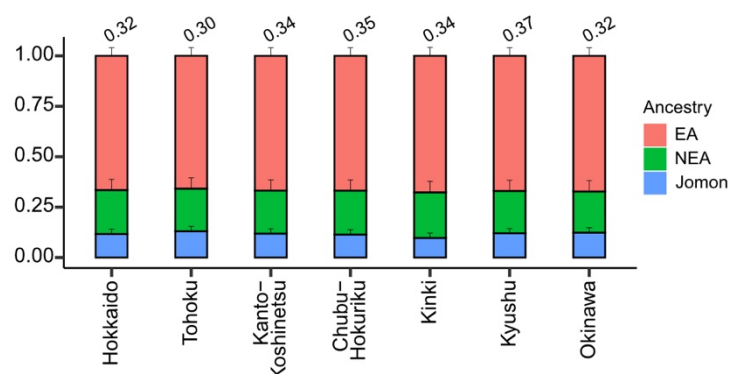


Supplementary Information
Genetic Legacy of Ancient Hunter-Gatherer Jomon in Japanese Populations

Kenichi Yamamoto et al.

Corresponding to Shigeki Nakagome (nakagoms@tcd.ie) and
Yukinori Okada (yokada@sg.med.osaka-u.ac.jp).

Supplementary Figures



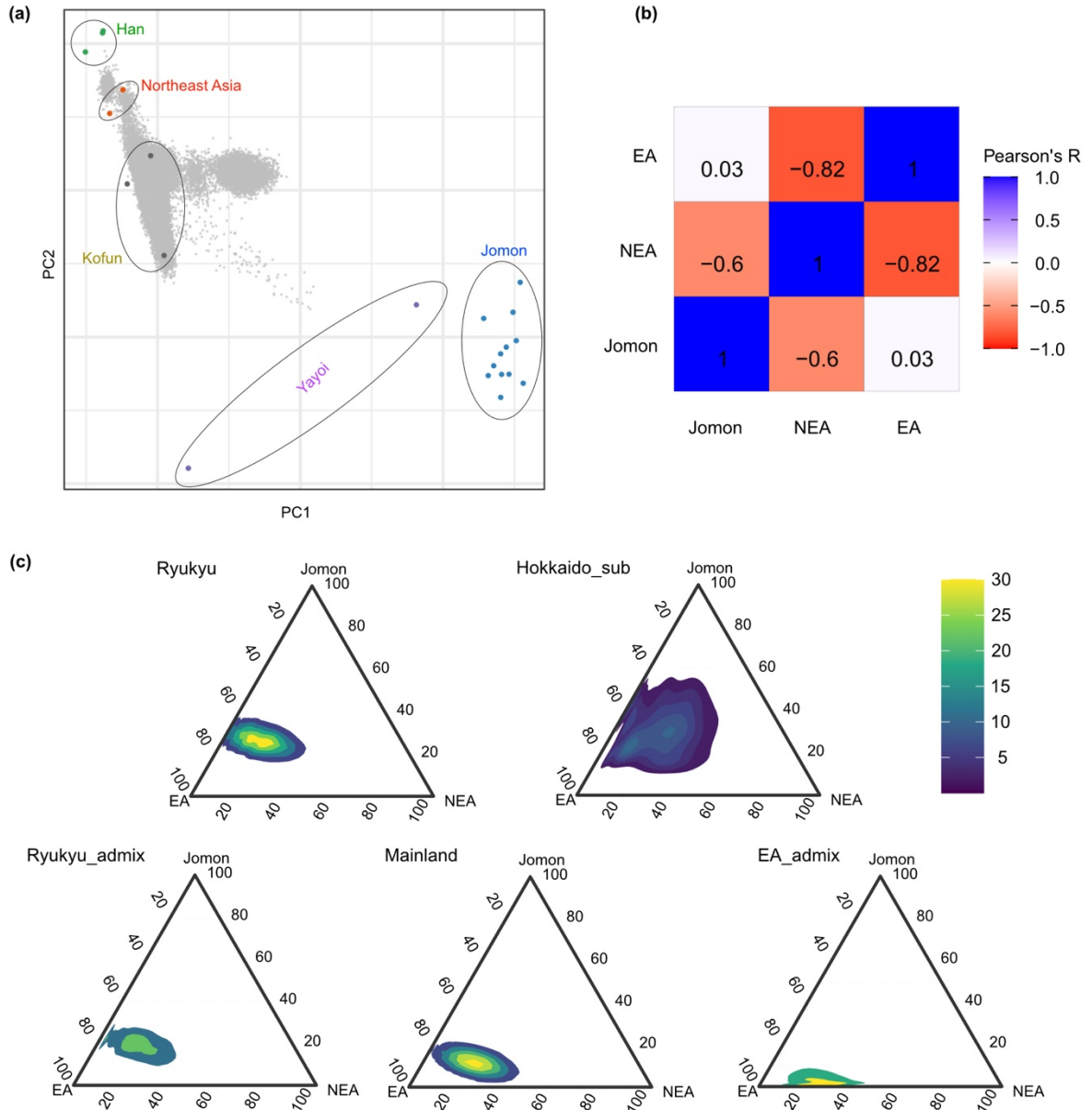
Supplementary Fig. 1. Variation in tripartite ancestry structures across groups of individuals in the Mainland cluster based on PCA.

Bar plots show the proportions of three distinct ancestors. Individuals within the Mainland cluster of BBJ ($n = 159,642$) are split into seven different regions based on their registered location (Hokkaido: $n = 7,829$, Tohoku: $n = 10,991$, Kanto-Koshinetsu: $n = 93,827$, Chubu-Hokuriku: $n = 9,403$, Kinki: $n = 24,365$, Kyushu: $n = 12,068$, Okinawa: $n = 284$). Proportions in the bar plots represent the mean values of three ancestral components estimated using qpAdm, with error bars indicating standard errors. Values at the top of each bar indicate the tail probabilities of the tripartite model for each group. EA, East Asian; NEA, Northeast Asian.



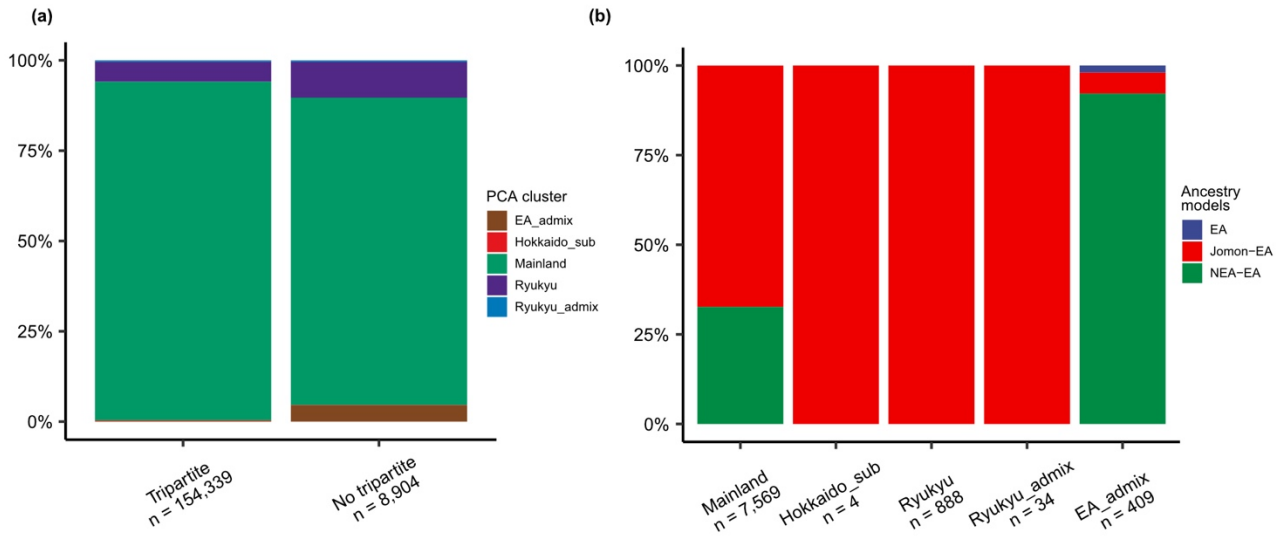
Supplementary Fig. 2. Genetic affinities of Jomon with present-day populations in the SGDP panel, 1000 Genomes Project (1KG), and Biobank Japan.

