Genomic insights into the cold adaptation of human populations in the Amur River Basin Supplementary Materials

Siqin Chen[1, §], Hao Chen[2, §], Zhaoqing Yang[3,§], Yang Gao[1,4], Kuiting Tao[1], Jiayou Chu[3]*, Shuhua Xu[1*]

[1]State Key Laboratory of Genetic Engineering, Human Phenome Institute, Zhangjiang Fudan International Innovation Center, Center for Evolutionary Biology, School of Life Sciences, Department of Liver Surgery and Transplantation Liver Cancer Institute, Zhongshan Hospital, Fudan University, Shanghai 200032, China

[2]Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China;

[3]Department of Medical Genetics, Institute of Medical Biology, Chinese Academy of Medical Sciences, Kunming 650118, China

[4]Ministry of Education Key Laboratory of Contemporary Anthropology, Fudan University, Shanghai 201203, China;

[§]These authors contributed equally to this work.

[*]Correspondence: xushua@fudan.edu.cn (S.X.), chujy@imbcams.com.cn (J.C.)

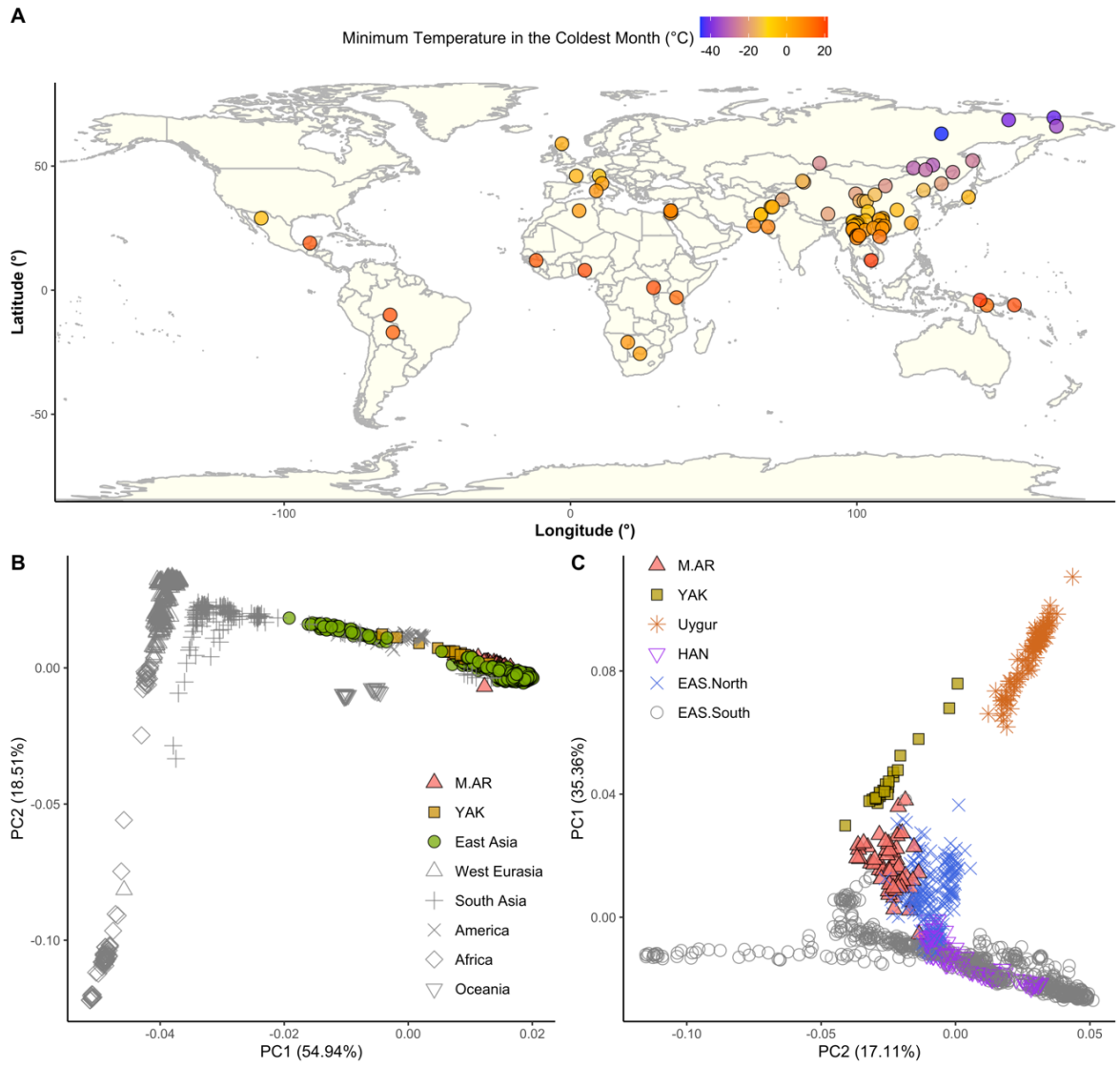

Fig. S1. Principal component analysis (PCA) in the global context.

A: Geographic locations of global populations used for PCA. The color of the dot represents the minimum temperature of the coldest month (°C), ranging from -45.5°C to 21.6°C.

B and C: PCA with 170,943 SNPs of (B) the global populations (1,721 individuals), and (C) East Asians and YAK (1,193 individuals). EAS.North: other northern East Asians; EAS.South: southern East Asians.

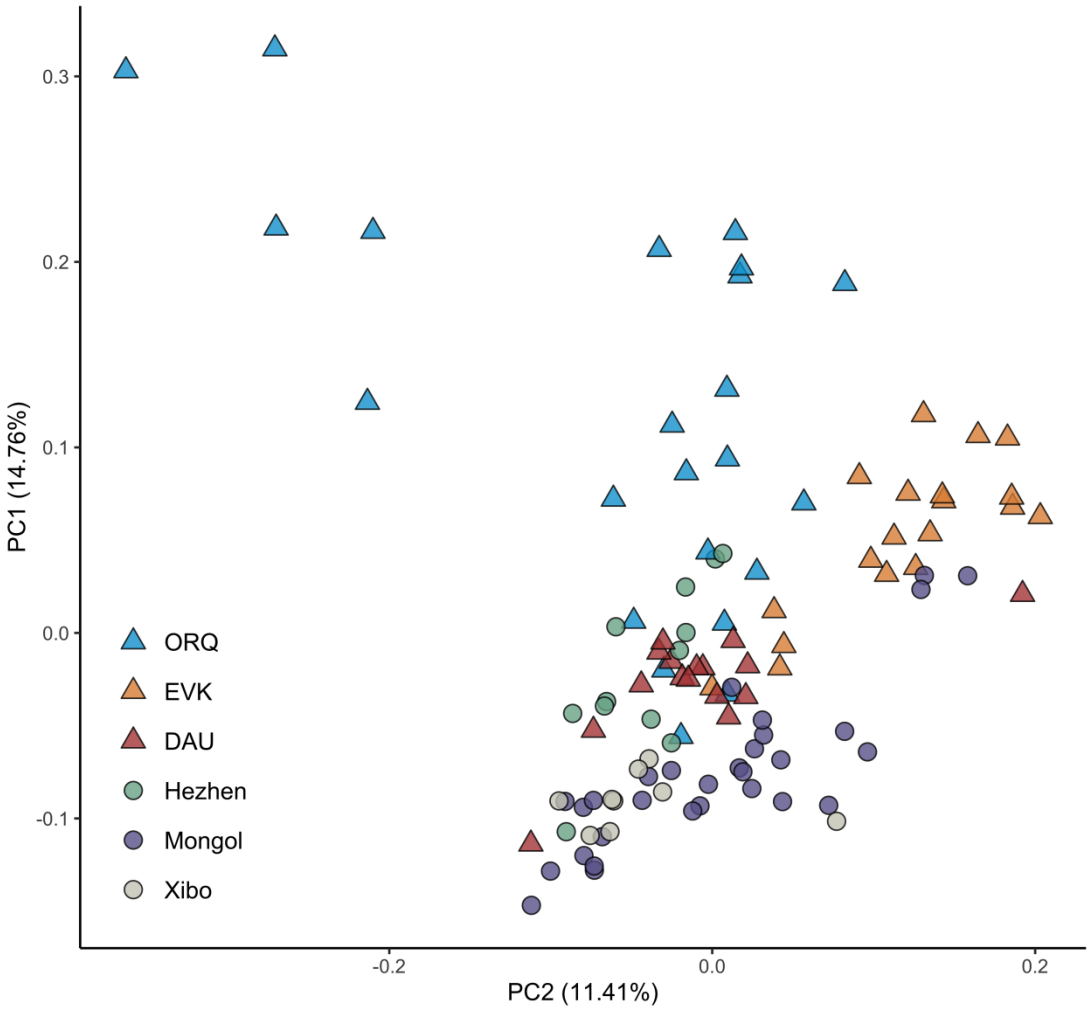

Fig. S2. PCA of M.AR populations and the Mongol, Hezhen, and Xibo populations.

PCA with 170,943 SNPs of three M.AR populations and the Mongol, Hezhen, and Xibo populations (109 individuals).

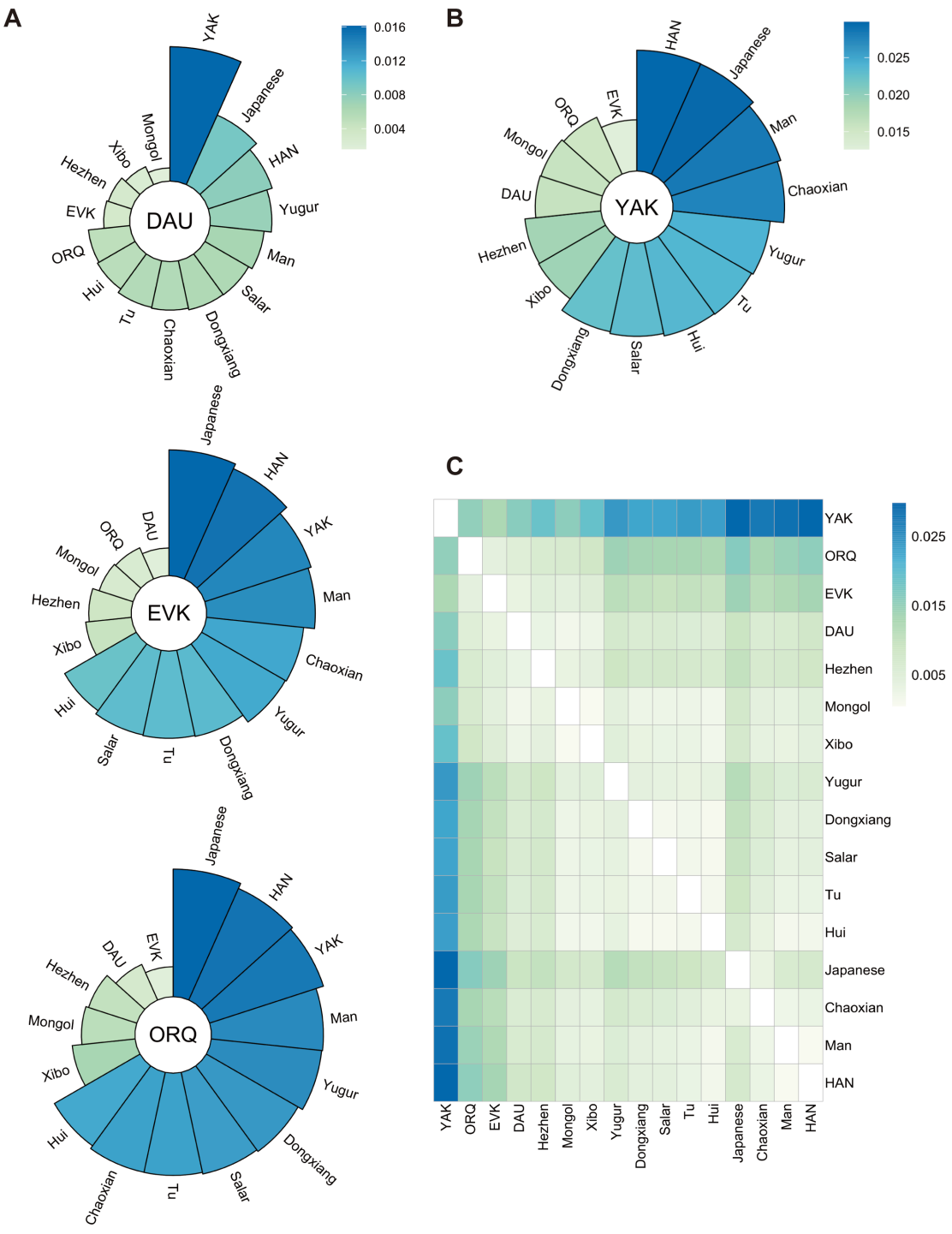

Fig. S3. Population differentiation measured by the unbiased fixation index (F_{ST}).

A: F_{ST} between each M.AR and other populations in northern East Asia and Siberia.

B: F_{ST} between YAK and all northern East Asian populations.

C: Heatmap of pairwise F_{ST} among populations in northern East Asia and Siberia.

