Supplementary Table 1: SNP filtering applied to the Japanese and HapMap dataset

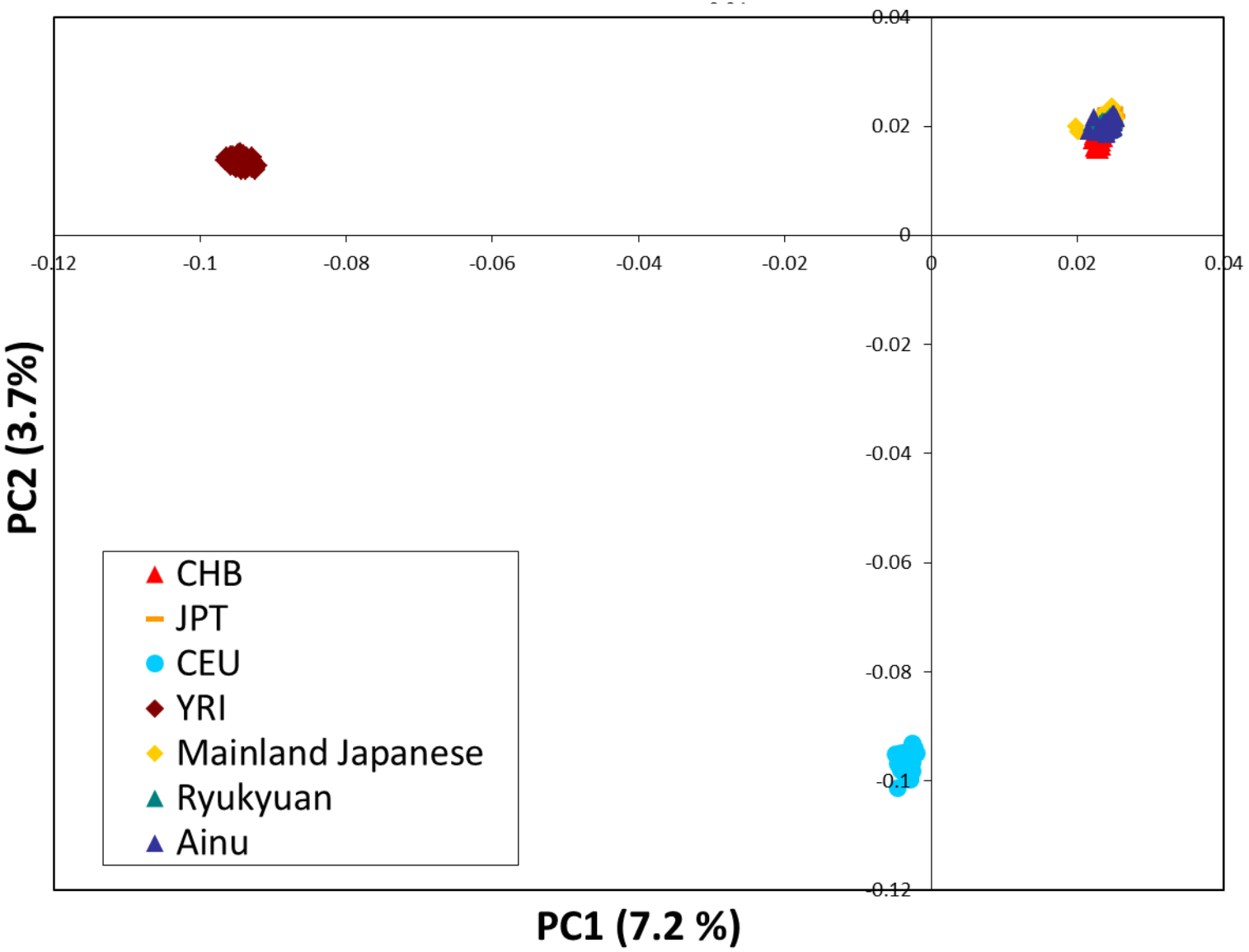
Population	No. samples	Number of SNPs omitted		Remaining SNP
		Genotyping call rate (<95%)	HWE (p<0.001)	
Ainu	36	212,448[†]	449	655,788
Ryukyu	35	29,874	538	837,845
Mainland Japanese	198	17,169	1888	849,200
CHB	42	3,069	336	864,852
JPT	45	5,004	446	862,807
CEU	89	4,887	514	862,856
YRI	89	4,780	706	862,771

[†]Includes SNP omitted based on confidence scores

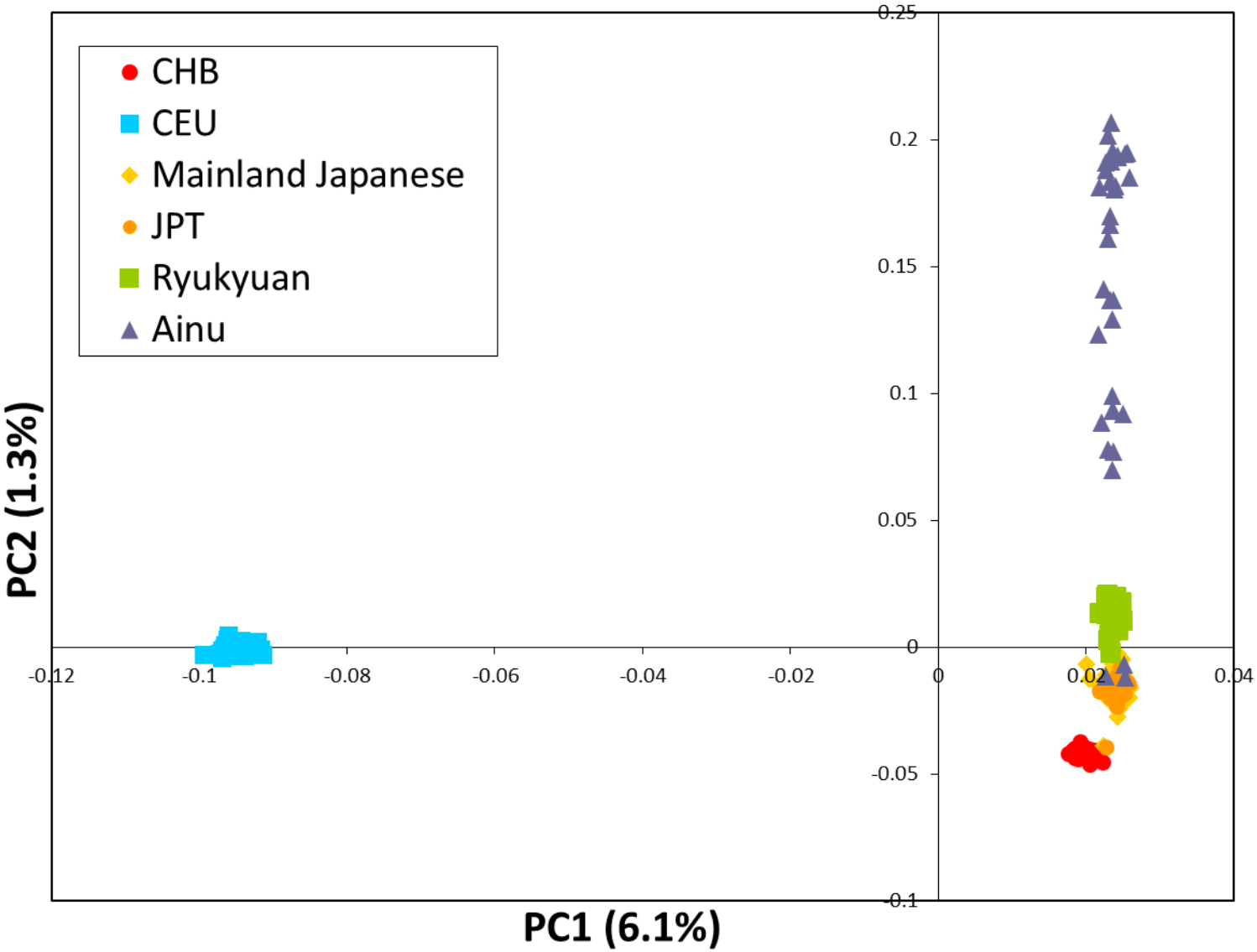
Supplementary Table 2. East Asian populations from HDGP-CEPH and PASNP datasets which were merged with the Japanese-HapMap datasets