Bar plots represent Z-scores from f_4 -tests with the form of $f_4(\text{Mbuti}, \text{Jomon}; \text{Han}, X)$, where X is a population from the SGDP dataset, 1KG, or sub-populations within the BBJ dataset. Positive or negative Z-scores are colored in right blue or pink. If the Z score is greater than 3, the bar is colored in green. The target populations are grouped by: Southeast Asia, South Asia, Oceania, Siberia, East Asia in SGDP; AFR, AMR, EUR, EAS, SAS in 1KG; JPT, CHS, CHB, CDX, and KHV in EAS in 1KG; Hokkaido_sub, Ryukyu, Ryukyu_admix, Mainland, and EAS_admix in Biobank Japan.



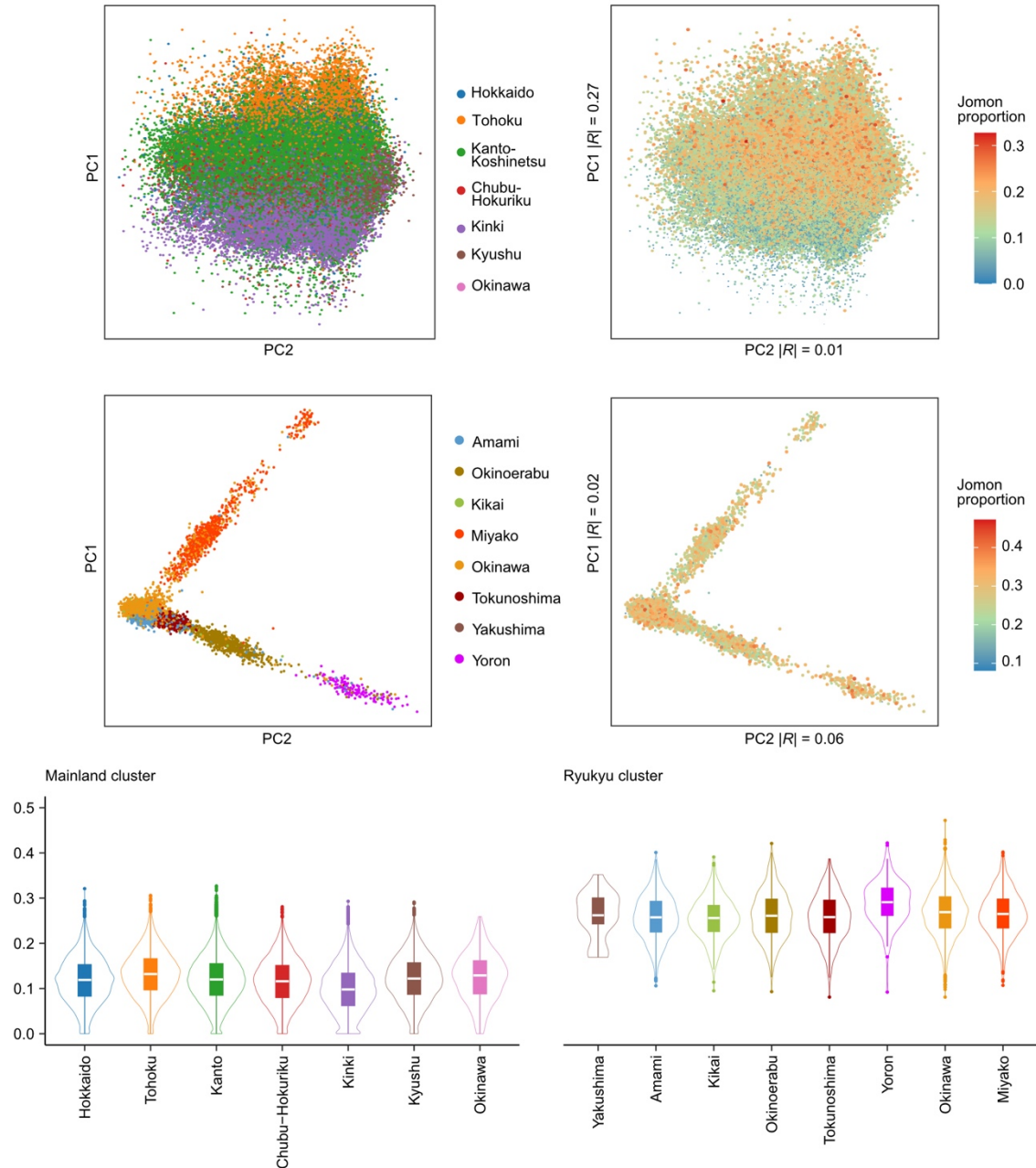
Supplementary Fig. 3. Positional and proportional associations among three ancestors of modern Japanese.

a) Ancient Japanese and those who represent continental ancestors in the tripartite model are projected onto the PCA of present-day Japanese individuals (shown by gray dots). **b)** A correlation matrix of three ancestors is measured from their proportions among all BBJ individuals. **c)** Ternary plots for the composition of three ancestors in the genetic clusters defined by PCA. EA, East Asian ancestry; NEA, Northeast Asian ancestry.



