (Beall, 2006) has revealed new insight into the physiology of hypoxia with a broad range of implications in medically relevant fields (Grocott et al., 2007; Oosthuyse et al., 2001; Rankin and Giaccia, 2008; Talks et al., 2000; Zhong et al., 1999), including intensive care treatment (McKenna and Martin, 2016) and tumorigenesis (Rankin and Giaccia, 2008). Another possible system of human adaptation to extreme environments with implications for hypoxia research is that of humans who engage in breath-hold diving.

The Bajau people, often referred to as Sea Nomads, have lived an entirely marine-dependent existence, traveling the Southeast Asian seas on houseboats for over 1,000 years (Sather, 1997). Their marine hunter-gatherer existence depends notably on the food they collect through free diving. They are renowned for their extraordinary abilities, diving to depths of over 70 m with nothing more than a set of weights and a pair of wooden goggles (Schagatay, 2014) and spending 60% of their daily working time underwater (Schagatay et al., 2011). The unique lifestyle of the Bajau relies on a number of cultural traits and technical innovations, but may also be facilitated by physiological adaptations to diving and diving-induced hypoxia (Clifton and Majors, 2012; Sopher, 1965). Humans, like other diving mammals, have a diving response induced by apnea and cold-water facial immersion (Thornton and Hochachka, 2004; Sterba and Lundgren, 1988). Physiological effects of this response include bradycardia, which lowers oxygen consumption (Elsner et al., 1966; Ferrigno et al., 1997; Kooyman and Campbell, 1972; Lin et al., 1972, 1983); peripheral

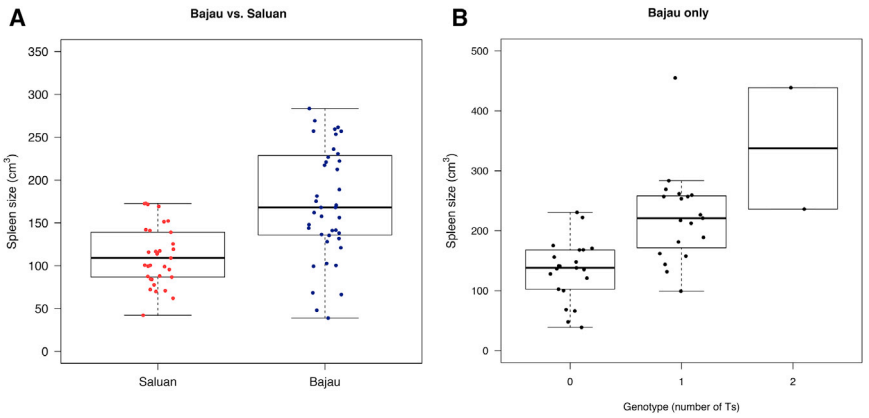


Figure 1. Bajau versus Saluan Spleen Size Distributions and Bajau Spleen Sizes Stratified by Imputed Genotypes of the SNP rs3008052

(A) Distributions and boxplots of spleen sizes in all not closely related Saluan (n = 33) in red and Bajau (n = 43) in blue. The thick black center lines represent the medians and the lower and upper limits of the boxes represent the 25th and 75th percentiles, respectively. Outliers are not shown.

(B) Untransformed spleen sizes stratified by imputed genotypes of rs3008052 for Bajau individuals only (n = 43).

vasoconstriction, which selectively redistributes blood flow to the organs most sensitive to hypoxia (Lin et al., 1983; Zapol et al., 1979); and contraction of the spleen, which injects a supply of oxygenated red blood cells into the circulatory system (Hurford et al., 1996; Stewart and McKenzie, 2002).

Splenic contraction as a component of the human diving response was first observed in the Ama, a group of Japanese pearl divers (Hurford et al., 1990) and is induced by a catecholamine-mediated alpha-2 adrenoreceptor response (Foster and Sheel, 2005). A single contraction expels ~160 mLa of red blood cells, causing a hemoglobin increase that corresponds to a 2.8%–9.6% increase in oxygen content (Stewart and McKenzie, 2002). It has therefore been hypothesized that the purpose of this contraction is to provide an oxygen boost, prolonging dive time (Hochachka, 1986). In a study of diving seal species, a positive correlation was observed between maximum dive time and spleen mass (Mottishaw et al., 1999), suggesting that spleen size could be an important trait affecting diving time. However, the relationship between spleen size and dive capacity has never before been examined at the genetic level. In fact, very little is known about the genetic basis of the diving response in humans: only one study has ever claimed to show a genetic variant that directly influences the dive response (Baranova et al., 2017). Bradykinin receptor B2 (BDKRB2), a signal peptide associated with both vasodilation and vasoconstriction, was suggested to affect peripheral vasoconstriction induced by the diving response in this study (Baranova et al., 2017).

It is entirely unknown whether the Sea Nomads are genetically adapted to their extreme lifestyle. The only trait that has been investigated in populations with a lifestyle dependent on diving is the superior underwater vision of Thai Sea Nomad children (Gislén et al., 2003). However, this was later shown to be a plastic response to training via repeated diving, replicable in a European cohort (Gislén et al., 2006). Here, we used a two-pronged approach to address the question of potential genetic adaptations in the Bajau. First, we performed a scan of their genomes for signatures of selection to identify genes that have been uniquely targeted by natural selection in the Bajau. Second, we examined if any of the candidate loci are associated with spleen size, one of the most relevant candidate traits for adaptation to free diving and hypoxia tolerance.

RESULTS

Spleen Size Difference in the Bajau

We first set out to identify if there is evidence that the Bajau have larger spleens than their close geographic neighbors, the Saluan, who interact minimally with the marine environment. We identified two seaside villages ~25 km apart in the Central Sulawesi peninsula of Indonesia: Jaya Bakti and Koyoan, primarily inhabited by ethnic Bajau and Saluan populations, respectively. We recruited 59 Bajau and 34 Saluan individuals to participate in the study, and from each individual we collected saliva samples for DNA and spleen measurements using a portable ultrasound machine. 16 Bajau individuals and 1 Saluan were found to be closely related to others from their respective communities based on genetic data. These individuals were excluded from all further analyses as most are based on the assumption that the individuals analyzed are not closely related. We made ultrasound measurements in two planes such that we were able to calculate spleen volumes according to the methodology outlined in Yetter et al. (2003) that best correlates with volumes obtained using a computed tomography (CT) scan. We used these measurements to compare spleen sizes in the two populations, revealing a clear visual difference, with the mean spleen size being higher among the Bajau (Figure 1). This difference was statistically significant (Welch two-sample t test, $p = 3.538e{-}07$). Notably, this difference is not significant when comparing Bajau divers to Bajau non-divers ($p = 0.2663$), suggesting the difference between the Bajau and Saluan is not simply driven by the fact that more Bajau individuals are divers. However, factors other than whether the individuals are divers may affect the results of the test (see STAR Methods and Figure S1 for details). We therefore also tested for a difference in spleen size between Bajau and Saluan using a linear model that allowed us to take additional factors into account. Specifically, we included gender, age, weight, height, and whether the individuals are divers as covariates. The results of this test also indicated that Bajau have significantly larger spleens than the Saluan, even when correcting for several potentially confounding factors ($p = 0.0438$, $\beta = 44.40$, SE 21.62, see STAR Methods for details). These results suggest a physiological difference between the Bajau and the Saluan that is not solely attributable to a plastic response of the spleen to diving

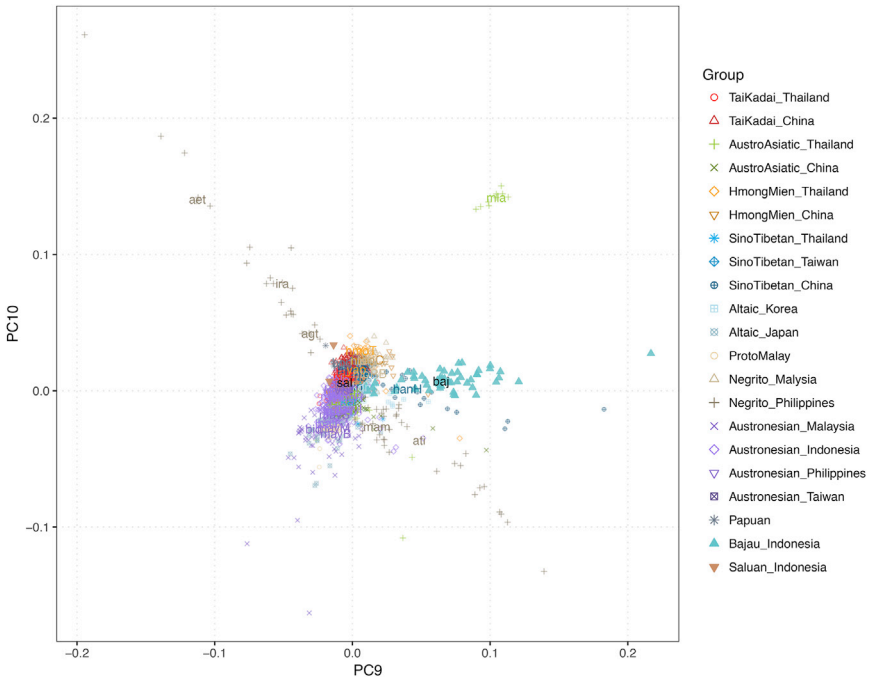


Figure 2. PCA Showing the Genetic Relationship of the Bajau and Saluan to Other Pan-Asia Populations

The 43 Bajau and 33 Saluan individuals appear to be most closely related to each other (see Figure S5 for additional PCs). In general, they cluster with other Austronesian people in the region but have some affinity to Melanesia on a gradient related to geographic proximity (see also Lipson et al., 2014).

See also Figure S3.

activity. While other unknown environmental factors could potentially explain the observed difference between the groups, genetic factors remain a possibility.

Population Demographic Analyses

Next, we generated low-depth whole-genome sequencing data for the same Bajau and Saluan individuals and analyzed the data using methods that take genotype uncertainty into account by working directly on genotype likelihoods (Cheng et al., 2016; Korneliussen et al., 2014; Skotte et al., 2012). For the purpose of population genetics analyses, we merged this data with data from the Pan-Asian genome project (Ngamphiw et al., 2011) (described in STAR Methods). This panel contains individuals from a variety of Pan Asian populations, but with a limited number of SNPs (~50 k).

We performed a principal component (PC) analysis that suggests the Bajau are genetically closer to the Saluan than to most other Asian populations, with the possible exceptions being their geographic neighbors (e.g., the Toraja) who also live on the island of Sulawesi (Figure 2). Across populations nearest to the Bajau and Saluan, we observe a gradient of Melanesian to Austronesian ancestry, with an increasing Melanesian component associated with closer geographic proximity to Melanesia. Lipson et al. (2014) suggested that the ancestral components of most of these populations could be modeled as an admixture between speakers of different language strata and substrata more than 1,000 years ago. Combined with our results, it seems that one of the main drivers of differentiation between the Bajau and the Saluan, indicated by the Bajau's separation in the direction of the Manggarai Rampasasa population, could be an increased Austro-Asiatic component (discussed further in STAR Methods with additional results in Figures S2 and S3).

This component can also be seen in admixture analyses, present at varying levels throughout the Bajau individuals when K >7 as a salmon-colored component (Figure 3, for full admixture analysis see Figure S4) that is much less abundant in the Saluan individuals. We then modeled the joint distribution of allele frequencies across ancestry components as a multivariate Gaussian, similarly to the models in TREEMIX (Pickrell and Pritchard, 2012) and Bayenv (Günther and Coop, 2013). From that, we estimated the population tree most compatible with the inferred covariance matrix for K = 19, where the Bajau and Saluan each receive their own unique component (see Figure 3, details on inference procedures in Cheng et al. [2016]).

We estimated the joint demographic history, including the divergence time, of the Bajau and Saluan populations using *fastsimcoal2* (Excoffier et al., 2013). This method optimizes a composite likelihood of the observed joint site-frequency spectrum (SFS) across a high-dimensional space of possible demographic models. We assume these two populations diverged from an ancestral population and then underwent exponential contraction or growth and migration between the two populations (at possibly asymmetrical rates) until the present. We found a model compatible with the data that has a divergence time of ~16 kya, with subsequent high migration from Bajau to Saluan and low migration from Saluan to Bajau (for details see STAR Methods). We note that the estimate of 16 kya may reflect the divergence of old admixture components shared in different proportions by the Saluan and the Bajau, similarly to, for example, European populations being closely related to each other but differing in the proportion of ancient admixture components.

Selection Scan

The results of our population genetics analyses suggest that a selection scan comparing Bajau and Saluan with a more divergent population group as an out-group, such as the Han Chinese, would be appropriate to detect Bajau-specific positive selection. Because our data are of relatively low coverage (on average 5×) and because of the absence of good references panels of haplotypes for the Bajau, computational haplotype phasing for our data would likely be unreliable. Furthermore, previous studies have shown that methods based on local patterns

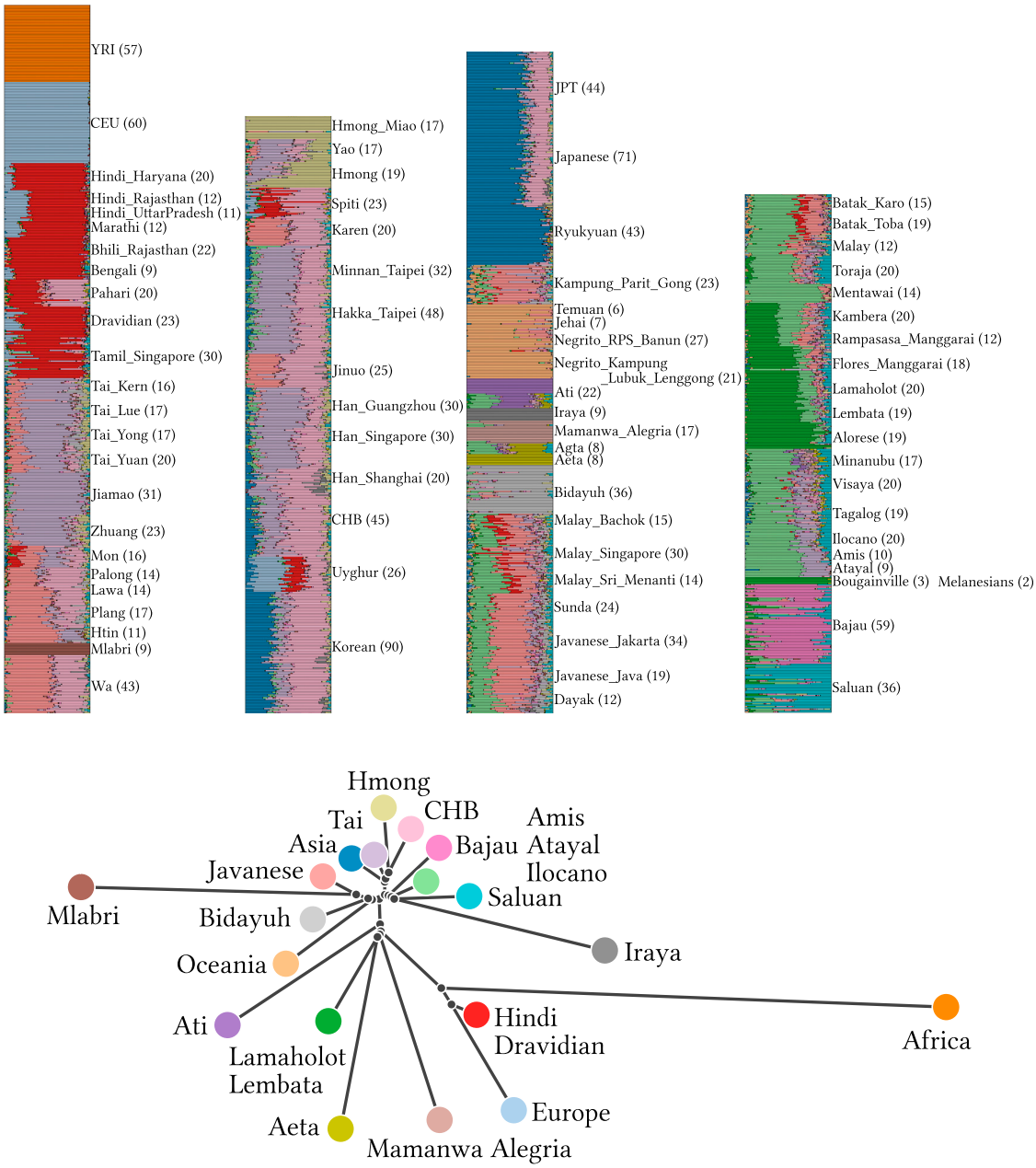


Figure 3. Pan-Asian Admixture and Tree Estimate for *K* = 19, Where Bajau and Saluan Receive Their Own Unique Components
See also Figure S4.