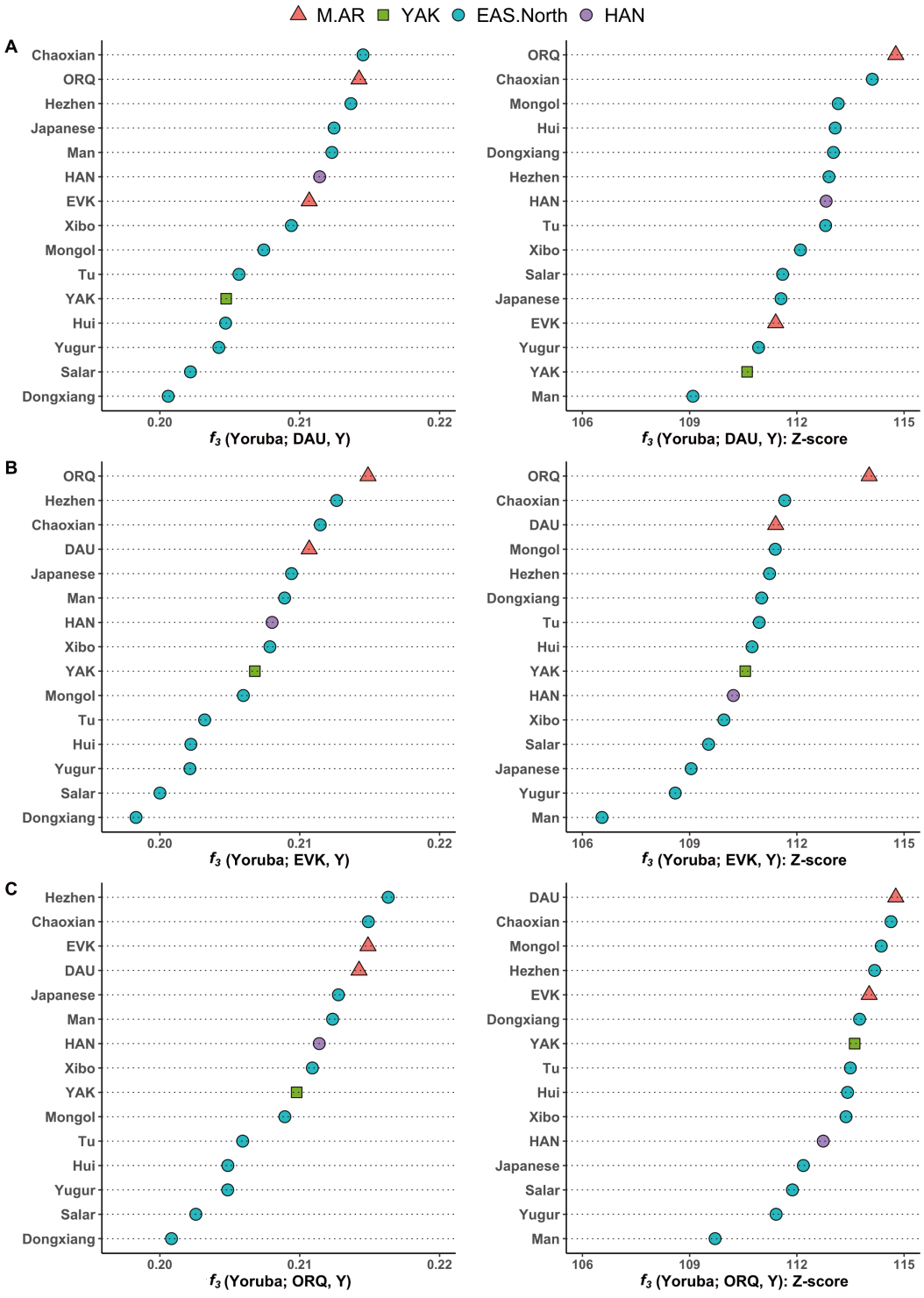

Fig. S4. Population differentiation measured by the outgroup f_3 statistic.

The outgroup f_3 statistic is in the form of f_3(Yoruba; X, Y), assuming Y is the population from other northern East Asian populations and X is (A) DAU, (B) EVK, and (C) ORQ. The higher f_3 value and Z-score indicate a closer genetic distance between two populations.

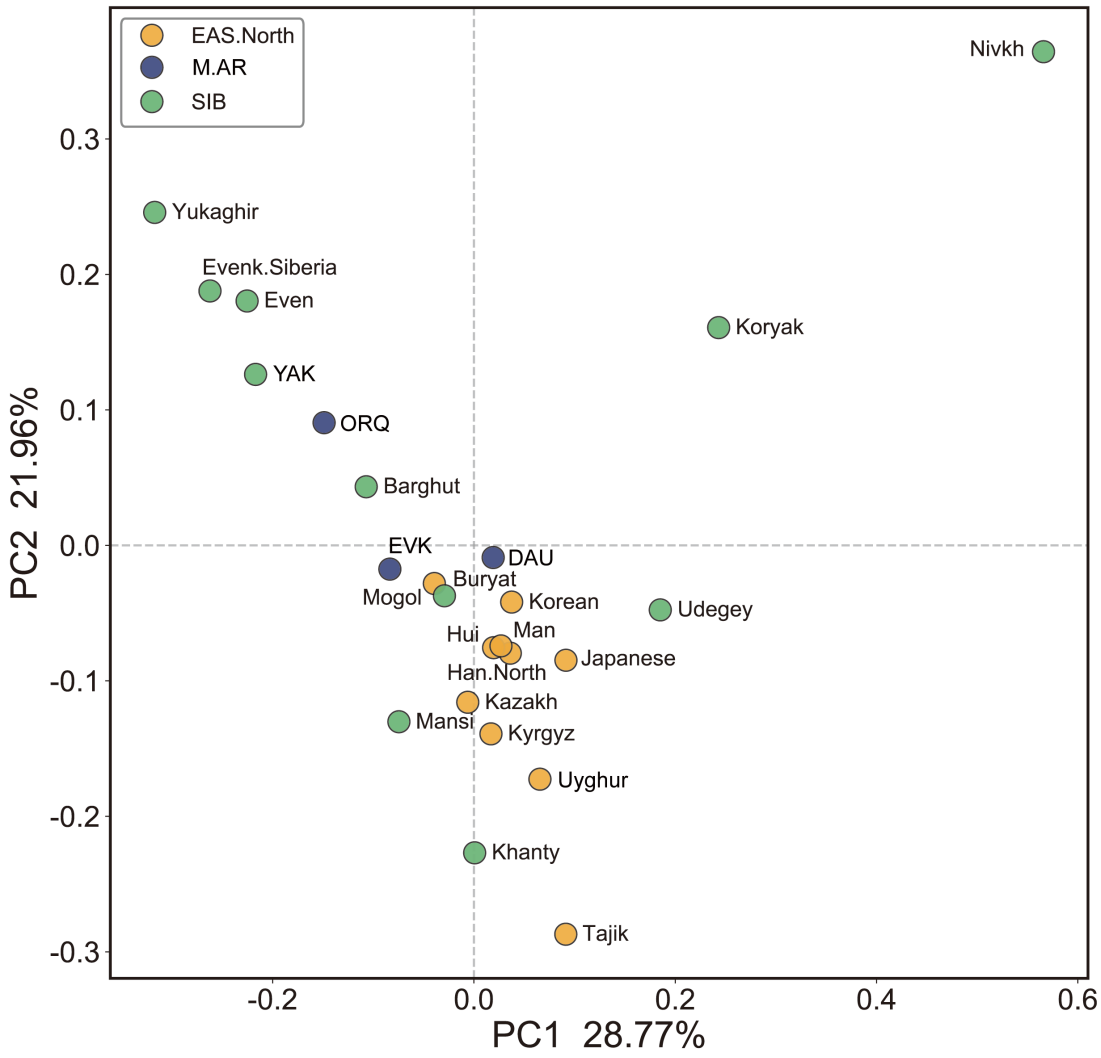

Fig. S5. PCA based on mtDNA haplogroup frequency of M.AR, northern East Asian and Siberian populations.

Genetic structure at the maternal level is measured by mtDNA haplogroup frequency of populations in northern East Asia and Siberia. EAS.North: other northern East Asians; SIB: Siberians.

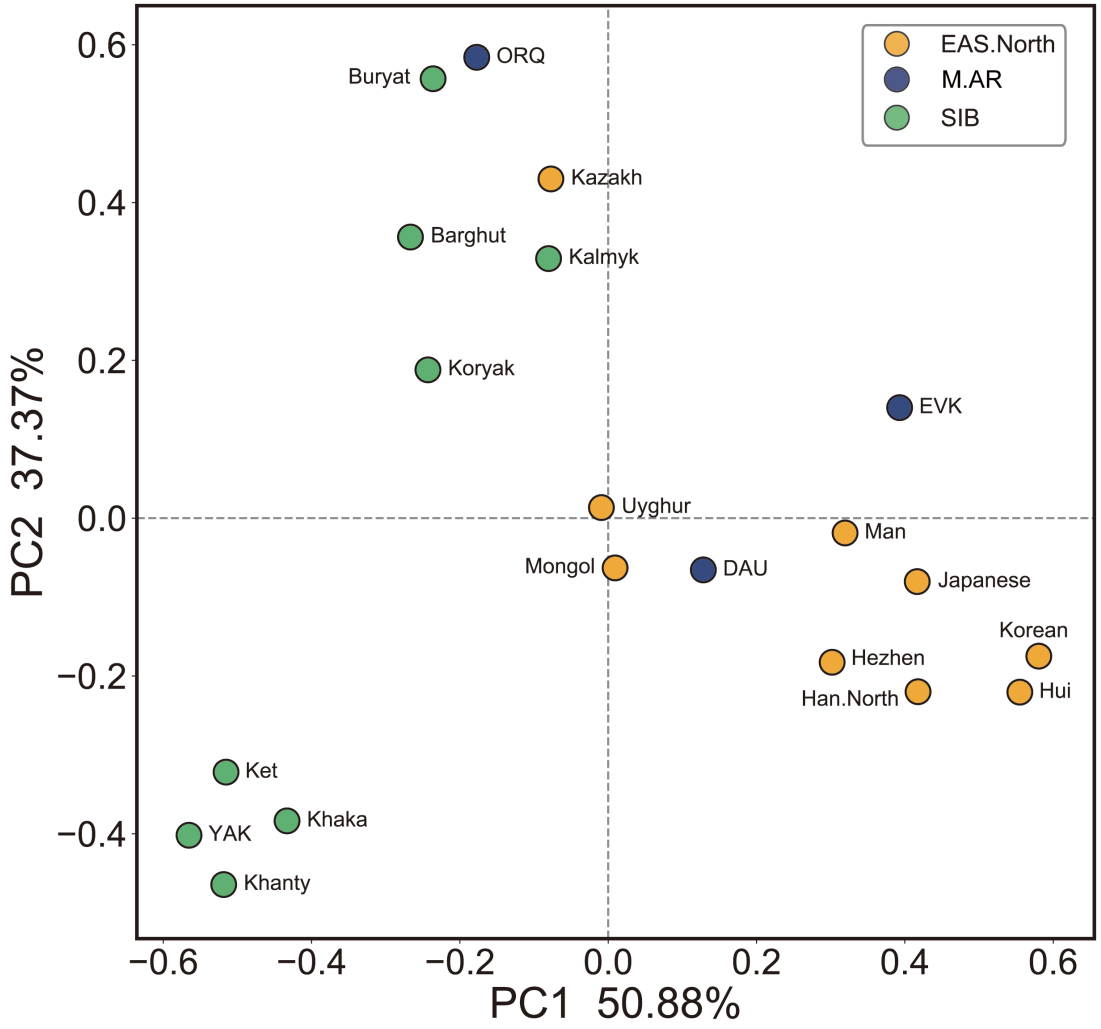

Fig. S6. PCA based on NRY haplogroup frequency of M.AR, northern East Asian and Siberian populations.

Genetic structure at the paternal level is measured by NRY haplogroup frequency of populations in northern East Asia and Siberia. EAS.North: other northern East Asians; SIB: Siberians.

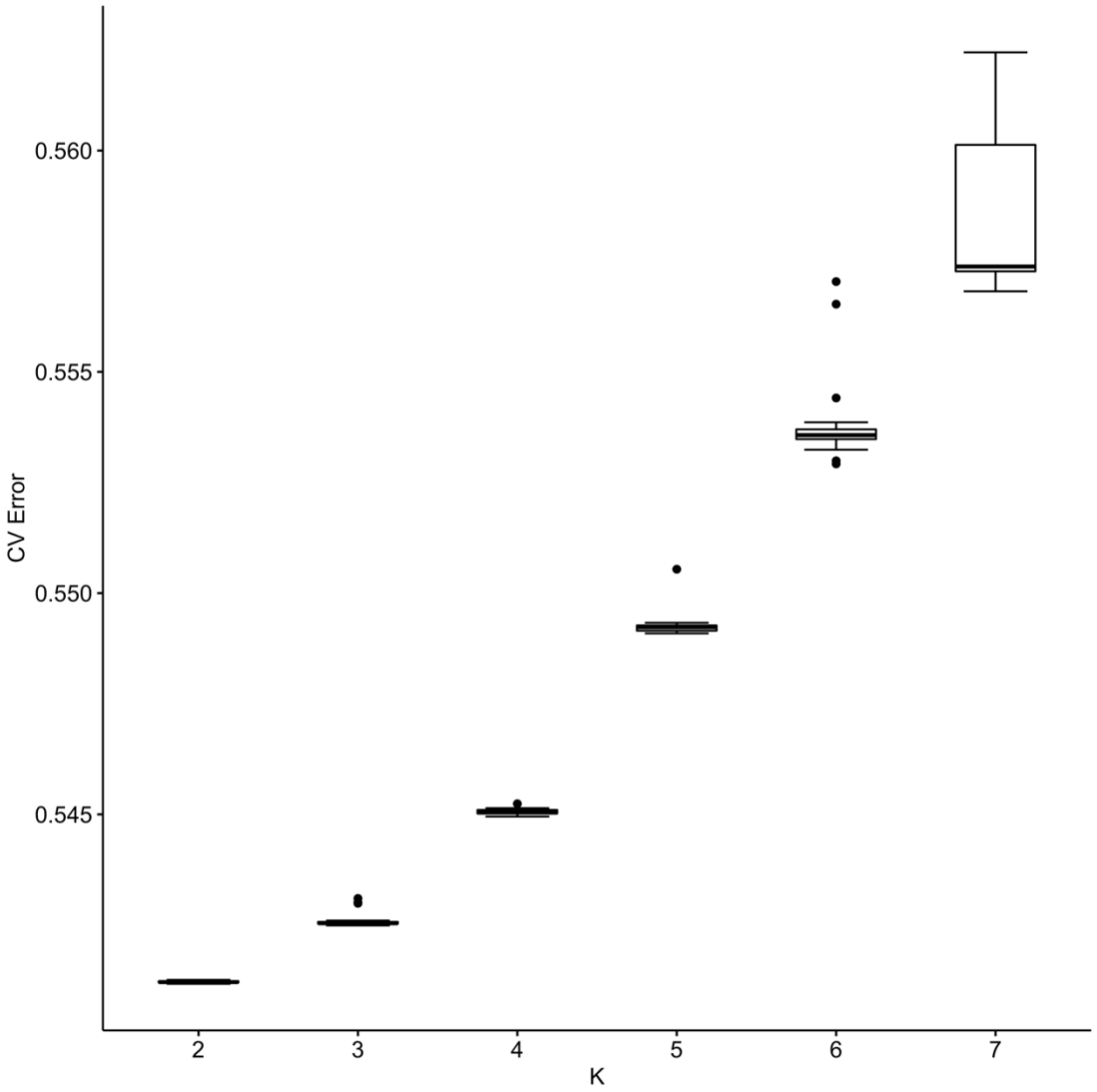

Fig. S7. Cross-validation (CV) error of *ADMIXTURE* analysis.

The CV-error scores for K=2 to 7 over the 20 iterations of *ADMIXTURE* analysis performed in the context of northern East Asia and Siberia. The best K is 2 with the lowest CV error.

