HUMAN GENETICS

Decoding triancestral origins, archaic introgression, and natural selection in the Japanese population by whole-genome sequencing

Xiaoxi Liu[1,2], Satoshi Koyama[3,4,5], Kohei Tomizuka[1], Sadaaki Takata[6], Yuki Ishikawa[1], Shuji Ito[1,7,8], Shunichi Kosugi[1], Kunihiko Suzuki[6], Keiko Hikino[9], Masaru Koido[1,10], Yoshinao Koike[1,7,11], Momoko Horikoshi[12], Takashi Gakuhari[13], Shiro Ikegawa[7], Kochi Matsuda[14,15], Yukihide Momozawa[6], Kaoru Ito[3], Yoichiro Kamatani[1,10], Chikashi Terao[1,2,16]*

We generated Japanese Encyclopedia of Whole-Genome/Exome Sequencing Library (JEWEL), a high-depth whole-genome sequencing dataset comprising 3256 individuals from across Japan. Analysis of JEWEL revealed genetic characteristics of the Japanese population that were not discernible using microarray data. First, rare variant–based analysis revealed an unprecedented fine-scale genetic structure. Together with population genetics analysis, the present-day Japanese can be decomposed into three ancestral components. Second, we identified unreported loss-of-function (LoF) variants and observed that for specific genes, LoF variants appeared to be restricted to a more limited set of transcripts than would be expected by chance, with *PTPRD* as a notable example. Third, we identified 44 archaic segments linked to complex traits, including a Denisovan-derived segment at *NKX6-1* associated with type 2 diabetes. Most of these segments are specific to East Asians. Fourth, we identified candidate genetic loci under recent natural selection. Overall, our work provided insights into genetic characteristics of the Japanese population.

INTRODUCTION

Whole-genome sequencing (WGS) datasets are invaluable resources for human genetic and biomedical research (*1*). Through comprehensive profiling of genetic variants, WGS data have enabled various in-depth analyses. These analyses have yielded insights into the characteristics of human genome variation (*2*), unveiled complex histories of human populations (*3*, *4*), and shed light on the processes of evolutionary adaptation and positive selection (*5*, *6*). In terms of application in genetics, WGS datasets are indispensable for imputation analysis. Large-scale WGS datasets have made it possible to construct multiethnic or population-specific reference panels (*7*, *8*). By accurately inferring ungenotyped variants from microarray data, imputation analysis effectively boosts the power of genome-wide association studies (GWASs), enables fine-mapping, and facilitates transethnic meta-analysis (*9*). Furthermore, WGS datasets provide a rich source of variants, including those that are rare, specific to certain populations, or predicted to be deleterious or loss of function (LoF) (*10*). These variants can be investigated not only for associations with various diseases but also for the effects of human knockouts, providing opportunities to identify their functional roles in both physiological and pathological processes and hence to explore the possibilities as targets for drug development (*11*, *12*). Accordingly, WGS datasets are essential to precise genetic analysis and the development of personalized medicine.

Currently, large-scale population-wide WGS data have been disproportionately represented by individuals of European descent, and substantial contributions have been made by projects such as U.K. Biobank (*13*), FinnGen (*14*), deCODE (*15*), among others. The Eurocentric imbalance in genomic data could result in unequal benefits of precision medicine and raise health disparity concerns (*16*). For example, polygenic risk scores often showed several times greater accuracy for individuals with European ancestry compared to other ancestries (*17*). Recognizing the importance of capturing the broader spectrum of human genetic variation to implement personalized medicine tailored for a specific population, concerted efforts have been made to sequence samples in more diverse ethnic groups such as in Trans-Omics for Precision Medicine and in *All of Us* project (*18*, *19*). In this context, noteworthy progress has also been made in generating WGS data from East Asian (EA) populations. Key initiatives such as GenomeAsia 100K (*20*), SG10K consortium (*21*), ChinaMap project (*22*), and Westlake BioBank for Chinese have been established (*23*). These efforts collectively uncover a wider range of genetic variants in EA populations, thereby enriching our understanding of this region's genetic diversity. Regarding WGS data from the Japanese population, notable efforts have been made by the Tohoku Medical Megabank (ToMMo) project (*24*). Nagasaki *et al.* (*25*) conducted a

[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]Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [4]Medical and Population Genetics and Cardiovascular Disease Initiative, Broad Institute of Harvard and MIT, Boston, MA, USA. [5]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [6]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [7]Laboratory for Bone and Joint Diseases, RIKEN Center for Medical Sciences, Tokyo, Japan. [8]Department of Orthopedic Surgery, Faculty of Medicine, Shimane University, Izumo, Japan. [9]Laboratory for Pharmacogenomics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [10]Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [11]Department of Orthopedic Surgery, Hokkaido University Graduate School of Medicine, Sapporo, Japan. [12]Laboratory for Genomics of Diabetes and Metabolism, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [13]Institute for the Study of Ancient Civilizations and Cultural Resources, College of Human and Social Sciences, Kanazawa University, Kanazawa, Japan. [14]Laboratory of Genome Technology, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [15]Laboratory of Clinical Genome Sequencing, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [16]The Department of Applied Genetics, The School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan.
*Corresponding author. Email: chikashi.terao@riken.jp

WGS of 1070 Japanese individuals recruited from the northeastern area of Japan. This study identified rare genetic variants and structural variants (SVs) and generated a Japanese-specific reference panel. Subsequent sequencing efforts from ToMMo and others have continued, and summary-level allele frequencies (AFs) based on WGS of 3500 and 8300 Japanese individuals have been reported (*26*, *27*). In addition, AF data based on a continually increasing number of individuals are available in the Japanese Multi-Omics Reference Panel database and the TogoVar database (*27*, *28*). These datasets offer valuable information as a catalog of genetic variations in the Japanese population and are important for variant interpretation in the context of genetic counseling. Recently, the National Center Biobank Network has released WGS data from 9287 individuals to aim primarily for use as common control samples, further enriching the Japanese genetic data resource (*29*).

Here, we generated Japanese Encyclopedia of Whole-Genome/Exome Sequencing Library (JEWEL), a comprehensive WGS using samples from Biobank Japan (BBJ)—one of Japan's largest biobanks and a leading entity in biobank research across Asia (note S1) (*30*, *31*). Differing from ToMMo, which is based on the general population in the northeastern area of Japan, BBJ was established as a nationwide patient-based biobank to advance the genomic medicine research (*32*). JEWEL, by sampling from diverse geographic regions, aims to better capture the genetic diversity of the Japanese. Principal components analysis (PCA) has identified a dual population structure of Japanese consisting of the main-island cluster and Ryukyu cluster, and recent studies have highlighted substantial genetic heterogeneity within main-island Japanese (*33*–*35*). Using WGS, JEWEL offers an opportunity to further explore the fine-scale population structure. In addition, extensive efforts have been made to collect and curate deep phenotypes through a review of medical records, follow-up surveys, and examinations in BBJ. These include primary and secondary disease diagnoses, longitudinal clinical test results, past medical history, family history, and survival information. As a result, JEWEL is enriched with potentially pathogenic variants associated with diseases, and detailed clinical information permits targeted examination of carriers of particular interest. In this study, we present in-depth analyses that includes a reexamination of the genetic structure using both common and rare variants, characterization of LoF variants and human knockouts, and identification of archaic segments likely introgressed from Neanderthals or Denisovans. Last, we attempted to identify genetic loci potentially targeted by selection in the Japanese population.

RESULTS

Characteristics of the JEWEL WGS dataset

A total of 3256 individuals, enrolled from medical institutes in seven geographic regions across Japan, were sequenced to generate JEWEL. These regions include Hokkaido, Tohoku, Kanto, Chubu, Kansai, Kyushu, and Okinawa, which are hereafter referred to as North, Northeast, East, Central, West, South, and Okinawa (see Materials and Methods and Fig. 1A). All regions except for Okinawa are located on the main islands of the Japanese Archipelago, commonly known as Hondo, while the term Okinawa in this study indicates the Ryukyu islands. The relative sample size proportionally reflects the population sizes of these regions in Japan (table S1). Sequencing was performed according to standard Illumina protocols, and an average WGS coverage depth of 25.6× was achieved. Variant calling was conducted according to established Genome Analysis Toolkit (GATK) best practices (see Materials and Methods and note S2 for details). The final dataset consisted of 45,586,919 single-nucleotide variants and 9,113,420 insertions or deletions (indels) from 23 chromosomes. We observed that 61 and 40% of variants were not registered in the Genome Aggregation Database (gnomAD) and ToMMo, respectively (*26*, *36*) (table S2); 15,410,953 (32.7%) variants were only observed in JEWEL. Compared to microarray genotyping data, a high genotype concordance rate of 99.971% was obtained (see Materials and Methods). Using 42,389,421 biallelic autosomal single-nucleotide variants, we estimated the ratio of transition to transversion (Ti/Tv) to be 2.11, which was in line with recent large-scale WGS analyses (*21*, *22*) (tables S2 and S3). These results confirmed that JEWEL dataset is of high quality in various aspects, allowing for a deeper analysis of the genetic characteristics of this population.

Triancestral origins of the Japanese population

We first conducted a conventional PCA based on 184,036 independent pruned common variants (see Materials and Methods). Consistent with previous studies, the analysis replicated the classic "dual-cluster" structure consisting of Okinawa and the Hondo clusters (Fig. 1B) (*33*, *35*, *37*). We hypothesized that rare variants might be more informative in revealing the population structure, and we conducted a PCA–Uniform Manifold Approximation and Projection (PCA-UMAP) analysis, which exclusively used 1,835,116 independent pruned rare variants (see Materials and Methods). The analysis uncovered an unprecedentedly fine structure of the Japanese population (Fig. 1C). This structure, resembling a "hummingbird," not only recapitulated patterns obtained from PCA based on common variants but also highlighted several notable features. Specifically, we observed (i) a clearer separation among subregions of Hondo and a clearer distinction of Okinawa cluster from Hondo cluster, (ii) Northeast individuals clustered in a thin, narrow area, and (iii) additional subclusters of individuals from West and South (figs. S1 and S2 and note S3).

To gain a deeper insight into the population structure, we performed an unsupervised ADMIXTURE analysis based on common variants (see Materials and Methods and note S4). To determine the optimal K value, we used Structure Selector, a method demonstrated to exhibit superior performance compared to other estimators (*38*). In this analysis, all four metrics support the K value of three as the optimal number of ancestral components (fig. S3). In addition, we used badMIXTURE to evaluate the goodness of fit and observed no systematic pattern of large residuals, indicating an overall good fit at $K = 3$ (fig. S4) (*39*). Therefore, our data suggested that the Japanese population could be best modeled by admixtures of three ancestral components (hereafter K1 to K3). K1 to K3 were the highest in Okinawa, Northeast, and West, respectively (Fig. 1D and table S4). K1 (Okinawa) component maintains a relatively stable fraction of around 12% in Hondo subgroups, except for South (which is a region adjacent to Okinawa), with a higher proportion of 22%. K2 (Northeast) and K3 (West) components showed a cline from West to East. We also conducted the ADMIXTURE analysis using both common and rare variants and observed consistent results, with additional detail from Okinawa (note S4).

We observed significant correlations between K values and PCA-UMAP values, despite the former obtained from the analysis of common variants and the latter from rare variant analysis. This finding seemed to offer additional support for $K = 3$. Specifically, UMAP1 is significantly correlated with K2/K3 (Pearson coefficient = −0.69 with

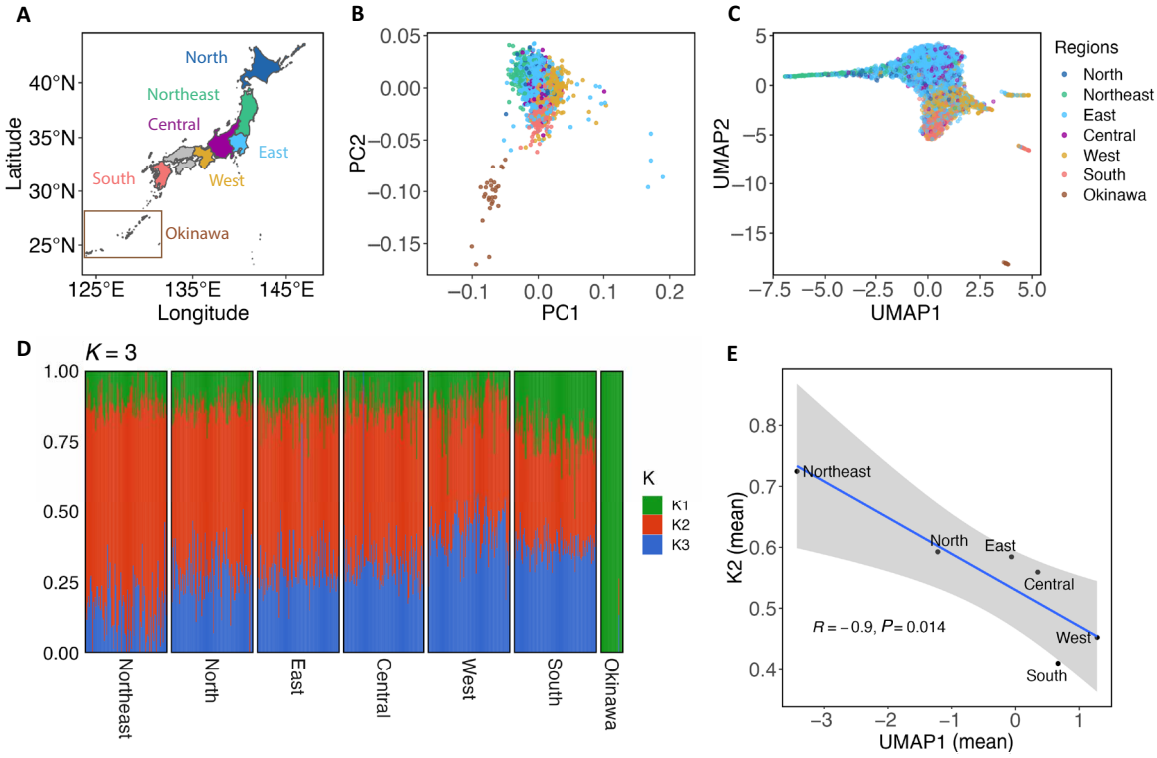

Fig. 1. Fine-scale genetic structure of the modern Japanese and its three ancestry origins. (**A**) Geographic regions in Japan from which the samples were recruited are described. These regions include the Japan archipelago, commonly known as Hondo, and the Ryukyu archipelago, which is termed as Okinawa in this study. The number of individuals from each region is provided in table S1. (**B**) PCA analysis based on common variants with a minor AF (MAF) ≥ 0.01. Individuals are colored according to their recruitment regions. (**C**) Rare variant–based PCA-UMAP analysis (0.001 ≤ MAF < 0.01) is displayed. (**D**) ADMIXTURE analysis with *K* set to 3. For regions other than Okinawa, 100 individuals were randomly selected and plotted. All 28 individuals from Okinawa were included in the plot. K1 represents Okinawa, while K2 and K3 are the highest in the Northeast and West, respectively. (**E**) UMAP1 is negatively correlated with the fraction of K2 ancestry. The correlations between each combination of UMAP and *K* are presented in fig. S5.

K2 and 0.60 with K3, $P < 2.2 \times 10^{-16}$ for both). This correlation pattern can also be clearly visualized by aggregating samples according to their respective regions (Fig. 1E and fig. S5). We additionally analyzed *K* values in the context of geography and found that the proportions of Okinawa (K1) and Northeast (K2) ancestries are correlated with geographic longitude. In contrast, the correlation with West (K3) is less pronounced and not statistically significant (fig. S6).

We attempted to gain hints about the potential ancestral origins of K1 to K3. Previous studies have suggested that Japanese carry Jomon and EA ancestry (represented by Han Chinese) (*34*, *40*). Recently, the presence of Northeast Asian (NEA) ancestry has been proposed on the basis of analyses of ancient genomes (*41*, *42*). In this context, we analyzed our data together with modern and ancient genetic data of Jomon, EA, and NEA. Using f_4 ratio statistic, we estimated that Okinawa had the highest Jomon ancestry (28.5%), followed by Northeast (18.9%), and the lowest in West (13.4%) (see Materials and Methods and table S5). These results align with prior studies demonstrating a high genetic affinity between Jomon and Okinawa people (*43*, *44*). Next, on the basis of outgroup f_3 statistic, we observed that individuals from West had the highest shared genetic drift with Han Chinese (table S6). We then used f_4 statistic in the form of f_4 (Mbuti, ancient genome; Northeast, West) to evaluate differential genetic affinities between Northeast and West, in relation to ancient genomes reported from China, Korea, and Japan (*41*, *44*–*47*). Our results indicated a significantly closer relationship between West and ancient Chinese groups around the Yellow River (YR) or upper YR region, specifically in the Middle Neolithic (MN) and Late Neolithic periods (table S7). In contrast, individuals of Northeast showed significantly higher genetic affinities with Jomon and ancient Japanese genome from Miyako Island in Okinawa (which had a high Jomon proportion) and ancient Koreans from the Three Kingdoms (TK) period (Korea-TK_2) (fourth to fifth century CE) (table S7). These results align with reports indicating that ancient Japanese in the Yayoi period and certain ancient Korean groups had a high proportion of Jomon ancestry (*42*, *47*).