of allele frequency differentiation are highly powerful for detecting local adaptation (Fumagalli et al., 2015; Yi et al., 2010). We therefore merged our sequencing data from the Bajau and Saluan with Han Chinese genomes from the 1000 Genomes Project (Auton et al., 2015) and performed a genome-wide selection scan using a new method for detecting local selection, akin to the PBS statistic (Yi et al., 2010) but based on an explicit likelihood model and adjusted to account for admixture and differing ancestral components (Cheng et al., 2016). The "selscan" program provided in the software suite Ohana (Cheng et al., 2016) allowed us to detect SNPs that deviate strongly in the Bajau from the genome-wide covariance structure using a likelihood ratio test. For each SNP, we introduced a scalar variable that is multiplied onto the variance associated with the Bajau population. This established two nested likelihood models: one that assumes the Bajau-specific allele frequency change can be predicted from the genome-wide covariance pattern and one that allows for a larger change in the Bajau (and hence a higher variance in the Bajau component) than expected from the genome-wide pattern. The resulting likelihood ratio test was used to

Table 1. Top 25 Candidate SNPs from the Selection Scan

chr	pos	rsid	Gene	LL Ratio	CADD Score	Association p Value
1	62249296	rs55870274	INADL	16.297	1.057	0.9525
1	77909616	rs3104465	AK5	15.061	0.892	0.5949
1	206325655	rs55944445	CTSE	17.487	4.756	0.5229
2	97627143	rs182631728	FAM178B	20.867	4.807	0.8554
2	98018288	rs192879353	x	17.746	4.086	0.8486
2	109634804	rs7568610	x	16.687	2.950	0.5461
2	116902921	rs62157649	x	18.585	0.622	0.6934
3	176151090	rs2971468	x	17.509	0.952	0.3095
5	132568207	rs186675869	FSTL4	15.672	4.232	0.2340
5	168553990	rs28544477	SLIT3	17.281	6.511	0.1376
6	31846450	rs117631350	x	18.447	4.358	0.1737
6	166066982	rs3008052	PDE10A	14.960	0.718	0.0003
7	5448302	rs10271391	TNRC18	15.301	1.193	0.0550
7	101787624	rs382934	CUX1	16.358	0.300	0.9665
8	54530665	rs7465617	x	17.865	5.583	0.0295
9	71927931	rs77280170	x	15.932	12.00	0.2562
10	18892382	rs7077786	NSUN6	20.050	1.042	0.6723
10	129682249	rs10765177	CLRN3	16.293	0.748	0.0064
14	78704596	rs10483896	x	17.681	22.80	0.4437
14	96563853	rs7158863	x	28.363	0.230	0.3766
15	82070184	rs118149708	x	16.625	9.838	0.3545
16	65173017	rs74847621	x	20.083	3.395	0.6385
17	848217	rs12936224	NXN	15.959	1.129	0.3151
19	13387904	rs16030	CACNA1A	16.156	8.527	0.4952
19	52144411	rs2167420	x	21.816	1.840	0.4598

LL ratio, log likelihood ratio from the selection scan for each SNP; CADD score, putative disruptiveness of each SNP as predicted by the combined annotation-dependent depletion online tool (PHRED-scaled); association p value, p values from testing for association with spleen size using sequencing depth, height, weight, age, sex, diving, and the first 5 PCs as covariates. When a SNP falls inside a genic region, the relevant gene is listed.

identify loci in which the Bajau experienced a larger-than-expected change in allele frequency as compared to the prediction from the genome-wide pattern, which is a signature of selection (see STAR Methods and Figure S5 for more details).

We note that selection scans with similar designs and similarly small sample sizes have been successful in identifying variants underlying important physiological adaptations in other human populations, such as the Inuit (Fumagalli et al., 2015) and Tibetans (Yi et al., 2010). Whereas genome-wide association studies (GWAS) must correct for testing at all sites in the genome, the power of using natural selection to elucidate function is that the only the sites identified by the selection scan are tested for phenotypic associations, thereby alleviating the burden of multiple testing.

Remarkably, the top hit of our selection scan (Table 1) is SNP rs7158863, located just upstream of BDKRB2, the only gene thus far suggested to be associated with the diving response in humans (Baranova et al., 2017). Genetic variation in this gene is thought to be associated with increased peripheral vasoconstriction, which helps preferentially oxygenate important tissues like the brain, heart, and lungs, thereby potentially increasing dive time. This result strongly suggests that the Bajau harbor genetic variation that has been targeted by selection related to phenotypes of importance for diving.

While some of the selection signals uniquely present in the Bajau may be related to other environmental factors, such as the pathogens, several of the other top hits also fall in candidate genes associated with traits of possible importance for diving. Examples include FAM178B, which encodes a protein that forms a stable complex with carbonic anhydrase, the primary enzyme responsible for maintaining carbon dioxide/bicarbonate balance, thereby helping maintain the pH of the blood (and preventing the build-up of carbon dioxide) (Drew et al., 2017); CACNA1A, which is involved in the regulation of the release of the excitatory neurotransmitter glutamate (Catterall, 1998) and the response to hypoxic conditions (Wang et al., 2005); and PDE10A, a cyclic nucleotide phosphodiesterase involved in the regulation of smooth muscle contraction, including that of the muscle surrounding the spleen (Exton, 1981).

We also assessed the functional importance of the top candidate SNPs using the combined annotation-dependent depletion online tool (CADD v1.3) (Kircher et al., 2014). We found several of these to be located in regulatory regions, including the lead CANA1A SNP, which is in a highly conserved position and lies

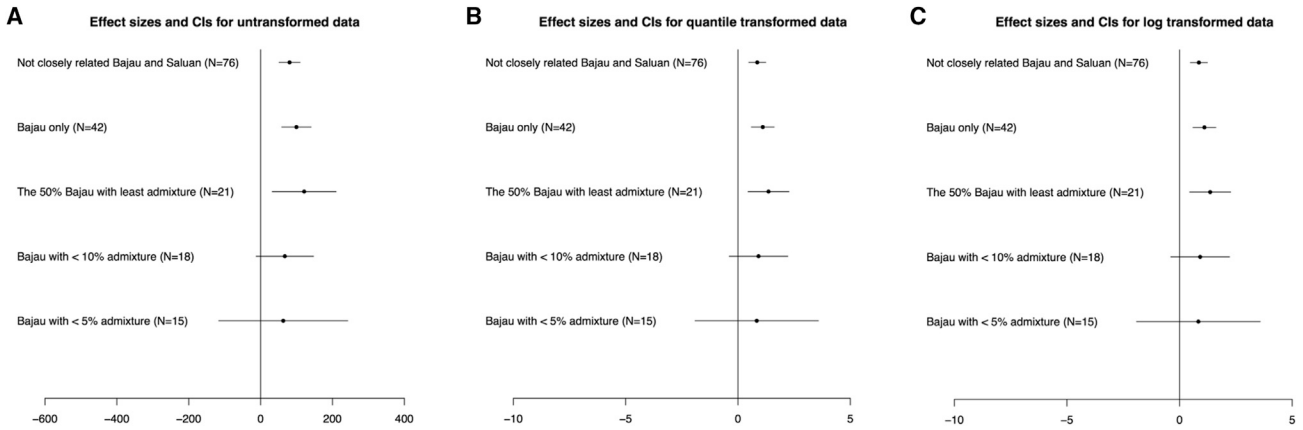


Figure 4. Effect Size Estimates Obtained from Imputed Data Using *lm*

(A–C) Effect size estimates for five different subsets of the data datasets obtained using (A) untransformed spleen sizes, (B) spleen sizes quantile transformed to a normal distribution, and (C) log transformed spleen sizes quantile transformed to a normal distribution.

inside a transcription factor binding site in an open chromatin region (see STAR Methods for details).

Exploring Signals of Adaptive Introgression

In order to investigate the possible origins of the selected alleles identified by our selection scan, we applied statistics that are sensitive to adaptive introgression (AI) from archaic humans. We applied these statistics over 100-kb windows of the genome with a 20-kb overlap (Martin et al., 2015; Racimo et al., 2017) using the maximum-likelihood population-frequency estimates obtained from ANGSD (Korneliussen et al., 2014). We used Yoruba (YRI) as the non-introgressed outgroup and either the Altai Neanderthal (Prüfer et al., 2014) or the Denisovan (Meyer et al., 2012) genomes as the introgressing source. We tested the Bajau as the target population, as well as Han Chinese and Saluan for comparison. Then, we looked for overlaps between the top regions in this scan and the top candidates identified in the *Ohana* selection scan.

We identified one region overlapping chr2:97627143, which falls in the gene FAM178B, that falls in the 99% quantile of the genome-wide distribution for the f_D statistic (Martin et al., 2015). Of the populations considered, this region exclusively stands out in the Bajau, and the signal appears strongest when using Denisova as source. Notably, this region was also proposed as a candidate for Denisovan introgression in Oceanic populations by Sankararaman et al. (2016). Two additional regions, overlapping chr1:62249296 and chr2:116902921, also had values of the Q95 and U20 statistics (Racimo et al., 2017) that are in the 99% quantile of the genome-wide distribution for both Denisova and Neanderthal. However, unlike the FAM178B region, these signals are also present in Han Chinese and the Saluan and are thus not specific to the Bajau. Noticeably, there is no evidence for archaic introgression in PDE10A or BDKRB2.

Association Testing

To investigate if any of the 25 top SNPs from the *Ohana* selection scan are associated with spleen size, we tested for association using a score test, which takes into account genotype uncertainty that is inherent in low-depth sequencing data (Skotte et al., 2012). In these tests, we included sequencing depth, height, weight, age, sex, diving, and the first 5 PCs from a PCA analysis that only included the individuals used in the test as covariates (see STAR Methods for details). We chose to include the PCs in order to correct for population structure. Only one SNP was significantly associated with spleen size at the Bonferroni-corrected 5% significance level (0.002): rs3008052, located in the gene Phosphodiesterase 10A (PDE10A) (Table 1). Notably, the association at this SNP is robust to a number of factors, including different transformations of the spleen sizes as well as additional corrections for population structure other than including PCs as covariates (see STAR Methods and Figure S6). The effect of the allele that has been favored in the Bajau is for an increase in spleen size (Figure 1; for more details see STAR Methods), and the effect size appears consistent across data subsamples and transformations (Figure 4).

PDE10A encodes a cyclic nucleotide phosphodiesterase that can hydrolyze both cAMP and cGMP (Fujishige et al., 1999). By regulating the concentration of cyclic nucleotides, it plays a pivotal role in signal transduction, including alpha-adrenergic responses like the contraction of the smooth muscle surrounding the spleen (Exton, 1981). Our top SNP, rs3008052, is associated with genetic variation affecting expression of PDE10A (Figure 5). PDE10A is most highly expressed in neuronal tissue, but the strongest association for rs3008052 is with expression in the thyroid gland ($p = 0.000016$) (Lonsdale et al., 2013), another organ with cAMP-mediated hormone release. The most likely mode of action is therefore increased spleen size as a consequence of increased thyroid hormone levels.

PDE10A and Spleen Size

To further investigate the mode of action of PDE10A, we examined the top spleen-size associated SNPs from the Bajau data in the Global Biobank Engine (Global Biobank Engine, Stanford, CA; https://biobankengine.stanford.edu/, October 2017). This engine contains a vast amount of case-control association

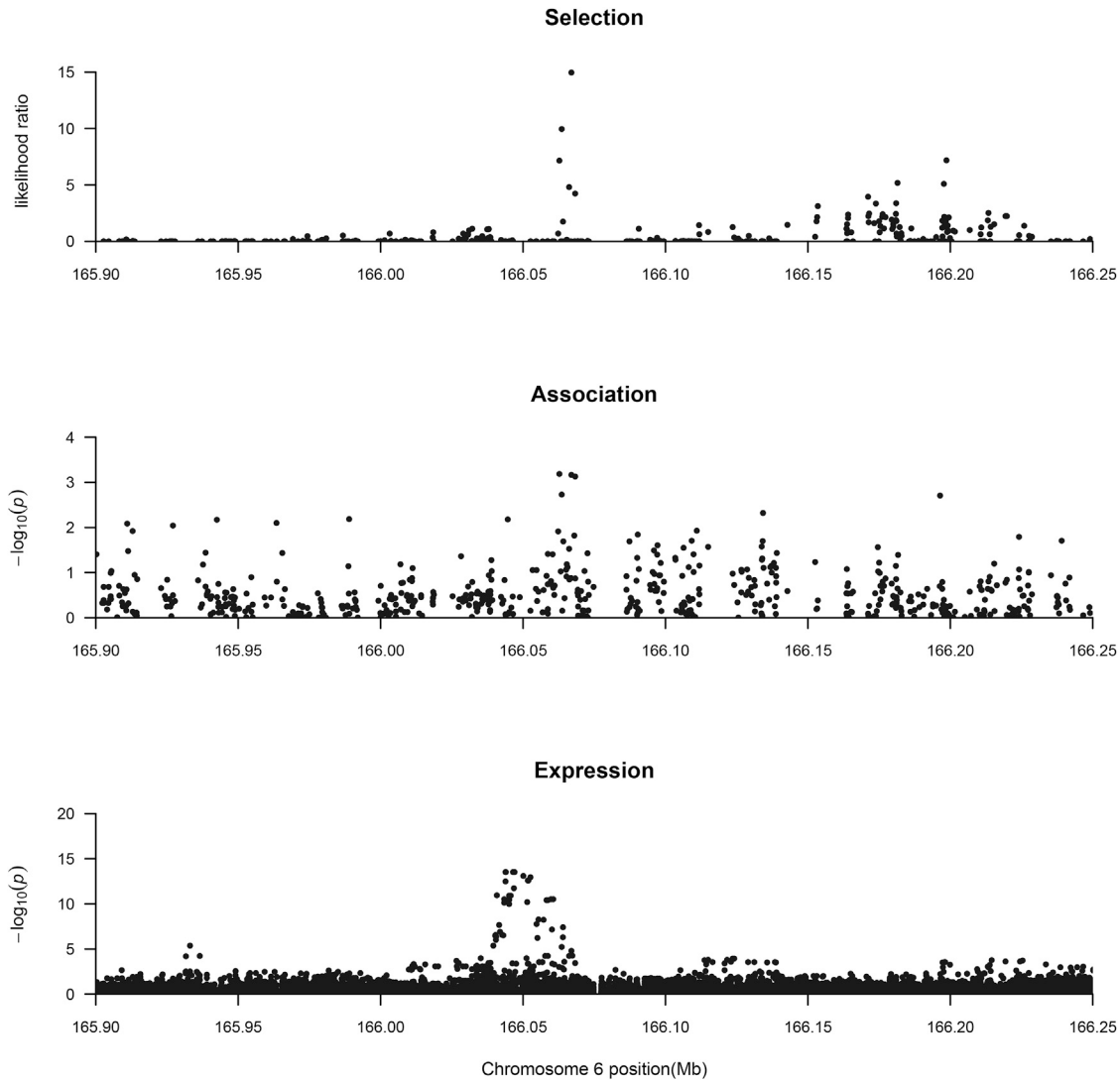


Figure 5. A Concurrent Peak in the Selection, Association, and Expression Scans Marks the Location of the SNP of Interest in PDE10A, rs3008052

Top: genomic position mapped against the log likelihood ratio, a measure of evidence of selection. In the middle panel, position is plotted against $-\log_{10}$(p value) from testing for association between the SNPs and spleen size. Bottom: position is plotted against $-\log_{10}$(p value) from testing for association between the SNPs and thyroid-specific expression of PDE10A. The expression p values are from GTEx (Lonsdale et al., 2013).

results from the UK Biobank hospital in-patient health-related outcomes summary information data. Unfortunately, our top SNP is not present in this data, so we examined SNPs in high linkage disequilibrium (LD) with our SNP; the three most highly correlated SNPs in the surrounding 1-MB region (rs2983527, rs3008050, and rs3008049 with r^2 values of 0.8501, 0.6402, and 0.6140, respectively). We found that all three SNPs are significantly associated with hypothyroidism at the 5% significance level after Bonferroni correction ($p = 0.0017$, $p = 0.00011$, and $p = 0.0043$), and the allele favored in the Bajau leads to a decrease in hypothyroidism in all three SNPs.

