Natural Selection Signatures in the Hondo and Ryukyu Japanese Subpopulations

Xiaoxi Liu,[1,2,‡] Masatoshi Matsunami,[3,‡] Momoko Horikoshi,[4] Shuji Ito,[1] Yuki Ishikawa,[1] Kunihiko Suzuki,[5] Yukihide Momozawa,[5] Shumpei Niida,[6] Ryosuke Kimura,[7] Kouichi Ozaki,[8] Shiro Maeda,[3,9] Minako Imamura,[3,9,*,§] and Chikashi Terao[1,2,10,*,§]

[1]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[2]Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan

[3]Department of Advanced Genomic and Laboratory Medicine, Graduate School of Medicine, University of the Ryukyus, Nishihara-Cho, Japan

[4]Laboratory for Genomics of Diabetes and Metabolism, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[5]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[6]Core Facility Administration, Research Institute, National Center for Geriatrics and Gerontology, Obu, Japan

[7]Department of Human Biology and Anatomy, Graduate School of Medicine, University of the Ryukyus, Nishihara-Cho, Japan

[8]Medical Genome Center, Research Institute, National Center for Geriatrics and Gerontology, Obu, Japan

[9]Division of Clinical Laboratory and Blood Transfusion, University of the Ryukyus Hospital, Okinawa, Japan

[10]The Department of Applied Genetics, School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan

***Corresponding author:** E-mail: mimamura@med.u-ryukyu.ac.jp; E-mail: chikashi.terao@riken.jp.

[‡]These authors contributed equally to this work.

[§]These authors co-supervised this work.

Abstract

Natural selection signatures across Japanese subpopulations are under-explored. Here we conducted genome-wide selection scans with 622,926 single nucleotide polymorphisms for 20,366 Japanese individuals, who were recruited from the main-islands of Japanese Archipelago (Hondo) and the Ryukyu Archipelago (Ryukyu), representing two major Japanese subpopulations. The integrated haplotype score (iHS) analysis identified several signals in one or both subpopulations. We found a novel candidate locus at *IKZF2*, especially in Ryukyu. Significant signals were observed in the major histocompatibility complex region in both subpopulations. The lead variants differed and demonstrated substantial allele frequency differences between Hondo and Ryukyu. The lead variant in Hondo tags *HLA-A*33:03-C*14:03-B*44:03-DRB1*13:02-DQB1*06:04-DPB1*04:01*, a haplotype specific to Japanese and Korean. While in Ryukyu, the lead variant tags *DRB1*15:01-DQB1*06:02*, which had been recognized as a genetic risk factor for narcolepsy. In contrast, it is reported to confer protective effects against type 1 diabetes and human T lymphotropic virus type 1-associated myelopathy/tropical spastic paraparesis. The FastSMC analysis identified 8 loci potentially affected by selection within the past 20–150 generations, including 2 novel candidate loci. The analysis also showed differences in selection patterns of *ALDH2* between Hondo and Ryukyu, a gene recognized to be specifically targeted by selection in East Asian. In summary, our study provided insights into the selection signatures within the Japanese and nominated potential sources of selection pressure.

Key words: **natural selection, Japanese, Hondo, Ryukyu, iHS, FastSMC, *ALDH2*.**

Introduction

Approximately 100,000 yr ago, anatomically modern human populations started moving out of Africa and initiated a global migration (Klein 2008). The novel environments, which might be markedly different from their African origin, are believed to have driven selection for new traits important for human survival, and have left discernible signatures within the human genome. With the advent of high-throughput genotyping technology and analytical methodologies, dozens of genetic loci have been linked to the natural selection (Benton et al. 2021). These loci reflect adaptations to various challenges such as high-altitude (Yang et al. 2017), novel pathogens (Kwiatkowski 2005), changes of food sources (Mathieson and Mathieson 2018), etc. Detection of positive selection for each local population not only greatly enhances our understanding of the adaptive evolution, but also has important medical implications (Vasseur and Quintana-Murci 2013).

Received: March 27, 2023. **Revised:** September 20, 2023. **Accepted:** October 06, 2023

In the Japanese population, genetic loci harboring *ALDH2* (Oota et al. 2004), *EDAR* (Fujimoto et al. 2008; Kimura et al. 2009), and major histocompatibility complex (*MHC*) region (Kawashima et al. 2012) were reported to be under positive selection based on the candidate gene approach. By applying the singleton density score method to 2,234 Japanese whole-genome sequencing (WGS) data, a genome-wide scan of very recent selection signature detected four loci including the ADH cluster, MHC region, *BRAP-ALDH2*, and *SERHL2* (Okada et al. 2018). A recent study using large-scale microarray data of 170,882 subjects from the Biobank Japan (BBJ) reported 29 candidate loci by the ascertained sequentially Markovian coalescent (ASMC) method, and two loci including ADH cluster and MHC by the integrated haplotype score (iHS) method, which remarkably expanded the number of candidate genes subjected to natural selection (Yasumizu et al. 2020). In the context of recent advances, several critical issues have not been adequately addressed yet. First, the principal component analysis (PCA) demonstrated a dual population structure in the Japanese: the Hondo (literally translated as main-islands) cluster on the Japanese Archipelago and the Ryukyu cluster on the Ryukyu Archipelago (Yamaguchi-Kabata et al. 2008; Sakaue et al. 2020). The different peopling histories of Hondo and Ryukyu populations, leading to varying levels of admixture between neolithic Jomon and Yayoi ancestral groups, have been proposed to underlie this unique population structure (Japanese Archipelago Human Population Genetics Consortium 2012; Bendjilali et al. 2014; Koganebuchi and Kimura 2019) (Supplementary Notes S1 and S2). Genetic differentiation was observed among island groups of the Ryukyu Archipelago, and there is little genetic affinity between aboriginal Taiwanese and any of the Ryukyu people despite the geographical proximity (Sato et al. 2014). In the essential context of Japan's dual population structure, it is worth noting that previous studies, though not primarily focused on the Japanese population but including it in their scope, have predominantly relied on individuals recruited from Hondo (Voight et al. 2006; Sabeti et al. 2007; Johnson and Voight 2018). Moreover, those studies that do focus on the Japanese did not distinguish adequately between Hondo and Ryukyu (Okada et al. 2018; Yasumizu et al. 2020). Large-scale analyses specifically targeting the Ryukyu subpopulation are notably absent (Liu et al. 2017). This represents a substantial knowledge gap in our understanding of natural selection within the Japanese population. Second, there is geographic variation in disease prevalence, pathogen subtype, and infection rate of pathogens between Hondo and Ryukyu (Hayashi et al. 1990; Fukiyama et al. 2000; Aoki et al. 2008; Takeuchi et al. 2014). For example, strains of *Helicobacter pylori*, which are associated with the development of gastric cancer, differ between Hondo and Ryukyu; the strains specific to Ryukyu are less virulent (Suzuki et al. 2022). Another example is a notably higher prevalence of Human T lymphotropic virus type 1 (HTLV-1) infection in Ryukyu than in Hondo, which can result in a range of clinical manifestations (Morikawa et al. 1988; Watanabe 2011; Iwanaga 2020). It would be intriguing to examine whether there are any differences in selection profiles between Hondo and Ryukyu, for which one possible explanation is provided by varying environmental factors. Third, replication studies to confirm the previously reported signals remain much lacking. Finally, studies with DNA microarrays designed to contain more East Asian-specific probes hold the promise to capture novel signals and may further advance our understanding of the natural selection signatures in the Japanese (Kawai et al. 2015).

To address the above issues, we carried out genome-wide selection scans using data of 20,366 individuals who were recruited from both Hondo and Ryukyu and had been genotyped on the Illumina Infinium Asian Screening Array (ASA), an array specifically designed to contain East Asian-specific variants. In order to make our results comparable with the previous study (Yasumizu et al. 2020), two methods: iHS (Voight et al. 2006) and FastSMC (Nait Saada et al. 2020) were applied. The iHS detects recent and ongoing selection signals, such as soft sweep based on phased haplotype information. Additionally, FastSMC offers an alternative approach for identifying candidate regions potentially targeted by selection through the inference of identity-by-descent (IBD) sharing (Palamara et al. 2018; Nait Saada et al. 2020).

Results

Fine-scale Genetic Structure of the Japanese Population, Especially the Ryukyu Cluster

A total of 20,366 individuals from two cohorts in Japan were analyzed in this study. The first cohort consisted of 13,753 participants at the National Center for Geriatrics and Gerontology (NCGG) Biobank in Japan (Shigemizu et al. 2021). The second cohort consisted of 6,613 participants who were recruited at Okinawa Prefecture through the Okinawa Bioinformation Bank (OBi) Project in which detailed geographical information of origins (including islands in Okinawa Prefecture) are available in some participants (Matsunami et al. 2021) (Supplementary Table S1; Supplementary Note S2). We merged the two cohorts and a total of 622,926 SNPs were available for analysis. For quality control (QC) (see Methods), we first removed 65,826 variants that have a call rate less than 97% or a Hardy–Weinberg equilibrium *P*-value (HWE-P) $< 1 \times 10^{-6}$. Then we removed 745 samples with a sample call rate of less than 97%. Furthermore, we excluded 1,585 closely related individuals with a shared IBD ($\hat{\pi}$) >=0.25. Based on the PCA, we additionally removed 45 samples whose PC1 or PC2 showed 3 standard deviations (SD) from the mean PC values, 25 samples overlapping with the Chinese cluster, and 67 samples overlapping with the Korean cluster. Because deviations from HWE can occur when examining data comprising distinct subpopulations, we conducted additional analyses after the PCA–UMAP (Uniform Manifold Approximation and Projection) (shown in a later section). We recalculated HWE for SNPs that failed the QC, separately for each subgroup. Through this approach, we were able to recover 3,033 variants that only significantly deviated from HWE when the whole dataset was considered. The final dataset consisted of 17,932 participants with 560,133 variants with an average SNP call rate of 99.85%.