Population ID	Ethnicity	n	Dataset	Geographical origin
Dai	Dai	10	HGDP-CEPH	China
Daur	Daur	9	HGDP-CEPH	China
Han	Han	44	HGDP-CEPH	China
Hezhen	Hezhen	9	HGDP-CEPH	China
Japanese	Japanese	28	HGDP-CEPH	Japan
Lahu	Lahu	8	HGDP-CEPH	China
Miaozu	Miaozu	10	HGDP-CEPH	China
Mongolia	Mongolia	10	HGDP-CEPH	China
Naxi	Naxi	8	HGDP-CEPH	China
Oroqen	Oroqen	9	HGDP-CEPH	China
She	She	10	HGDP-CEPH	China
Tu	Tu	10	HGDP-CEPH	China
Tujia	Tujia	10	HGDP-CEPH	China
Xibo	Xibo	9	HGDP-CEPH	China
Yakut	Yakut	25	HGDP-CEPH	Siberia
Yizu	Yizu	10	HGDP-CEPH	China
AX-AM	Ami	10	PASNP	Taiwan
AX-AT	Atayal	10	PASNP	Taiwan
CN-CC	Zhuang	26	PASNP	China
CN-HM	Hmong	26	PASNP	China
CN-UG	Ugyur	26	PASNP	China
CN-SH	Han	21	PASNP	China
CN-JN	Jinuo	29	PASNP	China
JP-ML	Mainland Japanese	71	PASNP	Japan
JP-RK	Ryukyu	49	PASNP	Japan
CN-WA	Wa	56	PASNP	China
KR-KR	Korean	90	PASNP	Korea
TW-HA	Han	80	PASNP	Taiwan
CN-JI	Jiamao	31	PASNP	China
CN-GA	Han	30	PASNP	China

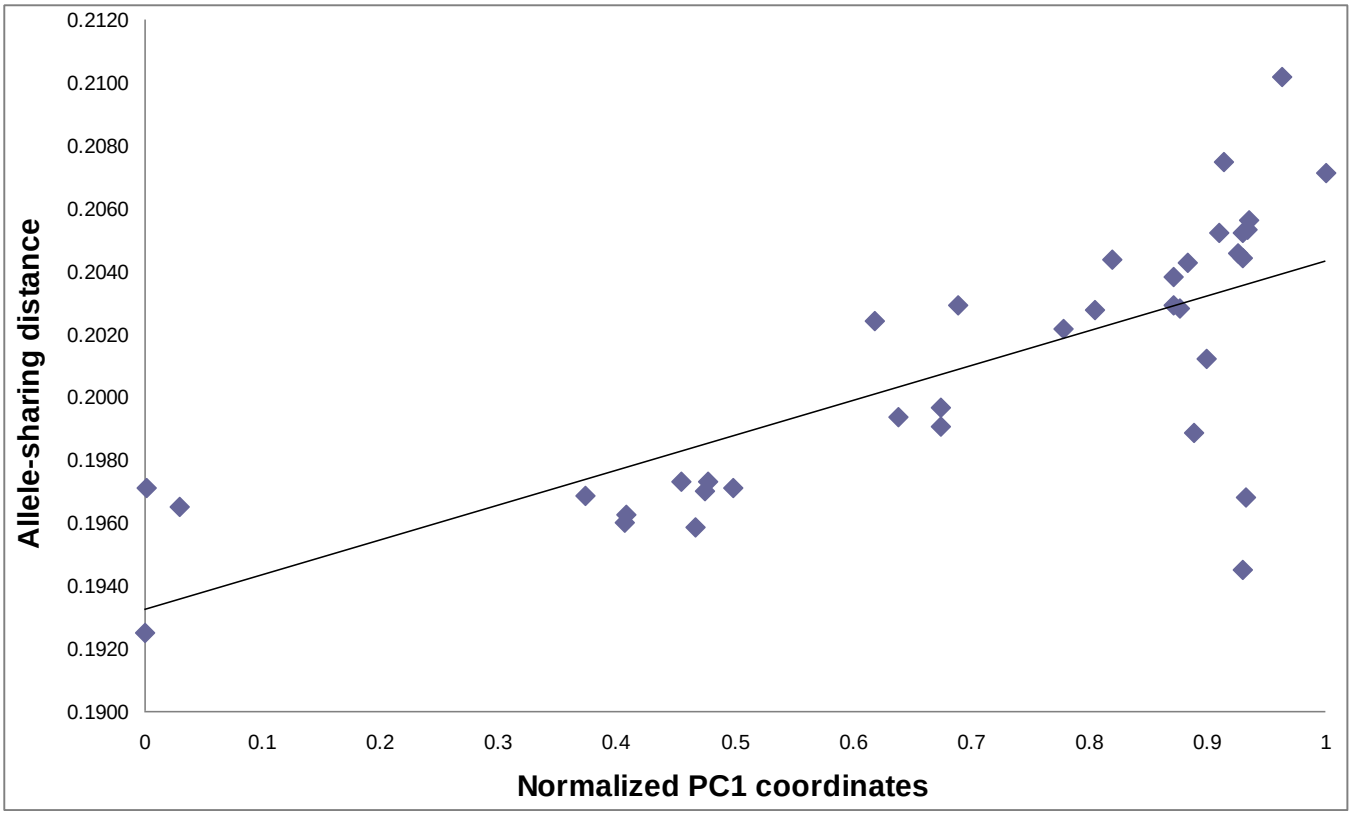
Supplementary Figure 1: PCA plot (PC1 and PC2) of three populations in the Japanese Archipelago and four populations of the HapMap dataset