We subsequently used qpAdm to estimate contributions of NEA, EA, and Jomon ancestries in each subgroup, following the approach described in prior studies (*41*, *48*) (see Materials and Methods). For this analysis, the Chinese Han was designated as representative of EA, while China_WLR_BA_o and China_HMMH_MN were grouped to represent NEA. The results revealed a generally good fit of the tripartite model to our dataset (table S8). The proportions and trends of Jomon ancestry estimated through qpAdm align with the findings from the f_4 ratio test, revealing the highest proportion in Okinawa (25%) and the lowest in the West (7.5%). Likely because of the low Jomon ancestry in West, we observed that EA ancestry is the highest in South rather than West. However, the fitting of this model for Northeast was rejected, indicated by an extreme *P* value ($P = 6.5 \times 10^{-4}$). Exploring

additional models, we found that Northeast could be alternatively modeled as a two-way admixture of Korea-TK_2 (68%) and Han (32%) (tables S8 and S9). Notably, among Hondo groups, Northeast showed the highest proportion of Korea-TK_2. For West, the initial three-way model that includes NEA, EA, and Jomon showed a better fit, as indicated by a lower chi-square value (9.14 compared to 11.8). Furthermore, the two-way admixture modeling involving combinations of Jomon, EA, and NEA proved to be unsuccessful (table S9). These multiple lines of evidence suggest that K1 and K3 may be linked to Jomon and EA ancestries. Although less clear, the ancestral origins of K2 could potentially be connected to ancient populations in Japan and the Korean Peninsula, such as Korea-TK_2.

Motivated by the above findings, we investigated whether this tri-ancestral framework could offer insights into the likely origins of Japanese founder mutations. We focused on two high-frequency pathogenic mutations associated with hereditary breast cancer among Japanese patients—the *BRCA1* Leu63Ter and the *BRCA2* c.5576_5579delTTAA frameshift mutation. The former is specific to the Japanese population and has a significantly higher frequency in Eastern Japan than in Western Japan (*49*). In contrast, the latter has a high frequency in Western Japan and has been reported in other Asian populations, including Chinese (*50*) and Korean (*51*). Plotting *BRCA1* Leu63Ter carriers in the PCA-UMAP showed that this mutation predominantly occurred in individuals with likely Northeastern ancestry, and its occurrence is significantly associated with UMAP1 ($P = 9.04 \times 10^{-6}$, logistic regression) (fig. S7). This pattern was not apparent when considering enrollment locations, as most carriers were recruited from East (seven of nine carriers were recruited from East, with the remaining two from North and Northeast). On the other hand, the *BRCA2* c.5576_5579delTTAA mutation was predominantly observed in individuals of West ancestry (fig. S7). Our data align with a recent study based on ~100,000 Japanese samples, showing that *BRCA1* Leu63Ter has the highest frequency in Northeast, while the *BRCA2* frameshift mutation is most frequent in West (*52*). Despite our much smaller sample size, the rare variant–based fine structure sheds insights into the likely origins of the two mutations in Japanese. The data suggested that the *BRCA1* Leu63Ter mutation likely originated in Northeast ancestry and spread to other regions. Since Japanese in West had a higher genetic affinity with Han Chinese, we speculate that this mutation may have been introduced to Japan from continental Asia. In addition, we explored whether *K* values are associated with quantitative phenotypes in JEWEL individuals based on linear regression. We found significant associations, particularly for total cholesterol ($P = 2.69 \times 10^{-13}$) and prothrombin time (PT; $P = 1.33 \times 10^{-12}$) with K1. Comparable *P* values of these traits with K2 were also observed (table S10).

LoF variants and human knockouts

JEWEL dataset allowed us to explore potentially clinically important protein-coding variants in Japan. In our analysis, we identified 18,481 LoF variants in 9045 genes, including 9780 LoF variants not registered in gnomAD or ToMMo (4.7K), with a substantial proportion of these being rare (Fig. 2A and table S11). These LoF variants are defined as variants that may cause premature stop codons (stop-gained), small-sized indels that shift the coding sequence (frameshift), or variants that change two immediately adjunct nucleotides to the splicing sites (splicing variants). Furthermore, we classified 177,112 synonymous variants and 306,923 missense variants, which affected 18,651 and 19,103 genes, respectively (Fig. 2B). Examination of LoF variants together with carriers' UMAP values identified 32 and 37 LoF variants, whose frequencies were significantly associated with UMAP1 and UMAP2 (false discovery rate < 5%), respectively (see Materials and Methods and table S12). We noticed that individuals from Northeast had the lowest average number of singleton coding variants compared to those from other regions (table S13). Since the sample size of Northeast is smaller than that in other Hondo regions, we conducted a random resampling analysis and confirmed that this observation is likely not attributable to sample size (table S14). We speculate that other factors, such as demographic history, especially population expansion, may be influencing this observation. Despite regional differences, the ratio between singleton missense and singleton synonymous variants (dN/dS) across regions was consistently close to 2, which is an observed ratio of de novo missense and synonymous variants reported in an in vivo study (*53*). Furthermore, consistent with observations in another report, this ratio negatively correlates with the AF, suggesting that many rare missense variants might be deleterious but remain in the gene pool (*54*). To further test this idea, we calculated the missense risk score by integrating annotations from 30 different annotation tools (see Materials and Methods). We observed that the missense risk score increased as the AF decreased ($P < 2.2 \times 10^{-16}$, Pearson correlation test). On average, singletons exhibited the highest risk scores (table S15). On the basis of the data above, missense variants that are rare in the general population could be prioritized for disease association analysis. This approach to prioritization could narrow down potential candidates, thereby increasing the likelihood of identifying a meaningful clinical connection.

JEWEL allowed us to further assess the potential applicability of LoF observed/expected upper-bound fraction (LOEUF) scores in the Japanese population. The LOEUF score was introduced as a metric to quantify a gene's tolerance to LoF variants, based on observed and expected counts of LoF variants in the gnomAD project (*36*). Given that individuals with EA ancestry constituted 7% of the gnomAD dataset, we are interested in testing whether LOEUF score is applicable to JEWEL. We observed that genes in the lowest LOEUF decile bin (indicating the highest intolerance to LoF variants) were least affected by LoFs (fig. S8). This supports the utility of LOEUF scores in stratifying genes highly intolerant to LoF variants. However, a discrepancy was found in the number of genes affected by LoF variants in top decile bins (fig. S8). Furthermore, we observed that the fraction of transcripts affected by LoF variants showed a significant positive correlation with LOEUF bins (Fig. 2C). Overall, these results support the generalizability of LOEUF score while acknowledging that there might be room for improvement in relation to LoF-tolerant genes.

Pathogenic variants and human knockouts are highly valuable for clinical research and drug development and may reveal human genotype-phenotype connections. We identified 371 ClinVar-registered pathogenic variants and 1723 unreported LoF variants in genes harboring pathogenic variants in ClinVar (note S5). We searched for human knockouts, defined as homozygotes or compound heterozygotes for LoF variants. Inspection of annotations and manual curation identified 23 human knockouts that are likely to be clinically relevant. We noted a carrier of compound heterozygous LoF variants in the *ABCC2* gene (see Materials and Methods and table S16). The LoF of this gene is known to cause Dubin-Johnson syndrome, an autosomal recessive liver disease related to hyperbilirubinemia (*55*, *56*). The syndrome is typically benign, and patients exhibit an increase in total bilirubin in the

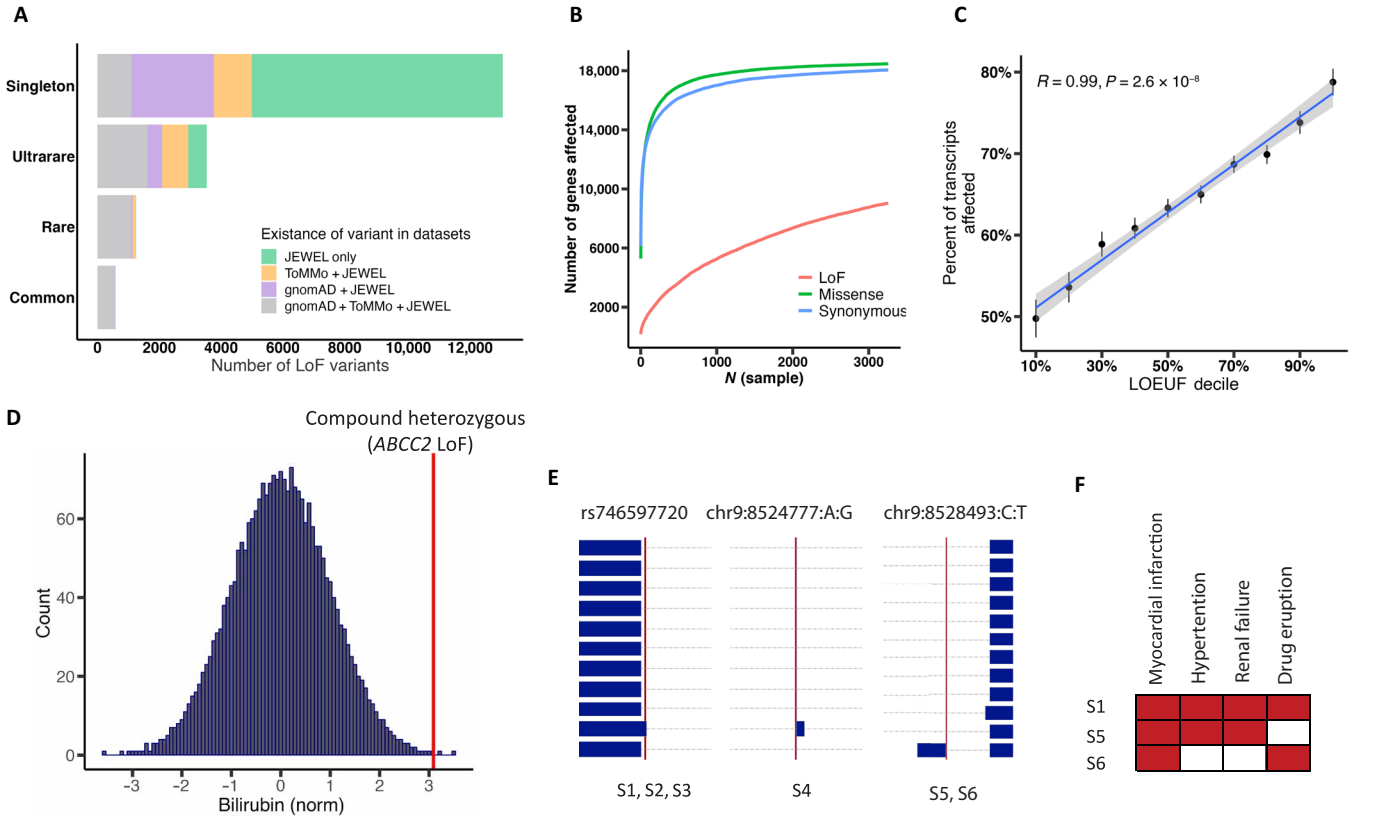

Fig. 2. LoF variants and human knockout in the JEWEL dataset. (**A**) Number of known and unregistered LoF variants compared with gnomAD database (v2.1.1) and ToMMo (4.7K). Variants are categorized into four AF bins. Common: MAF > 1%; rare, MAF < 1% and MAF ≥ 0.01%; ultrarare: minor allele count > 1 and MAF < 0.01%; singleton. (**B**) Cumulative number of genes affected by LoF, missense, and synonymous variants. (**C**) The average percentage of transcripts affected by LoF variants, categorized by the genes' LOEUF deciles. Genes that are highly intolerant to functional variation, as indicated by lower LOEUF deciles, have fewer affected transcripts compared to genes that are more tolerant. Error bars are included to indicate SEs. (**D**) The histogram of normalized total bilirubin levels among individuals in the JEWEL cohort. The red line highlights an individual with compound heterozygous LoF variants in the *ABCC2* gene, ranking third in the whole JEWEL dataset. This elevated level of total bilirubin is consistent with the clinical phenotype of Dubin-Johnson syndrome, which is caused by the inactivation of the *ABCC2* gene. (**E**) The plot presents data on six individuals carrying LoF variants in the *PTPRD* gene. The identifier for each LoF variant, either rsID or variant ID, is displayed at the top. Blue boxes represent exons for different transcripts, while the red lines mark the locations of these LoFs. Individual IDs carrying the LoF variants are indicated at the bottom. A zoomed-out perspective of the plot is presented in fig. S9. (**F**) The shared phenotypes among three *PTPRD* LoF carriers for whom comprehensive clinical data are available (S1, S5, and S6), with the names of the phenotypes provided for reference.

blood, leading to chronic jaundice. We obtained clinical history records and blood test results for this individual and confirmed the diagnosis of Dubin-Johnson syndrome and the clinical manifestation of hyperbilirubinemia (Fig. 2D). Furthermore, two of three individuals with homozygous LoF variants in *GJB2*, a gene associated with nonsyndromic sensorineural hearing loss, were confirmed to have hearing loss (*57*). These examples demonstrate that we can use JEWEL to identify likely underlying pathogenic variants responsible for diseases and to mine potentially clinically relevant genotype-phenotype connections.

In addition to conventional human knockout analyses presented above, we leveraged rich phenotyping data in JEWEL to examine individuals with heterozygous LoF variants in genes considered highly intolerant to LoF variants, as indicated by LOEUF scores. Focusing on genes that have multiple LoF variants, we identified six individuals with LoF variants in *PTPRD*, one of the top-ranked LOEUF genes (LOEUF = 0.11, rank = 271 among 19,704 genes), which encodes a receptor-like protein tyrosine phosphatase (Fig. 2E) (*58*). Detailed clinical information was obtained for three of the six individuals, who presented with several shared phenotypes, including myocardial infarction, kidney failure, hypertension, and drug eruption (Fig. 2F and table S17). The *PTPRD* gene has 13 transcripts with most exons being identical and shared among multiple transcripts. However, only two transcripts were affected by LoF variants, which is significantly fewer than would be expected by chance ($P = 0.005$, permutation test; see Materials and Methods, Fig. 2E, and fig. S9). We searched the literature for reported human knockout of *PTPRD*. A case report described a child carrying homozygous microdeletion of *PTPRD*, which was suspected to be associated with intellectual disability, trigonocephaly, and hearing loss (*59*). In addition, *Ptprd* knockout mice exhibit preweaning lethality with an incomplete penetrance (*60*). Given these data and the low LOEUF score, disruption of PTPRD protein might be highly deleterious. However, if LoFs affect only a limited number of transcripts or if the affected transcripts are of lesser functional importance, then the consequences might be more tolerable. Further genome-wide scanning identified additional

genes where LoF variants occurred in a restricted set of transcripts, including two more PTPR family genes, both of which are in the lowest LOEUF bin, *PTPRS* (LOEUF = 0.25, $P = 0.002$) and *PTPRM* (LOEUF = 0.23, $P = 0.009$) (table S18). The results suggest that phenotypic impacts of certain LoFs may be mitigated, even in genes that are generally intolerant to LoF. However, other factors such as nonrandom sampling or inaccurate annotation of LoF transcripts should also be considered. Further studies using WGS from either the Japanese population or other populations are needed. Seen as examples above, we highlight the necessity to integrate genetic information with in-depth clinical data to understand the full spectrum of gene functions when potentially disrupted by LoF. These findings also suggest that tolerability to LoF should be evaluated not only at the gene level but also at the transcript level.

Sequences introgressed from Neanderthals and Denisovans

EAs carry introgressed sequences from Denisovans and Neanderthals (*61*–*63*). However, the surveys of introgression have so far been restricted to a small number of samples in East Asia. To detect sequences likely introgressed from Neanderthals or Denisovans, we applied a recently developed probabilistic method, IBDmix, which does not use a modern reference population (see Materials and Methods). On an individual basis, the individual in JEWEL carries ~49 Mb of Neanderthal-derived sequences and 1.47 Mb of Denisovan-derived sequences (table S19). In total, we identified 3079 segments likely introgressed from Neanderthals and 210 segments likely introgressed from Denisovans, covering 772 and 31.46 Mb of the genome, respectively (Fig. 3A). Our results replicated 85% (2414 of 2843) of previously reported Neanderthal-introgressed segments based on the analysis of 104 Japanese in the 1000 Genomes project (1KGP) (fig. S10) (*63*). Notably, 47% (1439 of 3079) of Neanderthal-introgressed regions were not identified by the 1KGP Japanese in Tokyo, Japan (JPT) dataset, and 77% (1113 of 1439) of them were rare, with frequencies less than 5%. PCA of introgressed Neanderthal segments in JEWEL revealed no subregional differences (fig. S11). We compared Denisovan introgression in JEWEL to that in populations from the 1KGP dataset, as well as in Papuans and Philippine Ayta, both of which have a high proportion of Denisovan ancestry (*62*, *64*). The analysis revealed that the Denisovan-like segments in JEWEL significantly overlap with those in EA populations, while no statistical significance was found with those in Papuan and Philippine Ayta, indicating that Denisovan introgression in Japanese might be less relevant to that in Papuan and Philippine Ayta (table S20 and note S6).