Mirroring the geography of Japan, PCA based on 163,727 pruned tag SNPs separated our study samples into the Hondo and Ryukyu clusters (Fig. 1a and b), which was in alignment with previous studies (Yamaguchi-Kabata et al. 2008; Sakaue et al. 2020). The Hondo cluster lies in-between Chinese/Korean and Ryukyu clusters (Supplementary Fig. S1A), suggesting that selection signals in the Ryukyu cluster cannot be inferred by using only the Hondo cluster (and Chinese/Korean cluster). To validate that the PCA results were not influenced by technical batch effects because the genotyping was conducted independently from two cohorts, we performed PCA using only the OBi cohort (which had a sufficient number of samples from both Ryukyu and Hondo) and confirmed the projections with NCGG samples. The majority of NCGG samples were found in the expected Hondo cluster, indicating that the PCA results accurately reflect the population structure (Supplementary Fig. S2).

The PCA-UMAP analysis revealed a high-resolution fine-scale population structure in the Japanese, in which mainly five distinct clusters were identified (Fig. 1c) in addition to clear separation of Japanese subjects from other East Asian populations (Supplementary Fig. S1B). The NCGG samples mostly fell into a single cluster representing the Hondo cluster (Supplementary Table S1). The OBi subjects scattered across the other four clusters and we took advantage of 2,671 individuals with clear records of ancestry (all four grandparents born in the same island group) to infer the composition of non-Hondo clusters (Supplementary Table S2). We observed that these clusters correspond well to four major Ryukyu island groups, namely Okinawa, Yaeyama, Miyako, and Kerama/Kume-jima Islands. The geographic regions inferred from PCA-UMAP matched the birth records of 2,106 out of 2,195 samples (95.94%) in non-Hondo clusters and 475 out of 476 OBi samples (99.80%) in the Hondo cluster (Supplementary Table S2). Based on the clusters of PCA-UMAP, we redefined samples into five subpopulations: Hondo ($N = 13,533$), Okinawa-jima ($N = 1,747$), Miyako ($N = 1,406$), Yaeyama ($N = 807$), Kerama/Kume-jima ($N = 439$) (Supplementary Table S1).

Genetic Distance Inferred by the F_{ST} Analysis

We considered genetic distance using Hudson's fixation index (F_{ST}) for pairs of the five subpopulations (Table 1). Hondo showed larger F_{ST} values with each Ryukyu subpopulation than those among Ryukyu subpopulations, which is consistent with the PCA analysis. Miyako and Kerama/Kume-jima ($F_{ST} \pm \text{standard error} = 9.84 \times 10^{-4} \pm 3.53 \times 10^{-6}$) showed the highest F_{ST} value among the Ryukyu subpopulations. We computed P-values of F_{ST} for 415,141 variants which have a minor allele frequency (MAF) >= 1% in the combined dataset to characterize the genetic distance. We identified a genome-wide significant peak in the MHC region ($P < 1.20 \times 10^{-7}$, 0.05/415,141) with rs2071653 as the leading variant ($F_{ST} = 0.168$, $P = 1.05 \times 10^{-7}$) (Supplementary Fig. S3). Although no variant outside the MHC reached genome-wide significance after multiple testing adjustments, we scrutinized the nonsynonymous variants showing the highest F_{ST} values between the Hondo and Ryukyu. Notably, rs671 in *ALDH2* ranked at the top ($F_{ST} = 0.088$, $P = 1.16 \times 10^{-4}$) (Supplementary Table S3). The derived allele of rs671 (p.Glu504Lys) has been known to reduce enzyme activity of ALDH2 (Matsuo et al. 2006), leading to the alcoholic intolerance, and has significantly higher allele frequency (AF) in Hondo compared with Ryukyu (0.29 vs. 0.11, $P = 4.80 \times 10^{-236}$, Chi-squared test).

Shared Candidate Loci Influenced by Selection in Hondo and Ryukyu Subpopulations

We conducted iHS analysis in Hondo and Ryukyu separately and detected two candidate genetic loci potentially affected by selection at the genome-wide significance ($P_{iHS} < 6.33 \times 10^{-8}$, 0.05/394,906/2, see Methods), including the MHC region and *IKZF2* (Fig. 2 and Table 2). The Quantile-Quantile (QQ) plots indicate there is no systematic bias stratified across different frequency bins (Supplementary Fig. S4). We confirmed that both loci do not overlap with any known structural variants (SVs) (see Methods). A signal from MHC (6p22) was observed in both Hondo and Ryukyu. The signal from *IKZF2* (2q34) in Ryukyu ($P_{iHS} = 2.15 \times 10^{-8}$) slightly fall short of the genome-wide significance in Hondo ($P_{iHS} = 5.25 \times 10^{-7}$, Table 2). We also noticed subpeaks including *ALDH2* (12q24.12), which had rs671 as the lead variant ($P_{iHS} = 1.43 \times 10^{-7}$ in Hondo) and *ADH* (4q23). In comparison with the previous iHS analysis based on BBJ samples (Yasumizu et al. 2020), we could replicate all two reported loci: MHC and *ADH*, while the significance in *IKZF2* and the suggestive signal in *ALDH2* were newly identified in this study. We also performed iHS for each Ryukyu subpopulation and found generally consistent results with those in the entire Ryukyu population (Supplementary Fig. S5).

Inspection of HLA Alleles Under Selection in the MHC region

Although statistical significance at the MHC region was observed in both Hondo and Ryukyu by iHS, the lead variants were different. Additionally, these variants showed substantial differences in allele frequencies (Table 2). To identify the HLA alleles potentially tagged by the lead variants, we carried out HLA imputation. HLA imputation has been demonstrated to accurately predict HLA alleles at a high resolution and has been commonly used as an alternative to traditional HLA typing for refining genetic association signals in the MHC region (Luo et al. 2021). This method has been instrumental in pinpointing HLA alleles

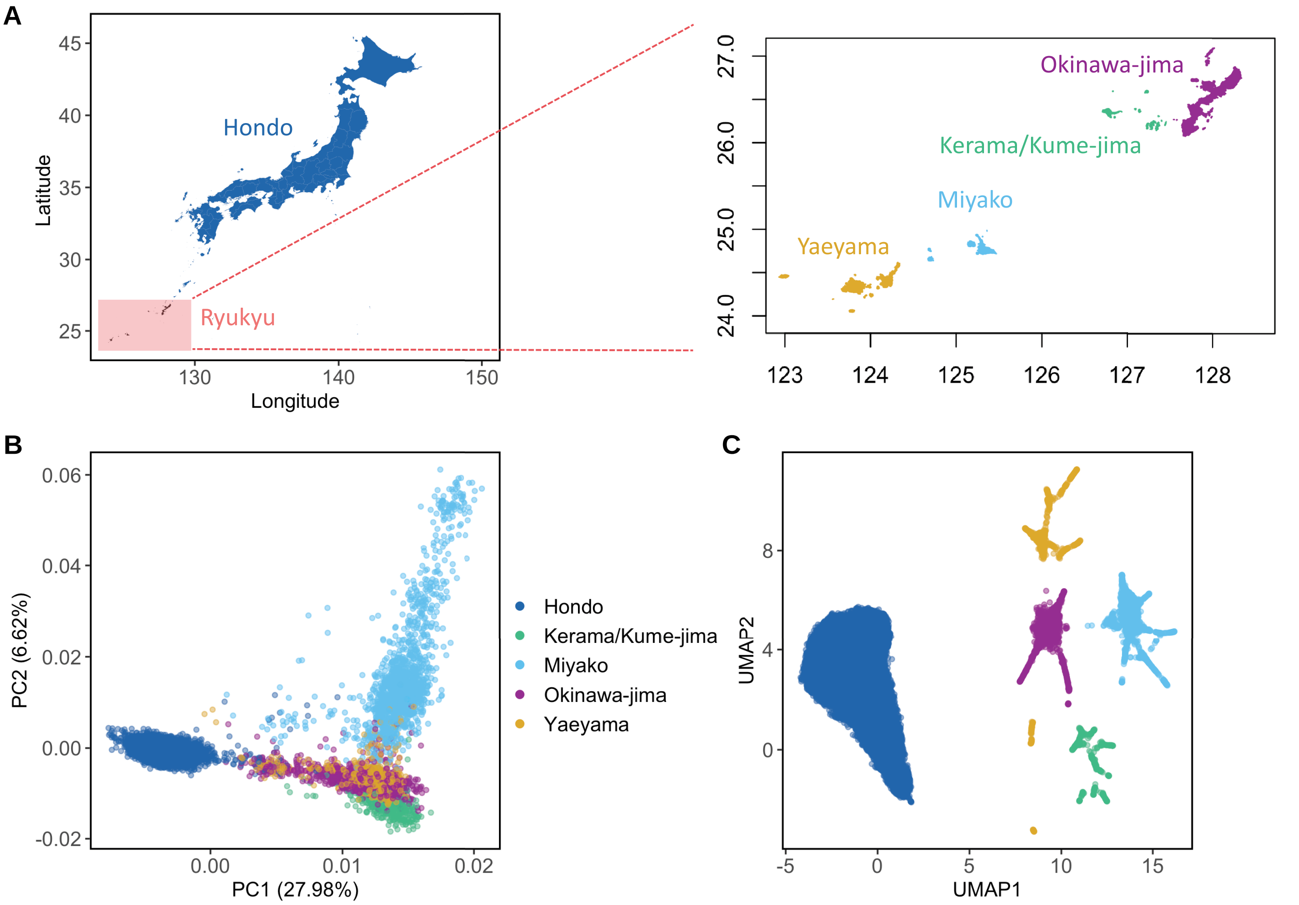

Fig. 1. Fine-scale population structure of the Japanese population. a) The geography of the Japanese archipelago and the Ryukyu archipelago. On the right, a zoomed-in view of the Ryukyu archipelago is displayed, consisting of four major island groups: Okinawa-jima, Kerama/kume-jima, Miyako, and Yaeyama. b) The PCA plot of 17,932 Japanese samples. c) The PCA-UMAP plot of 17,932 Japanese samples. The geographic region associated with each PCA-UMAP cluster were inferred based on individuals with clear ancestry records (refer to Supplementary Table S2). Samples located within the same PCA-UMAP cluster were regarded as originating from the same region. Note that the colors assigned to the samples in the PCA plot (panel b) were consistent with the colors assigned to the samples in the PCA-UMAP plot (panel c) to ensure coherence and clarity.