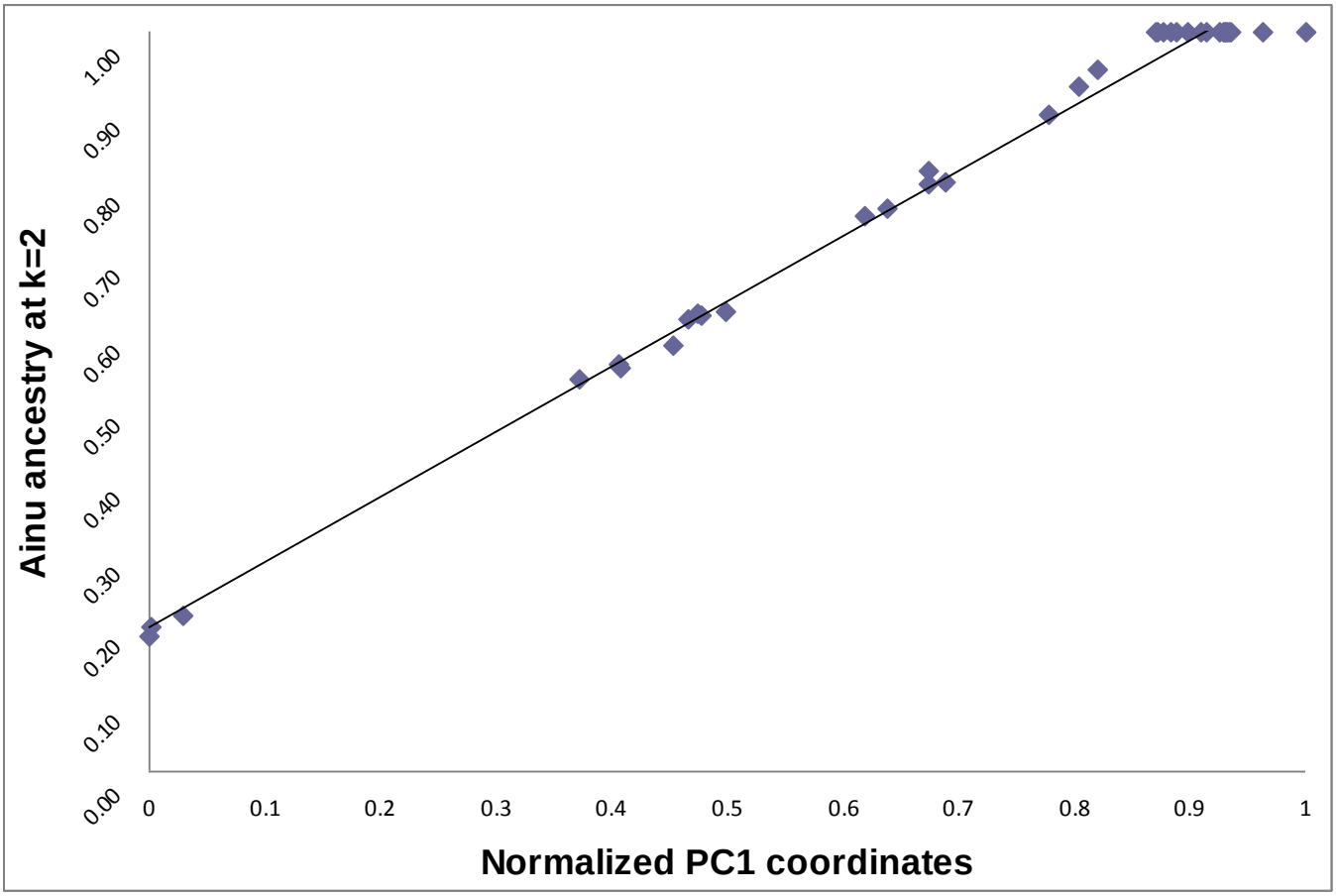
Supplementary Figure 2: PCA plot (PC1 and PC2) of three populations in the Japanese Archipelago and three populations of the HapMap dataset (African population was excluded)



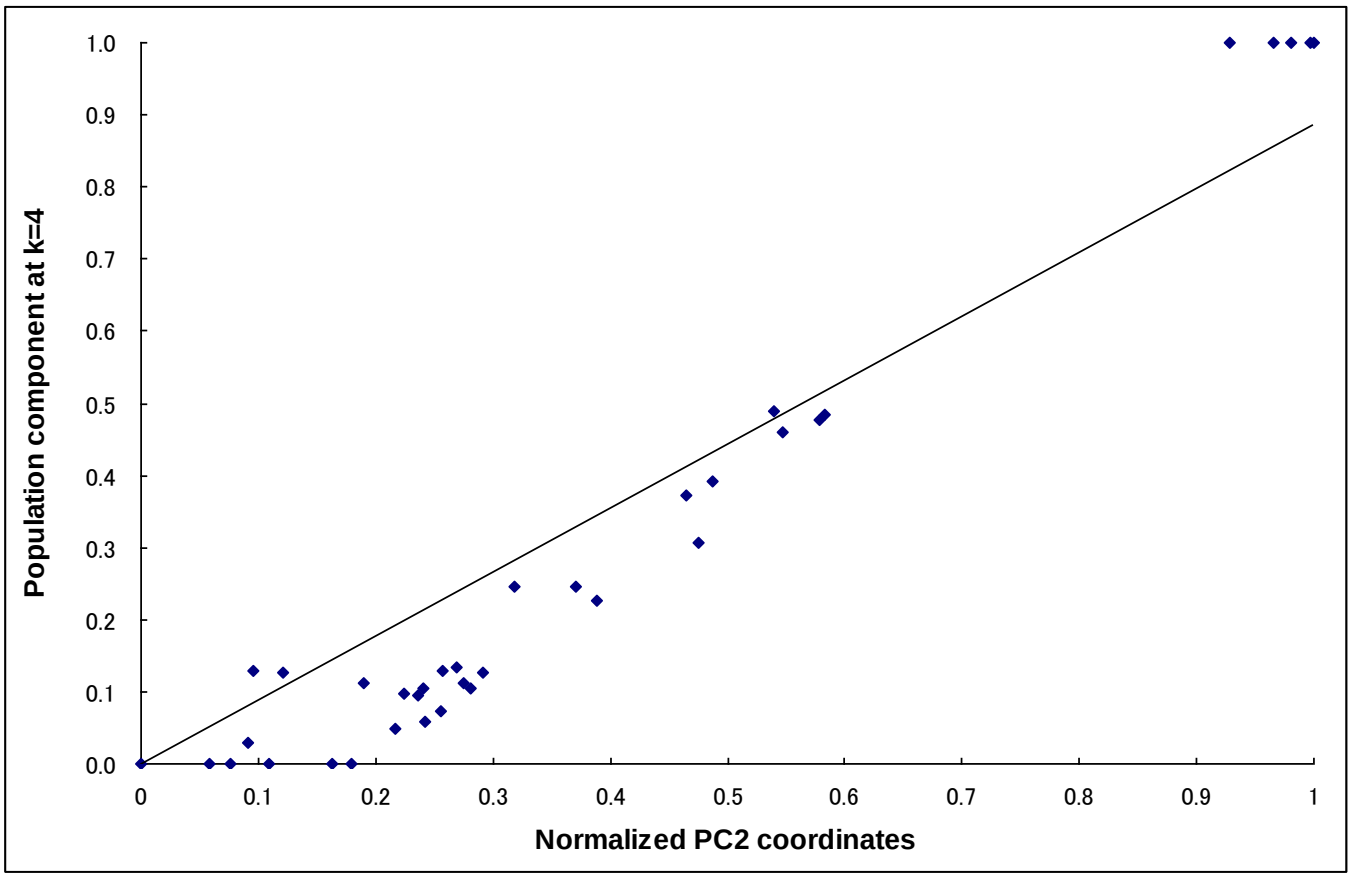
Supplementary Figure 3. Correlation between PC1 coordinates from Figure 1A (X-axis) and allele sharing distances (Y-axis) between Ainu and Mainland Japanese individuals. PC1 coordinates have been normalized to range from 0 to 1, so that 0 represents the closest proximity to the Mainland Japanese cluster. Pairwise allele sharing distance between Ainu and Mainland Japanese individuals was computed using AWclust[1] and was averaged to obtain values on the y-axis.