Subsequently, we examined the phenotypic effects of the identified introgressed sequences on 106 traits based on GWAS summary statistics generated from BBJ (see Materials and Methods). We identified 44 archaic segments associated with 49 phenotypes (2 from Denisovans and 42 from Neanderthals). Among these, 43 associations have not been reported in comparison to a previous study (*65*). We validated 39 of 44 archaic segments by an alternative method SPrime and confirmed that 5 segments not detected by SPrime showed a high matching rate with the Neanderthal genome (see Materials and Methods) (*62*). The Denisovan-inherited segment at *POLR3E* was associated with height. The segment at *NKX6-1* was

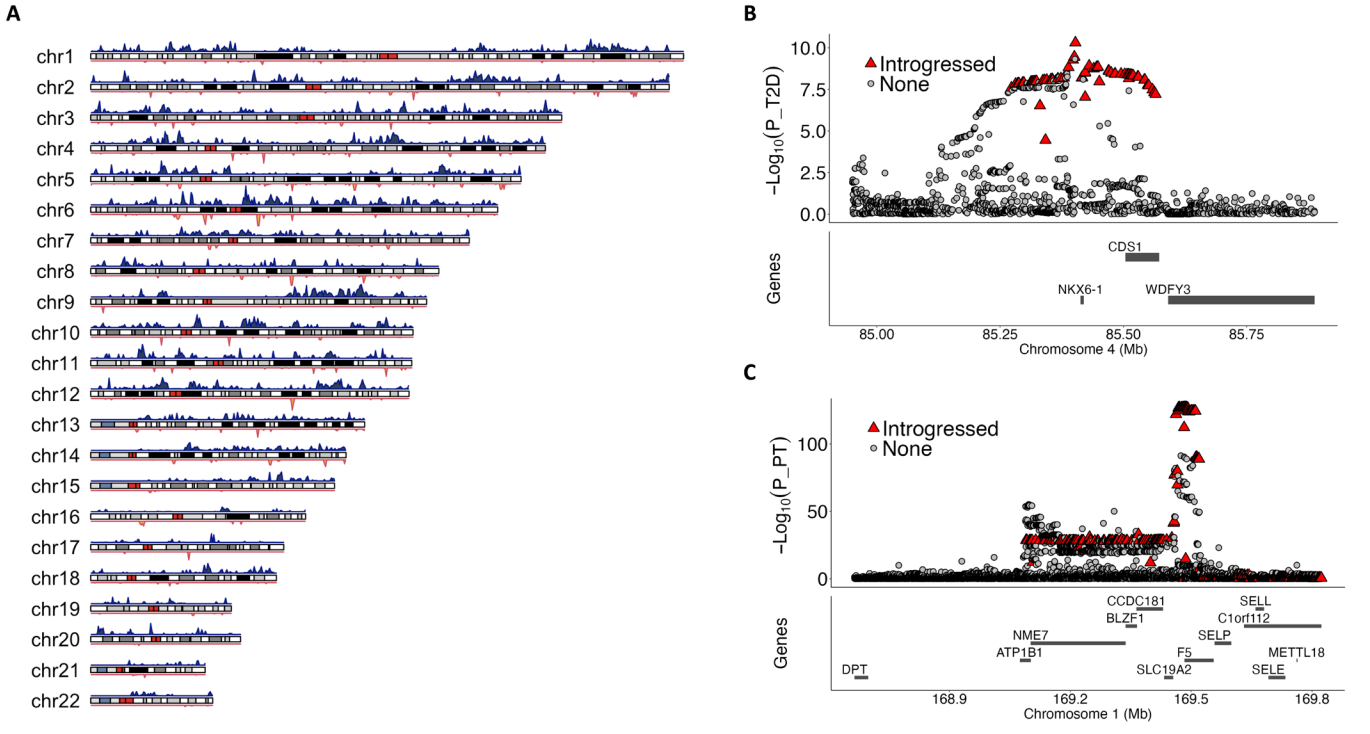

Fig. 3. Introgressed sequences from archaic Neanderthals or Denisovans in the Japanese population. (**A**) Density plot illustrating the distribution of introgressed sequences across each chromosome. The upper track, shown in blue, represents sequences likely introgressed from Neanderthals, while the lower track displays sequences originating from Denisovans. (**B**) Variants likely introgressed from Denisovans in the *NKX6-1* locus are associated with T2D in the Japanese population. The triangle indicated the introgressed variants, and the gray dots indicated the nonintrogressed variants. (**C**) Introgressed variations from the Neanderthals in the *F5* gene are associated with PT.

associated with type 2 diabetes (T2D) (Fig. 3B and Table 1). The *NKX6-1* segment has also been identified in other populations, including Papuans, Chinese [Han Chinese in Beijing (CHB) and Han Chinese South (CHS)], and Finnish (*62*). Moreover, archaic variants in this segment were found to be associated with T2D using GWAS data obtained from the FinnGen project ($P_{min} = 8.65 \times 10^{-10}$ at rs75560957) (*14*). For Neanderthal-derived segments, we observed 11 segments associated with seven diseases—T2D, coronary artery disease (CAD), stable angina pectoris (SAP), atopic dermatitis (AD), Graves' disease (GD), prostate cancer (PrCa), and rheumatoid arthritis (RA) (Table 1). A pathway analysis identified "regulation of insulin secretion" as the top associated pathway ($P = 1.9 \times 10^{-4}$). At the *ADAMTS7* locus, the lead introgressed single-nucleotide polymorphism (SNP), rs11639375, was reported to be protective against CAD and SAP. While this SNP is observed in all major populations with high frequencies, upon further examination, it appears that rs11639375 in Japanese resides within a haplotype that is likely to have been introgressed from Neanderthals. The haplotype comprises 39 potentially archaic variants that exhibit a strong linkage disequilibrium (LD) with rs11639375 ($r^2 > 0.7$). These variants are exclusive to EA and Latino Americans and are either absent or present at extremely low frequencies in other population groups (table S21). These data may suggest that this protective variant rs11639375 was once lost to EA and later restored through introgression. However, further analysis is needed to substantiate this hypothesis (note S7).

We observed that a causal variant for AD, rs12637953, located in the *CCDC80* locus, is likely to have been inherited from Neanderthals. This variant was implicated as potentially functional via decreasing expression levels of an enhancer in $CD1a^+$ Langerhans cells and skin epidermis cells by machine learning in silico prediction and was further experimentally validated (*66*, *67*). The introgressed segment at the *GLP1R* locus deserves attention. Variants at this locus were shown to be associated with T2D in a large-scale Japanese GWAS ($n = 191{,}764$), but not in European GWAS ($N = 159{,}208$), as previously reported (*68*). Through our analysis, we identified that the lead variants likely have archaic origins, specifically from Neanderthals. Further analyses using 1KGP data showed that this introgressed segment is present in Asians but absent in Europeans, which could account for the discrepancies in GWAS signals. In addition to archaic segments associated with diseases, we identified 37 distinct segments associated with 35 quantitative traits (table S22). As an example, archaic variants of the coagulation factor V (*F5*) gene showed positive associations with the bleeding trait (PT) (Fig. 3C). Notably, the same segment is associated with PT in the Icelandic population (*69*). We also confirmed that the Neanderthal-derived segment reported to be associated with severe COVID-19 (chr3: 45,859,651 to 45,909,024) was not detected in JEWEL (*70*). Last, the significant introgressed variants exhibited distinct population specificity in EAs compared to Europeans (fig. S12). The AFs were significantly higher in JEWEL compared to Europeans ($P = 4.66 \times 10^{-8}$, paired *t* test), and the

Table 1. Introgressed segments associated with disease phenotypes in the Japanese population.

Introgressed segment	Lead archaic SNP	Reported *P*	Disease	Beta	Origin	Gene
chr4: 85200961–85426528	4:85301870:T:C	4.91×10^{-11}	T2D	−0.134	Denisovan	*NKX6-1*
chr1: 39932346–40124123	1:39981740:G:A	3.16×10^{-8}	T2D	0.062	Neanderthal	*BMP8A*
chr1: 160151058–160608637	1:160419940:A:G	3.29×10^{-13}	GD	0.470	Neanderthal	*VANGL2*
chr2: 164906091–165538059	2:165381518:A:G	6.69×10^{-10}	T2D	−0.172	Neanderthal	*GRB14*
chr2: 173140874–173598206	2:173321791:T:G	5.07×10^{-12}	PrCa	−0.175	Neanderthal	*ITGA6*
chr3: 23163800–23502216	3:23210938:C:G	3.33×10^{-15}	T2D	0.100	Neanderthal	*UBE2E2*
chr3: 111531421–113933832	3:112394029:T:C	2.88×10^{-14}	AD	1.248	Neanderthal	*CCDC80*
chr6: 38249704–39053462	6:39037662:G:C	1.09×10^{-17}	T2D	−0.092	Neanderthal	*GLP1R*
chr10: 63625277–64526183	10:64063077:T:C	1.26×10^{-8}	RA	0.212	Neanderthal	*ZNF365*
chr12: 31070734–32216996	12:31441179:A:C	4.14×10^{-25}	T2D	0.112	Neanderthal	*FAM60A*
chr15: 78635757–79216385	15:79019990:C:T	3.79×10^{-10}	SAP	−0.078	Neanderthal	*ADAMTS7*
chr15: 78635757–79216385	15:79026723:G:A	2.90×10^{-15}	CAD	−0.079	Neanderthal	*ADAMTS7*

median AF in the Japanese population is 21.5 times that of the AF in the European population.

Evolutionary selection profile in the Japanese population

We conducted genome-wide scans to detect candidate genomic loci that were likely subject to selection in the Japanese population with two methods: integrated haplotype score (iHS) analysis and FastSMC. The iHS method is effective at identifying selective sweeps based on phased haplotype information (*71*). FastSMC is an extension of the ASMC algorithm designed to rapidly identify pairwise identical-by-descent (IBD) regions at a specified coalescence time. By inferring IBD sharing, the analysis could identify regions that are overinherited from a limited number of common ancestors, potentially indicating recent positive selection (e.g., a quick frequency rise of a favorable haplotype) (*72*). By iHS, we identified three loci under positive selection at the genome-wide significance threshold ($P_{iHS} = 8.24 \times 10^{-9}$), including major histocompatibility complex (MHC), alcohol dehydrogenase (*ADH*) cluster, and *ALDH2* (Table 2 and Fig. 4A). The quantile-quantile plot indicated that there was no systematic bias (fig. S13). We further explored potential regional differences in the selection profile across five representative regions: West, East, Northeast, South, and Okinawa. We observed similar selection profiles across Hondo regions. However, note that the signals of *ADH* cluster and *ALDH2* were relatively weaker in Okinawa and did not reach genome-wide significance (fig. S14 and table S23). These differences could be due to a limited sample size of Okinawa or maybe varying selection pressures, necessitating further study. In addition, we used the FastSMC method as a complementary approach to validating the signals observed in iHS. We initially assessed the fit of the density recent coalescence (DRC) statistic. The density plot and the quantile-quantile plot for the empirical null model indicated that gamma fitting was generally well fitting, although it may not handle large DRC values well, leading to conservative approximate *P* values (fig. S15). In total, this method identified four candidate loci potentially targeted by selection in the past 50 generations, which include three loci significant in iHS (*ADH*, *ALDH2*, and MHC), and a candidate locus 2p25.3 (Table 3 and Fig. 4B). These three loci (*ADH*, *ALDH2*, and MHC) were also detected using the singleton density score (SDS) method in a previous study (*73*), further substantiating the presence of strong selection pressure on the autoimmune system and alcohol-metabolizing pathway for the Japanese population.

DISCUSSION

In this study, we generated JEWEL, a dataset consisting of clinical and WGS data from 3256 Japanese individuals across seven different regions in Japan. This comprehensive genetic dataset enables us to delve into uncharted territories concerning population and medical genetics of the Japanese population. We highlight several unique aspects of this study. Our analysis revealed fine population structure of the Japanese, echoing and lending supports to the "tripartite origins" model. We showcased potential clinical usages of JEWEL and examined the genetic legacy of Neanderthals and Denisovans in the Japanese and investigated their associations with various phenotypes, which constitutes the largest non-European analysis to date. Furthermore, the identification of genomic loci under recent selection enriched our understanding of adaptive evolution in the Japanese population.

The rich source of variants and comprehensive inclusion of samples across Japan in JEWEL, combined with PCA-UMAP and population genetics analyses, enabled us to construct a more refined Japanese population structure and propose triancestral origins of the Japanese population. Compared to the prior PCA-UMAP analysis that used array data from BBJ, our analysis, which is based on rare variants from WGS, offers enhanced resolution for distinguishing Japanese in Hondo (*35*). We reason that this is because rare variants typically emerged more recently than common ones and could be more informative in revealing the fine-scale genetic structure. In our current analysis, all Okinawa individuals were grouped into a single cluster in PCA-UMAP. This is likely due to the limited sample size, which may not capture the known genetic heterogeneity among subpopulations from different island groups within Okinawa (*74*). By incorporating samples from diverse regions of Japan, our study reveals genetic heterogeneity in Hondo Japanese, which align well with a recent study that examined array data from 11,069 individuals across all 47 Japanese prefectures (*34*). Moreover, our study provides additional insights into the potential ancestral components of the Japanese population, which we believe may be enhanced by the unbiased selection of SNPs from WGS (note S8).

Concerning ancestral origins of the Japanese population, we recommend that our data should be interpreted in the context of existing models, including the widely accepted "dual structure" model and the recently proposed tripartite origins model. The dual structure model, which suggested that the modern Japanese population formed by the admixture of indigenous hunter-gatherer Jomon people and rice-farming Yayoi migrants from continental Asia, has been extensively studied and is considered as the primary working hypothesis (*75*–*77*). A refined model, named "innerdual structure" proposed that genetic variations exist between "Central Axis" inland regions and "Periphery" coastal areas, influenced by multiple migration waves (*78*). A recent study of ancient genomes from Yayoi and Imperial Kofun periods introduced a further refined model, suggesting that the Japanese population may have three ancestral origins: Jomon, NEA, and EA (*41*). This is an intriguing hypothesis that specifically suggests the likely origins of continental ancestry. One limitation, however, is that the number of ancient genome samples, particularly those from Yayoi and Kofun periods, remains limited. As a result, some uncertainty persists, and the hypothesis has yet to be fully validated. The presence of Jomon and EA genetic component

Table 2. Significant loci under positive selection detected by iHS analysis. BP, base pair position; DAF, derived AF; CHR, chromosome.