To date, there have been no reports on the link between levels of thyroid hormones (thyroxine [T4], which is converted to the active form of triiodothyronine [T3]) and spleen size in humans. However, levels of thyroid hormones (TH) have been strongly linked to spleen size in mice. In a number of studies, the effect of TH on spleen size has been investigated via the gene *Pax8*, which controls the production of follicles in the developing thyroid. *Pax8*$^{-/-}$ knockout mice, exhibiting deep congenital hypothyroidism (a nearly complete absence of T4 and T3), showed drastic reduction in spleen size (Angelin-Duclos et al., 2005; Flamant et al., 2002). Interestingly, the small spleen phenotype was shown to be partially reversible through TH injection (Angelin-Duclos et al., 2005). In further support of the link between TH and spleen size, mice treated with T4 to artificially induce hyperthyroxinemia have shown significant increases in spleen weight over periods of 8 and 32 weeks ($p < 0.001$ for both time periods) (Watanabe et al., 1995).

To further test the hypothesis of an association between the selected SNPs in PDE10A and thyroid hormone levels, we turned to the 500-FG cohort from the Human Functional Genomics Project (ter Horst et al., 2016), which has been genotyped using a SNP chip (Illumina HumanOmniExpressExome-8 v1.0) and for which T4 and thyroid-stimulating hormone (TSH) concentrations have been measured (see STAR Methods for details). We note that this cohort comes from a European population that is highly diverged from the Bajau, with a potentially different haplotype structure, and that validation in this cohort may therefore be difficult. Nonetheless, we tested for association using a linear regression technique as implemented in the Matrix eQTL R package (Shabalin, 2012). The method was applied to imputed "dosage" SNP data, where in imputation the genotypes are not discretely called, but are instead given values between 0 and 2 in each individual. Individuals that were related or had non-European ancestry were removed, and the tests were corrected for age, gender, BMI, and oral contraceptive usage by including these factors as covariates. We examined four SNPs: the lead SNP, rs3008052, and three other SNPs in high LD in Europeans with rs3008052 (rs2983527, rs3008050, and rs3008049). All of these SNPs have significant associations with hypothyroidism in the Global Biobank Engine data. For the top PDE10A SNP, rs3008052, we found a clear, significant association ($p = 0.0017$), with the allele at higher frequency in the Bajau associated with elevated T4 circulating plasma concentrations. We also found significant associations at the high-LD SNPs as well as negative associations with TSH concentrations (see Figure S7). The negative correlation between the associations with T4 and TSH concentrations are expected due to the well-known negative feedback between T4 and TSH.

There are no GWAS studies of spleen size conducted in any other populations that would allow us to directly validate the association with spleen size. However, abdominal MRI scans have been carried out in a study to assess the incidence of liver steatosis in the 300-OB cohort (see STAR Methods for details about the cohort). We were able to estimate spleen sizes of the individuals in this cohort from the MRI images, and we tested spleen size association for the same four SNPs as used in the T4 association test. We should not necessarily expect an association in this cohort, because of its different genetic background, but rs3008050 and rs3008049 were marginally associated with spleen size, matching the association found in the Bajau, with p values of borderline significance ($p = 0.05083$ and $p = 0.06493$). Unfortunately, rs3008050 was not present in the Bajau data. However, rs3008049 is present and has a highly significant association with spleen size ($p = 0.00043$) in the Bajau. The combined p value from the two studies (using Fisher's method) for rs3008049 is 2.79e−05, providing strong evidence for an association with spleen size in the region with evidence of natural selection in PDE10A. Even upon Bonferroni correction for the fact that we have investigated two SNPs with overlap between the two datasets (rs3008052 and rs3008049), and the original screening was based on 25 SNPs in 25 genes, the results remain significant ($p = 0.00073$).

Evolutionary History of the Selected Loci

The SNP rs3008049, which has been under selection in the Bajau, also has a wide geographic distribution. While it occurs at a high frequency in Bajau (37.1%), it also segregates at appreciable frequencies in both Saluans (6.7%) and Han Chinese (3.0%). The ubiquity of this SNP at some frequency across these populations suggests that selection may have occurred when rs3008049 was a standing variant, segregating neutrally in all of these populations prior to selection in the Bajau. For rs7158863, our top hit in the selection scan, we estimate this allele to be at 18.3% in the Bajau and <1%, or essentially non-segregating, in both the Saluans and Han Chinese.

In order to test whether the observed inter-population allele frequency differences are due to drift or sampling error rather than selection, we used the inferred isolation-with-migration demographic model to simulate the evolution of neutral alleles under selection from a standing variant. We found that these frequency differences are unlikely due to drift or sampling error ($p < 0.001$ and $p < 0.024$ for PDE10A and BDKRB2, respectively, see STAR Methods for details). We then tested the extent to which selection may have driven these allele frequency changes by quantifying the strength and timing of this potential selective event.

In order to explicitly fit a model of selection on standing variation to these allele frequencies, we simulated allele frequency trajectories under the same isolation-with-migration model, varying the values of three parameters: (1) s, the selection coefficient, (2) t, the time at which selection began to act on the allele, and (3) f, the frequency of the allele at the divergence time. We found that the PDE10A allele has a frequency difference compatible with a moderately strong selection coefficient of $s = 0.005$ acting on a standing variant ($f = 0.02$) starting roughly at the time of divergence ($t = 15.4$ kya) (Figure 6). These estimates, suggesting a sweep from a standing variant, are concordant with the frequency of the allele in panels from a number of other populations, such as Han Chinese, whose split with the Bajau predates the Bajau-Saluan split. We also find the BDKRB2 allele has a frequency difference compatible with a strong selection coefficient of $s = 0.01$ acting on a low-frequency variant ($f = 0.0001$) also starting roughly at the time of divergence ($t = 15.4$ kya). It is important to note that because the lead variants may not be the causal variants, this might affect our estimates and skew our conclusions toward selection from standing variation rather than from *de novo* mutation. As an additional check, we therefore examined the width of the likelihood ratio peaks for the lead SNP in PDE10A and BDKRB2 by calculating the distance between the two most distant SNPs that define the sweep region. We found the peak widths to be extremely short (5,583 bp and 2,825 bp, respectively), providing additional evidence of a sweep from standing variation.

DISCUSSION

In this study, we identify multiple candidate genes for adaptation to breath-hold diving in the Bajau. We have investigated one of the candidate genes, PDE10A, in detail, and the rest remain promising targets for future studies.

In the region of PDEA10, we find that the SNPs at the selection scan peak are associated with thyroid function and spleen size. Because thyroid hormones regulate normal erythropoiesis during early postnatal development (Angelin-Duclos et al., 2005),

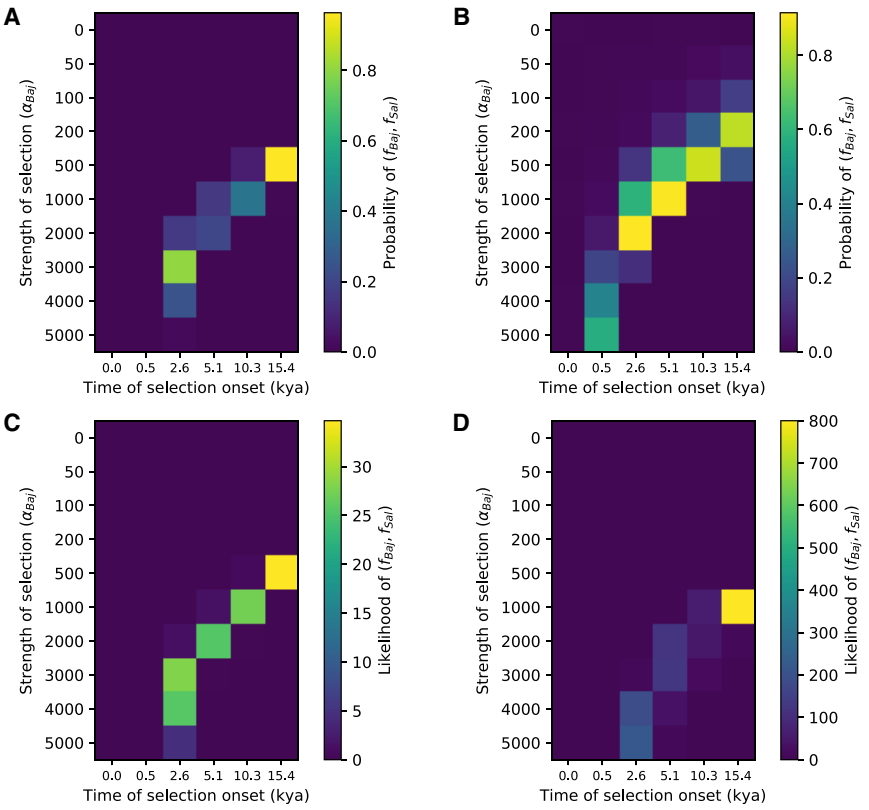


Figure 6. Estimated Strength and Timing of Selection for the Selected Alleles in PDE10A and BDKRB2

(A and B) The probabilities for the (A) PDE10A and (B) BDKRB2 alleles, fixing f = 0.03, 0.0001, respectively.

(C and D) The scaled aforementioned probabilities using the prior probability of a neutral allele segregating at frequency f (p(f) $\propto$ 1/f), and fixing f = 0.02, 0.0001 for (C) *PDE10A* and (D) *BDKRB2*, respectively.

the observed large spleen phenotype in the Bajau may be indicative of higher volume of erythrocytic cells. The selected genotype could provide a 2-fold phenotypic advantage of both an increased quantity of oxygenated cells and a larger reservoir in which to store them. Alternatively, the larger spleen size may simply be a secondary effect of the increased cell volume. Regardless, the resulting physiological change seems to have provided a functional adaptation to the conditions of acute hypoxia that is characteristic of breath-hold diving.

Most evidence suggests that the adaptations we observe in the Bajau come from standing variation. First, the candidate variants in genes such as PDE10A and BDKRB2 have a broad geographic distribution. Although we may not have identified the causal variants in these genes, the selection signals does not seem to be associated with variants private to the Bajau. Our model-based inferences are also compatible with selection acting on standing variation. In addition, the variant in FAM178B may originate from adaptive introgression, possibly from Denisovans, while the PDE10A and BDKRB2 variants likely do not.

Overall, our results suggest that the Bajau have undergone unique adaptations associated with spleen size and the diving response, adding new examples to the list of remarkable genetic adaptations humans have experienced in recent evolutionary history. Similar to other of the most extreme adaptations human have experienced, such as adaptation to diet associated with pastoralism (Ranciaro et al., 2014) or shifts in environmental availability of food resources (Fumagalli et al., 2015), these genetic adaptations have emerged as a consequence of new cultural practices, illustrating that human culture and biology have been co-evolving for thousands of years.

Furthermore, as both of the adaptations in the Bajau described above relate to hypoxia tolerance, they may be of significant medical relevance (e.g., by providing a new understanding of the relationship between hypoxia, thyroid function, cell volume, and spleen size). Prior studies using locally adapted populations to investigate the underlying genetics of hypoxia tolerance and related physiological processes in humans have focused nearly exclusively on high altitude populations (reviewed in McKenna and Martin, 2016). The Bajau, and other diving-dependent populations, exhibit an entirely new system of human physiological adaptations to environmental conditions not previously explored.

STAR★METHODS

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

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Indonesian samples
  - European samples
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Read processing
  - Error rate estimation
  - Relatedness estimation
  - Population genetics analyses
  - Testing for a difference in spleen size between Bajau and Saluan
  - Bajau selection analysis using Ohana
  - Functional annotation of selected SNPs
  - Spleen size association testing
  - Thyroid hormone association testing
  - Estimating the strength and timing of selection
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and two tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.03.054.

A video abstract is available at https://doi.org/10.1016/j.cell.2018.03.054#mmc3.

ACKNOWLEDGMENTS

We thank Richard Dougherty for medical imaging consultation, Yani Mile and Mardiyanto Saahi for assistance in Indonesia, and the residents of Jaya Bakti and Koyoan for their participation and support, especially Kepala Desa Hasan and Pai Bayubu. We would also like to acknowledge GenomeDK HPC Hub at Aarhus University and its staff, the staff of the Danish National High-throughput DNA Sequencing Centre, the laboratory technicians of the Centre for GeoGenetics, and Kurt Kjær for assistance with graphic design. The authors would like to thank the Rivas lab for making the Global Biobank Engine resource available. We gratefully acknowledge Marinette van der Graaf for the assistance with the MRIs and Vincenzo Positano of the CNR Institute of Clinical Physiology, Pisa, Italy, for providing us with the HIPPO FAT. This project was supported by the Lundbeck Foundation, the Danish National Research Foundation (DNRF94), the Danish National Science Foundation (FNU 109931), and KU2016. The 500FG study from the HFGP is supported by an European Research Council (ERC) Consolidator grant (ERC 310372) to M.G.N. The 300-OB study was supported by an IN-CONTROL CVON grant (CVON2012-03) from the Netherlands Heart Foundation to M.G.N. and L.A.B.J. M.G.N. was supported by a Spinoza grant of the Netherlands Organization for Scientific Research. E.W. thanks St. Johns College, Cambridge, for providing a stimulating environment for scientific discussion.