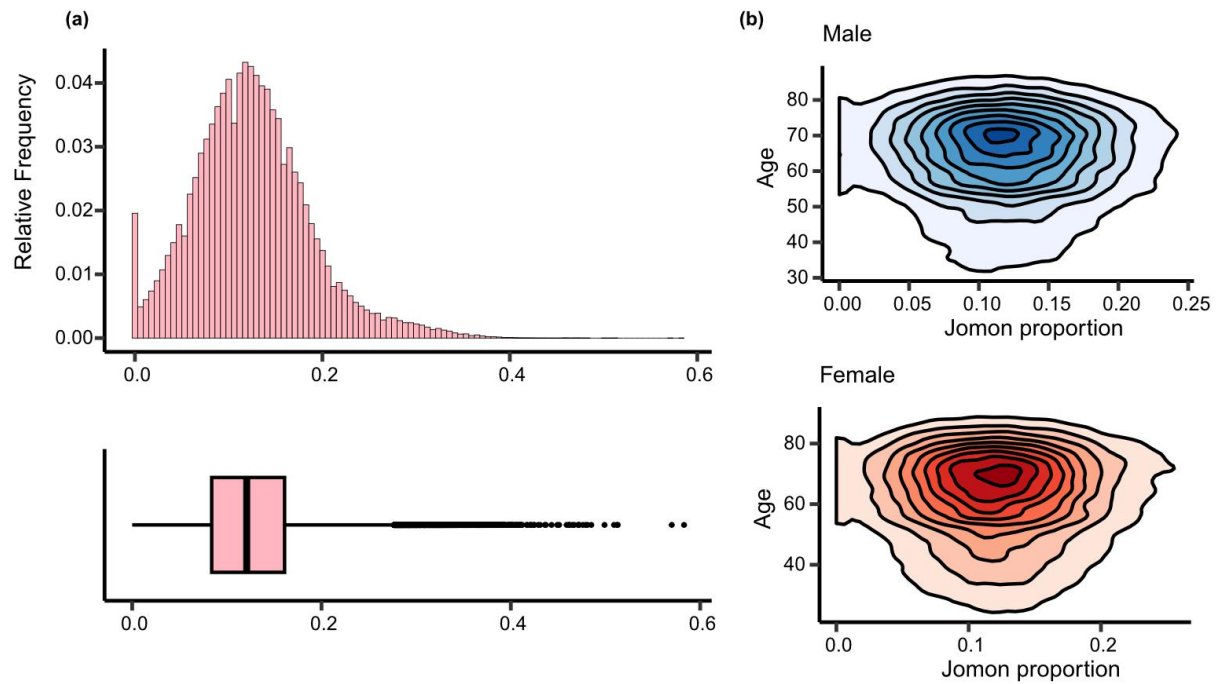
Supplementary Fig. 4. Comparisons of admixture scenarios in different geographic contexts.

a) Bar plots show the composition of PCA clusters within individuals supporting either the tripartite model (154,339 individuals labeled as "Tripartite") or alternative scenarios (8,904 individuals labeled as "No tripartite"). In the tripartite model, each proportion of the five clusters (EA_admix, Hokkaido_sub, Mainland, Ryukyu, and Ryukyu_admix) is as follows: 0.4%, 0.08%, 93.7%, 5.5%, and 0.4%. The proportions in the non-tripartite models are 4.6%, 0.04%, 85.0%, 10.0%, and 0%. **b)** Bar plots represent a detailed breakdown of "No tripartite" individuals by their PCA-based groups, together with the composition of admixture models (i.e., dual ancestors: Jomon and East Asia or Northeast Asia and East Asia; single ancestor: East Asia).



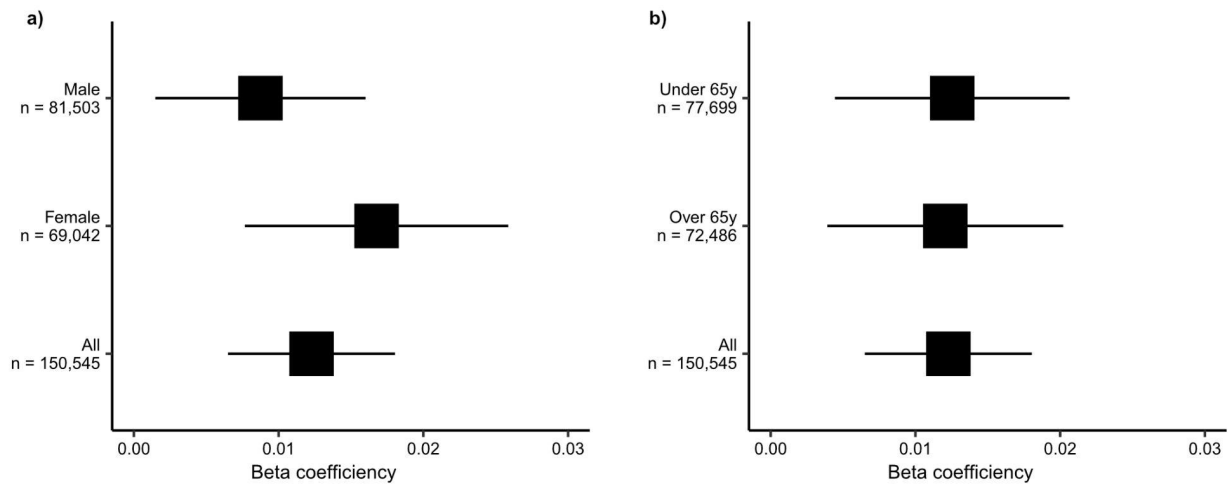
Supplementary Fig. 5. Projection of the proportions of Jomon ancestry onto the PCA plots based on the Mainland or Ryukyu clusters within BBJ.

The proportion of Jomon ancestry is projected onto the PCA plots based on (top) individuals from the Mainland ($n = 152,148$) or (middle) individuals from the Ryukyu clusters ($n = 8,112$). $|R|$ represents the absolute values of correlation coefficients between the proportion of Jomon ancestry and PCs. (bottom) Boxplots represent variation in Jomon ancestry across individuals from different geographic contexts within the Mainland or Ryukyu clusters (Yakushima; $n = 19$, Amami; $n = 1,314$, Kikai; $n = 465$, Okinoerabu; $n = 726$, Tokunoshima; $n = 415$, Yoron; $n = 148$, Okinawa; $n = 4,311$, Miyako; $n = 714$). The boxes denote the interquartile range (IQR) and the median is shown as white horizontal bars; whiskers extend to 1.5 times the IQR; outliers are shown as individual points. PC, Principal component



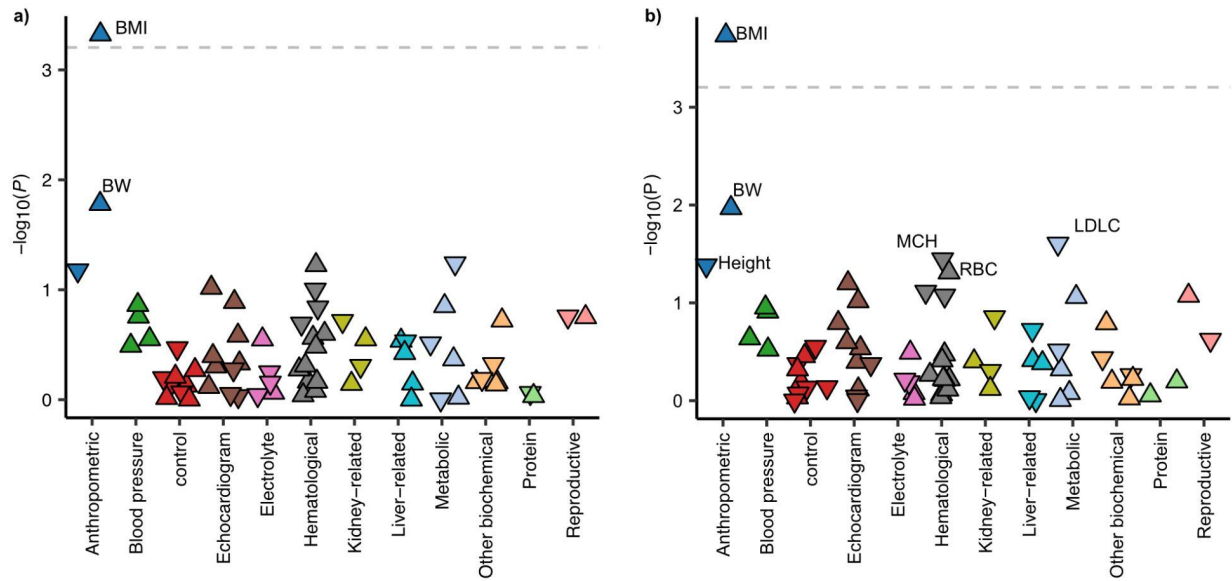
Supplementary Fig. 6. Variation in the proportion of Jomon ancestry among BBJ participants.