Table 1 Pair-wise F_{ST} of Japanese subpopulations defined by PCA-UMAP analysis

Population 1	Population 2	F_{ST}	SE
Hondo	Okinawa-jima	2.75E−03	6.73E−06
Hondo	Miyako	3.35E−03	7.97E−06
Hondo	Yaeyama	2.95E−03	7.38E−06
Hondo	Kerama/Kume-jima	3.49E−03	8.72E−06
Okinawa-jima	Miyako	4.05E−04	1.49E−06
Okinawa-jima	Yaeyama	2.53E−04	1.55E−06
Okinawa-jima	Kerama/Kume-jima	4.77E−04	2.39E−06
Miyako	Yaeyama	4.99E−04	2.12E−06
Miyako	Kerama/Kume-jima	9.84E−04	3.53E−06
Yaeyama	Kerama/Kume-jima	7.61E−04	3.43E−06

SE, stand error.

or amino acid associated with numerous immune-related diseases (Raychaudhuri et al. 2012; Hu et al. 2015). This would also be the case to delineate selection signals. By this approach, we observed the lead HLA SNP in Hondo rs6907458 tags an extended haplotype: *HLA-A*33:03-C*14:03-B*44:03-DRB1*13:02-DQB1*06:04-DPB1*04:01*, and it has the strongest linkage disequilibrium (LD) with *HLA-DQB1*06:04* ($r^2 = 0.94$), while in modest LD with the previously reported *HLA-DPB1*04:01* ($r^2 = 0.48$) (Yasumizu et al. 2020) (Supplementary Table S4, Methods). In Ryukyu, the lead SNP rs9268199 tags *DQB1*06:02-DRB1*15:01* ($r^2 = 0.79$). We noticed the HLA alleles under selection were among the alleles showing the biggest differences in frequencies between Hondo and Ryukyu (Supplementary Fig. S6 and Supplementary Table S5). Given that the P_{iHS} is an approximate value and differences in P_{iHS} of the lead variants can result from the normalization of iHS across different AF bins, we further examined both raw and normalized iHS for variants in the HLA region. Our analysis demonstrated rs6907458 and rs9268199 showed comparable scores, suggesting that these variants were likely subjected to selection in both subpopulations. Therefore, the differences in *P*-values should not be considered as evidence for differential selection (Supplementary Fig. S7).

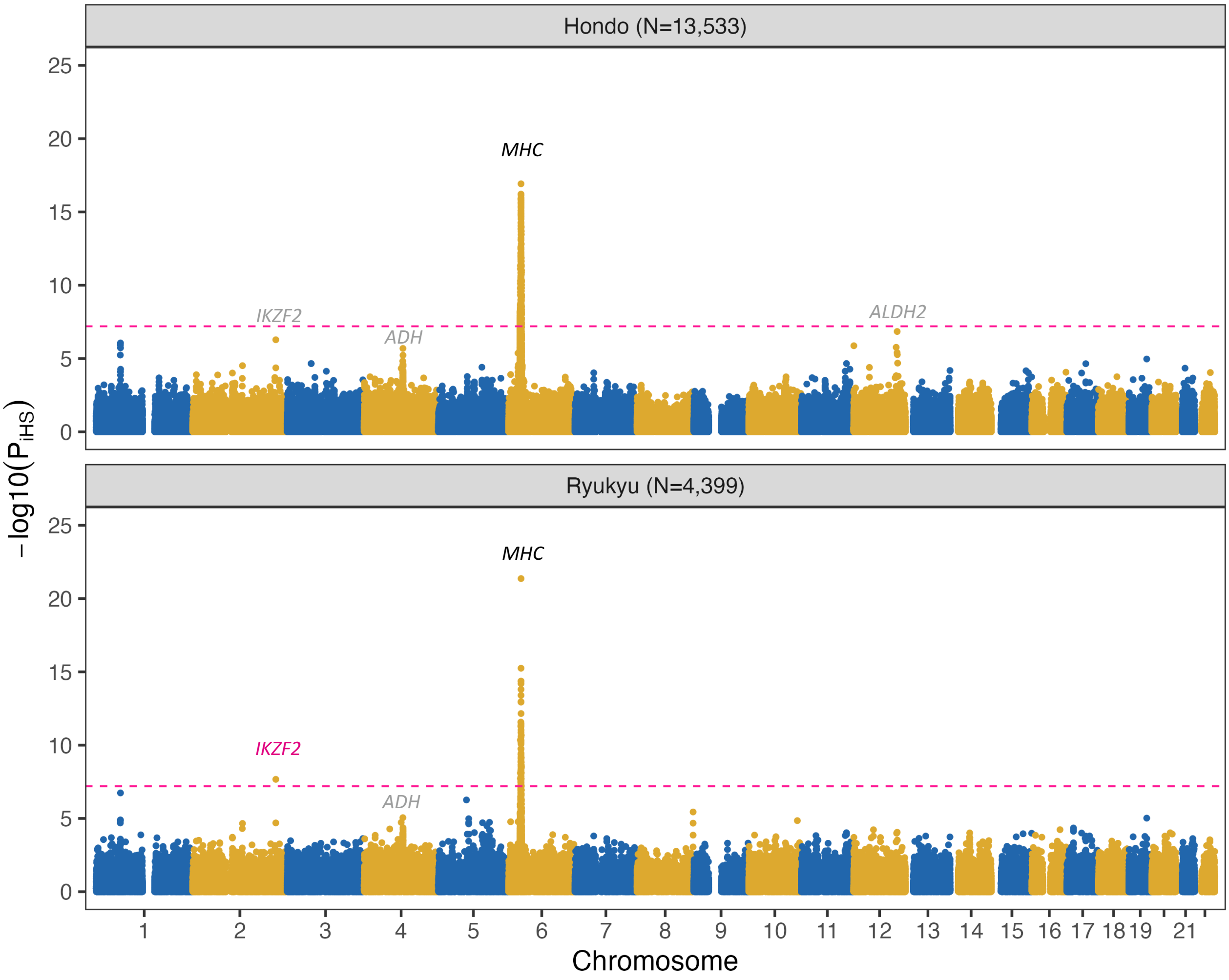

Fig. 2. Genetic loci under positive selection in the Japanese population based on iHS analysis in Hondo (top) and Ryukyu (bottom). The −log10(P_{iHS}) value (y axis) and the chromosomal position (x axis) of each SNP are plotted across the genome. The red dashed line indicates the Bonferroni-corrected genome-wide significance threshold ($P < 6.33 \times 10^{-8}$). Candidate genes previously reported to be involved in positive selection in the Japanese population are colored in black, while novel genes are marked in red. We have also identified subpeaks, including *ADH*, *ALDH2* (in Hondo only), and *IKZF2* (in Hondo only), where the P_{iHS} value slightly falls short of the genome-wide significance. These subpeaks are represented in grey.

A Novel Signal at *IKZF2* Highlighted a Population-specific Variant

A genome-wide significant natural selection signal driven by rs77756144 (2q34) was detected in Ryukyu with a comparable approximate P_{iHS} value in Hondo. The rs77756144 lies in-between the gene *IKZF2* and *SPAG16*, among which the *IKZF2* appears to be a probable candidate targeted for selection given its involvement in HTLV-1 infection related leukemia. *IKZF2* belongs to the Ikaros transcription factor family and played a critical role in lymphocyte development (Park et al. 2019) and the genetic region spanning *IKZF2* is frequently deleted in the HTLV-1 associated adult T cell leukemia/lymphoma (ATL). We observed an unusual AF pattern for rs77756144 in the gnomAD dataset (Karczewski et al. 2020), Northeast Asian Reference Database (NARD) (Yoo et al. 2019), and GenomeAsia 100K dataset (GenomeAsia100K Consortium et al. 2019). The allele was common in East Asian (AF ∼ 0.08) and Latino/admixed American populations (AF ∼ 0.135), but rare in European (AF = 0.001), African (AF = 0.002), and South Asian (AF = 0.01) (Supplementary Table S6). This suggests the derived allele may have originated in the Asian lineage before the split of East Asian and Native American. In our dataset, the AF is 0.186 in Hondo and strikingly reaches 0.31 in Ryukyu. We attempted to investigate if rs77756144 was a potential eQTL in the GTEx database. However, this variant was excluded from the analysis due to its low frequency among GTEx subjects (since most GTEx subjects had European ancestry).

Validation of the iHS Signals by the Public-available 1KGP Data

We validated implicated iHS signals using the external 1000 Genomes Project (1KGP) WGS data. The extended haplotype homozygosity (EHH) analysis showed longer

Table 2 Lead variants of candidate significant loci in Hondo and Ryukyu uncovered by iHS analysis

CHR	SNP	BP	Gene	Hondo			Ryukyu		
				DAF	Normalized iHS	P_{iHS}	DAF	Normalized iHS	P_{iHS}
6	rs6907458	32138545	*MHC*	0.083	8.55	1.20E−17	0.018	5.36	8.20E−08
6	rs9268199	32278635		0.093	7.05	1.78E−12	0.190	9.66	4.33E−22
2	rs77756144	214137825	*IKZF2*	0.186	5.02	5.25E−07	0.314	5.60	2.15E−08

Bold: $P < 6.33E{-}08$.
CHR, chromosome; SNP, single-nucleotide polymorphism; BP, base pair position; DAF, derived allele frequency.