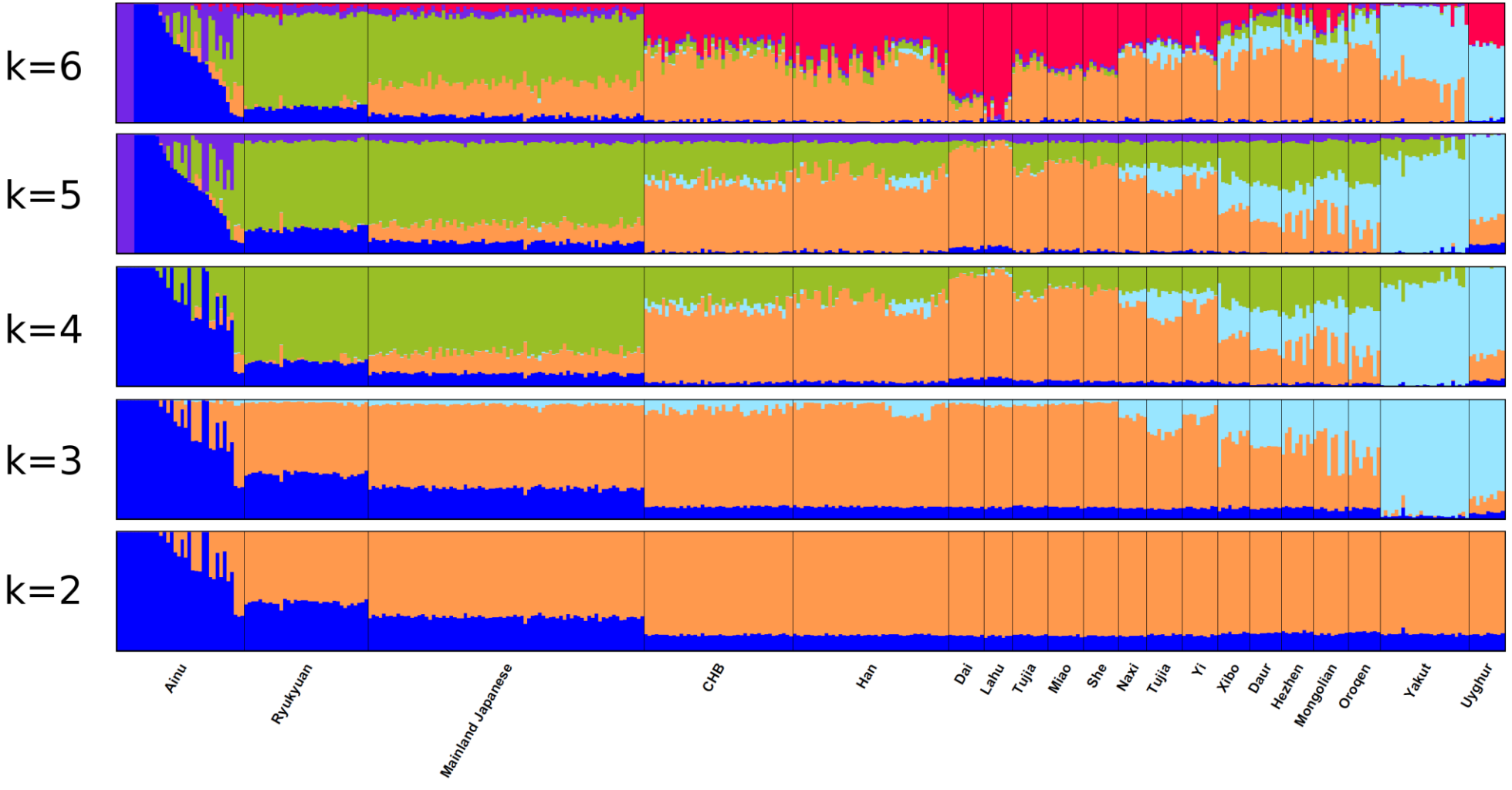
Supplementary Figure 4. Correlation between PC1 coordinates from Figure 1A (X-axis) and proportions of Ainu ancestry from *frappe* analysis at k=2 (blue-colored ancestry component in Fig. 2). PC1 coordinates have been normalized to range from 0 to 1, so that 0 represents the closest proximity to the Mainland Japanese cluster.



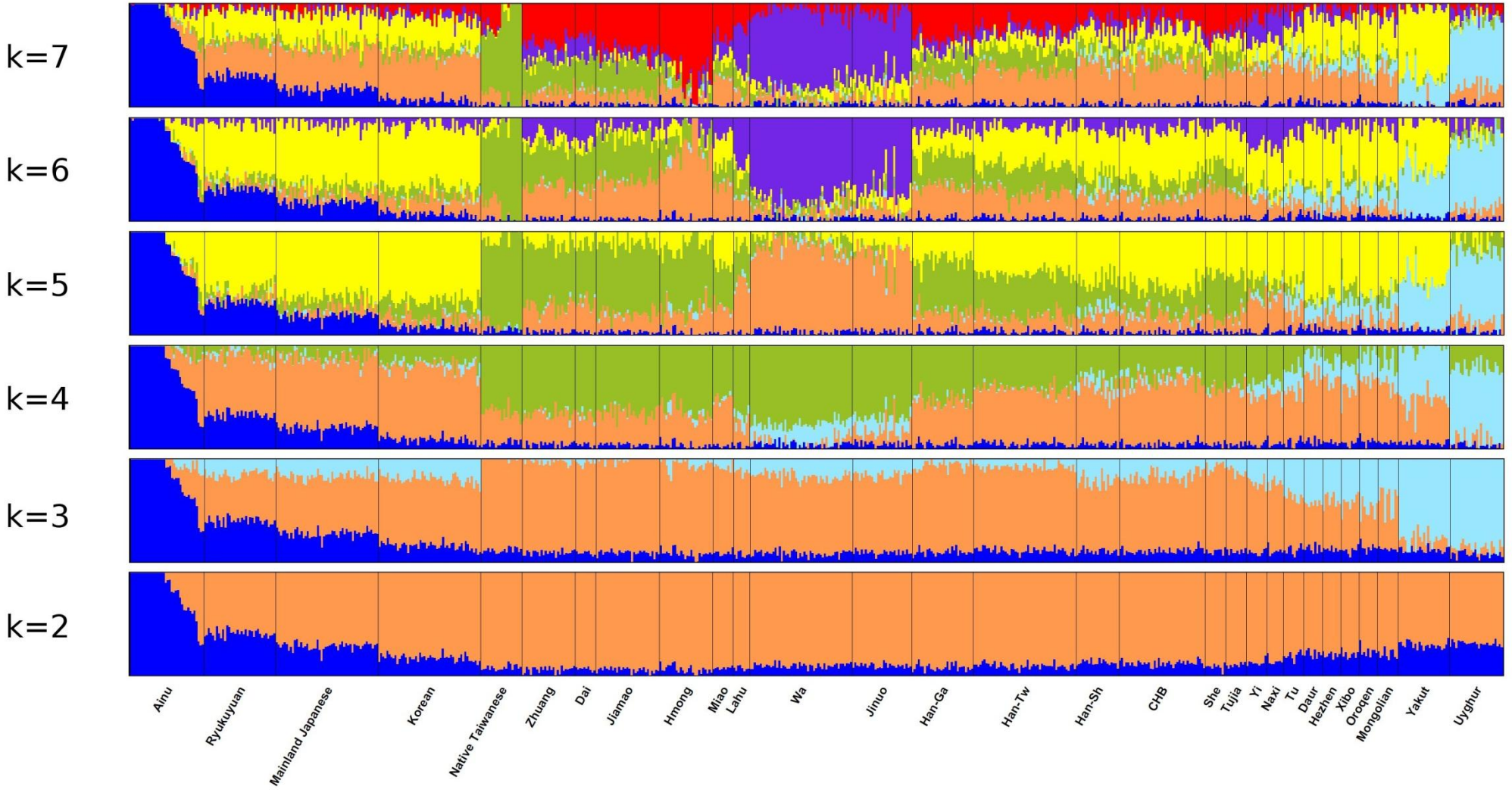
Supplementary Figure 5. Correlation between PC2 coordinates from Figure 1A (X-axis) and proportions of Ainu ancestry from *frappe* analysis at k=4 (purple-colored ancestry component in Fig. 2). PC2 coordinates have been normalized to range from 0 to 1, so that 1 represents the maximum value of PC2.