a) The distribution of Jomon ancestry in all participants ($n = 163,243$). A box denotes the interquartile range (IQR), with a median shown as a black horizontal bar; whiskers extend to 1.5 times the IQR; outliers are shown as individual points. **b)** The density plots between the Jomon proportion and age, stratified by male and female.



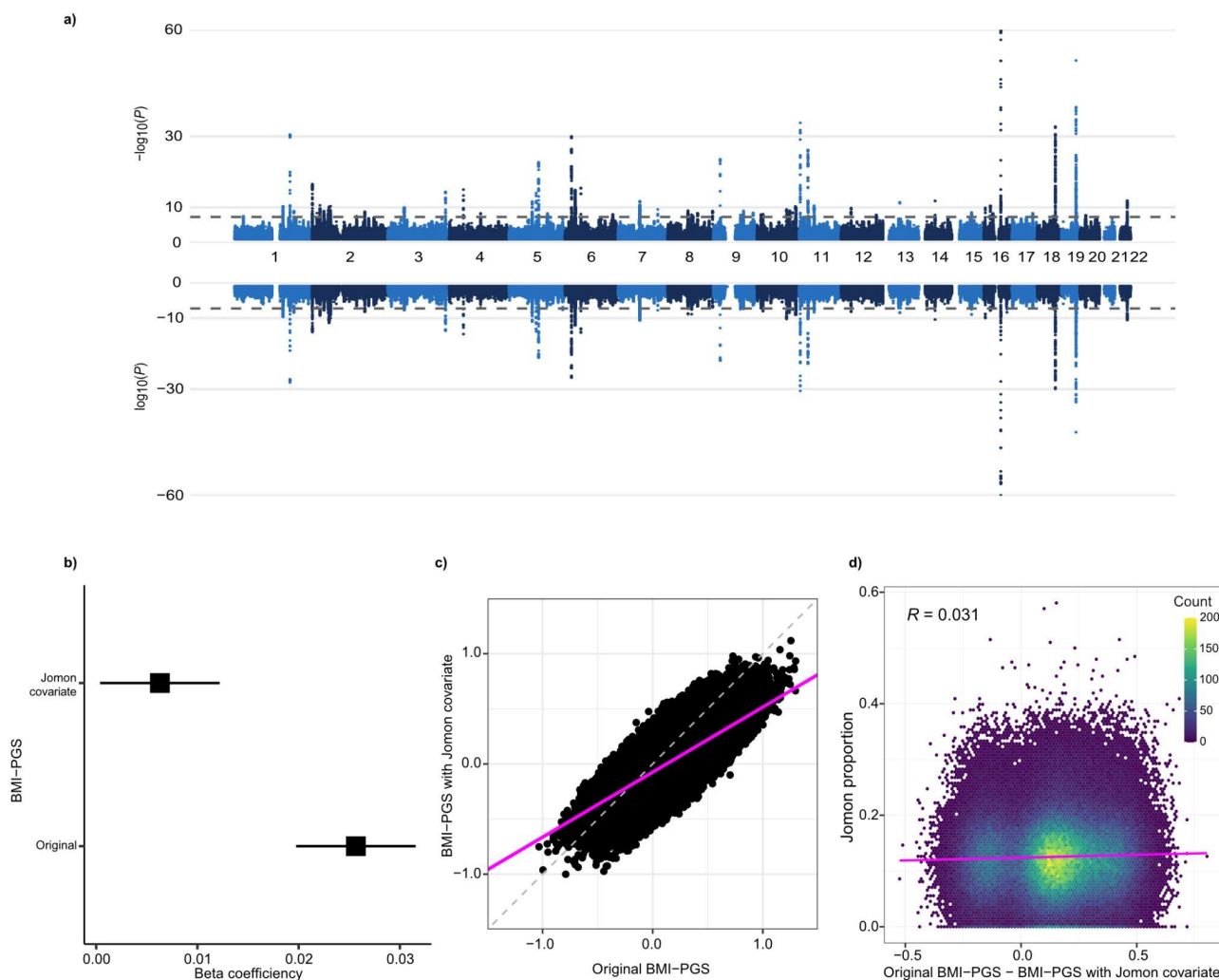
Supplementary Fig. 7. Associations between BMI and Jomon proportions, stratified by sex and age groups.

Forest plots represent the effect size of Jomon ancestry on BMI, stratified by **a)** sex (male or female) or **b)** age groups (older or younger than 65 years old, which is a mean age among the BBJ participants). Squares represent the point estimates, while error bars indicate 95% confidence intervals.



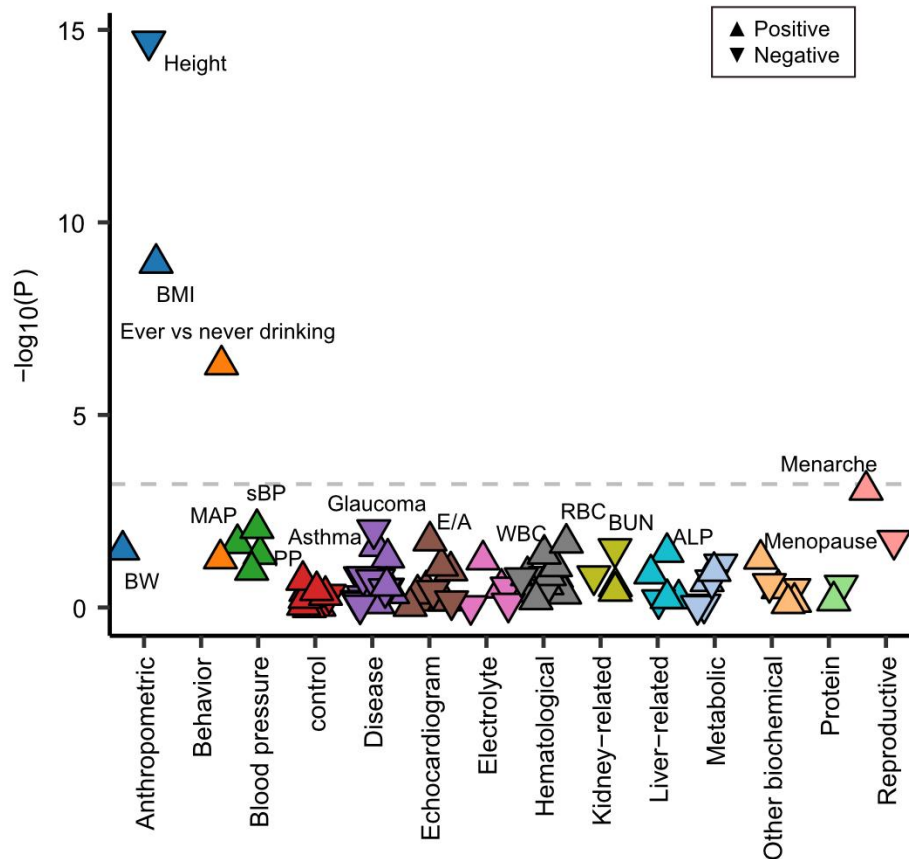
Supplementary Fig. 8. Associations of Jomon ancestry with phenotypes (80 complex traits) corrected for covariates.

