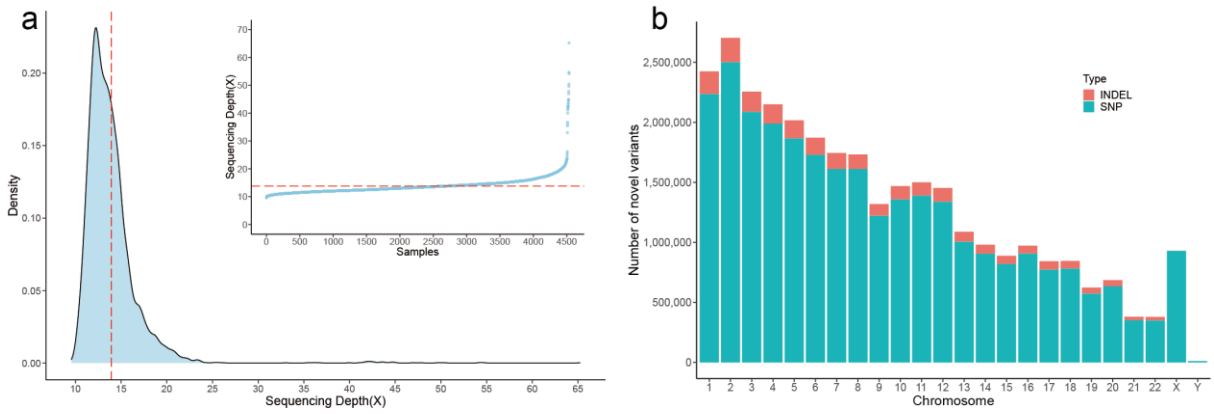
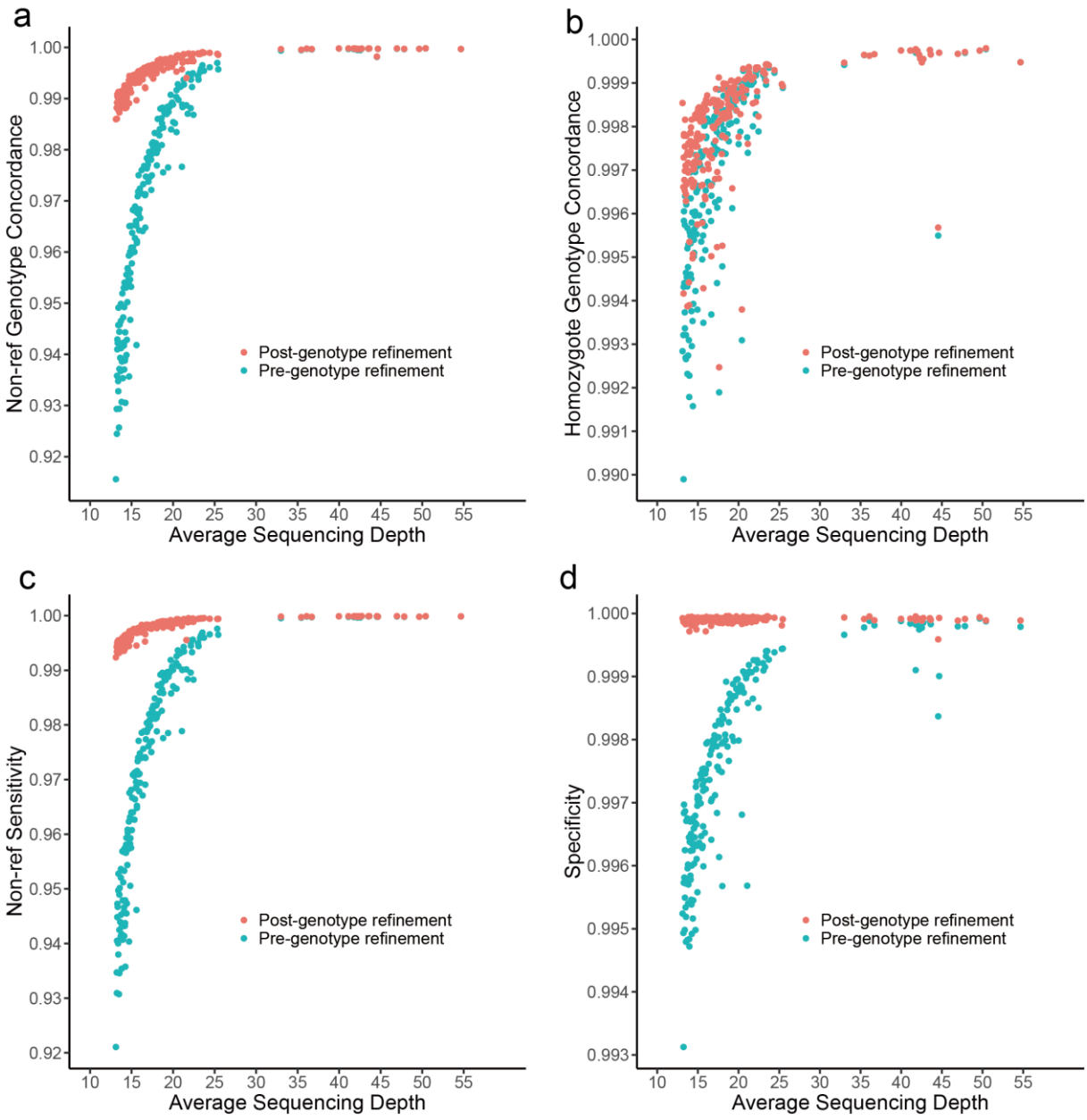
Extended Data