Table 3 Candidate loci detected to be under significant positive selection by FastSMC in Hondo and Ryukyu within the past 150 generations

CHR	Position (Mb)	Cytoband	*PDRC150*	Candidate Gene(s)	Group
2	108.623 to 108.927	2q12.3	1.61E−10	*EDAR*	Hondo
4	97.783 to 101.940	4q23	1.77E−33	ADH cluster	
6	24.401 to 36.091	6p21	3.40E−50	*MHC*	
12	110.271 to 113.954	12q24	7.25E−38	*ALDH2*	
2	108.676 to 108.927	2q12.3	1.05E−07	*EDAR*	Ryukyu
2	177.549 to 177.839	2q31.1	8.39E−09	HOXD gene cluster	
4	98.596 to 100.505	4q23	1.34E−16	ADH cluster	
5	130.414 to 131.774	5q23.3-q31.1	1.12E−07	*SLC22A5*	
6	31.058 to 32.658	6p21	2.59E−08	*MHC*	
10	66.619 to 67.091	10q21.3	3.39E−08	*JMJD1C, CTNNA3*	

Density of Recent Coalescence statistic within the past 150 generations.
CHR, chromosome; DRC150.

haplotype for rs6907458 (lead MHC variant in Hondo) in the Japanese compared with Chinese Han Beijing, the European and African populations (Supplementary Fig. S8). For *IKZF2* locus, we confirmed the lead variant rs77756144 was monomorphic in Europeans and African, while the derived haplotype extended longer in the Japanese compared with Chinese (Supplementary Fig. S9), which might indicate a higher level of selection pressure.

Selection Profiles Revealed by FastSMC Analysis

To explore potential selection signals, we conducted FastSMC analysis at three different timescales within the past 150, 50, and 20 generations in Hondo and Ryukyu. Similarly, we conducted FastSMC for each Ryukyu subpopulation as a subanalysis. A timescale of 150 generations was chosen in order to make the results comparable with a previous study that employed the ASMC method (Yasumizu et al. 2020). We selected timescale of 50/20 generations to detect recent selection signatures. This is based on evidence that individuals from the Kofun era (250–538 AD) had similar genetic makeup to present-day Hondo Japanese (Cooke et al. 2021), indicating limited genetic changes in Hondo since then. The density plots and QQ plots depicting the empirical null model suggested the Gamma fitting in general is reasonable but might not handle large Density of Recent Coalescence (DRC) statistic values well, which would lead to conservative approximate *P*-values (Supplementary Fig. S10). After excluding loci that overlapped with known SVs or were flanked by segmental duplications, as these loci are prone to false-positive findings (see Methods and Supplementary Table S7), we identified 4 and 6 candidate loci, in Hondo and Ryukyu, respectively, that surpassed the genome-wide significance threshold based on the DRC_{150} statistic, which indicate these loci may have been influenced by selection within the past 150 generations (Table 3 and Fig. 3). Overall, we detected similar landscape of positive selection pressure between the Hondo and Ryukyu subpopulations. We observed signals from the MHC, the ADH cluster, and *EDAR* in both subpopulations, *ALHD2* in Hondo, and two novel loci: the HOXD cluster and *JMJD1C/CTNNA3*, specifically in Ryukyu. One difference was the presence of the *ALDH2* signal only in Hondo, which seemed to be consistent to the iHS results. Both *ALDH2* and *ADH* are key genes in alcohol metabolism and have been reported to be under positive selection in Japanese and Chinese populations (Okada et al. 2018; Yasumizu et al. 2020; Cong et al. 2022). To confirm the lack of a significant peak at *ALDH2* in Ryukyu was not due to a smaller sample size, we conducted the FastSMC to calculate the DRC_{150} statistic by down-sampling Hondo subjects to match the size of Ryukyu ($N = 4{,}458$), in which the *ALDH2* peak could be readily detected (Supplementary Fig. S11). Additionally, it is worth noting that all peaks in the initial analysis were detected with a smaller but comparable approximate *P*-value (Supplementary Table S8). This suggests that a smaller sample size may be sufficient for FastSMC analysis. We further inferred the selection signatures during the past 50 and 20 generations and observed differences in the selection profiles between Hondo and Ryukyu (Supplementary Figs. S12 and S13; Supplementary Table S9). There were differences in the selection signal on the MHC region between Hondo and Ryukyu over

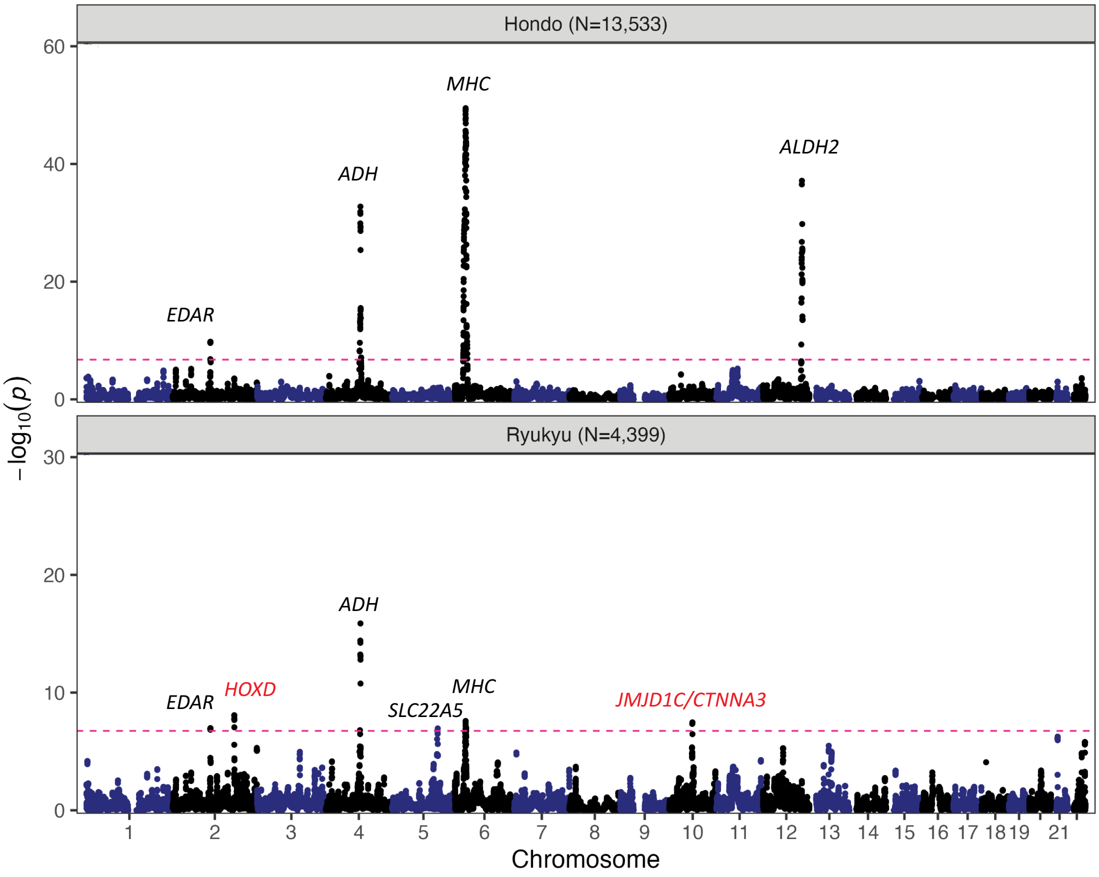

Fig. 3. Candidate loci influenced by positive selection of the Japanese population in the past 150 generations based on FastSMC analysis in Hondo (top) and Ryukyu (bottom). The −log10(P_{DRC150}) value (*y* axis) and the chromosomal position (*x* axis) of each binned region (0.05 cM) are plotted across the genome. The red dashed line represents the genome-wide significance threshold, which was obtained after Bonferroni correction for the number of bins of regions, subpopulations, and timescale ($P < 1.80 \times 10^{-7}$). Previously reported candidate genes linked with positive selection are indicated in black, while novel ones are highlighted in red.

the past 50 and 20 generations. Additionally, two alcohol-related gene regions, *ALDH2* and *ADH*, demonstrated continued evidence of selection in Hondo. Furthermore, we noticed a signal from *SLC22A5* was present in Ryukyu but not in Hondo. The *SLC22A5* gene encodes a transporter that participates in the uptake and recycling of carnitine, and this gene had been previously reported to be involved in selection in Europeans (Rees et al. 2020). We conducted a subanalysis focusing on four Ryukyu subpopulations and found that the results were generally consistent with the overall findings in Ryukyu (Supplementary Figs. S14 to S18; Supplementary Table S10). Moreover, we observed an additional peak including *TAFA5*, 21q21.1, and *COLEC11* specific to certain subpopulations (Supplementary Figs. S16 and S18).

Replication of FastSMC Signals in Independent Dataset

We conducted a replication study for significant signals observed in Hondo by FastSMC using an independent dataset of 12,103 Japanese individuals (genotyped on Illumina HumanOmniExpressExome BeadChip (OEE) microarray) (Supplementary Note S3). All significant loci detected in Hondo (*ADH*, *MHC*, *ALDH2*, and *EDAR*) were replicated with $P_{DRC150} < 2.32 \times 10^{-5}$ (Supplementary Fig. S19). Finally, we confirmed that 28 out of 29 reported loci, based on the ASMC method in a previous study with independent samples from the BBJ dataset, showed a $P_{DRC150} < 0.05$ in our replication dataset (Supplementary Table S11) (Yasumizu et al. 2020). The high replication rate demonstrates the robustness of the analysis and may indicate that the detection of novel candidate targets could be a result of including Ryukyu samples or East Asian-specific probes in the ASA.