Quantitative measures for the traits are regressed out by covariates including PCs before being tested for associations with Jomon ancestry based on a generalized linear model in **a)** the entire BBJ participants ($n = 163,243$) and **b)** Mainland participants ($n = 152,148$). A gray dashed line represents the statistical significance based on the Bonferroni correction ($P < 0.05/80 = 6.3 \times 10^{-4}$). The traits labeled on the plots are nominally significant in their association with Jomon ancestry at $P < 0.05$. P -values are computed by linear regression or logistic regression. All statistical tests are two-sided and unadjusted for multiple comparisons. BMI, Body mass index; BW, Body weight; MCH, mean corpuscular hemoglobin; RBC, Red blood cell; LDLC; low-density lipoprotein cholesterol.



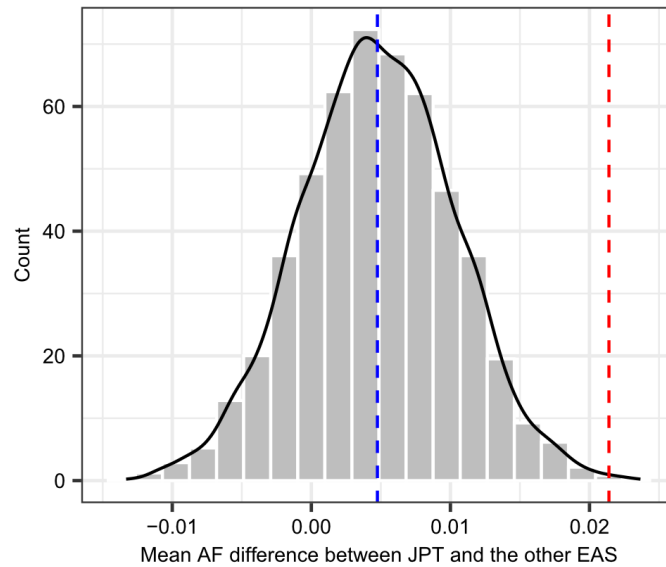
Supplementary Fig. 9. Assessing impacts of Jomon ancestry on GWAS for BMI (BMI-GWAS) and polygenic scores of BMI (BMI-PGS).

a) A Miami-plot of the original BMI-GWAS (top) and the BMI-GWAS including Jomon proportions as a covariate (bottom). P -values are computed by fastGWA. All statistical tests are two-sided and unadjusted for multiple comparisons. **b)** Forest plots contrast the effect sizes of Jomon ancestry on the original BMI-PGS (labeled as "Original ") with the effect on BMI-PGS after correction for Jomon ancestry ("Jomon covariate"). Squares represent the point estimates, and error bars indicate 95% confidence intervals. **c)** Scatter plots compare the values of the original BMI-PGS (i.e., GWAS results without accounting for Jomon ancestry) with those PGS derived from GWAS that incorporate Jomon proportions as a covariate. A dashed line coloured by gray represents $y = x$. **d)** Plots show the association of the difference in BMI-PGS between GWAS with and without using Jomon ancestry as a covariate with the Jomon proportions. A correlation coefficient ($R = 0.031$; $P = 8.5 \times 10^{-37}$) is measured by Pearson's method. P -value reflects a two-sided test.



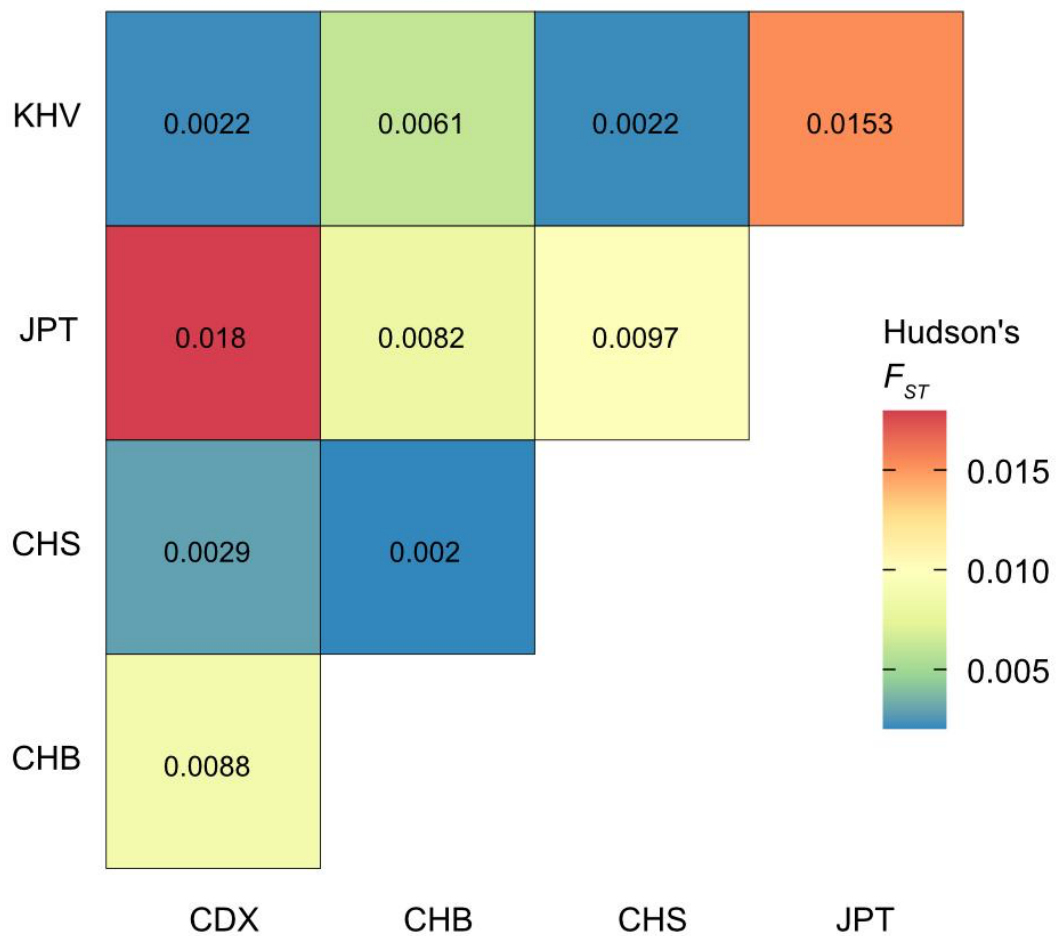
Supplementary Fig. 10. Associations of Jomon ancestry with complex traits without accounting for PCs as covariates.