Extended Data Fig. 1 The sequencing depth and novel variants in WBBC cohort.

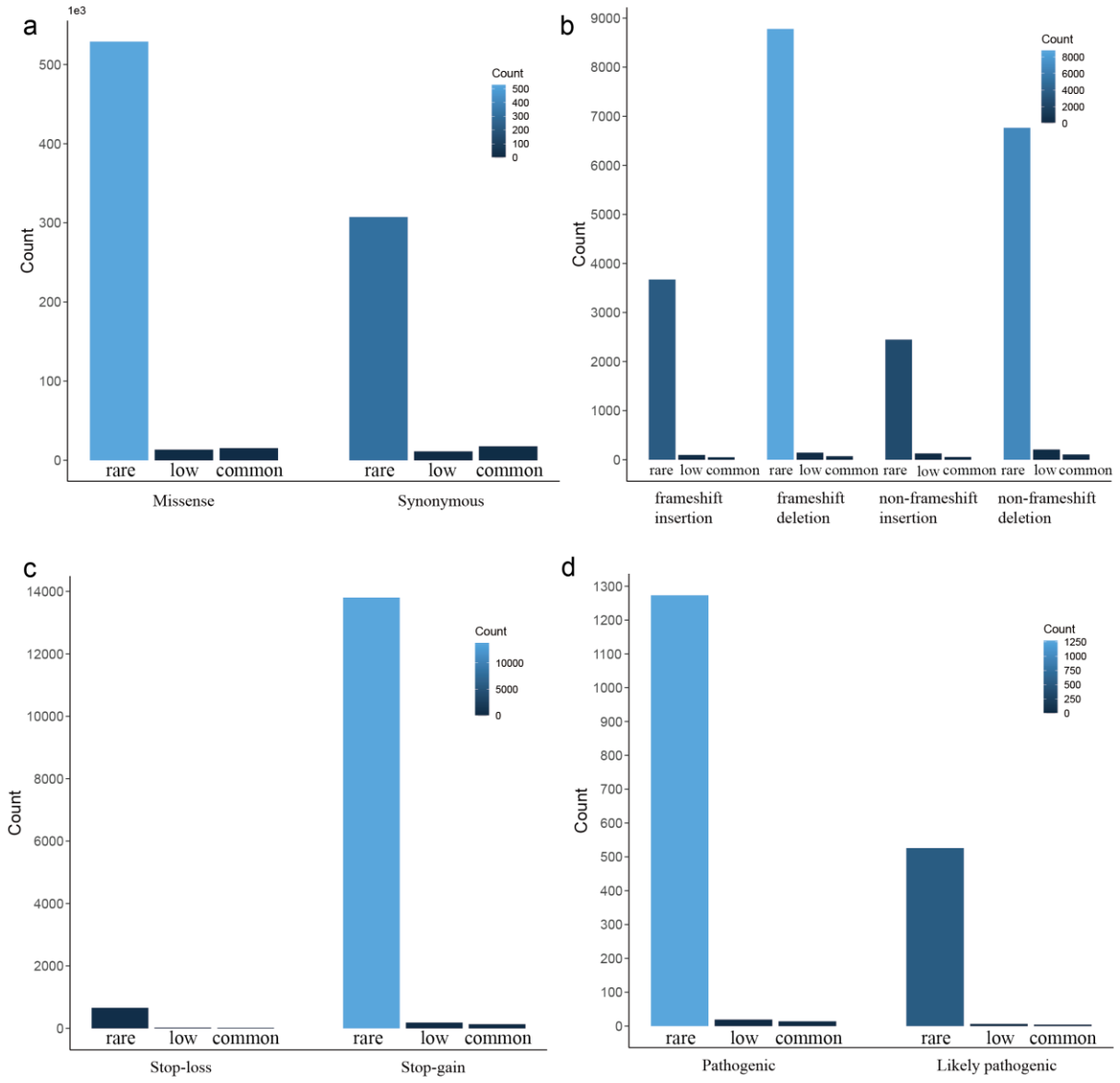
a, The sequencing depth of all 4,489 samples. The central red dot lines are the median. The inner chart represents the sequencing depth by sorted samples. The area plot indicates the density of sequencing depth.