AUTHOR CONTRIBUTIONS

M.A.I. conceived of the project. M.A.I. and E.W. initiated the project. M.A.I., R.N., and E.W. designed the study. S.S. provided logistical support in Indonesia. M.A.I. and P.d.B.D. collected the samples and prepared the DNA for sequencing with A.S.-O. S.R. processed the sequencing data. R.N. oversaw all computational analyses, which were performed by I.M., J.C., T.S.K., M.S., A.J.S., F.R., and M.A.I. M.G.N. and L.A.B.J. designed and coordinated the 300-OB and 500-FG studies within the HFGP, for which I.C.L.v.d.M. and R.t.H. collected the samples and analyzed the data. M.A.I., R.N., and E.W. wrote the manuscript with input from all authors. R.N. and E.W. supervised the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 21, 2017
Revised: January 1, 2018
Accepted: March 21, 2018
Published: April 19, 2018

REFERENCES

Aguirre-Gamboa, R., Joosten, I., Urbano, P.C.M., van der Molen, R.G., van Rijssen, E., van Cranenbroek, B., Oosting, M., Smeekens, S., Jaeger, M., Zorro, M., et al. (2016). Differential effects of environmental and genetic factors on T and B cell immune traits. Cell Rep. *17*, 2474–2487.

Allentoft, M.E., Sikora, M., Sjögren, K.G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of bronze age Eurasia. Nature *522*, 167–172.

Angelin-Duclos, C., Domenget, C., Kolbus, A., Beug, H., Jurdic, P., and Samarut, J. (2005). Thyroid hormone T3 acting through the thyroid hormone α receptor is necessary for implementation of erythropoiesis in the neonatal spleen environment in the mouse. Development *132*, 925–934.

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., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Baranova, T.I., Berlov, D.N., Glotov, O.S., Korf, E.A., Minigalin, A.D., Mitrofanova, A.V., Ahmetov, I.I., and Glotov, A.S. (2017). Genetic determination of the vascular reactions in humans in response to the diving reflex. Am. J. Physiol. Heart Circ. Physiol. *312*, H622–H631.

Beall, C.M. (2006). Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. Integr. Comp. Biol. *46*, 18–24.

Beall, C.M., Cavalleri, G.L., Deng, L., Elston, R.C., Gao, Y., Knight, J., Li, C., Li, J.C., Liang, Y., McCormack, M., et al. (2010). Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. Proc. Natl. Acad. Sci. USA *107*, 11459–11464.

Catterall, W.A. (1998). Structure and function of neuronal Ca2+ channels and their role in neurotransmitter release. Cell Calcium *24*, 307–323.

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

Cheng, J.Y., Mailund, T., and Nielsen, R. (2016). Ohana, a tool set for population genetic analyses of admixture components. bioRxiv. https://doi.org/10.1101/071233.

Clifton, J., and Majors, C. (2012). Culture, conservation, and conflict: perspectives on marine protection among the Bajau of Southeast Asia. Soc. Nat. Resour. *25*, 716–725.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J.K. (2010). Using environmental correlations to identify loci underlying local adaptation. Genetics *185*, 1411–1423.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Drew, K., Lee, C., Huizar, R.L., Tu, F., Borgeson, B., McWhite, C.D., Ma, Y., Wallingford, J.B., and Marcotte, E.M. (2017). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Mol. Syst. Biol. *13*, 932.

Elsner, R., Franklin, D.L., Van Citters, R.L., and Kenney, D.W. (1966). Cardiovascular defense against asphyxia. Science *153*, 941–949.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. PLoS Genetics *9*.

Exton, J.H. (1981). Molecular mechanisms involved in alpha-adrenergic responses. Mol. Cell. Endocrinol. *23*, 233–264.

Ferrigno, M., Ferretti, G., Ellis, A., Warkander, D., Costa, M., Cerretelli, P., and Lundgren, C.E. (1997). Cardiovascular changes during deep breath-hold dives in a pressure chamber. J. Appl. Physiol. *83*, 1282–1290.

Flamant, F., Poguet, A.L., Plateroti, M., Chassande, O., Gauthier, K., Streichenberger, N., Mansouri, A., and Samarut, J. (2002). Congenital hypothyroid Pax8(-/-) mutant mice can be rescued by inactivating the TRalpha gene. Mol. Endocrinol. *16*, 24–32.

Foster, G.E., and Sheel, A.W. (2005). The human diving response, its function, and its control. Scand. J. Med. Sci. Sports *15*, 3–12.

Fujishige, K., Kotera, J., Michibata, H., Yuasa, K., Takebayashi, S., Okumura, K., and Omori, K. (1999). Cloning and characterization of a novel human phosphodiesterase that hydrolyzes both cAMP and cGMP (PDE10A). J. Biol. Chem. *274*, 18438–18445.

Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M.E., Korneliussen, T.S., Gerbault, P., Skotte, L., Linneberg, A., et al. (2015). Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science *349*, 1343–1347.

Gislén, A., Dacke, M., Kröger, R.H., Abrahamsson, M., Nilsson, D.-E., and Warrant, E.J. (2003). Superior underwater vision in a human population of sea gypsies. Curr. Biol. *13*, 833–836.

Gislén, A., Warrant, E.J., Dacke, M., and Kröger, R.H. (2006). Visual training improves underwater vision in children. Vision Res. *46*, 3443–3450.

Grocott, M., Richardson, A., Montgomery, H., and Mythen, M. (2007). Caudwell Xtreme Everest: a field study of human adaptation to hypoxia. Crit. Care *11*, 151.

Günther, T., and Coop, G. (2013). Robust identification of local adaptation from allele frequencies. Genetics *195*, 205–220.

Hochachka, P.W. (1986). Balancing conflicting metabolic demands of exercise and diving. Fed. Proc. *45*, 2948–2952.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods *9*, 473–476.

Holewijn, S., den Heijer, M., Swinkels, D.W.H., Stalenhoef, A.F., and de Graaf, J. (2010). Apolipoprotein B, non-HDL cholesterol and LDL cholesterol for identifying individuals at increased cardiovascular risk. J. Intern. Med. *268*, 567–577.

R Development Core Team (2008). R: A language and environment for statistical computing (R Foundation for Statistical Computing).

Hurford, W.E., Hong, S.K., Park, Y.S., Ahn, D.W., Shiraki, K., Mohri, M., and Zapol, W.M. (1990). Splenic contraction during breath-hold diving in the Korean ama. J. Appl. Physiol. *69*, 932–936.

Hurford, W.E., Hochachka, P.W., Schneider, R.C., Guyton, G.P., Stanek, K.S., Zapol, D.G., Liggins, G.C., and Zapol, W.M. (1996). Splenic contraction, catecholamine release, and blood volume redistribution during diving in the Weddell seal. J. Appl. Physiol. *80*, 298–306.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Kooyman, G.L., and Campbell, W.B. (1972). Heart rates in freely diving Weddell seals, Leptonychotes weddelli. Comp. Biochem. Physiol. A *43*, 31–36.

Korneliussen, T.S., and Moltke, I. (2015). NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. Bioinformatics *31*, 4009–4011.

Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. BMC Bioinformatics *15*, 356.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, arXiv:1303.3997.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lin, Y.C., Matsuura, D.T., and Whittow, G.C. (1972). Respiratory variation of heart rate in the California sea lion. Am. J. Physiol. *222*, 260–264.

Lin, Y.C., Shida, K.K., and Hong, S.K. (1983). Effects of hypercapnia, hypoxia, and rebreathing on circulatory response to apnea. J. Appl. Physiol. *54*, 172–177.

Lipson, M., Loh, P.-R., Patterson, N., Moorjani, P., Ko, Y.-C., Stoneking, M., Berger, B., and Reich, D. (2014). Reconstructing Austronesian population history in Island Southeast Asia. Nat. Commun. *5*, 4689.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

Martin, S.H., Davey, J.W., and Jiggins, C.D. (2015). Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol. Biol. Evol. *32*, 244–257.

McKenna, H., and Martin, D. (2016). Surviving physiological stress: Can insights into human adaptation to austere environments be applied to the critical care unit? Trends Anaesthesia Crit. Care *11*, 6–13.

Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.

Mottishaw, P.D., Thornton, S.J., and Hochachka, P.W. (1999). The diving response mechanism and its surprising evolutionary path in seals and sea lions. Am. Zool. *39*, 434–450.

Netea, M.G., Joosten, L.A., Li, Y., Kumar, V., Oosting, M., Smeekens, S., Jaeger, M., Ter Horst, R., Schirmer, M., Vlamakis, H., et al. (2016). Understanding human immune function using the resources from the Human Functional Genomics Project. Nat. Med. *22*, 831–833.

Ngamphiw, C., Assawamakin, A., Xu, S., Shaw, P.J., Yang, J.O., Ghang, H., Bhak, J., Liu, E., and Tongsima, S.; HUGO Pan-Asian SNP Consortium (2011). PanSNPdb: the Pan-Asian SNP genotyping database. PLoS ONE *6*, e21451.

Oosthuyse, B., Moons, L., Storkebaum, E., Beck, H., Nuyens, D., Brusselmans, K., Van Dorpe, J., Hellings, P., Gorselink, M., Heymans, S., et al. (2001). Deletion of the hypoxia-response element in the vascular endothelial growth factor promoter causes motor neuron degeneration. Nat. Genet. *28*, 131–138.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., et al. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature *499*, 74–78.

Peng, Y., Yang, Z., Zhang, H., Cui, C., Qi, X., Luo, X., Tao, X., Wu, T., Ouzhuluobu, Basang, et al. (2011). Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. Mol. Biol. Evol. *28*, 1075–1081.

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

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Racimo, F., Marnetto, D., and Huerta-Sánchez, E. (2017). Signatures of archaic adaptive introgression in present-day human populations. Mol. Biol. Evol. *34*, 296–317.

Ranciaro, A., Campbell, M.C., Hirbo, J.B., Ko, W.-Y., Froment, A., Anagnostou, P., Kotze, M.J., Ibrahim, M., Nyambo, T., Omar, S.A., and Tishkoff, S.A. (2014). Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am. J. Hum. Genet. *94*, 496–510.

Rankin, E.B., and Giaccia, A.J. (2008). The role of hypoxia-inducible factors in tumorigenesis. Cell Death Differ. *15*, 678–685.

Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Curr. Biol. *26*, 1241–1247.

Sather, C. (1997). The Bajau Laut: Adaptation, History, and Fate in a Maritime Fishing Society of South-Eastern Sabah (Oxford University Press).

Schagatay, E. (2014). Human breath-hold diving ability and the underlying physiology. Hum. Evol. *29*, 125–140.

Schagatay, E., Lodin-Sundström, A., and Abrahamsson, E. (2011). Underwater working times in two groups of traditional apnea divers in Asia: the Ama and the Bajau. Diving Hyperb. Med. *41*, 27–30.

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes *9*, 88.

Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353–1358.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Simonson, T.S., Yang, Y., Huff, C.D., Yun, H., Qin, G., Witherspoon, D.J., Bai, Z., Lorenzo, F.R., Xing, J., Jorde, L.B., et al. (2010). Genetic evidence for high-altitude adaptation in Tibet. Science *329*, 72–75.

Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., Gilbert, M.T., Götherström, A., and Jakobsson, M. (2012). Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science *336*, 466–469.

Skotte, L., Korneliussen, T.S., and Albrechtsen, A. (2012). Association testing for next-generation sequencing data using score statistics. Genet. Epidemiol. *36*, 430–437.

Sopher, D.E. (1965). The Sea Nomads: A Study of the Maritime Boat People of Southeast Asia (National Museum).

Sterba, J.A., and Lundgren, C.E. (1988). Breath-hold duration in man and the diving response induced by face immersion. Undersea Biomed. Res. *15*, 361–375.

Stewart, I.B., and McKenzie, D.C. (2002). The human spleen during physiological stress. Sports Med. *32*, 361–369.

Talks, K.L., Turley, H., Gatter, K.C., Maxwell, P.H., Pugh, C.W., Ratcliffe, P.J., and Harris, A.L. (2000). The expression and distribution of the hypoxia-inducible factors HIF-1alpha and HIF-2alpha in normal human tissues, cancers, and tumor-associated macrophages. Am. J. Pathol. *157*, 411–421.

ter Horst, R., Jaeger, M., Smeekens, S.P., Oosting, M., Swertz, M.A., Li, Y., Kumar, V., Diavatopoulos, D.A., Jansen, A.F.M., Lemmers, H., et al. (2016). Host and environmental factors influencing individual human cytokine responses. Cell *167*, 1111–1124.

Thornton, S.J., and Hochachka, P.W. (2004). Oxygen and the diving seal. Undersea Hyperb. Med. *31*, 81–95.

Wang, V., Davis, D.A., Haque, M., Huang, L.E., and Yarchoan, R. (2005). Differential gene up-regulation by hypoxia-inducible factor-1alpha and hypoxia-inducible factor-2alpha in HEK293T cells. Cancer Res. *65*, 3299–3306.

Watanabe, K., Iwatani, Y., Hidaka, Y., Watanabe, M., and Amino, N. (1995). Long-term effects of thyroid hormone on lymphocyte subsets in spleens and thymuses of mice. Endocr. J. *42*, 661–668.

Wuren, T., Simonson, T.S., Qin, G., Xing, J., Huff, C.D., Witherspoon, D.J., Jorde, L.B., and Ge, R.L. (2014). Shared and unique signals of high-altitude adaptation in geographically distinct Tibetan populations. PLoS ONE *9*, e88252.

Xiang, K., Ouzhuluobu, Peng, Y., Yang, Z., Zhang, X., Cui, C., Zhang, H., Li, M., Zhang, Y., Bianba, et al. (2013). Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. Mol. Biol. Evol. *30*, 1889–1898.

Xu, S., Li, S., Yang, Y., Tan, J., Lou, H., Jin, W., Yang, L., Pan, X., Wang, J., Shen, Y., et al. (2011). A genome-wide search for signals of high-altitude adaptation in Tibetans. Mol. Biol. Evol. *28*, 1003–1011.

Yang, J., Jin, Z.B., Chen, J., Huang, X.F., Li, X.M., Liang, Y.B., Mao, J.Y., Chen, X., Zheng, Z., Bakshi, A., et al. (2017). Genetic signatures of high-altitude adaptation in Tibetans. Proc. Natl. Acad. Sci. USA *114*, 4189–4194.

Yetter, E.M., Acosta, K.B., Olson, M.C., and Blundell, K. (2003). Estimating splenic volume: sonographic measurements correlated with helical CT determination. AJR Am. J. Roentgenol. *181*, 1615–1620.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.

Zapol, W.M., Liggins, G.C., Schneider, R.C., Qvist, J., Snider, M.T., Creasy, R.K., and Hochachka, P.W. (1979). Regional blood flow during simulated diving in the conscious Weddell seal. J. Appl. Physiol. *47*, 968–973.