Discussion

Here we performed large-scale genome-wide scans to identify positive selection signals in the Japanese population using the DNA microarray designed to contain many EAS-specific variants. Precise genotype information of EAS-specific variants enabled us to show previously unseen genetic structures in the Japanese and uncovered four major island groups in the Ryukyu cluster. We identified one novel candidate locus at *IKZF2* by iHS and additional 8 candidate loci which might be influenced by positive selection within the past 20–150 generations. We found different selection signals in the MHC region and *ALDH2* region between the two subpopulations.

The iHS analysis identified significant variants in the MHC region that tags specific HLA alleles. In Hondo, the lead variant rs6907458 has the strongest LD with *DQB1*06:04* and tags an extended haplotype *HLA-A*33:03-C*14:03-B*44:03-DRB1*13:02-DQB1*06:04-DPB1*04:01*, which spans from the MHC class I to class II region. This haplotype has also been previously implicated in positive selection within the Japanese population, with specific attention given to *DPB1*04:01* (Kawashima et al. 2012; Yasumizu et al. 2020). It is noteworthy that the *HLA-A*33:03-C*14:03-B*44:03-DRB1*13:02-DQB1*06:04-DPB1*04:01* haplotype was reported to be the second most frequent haplotype in the mainland Japanese, and the most frequent haplotype in Korean, but rare or not observable in other East Asian populations (Nakaoka and Inoue 2015; Zhou et al. 2015; Park et al. 2016). Given its population specificity, the long haplotype range and the constituent alleles are in strong LD, it has been speculated that this haplotype originated in Korean Peninsula, and then likely spread to Japan Hondo, potentially during the Yayoi period, followed by a rapid expansion (Nakaoka and Inoue 2015). This speculation seems to be compatible with the inference based on the FastSMC method, which revealed strong signals in the MHC region in Hondo especially within the past 20–50 generations. We reported a MHC signal in Ryukyu, in which the lead positively selected MHC variant rs9268199 tags two HLA alleles: *HLA-DRB1*15:01* and *HLA-DQB1*06:02*, which are in tight LD within the MHC class II region. The *DRB1:15:01* exhibited the highest frequency among Oceania, southeast Asia based to the survey of HLA frequency worldwide (http://www.allelefrequencies.net/) (Isshiki and Ohashi 2020).

Through a literature review, we found some examples that HLA alleles potentially under selection pressure are associated with various medical conditions, as reported in genetic association studies (Nishida et al. 2016; Pugliese et al. 2016; Yasunami et al. 2017; Penova et al. 2021). Some of these alleles could potentially offer protection against specific viral infections or conditions related to viral infections (Nishida et al. 2016; Penova et al. 2021). This suggests that viruses may be one of the contributing factors for the observed selection signals. The HLA allele *DQB1*06:04*, which is most strongly tagged by lead variant rs6907458 in Hondo, had been reported with protective effects with Hepatitis B virus (HBV) infection in a previous study (reported $P = 2.73 \times 10^{-5}$, $\mathrm{OR} = 0.44$) (Nishida et al. 2016). The estimated prevalence of chronic HBV infection in Japan in 2016 was around 0.6% (Razavi-Shearer et al. 2018), while historically, the prevalence of chronic HBV infection was much higher before the HBV vaccination program was implemented nationwide in 1986. As such, there may be a potential connection between the positive selection of *DQB1*06:04* and HBV infection. For *DRB1*15:01-DQB1*06:02*, it has been recognized as the most potent genetic factor for narcolepsy (Miyagawa and Tokunaga 2019). Recent research has also reported an association of this haplotype with an increased risk of systemic lupus erythematosus (SLE) and a protective effect against type 1 diabetes (T1D) (Pugliese et al. 2016; Kawasaki et al. 2023). In addition, a GWAS of HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP) patients and asymptomatic HTLV-1 carriers identified *HLA-DRB1*15:01* (reported $P = 1.06 \times 10^{-5}$, $\mathrm{OR} = 0.59$) and *HLA-DQB1*06:02* (reported $P = 1.78 \times 10^{-6}$, $\mathrm{OR} = 0.43$) as top protective alleles (Penova et al. 2021). This raises the possibility of a potential relationship between the HTLV-1 and the selection signal. The HTLV-1 infection is prevalent in Japan (Watanabe 2011). In particular, Ryukyu and the southwestern part of Hondo have the highest HTLV-1 infection rate (Iwanaga 2020). The HTLV-1 infection is largely latent but can reactivate and lead to severe symptoms such as HAM/TSP which lead to neuronal damage or even life-threatening ATL. The *HLA-DRB1*15:01-DQB1*06:02* haplotype has been reported to be under positive selection in Papua New Guinea, a region known for its high prevalence of HTLV-1 (Gessain and Cassar 2012). Additionally, it is noteworthy that Oceania and Southeast Asia, where the haplotype exhibits a relatively higher frequency, are recognized for their elevated HTLV-1 prevalence (Gessain and Cassar 2012). The haplotype frequency of *DRB1*15:01* and *DQB1*06:02* is significantly higher in Ryukyu compared with that of Hondo (0.191 vs. 0.071, $P = 1.33 \times 10^{-179}$). While this may mirror the differences in HTLV-1 prevalence between Ryukyu and Hondo, we should not overinterpret the potential correlation between HLA allele frequency and HTLV-1 prevalence. Factors other than these pathogens, diseases or demographic history could also contribute to the difference in HLA allele frequencies and further accumulation of investigations is essential. Moreover, the implicated HLA alleles may have associations with other pathogens. For example, *DRB1*15:01* has been reported as a co-receptor for the Epstein–Barr virus (Menegatti et al. 2021), which implies that this allele could potentially interact with other known or unidentified viruses.

In addition to MHC, the iHS analysis identified another strong signal at *IKZF2* which might be linked with the HTLV-1. The IKZF2 has been recently shown as a key regulator of T cell development, which maintains the stem cell self-renewal and suppresses myeloid linage differentiation by modulating chromatin accessibility (Park et al. 2019). In a multiomics study consisted of 426 ATL cases, intragenic deletions and inversions of *IKZF2* were observed as one of the most frequent genetic alterations in ATL (Kataoka et al. 2015), suggesting it might play a critical role in pathogenesis of ATL. It is still unknown whether rs77756144 has any protective effects against HTLV-1 infection outcomes, and further research is needed to uncover the biological significance of this candidate selection signal.

Based on a recently developed method FastSMC, we detected additional candidate targets that may have been influenced by selection in the Japanese population at three different time frames. The identified loci include *EDAR*, *ADH* cluster, and *ALDH2*, three well-known East Asian-specific loci targeted by positive selection (Fujimoto et al. 2008; Kimura et al. 2009; Okada et al. 2018; Yasumizu et al. 2020). These *ADH* and *ALDH2* are related to alcohol metabolism and missense variants rs1229984 (Arg48His) in *ADH1B* and rs671 (Glu504Lys) in *ALDH2*, which make carriers less tolerant to alcohol, were shown to be favored by selection (Yasumizu et al. 2020). The FastSMC analysis, focusing on the past 20–150 generations, detected a signal at *ADH* in both Hondo and Ryukyu, whereas a signal for *ALDH2* was observed only in Hondo and not in Ryukyu. This difference warrants further investigation in future analyses. The seemingly different selection profiles for *ADH* and *ALDH2* may be consistent with a recent study, which suggests that the onset of positive selection for *ADH* occurred approximately 12,500 yr earlier than that for *ALDH2*. Specifically, while positive selection on *ADH* may have begun around 20,000 yr ago, that selection on *ALDH2* is estimated to have started about 7,500 yr ago in East Asian populations (Kawai et al. 2023). The reasons for the positive selection of these alcohol-related genes remain unknown, but it has been proposed that they may be related to the large-scale adoption of rice cultivation in East Asia (Supplementary Note S4).

We highlight several aspects where caution should be exercised. First, genome-wide significant signals from MHC based on $\mathrm{DRC}_{50/20}$ were observed in Hondo but not in Ryukyu. This result requires careful interpretation; it does not necessarily indicate a lack of natural selection in the MHC in Ryukyu within the past 50 and 20 generations, a conclusion that would contradict the iHS results. Instead, the HLA alleles under selection in Hondo may originate differently given the peopling history of Hondo as we previously discussed. Second, it should be emphasized that difference in the strength of signature should not be interpreted as differential selection. The detection of a selection signature depends on multiple factors such as allele frequency and the population's demographic history, and potential technical artifact (Meyer et al. 2006; Nielsen et al. 2007; Huber et al. 2014). Thus, lack of significance or diminished evidence in a subpopulation should not be taken as sufficient proof of differential selection. Third, we recommend conducting replication analyses of the newly identified candidate loci in future studies, utilizing independent datasets.

In summary, we presented genome-wide scans of selection in the Japanese population with individuals collected from both Hondo and Ryukyu. We highlighted selection signatures in the MHC region. The selection signal in Ryukyu might be linked to specific diseases or viral infections such as HTLV-1, although further analysis will be required to validate this hypothesis. The FastSMC analysis detected signals for the *ADH1B* and *ALDH2* genes for the Hondo and Ryukyu populations, and nominated novel candidate selection targets for future studies. Our study highlights the significance and need for broadening previous genetic studies to include individuals from a range of genetic backgrounds. Furthermore, it underscores the value of considering population-specific variants and connecting selection signals with epidemiological characteristics.

Methods

Samples and Genotype Data

Samples analyzed in this study were obtained from two cohorts. The first cohort consisted of 13,753 participants who were recruited from the NCGG Biobank of Japan. The second one consisted of 6,613 participants who were recruited at Okinawa Prefecture through the OBi Project. All genomic DNA samples were genotyped by Illumina Infinium ASA v1.0. Origins of individual participants were surveyed in OBi Project and information of the birthplace (islands) for their four grandparents were obtained by questionnaire. The geographic birthplace information was not available for the NCGG cohort. The genotyping was performed following the manufacturer's recommended protocols. After the merge of raw genotyping calls of the two cohorts, we conducted QC to remove variants that are low quality and samples with low call-rates, having first or second degree relative(s), or non-Japanese inferred by PCA.