b, The novel variants based on dbSNP build 151 and their distribution in each chromosome. Green represents the SNVs and red denotes the INDELs.



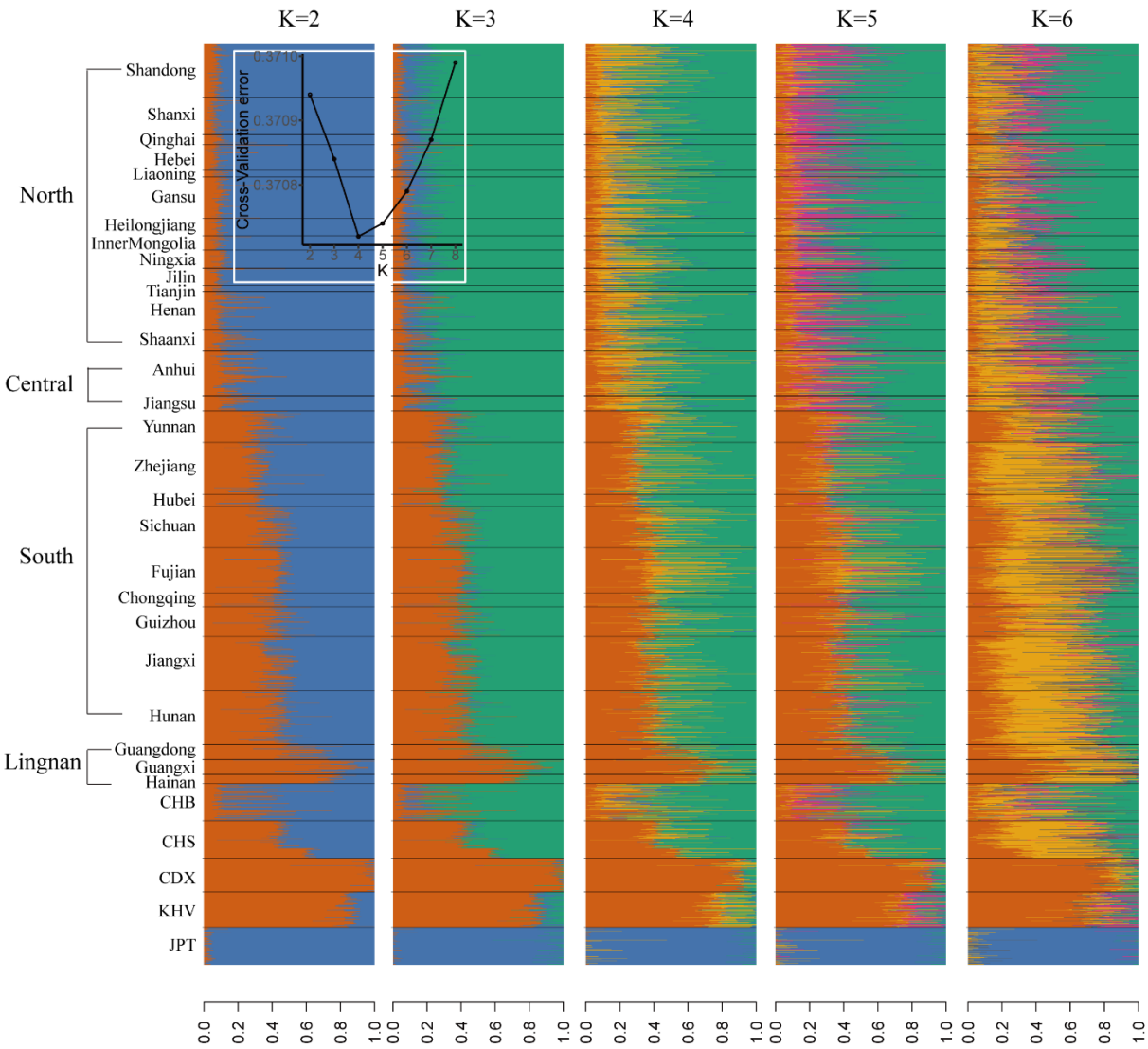
Extended Data Fig. 2 A comparison of the genotype concordance between WGS and SNP array data.