CHR	Position (Mb)	Cytoband	Lead SNP	DAF	Normalized iHS	P_{iHS}	Candidate gene
4	97.20–99.30	4q22.3	rs79395698	0.914619	−6.51255	7.39×10^{-11}	*ADH*
6	30.56–32.91	6p21.32	rs139510765	0.101505	6.19468	$5.84 \times 10-^{10}$	*MHC*
12	110.91–112.82	12q24.1	rs77768175	0.293151	6.41108	1.44×10^{-10}	*ALDH2*

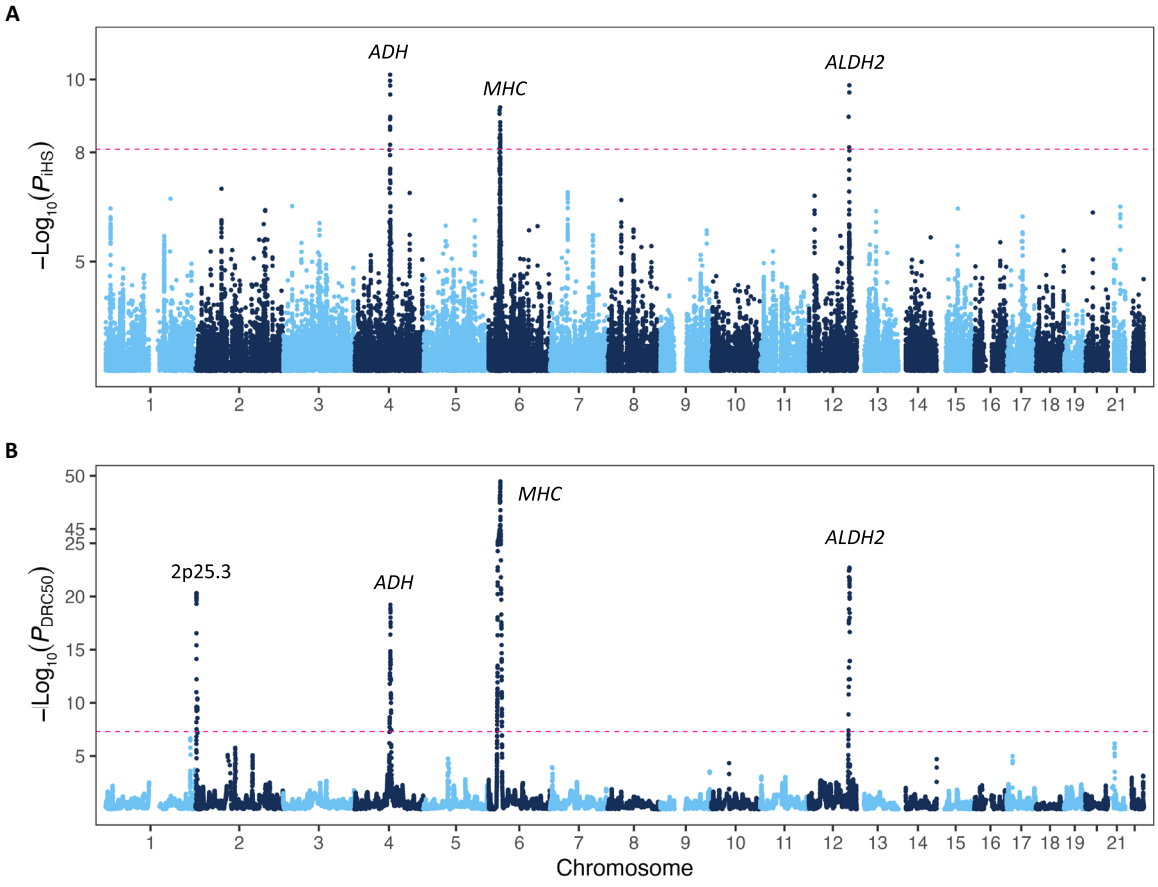

Fig. 4. Positive selection signals in the Japanese population based on iHS and FastSMC analysis. (**A**) Manhattan plot showing the genome-wide distribution of P_{iHS} for variants in autosomes in the iHS analysis. The red horizontal dashed line indicates the genome-wide significance threshold of $P_{iHS} = 8.24 \times 10^{-9}$. (**B**) Manhattan plot of the FastSMC analysis; the genome-wide significance threshold is set at $P_{DRC50} = 5 \times 10^{-8}$.

Table 3. Candidate loci detected to be under significant positive selection by FastSMC within the past 50 generations. DRC50, density of recent coalescence statistic within the past 50 generations.

CHR	Position (Mb)	Cytoband	*P*DRC50	Candidate gene(s)
2	2.79–5.27	2p25.3	4.91×10^{-21}	*ADI1/COLEC11*
4	99.71–100.06	4q23	6.18×10^{-20}	ADH cluster
6	27.05–32.74	6p21	3.30×10^{-50}	*MHC*
12	113.15–113.43	12q24	1.97×10^{-23}	*ALDH2*

(e.g., Han Chinese) has been proposed to explain the dual-cluster pattern observed in PCA of the Japanese population. In line with this, the current study and previous research indicate that Okinawa has a higher genetic affinity to Jomon, while West, or regions near West, is genetically closer to the Chinese compared to other regions in Hondo (*33*, *34*, *40*). The qpAdm analysis offers further insights into potential ancestral origins of the Japanese population. We observed a reasonable fit of the tripartite model, involving Jomon, EA, and NEA, across our dataset, with the exception of Northeast. Crucially, two-way models using pairwise combinations of Jomon, EA, and NEA did not yield successful results. This outcome adds further supports for the triancestral model and indicates that the traditional "dual-structure" model may be insufficient. The observation that West had a closer genetic affinity to Chinese is potentially associated with a substantial influx of people with EA ancestry during the post-Yayoi period, with historical evidence indicating continued migration from the Korean Peninsula through the Kofun and Nara periods (250 to 794 CE) (*76*, *79*). This continued influx may have played a role in the formation of Japan's first centralized imperial state during the Kofun period, which was established in West (in present-day Nara Prefecture) (*80*). This period also witnessed a substantial technological and cultural influx, characterized by Chinese influence.

This is apparent in the comprehensive adoption of Chinese-style legitimation, language, and educational systems (*81*).

In our analysis, we observed K2, which is the highest in the present-day Japanese of Northeast, may serve as an additional genetic origin alongside Jomon and EA ancestries. We observed that this component has a significantly higher genetic affinity with Jomon and ancient Korean genomes in the TK era compared to West. The Northeast could be explained by a two-way admixture model using either Korea-TK_2 and Han instead of a triancestral model. It should be noted that Korea-TK_2 can be modeled as either 66% China_WLR_BA and 34% Jomon ancestry or by a triancestral model of 32% NEA, 43% EA, and 25% Jomon (*47*). These data may suggest a potential link between Northeast and NEA, although additional evidence is required to substantiate this connection. Historical records indicate that Northeast was inhabited by the so-called Emishi people, literally translated as "shrimp barbarians" (*82*). The origin of Emishi is somehow understudied and remains a matter of debate, but it was proposed that they might be related to NEA (*83*, *84*). In addition, it has been suggested that the Emishi people might have spoken a distinct Japonic language, akin to the historical Izumo dialect (*85*). Furthermore, despite the geographical distance between Northeast and South—specifically, Northern Kyushu, where evidence suggests that rice farming was first introduced in Japan (*86*)—it has been reported that local groups in the northern part of Northeast exclusively adopted rice during the early Yayoi period (*87*). This connection may be facilitated by human movements along the coastline of the Sea of Japan, potentially suggesting a link between the Northeast and the adoption of rice farming during the Yayoi period. Note that although the two-way fit model, using Korea-TK_2 and Han, demonstrates an acceptable fit, it implies the introduction of Jomon ancestry into the Northeast by continental immigrants, which seemed to be inconsistent with historical context (*76*). The unsuccessful fitting of the triancestral model could result from a higher proportion of Jomon ancestry in Northeast, possibly due to admixture with local populations with greater Jomon ancestry or owing to the limitations of our reliance on the precompiled Allen Ancient DNA Resource (AADR) dataset, which includes only 1240K SNP sites. The additional filtering on transversion sites further reduced the number of SNPs available for analysis. Ideally, this limitation would be addressed by processing directly raw sequencing alignment data; however, this extensive analysis is beyond the scope of the current study. Furthermore, the f_4 analysis did not pinpoint a specific ancestral source among ancient NEA populations for Northeast. This important matter warrants future investigation, optimally involving new and more broadly and densely sampled ancient genomes from NEA. Last, we propose that genetic evidence be examined together with data from other domains, such as archaeology, culture, and linguistics. This interdisciplinary approach can enhance our understanding of the mysterious prehistory of the Japanese population. In addition, it should be acknowledged that both dual structure and tripartite origins models represent simplifications, although the latter may offer several advantages (note S9). The actual population history may be more complex and require further analysis.

In addition to the population structure analysis, we extensively analyzed coding variants in JEWEL. We observed that LoF variants in a set of genes were restricted to limited transcripts than expected by chance; sometimes, the genes are highly constrained, and carriers with those LoF variants displayed shared clinical phenotypes. A previous study has shown that more accurate transcript-level annotation could be achieved by incorporating isoform expression data (*88*). Our results suggest that WGS data offer a potential opportunity to develop a new metric or score of the constraint spectrum by comparing the intolerance of LoF across transcripts within a given gene. We have demonstrated that the extensive clinical information available in JEWEL can be effectively used to uncover potential link between genotype and phenotype.

We reported archaic-introgressed variants are associated with a broad range of phenotypes, including immune and metabolic phenotypes in present-day Japanese. It has been shown that introgressed Denisovan sequences at the *EPAS1* locus have helped Tibetans adapt to high-altitude environments (*89*). However, beyond a few specific examples such as *EPAS1*, the impacts of Denisovan introgression on human phenotypes remain less understood, particularly in comparison with introgression from Neanderthal (*90*). In this context, we have shown that Denisovan-derived segments at *NKX6-1* and *POLR3E* are associated with T2D and height, respectively. A previous study had reported the likely Neanderthal-introgressed segments associated with disease phenotypes using publicly available BBJ GWAS sumstats and precalled archaic variants (*65*). Our study replicated all reported findings and reported 43 additional associations, which greatly expanded the number of introgressed linked with phenotypes and enhanced our understanding of the phenotypic impact of archaic sequence in the Japanese population. In particular, the association between Neanderthal-derived variants of *GLP1R* and T2D is intriguing, considering population specificity and the development of oral semaglutide, a glucagon-like peptide-1 (GLP-1) analog, for treating T2D (*91*). Future research could investigate whether individuals with these archaic variants respond differently to semaglutide treatment and explore the presence of additional archaic segments that could be potential targets for drug discovery. In addition to this specific example, we have demonstrated that overall significant introgressed variants exhibit population specificity in EAs compared to Europeans, which suggests that these archaic variant–phenotype associations might be missed when only examining European data.

Our selection analysis complements genome-wide scans for recent selection signatures in the Japanese population, using methods including SDS and ASMC. In a study based on 170,882 individuals from the BBJ, 29 candidate loci were suggested to be under selection in the past 150 generations based on DRC_{150} statistics using ASMC. In addition, two loci, including the ADH cluster and MHC, were identified by the iHS method (*92*). However, the selection profile within a more recent time frame using DRC-based statistics has yet to be explored. Our analysis indicated that *MHC*, *ADH*, and *ALDH2* are under recent positive selection according to iHS, FastSMC, and previously reported SDS analysis. There are potential differences in ADH/*ALDH2* signals between Okinawa and Hondo groups, and this may warrant further analysis (note S10). We also observed a candidate locus at 2p25.3. While several genes in this locus warrant consideration as candidate genes, we recommend conducting further replication analyses before focusing on any specific gene.

In summary, our study has unveiled genetic characteristics of the Japanese population that were not previously discernible with microarray data. The extensive dataset created in this study also serves as a reference for future genetic research within and beyond the Japanese population. The study emphasized potential applications of WGS in personalized medicine and other clinical settings and highlighted the importance of extending WGS to diverse populations to decode genetic characteristics and better understand human history in a population-specific manner.

MATERIALS AND METHODS

WGS and variant calling

Briefly, sequencing was done at two different depths: (i) 1502 individuals were sequenced at a ~30× (mean, 32.3; median, 31.8) with Illumina HiSeq 2500 (rapid mode or V4) or Illumina HiSeq X Five platform; (ii) 1786 individuals were sequenced at a ~20× depth (mean, 19.9; median, 19.5) using the Illumina HiSeq X Five platform. Sequencing libraries were prepared using standard Illumina protocols and paired-end sequencing was conducted (2 × 125, 2 × 150, or 2 × 160 bp). After sequencing, we performed sample quality control (QC) to remove low-quality sequenced and closely related individuals. In total, 32 of 3288 individuals were excluded, leaving 3256 samples (note S2). After alignment of reads to a human reference (hg19) using BWA-MEM (v0.7.5 or v0.7.13) and removal of duplicated reads, we conducted the joint genotyping calling following the best practice proposed by GATK (v3.2-2). We performed further SNP QC with the following exclusion criteria: (i) read depth (DP) < 5; (ii) genotype quality (GQ) < 20; (iii) DP > 60 or GQ > 95; (iv) variants failed the variant quality score recalibration filtering. Detailed procedures of WGS have also been described previously (*73*). Among 3256 individuals, array-based genotyping data of 3157 individuals were available. These individuals were genotyped using Illumina Human OmniExpress Exome BeadChip or a combination of Illumina HumanOmniExpress and HumanExome BeadChips. We compared the genotyping concordance rate for QC-passed SNPs, which has a call rate ≥ 99% and a Hardy-Weinberg equilibrium P value (P_{HWE}) ≥ 1×10^{-6}. Because sequencing depths are different for two subcohorts at 20× and 30×, we examined metrics related to genotyping quality, including Ti/Tv, concordance, heterozygosity rate, and number of singleton coding variants per individual. We observed comparable values among the two cohorts (table S3). After the removal of singletons, phasing was conducted by Eagle (v2.4.1) for all biallelic variants on each chromosome using the default parameters (*93*).

Population structure and population genetic analysis

PCA was conducted by PLINK (v1.9) based on pruned common or rare variants. We defined common variants as those with a minor AF (MAF) ≥ 0.01 and rare variants as those with an MAF between 0.001 and 0.01. We performed pruning for both categories of variants to select tag SNPs by PLINK with the following parameter: --indep 500, 50, 0.2. Variants in MHC region (chr6: 25 to 34 Mb, hg19) were excluded from the analysis. A total of 184,036 common and 1,835,116 rare variants were obtained after pruning and were used for PCA. The UMAP analysis of the top 20 PCs from rare variant–based PCA was conducted with the UMAP package (v1.1) in R (version 3.1). ADMIXTURE (v1.3.0) was used for the admixture analysis based on the 184,036 pruned common variants (*94*). To determine the optimal K value, we used the Structure Selector software (*38*). To avoid unbalanced sample selection, we randomly selected 50 samples from each region (excluding Okinawa, for which we included all 28 samples) and conducted admixture analysis from $K = 2$ to 6 with three repetitions at each run. In addition, we used the badMIXTURE to visualize the model fitting according to the recommended analysis procedures (*39*).

Using ADMIXTOOLS (v7.0.2) and admixr package, we computed f_4 and f_3 statisitcs (*95*). We calculated f_4 ratio with the form f_4 (a: Chinese Dai in Xishuangbanna; b: CHB; x: target population; c: Jomon; o: Yoruba in Ibadan, Nigeria). In this formula, "a" represents a population related to "b" but not involved in the admixture. On the other hand, b and "c" are the source populations contributing to the admixture, and x is the target admixed population. Last, "o" serves as the outgroup population. The a and 1-a reflect the proportion of admixture from CHB and Jomon. To ensure equal sample sizes, we selected ~30 individuals from each region considering PCA-UMAP information and merged with the AADR dataset (V54.1.p1), matching the "1240K" variants in the AADR panel. Only transversion sites were used for the analysis. We used Jomon individuals labeled "Japan_HG_Jomon" from a previous study for the f_4 ratio test (*41*). Furthermore, we computed outgroup f_3 statistics with the form f_3 (a: target population; b: Chinese Han; o: Papuan). The statistics reflects shared genetic drift between two source population a and b, and large values indicates greater shared genetic drift and thus. We set subregion groups in JEWEL as a, Han as b, and Yoruba as o. We also calculated f_4 statistics using the formula f_4 (Mbuti, ancient genome; Northeast, West) and focused on results supported by more than 50,000 SNPs. We included ancient genomes from China, Korea, and Japan from previous studies (*41*, *42*, *45*, *47*, *96*). We defined NEA by grouping China_WLR_BA_o and China_HMMH_MN, as used in the previous study (*41*). In addition, following the previous report, we defined Korea-TK_2 group as AKG_10203 and AKG_10207. This group has been shown to be more closely related to present-day Japanese and other ancient Japanese groups with high Jomon ancestry (*47*). We conducted a qpAdm analysis (qpAdm version 1520) to model the three-way or two-way admixture following the configuration outlined in a previous study (*41*). We used a set of nine Eurasian populations as right group, comprising Sardinian ($n = 3$), Kusunda ($n = 2$), Papuan ($n = 14$), Dai ($n = 4$), Ami ($n = 2$), Naxi ($n = 3$), Tianyuan ($n = 1$), Chokhopani ($n = 1$), and Mal'ta ($n = 1$), with the option set to "allsnps: YES."

Identification of LoF variants

We performed variant annotation for all biallelic variants using the software VEP (v87) and the LoF transcript effect estimator package (*36*). For missense variants, we incorporated annotations or in silico predictions from 30 different tools, and the risk score was the sum of the number of tools supporting the variant to be deleterious (*97*). We defined LoF variants as those that cause premature stop codons (stop-gained), small-sized indels that shift the coding sequence (frameshift), or variants changing two immediately adjunct nucleotides to the splicing sites (splicing variants). Using LoF transcript effect estimator, we filter high-confidence LoF variants by filtering out those likely annotation artifacts (e.g., the LoF variants in the 3′ end of the transcripts). For LoF variants with a minor allele count ≥ 3, we examined whether the occurrence of the LoF variant is associated with UMAP1/2 by logistic regression analysis using R (v3.1).

Human knockouts

We screened individuals carrying any rare homozygous LoF variant(s) (MAF < 0.01) or rare compound heterozygous LoF variants. We restricted this analysis to LoF variants in genes that contain pathogenic variants in the ClinVar database (v20201208) (www.ncbi.nlm.nih.gov/clinvar/). To identify potential compound heterozygotes, we filtered for instances where multiple LoF variants were present within the same gene of the same individual, and we examined the phased haplotypes. For all candidate human knockouts, we performed manual curation and visually examined the raw alignment reads by the Integrative Genomics Viewer. To identify genes in which LoF variants occurred in fewer transcripts than expected by chance, we selected

4192 genes that had more than one LoF variant and performed a simplified permutation-based test. For a gene with N LoF variants, we summed the actual number of transcripts affected by these N LoFs, denoted as J. Next, we randomly selected N positions within the gene's coding region based on GENCODE gene annotation (v19) (www.gencodegenes.org/human/), and we counted the total number of transcripts overlapped with N positions, denoted as K. We repeated this procedure 1000 times to obtain a list of values from K_1 to K_{1000}. The empirical permutation P value was calculated as the rank of J among the 1000 K values, sorted in ascending order.