Zhong, H., De Marzo, A.M., Laughner, E., Lim, M., Hilton, D.A., Zagzag, D., Buechler, P., Isaacs, W.B., Semenza, G.L., and Simons, J.W. (1999). Overexpression of hypoxia-inducible factor 1alpha in common human cancers and their metastases. Cancer Res. *59*, 5830–5835.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Saliva sample	Healthy volunteers	1104-1106, 1201, 1202, 1204-1208, 1301, 1302, 1304, 1306-1309, 1401-1419, 1501-1510, 1512-1516, 1518, 1519, 1601, 1602, 1701, 1702, 1705-1707, 2301, 2302, 2401-2408, 2410-2413, 2415-2425, 2427, 2428, 2430, 2501, 2502, 2504-2507
Deposited Data		
Raw data, Bajau and Saluan	This paper	EGAS00001002823
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
1000 Genomes Han Chinese	Genome Project Consortium	http://www.internationalgenome.org/data
PanAsian SNP data	Ngamphiw et al., 2011	http://www4a.biotec.or.th/PASNP
Software and Algorithms		
BWA	Li, 2013	http://bio-bwa.sourceforge.net/
Samtools	Li et al., 2009	http://samtools.sourceforge.net/
GATK	DePristo et al., 2011	https://software.broadinstitute.org/gatk/
ANGSD	Korneliussen et al., 2014	http://www.popgen.dk/angsd/index.php/ANGSD
NGSRelate	Korneliussen and Moltke, 2015	http://www.popgen.dk/software/index.php/NgsRelate
Fastsimcoal	Excoffier et al., 2013	http://cmpg.unibe.ch/software/fastsimcoal2/
Ohana	Cheng et al., 2016	https://github.com/jade-cheng/ohana
Plink	Chang et al., 2015	http://zzz.bwh.harvard.edu/plink/
CADD	Kircher et al., 2014	http://cadd.gs.washington.edu/
ENCODE	National Human Genome Research Institute	https://www.encodeproject.org/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information should be directed to and will be fulfilled by the Lead Contact, Rasmus Nielsen (rasmus_nielsen@berkeley.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Indonesian samples were taken according to a protocol approved by The Developing-Country Committee of The Danish National Committee on Health Research Ethics. All participants provided informed consent for their participation in this study. The 59 ethnic Bajau and 34 ethnic Saluan individuals analyzed in this study were from the Indonesian villages of Jaya Bakti and Koyoan, respectively. Both villages are found on the eastern tip of the Central Sulawesi peninsula. The individuals had an average age of 41.1 (ranging from 18-85), and were comprised of 75.3% men and 24.7% women.

European data came from two cohorts, the 500FG cohort and the 300-Obesity (300-OB) cohort. Both studies were approved by the Ethical Committee of Radboud University Nijmegen, and all participants received detailed printed and oral information and provided written informed consent. Experiments were conducted according to the principles expressed in the Declaration of Helsinki. The 500FG cohort consisted of 534 healthy individuals of Western-European genetic ancestry (ter Horst et al., 2016). The inclusion of the volunteers took place between 8/2013 and 12/2014 at the Radboud University Medical Center, the Netherlands. Volunteers had a mean age of 28.5 ± 13.9 years, 44.5% were men and 55.5% women. The 300-OB cohort included 302 individuals aged between 55 to 80 at the Radboud University Medical Center in the period between 2014 and 2016 (Netea et al., 2016). All subjects had a BMI above 27 kg/m^2 and most had also participated in the Nijmegen Biomedical Study – Non-Invasive Measurements of Atherosclerosis 1 (NBS-NIMA1) study, a population-based survey of Nijmegen residents (Holewijn et al., 2010). Among the patients 45% were men and 55% women.

METHOD DETAILS

Indonesian samples

Spit samples were collected using the Oragene DNA kit from DNA Genotek.

Spleen measurements were performed using the SonoScape A6 Portable Ultrasound System with a C321 2-5MHz Convex Transducer. Spleen measurements were made in the transverse and longitudinal planes in order to calculate spleen volume according to the methodology outlined in Yetter et al. (2003) shown to have the highest correlation with values obtained through a computed tomography (CT) scan. All spleen measurements were taken by the same researcher to ensure consistent measurements.

Two other phenotypic measurements, height and weight, were obtained at the same time the DNA samples and spleen measurements were taken. Through a brief interview with each participant, we obtained demographic information including ethnicity, age, gender, and diving history. An individual was defined as a diver if they engaged in breath-hold diving on average at least three times a week, with the last dive having occurred within the last week. Non-divers were those who had never engaged in frequent or non-recreational diving (thereby excluding formerly frequent divers). There were no ambiguous cases in which an individual fell between diver and non-diver.

Genomic DNA was extracted from spit samples using the prepIT L2P extraction kit from DNA Genotek. Genomic DNA extracts were quantified using a Qubit™ dsDNA High-Sensitivity Assay. Aliquots containing 75 ng of DNA were transfered to Covaris micro-TUBE-15 and fragmented down to 550 bp on a Covaris M220 ultrasonicator (duty factor 20%, 50 cycles per burst, 23 s). Sequencing libraries were built using the TruSeq Nano DNA Library Preparation Kit on an Illumina NeoPrep instrument, following manufacturer's instructions, but without normalization step. Each library preparation session included one negative control. Sequencing Libraries were checked for size distribution and molarity on an Agilent BioAnalyser 2100 instrument, using the High-Sensitivity DNA Kit, and pooled equimolar (8 to 16 libraries per pool). Each pool was sequenced 125 Paired-End over one or two lanes on the Illumina HiSeq2500 (version 4 chemistry). Samples were sequenced to an average depth of 5x.

European samples

In the 500FG cohort, blood samples were collected by venipuncture from the cubital vein in the morning, and a large set of circulating mediators, including steroid and thyroid hormones, were measured as previously reported (ter Horst et al., 2016).

In the 300-OB cohort, abdominal MRIs were performed to assess the prevalence of hepatic steatosis and the abdominal fat distribution. All MRI examinations were performed using a 3.0 T Magnetom Trio or Skyra (Siemens, Erlangen, Germany). Subjects were examined in the supine position with their arms positioned parallel to the lateral sides of the body. For each subject, a series of thirty T1-weighted TIRM axial MR images of 5 mm each were acquired from the liver region. The images acquired were retrieved from the MR scanner in DICOM (Digital Imaging and Communications in Medicine) format, and analyzed with software developed in the IDL 6.0 environment, called HIPPO FAT (version 1.3, V. Positano[22]). In HIPPO FAT all thirty slices were assessed for the spleen size. Of the slice with visually the largest spleen size, contour lines were placed manually and the size was calculated automatically. The intra-class correlation coefficients of the spleen size was 0.995.

QUANTIFICATION AND STATISTICAL ANALYSIS

Read processing

The sequences were basecalled using the Illumina software CASAVA-1.8.2 and de-multiplexed using a full match of the 6 nucleotide index incorporated during library adaptor ligation. The reads were trimmed using AdapterRemoval-2.1.3 (Schubert et al., 2016) for adaptor sequences and leading/trailing stretches of Ns. Additionally, bases with quality of 2 or less were removed by trimming from the 3′ and only reads larger than 30bp were kept. Retained read pairs after trimming were aligned using BWA mem-0.7.10 (Li, 2013) to GRCh37 and processed using samtools-1.3.1 (Li et al., 2009) removing reads with a mapping quality lower than 30 and merged to libraries. Hereafter duplicates were marked using picard-1.127 MarkDuplicates (https://broadinstitute.github.io/picard/), libraries merged to sample level and realigned using GATK-3.3.0 (DePristo et al., 2011) with Mills and 1000G gold standard indels. Finally, realigned bams had the md-tag updated and extended BAQs calculated using samtools calmd. Read depth and coverage were determined using pysam (https://code.google.com/archive/p/pysam/) and BEDtools (Quinlan and Hall, 2010).

Error rate estimation

To estimate error rates, we analyzed each of the samples of interest separately using a method in ANGSD (Korneliussen et al., 2014) that is based on comparison to an error free genome. Briefly explained, the method exploits the fact that if the analyzed sample has no errors it should have the same expected number of derived alleles as an error free genome, and that any observed excess of derived alleles in the analyzed sample compared to an error free genome therefore must be due to errors in the analyzed sample. The method thus bases its error estimates on the excess in derived alleles in the analyzed sample compared to a given error free genome. Since no genomes are entirely error free, the estimates in practice become estimates of excess error compared to a high-quality genome, i.e., relative error rates. However, if a high-quality genome is used, the estimates should be close to absolute error rates (for more detail about the method see Orlando et al., 2013).

We used the chimp genome from the hg19 multiz46 alignment as an out-group for assessing what alleles are derived. As the high-quality genome, we used NA12778 low-coverage genome from the 1000 Genomes Project's May 2013 release. Before the analyses, we filtered the data from this genome so only reads with MapQ > 35 and bases with baseQ > 35 were used to ensure that the data used were indeed of high-quality. Also, we only included data from genomic positions where there is coverage on both the chimp, the sample of interest, and the high-quality genome. We used all reads from the samples of interest as opposed to just sampling one per site. However, before performing the estimation we filtered away reads with MapQ < 30 and bases with baseQ < 20 to mimic the filtering we have used in the many of the analyses performed in this study.

None of the base type specific error rates stand out for any of the individuals and all overall estimates are below or equal to 0.05% suggesting that all the genomes have low error rates and can be used in downstream analyses.

Relatedness estimation

Several of the analyses we performed in this study are based on an assumption that the individuals analyzed are not closely related. We therefore performed analyses to infer to what extent the sequenced individuals are related. More specifically, we estimated the three relatedness coefficients k0, k1, and k2 for each pair of individuals within each of the 2 populations and based on these we made a list of individuals to exclude from analyses of individuals that are not closely related. The coefficients k0, k1, and k2 here denote the proportions of the genome where the pair of individuals analyzed share 0, 1 and 2 alleles identical by descent, respectively. Importantly, the less related a pair of non-inbred individuals is, the higher k0 is expected to be, with k0 being equal to 1 for completely unrelated individuals. We used k0 below 0.75 as a threshold for being closely related, corresponding to the expected value for k0 for first cousins.

Since many of the individuals are sequenced to quite low depth, we used the program NGSrelate (Korneliussen and Moltke, 2015) to infer k0, k1, and k2 for each pair. NGSrelate is a maximum-likelihood based program that allows inference of the three relatedness coefficients from genotype likelihoods instead of called genotypes and through that it takes into account the uncertainty of the true genotypes, which is inherent to low-depth sequencing data.

When running NGSrelate, we used standard EM for optimization, allowed up to 500,000 EM iterations for each pair and used a stopping criterion of 1e-12 difference in likelihood between two consecutive EM-iterations. Furthermore, we ran each analysis 10 times with different random number seeds to be able to assess convergence of the EM-algorithm. For all pairs, the difference in log likelihood between the 10 NGSrelate solutions was less than 0.002, suggesting convergence was reached.

Since several of the individuals reported themselves to be admixed and NGSrelate, like most relatedness estimation method, assume the individuals are not admixed, we ran the analyses both of the full dataset and of the subsets consisting of the 49 Bajau and 23 Saluan individuals that were *self-reported to be unadmixed* (here termed "unmixed"). We used the latter to check to what extent the potential admixture affected the estimates of the former dataset.

Once we had the estimates of the relatedness coefficients, we identified all pairs of close relatives defined as pairs with a k0 estimate below 0.75. Next, we iteratively removed the individual with the highest number of close relatives until no pair of close relatives was left.

For these analyses, we only included data from autosomal sites that overlapped with the PanAsia dataset, described further below. For each of the two populations we used ANGSD (Korneliussen et al., 2014) to estimate allele frequencies and genotype likelihoods from reads with MapQ < 30 and bases with baseQ < 20 for the selected sites. This was done using the allele information from the SNP chip (-doMajorMinor 3) and with the SAMtools genotype likelihood model (-gl 1). For each of the two populations, we performed the allele frequency estimation and genotype likelihood estimation for both for all samples from the population and for the subset of "unmixed" individuals.

The results from the analyses of all individuals are shown in Figure S2. These estimates revealed numerous close familial relationships including parent-offspring and full siblings. Our subsequent filtering approach resulted in a list of samples to remove from analyses that should not contain data from closely related individuals. This list contains 16 Bajau and 1 Saluan individuals.

Population genetics analyses

To investigate the population affinities of Bajau and Saluan individuals in a broader context of regional populations, we analyzed them together with published SNP array dataset of 79 Asian populations (Ngamphiw et al., 2011). Since our study individuals are a mixture of both low and high coverage individuals, we did not attempt to call diploid genotypes, but rather represented each individual by a randomly sampled nucleotide from a high-quality read at each of the 50,796 autosomal SNP position in the panel. This is an established approach for genetic ancestry analyses on datasets with low genomic coverage and/or quality (e.g., ancient DNA studies (Allentoft et al., 2015; Skoglund et al., 2012).

We also carried out analyses of genetic drift sharing using outgroup f_3 statistics, using the HapMap Yoruban individuals as outgroup. The pairwise matrix of f_3 sharing was then converted into a distance matrix by calculating 1- f_3, after normalizing all values to the range [0,1], followed by multidimensional scaling using the 'cmdscale' function in R (R Development Core Team, 2008). Results are shown in Figures S2 and S3.

Principal component analysis (PCA) was carried out using the '–pca' option in PLINK (Chang et al., 2015). The first 2 PCs are shown in the main text, additional PCS are shown in Figure S3.

We used the Ohana tool suite (Cheng et al., 2016) to infer global ancestry and the covariance structure of allelic frequencies among populations. We analyzed a range of values of *K*, where *K* is the number of ancestry components. For each value of *K*, we used 32 independent executions with different random seeds, and we only report the ones that reached the best likelihood for each *K* (see Figure S4).

To estimate demographic model parameters, FastSimCoal requires the user to specify the topology of the model; in our case, we assume that Bajau and Saluan share an ancestral population of constant effective population size, which at some point split into the modern Bajau and Saluan demes. We also assume these demes each grew or contracted exponentially up to the date of sampling, and we allow migration between these two demes at constant and potentially asymmetrical rates.

As input, FastSimCoal takes the observed joint site frequency spectrum (jSFS) of the two populations. In practice, FastSimCoal may produce unstable estimates of demographic parameters due to the large space of parameters for which the observed jSFS is likely. To combat this, as is recommended for best practice, we ran FastSimCoal 100 times and chose the model with the highest likelihood.

From f_3 statistics, we find that both Bajau and Saluan populations show similar drift sharing profiles, exhibiting highest sharing with Austronesian-speaking populations from Taiwan (Ami, Atayal; Figure S2). This corresponds to previously reported shared ancestry as a consequence of the 'Austronesian expansion', which likely originated from Taiwan (Lipson et al., 2014). Genetic clustering based on pairwise f_3 values places both Bajau and Saluan close to other Austronesian-speaking populations from Indonesia and the Philippines. As seen in the PCA (Figure 2, additional PCs in Figure S3), Indonesian populations are arranged on a cline between Taiwanese Austronesian populations and Papuans, with populations in closer geographic proximity to Oceania also showing greater genetic affinity. We also observe a subtle shift of some Indonesian populations including the Bajau toward mainland Southeast Asian populations. Results from the admixture analysis (Figure S4) suggest that this reflects low levels of shared ancestry with Austro-Asiatic speaking languages, which has previously been reported in another study (Lipson et al., 2014).