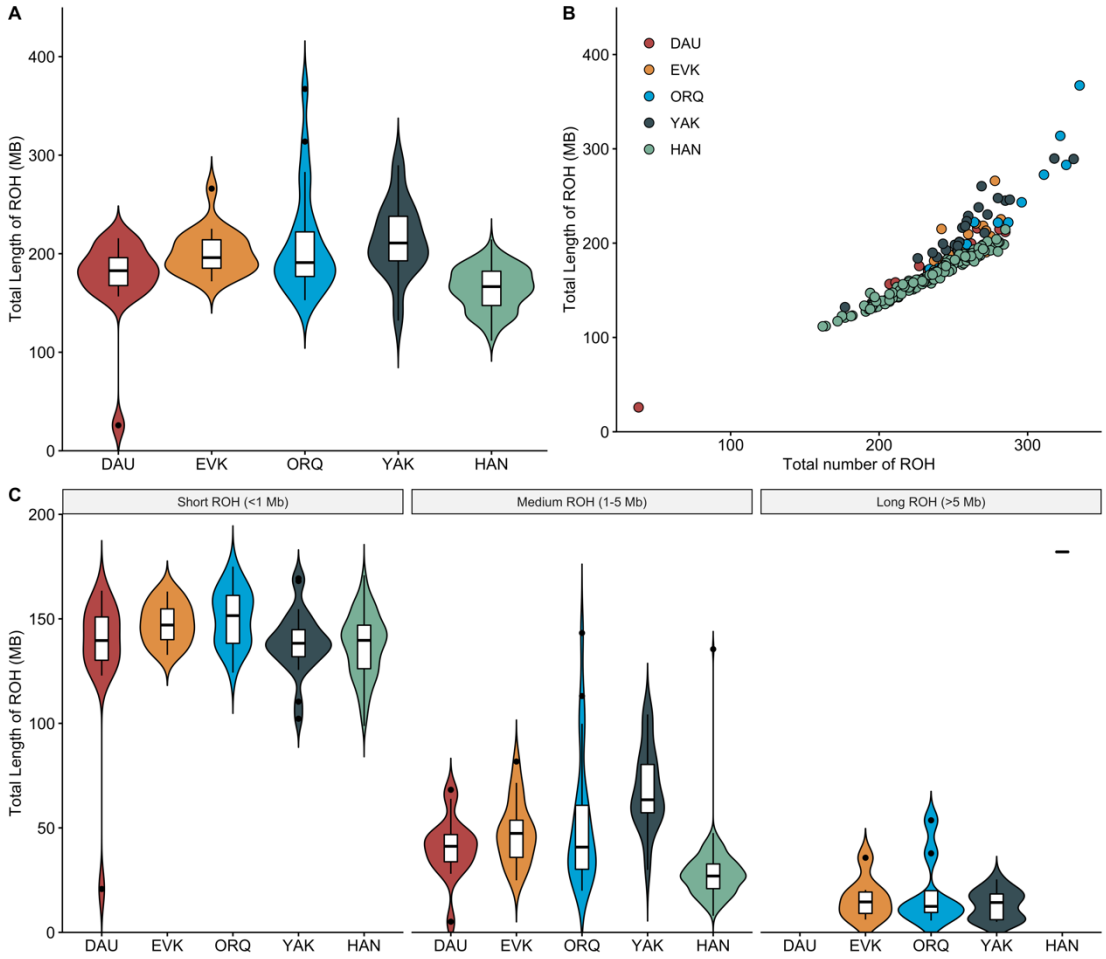

Fig. S8. Runs of homozygosity (ROH) of M.AR populations, YAK, and HAN.

A: The total length of ROH in M.AR populations, YAK, and HAN.

B: Comparison of the total number of ROH (x-axis) and the total length of ROH (y-axis) at the individual level.

C: The total length of short (< 1Mb), medium (1-5 Mb), and long ROH (> 5 Mb) in M.AR populations, YAK, and HAN.

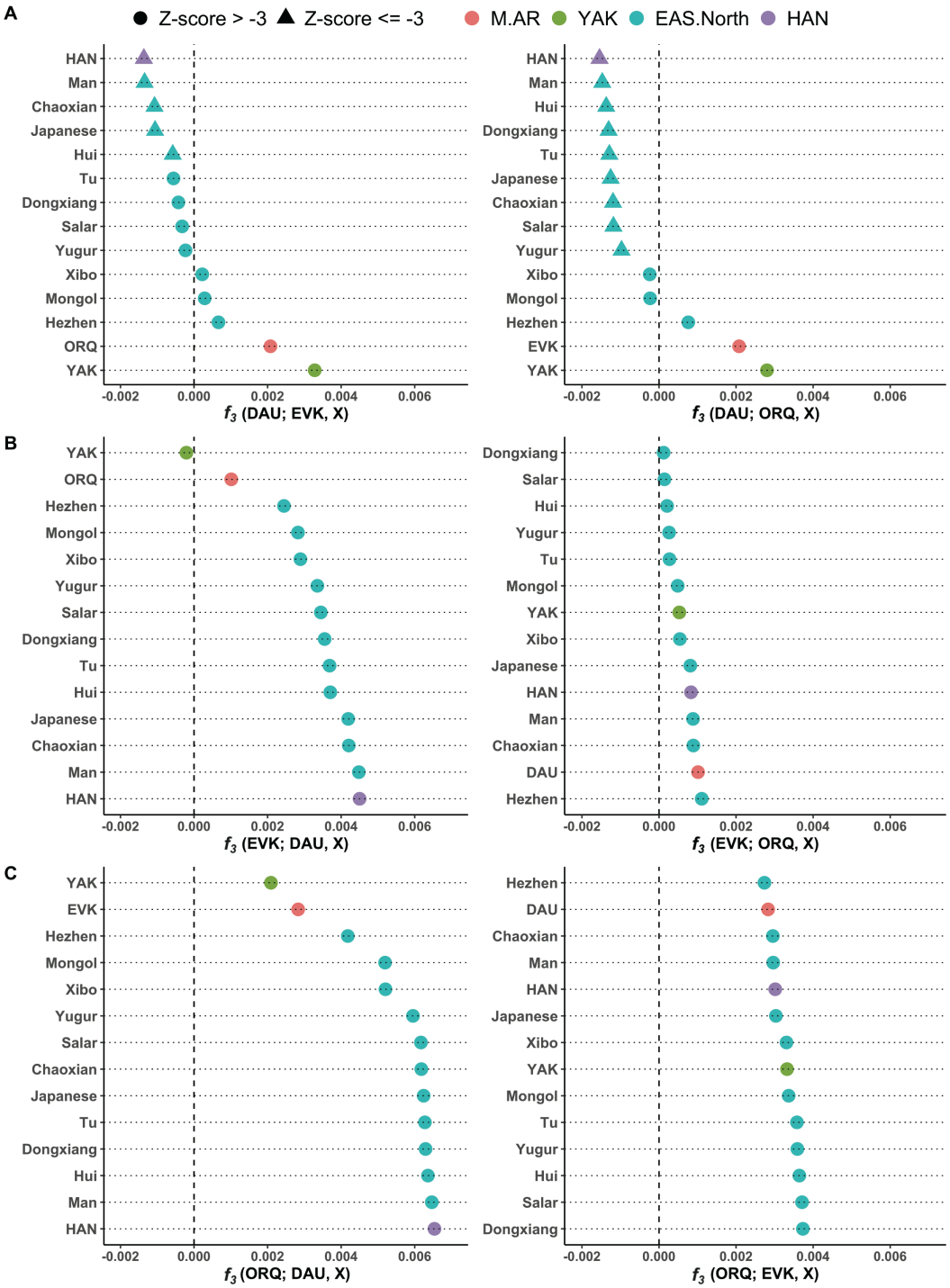

Fig. S9. Potential gene introgression in M.AR estimated by the f_3 statistic.

The f_3 statistic is in the form of f_3(X; Y, Z), where X represents the target M.AR populations, including (A) DAU, (B) EVK, and (C) ORQ, Y represents each of the other two M.AR populations, and Z represents other populations in northern East Asia and Siberia. Significantly negative values of the f_3 statistic ($f_3 < 0$ and Z-score <= -3) indicate that X is a mixture of two populations Y and Z.

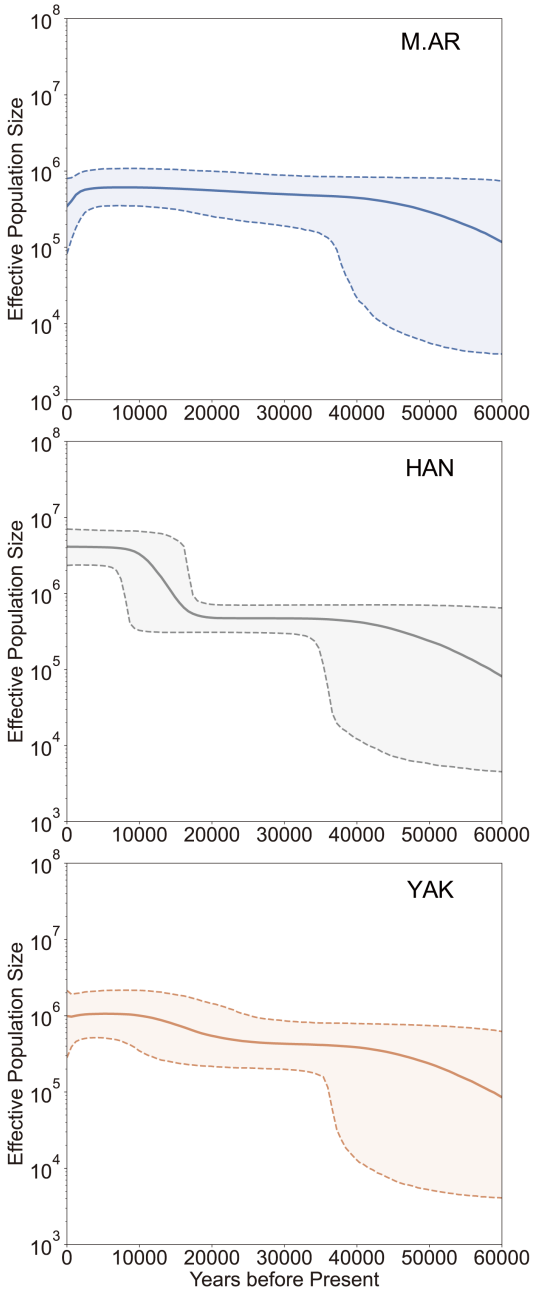

Fig. S10. Bayesian skyline plot of M.AR, YAK, and HAN based on mtDNA .

Bayesian skyline plot based on mtDNA showing effective population size of M.AR, YAK, and HAN. The colored light region indicates the 95% confidence interval (CI) for estimations.

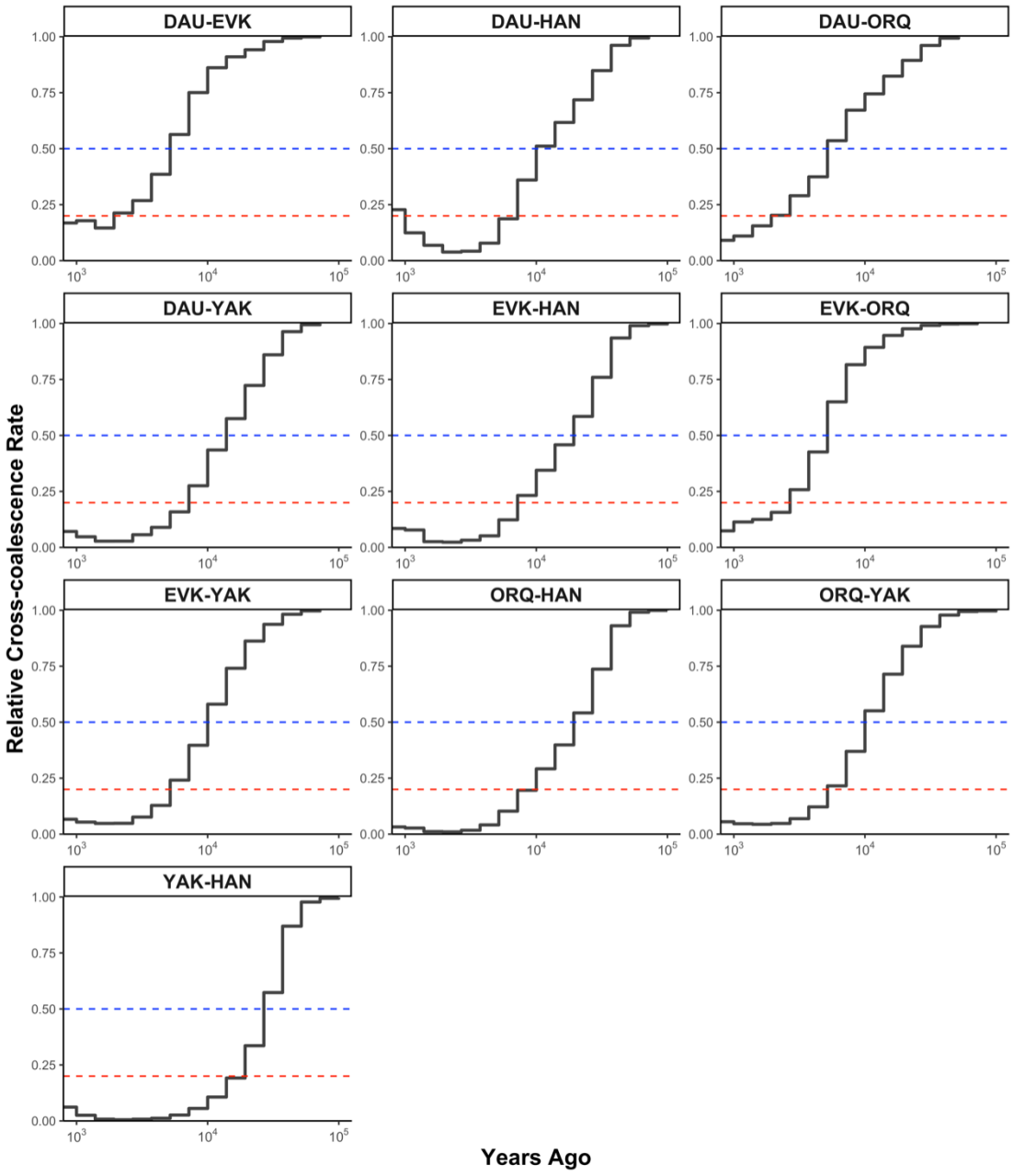

Fig. S11. Estimates of the pairwise divergence of three M.AR populations, YAK, and HAN.

An autosomal mutation rate used in the estimates was 1.25×10^{-8} per base-pair per generation, with 25 years assumed per generation. The mid-point (0.5, blue dashed line) was considered as the start of separation, and the point of 0.2 (red dashed line) was considered the time when the two populations were separated.