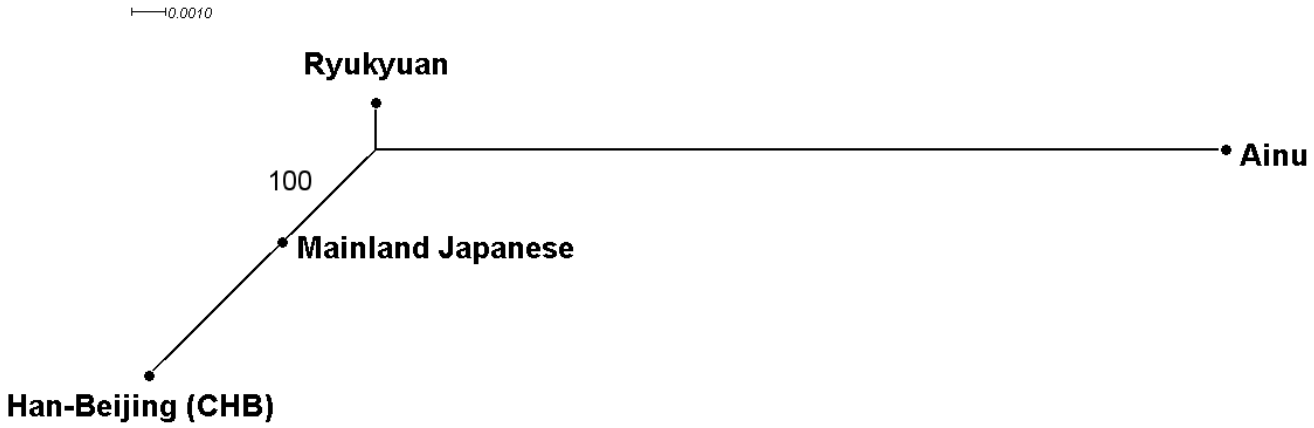
Supplementary Figure 6: Results of frappe analysis using the Japanese populations dataset merged with HGDP-CEPH populations listed in Supplementary Table 2. Results from k=2 to k=6 are shown.



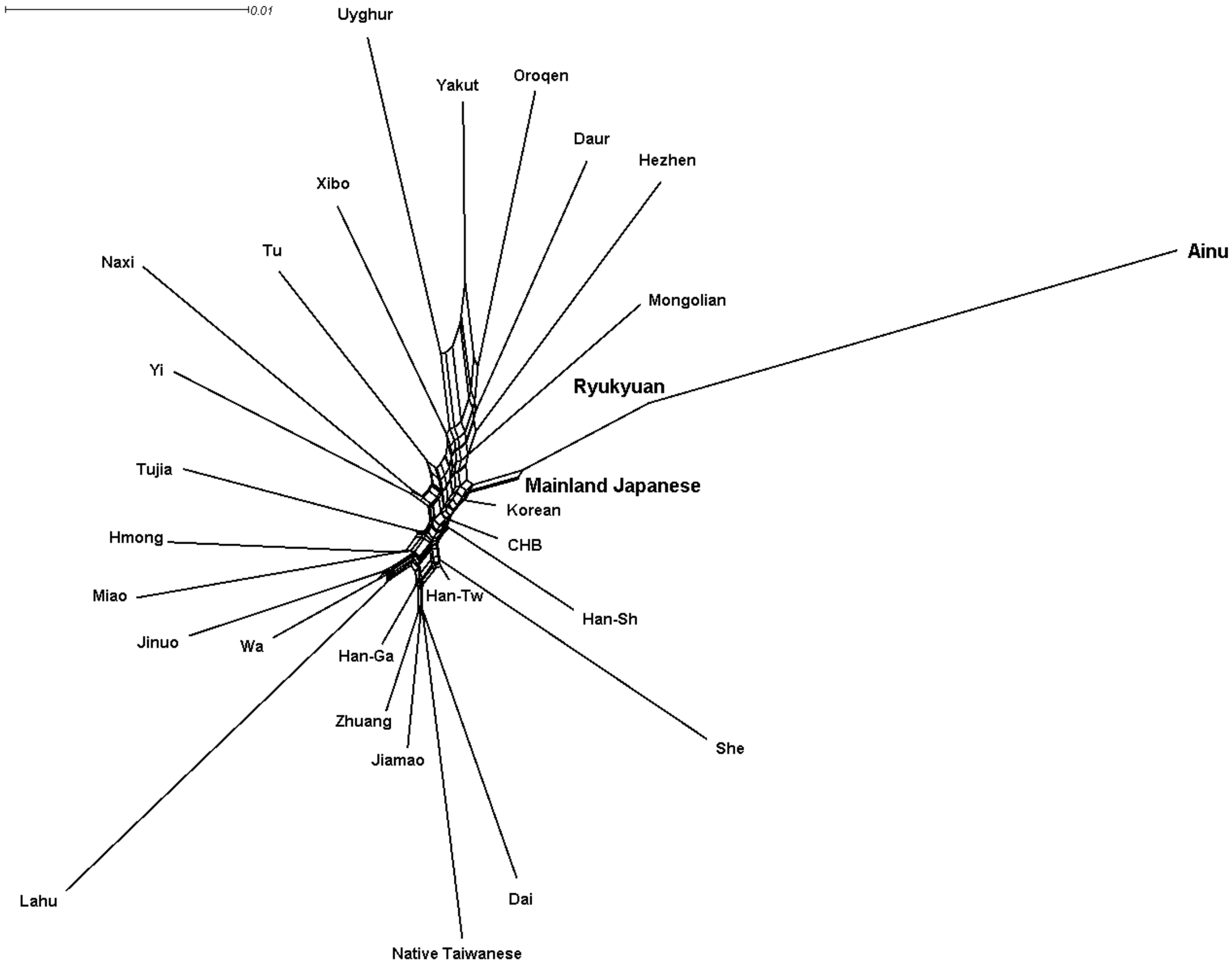
Supplementary Figure 7: Results of frappe analysis using the Japanese population dataset merged with HGDP-CEPH and PASNP populations listed in Supplementary Table 2. Results from k=2 to k=7 are shown.