PCA and PCA-UMAP

The PCA was conducted to explore and visualize underlying population structure. To make appropriate inferences, we included additional samples from two public datasets: (1) 2,504 samples in the 1KGP (1000 Genomes Project Consortium 2012) and (2) 85 Koreans from the Korean Personal Genome Project (KPGP) (http://kpgp.kr/). For the former, we obtained the high-depth WGS released by the New York Genome Center (NYGC) (https://www.internationalgenome.org/data-portal/data-collection/30x-grch38), and extracted SNPs included in the ASA and manually lift-down to hg19 based on the manifest files provided by Illumina (https://jp.support.illumina.com/downloads/infinium-asian-screening-array-v1-0-product-files.html). The genotype data of Korean was downloaded from the following site: http://camda2021.bioinf.jku.at/kpg_prepro. After the merge of all samples, we pruned the non-HLA variants and selected pruned common variants (MAF >= 5%; pruning parameter: index-pair 100, 10, 0.2), and the top 20 PC scores were calculated by PLINK v1.9 (Chang et al. 2015). To gain a higher resolution of the population structure, we further applied UMAP) analysis for the top 20 PCs, using the R package umap (0.2.7.0) (McInnes et al. 2018).

F_{ST} Calculation

We calculated the Hudson's F_{ST} value of each SNP using the KRIS software package. (version 1.1.1, R version 3.6.2; https://rdrr.io/cran/KRIS/). We used modified function "fst.hudson" for calculation of F_{ST} with standard error.

Since the distribution of weighted mean F_{ST} theoretically follows the Chi-square distribution (Hartl and Clark 1997), we fitted these two distributions and calculated the approximate one-sided *P*-values. To correct the multiple testing, we adopted Bonferroni adjusted genome-wide significance level at $P = 1.20 \times 10^{-7}$ (0.05/415,141 SNPs).

Integrated Haplotype Score Analysis

We restricted the analysis to variants whose ancestral allele is supported by chimp and macaque. Ancestral states of each SNP were inferred using the ancestral hg19 genome provided by the 1000 genomes consortium, based on the human–chimp–macaque alignment (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/) (Flicek et al. 2012). To ensure accurate phasing and iHS analysis, we utilized the Japanese-specific recombination map, which was created using the 1KGP JPT WGS data (Spence and Song 2019) (https://github.com/popgenmethods/pyrho#human-recombination-maps). Since the Ryukyu population was not present in the 1KGP dataset, we used the same recombination map of JPT. The SNP phasing was done for each chromosome using Eagle v2.4.1 with the default parameters (Loh et al. 2016). We excluded variants which have a MAF < 1% or whose ancestral alleles were not determined. We interpolated the physical position for each variant. We then conducted iHS analysis using the software selscan (v1.3) (Szpiech and Hernandez 2014) for the Hondo and Ryukyu clusters inferred by PCA-UMAP. The normalized iHS (standardized Z scores) were obtained by normalization under 100 MAF bins. Approximate *P*-values were calculated by fitting the normalized iHS scores, assuming a normal distribution. To correct the multiple testing, we adopted Bonferroni adjusted genome-wide significance level at approximate $P_{iHS} = 6.33 \times 10^{-8}$ (0.05/394,906/2, which is the number of variants that received an iHS score and two subpopulations) in the main analysis, and approximate $P_{iHS} = 3.17 \times 10^{-8}$ (0.05/394,906/4) in the subanalysis for each Ryukyu cluster. EHH around the selected SNPs were calculated and plotted using the rehh software package (Gautier and Vitalis 2012) (version 3.2.2, R version 3.6.2) with the public available 1KGP data (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502).

HLA Imputation

To understand variations in HLA alleles depending on regional populations defined in PCA-UMAP, we conducted HLA imputation with the HLA-TAPAS pipeline (Luo et al. 2021). In brief, we extracted 7,885 SNPs of the HLA region (chr6:25MB-34MB, hg19). Based on a reference panel build from 1KGP dataset, the HLA alleles were imputed up to four digits using an enhanced version of SNP2HLA (Jia et al. 2013). We then calculated the frequencies of HLA alleles in each population. We also calculated the LD *r*-squared (*r*2) between the top SNPs under selection and HLA alleles using PLINK v1.9.

FastSMC Analysis

In addition to iHS, we attempted to use FastSMC (Nait Saada et al. 2020), an algorithm designed for Identical-By-Descent (IBD) detection and estimation of coalescent times for each IBD, to specifically identify recent selection signals in the Japanese population. By analyzing locus-specific IBD sharing patterns, we calculated DRC within past 20, 50, and 150 generations (DRC_{20}, DRC_{50}, and DRC_{150}) from IBD quality scores. We summarized the mean value for each sliding window at a size of 0.05 centimorgan (cM). The decoding file was prepared from the JPT demographic and allele frequency file. The higher DRC values are expected to be found in the genomic loci that were shared by many subjects but inherited from limited number of common ancestors, which might be due to the recent positive selection. We fitted a Gamma distribution to the estimated DRC values using the neutral regions in the genome. We first excluded genetic loci that have been reported in the literature to be under positive selection in the Japanese population, and we iteratively removed regions that showed evidence of being targeted by selection based on the DRC statistic (Fujimoto et al. 2008; Okada et al. 2018; Yasumizu et al. 2020). Based on this null model, we derived approximate one-sided *P*-values. To correct the multiple testing, we adopted Bonferroni adjusted genome-wide significance level at approximate $P_{DRC} = 1.80 \times 10^{-7}$ (0.05/46,306 regions/2 subpopulations/3 time points) in the main analysis, and at approximate $P_{DRC} = 9.00 \times 10^{-8}$ (0.05/46,306 regions/4 Ryukyu subpopulations/3 time points) in the subanalysis.

Inspecting and Filtering Potential False-positive Significant Loci

Extended LD and shared haplotypes can also result from the underlying presence of SV, which may suppress recombination in heterozygote carriers (Fishman et al. 2013; Morgan et al. 2017). To examine whether any loci deemed to have genome-wide significance by iHS or DRC statistic overlap with known SVs, we examined the Human Genome Structural Variation Consortium (HGSVC) Phase 2 dataset (Ebert et al. 2021). We obtained the called SV data of 3,202 deep-coverage samples processed by the PanGenie genotyping pipeline (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/PanGenie_results/pangenie_merged_multi_nosnvs.vcf.gz). Given that the genome build of the SV data is in hg38, we liftovered the genomic positions for all significant loci using liftover tool from the UCSC genome browser (https://genome.ucsc.edu/cgi-bin/hgLiftOver). We extracted common SVs with length >=10 kb and an AF >= 1% and in Asians. To detect any significant loci overlapping with the SVs, we employed the intersectBed tool from Bedtools (v2.3). Specifically, we considered loci that had an overlap of at least 10% of the length of the SV. Utilizing the “Segmental Dups” track and “Reference Assembly Fix Patch Sequence Alignments” track, both available via the UCSC genome browser (https://genome.

ucsc.edu), we identified and subsequently excluded significant loci that were either encompassed by or flanked by segmental duplications or contained an excessive number of reference "fix patches'.

Replication Study

We additionally included 12,103 Japanese samples from the BBJ that were genotyped by Illumina OmniExpressExome assay. These samples have not been previously analyzed and were independent of 170,882 BBJ samples used in the previous genome-wide scan for selection signals in the Japanese population (Yasumizu et al. 2020). This dataset was pre QCed and used in a previous study (Ito et al. 2023). We obtained the phased VCF files (Supplementary Note S3). The FastSMC analysis was conducted in the same manner as the previously described for ASA data. The stringent Bonferroni adjusted genome-wide significance levels were set at $P = 1.13 \times 10^{-6}$ (0.05/44,294 regions).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank all participants who donated DNA samples and clinical information for the National Center for Geriatrics and Gerontology (NCGG) Biobank, Okinawa Bioinformation Bank (OBi) Project, and BioBank Japan (BBJ) Project. We would like to express our gratitude to the staff at all study sites for their assistance in sample collection. We extend our gratitude to Prof. Takuya Fukushima (University of the Ryukyus) for his valuable guidance in interpreting the potential social implications arising from our analytical results.

Funding

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI (20H00462 to C.T., 18H05506 to M.M.) and by the Japan Agency for Medical Research and Development (AMED) (21ek0109555, 21tm0424220, and 21ck0106642 to C.T.). S.M. is supported by a grant from the promotion project of innovation-eco system collaborating organization from the Okinawa prefecture.

Conflict of interest statement. None declared.

Data Availability

The summary statistics of the iHS and FastSMC analysis is available from the Japanese ENcyclopedia of GEnetic associations by RIKEN (JENGER) website (http://jenger.riken.jp/).

References

Aoki R, Karube K, Sugita Y, Nomura Y, Shimizu K, Kimura Y, Hashikawa K, Suefuji N, Kikuchi M, Ohshima K. Distribution of malignant lymphoma in Japan: analysis of 2260 cases, 2001–2006. *Pathol Int*. 2008:**58**(3):174–182. https://doi.org/10.1111/j.1440-1827.2007.02207.x.

Bendjilali N, Hsueh W-C, He Q, Willcox DC, Nievergelt CM, Donlon TA, Kwok P-Y, Suzuki M, Willcox BJ. Who are the Okinawans? Ancestry, genome diversity, and implications for the genetic study of human longevity from a geographically isolated population. *J Gerontol Ser A: Biol Sci Med Sci*. 2014:**69**(12):1474–1484. https://doi.org/10.1093/gerona/glt203.