a, Non-reference genotype concordance. **b,** homozygote genotype concordance. **c,** non-reference sensitivity. **d,** specificity. The LD-based genotype refinement was conducted by BEAGLE software version 5.1. The red dots indicate the average proportion of post-genotype refinement and the green dots denote the raw genotype calls.



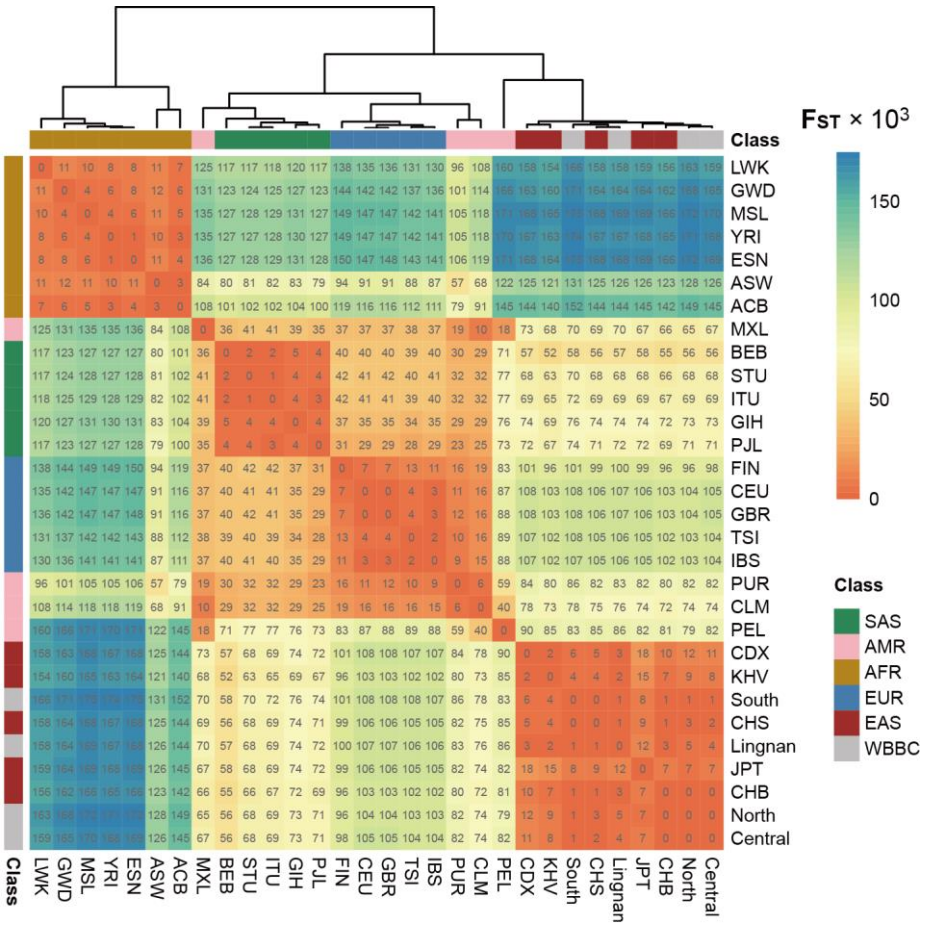
Extended Data Fig. 3 Functional annotation and distribution of variants in three frequency bins.