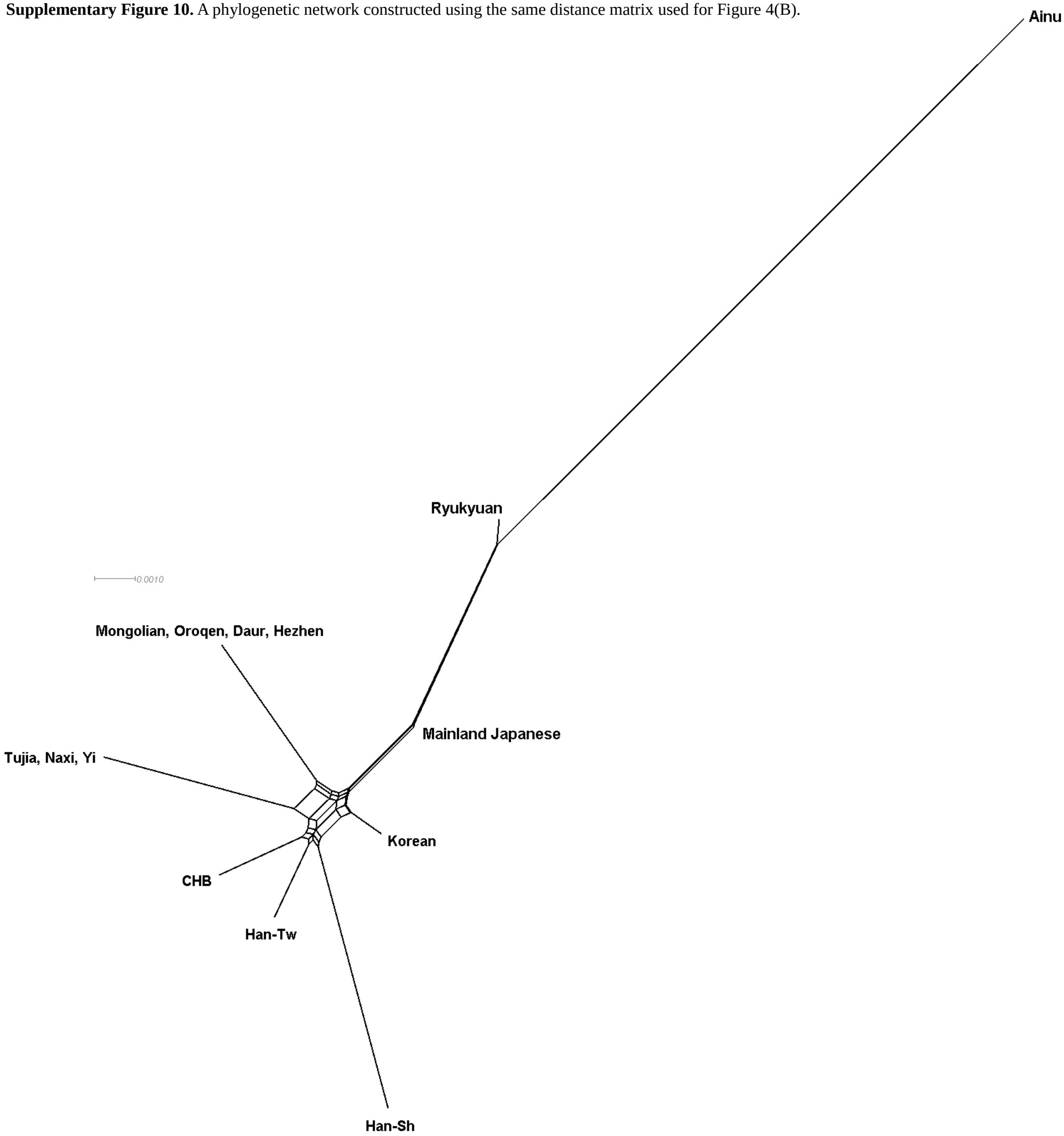
Supplementary Figure 8: Maximum-likelihood tree using approximately 600,000 SNP in the three Japanese populations and HapMap Han Chinese from Beijing.



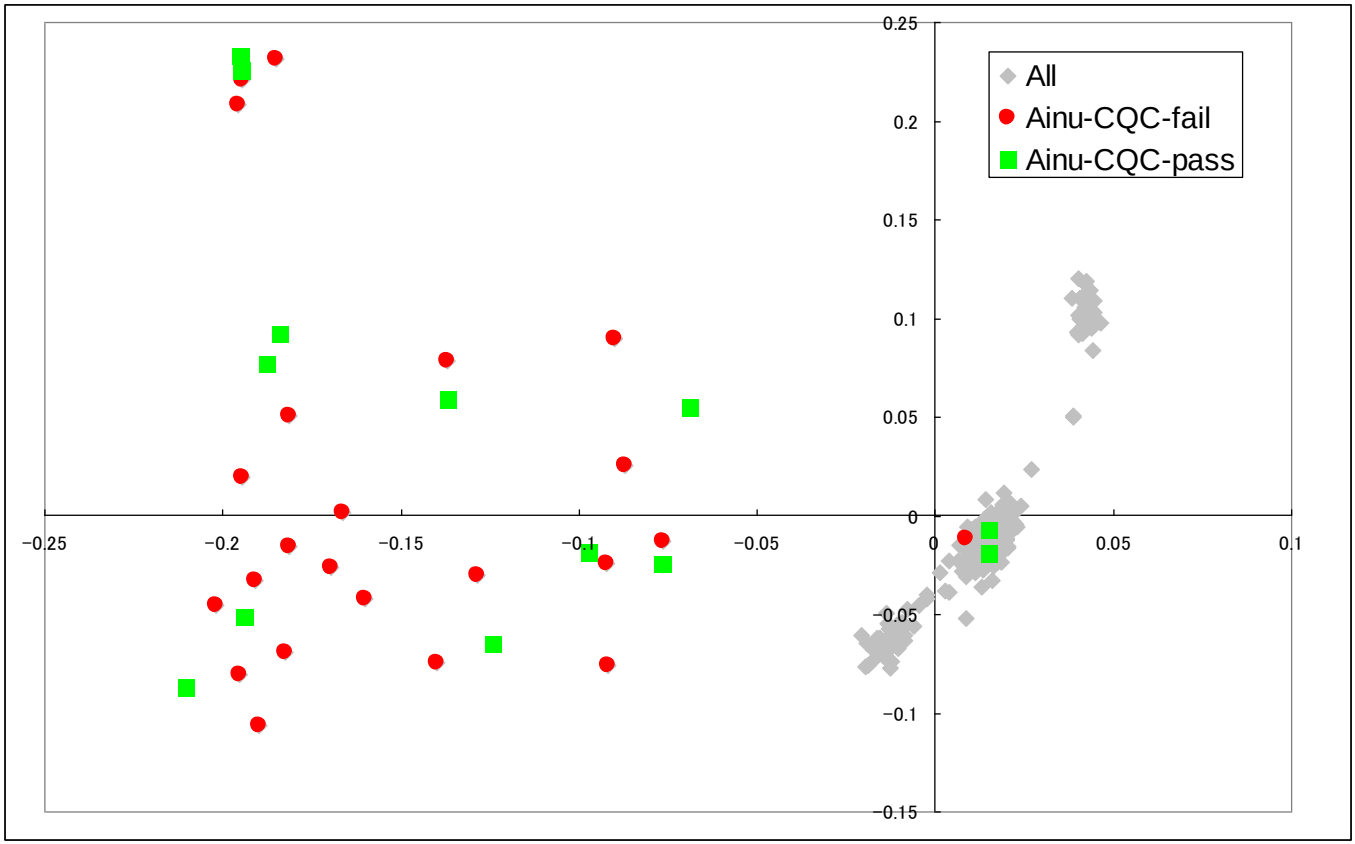
Supplementary Figure 9. A phylogenetic network constructed using the same distance matrix used for Figure 4(A).