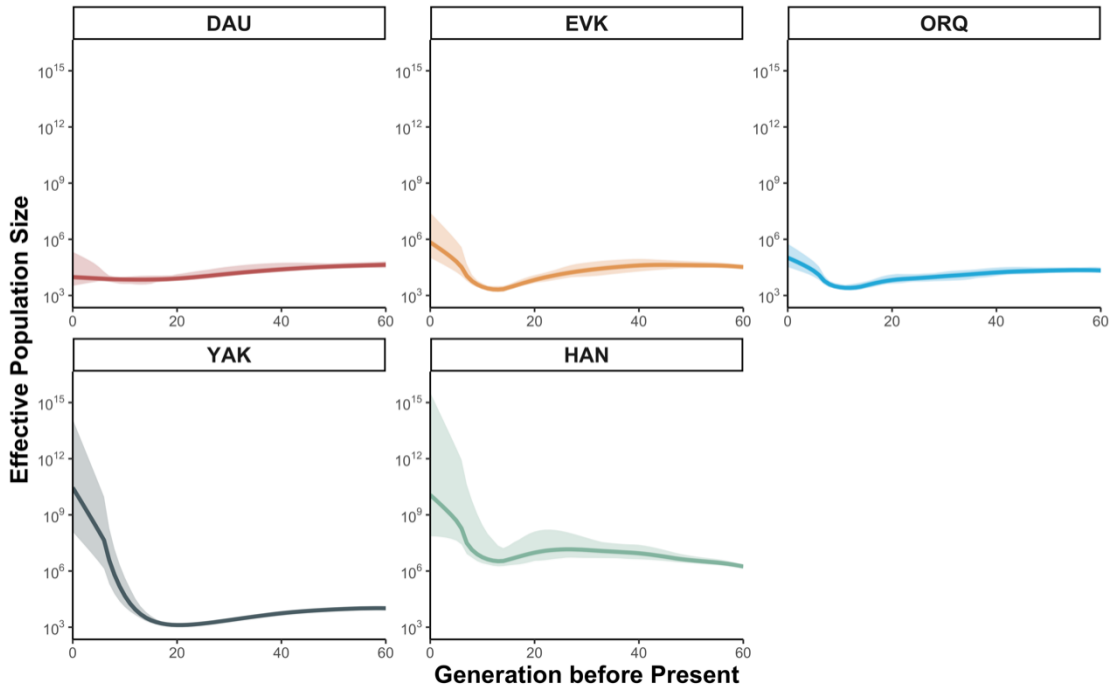

Fig. S12. Recent population demography of three M.AR populations, YAK, and HAN. Estimates of recent effective population size based on IBD sharing using *IBDNe*. The recent effective population sizes within 60 generations of YAK and HAN were used for comparison with M.AR populations. The colored light region indicates the 95% CI.

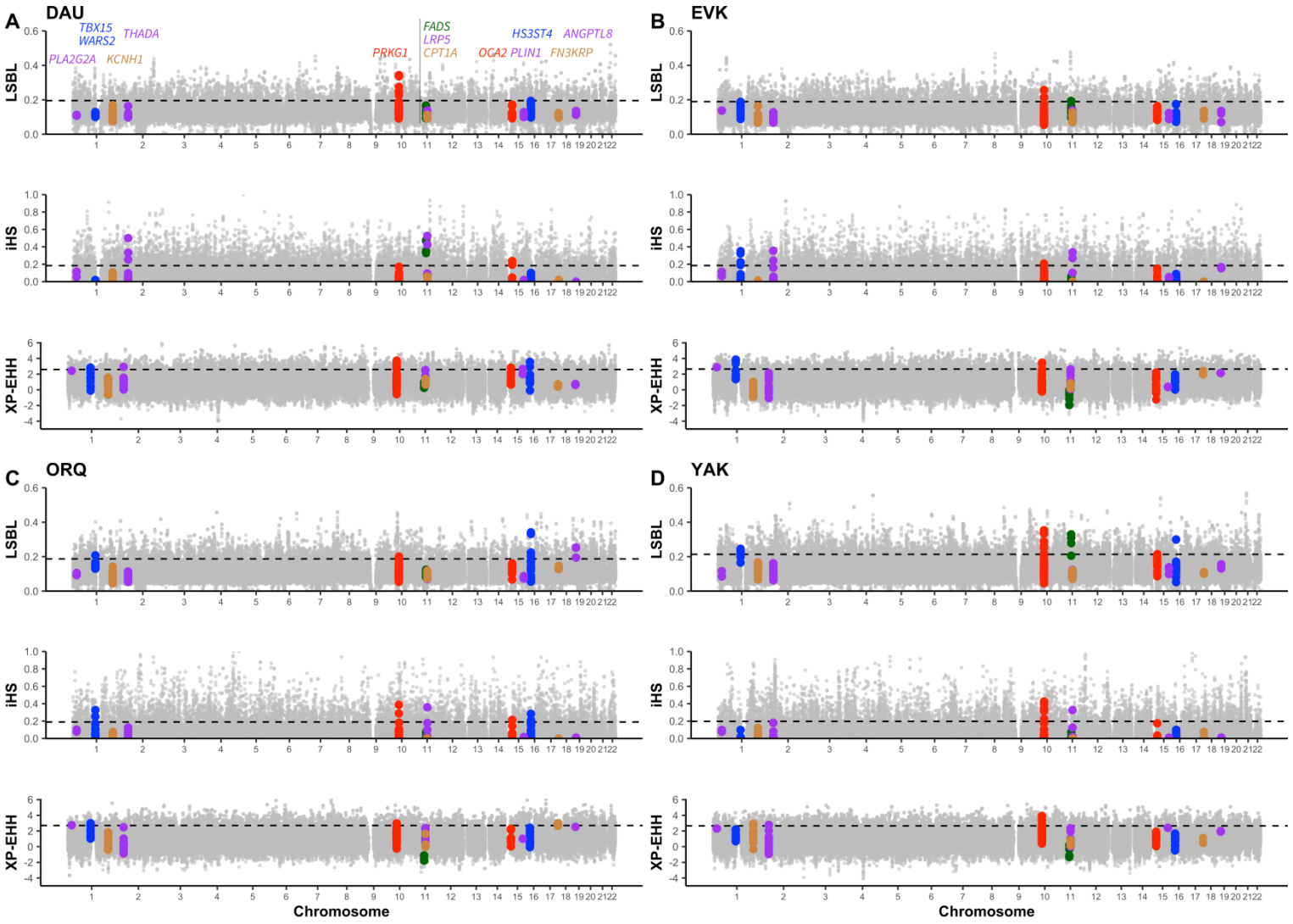

Fig. S13. Selection tests on reported cold adaptation related (CAR) genes using three statistical methods (LSBL, iHS, and XP-EHH).

For each M.AR population and YAK, three Manhattan plots show the LSBL, iHS, and XP-EHH values across the whole genome from top to bottom. LSBL was calculated using HAN and the French population as reference populations. XP-EHH was calculated using HAN as the reference population. Each point indicates one 50-kb window. The black dashed line indicates the significance threshold (empirical *P* value = 0.05). The colored points highlight the windows that overlap with reported CAR genes.

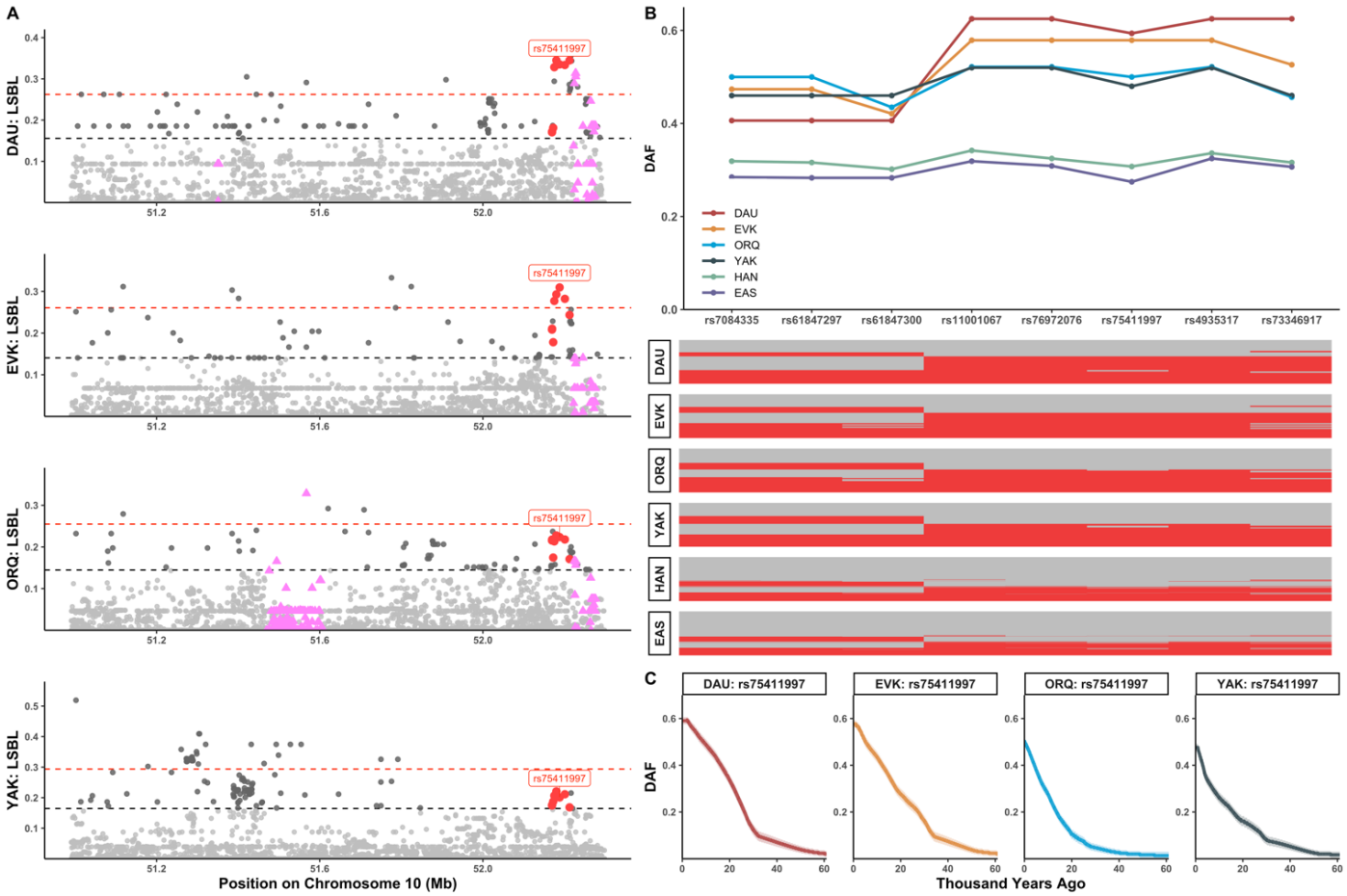

Fig. S14. LSBL local distribution and haplotype pattern within *PRKG1*.

A: Local distribution of LSBL values within *PRKG1* in each M.AR population and YAK. The black and red dashed lines indicate the top 1% and 0.1% threshold of the empirical distribution of LSBL, respectively. Each point represents one SNP, and each purple triangle represents one archaic ancestry informative marker (aAIM). Red dots represent eight highly differentiated SNPs in both M.AR and YAK, of which rs75411997 showed the strongest selection signal evidenced by *RELATE*.

B: The haplotype consisted of eight highly differentiated SNPs in both M.AR and YAK. The upper plot shows the DAF value of each variant in each population. The lower plot shows the haplotype distribution in each population, and red and grey represent the derived and ancestral alleles, respectively. EAS: East Asians except M.AR and HAN.

C: The allele frequency trajectories across time in each M.AR population and YAK. According to the *RELATE* results, rs75411997 showed the strongest selection signal in each population, so we used it to tag this haplotype. The solid line indicates the frequency trajectory of rs75411997, as inferred by *CLUES*, and the colored region indicates the 95% CI.

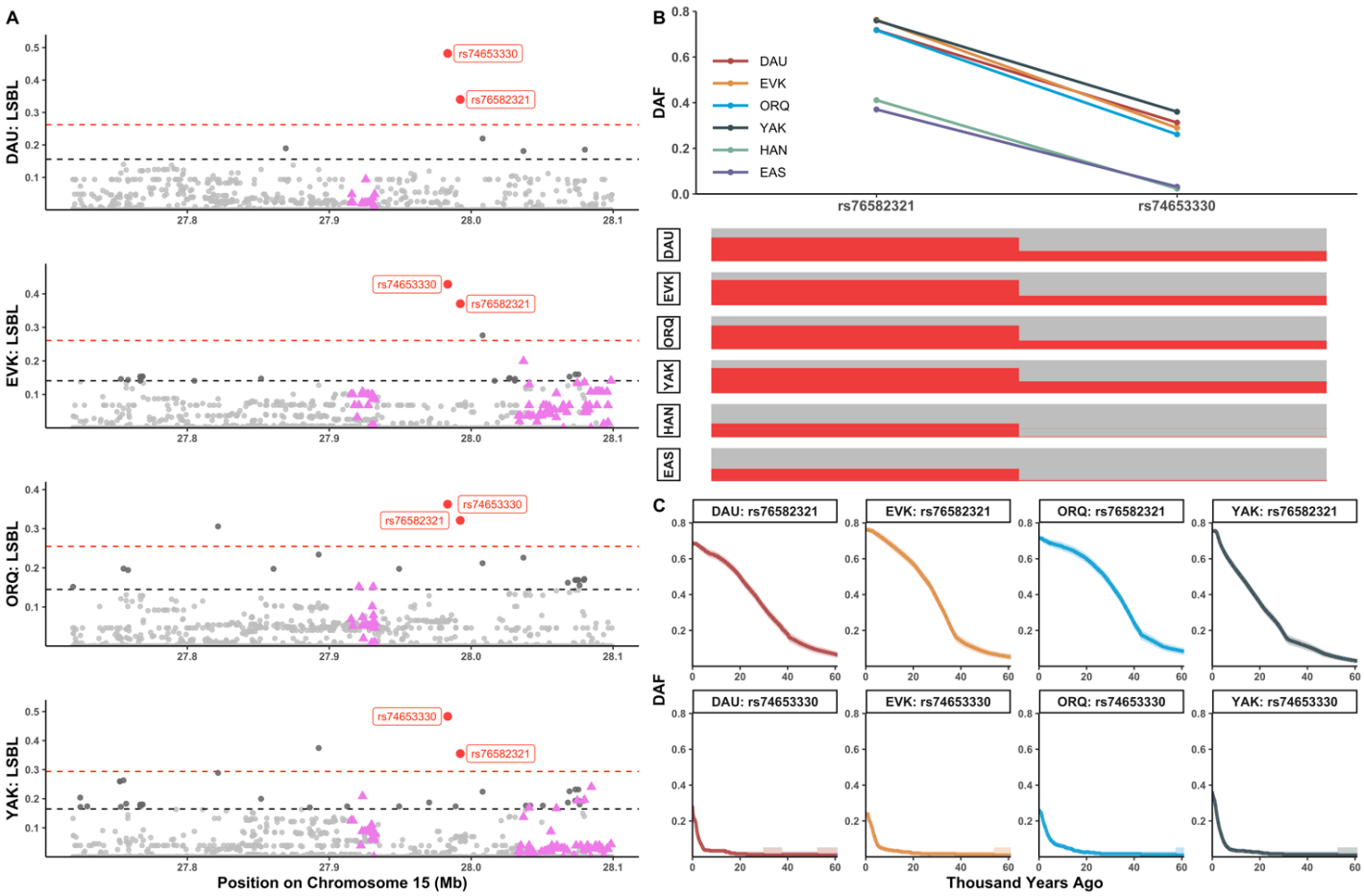

Fig. S15. LSBL local distribution and haplotype pattern within *OCA2*.