Our results from *FastSimCoal* indicate that the most compatible model has an ancestral effective population sizes of 48670. Assuming an average generation time of 25 years, the estimated divergence time is 16.1 kya. Following divergence, the modern effective population sizes are estimated to have contracted/grown at rates of −0.009%, +0.040% in Bajau and Saluan, respectively, such that their modern effect population sizes are 51267 and 405050, respectively. We also estimate highly asymmetrical migration rates, suggesting a higher per-generation migration rate out of Bajau to Saluan than vice versa (5.48, 1.19e-4, respectively, in population-scaled units of $4N_{Baj}m$). Simulating the jSFS under the chosen model, we confirm the model is compatible with the data. We simulated 1000 1Mbp regions with mutation and recombination rates of μ = 2.5e-8, r = 1.15e-8 per generation per nucleotide using *MSMS.*

Testing for a difference in spleen size between Bajau and Saluan

Before performing a selection scan, we first wanted to test if there is a difference in spleen size between the Bajau and Saluan in order to determine if this is an appropriate target for our selection investigation. As an exploratory test of the data, we performed a Welch two-sample t test comparing the spleen sizes of the Bajau to those of the Saluan. This yielded a p value of 3.538e−07, indicating the spleens of the Bajau are significantly larger than those of the Saluan. Importantly, we also performed the Welch two-sample t test within the Bajau, comparing divers with non-divers, and we found no statistically significant difference in spleen size (p value: 0.2663). This result suggests that the observed difference in spleen size between Bajau and Saluan is not attributable to a plastic response of the spleen to diving. However, other factors than whether the individuals are divers may affect the results of the Welch two sample t test. We therefore proceeded with a linear model, which allows us to account for additional factors, like age and gender as well.

First, we used the R function *lm* to fit a linear model of the form

$$y = a + \beta_{pop}\, population + \beta_{gender}\, gender + \beta_{diver}\, diver + \beta_{age}\, age + \beta_{weight}\, weight + \beta_{height}\, height \quad (1)$$

where *y* is the spleen size, α is the intercept, *population* is a binary indicator, which takes the value 0 for Saluan individuals and 1 for Bajau individuals and *diver* is a binary indicator of whether the individuals are divers.

We did this with the goal of testing if β_{pop} is significantly different from 0, which would indicate that the Bajau and Saluan have different spleen sizes when taking into account gender, age, height, weight and whether the individuals are divers. We fitted the model using two different datasets to be able to assess if the results are affected by admixture:

Dataset 1: all samples that are not closely related
Dataset 2: dataset 1 with samples estimated to have above 5% admixture removed

After fitting the model using these two datasets we checked if the residuals were normal distributed by making QQ-plots and running a Shapiro-Wilks test of normality, because the p values provided by *lm* are based on the assumption of the residuals being normal distributed.

Because the Shapiro-Wilks test led to the rejection of the null hypothesis of the residuals being normal distributed for the first of the two datasets, we did not base our conclusions on the p values provided by *lm.* Instead we performed permutation tests to achieve p values using the function *lmp* in the R package lmPerm. When using this function, we set the stopping parameter *Ca* to 0.001 and

the *maxIter* parameter to 10^9, meaning that the permutation sampling was set to stop when the estimated standard deviation fell below 0.1 of the estimated p value, or 10^9 permutations had been sampled (the criterion reached first is used). In practice, the function stopped before reaching 10^9 iterations in all our analyses.

Second, we performed the exact same analyses using estimated Bajau ancestry proportions as the *population* predictor to explore how this would affect the results.

The results of fitting the model in Equation 1 with the two datasets described in the data section can be found in Table S1. This includes p values for the null hypothesis of β_{pop} being 0, which is provided by the standard *lm* function in R (''lm p-value'' in Table S1). However, we note that both QQ-plots and the results of the Shapiro-Wilks test for normality suggest that the residuals for the model fitted with data from all not closely related individuals are not normal distributed (Figure S1). This means that one has to be careful using the p values provided by *lm*, since these are based on an assumption of the residuals being normal distributed. We therefore also performed permutation testing of the same null hypothesis (''permutation p-value'' in Table S1) and used that as a base for our conclusions instead.

This approach resulted in a β_{pop} estimate significantly above 0 for both datasets, suggesting that the Bajau have a higher mean spleen size than Saluan even when correcting for gender, age, height, weight and whether the individuals are divers. And importantly, the fact that the result holds up for the subset of the individuals with less than 5% admixture, suggests that the results achieved from all not closely related individuals is not simply an artifact of admixture.

The results of fitting the model in Equation 1 with the two datasets using estimated Bajau ancestry proportion instead of a simply binary variable as population predictor can be found in Table S1. This includes p values for the null hypothesis of β_{pop} being 0, which is provided by the standard *lm* function in R (''lm p-value'' in Table S1). However, also for this predictor we note that both QQ-plots and the results of Shapiro-Wilks test for normality suggest that the residuals for the model fitted with data from all not closely related individuals are not normal distributed (Figure S1). We therefore also performed permutation testing of the same null hypothesis (''permutation p-value'' in Table S1).

Importantly, this approach resulted in a β_{pop} estimate above 0 for both datasets, but only significantly so (p value < 0.05) for the subset of the individuals with less than 5% admixture. For the larger admixed dataset that includes admixed individuals, the p value is slightly above 0.05. However, we note that admixture reduces the power to detect population differences if they are caused by genetics.

Bajau selection analysis using Ohana

The original selection analysis dataset contained 153 samples: 59 Bajau, 34 Saluan, and 60 CHS (Han Chinese) samples from the 1000 Genomes project (Auton et al., 2015). After removing closely related individuals from the Bajau and Saluan sets, as described previously, the dataset contained 136 individuals: 43 Bajau samples, 33 Saluan samples, and 60 CHS.

We used ANGSD (Korneliussen et al., 2014) to estimate genotype likelihoods from reads with MapQ < 30 and bases with baseQ < 20. This was done using allele information inferred directly from likelihoods (-doMajorMinor 1) and the SAMtools genotype likelihood model (-gl 1). SNPs were filtered for deviations from Hardy Weinberg Equilibrium (-hwe_pval 5e-7), strand bias (-sb_pval 5e-7), and the Wilcox rank sum test for qscore bias (-qscore_pval 5e-9). The full dataset contained genotype likelihoods for 2,665,716 markers. After masking using the 1000 Genomes filter (Auton et al., 2015), it contained 2,333,499 markers resulting in a 12.5% reduction rate. We performed analyses both with and without the 1000 Genomes masking.

In annotating markers, we assigned RSIDs to physical genome locations by comparing to them to dbSNP build 147 [Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 147), 2016]. This dbSNP build contained 154,822,082 markers, of which 2,660,285 were among our 2,665,716 markers. 5,431 markers (0.2%) did not have assigned RSIDs, but none of these markers were identified in the selection scan as potential candidates for being targets of selection.

To estimate genome-wide admixture proportions and the correlation structure and population tree needed for the selection scan, we used the same dataset as for the selection scan. However, to reduce the computational burden in performing these analyses, we down-sampled the dataset from the full 2,333,499 markers to 100,000 randomly chosen markers.

The admixture and tree analyses were performed in the same manner as described earlier in the text, but using the Bajau, Saluan, and Han dataset described above. We performed the analyses both for $K = 2$ and $K = 3$, but only used the results from $K = 3$ in the downstream analysis (Figure S5).

We used the ''selscan'' program provided in Ohana to detect SNPs that deviate strongly in the Bajau from the genome-wide covariance structure using a likelihood ratio test. For each SNP we introduce a scalar variable that is multiplied onto the variance associated with the Bajau population. This establishes two nested likelihood models, one which assumes the Bajau specific allele frequency change is as predicted from the genome-wide covariance pattern, and one which allows a larger change in the Bajau and hence a higher variance in the Bajau component than expected from the genome-wide pattern (Figure S5). The resulting likelihood ratio test can be used to identify loci in which the Bajau population has experienced a larger than expected change in allele frequency as compared to the prediction from the genome-wide pattern, which is a signature of selection. This method is inspired by a number of similar, recently developed methods that use a Gaussian distribution as an approximation to model the distribution of allele frequencies among populations (Coop et al., 2010; Pickrell and Pritchard, 2012).

We removed signals driven by a single SNP that might be due to base-calling or mapping errors, and identified potential causal SNPs by only retaining SNPs for which: 1) the SNP does not have any neighbor within ± 100kb (200kb window) that has a higher likelihood ratio; 2) the SNP is not the only one within ± 100kb (200kb window) that is among the most extreme 1% likelihood ratios. Two example peaks from the results are shown in Figure S5.

Functional annotation of selected SNPs

We assessed how potentially disruptive each of the top 25 SNPs identified in the selection scan were through a query of the Combined Annotation Dependent Depletion online tool (CADD v1.3) (Kircher et al., 2014). We then focused on the top 5 most disruptive SNPs to determine the nature of their disruptive score by manually inspecting their annotations and visualizing their corresponding ENCODE segmentation tracks (ENCODE Project Consortium, 2012) in the Washington University Epigenomics Browser (v.44.1).

The CADD score for each SNP is listed in Table 1 in the main text. The SNP with the largest CADD score (22.8, PHRED-scaled), rs10483896, lies in a Segway repressor region (Hoffman et al., 2012) upstream of NRXN3 and is expressed preferentially in brain tissues. The position is highly conserved across primates, mammals, and vertebrates (PhastCons scores excluding human genome = 0.98, 1 and 1, respectively (Siepel et al., 2005)) with a high GERP++ rejected substitution score (5.26) (Davydov et al., 2010). The SNP with the second highest CADD score (12), rs77280170, is in an intergenic region. It is not highly conserved and lies in a heterochromatic region of low regulatory activity. The SNP rs118149708, with the third highest CADD score (9.838), is in a Segway repressor region in an intron of RP11-499F3.2. The SNP with the fourth highest CADD score (8.527), rs16030, is a highly conserved synonymous change that also lies inside a transcription factor binding site in an open chromatin region. It is located in the gene CACNA1A. Visual inspection in the Epigenomics Browser suggests the site is in an enhancer region that overlaps an exon of the gene. Finally, the SNP rs28544477, with the fifth highest CADD score (6.521), is in a regulatory feature (Segway FAIRE-seq region) but is not highly conserved.

Spleen size association testing

Initially, we tested for association between spleen size and each of the 25 SNPs with the highest score in the selection scan using the score statistic based method by Skotte et al. (2012) as implemented in ANGSD. We chose this method for two reasons. First, it bases its test on posterior genotype probabilities and uses these to take into account the uncertainty of the genotype, which is inherent to low-depth sequencing data. Second, this method is based on a generalized linear model framework. It therefore allows inclusion of covariates in the tests and thus makes it possible to correct for potentially confounding factors, like admixture.

When performing these score statistics based tests, we assumed an additive effect model and included the five first principle components (PCs) as covariates to correct for admixture and other population structure. Additionally, we included age, gender, height, weight, sequencing depth, and whether the individuals are divers as covariates. Hence we assumed the linear model:

$$y = a + \beta_{genotype}\, genotype + \beta_{gender}\, gender + \beta_{diver}\, diver + \beta_{age}\, age + \beta_{weight}\, weight + \beta_{height}\, height + \beta_{sequencing\ depth}\, sequencing\ depth + \beta_{PC1}\, PC1 + \beta_{PC2}\, PC2 + \beta_{PC3}\, PC3 + \beta_{PC4}\, PC4 + \beta_{PC5}\, PC5$$

where y is spleen size and *genotype* is coded as 0, 1, or 2 corresponding to the number of copies of the selected allele carried by an individual. The test we performed for each of the 25 SNPs was whether $\beta_{genotype}$ is significantly different from 0. When assessing significance of the test results, we used a Bonferroni corrected p value threshold of 0.05/25 = 0.002.

We note that an underlying assumption in model used is that the residuals are normal distributed. Unfortunately, ANGSD does not provide any possibility to investigate whether this assumption is violated. Therefore, we performed the tests not only using the raw spleen size measurements, but also using the spleen size quantile transformed to a normal distribution and on the spleen sizes log-transformed to be as sure as possible that any potential association signal is not simply caused by a violation of the underlying normality assumption.

To follow up on the results from the initial tests, we performed two additional analyses, both with the purpose of investigating whether the initial results are affected by admixture (despite the fact that we included the first five PCs as covariates). We did this because it is well known that admixture and population structure in general can lead to false positives.

First, we applied the same test to a large number of additional SNPs from across the genome and based on the p values from these tests we calculated genomic control inflation factors to assess if there is a general inflation in the –log(P)-values due to population structure.

Second, we performed association tests very similar to the initial tests with the only difference being that we did not include any PC as covariates. Instead, we employed genomic control to correct for population structure. We did this to investigate if the results are robust to the approach of correcting for population structure. We based the genomic control correction on results from the same set of additional SNPs as in the previous analysis.

All previous score statistic based analyses provided evidence of an association between spleen size and one of the tested 25 SNPs. To further investigate this signal, we used the same linear model as in the initial analyses but this time fitted using the *lm* function in R based on imputed data. We did this using the following subsets of the data:

1) The same individuals as in the initial analyses (purpose: to investigate if the association signal from the initial analyses is an artifact of using the score test method)

2) The subset of individuals that are Bajau, excluding the one individual that, according the admixture analyses, has very little (if any) Bajau ancestry (purpose: to investigate if the association result from the first analyses is an artifact of the fact that the Saluan individuals were included in the initial analyses)
3) Three subsets of the Bajau individuals with different, limited amounts of non-Bajau ancestry (purpose: to further ensure the association signal is not an artifact of population structure unaccounted for in prior analyses)

Importantly, all of these *lm* analyses also gave us a chance to investigate the direction of the effect, which is not possible using the score-based test.

After fitting the models, we checked if the residuals were normal distributed by running a Shapiro-Wilks test of normality, because the results provided by *lm* are based on the assumption of the residuals being normal distributed.

All the score statistic based tests were applied to data from both Bajau and Saluan individuals. Before performing the tests all close relatives were filtered away, which left us with 43 Bajau and 33 Saluan with spleen size measurements. All the imputation based tests were applied to data from these same individuals and subsets of these, based on their amount of estimated Bajau ancestry.

The analyses on which we based our relatedness filtering are described previously, as well as the analyses on which we based our admixture filtering.

All the score statistic-based tests were performed on posterior genotype likelihood, which were estimated using ANGSD (Korneliussen et al., 2014) with the SAMtools genotype likelihood model (-gl 1), and with major and minor alleles were inferred from the genotype likelihoods (-doMajorMinor 1). Only reads with MapQ > 35 and bases with baseQ > 35 were used to ensure that the data used were indeed of high-quality. Sites were restricted to those deemed accessible according to the 1000 Genomes accessibility mask.

For the imputation-based tests, we imputed data for the SNP of interest using the same input and settings as for the score statistic-based tests.

For phenotypes, we used the spleen size, weight, and height measurements obtained at the same time as DNA material was sampled. For age, gender, and diver (yes/no) we used values obtained through an interview also performed at the same time as the DNA material was sampled.

The PCs were calculated using genetic data from the individuals included in the association tests only.

We tested for association between spleen size and each of the 25 top SNPs from the selection scan using a score statistic based test. One SNP, rs3008052 in the gene PDE10A, showed evidence of being associated (Table 1 from the main text). Notably, this result is consistent across three different transformations of the spleen size measurements, suggesting that the result is not simply an artifact of a potential violation of the normality assumption underlying the method used.

In the initial tests, we included the five first PCs as covariates to correct for population structure. To investigate if the results we obtained from these tests were affected by population structure (despite this inclusion of PCs), we performed the same test on a large set of SNPs across the genome and, based on the p values from these tests, we calculated genomic inflation factors and made QQ plots (Figure S6). This results suggested that the –log(p values) are in general not markedly inflated due to population. Additionally, QQ-plots of the 25 top SNPs from the selection scan suggest that the p values for these do not in general seem to be inflated either, despite the fact that these SNPs are characterized by having very different allele frequencies in Bajau compared to other populations (Figure S6).

We also performed score statistic-based association testing using genomic control instead of PCs to correct for population structure. This gave rise to the even lower p values for rs3008052 (3.8E-05, 3.0E-05, and 7.2E-05 for Untransformed, Qnorm transformed, and log transformed data, respectively). Hence, all in all, we did not detect any evidence that the observed association signal is an artifact of population structure.