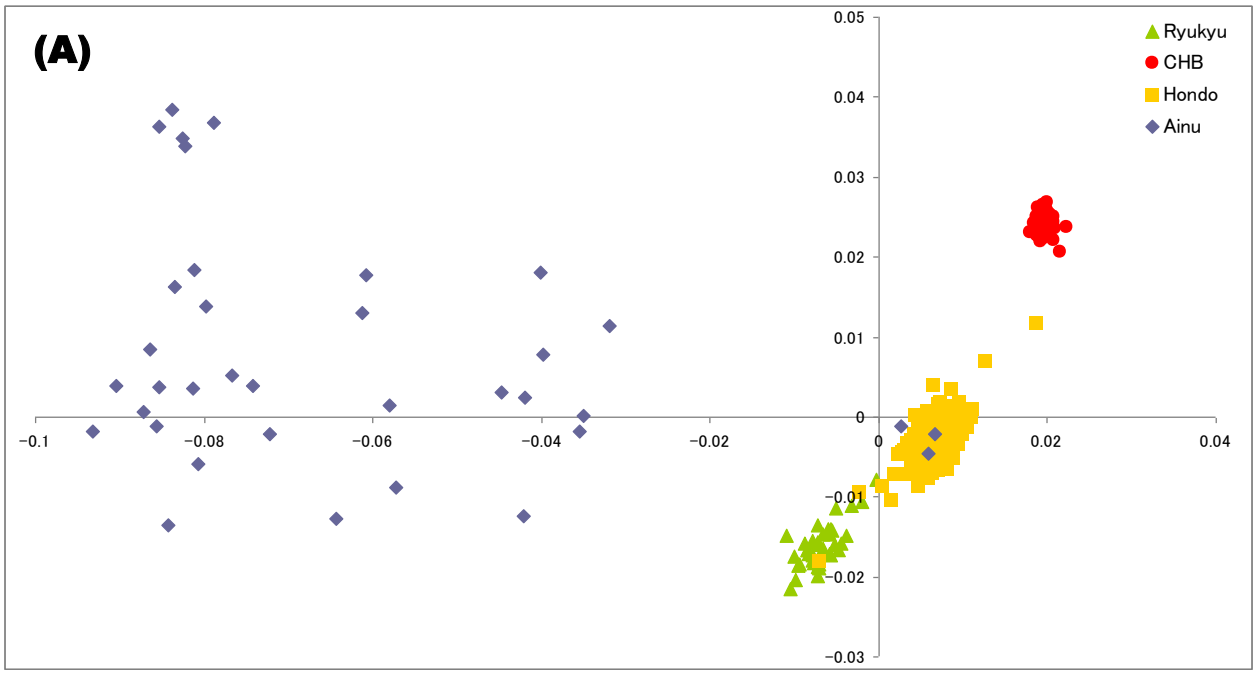
Supplementary Figure 10. A phylogenetic network constructed using the same distance matrix used for Figure 4(B).

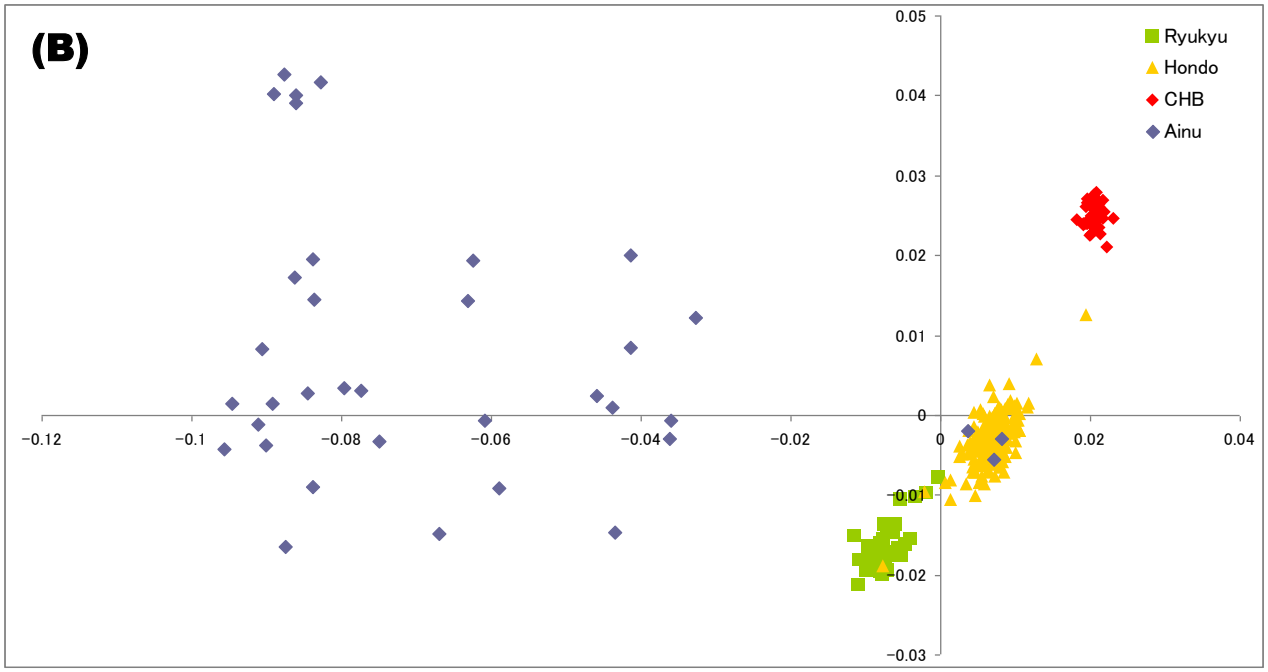


Supplementary Figure 11. Distribution of Affymetrix Contrast QC (CQC) values for 36 Ainu samples on PCA plot of Figure 2(A). Samples with QCQ values below 0.4 are interpreted as having failed this QC step as recommended by Affymetrix.