Detection of introgressed sequences and variants

To identify sequences that are likely introgressed from Neanderthals or Denisovans, we applied a recently developed computational method IBDMix (*63*). In contrast to other methods, IBDMix used an archaic reference genome to infer the introgression segments. We conducted following filtering steps for the introgression analysis. Briefly, for Neanderthal and Denisovan genomes, minimal filter masks were applied (obtained from https://bioinf.eva.mpg.de/). For human genome sequences, we applied the 1KGP accessibility mask (downloaded from: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20140520.strict_mask.autosomes.bed). We masked out sequences within 5 bp of the indels and only autosomes were analyzed. The called introgressed sequences with a logarithm of the odds ratio for linkage score ≥ 4 and a length ≥ 50 kb were retained for downstream analyses. To identify introgressed variants and exclude those misclassified due to incomplete lineage sorting, we focused on high-confidence introgressed segments. Briefly, we obtained phased haplotypes for each introgressed segment and calculated match rates to the Neanderthal and Denisovan genomes, excluding variant sites with unknown archaic status. The high-confidence Denisovan segments were defined with a Denisovan match rate ≥ 0.5 and a Neanderthal match rate < 0.5. The high-confidence Neanderthal segments were defined as a Neanderthal match rate ≥ 0.7. The variants observed in over half of introgressed haplotypes were selected as likely introgressed variants. We screened introgressed genetic variants to determine their association with both disease and quantitative traits. We used summary statistics from previous studies encompassing 42 diseases and 64 quantitative traits, all based on the BBJ dataset (*98*). We filtered associations that surpassed the genome-wide significance level at 5×10^{-8}. To exclude association due to LD with nonarchaic variants, for all loci in which the archaic variant was not the lead variant, we calculated the r^2 between lead archaic variant and lead GWAS variant and removed those pairs with $r^2 < 0.9$. We conducted a comparison of introgression segments with those previously reported within the Japanese population based on 1KGP JPT data (*63*). In addition, we used an alternative method, SPrime, to validate introgressed segments that exhibited significant phenotypic associations. For this analysis, we set Yoruba in Ibadan, Nigeria as the outgroup and followed default parameters. For segments likely introgressed from Denisovans, we conducted an enrichment analysis to ascertain whether these segments significantly overlap with those reported in a previous study (*62*). Populations with fewer than 30 segments detected were excluded, as this could result from indirect inheritance. The 1KGP data for this analysis were obtained from the following link: https://data.mendeley.com/datasets/y7hyt83vxr/1. The Philippine Ayta data were obtain from a previous study (*68*). We used Bedtools "fisher" utility to perform the enrichment analysis (https://bedtools.readthedocs.io). We conducted a pathway analysis to identify the biological pathways that exhibited enrichment in genes containing archaic variants associated with diseases. This analysis was carried out using Enrichr. (https://maayanlab.cloud/Enrichr/). To confirm the association between archaic variants in *NKX6-1* locus and T2D, we obtained the GWAS summary statistics from the FinnGen database (https://r9.finngen.fi/) (*14*).

Analysis related to natural selection

The selscan software (v1.3.0) was used to calculate the iHS scores using the default parameters (*99*). We restricted the analysis of autosomal biallelic variants with an MAF ≥ 0.01. Next, on the basis of the human-chimp-macaque alignment provided by Ensembl (downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/), we retained variants of whose ancestral allele was present in chimpanzee or macaque. We used a Japanese-specific recombination map created using the 1KGP JPT dataset to calculate iHS (*100*). We normalized unstandardized iHS scores across 100 AF bins. The approximate P_{iHS} values were calculated by fitting normalized iHS scores, assuming a normal distribution. The genome-wide significance threshold was determined at 8.24×10^{-9} based on Bonferroni correction (0.05/6,066,864 variants). To exclude potentially false positive signals, we removed loci showing extremely high or low recombination rates, loci containing only a single significant variant, and loci with segmental duplication in the nearby region. We conducted a subanalysis on samples from five representative regions: West, East, Northeast, South, and Okinawa, using the same procedure. For Okinawa, we limited the analysis to variants with an AF of 5% or higher, owing to the limited sample size. In addition to iHS, we used FastSMC to identify genomic loci likely targeted by selection in the Japanese population. Since the method was developed and tuned using microarray data, we extracted a superset of variants included in the Illumina HumanOmniExpressExome BeadChip array, the Illumina Infinium Asian Screening Array, and the Affymetrix Japonica array. By analyzing the locus-specific IBD sharing patterns, the DRC within the past 50 generations (DRC_{50}) was calculated by FastSMC. We summarized the mean DRC_{50} for each sliding window at a size of 0.05 centimorgan. The decoding file was prepared from the 1KGP JPT demographic and AF file. We then fitted a Gamma distribution to the averaged DRC_{50} values using the neutral regions in the genome. We excluded genetic loci reported to be under positive selection in the Japanese population based on genome-wide analysis (*73*, *92*). We further iteratively removed regions that showed evidence of being targeted by selection based on the DRC statistic. On the basis of this null model, we derived approximate one-sided P values. The genome-wide significance threshold was set at 5×10^{-8} for P_{DRC50}. To assess the overlap between loci identified as genome-wide significant by iHS or DRC statistics and known SVs, we analyzed the phase 2 dataset from the Human Genome Structural Variation Consortium (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/PanGenie_results/pangenie_merged_multi_nosnvs.vcf.gz).

Supplementary Materials

This PDF file includes:

Notes S1 to S10
Figs. S1 to S18
Legends for tables S1 to S24
References

Other Supplementary Material for this manuscript includes the following:
Tables S1 to S24

REFERENCES AND NOTES

1. O. Bocher, C. J. Willer, E. Zeggini, Unravelling the genetic architecture of human complex traits through whole genome sequencing. *Nat. Commun.* **14**, 3520 (2023).
2. H. Jónsson, P. Sulem, B. Kehr, S. Kristmundsdottir, F. Zink, E. Hjartarson, M. T. Hardarson, K. E. Hjorleifsson, H. P. Eggertsson, S. A. Gudjonsson, L. D. Ward, G. A. Arnadottir, E. A. Helgason, H. Helgason, A. Gylfason, A. Jonasdottir, A. Jonasdottir, T. Rafnar, M. Frigge, S. N. Stacey, O. T. Magnusson, U. Thorsteinsdottir, G. Masson, A. Kong, B. V. Halldorsson, A. Helgason, D. F. Gudbjartsson, K. Stefansson, Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
3. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. Van Driem, P. De Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
4. J. Choin, J. Mendoza-Revilla, L. R. Arauna, S. Cuadros-Espinoza, O. Cassar, M. Larena, A. M.-S. Ko, C. Harmant, R. Laurent, P. Verdu, G. Laval, A. Boland, R. Olaso, J.-F. Deleuze, F. Valentin, Y.-C. Ko, M. Jakobsson, A. Gessain, L. Excoffier, M. Stoneking, E. Patin, L. Quintana-Murci, Genomic insights into population history and biological adaptation in Oceania. *Nature* **592**, 583–589 (2021).
5. Y. Field, E. A. Boyle, N. Telis, Z. Gao, K. J. Gaulton, D. Golan, L. Yengo, G. Rocheleau, P. Froguel, M. I. McCarthy, J. K. Pritchard, Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
6. K. E. Johnson, B. F. Voight, Patterns of shared signatures of recent positive selection across human populations. *Nat. Ecol. Evol.* **2**, 713–720 (2018).
7. S. M. Carthy, S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood, A. Teumer, H. M. Kang, C. Fuchsberger, P. Danecek, K. Sharp, Y. Luo, C. Sidore, A. Kwong, N. Timpson, S. Koskinen, S. Vrieze, L. J. Scott, H. Zhang, A. Mahajan, J. Veldink, U. Peters, C. Pato, C. M. van Duijn, C. E. Gillies, I. Gandin, M. Mezzavilla, A. Gilly, M. Cocca, M. Traglia, A. Angius, J. Barrett, D. I. Boomsma, K. Branham, G. Breen, C. Brummet, F. Busonero, H. Campbell, A. Chan, S. Chen, E. Chew, F. S. Collins, L. Corbin, G. D. Smith, G. Dedoussis, M. Dorr, A.-E. Farmaki, L. Ferrucci, L. Forer, R. M. Fraser, S. Gabriel, S. Levy, L. Groop, T. Harrison, A. Hattersley, O. L. Holmen, K. Hveem, M. Kretzler, J. Lee, M. M. Gue, T. Meitinger, D. Melzer, J. Min, K. L. Mohlke, J. Vincent, M. Nauck, D. Nickerson, A. Palotie, M. Pato, N. Pirastu, M. M. Innis, B. Richards, C. Sala, V. Salomaa, D. Schlessinger, S. Schoenheer, P. E. Slagboom, K. Small, T. Spector, D. Stambolian, M. Tuke, J. Tuomilehto, L. Van den Berg, W. Van Rheenen, U. Volker, C. Wijmenga, D. Toniolo, E. Zeggini, P. Gasparini, M. G. Sampson, J. F. Wilson, T. Frayling, P. de Bakker, M. A. Swertz, S. M. Carroll, C. Kooperberg, A. Dekker, D. Altshuler, C. Willer, W. Iacono, S. Ripatti, N. Soranzo, K. Walter, A. Swaroop, F. Cucca, C. Anderson, M. Boehnke, M. I. Mc Carthy, R. Durbin, G. Abecasis, J. Marchini, for the Haplotype Reference Consortium, A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
8. S.-K. Yoo, C.-U. Kim, H. L. Kim, S. Kim, J.-Y. Shin, N. Kim, J. S. W. Yang, K.-W. Lo, B. Cho, F. Matsuda, S. C. Schuster, C. Kim, J.-I. Kim, J.-S. Seo, NARD: Whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Med.* **11**, 64 (2019).
9. S. Das, G. R. Abecasis, B. L. Browning, Genotype imputation from large reference panels. *Annu. Rev. Genomics Hum. Genet.* **19**, 73–96 (2018).
10. The UK10K Consortium, The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
11. F. S. Alkuraya, Natural human knockouts and the era of genotype to phenotype. *Genome Med.* **7**, 48 (2015).
12. E. V. Minikel, K. J. Karczewski, H. C. Martin, B. B. Cummings, N. Whiffin, D. Rhodes, J. Alföldi, R. C. Trembath, D. A. van Heel, M. J. Daly, Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, S. L. Schreiber, D. G. MacArthur, Evaluating drug targets through human loss-of-function genetic variation. *Nature* **581**, 459–464 (2020).
13. B. V. Halldorsson, H. P. Eggertsson, K. H. S. Moore, H. Hauswedell, O. Eiriksson, M. O. Ulfarsson, G. Palsson, M. T. Hardarson, A. Oddsson, B. O. Jensson, S. Kristmundsdottir, B. D. Sigurpalsdottir, O. A. Stefansson, D. Beyter, G. Holley, V. Tragante, A. Gylfason, P. I. Olason, F. Zink, M. Asgeirsdottir, S. T. Sverrisson, B. Sigurdsson, S. A. Gudjonsson, G. T. Sigurdsson, G. H. Halldorsson, G. Sveinbjornsson, K. Norland, U. Styrkarsdottir, D. N. Magnusdottir, S. Snorradottir, K. Kristinsson, E. Sobech, H. Jonsson, A. J. Geirsson, I. Olafsson, P. Jonsson, O. B. Pedersen, C. Erikstrup, S. Brunak, S. R. Ostrowski, D. B. D. S. G. Consortium, G. Thorleifsson, F. Jonsson, P. Melsted, I. Jonsdottir, T. Rafnar, H. Holm, H. Stefansson, J. Saemundsdottir, D. F. Gudbjartsson, O. T. Magnusson, G. Masson, U. Thorsteinsdottir, A. Helgason, H. Jonsson, P. Sulem, K. Stefansson, The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
14. M. I. Kurki, J. Karjalainen, P. Palta, T. P. Sipilä, K. Kristiansson, K. M. Donner, M. P. Reeve, H. Laivuori, M. Aavikko, M. A. Kaunisto, A. Loukola, E. Lahtela, H. Mattsson, P. Laiho, P. D. B. Parolo, A. A. Lehisto, M. Kanai, N. Mars, J. Rämö, T. Kiiskinen, H. O. Heyne, K. Veerapen, S. Rüeger, S. Lemmelä, W. Zhou, S. Ruotsalainen, K. Pärn, T. Hiekkalinna, S. Koskelainen, T. Paajanen, V. Llorens, J. Gracia-Tabuenca, H. Siirtola, K. Reis, A. G. Elnahas, B. Sun, C. N. Foley, K. Aalto-Setälä, K. Alasoo, M. Arvas, K. Auro, S. Biswas, A. Bizaki-Vallaskangas, O. Carpen, C.-Y. Chen, O. A. Dada, Z. Ding, M. G. Ehm, K. Eklund, M. Färkkilä, H. Finucane, A. Ganna, A. Ghazal, R. R. Graham, E. M. Green, A. Hakanen, M. Hautalahti, Å. K. Hedman, M. Hiltunen, R. Hinttala, I. Hovatta, X. Hu, A. Huertas-Vazquez, L. Huilaja, J. Hunkapiller, H. Jacob, J.-N. Jensen, H. Joensuu, S. John, V. Julkunen, M. Jung, J. Junttila, K. Kaarniranta, M. Kähönen, R. Kajanne, L. Kallio, R. Kälviäinen, J. Kaprio, F. Gen, N. Kerimov, J. Kettunen, E. Kilpeläinen, T. Kilpi, K. Klinger, V.-M. Kosma, T. Kuopio, V. Kurra, T. Laisk, J. Laukkanen, N. Lawless, A. Liu, S. Longerich, R. Mägi, J. Mäkelä, A. Mäkitie, A. Malarstig, A. Mannermaa, J. Maranville, A. Matakidou, T. Meretoja, S. V. Mozaffari, M. E. K. Niemi, M. Niemi, T. Niiranen, C. J. O. Donnell, M. E. Obeidat, G. Okafo, H. M. Ollila, A. Palomäki, T. Palotie, J. Partanen, D. S. Paul, M. Pelkonen, R. K. Pendergrass, S. Petrovski, A. Pitkäranta, A. Platt, D. Pulford, E. Punkka, P. Pussinen, N. Raghavan, F. Rahimov, D. Rajpal, N. A. Renaud, B. Riley-Gillis, R. Rodosthenous, E. Saarentaus, A. Salminen, E. Salminen, V. Salomaa, J. Schleutker, R. Serpi, H.-Y. Shen, R. Siegel, K. Silander, S. Siltanen, S. Soini, H. Soininen, J. H. Sul, I. Tachmazidou, K. Tasanen, P. Tienari, S. Toppila-Salmi, T. Tukiainen, T. Tuomi, J. A. Turunen, J. C. Ulirsch, F. Vaura, P. Virolainen, J. Waring, D. Waterworth, R. Yang, M. Nelis, A. Reigo, A. Metspalu, L. Milani, T. Esko, C. Fox, A. S. Havulinna, M. Perola, S. Ripatti, A. Jalanko, T. Laitinen, T. P. Mäkelä, R. Plenge, M. M. Carthy, H. Runz, M. J. Daly, A. Palotie, FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
15. D. F. Gudbjartsson, H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson, A. Gylfason, S. Besenbacher, G. Magnusson, B. V. Halldorsson, E. Hjartarson, G. T. Sigurdsson, S. N. Stacey, M. L. Frigge, H. Holm, J. Saemundsdottir, H. T. Helgadottir, H. Johannsdottir, G. Sigfusson, G. Thorgeirsson, J. T. Sverrisson, S. Gretarsdottir, G. B. Walters, T. Rafnar, B. Thjodleifsson, E. S. Bjornsson, S. Olafsson, H. Thorarinsdottir, T. Steingrimsdottir, T. S. Gudmundsdottir, A. Theodors, J. G. Jonasson, A. Sigurdsson, G. Bjornsdottir, J. J. Jonsson, O. Thorarensen, P. Ludvigsson, H. Gudbjartsson, G. I. Eyjolfsson, O. Sigurdardottir, I. Olafsson, D. O. Arnar, O. T. Magnusson, A. Kong, G. Masson, U. Thorsteinsdottir, A. Helgason, P. Sulem, K. Stefansson, Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
16. A. B. Popejoy, S. M. Fullerton, Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
17. A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, M. J. Daly, Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
18. The All of Us Research Program Investigators, The "All of Us" research program. *N. Engl. J. Med.* **381**, 668–676 (2019).
19. D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. L. Faive, S.-B. Lee, X. Tian, B. L. Browning, S. Das, A.-K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, T. W. Blackwell, A. V. Smith, Q. Wong, X. Liu, M. P. Conomos, D. M. Bobo, F. Aguet, C. Albert, A. Alonso, K. G. Ardlie, D. E. Arking, S. Aslibekyan, P. L. Auer, J. Barnard, R. G. Barr, L. Barwick, L. C. Becker, R. L. Beer, E. J. Benjamin, L. F. Bielak, J. Blangero, M. Boehnke, D. W. Bowden, J. A. Brody, E. G. Burchard, B. E. Cade, J. F. Casella, B. Chalazan, D. I. Chasman, Y.-D. I. Chen, M. H. Cho, S. H. Choi, M. K. Chung, C. B. Clish, A. Correa, J. E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D. L. De Meo, S. K. Dutcher, P. T. Ellinor, L. S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S. M. Fullerton, S. Germer, M. T. Gladwin, D. J. Gottlieb, X. Guo, M. E. Hall, J. He, N. L. Heard-Costa, S. R. Heckbert, M. R. Irvin, J. M. Johnsen, A. D. Johnson, R. Kaplan, S. L. R. Kardia, T. Kelly, S. Kelly, E. E. Kenny, D. P. Kiel, R. Klemmer, B. A. Konkle, C. Kooperberg, A. Köttgen, L. A. Lange, J. Lasky-Su, D. Levy, X. Lin, K.-H. Lin, C. Liu, R. J. F. Loos, L. Garman, R. Gerszten, S. A. Lubitz, K. L. Lunetta, A. C. Y. Mak, A. Manichaikul, A. K. Manning, R. A. Mathias, D. D. M. Manus, S. T. M. Garvey, J. B. Meigs, D. A. Meyers, J. L. Mikulla, M. A. Minear, B. D. Mitchell, S. Mohanty, M. E. Montasser, C. Montgomery, A. C. Morrison, J. M. Murabito, A. Natale, P. Natarajan, S. C. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, N. Pankratz, G. M. Peloso, P. A. Peyser, J. Pleiness, W. S. Post, B. M. Psaty, D. C. Rao, S. Redline, A. P. Reiner, D. Roden, J. I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D. A. Schwartz,