To further investigate the signal of association between spleen size and rs3008052, we imputed genotypes for rs3008052 and analyzed these for several subsets of the data: the exact same dataset as in the initial, the subset consisting only of Bajau individuals, and three different subsets of these Bajau individuals with only limited amounts of non-Bajau ancestry.

First, we plotted the spleen sizes stratified by imputed genotype for all the datasets and all the transformations of the spleens sizes used in the initial tests. Visual inspection of these plots for all datasets showed a clear trend that, the more copies of the selected allele (T), the larger the spleen size, and thus an association between rs3008052 and spleen size in accordance with the initial results.

Next, we tested if these associations are statistically significant using the exact same model as in the initial score statistic based tests, but this time using the R function *lm* to fit the model and perform the tests. We note that for these analyses the residuals were in no case rejected to be normal distributed, suggesting that the underlying assumption of normality of *lm* is not evidently violated.

When performing these tests using data from the same individuals as in the initial score statistic based analyses, we got even lower p values than those obtained using the score statistic based test, suggesting that the initial association signal was not simply an artifact of using the score statistic based framework (Table S2). Similarly, when performing the test based only on the subset of individuals that are Bajau, we also get highly significant p values despite the much smaller sample size (n = 43 as opposed to n = 76), which suggests that the initial association signal is not an artifact of including Saluan individuals in the initial analysis either (Table S2).

When performing the test of three different subsets of the Bajau individuals with only limited amounts of non-Bajau ancestry, the results were more mixed. For the largest subset consisting of the 50% of the Bajau individuals with least non-Bajau ancestry (where all individual are estimated to have < 16% non-Bajau ancestry) there was still a marginally significant p value, despite a sample size of only 21. But this was not the case for the even smaller datasets consisting of Bajau individuals with less than 10% and 5% non-Bajau

ancestry. However, these two datasets are quite small (n = 18 and n = 15, respectively) and the effect sizes estimates obtained from them are consistent with the estimates obtained from all the larger datasets (main text Figure 4). Hence it seems likely that the lack of significance for these two datasets is due to their limited size, rather than a sign that the signals association signal observed in the larger more admixed dataset are caused by population admixture not corrected for by the PCs.

It is also worth noting that the effect size estimates are all consistent with a positive effect on spleen size of the selected allele (T) i.e., having more Ts is associated with having larger spleen size.

Thyroid hormone association testing

Using data from the 500FG cohort, we tested for association using a linear regression technique as implemented in the Matrix eQTL package (Shabalin, 2012). The method was applied to imputed ''dosage'' SNP data, where in imputation the genotypes are not discretely called, but rather values between 0 and 2 are given to each individual. Related and non-European ethnicity individuals were removed, and the tests were corrected for age, gender, BMI, and oral contraceptive usage by including these factors as covariate.

Thyroid hormone concentrations were scaled using an inverse rank transformation (IRT) algorithm, similar to the one applied in Aguirre-Gamboa et al. (2016). We examined four SNPs: the lead SNP, rs3008052, and three other SNPs in high LD in Europeans with rs3008052 (rs2983527, rs3008050, and rs3008049). All of these SNPs have significant associations with hypothyroidism in the Global Biobank Engine data.

We found clear, significant association for our investigated SNPs with the allele at higher frequency in the Bajau significantly associated with elevated T4 circulating plasma concentrations. Results for all tested SNPs are found along with boxplots illustrating the directionality of the associations in Figure S7.

Estimating the strength and timing of selection

For both the PDE10A and BDKRB2 alleles, we estimate the strength, timing, and frequency of the allele at the divergence time (*s, t, f,* respectively) by simulating joint allele frequency trajectories in Bajau and Saluan using the model inferred previously. We assume the allele to be segregating prior to the divergence time due to the presence of this allele in a number of other human panels, such as Han Chinese, whose common ancestors predate the Bajau-Saluan split.

Across a 3-D grid of values of these 3 parameters, at each gridpoint we simulate 1000 trajectories in *MSMS.* To estimate selection on a particular allele, we take a bin around its estimated allele frequencies in Bajau and Saluan, which we construct using a Clopper-Pearson 99% confidence interval on the allele's true population frequencies (*PDE10A*: Bajau 0.371 [0.24–0.515], Saluan 0.067 [0.013–0.188]; *BDKRB2*: Bajau 0.183 [0.089–0.311], Saluan 1e-05 [0–0.077]). As a heuristic for the likelihood of a particular value of (*s,t,f*), we calculate the proportion of simulations for which the trajectory lands in both the Bajau and Saluan bins.

For alleles simulated under neutrality (i.e., $s = 0$), we observe no hits in the bins for *PDE10A* under any combinations of (*t,f*), suggesting that the allele frequency discrepancy is unlikely under neutral evolution in the inferred demographic model ($p < 0.001$). For *BDKRB2,* the maximal probability across these neutral bins is 0.024, and thus we suggest that *BDKRB2*'s frequency discrepancy is unlikely under neutral evolution in the inferred demographic model ($p < 0.024$). Results are shown in main text Figure 6.

DATA AND SOFTWARE AVAILABILITY

The accession number for the newly generated genome data reported in this paper is European Genome-Phenome Archive: EGAS00001002823.

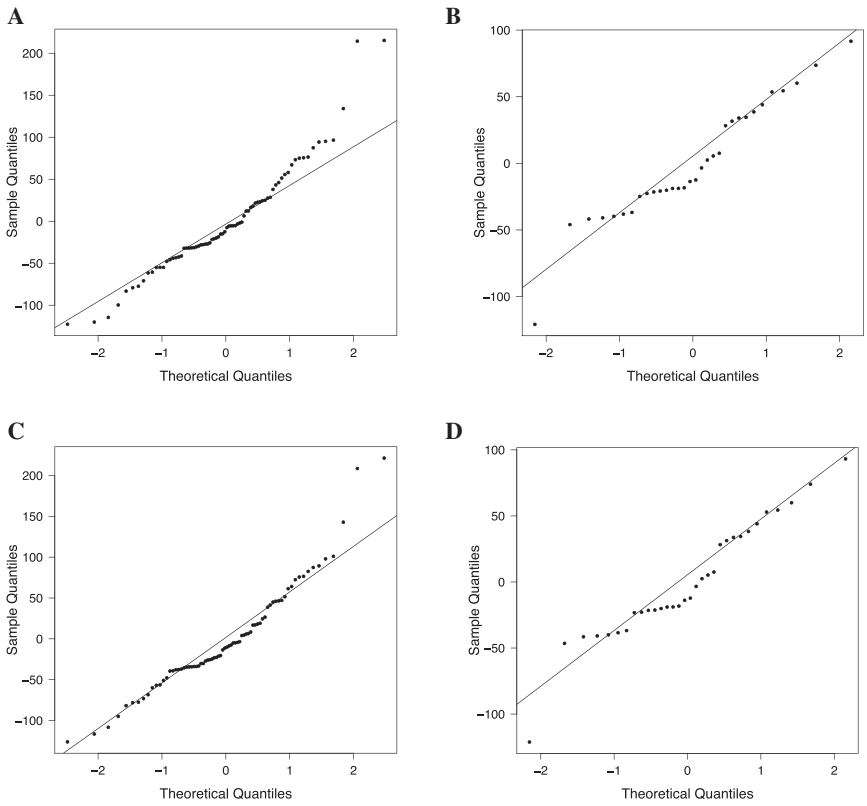


Figure S1. QQ-Plots for the Residuals from the Models Fitted with the *lm* Function, Related to Quantification and Statistical Analysis–Testing for a Difference in Spleen Size between Bajau and Saluan

(A–D) QQ plot for residuals obtained when fitting (A) the model in Equation 1 with dataset 1 and a binary population predictor, (B) the model in Equation 1 with dataset 2 and a binary population predictor, (C) the model in Equation 1 with dataset 1 and estimated Bajau ancestry proportion as population predictor, and (D) the model in Equation 1 with dataset 2 and estimated Bajau ancestry proportion as population predictor. The corresponding p values obtained from performing Shapiro-Wilks normality test are: 0.001879, 0.142, 0.001352 and 0.146, respectively.

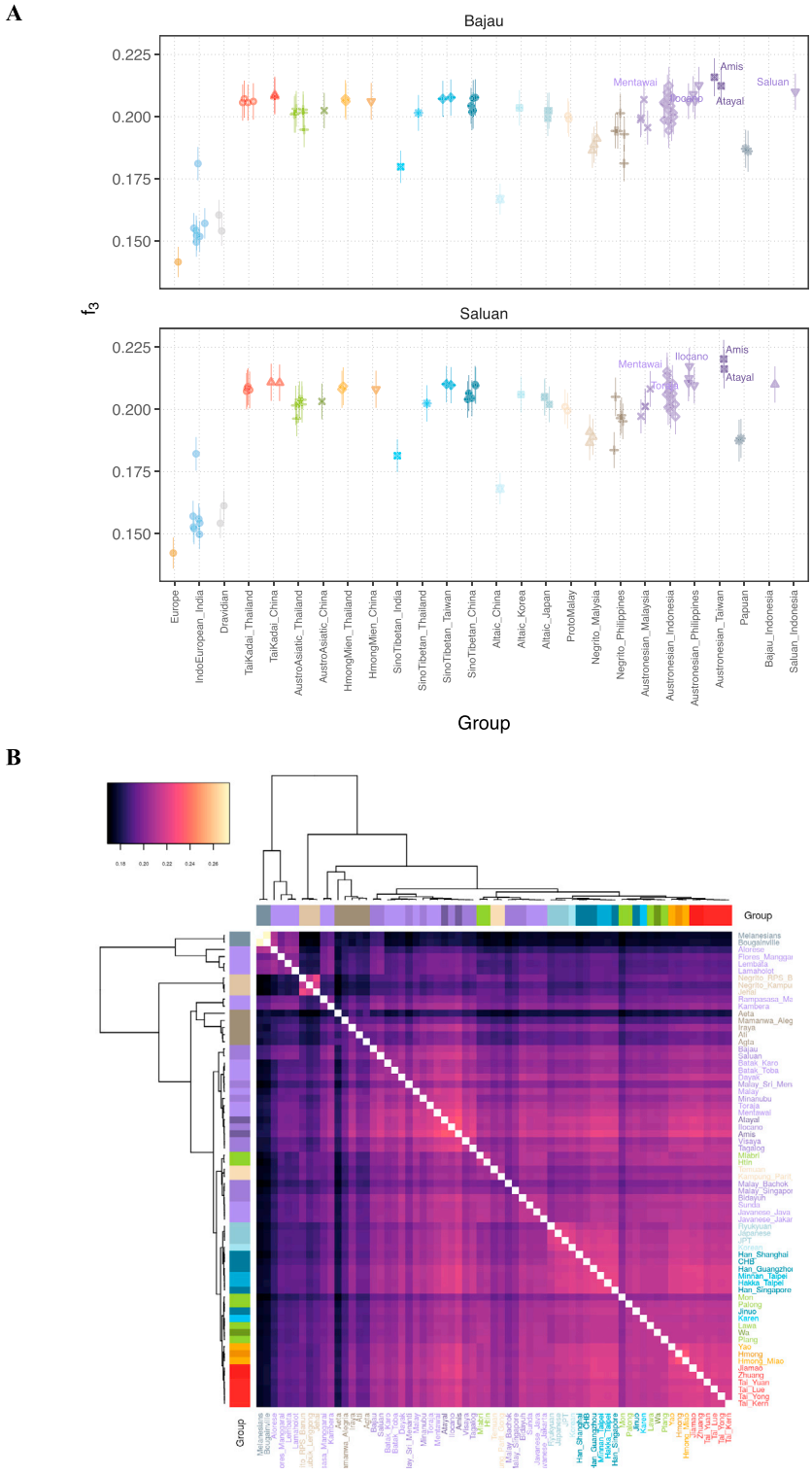


Figure S2. f_3 Statistics, Related to Quantification and Statistical Analysis–Population Genetics Analyses

(A) Distributions of outgroup f_3 statistics for Bajau and Saluan populations. Reference populations are grouped by linguistic family and geographical location. Labels in the panels indicate the top five populations with highest f_3 values.

(B) Heatmap of pairwise outgroup f_3 statistics: genetic drift sharing for all pairs of populations. Colored bars and labels indicate linguistic family of the respective population.

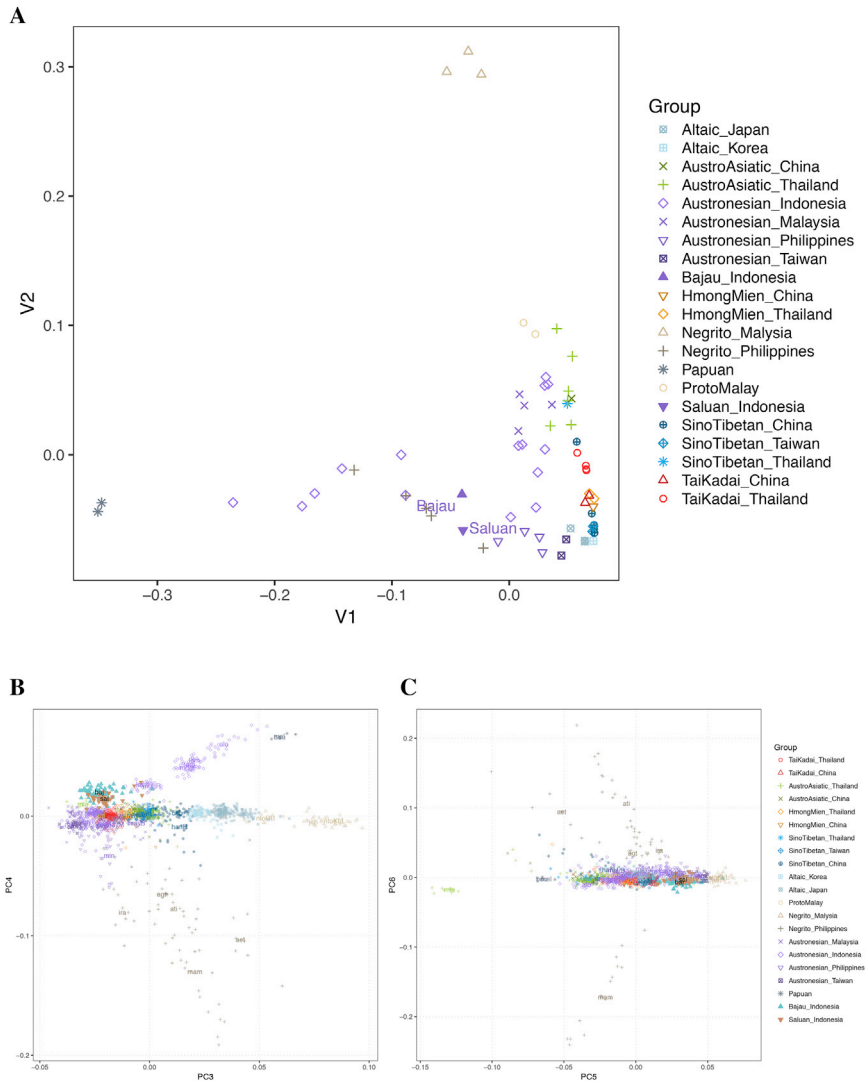


Figure S3. Bajau and Saluan Population Demography, Related to Quantification and Statistical Analysis–Population Genetics Analyses and Figure 2

(A) Multidimensional scaling of pairwise f_3. Plot shows the first two dimensions of an MDS based on the f_3 distance between pairs of populations.
(B and C) (B) Additional PCs in the PCA of Bajau, Saluan, and PanAsia populations, PCs 3 and 4, and (C) PCs 5 and 6.