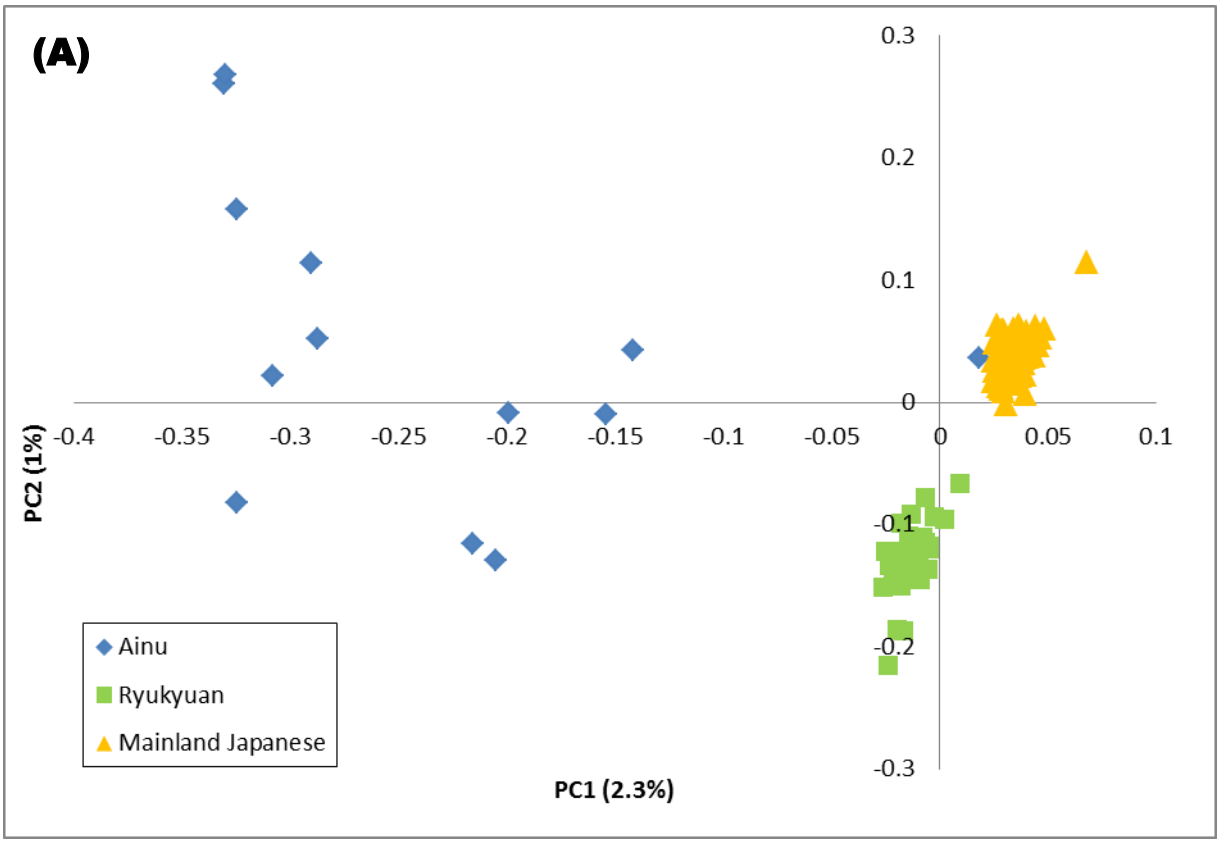


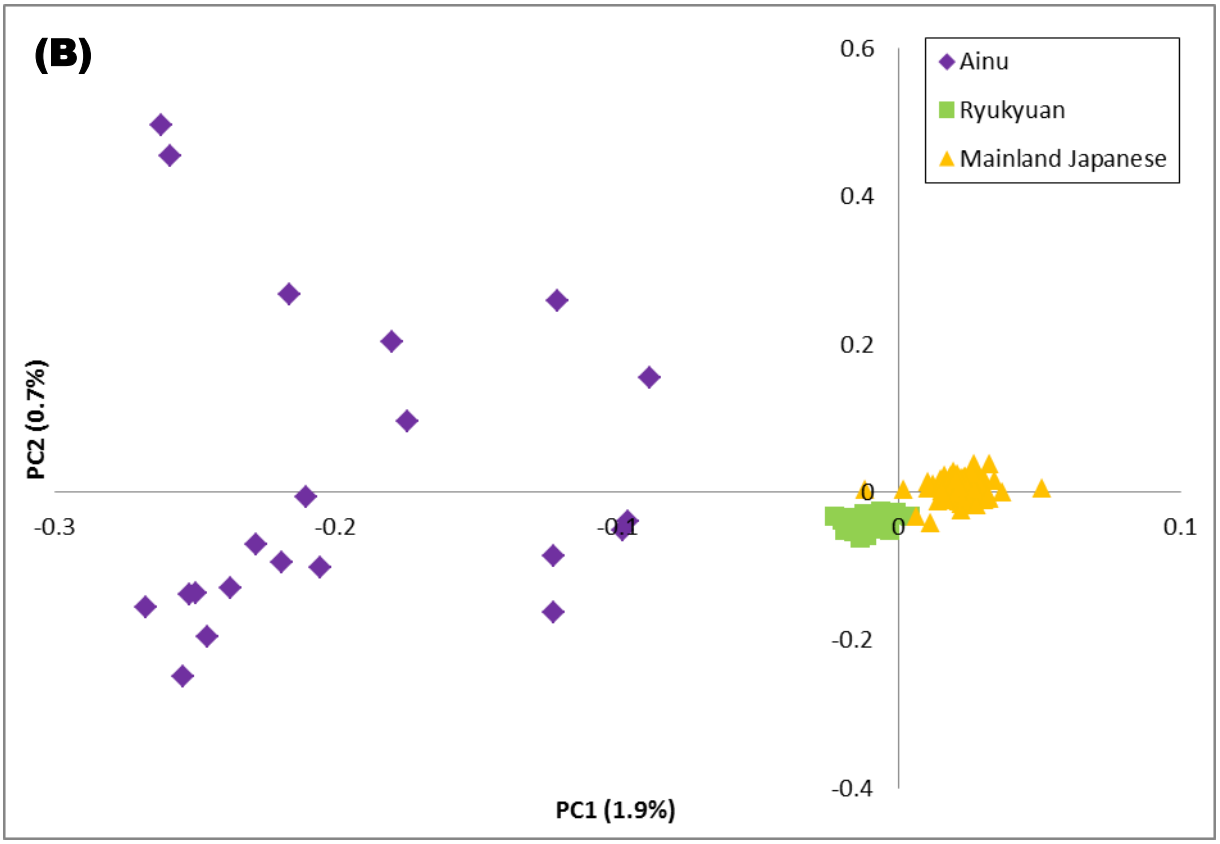
Supplementary Figure 12: PCA plots of Japanese and Han-Chinese populations using SNP sets obtained from different SNP genotyping call rate thresholds. A) 90% call rate (1106 SNP omitted, 684907 SNP remaining). B) 100% call rate (215061 SNP omitted, 470952 SNP remaining).





Supplementary Figure 13: A) PCA plot using 13 Ainu samples which passed Affymetrix cQC threshold of 0.4. A total number of 825,484 good quality SNPs were used. B) PCA plot using 23 Ainu samples which did not pass the Affymetrix cQC threshold of 0.4. Total number of SNPs used was 641,314.





Supplementary references

1. Gao, X. & Starmer, J. D. AWclust: point-and-click software for non-parametric population structure analysis. BMC Bioinformatics 9, 77 (2008).