J.-S. Seo, S. Seshadri, V. A. Sheehan, W. H. Sheu, M. B. Shoemaker, N. L. Smith, J. A. Smith, N. Sotoodehnia, A. M. Stilp, W. Tang, K. D. Taylor, M. Telen, T. A. Thornton, R. P. Tracy, D. J. Van Den Berg, R. S. Vasan, K. A. Viaud-Martinez, S. Vrieze, D. E. Weeks, B. S. Weir, S. T. Weiss, L.-C. Weng, C. J. Willer, Y. Zhang, X. Zhao, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K. M. Rice, S. S. Rich, E. K. Silverman, P. Qasba, W. Gan, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, G. J. Papanicolaou, D. A. Nickerson, S. R. Browning, M. C. Zody, S. Zöllner, J. G. Wilson, L. A. Cupples, C. C. Laurie, C. E. Jaquish, R. D. Hernandez, T. D. O'Connor, G. R. Abecasis, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

20. GenomeAsia100K Consortium, The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
21. D. Wu, J. Dou, X. Chai, C. Bellis, A. Wilm, C. C. Shih, W. W. J. Soon, N. Bertin, C. B. Lin, C. C. Khor, M. DeGiorgio, S. Cheng, L. Bao, N. Karnani, W. Y. K. Hwang, S. Davila, P. Tan, A. Shabbir, A. Moh, E.-K. Tan, J. N. Foo, L. L. Goh, K. P. Leong, R. S. Y. Foo, C. S. P. Lam, A. M. Richards, C.-Y. Cheng, T. Aung, T. Y. Wong, H. H. Ng, J. Liu, C. Wang, M. A. Ackers-Johnson, E. Aliwarga, K. H. K. Ban, D. Bertrand, J. C. Chambers, D. L. H. Chan, C. X. L. Chan, M. L. Chee, M. L. Chee, P. Chen, Y. Chen, E. G. Y. Chew, W. J. Chew, L. H. Y. Chiam, J. P. C. Chong, I. Chua, S. A. Cook, W. Dai, R. Dorajoo, C.-S. Foo, R. S. M. Goh, A. M. Hillmer, I. D. Irwan, F. Jaufeerally, A. Javed, J. Jeyakani, J. T. H. Koh, J. Y. Koh, P. Krishnaswamy, J. L. Kuan, N. Kumari, A. S. Lee, S. E. Lee, S. Lee, Y. L. Lee, S. T. Leong, Z. Li, P. Y. Li, J. X. Liew, O. W. Liew, S. C. Lim, W. K. Lim, C. W. Lim, T. B. Lim, C. K. Lim, S. Y. Loh, A. W. Lok, C. W. L. Chin, S. Majithia, S. Maurer-Stroh, W. Y. Meah, S. Q. Mok, N. Nargarajan, P. Ng, S. B. Ng, Z. Ng, J. Y. X. Ng, E. Ng, S. L. Ng, S. Nusinovici, C. T. Ong, B. Pan, V. Pedergnana, S. Poh, S. Prabhakar, K. M. Prakash, I. Quek, C. Sabanayagam, W. Q. See, Y. Y. Sia, X. Sim, W. C. Sim, J. So, D. K. N. Soon, E. S. Tai, N. Y. Tan, L. C. S. Tan, H. C. Tan, W. L. W. Tan, M. Tandiono, A. Tay, S. Thakur, Y. C. Tham, Z. Tiang, G. L.-X. Toh, P. K. Tsai, L. Veeravalli, C. S. Verma, L. Wang, M. R. Wang, W.-C. Wong, Z. Xie, K. K. Yeo, L. Zhang, W. Zhai, Y. Zhao, Large-Scale whole-genome sequencing of three diverse asian populations in singapore. *Cell* **179**, 736–749.e15 (2019).
22. Y. Cao, L. Li, M. Xu, Z. Feng, X. Sun, J. Lu, Y. Xu, P. Du, T. Wang, R. Hu, Z. Ye, L. Shi, X. Tang, L. Yan, Z. Gao, G. Chen, Y. Zhang, L. Chen, G. Ning, Y. Bi, W. Wang, The China MAP Consortium, The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.* **30**, 717–731 (2020).
23. P.-K. Cong, W.-Y. Bai, J.-C. Li, M.-Y. Yang, S. Khederzadeh, S.-R. Gai, N. Li, Y.-H. Liu, S.-H. Yu, W.-W. Zhao, J.-Q. Liu, Y. Sun, X.-W. Zhu, P.-P. Zhao, J.-W. Xia, P.-L. Guan, Y. Qian, J.-G. Tao, L. Xu, G. Tian, P.-Y. Wang, S.-Y. Xie, M.-C. Qiu, K.-Q. Liu, B.-S. Tang, H.-F. Zheng, Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat. Commun.* **13**, 2939 (2022).
24. S. Kuriyama, N. Yaegashi, F. Nagami, T. Arai, Y. Kawaguchi, N. Osumi, M. Sakaida, Y. Suzuki, K. Nakayama, H. Hashizume, G. Tamiya, H. Kawame, K. Suzuki, A. Hozawa, N. Nakaya, M. Kikuya, H. Metoki, I. Tsuji, N. Fuse, H. Kiyomoto, J. Sugawara, A. Tsuboi, S. Egawa, K. Ito, K. Chida, T. Ishii, H. Tomita, Y. Taki, N. Minegishi, N. Ishii, J. Yasuda, K. Igarashi, R. Shimizu, M. Nagasaki, S. Koshiba, K. Kinoshita, S. Ogishima, T. Takai-Igarashi, T. Tominaga, O. Tanabe, N. Ohuchi, T. Shimosegawa, S. Kure, H. Tanaka, S. Ito, J. Hitomi, K. Tanno, M. Nakamura, K. Ogasawara, S. Kobayashi, K. Sakata, M. Satoh, A. Shimizu, M. Sasaki, R. Endo, K. Sobue, The Tohoku Medical Megabank Project Study Group, M. Yamamoto, The Tohoku Medical Megabank Project: Design and mission. *J. Epidemiol.* **26**, 493–511 (2016).
25. M. Nagasaki, J. Yasuda, F. Katsuoka, N. Nariai, K. Kojima, Y. Kawai, Y. Yamaguchi-Kabata, J. Yokozawa, I. Danjoh, S. Saito, Y. Sato, T. Mimori, K. Tsuda, R. Saito, X. Pan, S. Nishikawa, S. Ito, Y. Kuroki, O. Tanabe, N. Fuse, S. Kuriyama, H. Kiyomoto, A. Hozawa, N. Minegishi, J. D. Engel, K. Kinoshita, S. Kure, N. Yaegashi, To MMo Japanese Reference Panel Project, M. Yamamoto, Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 8018 (2015).
26. S. Tadaka, F. Katsuoka, M. Ueki, K. Kojima, S. Makino, S. Saito, A. Otsuki, C. Gocho, M. Sakurai-Yageta, I. Danjoh, I. N. Motoike, Y. Yamaguchi-Kabata, M. Shirota, S. Koshiba, M. Nagasaki, N. Minegishi, A. Hozawa, S. Kuriyama, A. Shimizu, J. Yasuda, N. Fuse, Tohoku Medical Megabank Project Study Group, G. Tamiya, M. Yamamoto, K. Kinoshita, 3.5KJPNv2: An allele frequency panel of 3552 Japanese individuals including the X chromosome. *Hum. Genome Var.* **6**, 28 (2019).
27. S. Tadaka, E. Hishinuma, S. Komaki, I. N. Motoike, J. Kawashima, D. Saigusa, J. Inoue, J. Takayama, Y. Okamura, Y. Aoki, M. Shirota, A. Otsuki, F. Katsuoka, A. Shimizu, G. Tamiya, S. Koshiba, M. Sasaki, M. Yamamoto, K. Kinoshita, jMorp updates in 2020: Large enhancement of multi-omics data resources on the general Japanese population. *Nucleic Acids Res.* **49**, D536–D544 (2021).
28. N. Mitsuhashi, L. Toyo-oka, T. Katayama, M. Kawashima, S. Kawashima, K. Miyazaki, T. Takagi, TogoVar: A comprehensive Japanese genetic variation database. *Hum. Genome Var.* **9**, 44 (2022).
29. Y. Kawai, Y. Watanabe, Y. Omae, R. Miyahara, S.-S. Khor, E. Noiri, K. Kitajima, H. Shimanuki, H. Gatanaga, K. Hata, K. Hattori, A. Iida, H. Ishibashi-Ueda, T. Kaname, T. Kanto, R. Matsumura, K. Miyo, M. Noguchi, K. Ozaki, M. Sugiyama, A. Takahashi, H. Tokuda, T. Tomita, A. Umezawa, H. Watanabe, S. Yoshida, Y. Goto, Y. Maruoka, Y. Matsubara, S. Niida, M. Mizokami, K. Tokunaga, Exploring the genetic diversity of the Japanese population: Insights from a large-scale whole genome sequencing analysis. *PLOS Genet.* **19**, e1010625 (2023).
30. S. Koyama, K. Ito, C. Terao, M. Akiyama, M. Horikoshi, Y. Momozawa, H. Matsunaga, H. Ieki, K. Ozaki, Y. Onouchi, A. Takahashi, S. Nomura, H. Morita, H. Akazawa, C. Kim, J. Seo, K. Higasa, M. Iwasaki, T. Yamaji, N. Sawada, S. Tsugane, T. Koyama, H. Ikezaki, N. Takashima, K. Tanaka, K. Arisawa, K. Kuriki, M. Naito, K. Wakai, S. Suna, Y. Sakata, H. Sato, M. Hori, Y. Sakata, K. Matsuda, Y. Murakami, H. Aburatani, M. Kubo, F. Matsuda, Y. Kamatani, I. Komuro, Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat. Genet.* **52**, 1169–1177 (2020).
31. C. Terao, A. Suzuki, Y. Momozawa, M. Akiyama, K. Ishigaki, K. Yamamoto, K. Matsuda, Y. Murakami, S. A. McCarroll, M. Kubo, P.-R. Loh, Y. Kamatani, Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature* **584**, 130–135 (2020).
32. A. Nagai, M. Hirata, Y. Kamatani, K. Muto, K. Matsuda, Y. Kiyohara, T. Ninomiya, A. Tamakoshi, Z. Yamagata, T. Mushiroda, Y. Murakami, K. Yuji, Y. Furukawa, H. Zembutsu, T. Tanaka, Y. Ohnishi, Y. Nakamura, M. Kubo, M. Shiono, K. Misumi, R. Kaieda, H. Harada, S. Minami, M. Emi, N. Emoto, H. Daida, K. Miyauchi, A. Murakami, S. Asai, M. Moriyama, Y. Takahashi, T. Fujioka, W. Obara, S. Mori, H. Ito, S. Nagayama, Y. Miki, A. Masumoto, A. Yamada, Y. Nishizawa, K. Kodama, H. Kutsumi, Y. Sugimoto, Y. Koretsune, H. Kusuoka, H. Yanai, Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
33. Y. Yamaguchi-Kabata, K. Nakazono, A. Takahashi, S. Saito, N. Hosono, M. Kubo, Y. Nakamura, N. Kamatani, 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.* **83**, 445–456 (2008).
34. Y. Watanabe, M. Isshiki, J. Ohashi, Prefecture-level population structure of the Japanese based on SNP genotypes of 11,069 individuals. *J. Hum. Genet.* **66**, 431–437 (2021).
35. S. Sakaue, J. Hirata, M. Kanai, K. Suzuki, M. Akiyama, C. Lai Too, T. Arayssi, M. Hammoudeh, S. Al Emadi, B. K. Masri, H. Halabi, H. Badsha, I. W. Uthman, R. Saxena, L. Padyukov, M. Hirata, K. Matsuda, Y. Murakami, Y. Kamatani, Y. Okada, Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. *Nat. Commun.* **11**, 1569 (2020).
36. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
37. 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.* **57**, 787–795 (2012).
38. Y.-L. Li, J.-X. Liu, StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. *Mol. Ecol. Resour.* **18**, 176–177 (2018).
39. D. J. Lawson, L. van Dorp, D. Falush, A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.* **9**, 3258 (2018).
40. T. Jinam, Y. Kawai, Y. Kamatani, S. Sonoda, K. Makisumi, H. Sameshima, K. Tokunaga, N. Saitou, Genome-wide SNP data of Izumo and Makurazaki populations support inner-dual structure model for origin of Yamato people. *J. Hum. Genet.* **66**, 681–687 (2021).
41. N. P. Cooke, V. Mattiangeli, L. M. Cassidy, K. Okazaki, C. A. Stokes, S. Onbe, S. Hatakeyama, K. Machida, K. Kasai, N. Tomioka, A. Matsumoto, M. Ito, Y. Kojima, D. G. Bradley, T. Gakuhari, S. Nakagome, Ancient genomics reveals tripartite origins of Japanese populations. *Sci. Adv.* **7**, eabh2419 (2021).
42. M. Robbeets, R. Bouckaert, M. Conte, A. Savelyev, T. Li, D.-I. An, K. Shinoda, Y. Cui, T. Kawashima, G. Kim, J. Uchiyama, J. Dolińska, S. Oskolskaya, K.-Y. Yamano, N. Seguchi, H. Tomita, H. Takamiya, H. Kanzawa-Kiriyama, H. Oota, H. Ishida, R. Kimura, T. Sato, J.-H. Kim, B. Deng, R. Bjørn, S. Rhee, K.-D. Ahn, I. Gruntov, O. Mazo, J. R. Bentley, R. Fernandes, P. Roberts, I. R. Bausch, L. Gilaizeau, M. Yoneda, M. Kugai, R. A. Bianco, F. Zhang, M. Himmel, M. J. Hudson, C. Ning, Triangulation supports agricultural spread of the Transeurasian languages. *Nature* **599**, 616–621 (2021).
43. H. Kanzawa-Kiriyama, T. A. Jinam, Y. Kawai, T. Sato, K. Hosomichi, A. Tajima, N. Adachi, H. Matsumura, K. Kryukov, N. Saitou, K.-I. Shinoda, Late Jomon male and female genome