The associations are tested for 90 traits (80 complex traits and 10 dummy phenotypes) with a generalized linear model that does not account for PCs as covariates ($n = 163,243$). The direction of the triangle corresponds to the sign of the beta coefficient for each trait. The gray dashed line indicates the statistical significance based on the Bonferroni correction ($P < 0.05/80 = 6.3 \times 10^{-4}$). A label "control" includes 10 dummy phenotypes. All traits labeled on the plots have statistical significance with $P < 0.05$. P -values are computed by linear regression or logistic regression. All statistical tests are two-sided and unadjusted for multiple comparisons. BMI, Body mass index; BW, Body weight; sBP, Systolic blood pressure; MAP, Mean arterial pressure; PP, Pulse pressure; E/A, E/A ratio; WBC, White blood cell count; RBC, Red blood cell count; ALP, alkaline phosphatase.



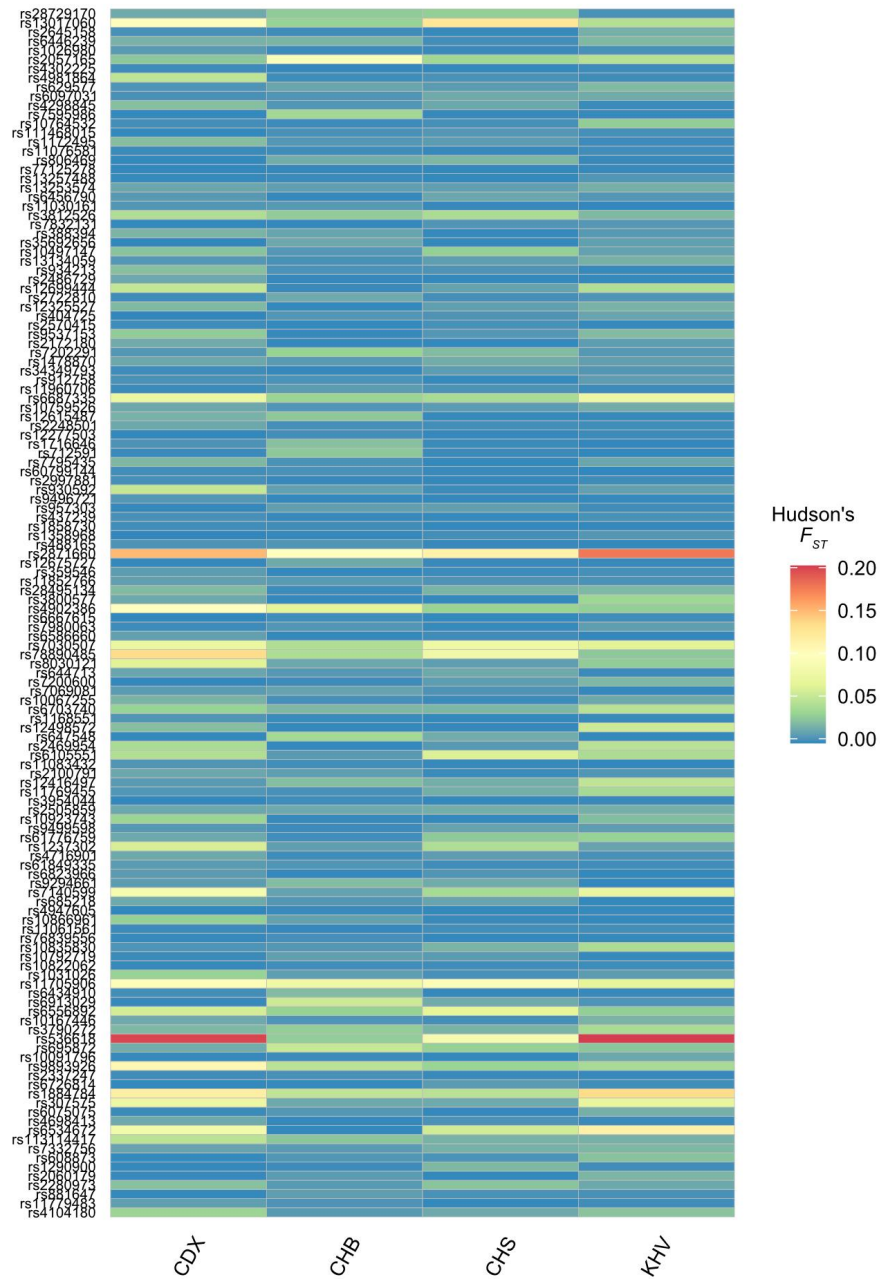
Supplementary Fig 11. Mean allele frequency difference of Jomon-related and non-Jomon related variants between JPT and the other East Asian populations within the 1000 Genome dataset.

A histogram shows the distribution of mean frequency difference for 5,000 sets of randomly sampled 132 non-Jomon-related SNPs, matched by the frequencies of Jomon-related SNPs. A blue dashed line represents the average of the distribution, while a red line indicates the mean AF difference observed in the Jomon-related variants.



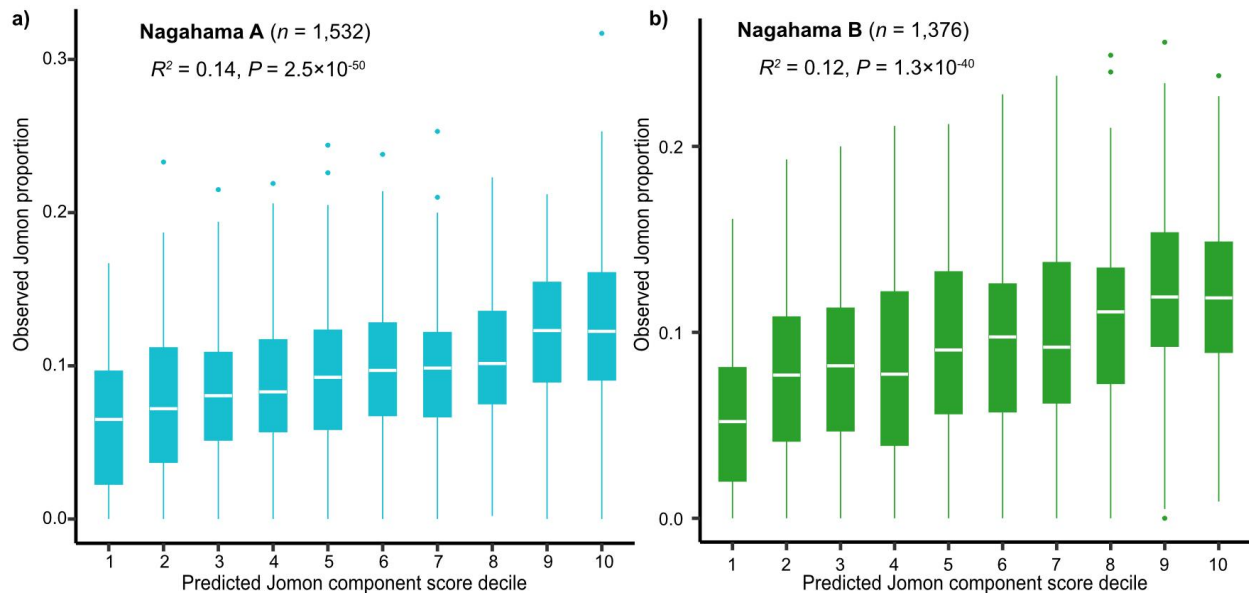
Supplementary Fig. 12. F_{ST} using 132 Jomon-related variants between all pairs of the 1KG EAS populations.