A: Local distribution of LSBL values within *OCA2* in each M.AR population and YAK. The black and red dashed lines indicate the top 1% and 0.1% threshold of the empirical distribution of LSBL, respectively. Each point represents one SNP, and each purple triangle represents one aAIM. Red dots represent two extremely highly differentiated SNPs (rs74653330 and rs76582321) in both M.AR and YAK.

B: The haplotype consisted of rs74653330 and rs76582321 in both M.AR and YAK. The upper plot shows the DAF value of each variant in each population. The lower plot shows the haplotype distribution in each population, and red and grey represent the derived and ancestral alleles, respectively. EAS: East Asians except M.AR and HAN.

C: The allele frequency trajectories across time in each M.AR population and YAK. The solid line indicates the frequency trajectory of each variant, as inferred by *CLUES*, and the colored region indicates the 95% CI.

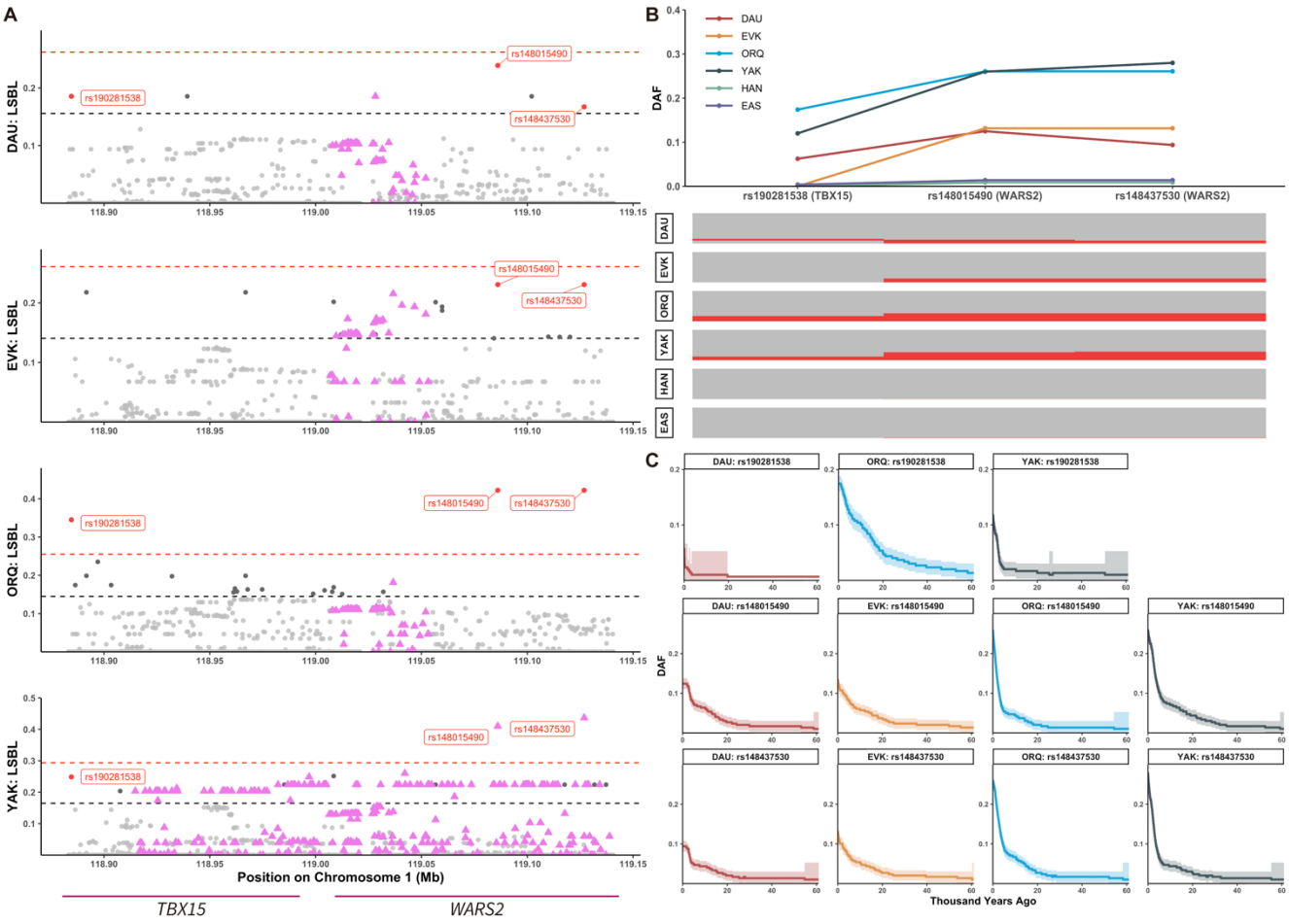

Fig. S16. LSBL local distribution and haplotype pattern within *TBX15* and *WARS2*.

A: Local distribution of LSBL values within *TBX15* and *WARS2* in each M.AR population and YAK. The black and red dashed lines indicate the top 1% and 0.1% threshold of the empirical distribution of LSBL, respectively. Each point represents one SNP, and each purple triangle represents one aAIM. Red dots represent three SNPs (rs190281538, rs148015490, and rs148437530) with the highest LSBL values in both M.AR and YAK. In particular, the derived allele of rs190281538 is absent in EVK, and rs148015490 and rs148437530 are aAIMs in YAK.

B: The haplotype consisted of rs190281538, rs148015490, and rs148437530 in both M.AR and YAK. The upper plot shows the DAF value of each variant in each population. The lower plot shows the haplotype distribution in each population, and red and grey represent the derived and ancestral alleles, respectively. EAS: East Asians except M.AR and HAN.

C: The allele frequency trajectories across time in each M.AR population and YAK. The solid line indicates the frequency trajectory of each variant, as inferred by *CLUES*, and the colored region indicates the 95% CI.

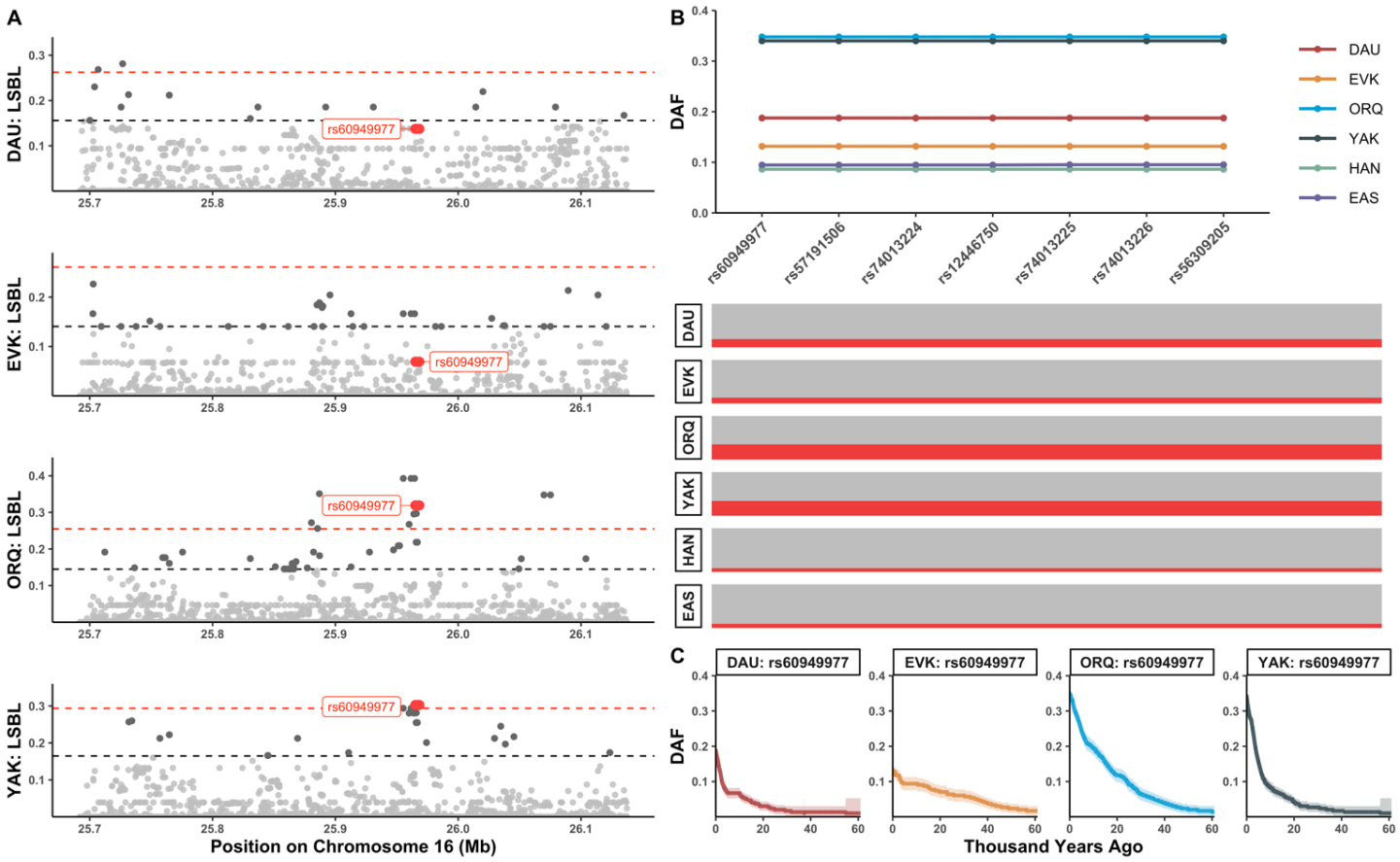

Fig. S17. LSBL local distribution and haplotype pattern within *HS3ST4*.

A: Local distribution of LSBL values within *HS3ST4* in each M.AR population and YAK. The black and red dashed lines indicate the top 1% and 0.1% threshold of the empirical distribution of LSBL, respectively. Each point represents one SNP, and red dots represent seven extremely highly differentiated SNPs in ORQ and YAK, of which rs60949977 showed the strongest selection signal evidenced by *RELATE*.

B: The haplotype consisted of seven extremely highly differentiated SNPs in ORQ and YAK. The upper plot shows the DAF value of each variant in each population. The lower plot shows the haplotype distribution in each population, and red and grey represent the derived and ancestral alleles, respectively. EAS: East Asians except M.AR and HAN.

C: The allele frequency trajectories across time in each M.AR population and YAK. According to the *RELATE* results, rs60949977 showed the strongest selection signal in each population, so we used it to tag this haplotype. The solid line indicates the frequency trajectory of rs60949977, as inferred by *CLUES*, and the colored region indicates the 95% CI.

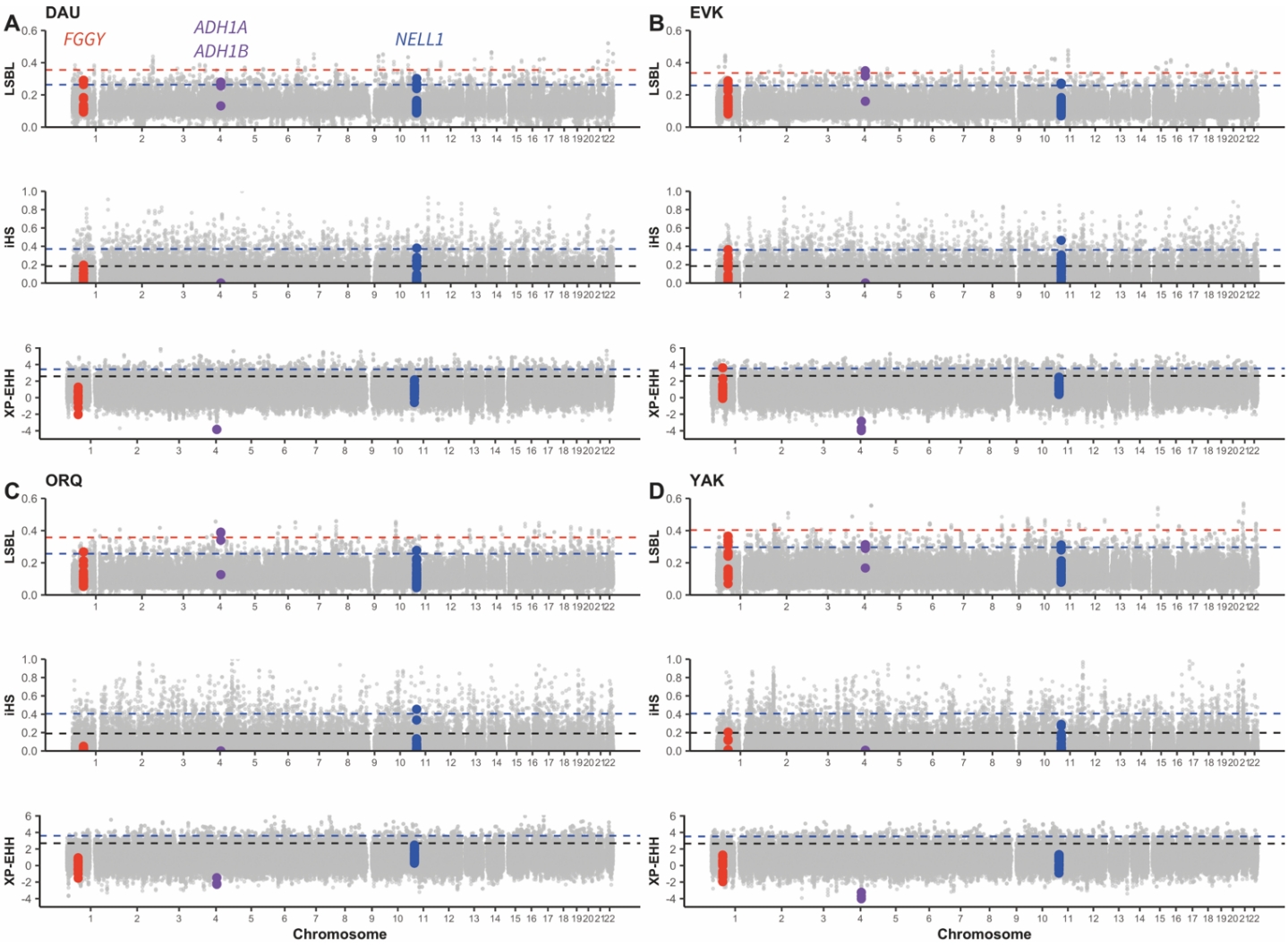

Fig. S18. Newly identified selection signals related to cold adaptation in M.AR populations.