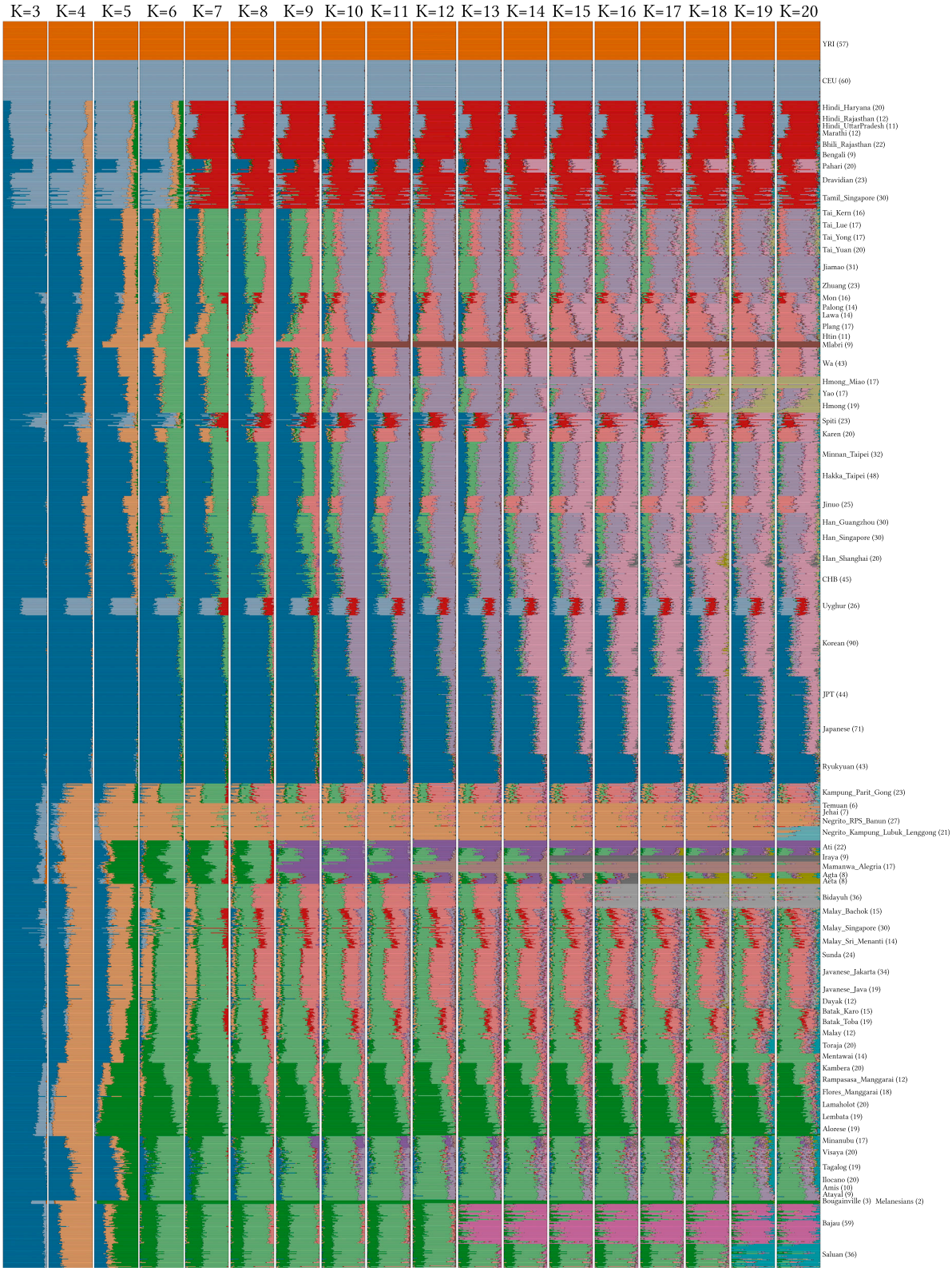


Figure S4. Pan-Asian Admixture Estimates, Related to Quantification and Statistical Analysis–Population Genetics Analyses and Figure 3
Component colors are matched across plots.

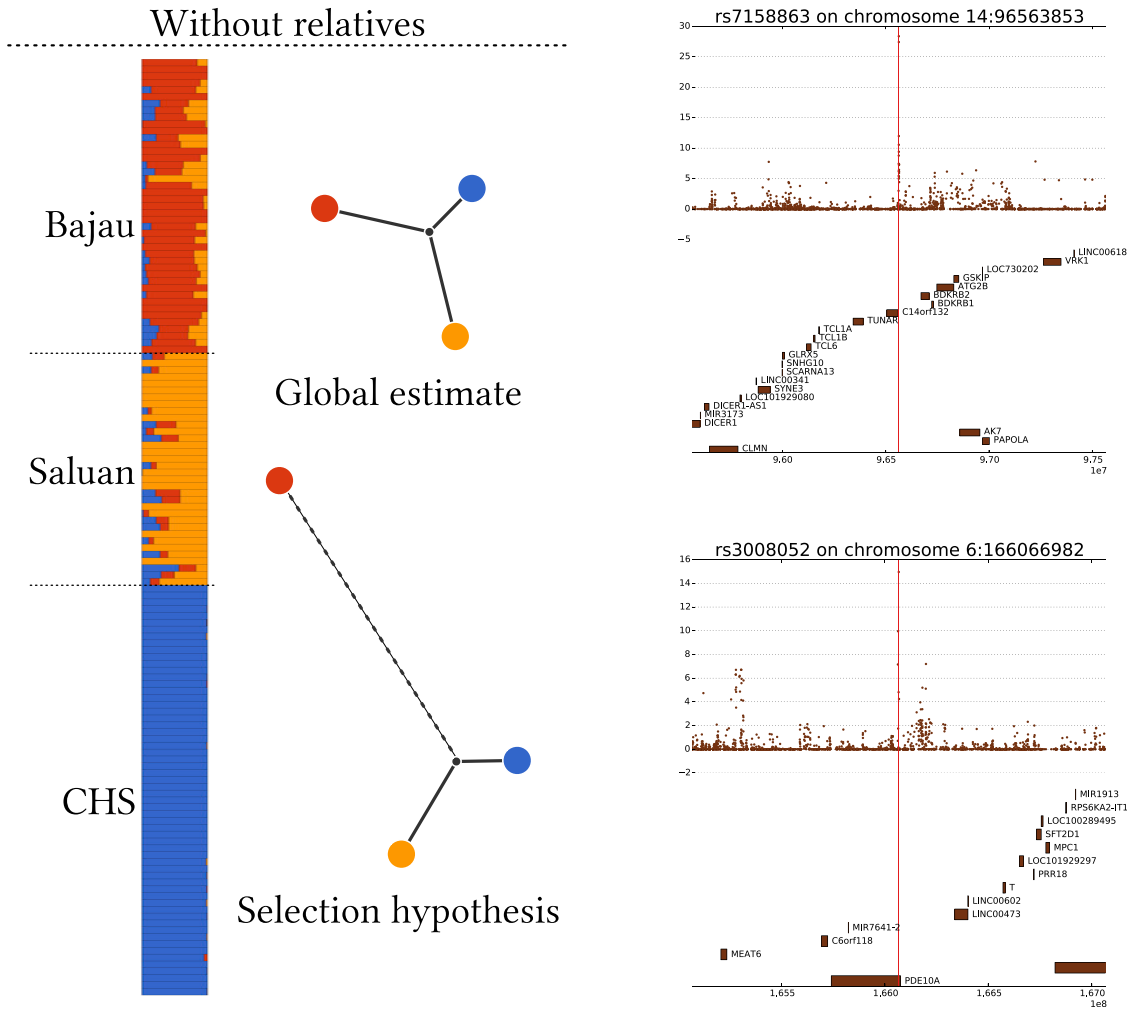


Figure S5. Selection Hypotheses and Example Resulting Peaks, Related to Quantification and Statistical Analysis–Bajau Selection Analysis Using Ohana

Using the selection suite from Ohana, selection hypotheses are constructed based on the structure and tree results from $K = 3$. The likelihood ratio test compares a model in which the Bajau component experiences faster allele frequency changes in a specific locus (selection hypothesis) than expected from the genome-wide distribution of allele frequency changes (global estimate). Two resulting peaks from the selection analysis are shown; the first peak, ranking #1, is located upstream of gene BDKRB2. The second peak, ranking #22, is located on the gene PDE10A.

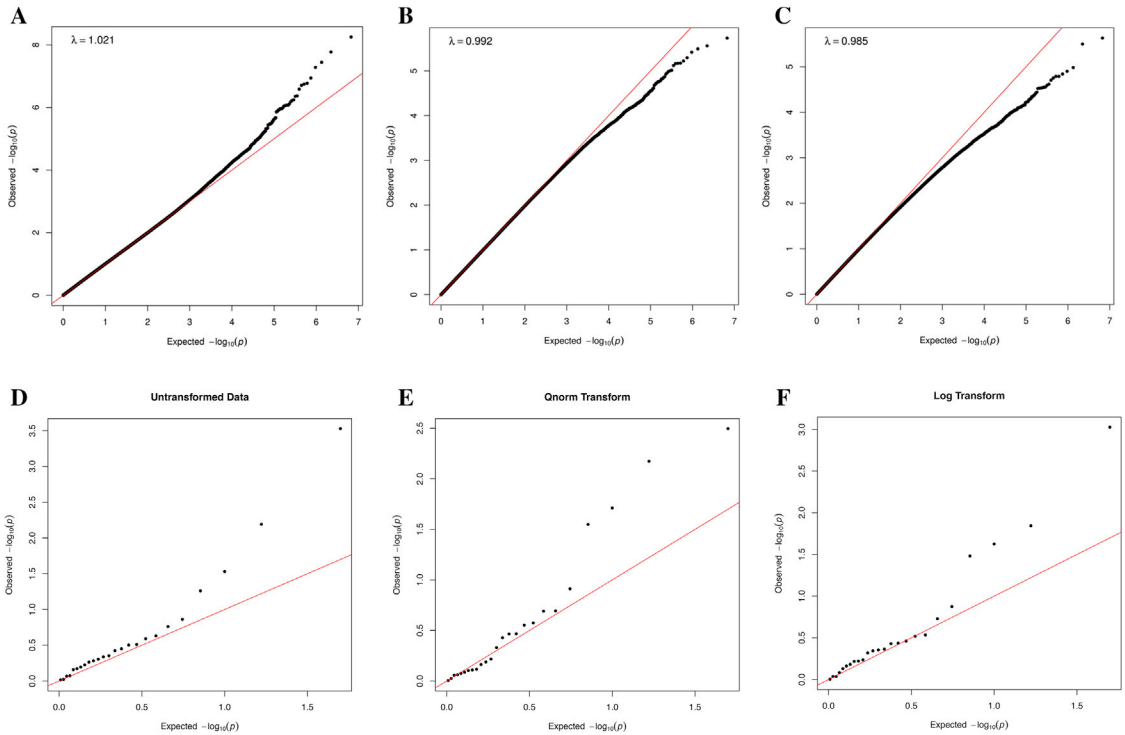


Figure S6. QQ-Plots for a Large Set of SNPs and the Top 25 SNPs from the Selection Scan, Related to Quantification and Statistical Analysis–Spleen Size Association Testing

QQ-plots using a score statistic based test and a model with the first five PCs included as covariates to correct for population structure.

(A) QQ-plot for testing all (1,536,467,431) SNPs for association with untransformed spleen sizes.

(B) QQ-plot for testing all SNPs for association with spleen sizes quantile transformed to a normal distribution.

(C) QQ-plot for testing all SNPs for association with log transformed spleen.

(D) QQ-plot for testing top 25 SNPs for association with untransformed spleen sizes.

(E) QQ-plot for testing top 25 SNPs for association with spleen sizes quantile transformed to a normal distribution.

(F) QQ-plot for testing top 25 SNPs for association with log transformed spleen.

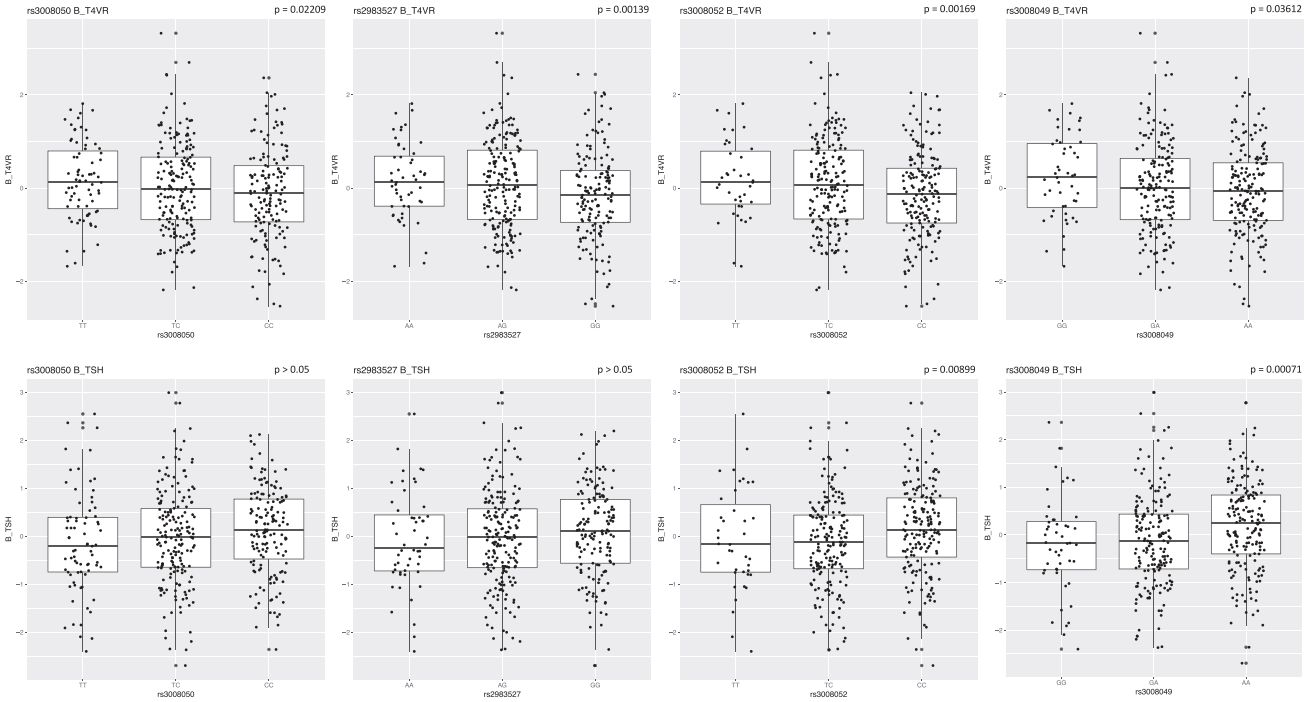


Figure S7. T4 and TSH Levels Stratified by Genotypes of the Lead PDE10A SNP (rs3008052) and the Three High Linkage Disequilibrium Proxy SNP Using Data from the 500FG Cohort, Related to Quantification and Statistical Analysis–Thyroid Hormone Association Testing

The allele favored in the Bajau is associated with elevated T4 circulating plasma concentrations and decreased levels of TSH.