Benton ML, Abraham A, LaBella AL, Abbot P, Rokas A, Capra JA. The influence of evolutionary history on human health and disease. *Nat Rev Genet*. 2021:**22**(5):269–283. https://doi.org/10.1038/s41576-020-00305-9.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci*. 2015:**4**(1):7. https://doi.org/10.1186/s13742-015-0047-8.

Cong P-K, Bai W-Y, Li J-C, Yang M-Y, Khederzadeh S, Gai S-R, Li N, Liu Y-H, Yu S-H, Zhao W-W, *et al.* Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat Commun*. 2022:**13**(1):2939. https://doi.org/10.1038/s41467-022-30526-x.

Cooke NP, Mattiangeli V, Cassidy LM, Okazaki K, Stokes CA, Onbe S, Hatakeyama S, Machida K, Kasai K, Tomioka N, *et al.* Ancient genomics reveals tripartite origins of Japanese populations. *Sci Adv*. 2021:**7**(38):eabh2419. https://doi.org/10.1126/sciadv.abh2419.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021:**372**(6537):eabf7117. https://doi.org/10.1126/science.abf7117.

Fishman L, Stathos A, Beardsley PM, Williams CF, Hill JP. Chromosomal rearrangements and the genetics of reproductive barriers in mimulus (monkey flowers). *Evolution*. 2013:**67**(9):2547–2560. https://doi.org/10.1111/evo.12154.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, *et al.* Ensembl 2012. *Nucleic Acids Res*. 2012:**40**(D1):D84–D90. https://doi.org/10.1093/nar/gkr991.

Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, *et al.* A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet*. 2008:**17**(6):835–843. https://doi.org/10.1093/hmg/ddm355.

Fukiyama K, Kimura Y, Wakugami K, Muratani H. Incidence and long-term prognosis of initial stroke and acute myocardial infarction in Okinawa, Japan. *Hypertens Res*. 2000:**23**(2):127–135. https://doi.org/10.1291/hypres.23.127.

Gautier M, Vitalis R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*. 2012:**28**(8):1176–1177. https://doi.org/10.1093/bioinformatics/bts115.

GenomeAsia100K Consortium; Wall JD, Stawiski EW, Ratan A, Kim HL, Kim C, Gupta R, Suryamohan K, Gusareva ES, Purbojati RW, *et al.* 2019. The GenomeAsia 100 K project enables genetic discoveries across Asia. *Nature*. **576**(7785):106–111. https://doi.org/10.1038/s41586-019-1793-z.

Gessain A, Cassar O. Epidemiological aspects and world distribution of HTLV-1 infection. *Front Microbiol*. 2012:**3**:388. https://doi.org/10.3389/fmicb.2012.00388.

Hartl DL, Clark AG. *Principles of population genetics*. Sunderland: Sinauer Associates; 1997.

Hayashi J, Kajiyama W, Noguchi A, Ikematsu H, Nomura H, Nakashima K, Morofuji M, Kashiwagi S. Marked decrease of hepatitis B virus infection among children in Okinawa, Japan. *Int J Epidemiol*. 1990:**19**(4):1083–1085. https://doi.org/10.1093/ije/19.4.1083.

Hu X, Deutsch AJ, Lenz TL, Onengut-Gumuscu S, Han B, Chen W-M, Howson JMM, Todd JA, de Bakker PIW, Rich SS, *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat Genet*. 2015:**47**(8):898–905. https://doi.org/10.1038/ng.3353.

Huber CD, Nordborg M, Hermisson J, Hellmann I. Keeping it local: evidence for positive selection in Swedish Arabidopsis thaliana. *Mol Biol Evol*. 2014:**31**(11):3026–3039. https://doi.org/10.1093/molbev/msu247.

Isshiki M, Ohashi J. Population specific positive selection acted on the HLA class II region in Papuans. *MHC*. 2020:**27**(2):53–58. https://doi.org/10.12667/mhc.27.53.

Ito S, Liu X, Ishikawa Y, Conti DD, Otomo N, Kote-Jarai Z, Suetsugu H, Eeles RA, Koike Y, Hikino K, *et al.* Androgen receptor binding sites enabling genetic prediction of mortality due to prostate cancer in cancer-free subjects. *Nat Commun*. 2023:**14**(1):4863. https://doi.org/10.1038/s41467-023-39858-8.

Iwanaga M. Epidemiology of HTLV-1 infection and ATL in Japan: an update. *Front Microbiol*. 2020:**11**:1124. https://doi.org/10.3389/fmicb.2020.01124.

Japanese Archipelago Human Population Genetics Consortium. The history of human populations in the Japanese archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. *J Hum Genet*. 2012:**57**(12):787–795. https://doi.org/10.1038/jhg.2012.114.

Jia X, Han B, Onengut-Gumuscu S, Chen W-M, Concannon PJ, Rich SS, Raychaudhuri S, de Bakker PIW. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One*. 2013:**8**(6): e64683. https://doi.org/10.1371/journal.pone.0064683.

Johnson KE, Voight BF. Patterns of shared signatures of recent positive selection across human populations. *Nat Ecol Evol*. 2018:**2**(4): 713–720. https://doi.org/10.1038/s41559-018-0478-6.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020:**581**(7809):434–443. https://doi.org/10.1038/s41586-020-2308-7.

Kataoka K, Nagata Y, Kitanaka A, Shiraishi Y, Shimamura T, Yasunaga J, Totoki Y, Chiba K, Sato-Otsubo A, Nagae G, *et al.* Integrated molecular analysis of adult T cell leukemia/lymphoma. *Nat Genet*. 2015:**47**(11):1304–1315. https://doi.org/10.1038/ng.3415.

Kawai Y, Mimori T, Kojima K, Nariai N, Danjoh I, Saito R, Yasuda J, Yamamoto M, Nagasaki M. Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals. *J Hum Genet*. 2015:**60**(10):581–587. https://doi.org/10.1038/jhg.2015.68.

Kawai Y, Watanabe Y, Omae Y, Miyahara R, Khor S-S, Noiri E, Kitajima K, Shimanuki H, Gatanaga H, Hata K, *et al.* 2023. Exploring the genetic diversity of the Japanese Population: insights from a large-scale whole genome sequencing analysis. *Genomics*. Available from: https://doi.org/10.1101/2023.01.23.525133, 25 January 2023, preprint: not peer reviewed.

Kawasaki A, Kusumawati PA, Kawamura Y, Kondo Y, Kusaoi M, Amano H, Kusanagi Y, Itoh K, Fujimoto T, Tamura N, *et al.* Genetic dissection of *HLA-DRB1*15:01* and XL9 region variants in Japanese patients with systemic lupus erythematosus: primary role for *HLA-DRB1*15:01*. *RMD Open*. 2023:**9**(2):e003214. https://doi.org/10.1136/rmdopen-2023-003214.

Kawashima M, Ohashi J, Nishida N, Tokunaga K. Evolutionary analysis of classical HLA class I and II genes suggests that recent positive selection acted on DPB1*04 : 01 in Japanese population. *PLoS One*. 2012:**7**(10):e46806. https://doi.org/10.1371/journal.pone.0046806.

Kimura R, Yamaguchi T, Takeda M, Kondo O, Toma T, Haneji K, Hanihara T, Matsukusa H, Kawamura S, Maki K. A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *Am J Hum Genet*. 2009:**85**(4):528–535. https://doi.org/10.1016/j.ajhg.2009.09.006.

Klein RG. Out of Africa and the evolution of human behavior. *Evol Anthropol*. 2008:**17**(6):267–281. https://doi.org/10.1002/evan.20181.

Koganebuchi K, Kimura R. Biomedical and genetic characteristics of the Ryukyuans: demographic history, diseases and physical and physiological traits. *Ann Hum Biol*. 2019:**46**(4):354–366. https://doi.org/10.1080/03014460.2019.1582699.

Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet*. 2005:**77**(2):171–192. https://doi.org/10.1086/432519.

Liu X, Lu D, Saw W-Y, Shaw PJ, Wangkumhang P, Ngamphiw C, Fucharoen S, Lert-Itthiporn W, Chin-Inmanu K, Chau TNB, *et al.* Characterising private and shared signatures of positive selection in 37 Asian populations. *Eur J Hum Genet*. 2017:**25**(4): 499–508. https://doi.org/10.1038/ejhg.2016.181.

Loh P-R, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK biobank cohort. *Nat Genet*. 2016:**48**(7):811–816. https://doi.org/10.1038/ng.3571.

Luo Y, Kanai M, Choi W, Li X, Sakaue S, Yamamoto K, Ogawa K, Gutierrez-Arcelus M, Gregersen PK, Stuart PE, *et al.* A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat Genet*. 2021:**53**(10):1504–1516. https://doi.org/10.1038/s41588-021-00935-7.

Mathieson S, Mathieson I. FADS1 and the timing of human adaptation to agriculture. *Mol Biol Evol*. 2018:**35**(12):2957–2970. https://doi.org/10.1093/molbev/msy180.

Matsunami M, Koganebuchi K, Imamura M, Ishida H, Kimura R, Maeda S. Fine-scale genetic structure and demographic history in the Miyako Islands of the Ryukyu Archipelago. *Mol Biol Evol*. 2021:**38**(5):2045–2056. https://doi.org/10.1093/molbev/msab005.

Matsuo K, Wakai K, Hirose K, Ito H, Saito T, Tajima K. Alcohol dehydrogenase 2 His47Arg polymorphism influences drinking habit independently of aldehyde dehydrogenase 2 Glu487Lys polymorphism: analysis of 2,299 Japanese subjects. *Cancer Epidemiol Biomarkers Prev*. 2006:**15**(5):1009–1013. https://doi.org/10.1158/1055-9965.EPI-05-0911.

McInnes L, Healy J, Melville J. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. Available from: https://arxiv.org/abs/1802.03426

Menegatti J, Schub D, Schäfer M, Grässer FA, Ruprecht K. HLA-DRB1*15:01 is a co-receptor for Epstein–Barr virus, linking genetic and environmental risk factors for multiple sclerosis. *Eur J Immunol*. 2021:**51**(9):2348–2350. https://doi.org/10.1002/eji.202149179.

Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history and natural selection in the human major histocompatibility complex loci. *Genetics*. 2006:**173**(4): 2121–2142. https://doi.org/10.1534/genetics.105.052837.

Miyagawa T, Tokunaga K. Genetics of narcolepsy. *Hum Genome Var*. 2019:**6**(1):4. https://doi.org/10.1038/s41439-018-0033-7.

Morgan AP, Gatti DM, Najarian ML, Keane TM, Galante RJ, Pack AI, Mott R, Churchill GA, de Villena FP-M. Structural variation shapes the landscape of recombination in mouse. *Genetics*. 2017:**206**(2):603–619. https://doi.org/10.1534/genetics.116.197988.

Morikawa K, Kuroda M, Tofuku Y, Uehara H, Akizawa T, Kitaoka T, Koshikawa S, Sugimoto H, Hashimoto K. Prevalence of HTLV-1 antibodies in hemodialysis patients in Japan. *Am J Kidney Dis*. 1988:**12**(3):185–193. https://doi.org/10.1016/S0272-6386(88)80120-3.

Nait Saada J, Kalantzis G, Shyr D, Cooper F, Robinson M, Gusev A, Palamara PF. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat Commun*. 2020:**11**(1):6130. https://doi.org/10.1038/s41467-020-19588-x.

Nakaoka H, Inoue I. Distribution of HLA haplotypes across Japanese archipelago: similarity, difference and admixture. *J Hum Genet*. 2015:**60**(11):683–690. https://doi.org/10.1038/jhg.2015.90.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. *Nat Rev Genet*. 2007:**8**(11):857–868. https://doi.org/10.1038/nrg2187.

Nishida N, Ohashi J, Khor S-S, Sugiyama M, Tsuchiura T, Sawai H, Hino K, Honda M, Kaneko S, Yatsuhashi H, *et al.* Understanding of HLA-conferred susceptibility to chronic hepatitis B infection requires HLA genotyping-based association analysis. *Sci Rep*. 2016:**6**(1):24767. https://doi.org/10.1038/srep24767.

Okada Y, Momozawa Y, Sakaue S, Kanai M, Ishigaki K, Akiyama M, Kishikawa T, Arai Y, Sasaki T, Kosaki K. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat Commun*. 2018:**9**(1):1631. https://doi.org/10.1038/s41467-018-03274-0.

Oota H, Pakstis AJ, Bonne-Tamir B, Goldman D, Grigorenko E, Kajuna SLB, Karoma NJ, Kungulilo S, Lu R-B, Odunsi K, *et al.* The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. *Ann Hum Genet*. 2004:**68**(2):93–109. https://doi.org/10.1046/j.1529-8817.2003.00060.x.

Palamara PF, Terhorst J, Song YS, Price AL. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nat Genet*. 2018:**50**(9): 1311–1317. https://doi.org/10.1038/s41588-018-0177-x.

Park H, Lee Y, Song EY, Park MH. HLA-A, HLA-B and HLA-DRB1 allele and haplotype frequencies of 10 918 Koreans from bone marrow donor registry in Korea. *Int J Immunogenet*. 2016:**43**(5):287–296. https://doi.org/10.1111/iji.12288.

Park S-M, Cho H, Thornton AM, Barlowe TS, Chou T, Chhangawala S, Fairchild L, Taggart J, Chow A, Schurer A, *et al.* IKZF2 drives leukemia stem cell self-renewal and inhibits myeloid differentiation. *Cell Stem Cell*. 2019:**24**(1):153–165.e7. https://doi.org/10.1016/j.stem.2018.10.016.

Penova M, Kawaguchi S, Yasunaga J, Kawaguchi T, Sato T, Takahashi M, Shimizu M, Saito M, Tsukasaki K, Nakagawa M, *et al.* Genome wide association study of HTLV-1–associated myelopathy/tropical spastic paraparesis in the Japanese population. *Proc Natl Acad Sci USA*. 2021:**118**(11):e2004199118. https://doi.org/10.1073/pnas.2004199118.

Pugliese A, Boulware D, Yu L, Babu S, Steck AK, Becker D, Rodriguez H, DiMeglio L, Evans-Molina C, Harrison LC, *et al.* HLA-DRB1*15:01-DQA1*01:02-DQB1*06:02 haplotype protects autoantibody-positive relatives from type 1 diabetes throughout the stages of disease progression. *Diabetes*. 2016:**65**(4): 1109–1119. https://doi.org/10.2337/db15-1105.

Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee H-S, Jia X, Alfredsson L, Padyukov L, Klareskog L, Worthington J, *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet*. 2012:**44**(3):291–296. https://doi.org/10.1038/ng.1076.

Razavi-Shearer D, Gamkrelidze I, Nguyen MH, Chen D-S, Van Damme P, Abbas Z, Abdulla M, Abou Rached A, Adda D, Aho I, *et al.* Global prevalence, treatment, and prevention of hepatitis B virus infection in 2016: a modelling study. *Lancet Gastroenterol Hepatol*. 2018:**3**(6):383–403. https://doi.org/10.1016/S2468-1253(18)30056-6.

Rees JS, Castellano S, Andrés AM. The genomics of human local adaptation. *Trends Genet*. 2020:**36**(6):415–428. https://doi.org/10.1016/j.tig.2020.03.006.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature*. 2007:**449**(7164):913–918. https://doi.org/10.1038/nature06250.

Sakaue S, Hirata J, Kanai M, Suzuki K, Akiyama M, Lai Too C, Arayssi T, Hammoudeh M, Al Emadi S, Masri BK, *et al.* Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. *Nat Commun*. 2020:**11**(1):1569. https://doi.org/10.1038/s41467-020-15194-z.

Sato T, Nakagome S, Watanabe C, Yamaguchi K, Kawaguchi A, Koganebuchi K, Haneji K, Yamaguchi T, Hanihara T, Yamamoto K. Genome-wide SNP analysis reveals population structure and demographic history of the Ryukyu islanders in the southern part of the Japanese archipelago. *Mol Biol Evol*. 2014:**31**(11):2929–2940. https://doi.org/10.1093/molbev/msu230.

Shigemizu D, Mitsumori R, Akiyama S, Miyashita A, Morizono T, Higaki S, Asanomi Y, Hara N, Tamiya G, Kinoshita K. Ethnic and trans-ethnic genome-wide association studies identify new loci influencing Japanese Alzheimer's Disease risk. *Transl Psychiatry*. 2021:**11**(1):151. https://doi.org/10.1038/s41398-021-01272-3.

Spence JP, Song YS. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Sci Adv*. 2019:**5**(10):eaaw9206. https://doi.org/10.1126/sciadv.aaw9206.

Suzuki R, Saitou N, Matsuari O, Shiota S, Matsumoto T, Akada J, Kinjo N, Kinjo F, Teruya K, Shimoji M, *et al.* Helicobacter pylori genomes reveal Paleolithic human migration to the east end of Asia. *iScience*. 2022:**25**(7):104477. https://doi.org/10.1016/j.isci.2022.104477.

Szpiech ZA, Hernandez RD. Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol*. 2014:**31**(10):2824–2827. https://doi.org/10.1093/molbev/msu211.

Takeuchi S, Esaki H, Furue M. Epidemiology of atopic dermatitis in Japan. *J Dermatol*. 2014:**41**(3):200–204. https://doi.org/10.1111/1346-8138.12331.

Vasseur E, Quintana-Murci L. The impact of natural selection on health and disease: uses of the population genetics approach in humans. *Evol Appl*. 2013:**6**(4):596–607. https://doi.org/10.1111/eva.12045.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol*. 2006:**4**(3):e72. https://doi.org/10.1371/journal.pbio.0040072.

Watanabe T. Current status of HTLV-1 infection. *Int J Hematol*. 2011:**94**(5):430–434. https://doi.org/10.1007/s12185-011-0934-4.

Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, Nakamura Y, Kamatani N. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am J Hum Genet*. 2008:**83**(4):445–456. https://doi.org/10.1016/j.ajhg.2008.08.019.

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.* Genetic signatures of high-altitude adaptation in Tibetans. *Proc Natl Acad Sci U S A*. 2017:**114**(16): 4189–4194. https://doi.org/10.1073/pnas.1617042114.

Yasumizu Y, Sakaue S, Konuma T, Suzuki K, Matsuda K, Murakami Y, Kubo M, Palamara PF, Kamatani Y, Okada Y. Genome-wide natural selection signatures are linked to genetic risk of modern phenotypes in the Japanese population. *Mol Biol Evol*. 2020:**37**(5):1306–1316. https://doi.org/10.1093/molbev/msaa005.

Yasunami M, Nakamura H, Tokunaga K, Kawashima M, Nishida N, Hitomi Y, Nakamura M. Principal contribution of HLA-DQ alleles, DQB1*06:04 and DQB1*03:01, to disease resistance against primary biliary cholangitis in a Japanese population. *Sci Rep*. 2017:**7**(1):11093. https://doi.org/10.1038/s41598-017-11148-6.

Yoo S-K, Kim C-U, Kim HL, Kim S, Shin J-Y, Kim N, Yang JSW, Lo K-W, Cho B, Matsuda F, *et al.* NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Med*. 2019:**11**(1):64. https://doi.org/10.1186/s13073-019-0677-z.

Zhou X-Y, Zhu F-M, Li J-P, Mao W, Zhang D-M, Liu M-L, Hei A-L, Dai D-P, Jiang P, Shan X-Y, *et al.* High-resolution analyses of human leukocyte antigens allele and haplotype frequencies based on 169,995 volunteers from the China bone marrow donor registry program. *PLoS One*. 2015:**10**(9):e0139485. https://doi.org/10.1371/journal.pone.0139485.