sequences from the Funadomari site in Hokkaido, Japan. *Anthropol. Sci.* **127**, 83–108 (2019).
44. T. Gakuhari, S. Nakagome, S. Rasmussen, M. E. Allentoft, T. Sato, T. Korneliussen, B. N. Chuinneagáin, H. Matsumae, K. Koganebuchi, R. Schmidt, S. Mizushima, O. Kondo, N. Shigehara, M. Yoneda, R. Kimura, H. Ishida, T. Masuyama, Y. Yamada, A. Tajima, H. Shibata, A. Toyoda, T. Tsurumoto, T. Wakebe, H. Shitara, T. Hanihara, E. Willerslev, M. Sikora, H. Oota, Ancient Jomon genome sequence analysis sheds light on migration patterns of early East Asian populations. *Commun. Biol.* **3**, 437 (2020).
45. C. Ning, T. Li, K. Wang, F. Zhang, T. Li, X. Wu, S. Gao, Q. Zhang, H. Zhang, M. J. Hudson, G. Dong, S. Wu, Y. Fang, C. Liu, C. Feng, W. Li, T. Han, R. Li, J. Wei, Y. Zhu, Y. Zhou, C.-C. Wang, S. Fan, Z. Xiong, Z. Sun, M. Ye, L. Sun, X. Wu, F. Liang, Y. Cao, X. Wei, H. Zhu, H. Zhou, J. Krause, M. Robbeets, C. Jeong, Y. Cui, Ancient genomes from northern China suggest links between subsistence changes and human migration. *Nat. Commun.* **11**, 2700 (2020).
46. C.-C. Wang, H.-Y. Yeh, A. N. Popov, H.-Q. Zhang, H. Matsumura, K. Sirak, O. Cheronet, A. Kovalev, N. Rohland, A. M. Kim, S. Mallick, R. Bernardos, D. Tumen, J. Zhao, Y.-C. Liu, J.-Y. Liu, M. Mah, K. Wang, Z. Zhang, N. Adamski, N. Broomandkhoshbacht, K. Callan, F. Candilio, K. S. D. Carlson, B. J. Culleton, L. Eccles, S. Freilich, D. Keating, A. M. Lawson, K. Mandl, M. Michel, J. Oppenheimer, K. T. Özdoğan, K. Stewardson, S. Wen, S. Yan, F. Zalzala, R. Chuang, C.-J. Huang, H. Looh, C.-C. Shiung, Y. G. Nikitin, A. V. Tabarev, A. A. Tishkin, S. Lin, Z.-Y. Sun, X.-M. Wu, T.-L. Yang, X. Hu, L. Chen, H. Du, J. Bayarsaikhan, E. Mijiddorj, D. Erdenebaatar, T.-O. Iderkhangai, E. Myagmar, H. Kanzawa-Kiriyama, M. Nishino, K.-I. Shinoda, O. A. Shubina, J. Guo, W. Cai, Q. Deng, L. Kang, D. Li, D. Li, R. Lin, R. S. Nini, L.-X. Wang, L. Wei, G. Xie, H. Yao, M. Zhang, G. He, X. Yang, R. Hu, M. Robbeets, S. Schiffels, D. J. Kennett, L. Jin, H. Li, J. Krause, R. Pinhasi, D. Reich, Genomic insights into the formation of human populations in East Asia. *Nature* **591**, 413–419 (2021).
47. P. Gelabert, A. Blazyte, Y. Chang, D. M. Fernandes, S. Jeon, J. G. Hong, J. Yoon, Y. Ko, V. Oberreiter, O. Cheronet, K. T. Özdoğan, S. Sawyer, S. Yang, E. M. Greytak, H. Choi, J. Kim, J.-I. Kim, C. Jeong, K. Bae, J. Bhak, R. Pinhasi, Northeastern Asian and Jomon-related genetic structure in the Three Kingdoms period of Gimhae, Korea. *Curr. Biol.* **32**, 3232–3244.e6 (2022).
48. N. P. Cooke, V. Mattiangeli, L. M. Cassidy, K. Okazaki, K. Kasai, D. G. Bradley, T. Gakuhari, S. Nakagome, Genomic insights into a tripartite ancestry in the Southern Ryukyu Islands. *Evol. Hum. Sci.* **5**, e23 (2023).
49. M. Sekine, H. Nagata, S. Tsuji, Y. Hirai, S. Fujimoto, M. Hatae, I. Kobayashi, T. Fujii, I. Nagata, K. Ushijima, K. Obata, M. Suzuki, M. Yoshinaga, N. Umesaki, S. Satoh, T. Enomoto, S. Motoyama, K. Tanaka, Japanese Familial Ovarian Cancer Study Group, Mutational analysis of BRCA1 and BRCA2 and clinicopathologic analysis of ovarian cancer in 82 ovarian cancer families: Two common founder mutations of BRCA1 in Japanese population. *Clin. Cancer Res.* **7**, 3144–3150 (2001).
50. Y. C. Kim, L. Zhao, H. Zhang, Y. Huang, J. Cui, F. Xiao, B. Downs, S. M. Wang, Prevalence and spectrum of BRCA germline variants in mainland Chinese familial breast and ovarian cancer patients. *Oncotarget* **7**, 9600–9612 (2016).
51. H. Kim, D.-Y. Cho, D. H. Choi, S.-Y. Choi, I. Shin, W. Park, S. J. Huh, S.-H. Han, M. H. Lee, S. H. Ahn, B. H. Son, S.-W. Kim, Korean Breast Cancer Study Group, B. G. Haffty, Characteristics and spectrum of BRCA1 and BRCA2 mutations in 3,922 Korean patients with breast and ovarian cancer. *Breast Cancer Res. Treat.* **134**, 1315–1326 (2012).
52. Y. Momozawa, R. Sasai, Y. Usui, K. Shiraishi, Y. Iwasaki, Y. Taniyama, M. T. Parsons, K. Mizukami, Y. Sekine, M. Hirata, Y. Kamatani, M. Endo, C. Inai, S. Takata, H. Ito, T. Kohno, K. Matsuda, S. Nakamura, K. Sugano, T. Yoshida, H. Nakagawa, K. Matsuo, Y. Murakami, A. B. Spurdle, M. Kubo, Expansion of cancer risk profile for BRCA1 and BRCA2 pathogenic variants. *JAMA Oncol.* **8**, 871–878 (2022).
53. K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnström, S. Mallick, A. Kirby, D. P. Wall, D. G. MacArthur, S. B. Gabriel, M. DePristo, S. M. Purcell, A. Palotie, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, R. A. Gibbs, G. D. Schellenberg, J. S. Sutcliffe, B. Devlin, K. Roeder, B. M. Neale, M. J. Daly, A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
54. L. Sundaram, H. Gao, S. R. Padigepati, J. F. McRae, Y. Li, J. A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon, J. Xu, S. Batzoglou, X. Li, K. K.-H. Farh, Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
55. C. C. Paulusma, M. Kool, P. J. Bosma, G. L. Scheffer, F. Ter Borg, R. J. Scheper, G. N. Tytgat, P. Borst, F. Baas, R. O. Elferink, A mutation in the human canalicular multispecific organic anion transporter gene causes the Dubin-Johnson syndrome. *Hepatology* **25**, 1539–1542 (1997).
56. K. Hashimoto, T. Uchiumi, T. Konno, T. Ebihara, T. Nakamura, M. Wada, S. Sakisaka, F. Maniwa, T. Amachi, K. Ueda, M. Kuwano, Trafficking and functional defects by mutations of the ATP-binding domains in MRP2 in patients with Dubin-Johnson syndrome. *Hepatology* **36**, 1236–1245 (2002).
57. R. L. Snoeckx, P. L. M. Huygen, D. Feldmann, S. Marlin, F. Denoyelle, J. Waligora, M. Mueller-Malesinska, A. Pollak, R. Ploski, A. Murgia, E. Orzan, P. Castorina, U. Ambrosetti, E. Nowakowska-Szyrwinska, J. Bal, W. Wiszniewski, A. R. Janecke, D. Nekahm-Heis, P. Seeman, O. Bendova, M. A. Kenna, A. Frangulov, H. L. Rehm, M. Tekin, A. Incesulu, H.-H. M. Dahl, D. du Sart, L. Jenkins, D. Lucas, M. Bitner-Glindzicz, K. B. Avraham, Z. Brownstein, I. del Castillo, F. Moreno, N. Blin, M. Pfister, I. Sziklai, T. Toth, P. M. Kelley, E. S. Cohn, L. Van Maldergem, P. Hilbert, A.-F. Roux, M. Mondain, L. H. Hoefsloot, C. W. R. J. Cremers, T. Löppönen, H. Löppönen, A. Parving, K. Gronskov, I. Schrijver, J. Roberson, F. Gualandi, A. Martini, G. Lina-Granade, N. Pallares-Ruiz, C. Correia, G. Fialho, K. Cryns, N. Hilgert, P. Van de Heyning, C. J. Nishimura, R. J. H. Smith, G. Van Camp, GJB2 mutations and degree of hearing loss: A multicenter study. *Am. J. Hum. Genet.* **77**, 945–957 (2005).
58. A. W. Stoker, Receptor tyrosine phosphatases in axon growth and guidance. *Curr. Opin. Neurobiol.* **11**, 95–102 (2001).
59. N. Choucair, C. Mignon-Ravix, P. Cacciagli, J. Abou Ghoch, A. Fawaz, A. Mégarbané, L. Villard, E. Chouery, Evidence that homozygous PTPRD gene microdeletion causes trigonocephaly, hearing loss, and intellectual disability. *Mol. Cytogenet.* **8**, 39 (2015).
60. T. Groza, F. L. Gomez, H. H. Mashhadi, V. Muñoz-Fuentes, O. Gunes, R. Wilson, P. Cacheiro, A. Frost, P. Keskivali-Bond, B. Vardal, A. McCoy, T. K. Cheng, L. Santos, S. Wells, D. Smedley, A.-M. Mallon, H. Parkinson, The International Mouse Phenotyping Consortium: Comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res.* **51**, D1038–D1045 (2023).
61. P. Qin, M. Stoneking, Denisovan ancestry in East Eurasian and Native American populations. *Mol. Biol. Evol.* **32**, 2665–2674 (2015).
62. S. R. Browning, B. L. Browning, Y. Zhou, S. Tucci, J. M. Akey, Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell* **173**, 53–61.e9 (2018).
63. L. Chen, A. B. Wolf, W. Fu, L. Li, J. M. Akey, Identifying and interpreting apparent Neanderthal ancestry in African individuals. *Cell* **180**, 677–687.e16 (2020).
64. M. Larena, J. McKenna, F. Sanchez-Quinto, C. Bernhardsson, C. Ebeo, R. Reyes, O. Casel, J.-Y. Huang, K. P. Hagada, D. Guilay, J. Reyes, F. P. Allian, V. Mori, L. S. Azarcon, A. Manera, C. Terando, L. Jamero, G. Sireg, R. Manginsay-Tremedal, M. S. Labos, R. D. Vilar, A. Latiph, R. L. Saway, E. Marte, P. Magbanua, A. Morales, I. Java, R. Reveche, B. Barrios, E. Burton, J. C. Salon, M. J. T. Kels, A. Albano, R. B. Cruz-Angeles, E. Molanida, L. Granehäll, M. Vicente, H. Edlund, J.-H. Loo, J. Trejaut, S. Y. W. Ho, L. Reid, K. Lambeck, H. Malmström, C. Schlebusch, P. Endicott, M. Jakobsson, Philippine Ayta possess the highest level of Denisovan ancestry in the world. *Curr. Biol.* **31**, 4219–4230.e10 (2021).
65. M. Dannemann, The population-specific impact of Neandertal introgression on human disease. *Genome Biol. Evol.* **13**, evaa250 (2021).
66. N. Tanaka, M. Koido, A. Suzuki, N. Otomo, H. Suetsugu, Y. Kochi, K. Tomizuka, Y. Momozawa, Y. Kamatani, Biobank. Japan. Project, S. Ikegawa, K. Yamamoto, C. Terao, Eight novel susceptibility loci and putative causal variants in atopic dermatitis. *J. Allergy Clin. Immunol.* **148**, 1293–1306 (2021).
67. M. Koido, C.-C. Hon, S. Koyama, H. Kawaji, Y. Murakawa, K. Ishigaki, K. Ito, J. Sese, N. F. Parrish, Y. Kamatani, P. Carninci, C. Terao, Prediction of the cell-type-specific transcription of non-coding RNAs from genome sequences via machine learning. *Nat. Biomed. Eng.* **7**, 830–844 (2023).
68. K. Suzuki, M. Akiyama, K. Ishigaki, M. Kanai, J. Hosoe, N. Shojima, A. Hozawa, A. Kadota, K. Kuriki, M. Naito, K. Tanno, Y. Ishigaki, M. Hirata, K. Matsuda, N. Iwata, M. Ikeda, N. Sawada, T. Yamaji, M. Iwasaki, S. Ikegawa, S. Maeda, Y. Murakami, K. Wakai, S. Tsugane, M. Sasaki, M. Yamamoto, Y. Okada, M. Kubo, Y. Kamatani, M. Horikoshi, T. Yamauchi, T. Kadowaki, Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nat. Genet.* **51**, 379–386 (2019).
69. L. Skov, M. Coll Macià, G. Sveinbjörnsson, F. Mafessoni, E. A. Lucotte, M. S. Einarsdóttir, H. Jonsson, B. Halldorsson, D. F. Gudbjartsson, A. Helgason, M. H. Schierup, K. Stefansson, The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. *Nature* **582**, 78–83 (2020).
70. H. Zeberg, S. Pääbo, The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**, 610–612 (2020).
71. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, A map of recent positive selection in the human genome. *PLOS Biol.* **4**, e72 (2006).
72. P. F. Palamara, J. Terhorst, Y. S. Song, A. L. Price, High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nat. Genet.* **50**, 1311–1317 (2018).
73. Y. Okada, Y. Momozawa, S. Sakaue, M. Kanai, K. Ishigaki, M. Akiyama, T. Kishikawa, Y. Arai, T. Sasaki, K. Kosaki, M. Suematsu, K. Matsuda, K. Yamamoto, M. Kubo, N. Hirose, Y. Kamatani, Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat. Commun.* **9**, 1631 (2018).
74. X. Liu, M. Matsunami, M. Horikoshi, S. Ito, Y. Ishikawa, K. Suzuki, Y. Momozawa, S. Niida, R. Kimura, K. Ozaki, S. Maeda, M. Imamura, C. Terao, Natural selection signatures in the Hondo and Ryukyu Japanese subpopulations. *Mol. Biol. Evol.* **40**, msad231 (2023).
75. K. Hanihara, Dual structure model for the population history of the Japanese. *Jpn. Rev.* **2**, 1–33 (1991).
76. M. J. Hudson, S. Nakagome, J. B. Whitman, The evolving Japanese: The dual structure hypothesis at 30. *Evol. Hum. Sci.* **2**, e6 (2020).