Rare allele (< 0.5%), low-frequency allele (0.5% ≤ AF ≤ 5%), and common allele (AF > 5%) are from left to right. The x-axis represents the functional categories, while the y-axis indicates the variants count. **a,** Missense and Synonymous variants. **b,** frameshift insertion, frameshift deletion, non-frameshift insertion, and non-frameshift deletion. **c,** Stoploss vs. Stopgain. **d,** Pathogenic and likely pathogenic variants annotated by ClinVar.



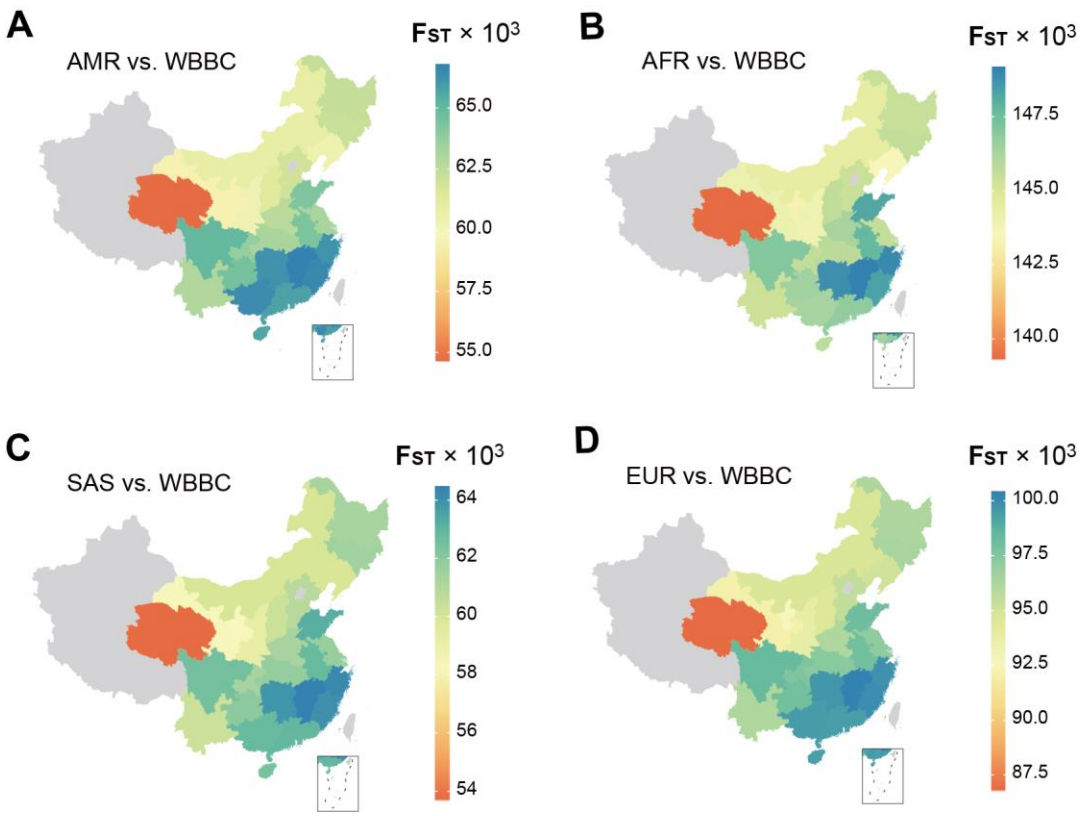
Extended Data Fig. 4 ADMIXTURE analysis of the Han Chinese, CHB, CHS, CDX, KHV and JPT individuals from the 1000 Genomes Projects.

The Han Chinese (2,056 individuals) was selected from our WBBC cohort. CHB (103 individuals) is a Han Chinese in Beijing. CHS (105 individuals) is Han Chinese from southern China. CDX (93 individuals) is Chinese Dai in Xishuangbanna. KHV (99 individuals) is Vietnamese from Kinh in Ho Chi Minh City. JPT (104 individuals) is Japanese in Tokyo. The provinces and regions are shown on the left.