For each M.AR population and YAK, LSBL was calculated using HAN and the French population as reference populations, and XP-EHH was calculated using HAN as the reference population. Each point indicates one 50-kb window. The black dashed line indicates the significance threshold (empirical *P* value = 0.05), and the blue and red dashed lines indicate the top 1% and 0.1% threshold of the empirical distribution of each statistic, respectively. The red, purple, and blue points represent the windows that overlap with *FGGY*, *ADH1A-ADH1B*, and *NELL1*, respectively.

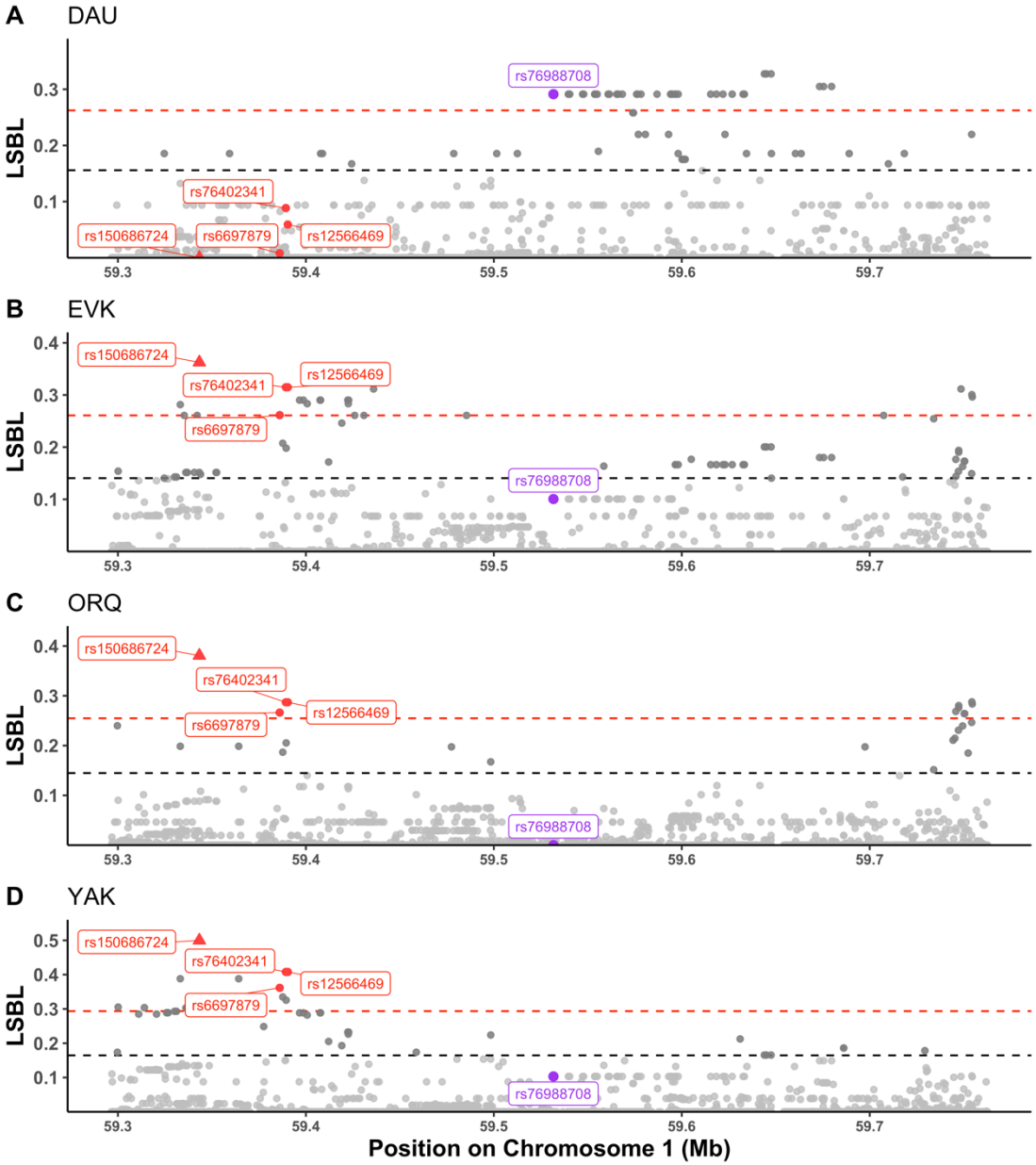

Fig. S19. LSBL local distribution within *FGGY* in M.AR populations and YAK.

For each M.AR population and YAK, LSBL was calculated using HAN and the French population as reference populations. The black and red dashed lines indicate the top 1% and 0.1% threshold of the empirical distribution of LSBL, respectively. Each point represents one SNP, and red dots represent four extremely highly differentiated SNPs in EVK (B), ORQ (C), and YAK (D), of which rs150686724 (the red triangle) showed the highest LSBL values and the strongest selection signals as evidenced by *RELATE*. However, in DAU (A), rs76988708 (the purple dot) showed the strongest selection signal.

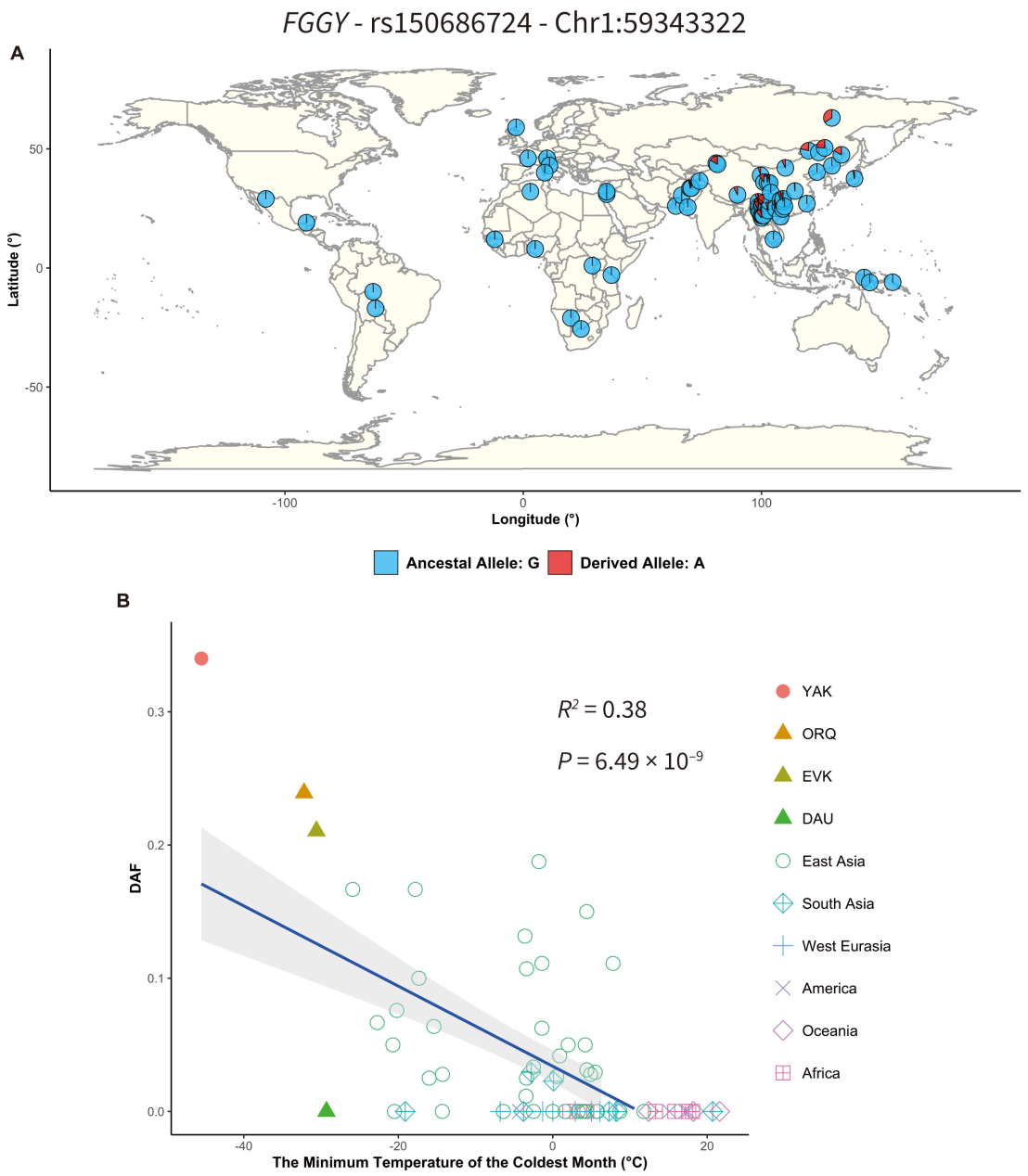

Fig. S20. Allele frequency of *FGGY*-rs150686724 in global populations and its correlation with temperature.

A: Distribution of the allele frequency of rs150686724 in global populations.

B: Correlation between the DAF of rs150686724 and the minimum temperature of the coldest month across all sampled populations.

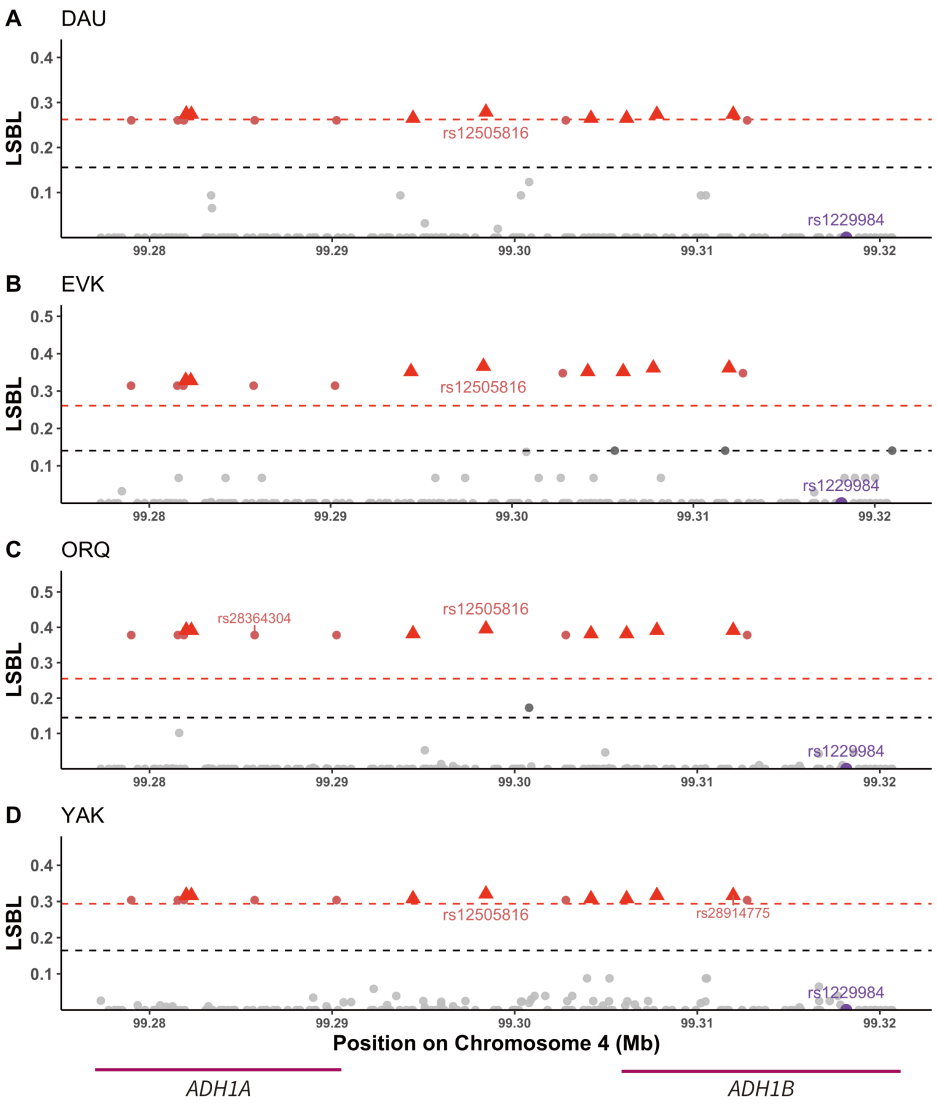

Fig. S21. LSBL local distribution within *ADH1A-ADH1B* in M.AR populations and YAK.

For each M.AR population and YAK, LSBL was calculated using HAN and the French population as reference populations. The black and red dashed lines indicate the top 1% and 0.1% threshold of the empirical distribution of LSBL, respectively. Each point represents one SNP. Red triangles represent nine extremely highly differentiated SNPs in both M.AR and YAK, of which rs12505816, located in the intergenic region, showed the highest LSBL values. The strongest selection signals evidenced by *RELATE* were revealed at rs12505816 in DAU and EVK, rs28364304 in ORQ, and rs28914775 in YAK, respectively. In addition, red dots represent the other seven extremely highly differentiated SNPs in EVK (B), ORQ (C), and YAK (D), and highly differentiated SNPs in DAU (A). The purple dot represents rs1229984.