77. N. Osada, Y. Kawai, Exploring models of human migration to the Japanese archipelago using genome-wide genetic data. *Anthropol. Sci.* **129**, 45–58 (2021).
78. T. A. Jinam, Y. Kawai, N. Saitou, Modern human DNA analyses with special reference to the inner dual-structure model of Yaponesian. *Anthropol. Sci.* **129**, 3–11 (2021).
79. K. Mizoguchi, *The Archaeology of Japan: From the Earliest Rice Farming Villages to the Rise of the State* (Cambridge Univ. Press, ed. 1, 2013).
80. G. Barnes, *State Formation in Japan: Emergence of a 4th-Century Ruling Elite* (Routledge, 2007).
81. C.-H. Huang, D. C. Kang, State formation in Korea and Japan, 400–800 CE: Emulation and learning, not bellicist competition. *Int. Organ.* **76**, 1–31 (2022).
82. K. F. Friday, Pushing beyond the Pale: The Yamato Conquest of the Emishi and Northern Japan. *J. Jpn. Stud.* **23**, 1–24 (1997).
83. K. Hanihara, Emishi, Ezo and Ainu: An anthropological perspective. *Jpn. Rev.* **1**, 35–48 (1990).
84. H. Matsumura, Y. Dodo, Dental characteristics of Tohoku residents in Japan: Implications for biological affinity with ancient Emishi. *Anthropol. Sci.* **117**, 95–105 (2009).
85. E. De Boer, M. A. Yang, A. Kawagoe, G. L. Barnes, Japan considered from the hypothesis of farmer/language spread. *Evol. Hum. Sci.* **2**, e13 (2020).
86. E. R. Crema, C. J. Stevens, S. Shoda, Bayesian analyses of direct radiocarbon dates reveal geographic variations in the rate of rice farming dispersal in prehistoric Japan. *Sci. Adv.* **8**, eadc9171 (2022).
87. E. Endo, C. Leipe, The onset, dispersal and crop preferences of early agriculture in the Japanese archipelago as derived from seed impressions in pottery. *Quat. Int.* **623**, 35–49 (2022).
88. B. B. Cummings, K. J. Karczewski, J. A. Kosmicki, E. G. Seaby, N. A. Watts, M. Singer-Berk, J. M. Mudge, J. Karjalainen, F. K. Satterstrom, A. H. O'Donnell-Luria, T. Poterba, C. Seed, M. Solomonson, J. Alföldi, M. J. Daly, D. G. MacArthur, Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).
89. E. Huerta-Sánchez, X. Jin, Asan, Z. Bianba, B. M. Peter, N. Vinckenbosch, Y. Liang, X. Yi, M. He, M. Somel, P. Ni, B. Wang, X. Ou, J. L. Huasang, Z. X. P. Cuo, K. Li, G. Gao, Y. Yin, W. Wang, X. Zhang, X. Xu, H. Yang, Y. Li, J. Wang, J. Wang, R. Nielsen, Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
90. M. Dannemann, J. Kelso, The contribution of neanderthals to phenotypic variation in modern humans. *Am. J. Hum. Genet.* **101**, 578–589 (2017).
91. L. B. Knudsen, J. Lau, The discovery and development of liraglutide and semaglutide. *Front. Endocrinol.* **10**, 155 (2019).
92. Y. Yasumizu, S. Sakaue, T. Konuma, K. Suzuki, K. Matsuda, Y. Murakami, M. Kubo, P. F. Palamara, Y. Kamatani, Y. Okada, Genome-wide natural selection signatures are linked to genetic risk of modern phenotypes in the Japanese population. *Mol. Biol. Evol.* **37**, 1306–1316 (2020).
93. P.-R. Loh, P. F. Palamara, A. L. Price, Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
94. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
95. M. Petr, B. Vernot, J. Kelso, admixr - R package for reproducible analyses using ADMIXTOOLS. *Bioinformatics* **35**, 3194–3195 (2019).
96. M. A. Yang, X. Fan, B. Sun, C. Chen, J. Lang, Y.-C. Ko, C. Tsang, H. Chiu, T. Wang, Q. Bao, X. Wu, M. Hajdinjak, A. M.-S. Ko, M. Ding, P. Cao, R. Yang, F. Liu, B. Nickel, Q. Dai, X. Feng, L. Zhang, C. Sun, C. Ning, W. Zeng, Y. Zhao, M. Zhang, X. Gao, Y. Cui, D. Reich, M. Stoneking, Q. Fu, Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**, 282–288 (2020).
97. S. J. Jurgens, S. H. Choi, V. N. Morrill, M. Chaffin, J. P. Pirruccello, J. L. Halford, L.-C. Weng, V. Nauffal, C. Roselli, A. W. Hall, M. T. Oetjens, B. Lagerman, D. P. vanMaanen, R. G. Center, G. Abecasis, X. Bai, S. Balasubramanian, A. Baras, C. Beechert, B. Boutkov, M. Cantor, G. Coppola, T. De, A. Deubler, A. Economides, G. Eom, M. A. R. Ferreira, C. Forsythe, E. D. Fuller, Z. Gu, L. Habegger, A. Hawes, M. B. Jones, K. Karalis, S. Khalid, O. Krasheninina, R. Lanche, M. Lattari, D. Li, A. Lopez, L. A. Lotta, K. Manoochehri, A. J. Mansfield, E. K. Maxwell, J. Mighty, L. J. Mitnaul, M. Nafde, J. Nielsen, S. O'Keeffe, M. Orelus, J. D. Overton, M. S. Padilla, R. Panea, T. Polanco, M. Pradhan, A. Rasool, J. G. Reid, W. Salerno, T. D. Schleicher, A. Shuldiner, K. Siminovitch, J. C. Staples, R. H. Ulloa, N. Verweij, L. Widom, S. E. Wolf, K. G. Aragam, K. L. Lunetta, C. M. Haggerty, S. A. Lubitz, P. T. Ellinor, Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. *Nat. Genet.* **54**, 240–250 (2022).
98. K. Ishigaki, M. Akiyama, M. Kanai, A. Takahashi, E. Kawakami, H. Sugishita, S. Sakaue, N. Matoba, S.-K. Low, Y. Okada, C. Terao, T. Amariuta, S. Gazal, Y. Kochi, M. Horikoshi, K. Suzuki, K. Ito, S. Koyama, K. Ozaki, S. Niida, Y. Sakata, Y. Sakata, T. Kohno, K. Shiraishi, Y. Momozawa, M. Hirata, K. Matsuda, M. Ikeda, N. Iwata, S. Ikegawa, I. Kou, T. Tanaka, H. Nakagawa, A. Suzuki, T. Hirota, M. Tamari, K. Chayama, D. Miki, M. Mori, S. Nagayama, Y. Daigo, Y. Miki, T. Katagiri, O. Ogawa, W. Obara, H. Ito, T. Yoshida, I. Imoto, T. Takahashi, C. Tanikawa, T. Suzuki, N. Sinozaki, S. Minami, H. Yamaguchi, S. Asai, Y. Takahashi, K. Yamaji, K. Takahashi, T. Fujioka, R. Takata, H. Yanai, A. Masumoto, Y. Koretsune, H. Kutsumi, M. Higashiyama, S. Murayama, N. Minegishi, K. Suzuki, K. Tanno, A. Shimizu, T. Yamaji, M. Iwasaki, N. Sawada, H. Uemura, K. Tanaka, M. Naito, M. Sasaki, K. Wakai, S. Tsugane, M. Yamamoto, K. Yamamoto, Y. Murakami, Y. Nakamura, S. Raychaudhuri, J. Inazawa, T. Yamauchi, T. Kadowaki, M. Kubo, Y. Kamatani, Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679 (2020).
99. Z. A. Szpiech, R. D. Hernandez, selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* **31**, 2824–2827 (2014).
100. J. P. Spence, Y. S. Song, Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Sci. Adv.* **5**, eaaw9206 (2019).
101. X. Liu, S. Takata, K. Ashikawa, T. Aoi, S. Kosugi, C. Terao, N. F. Parrish, K. Matsuda, H. Nakagawa, Y. Kamatani, M. Kubo, Y. Momozawa, Prevalence and spectrum of pathogenic germline variants in Japanese patients with early-onset colorectal, breast, and prostate cancer. *JCO Precis. Oncol.* **4**, 183–191 (2020).
102. S. Ito, X. Liu, Y. Ishikawa, D. D. Conti, N. Otomo, Z. Kote-Jarai, H. Suetsugu, R. A. Eeles, Y. Koike, K. Hikino, S. Yoshino, K. Tomizuka, M. Horikoshi, K. Ito, Y. Uchio, Y. Momozawa, M. Kubo, T. B. B. J. Project, A. Masumoto, A. Nagai, D. Obata, H. Yamaguchi, K. Muto, K. Takahashi, K. Yamaji, K. Yoshimori, M. Higashiyama, N. Sinozaki, S. Asai, S. Nagayama, S. Murayama, S. Minami, T. Suzuki, T. Morisaki, W. Obara, Y. Takahashi, Y. Furukawa, Y. Murakami, Y. Yamanashi, Y. Koretsune, Y. Kamatani, K. Matsuda, C. A. Haiman, S. Ikegawa, H. Nakagawa, C. Terao, Androgen receptor binding sites enabling genetic prediction of mortality due to prostate cancer in cancer-free subjects. *Nat. Commun.* **14**, 4863 (2023).
103. Y. Ishikawa, N. Tanaka, Y. Asano, M. Kodera, Y. Shirai, M. Akahoshi, M. Hasegawa, T. Matsushita, K. Saito, S. Motegi, H. Yoshifuji, A. Yoshizaki, T. Kohmoto, K. Takagi, A. Oka, M. Kanda, Y. Tanaka, Y. Ito, K. Nakano, H. Kasamatsu, A. Utsunomiya, A. Sekiguchi, H. Niiro, M. Jinnin, K. Makino, T. Makino, H. Ihn, M. Yamamoto, C. Suzuki, H. Takahashi, E. Nishida, A. Morita, T. Yamamoto, M. Fujimoto, Y. Kondo, D. Goto, T. Sumida, N. Ayuzawa, H. Yanagida, T. Horita, T. Atsumi, H. Endo, Y. Shima, A. Kumanogoh, J. Hirata, N. Otomo, H. Suetsugu, Y. Koike, K. Tomizuka, S. Yoshino, X. Liu, S. Ito, K. Hikino, A. Suzuki, Y. Momozawa, S. Ikegawa, Y. Tanaka, O. Ishikawa, K. Takehara, T. Torii, S. Sato, Y. Okada, T. Mimori, F. Matsuda, K. Matsuda, T. Amariuta, I. Imoto, K. Matsuo, M. Kuwana, Y. Kawaguchi, K. Ohmura, C. Terao, GWAS for systemic sclerosis identifies six novel susceptibility loci including one in the Fcγ receptor region. *Nat. Commun.* **15**, 319 (2024).
104. S. Kosugi, Y. Kamatani, K. Harada, K. Tomizuka, Y. Momozawa, T. Morisaki, C. Terao, Detection of trait-associated structural variations using short-read sequencing. *Cell Genomics* **3**, 100328 (2023).
105. A. Diaz-Papkovich, L. Anderson-Trocmé, S. Gravel, A review of UMAP in population genetics. *J. Hum. Genet.* **66**, 85–91 (2021).
106. Q. Liu, D. Wu, C. Wang, Identification of genomic regions distorting population structure inference in diverse continental groups. *Quant. Biol.* **10**, 287–298 (2022).
107. T. Sato, S. Nakagome, C. Watanabe, K. Yamaguchi, A. Kawaguchi, K. Koganebuchi, K. Haneji, T. Yamaguchi, T. Hanihara, K. Yamamoto, 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.* **31**, 2929–2940 (2014).
108. M. Matsunami, K. Koganebuchi, M. Imamura, H. Ishida, R. Kimura, S. Maeda, Fine-scale genetic structure and demographic history in the Miyako Islands of the Ryukyu Archipelago. *Mol. Biol. Evol.* **38**, 2045–2056 (2021).
109. K. Koganebuchi, K. Haneji, T. Toma, K. Joh, H. Soejima, K. Fujimoto, H. Ishida, M. Ogawa, T. Hanihara, S. Harada, S. Kawamura, H. Oota, The allele frequency of *ALDH2*Glu504Lys* and *ADH1B*Arg47His* for the Ryukyu islanders and their history of expansion among East Asians. *Am. J. Hum. Biol.* **29**, e22933 (2017).
110. Y. Watanabe, J. Ohashi, Modern Japanese ancestry-derived variants reveal the formation process of the current Japanese regional gradations. *iScience* **26**, 106130 (2023).
111. A. M. Chiu, E. K. Molloy, Z. Tan, A. Talwalkar, S. Sankararaman, Inferring population structure in biobank-scale genomic data. *Am. J. Hum. Genet.* **109**, 727–737 (2022).
112. X. Zhang, K. E. Witt, M. M. Bañuelos, A. Ko, K. Yuan, S. Xu, R. Nielsen, E. Huerta-Sanchez, The history and evolution of the Denisovan- *EPAS1* haplotype in Tibetans. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2020803118 (2021).
113. R. Torrance, The infrastructure of the gods: Izumo in the Yayoi and Kofun periods. *Jpn. Rev.* **29**, 3–38 (2016).
114. Y. Zhong, *The Origin of Modern Shinto in Japan: The Vanquished Gods of Izumo* (Bloomsbury Academic, Paperback edition, 2018).
115. H. Oota, A. J. Pakstis, B. Bonne-Tamir, D. Goldman, E. Grigorenko, S. L. B. Kajuna, N. J. Karoma, S. Kungulilo, R.-B. Lu, K. Odunsi, F. Okonofua, O. V. Zhukova, J. R. Kidd, K. K. Kidd, The evolution and population genetics of the ALDH2 locus: Random genetic drift, selection, and low levels of recombination. *Ann. Hum. Genet.* **68**, 93–109 (2004).
116. D. Goldman, M.-A. Enoch, Genetic epidemiology of ethanol metabolic enzymes: A role for selection, in *World Review of Nutrition and Dietetics*, A. P. Simopoulos, B. Childs, Eds. (S. Karger AG, 1990), vol. 63, pp. 143–160.
117. A. Espinosa, L. Yan, Z. Zhang, L. Foster, D. Clark, E. Li, S. L. Stanley, The bifunctional entamoeba histolytica alcohol dehydrogenase 2 (EhADH2) protein is necessary for amebic growth and survival and requires an intact C-terminal domain for both alcohol

dehydrogenase and acetaldehyde dehydrogenase activity. *J. Biol. Chem.* **276**, 20136–20143 (2001).
118. E. Strauss, Coenzyme A biosynthesis and enzymology, in *Comprehensive Natural Products II* (Elsevier, 2010), pp. 351–410.
119. A. Azam, M. N. Peerzada, K. Ahmad, Parasitic diarrheal disease: Drug development and targets. *Front. Microbiol.* **6**, 1183 (2015).
120. N. Tran-Thi, R. J. Lowe, J. M. Schurer, T. Vu-Van, L. E. MacDonald, P. Pham-Duc, Turning poop into profit: Cost-effectiveness and soil transmitted helminth infection risk associated with human excreta reuse in Vietnam. *PLOS Negl. Trop. Dis.* **11**, e0006088 (2017).
121. H. Takamiya, N. Nakamura, The beginning of agriculture in the Ryukyu archipelago. *S. Pac. Study* **41**, 1–34 (2021).
122. X. Zhang, A. Sun, J. Ge, Origin and spread of the ALDH2 Glu504Lys allele. *Phenomics* **1**, 222–228 (2021).
123. N. Matoba, M. Akiyama, K. Ishigaki, M. Kanai, A. Takahashi, Y. Momozawa, S. Ikegawa, M. Ikeda, N. Iwata, M. Hirata, K. Matsuda, Y. Murakami, M. Kubo, Y. Kamatani, Y. Okada, GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits. *Nat. Hum. Behav.* **4**, 308–316 (2020).
124. S. Sakaue, M. Akiyama, M. Hirata, K. Matsuda, Y. Murakami, M. Kubo, Y. Kamatani, Y. Okada, Functional variants in ADH1B and ALDH2 are non-additively associated with all-cause mortality in Japanese population. *Eur. J. Hum. Genet.* **28**, 378–382 (2020).

Acknowledgments: We express our gratitude to all volunteers enrolled in the BBJ project, as well as the doctors, medical staff, and research personnel at the participating hospitals and study sites. We thank A. Tajima (Kanazawa University), P. Qin (BGI), M. Hudson (Max Planck Institute), D. Falush (Institute Pasteur Shanghai), D. Lawson (University of Bristol), and S. Nakagome (Trinity College Dublin) for helpful advice. We thank M. Larena and M. Jakobsson (Uppsala University) for providing introgression data of Philippine Ayta. We appreciate constructive comments and suggestions from three anonymous reviewers. X.L. would like to thank A. Lysenko (University of Tokyo) for critical reading and editing of the initial draft. T.G. would like to express gratitude for the support provided by the Sakigake Project at Kanazawa University. Y.Ka. has received speaking honoraria from Illumina Japan. **Funding:** This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant (JP20H00462 to C.T.) and Japan Agency for Medical Research and Development (AMED) (JP21ek0109555, JP21tm0424220, JP21ck0106642, JP23ek0410114, and JP23tm0424225 to C.T. and JP18km0605001 to Y.Ka. and K.M.). **Author contributions:** The conceptual framework was developed by X.L., S.It., T.G., Y.M., K.I., M.H., and C.T. Methodological design was formulated by X.L., S.It., T.G., Y.M., M.H., and C.T. The curation of data was done by Y.M., Y.I., S.Kos., K.I., and C.T. Analytical pipeline was developed by X.L., S.Koy., K.S., S.It., and C.T. The formal analysis was conducted by X.L., T.G., K.S., S.Kos., S.It., and C.T. Visualization tasks were performed by X.L., S.Koy., K.H., M.H., and C.T. Validation was executed by X.L., S.T., and C.T. Research investigations were conducted by X.L., M.K., Y.M., K.T., K.S., S.T., K.I., and C.T. The original draft was authored by X.L., T.G., Y.Ko., S.Koy., K.H., K.I., M.H., and C.T. Review and editing of the draft and revisions were undertaken by X.L., T.G., Y.Ko., S.Koy., M.K., Y.M., Y.I., K.H., K.I., M.H., S.It., and C.T. The resources were acquired by Y.Ka., M.K., Y.M., S.Ik., S.T., K.M., and C.T. The funding was obtained by Y.Ka., K.M., and C.T. Supervision was conducted by S.Ik., M.H., and C.T. Project administration was done by C.T. All authors reviewed the results and approved the final version of the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** We shared the personal genomic data as Controlled-Access Data on the English website of the National Bioscience Database Center (NBDC) (https://humandbs.biosciencedbc.jp/en/). The WGS data, including raw FASTQ, BAM, and VCF files, were deposited into the NBDC Human Database under an umbrella research ID hum0014 (current version 30). The child IDs associated with the data are JGAS000381 and JGAS000114. The clinical information corresponding to the disclosed data could be applied through the BBJ official website (https://biobankjp.org/en/info/nbdc.html) if the application complies with Japanese privacy laws. Users are also encouraged to contact either BBJ or NBDC directly regarding any issues. The summary statistics from the iHS and FastSMC analysis, the called introgression segments by IBDMix and Sprime, and site VCF files containing the AF information are available from JENGER, a website of the Laboratory for Statistical and Translational Genetics at RIKEN IMS (http://jenger.riken.jp/en/data). All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 21 May 2023
Accepted 7 March 2024
Published 17 April 2024
10.1126/sciadv.adi8419