CDX, Chinese Dai; CHB, Han Chinese in Beijing; CHS, Han Chinese South; JPT, Japanese in Tokyo; KHV, Kinh in Ho Chi Minh City.



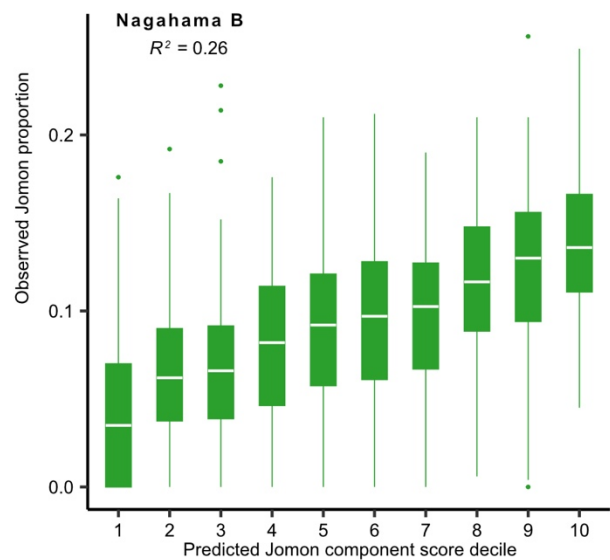
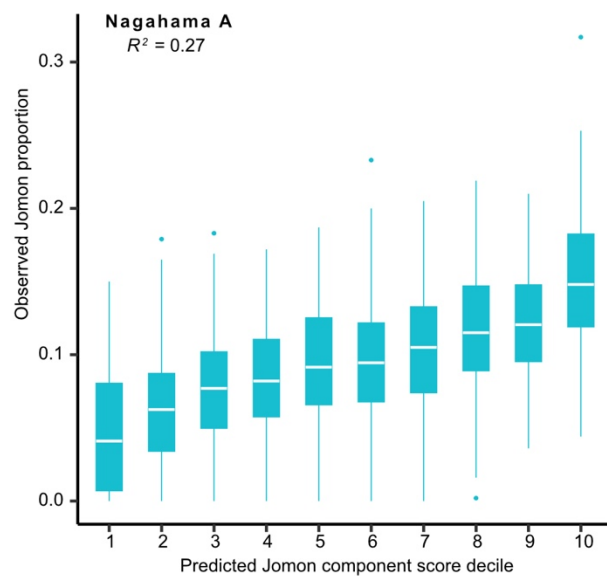
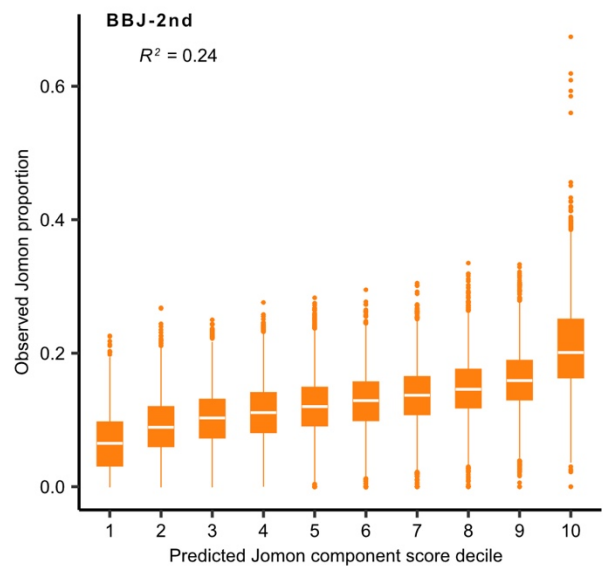
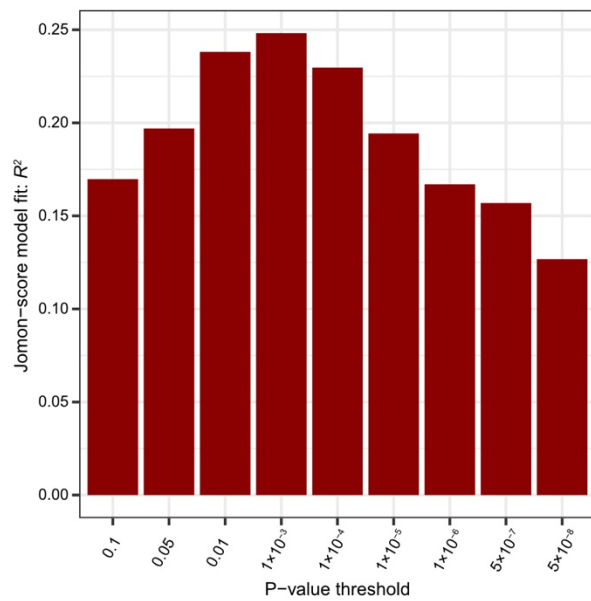
Supplementary Fig. 13. F_{ST} values for 132 Jomon-related variants between JPT and each of the other 1KG EAS populations.

CDX, Chinese Dai; CHB, Han Chinese in Beijing; CHS, Han Chinese South; JPT, Japanese in Tokyo; KHV, Kinh in Ho Chi Minh City.



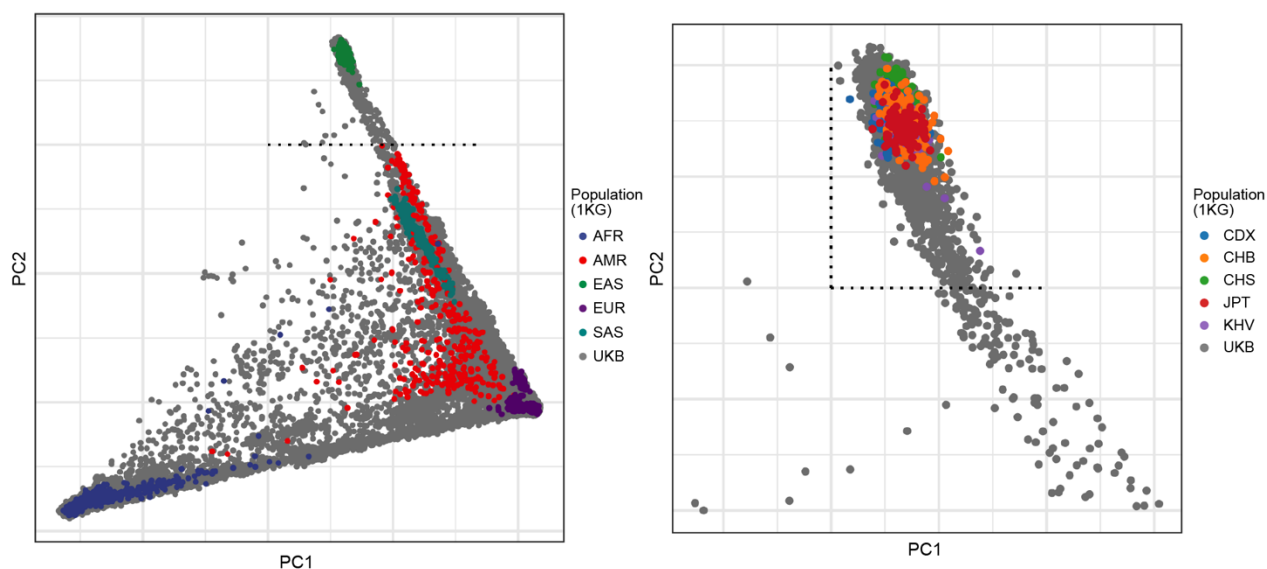
Supplementary Fig. 14. Predictive power of 132 Jomon-related variants using the Nagahama cohort.