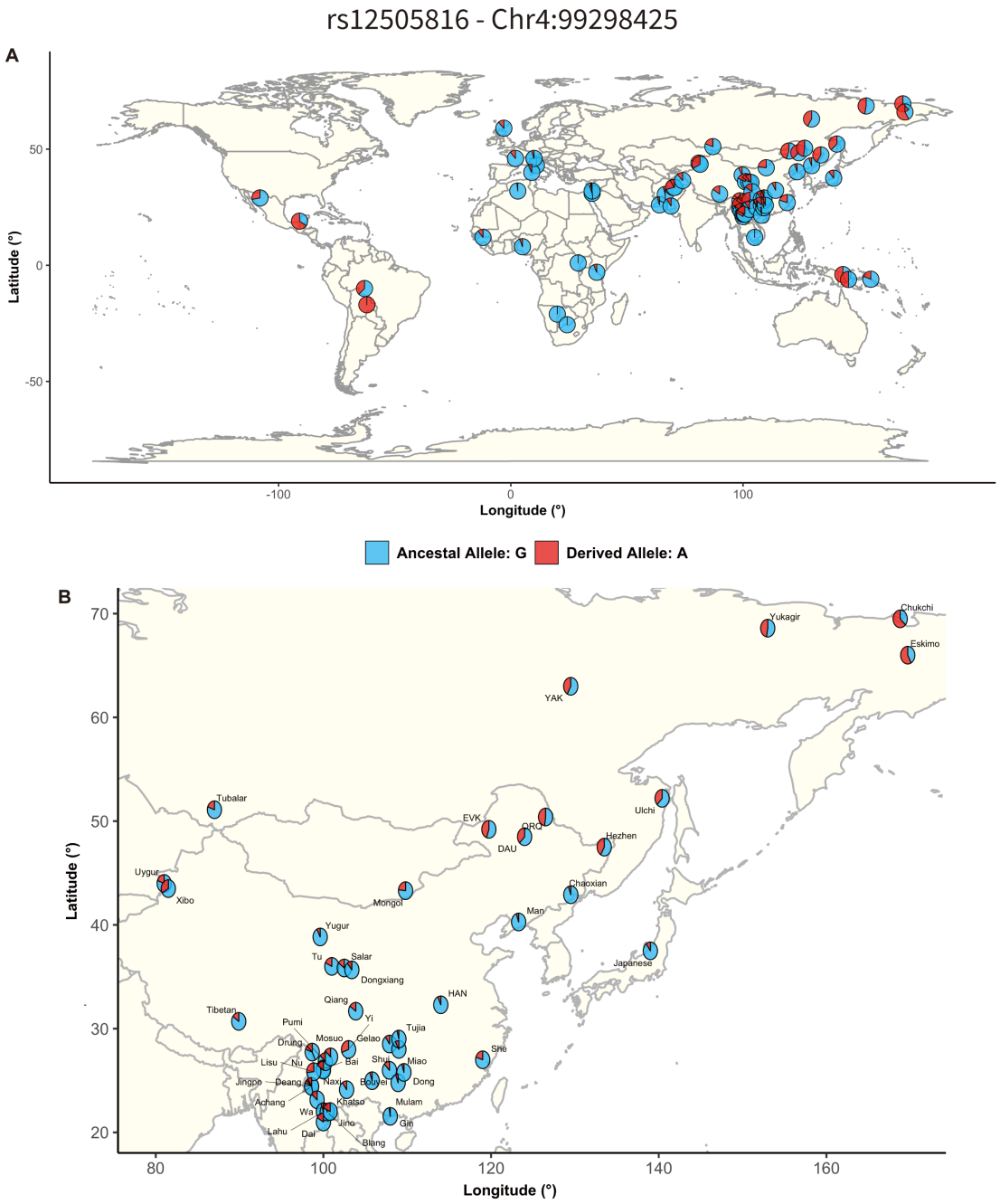

Fig. S22. Allele frequency of rs12505816 in global populations.

Allele frequency distribution of rs12505816 in the context of (A) global populations and (B) populations in East Asia and Siberia.

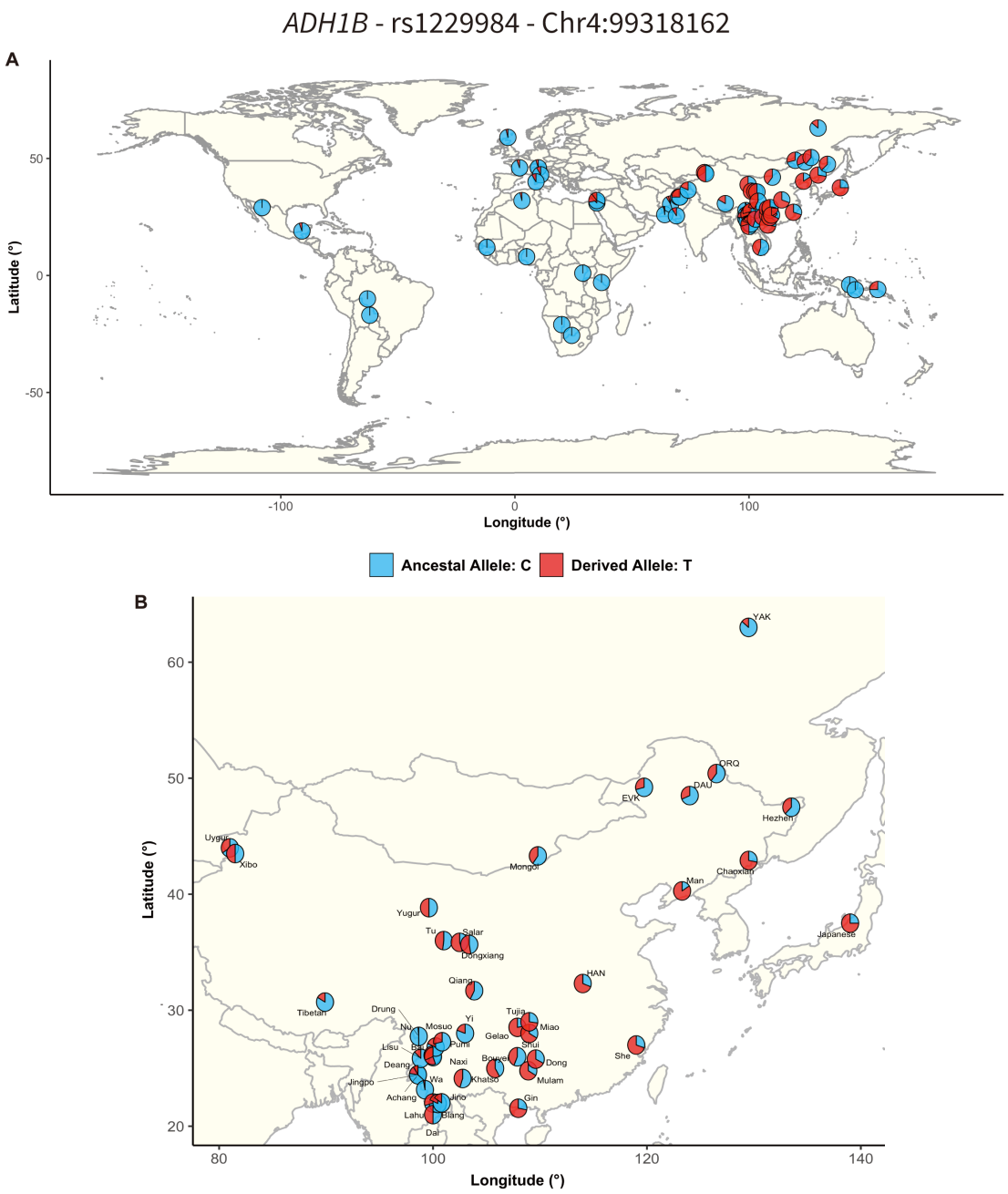

Fig. S23. Allele frequency of *ADH1B*-rs1229984 in global populations.

Allele frequency distribution of rs1229984 in the context of (A) global populations and (B) populations in East Asia and Siberia.

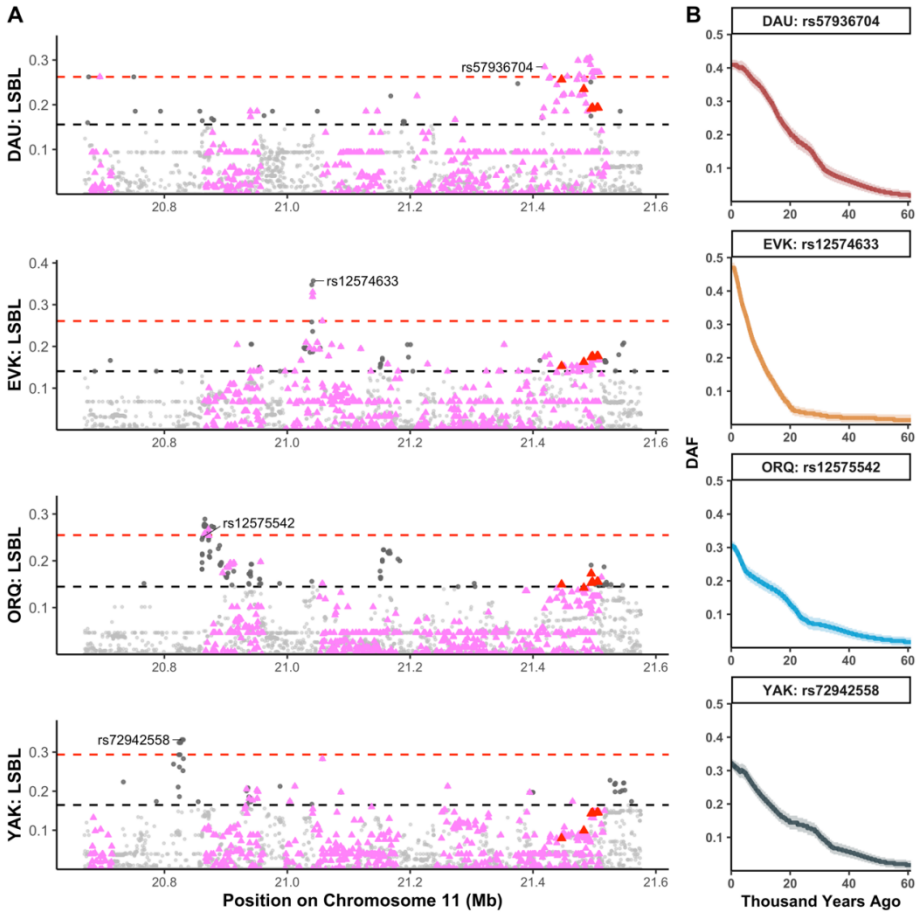

Fig. S24. LSBL local distribution and allele frequency trajectories of *NELL1*.

A: Local distribution of LSBL values within *NELL1* in each M.AR population and YAK. For each M.AR population and YAK, LSBL was calculated using HAN and the French population as reference populations. The black and red dashed lines indicate the top 1% and 0.1% threshold of the empirical distribution of LSBL, respectively. Each point represents one SNP, and each purple triangle represents one aAIM. No overlap was observed among the extremely highly differentiated SNPs across the four populations; thus, the strongest selection signals evidenced by *RELATE* were revealed at different variants in each population. However, upon relaxing the criteria, 16 highly differentiated SNPs (red triangles) were identified among three M.AR populations, which were also aAIMs.

B: The allele frequency trajectories across time in each M.AR population and YAK. Based on the *RELATE* results, rs57936704, rs12574633, rs12575542, and rs72942558 showed the strongest selection signal in DAU, EVK, ORQ, and YAK, respectively. The solid line indicates the frequency trajectory of these putative adaptive alleles, as inferred by *CLUES*, and the colored region indicates the 95% CI.

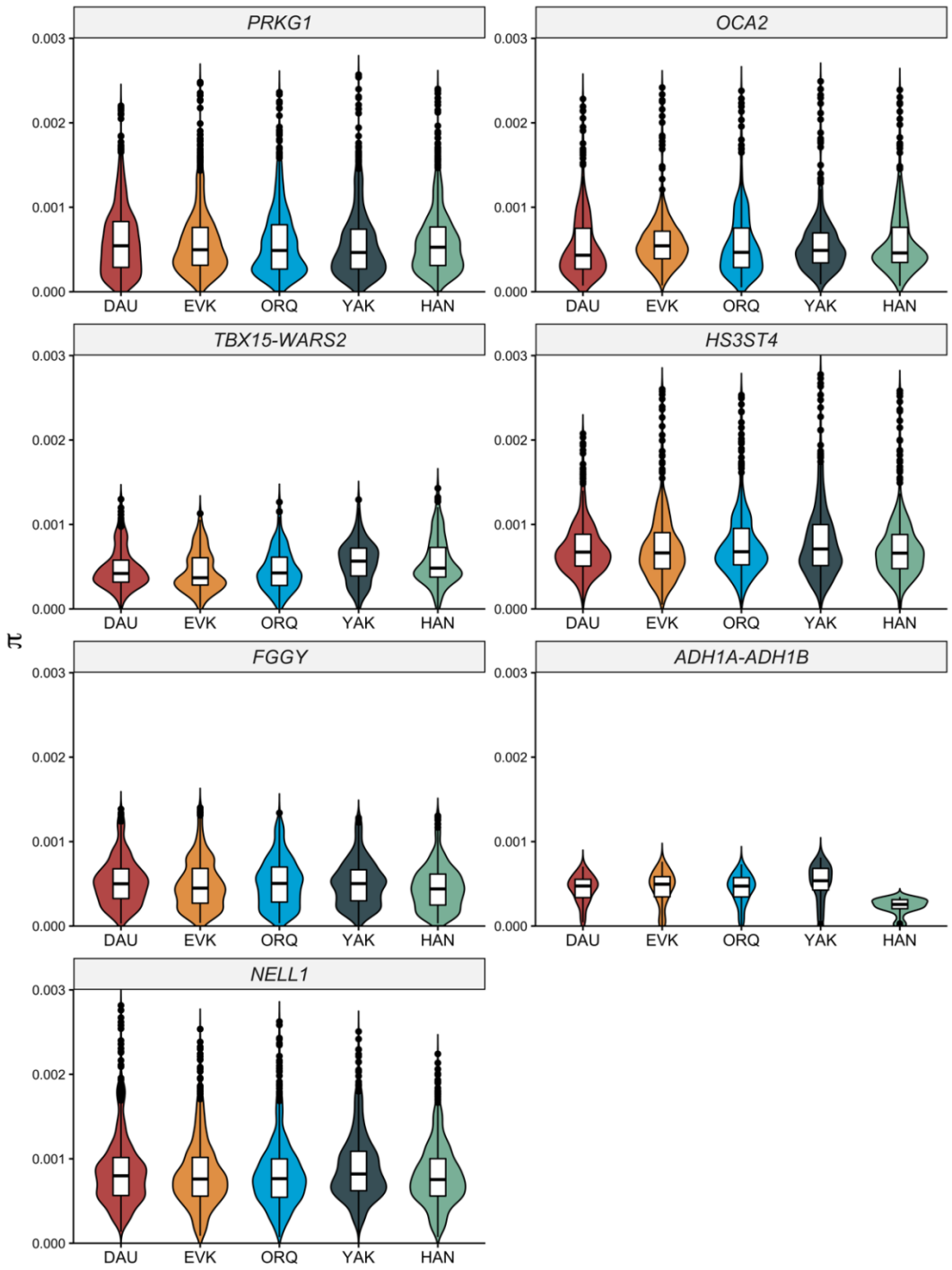

Fig. S25. Genetic diversity of CAR genes with selection signals measured by nucleotide differences.

Shown are nucleotide differences (π) of seven CAR genomic regions with selection signals identified in M.AR populations and YAK. For each genomic region, nucleotide differences of three M.AR populations, YAK, and HAN are presented. The higher π value indicates the higher genetic diversity of the genomic region.