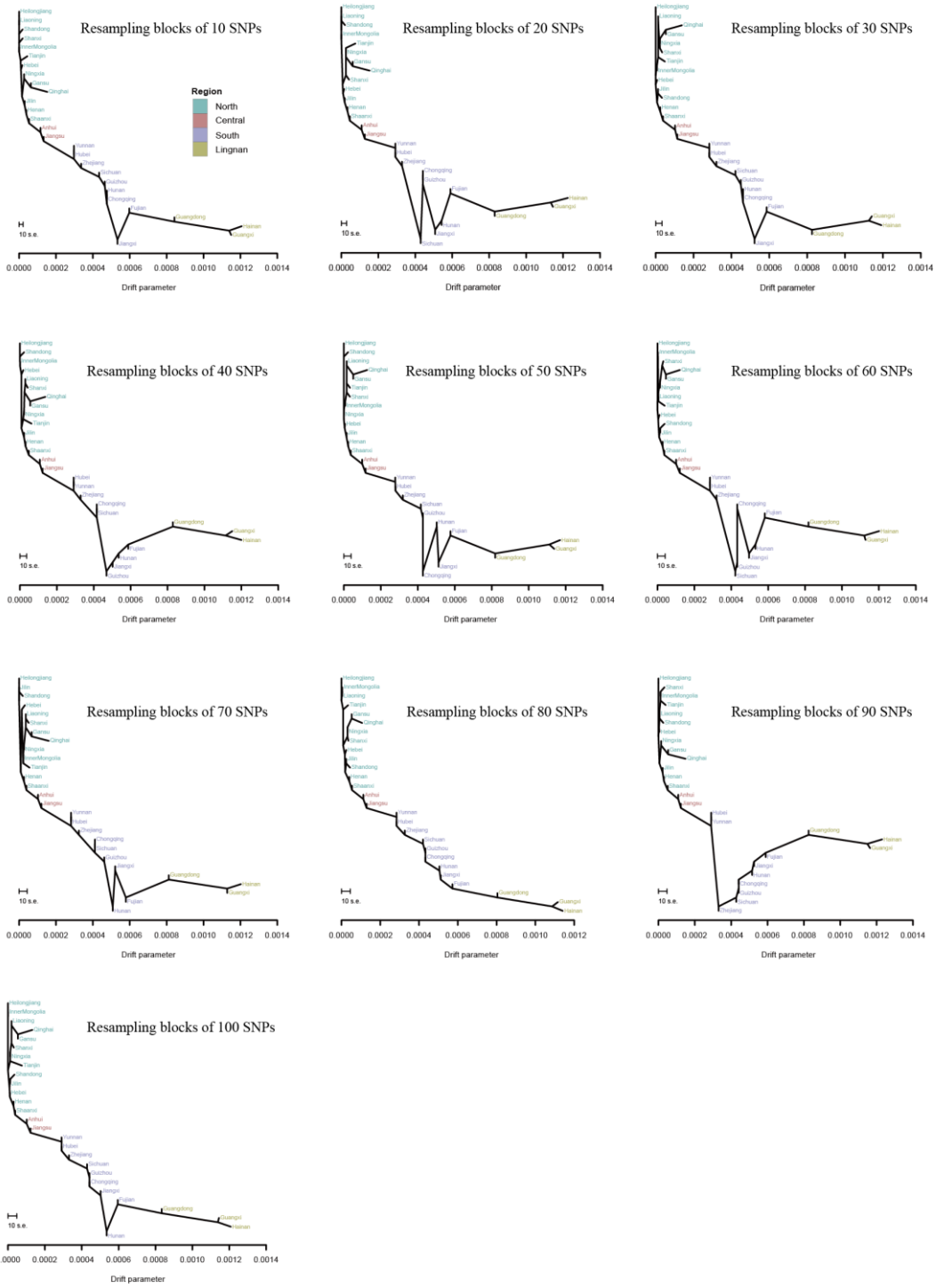
Extended Data Fig. 5 Pairwise F_{ST} between the WBBC and 1KG Project populations.

Numbers in each rectangle mean the pairwise F_{ST} value times 1,000. The bars on the top and left show the population classifications. SAS, AMR, AFR, EUR, and EAS are five continent-level ancestry groups of the 1KG Project.