The Nagahama cohort comprises two sub-cohorts: **a)** Nagahama A ($n = 1,532$) and **b)** Nagahama B ($n = 1,376$). The Jomon component prediction scores measured from all individuals within a cohort are split into the deciles. The boxplot shows the distribution of observed Jomon proportions in a given decile of the predicted scores. R^2 indicates the squared value of the correlation coefficient between the Jomon component prediction scores and the residuals of Jomon proportions regressed out with 10 PCs. P -values reflect two-sided tests. The boxes denote the interquartile range (IQR) and the median is shown as white horizontal bars; whiskers extend to 1.5 times the IQR; outliers are shown as individual points.



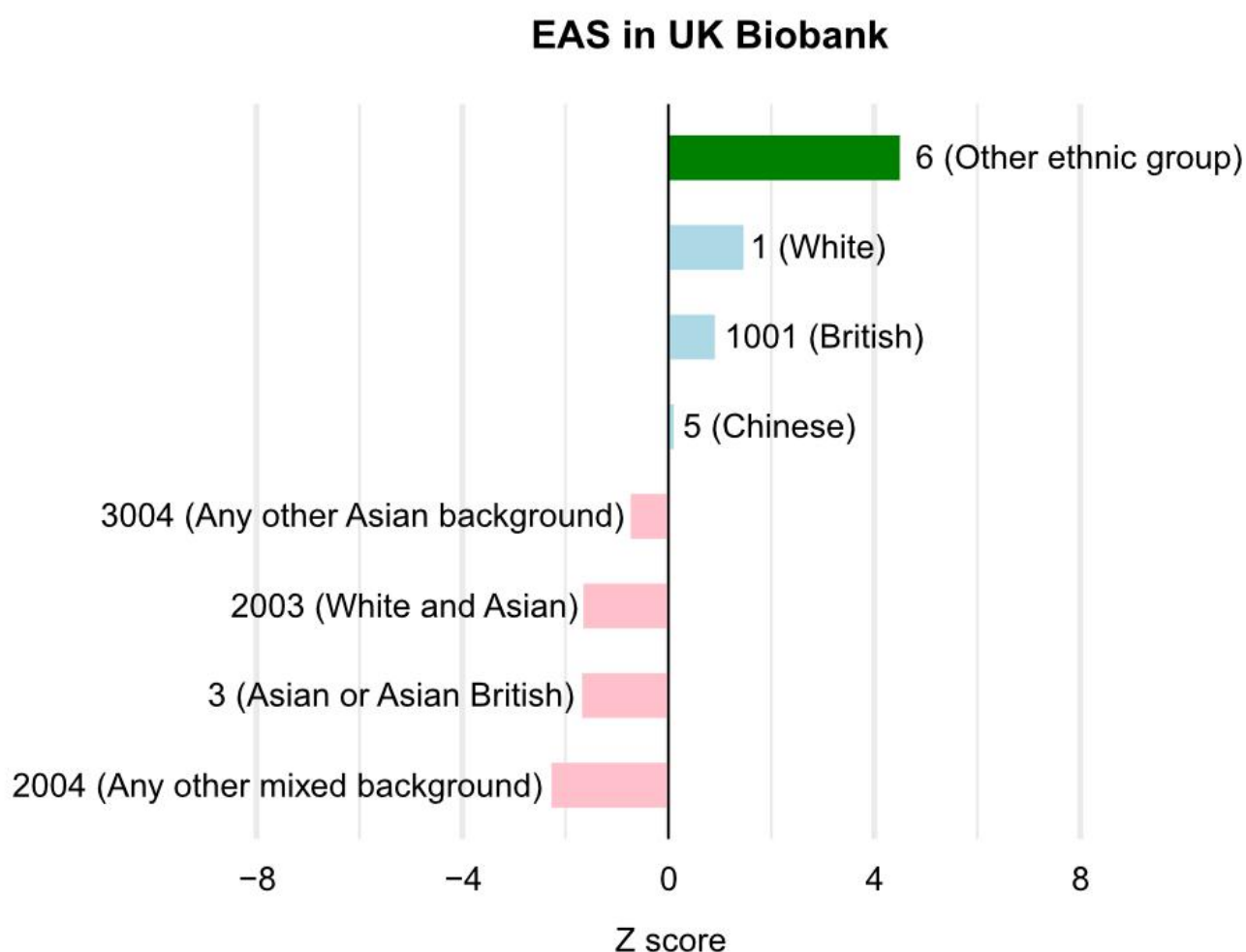
Supplementary Fig. 15. Predictive power of polygenic scores for Jomon ancestry.

Bar plots show the variances in Jomon ancestry explained by different sets of variants selected from GWAS on Jomon proportions using nine different p -value thresholds. A set of variants at a p -value cut-off of 10^{-3} shows the highest variance of 24.3%, which is validated with three independent cohorts: BBJ-2nd, Nagahama A, and Nagahama B. Boxplots show the distribution of observed Jomon proportions in a given decile of the predicted scores. R^2 indicates the squared value of the correlation coefficient between the Jomon component prediction scores and the residuals of Jomon proportions regressed out with 10 PCs. Boxes denote the interquartile range (IQR), with medians shown as white horizontal bars; whiskers extend to 1.5 times the IQR; outliers are shown as individual points.



Supplementary Fig. 16. Scatter plots of principal component analysis for UKB participants ($n = 407,128$) combined with 1KG.

The individuals in EAS subpopulations are visually curated based on PCA ($n = 2,286$). 1KG, 1000 Genomes Projects; AFR, African; AMR, South American; EAS, East Asian; EUR, European; SAS, South East Asian; UKB, UK Biobank; CDX, Chinese Dai; CHB, Han Chinese in Beijing; CHS, Han Chinese South; JPT, Japanese in Tokyo; KHV, Kinh in Ho Chi Minh City.



Supplementary Fig. 17. Genetic affinities of Jomon with populations with different ethnic backgrounds within UKB EAS.

Bar plots represent Z-scores from f_4 -tests with the form of $f_4(\text{Mbuti}, \text{Jomon}; \text{Han}, X)$, where X is a population from UKB EAS. Each bar plot represents a Z-score of a population with a particular ethnic background. Positive or negative Z scores are colored in right blue or pink. If the Z score is greater than 3, the bar is coloured in green. The numbers in the plots indicate the data-coding in the corresponding ethnic groups.