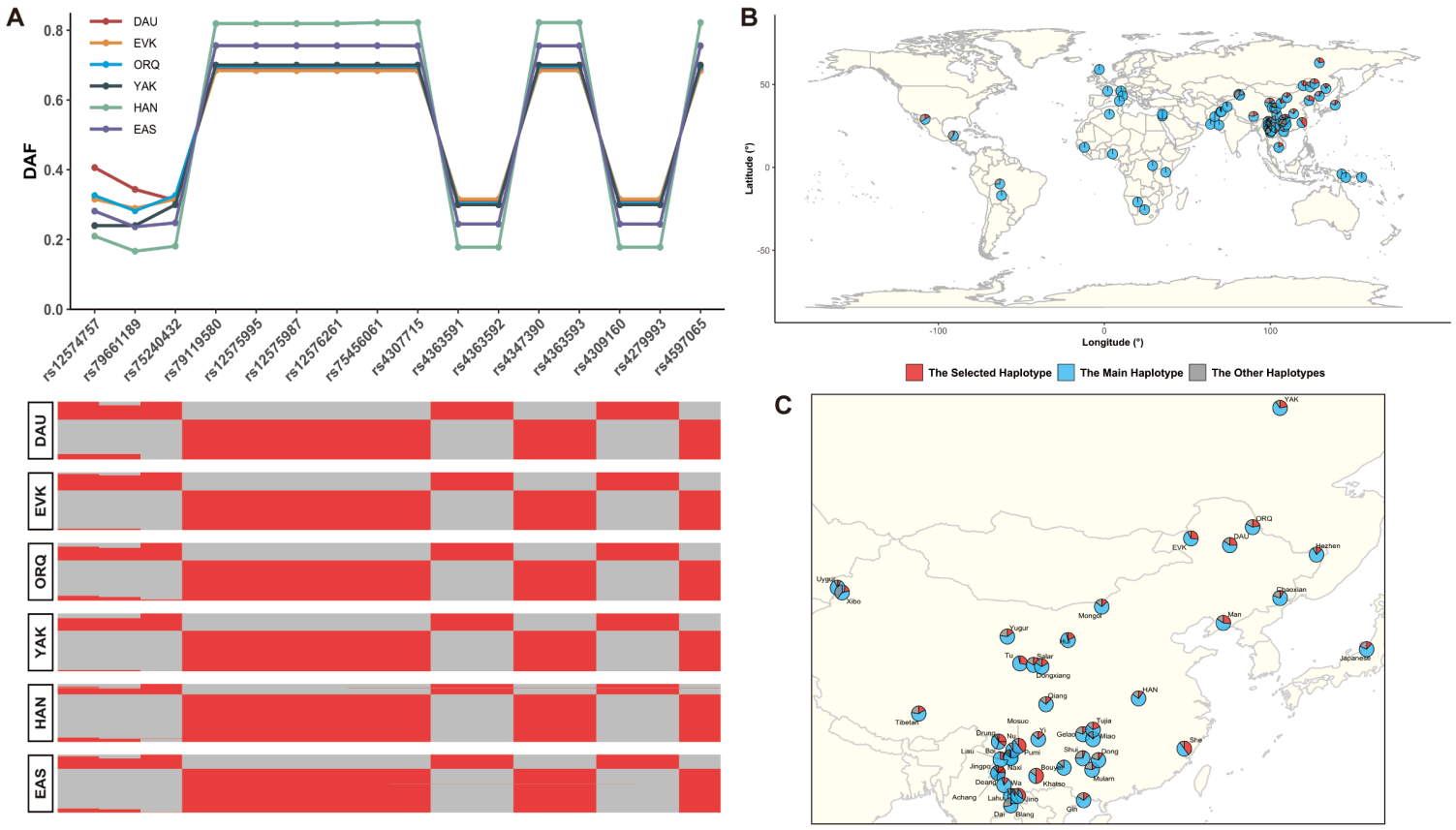

Fig. S26. Haplotype pattern and allele frequency of *NELL1*.

A: The haplotype consisted of 16 highly differentiated SNPs in three M.AR populations. The upper panel shows the DAF value of each variant in each population. The lower panel shows the haplotype distribution in each population, and red and grey represent the derived and ancestral alleles, respectively. EAS: East Asians except M.AR and HAN.

B and C: The frequency distributions of the haplotype in the context of global (B) and East Asia and Siberia (C).

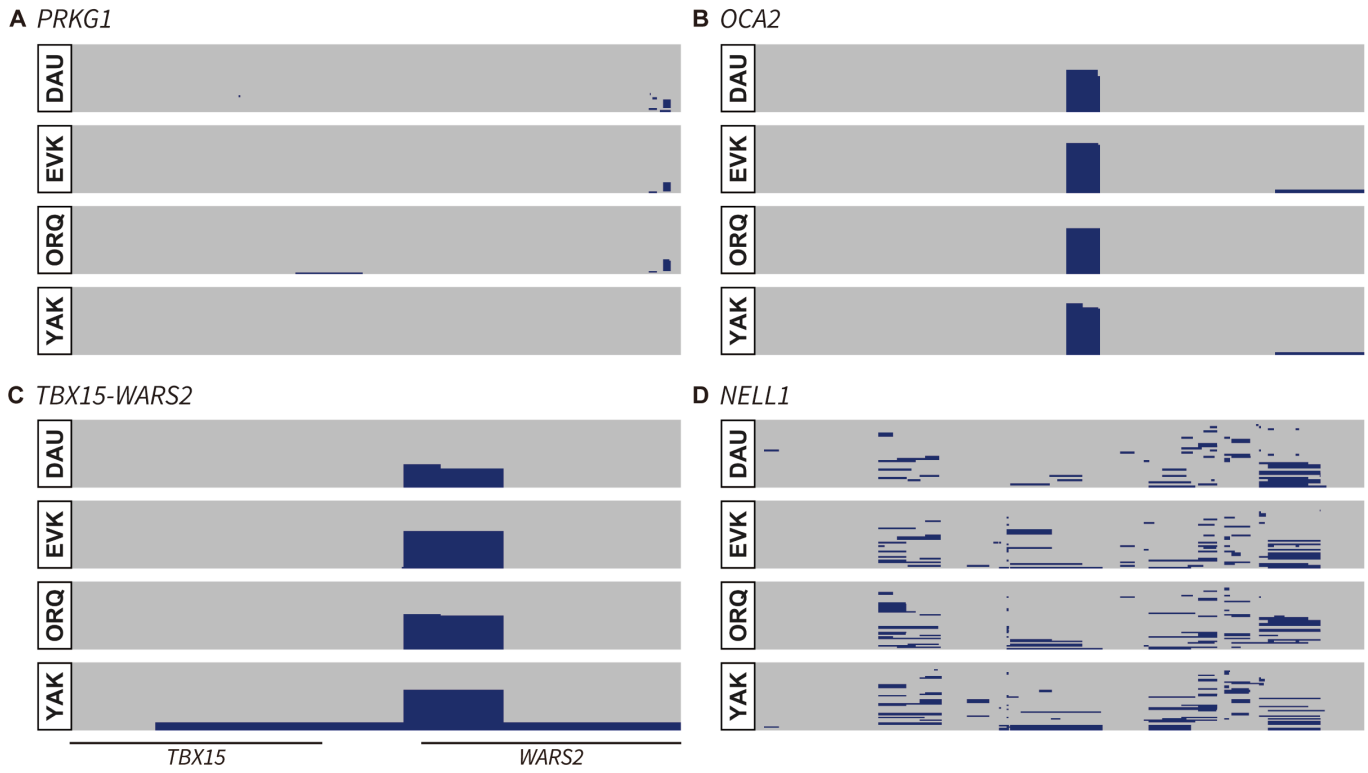

Fig. S27. Detected archaic segments within putative adaptive genes with archaic introgression signals in M.AR populations and YAK.

Detected archaic segments within *PRKG1* (A), *OCA2* (B), *TBX15-WARS2* (C), and *NELL1* (D) in three M.AR populations and YAK using *ArchaicSeeker* 2.0. Blue and grey represent archaic and non-archaic segments, respectively.

Table S3. mtDNA and NRY haplogroups of M.AR samples in this study.

Sample ID	Population	Gender	mtDNA Haplogroup	NRY Haplogroup
AAGC032558D	DWE	Male	G2a+152	O2a2b2a1a
AAGC032559D	DWE	Female	Z3d	
AAGC032560D	DWE	Male	D4	N1a1a1a1a
AAGC032571D	DWE	Female	D4q	
AAGC032572D	DWE	Female	F2a	
AAGC032665D	DWE	Male	Y1	N1a1a1a1a3a
AAGC032666D	DWE	Female	Z3d	
AAGC032667D	DWE	Male	G2a1	C2a1a1b1a2b
AAGC032668D	DWE	Male	C1a	O1b1a1a1a1b2a1a2a
AAGC032698D	DWE	Female	M7c1a1b	
AAGC032699D	DWE	Female	Z3d	
AAGC032028D	ORQ	Female	C4a1a+195	
AAGC032029D	ORQ	Female	D3	
AAGC032030D	ORQ	Female	D4b1	
AAGC032031D	ORQ	Female	D4f1	
AAGC032032D	ORQ	Male	D3	C2a1a1b1a2b
AAGC032033D	ORQ	Male	D3	N1a1a1a1a
AAGC032034D	ORQ	Female	B4b1a3a	
AAGC032067D	ORQ	Female	B4b1a3a	
AAGC032116D	ORQ	Male	A12a	C2b1a2a2b1a2
AAGC032117D	ORQ	Male	D3	C2a1a1b1a2b
AAGC032119D	ORQ	Female	A12a	
AAGC032120D	ORQ	Female	D3	
AAGC032121D	ORQ	Male	N	C2a1a1b1a2b
AAGC032122D	ORQ	Male	F1b1e	C2a1a1b1a2b
AAGC032123D	ORQ	Female	D4b2b2c	
AAGC032124D	ORQ	Male	A12a	C2a1a1b1a2b
AAGC032125D	ORQ	Male	G2b1a2	C2a1a1b1a2b
AAGC032126D	ORQ	Male	M7b1a1+(16192)	C2a1a1b1a2b
AAGC032127D	ORQ	Male	G2b1a2	C2a1a1b1a2b
AAGC032128D	ORQ	Male	M11a	C2b1b2
AAGC032129D	ORQ	Male	M11a	C2b1b2
AAGC032130D	ORQ	Male	D4	C2a1a1b1a2b
AAGC032312D	ORQ	Female	C4b	
AAGC032313D	ORQ	Female	D4l2a1	
AAGC032314D	ORQ	Male	D3	N1a1a1a1a
AAGC032332D	ORQ	Female	G2b1a2	

AAGC032333D	ORQ	Female	B4d1'2'3	
AAGC032334D	ORQ	Female	C4a1a+195	
AAGC032335D	ORQ	Male	F1c1a1	C2a1a1b1a2b
AAGC032584D	ORQ	Female	D4l1	
AAGC032616D	ORQ	Male	C5b1a	O2a1b1a1a1a1e1a2a
AAGC032113D	EVK	Male	C4b1	O2a2b1a2a1a1a1
AAGC032114D	EVK	Male	J1c4	O1b1a1a1a1b2a1a2a
AAGC032115D	EVK	Female	D4b2b	
AAGC032358D	EVK	Female	C4b3a	
AAGC032359D	EVK	Male	B5b2	C2a1a1b1a2b
AAGC032382D	EVK	Male	B4b1c2	C2a1a1b1a2b
AAGC032580D	EVK	Male	D4i2	O2a2b1a2a1a1b1b2b1
AAGC032581D	EVK	Male	H4a1a1a	C2a1a1b1a2b
AAGC032582D	EVK	Male	B4b1c2	O2a2b1a1a1c1a1a
AAGC032583D	EVK	Female	D4o2a	
AAGC032610D	EVK	Female	D4	
AAGC032611D	EVK	Male	H4a1a1a	O2a2b1a2a1a1b1b2b1
AAGC032612D	EVK	Female	G2b1a2	
AAGC032613D	EVK	Female	G1a1	
AAGC032614D	EVK	Male	A12a	C2a1a1b1a2b
AAGC032615D	EVK	Female	B4c1c1	
AAGC032706D	EVK	Male	B4c1c	O2a2b1a2a1a1b1b2b1
AAGC032707D	EVK	Female	C4a1a3b	
AAGC032708D	EVK	Female	H28a	
AAGC032730D	EVK	Female	D4b2b2b	

Table S4. Reported genes related to cold adaptation in Siberian, Inuit, and Alaskan populations.

Gene	Position (GRCh38)	Function Description	Selected Population	Evidence of Positive Selection	Citation
PLA2G2A	Chr1:19975431-19980434	Lipid Metabolism	Siberians	PBS	Hallmark et al. 2019
TBX15	Chr1:118883047-118989510	Adipocyte Differentiation	Greenlandic Inuit	PBS	Fumagalli et al. 2015
WARS2	Chr1:119031216-119140672	Energy Metabolism	Greenlandic Inuit	PBS	Fumagalli et al. 2015
KCNH1	Chr1:210678314-211134148	Adipose Tissue Production	Alaskans	PBS, iHS	Reynolds et al. 2019
THADA	Chr2:43230851-43596038	Energy Regulation and Metabolism	Southern Siberians	PBS, iHS, XP-EHH	Cardona et al. 2014
PRKG1	Chr10:50990888-52298350	Smooth Muscle Contraction	Central and Northeastern Siberians	PBS, iHS, XP-EHH	Cardona et al. 2014
FADS1-FADS2-FADS3	Chr11:61799627-61892224	Fatty Acid Metabolism	Greenlandic Inuit	PBS	Fumagalli et al. 2015
LRP5	Chr11:68298412-68449275	Energy Regulation and Metabolism	Northeastern Siberians	PBS, iHS, XP-EHH	Cardona et al. 2014
CPT1A	Chr11:68754620-68844277	Energy Regulation and Metabolism	Northeastern Siberians	PBS, iHS, XP-EHH; iHS, Tajima's D	Cardona et al. 2014; Clemente et al. 2014
OCA2	Chr15:27719008-28099315	Melanin Production	Alaskans	PBS, iHS	Reynolds et al. 2019
PLIN1	Chr15:89664367-89679367	Lipid Metabolism	Siberians	PBS	Hallmark et al. 2019
HS3ST4	Chr16:25691959-26137685	Heparan Sulfate Biosynthesis	Alaskans	PBS, iHS	Reynolds et al. 2019
FN3KRP	Chr17:82716706-82728013	Energy Metabolism	Greenlandic Inuit	PBS	Fumagalli et al. 2015
ANGPTL8	Chr19:11239619-11241943	Lipid Metabolism	Siberians	PBS	Hallmark et al. 2019

The following supplementary tables were provided as spreadsheets:

Table S1. Information of populations analyzed in this study.

Table S2. Information of samples used in this study.

Table S5. Putative adaptive variants within reported CAR genes with significant selection signatures in M.AR.

Table S6. Putative adaptive variants within newly identified CAR genes.