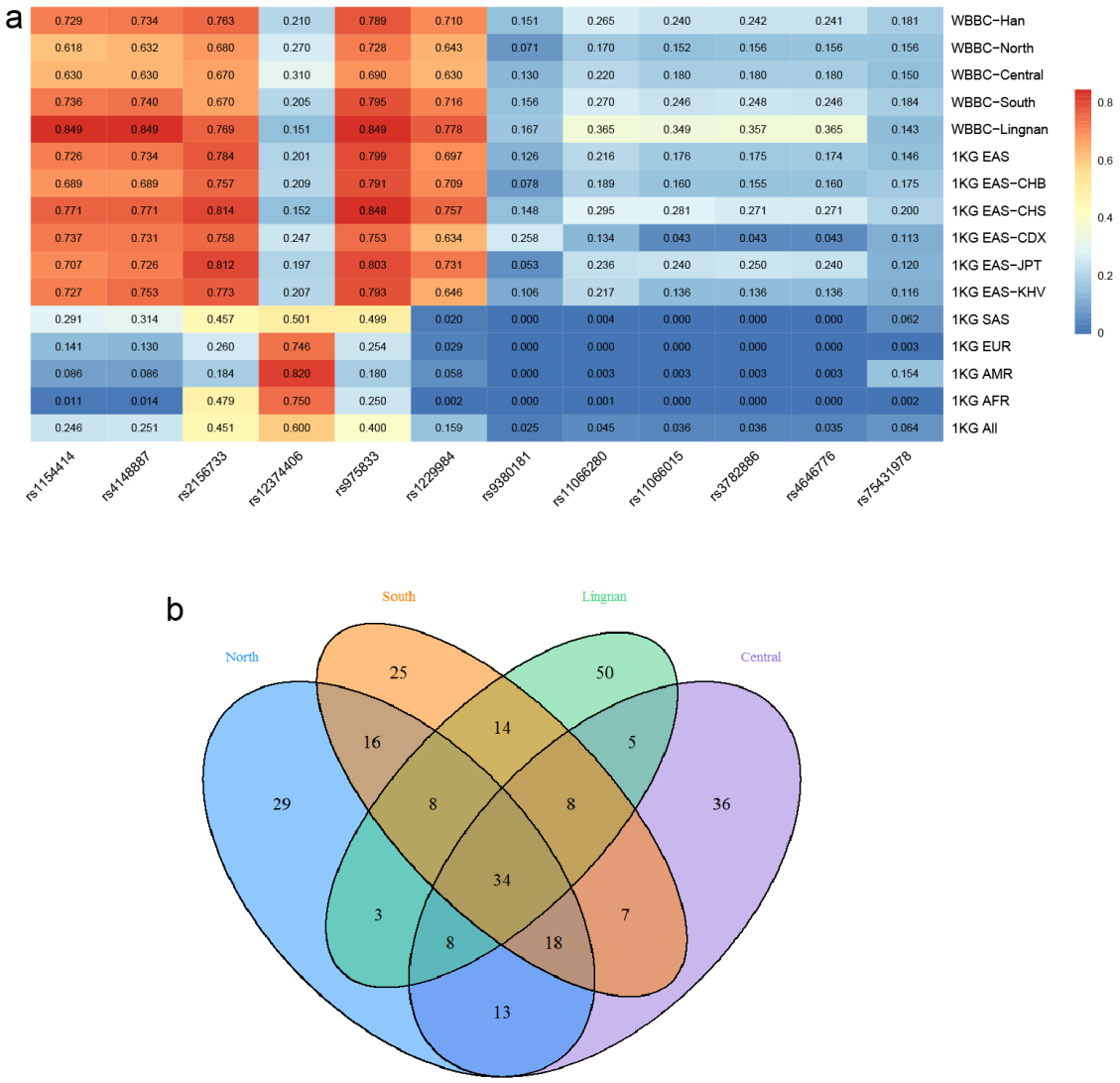


Extended Data Fig. 6 Geographic patterns of pairwise F_{ST} using the 1KG as a reference.

Using the four non-Chinese continent-level ancestry groups of the 1KG Project as the reference, including **a,** AMR, **b,** AFR, **c,** SAS, and **d,** EUR, we further investigated the geographic patterns of F_{ST} in the 27 administrative divisions. Regions in grey were not sampled.

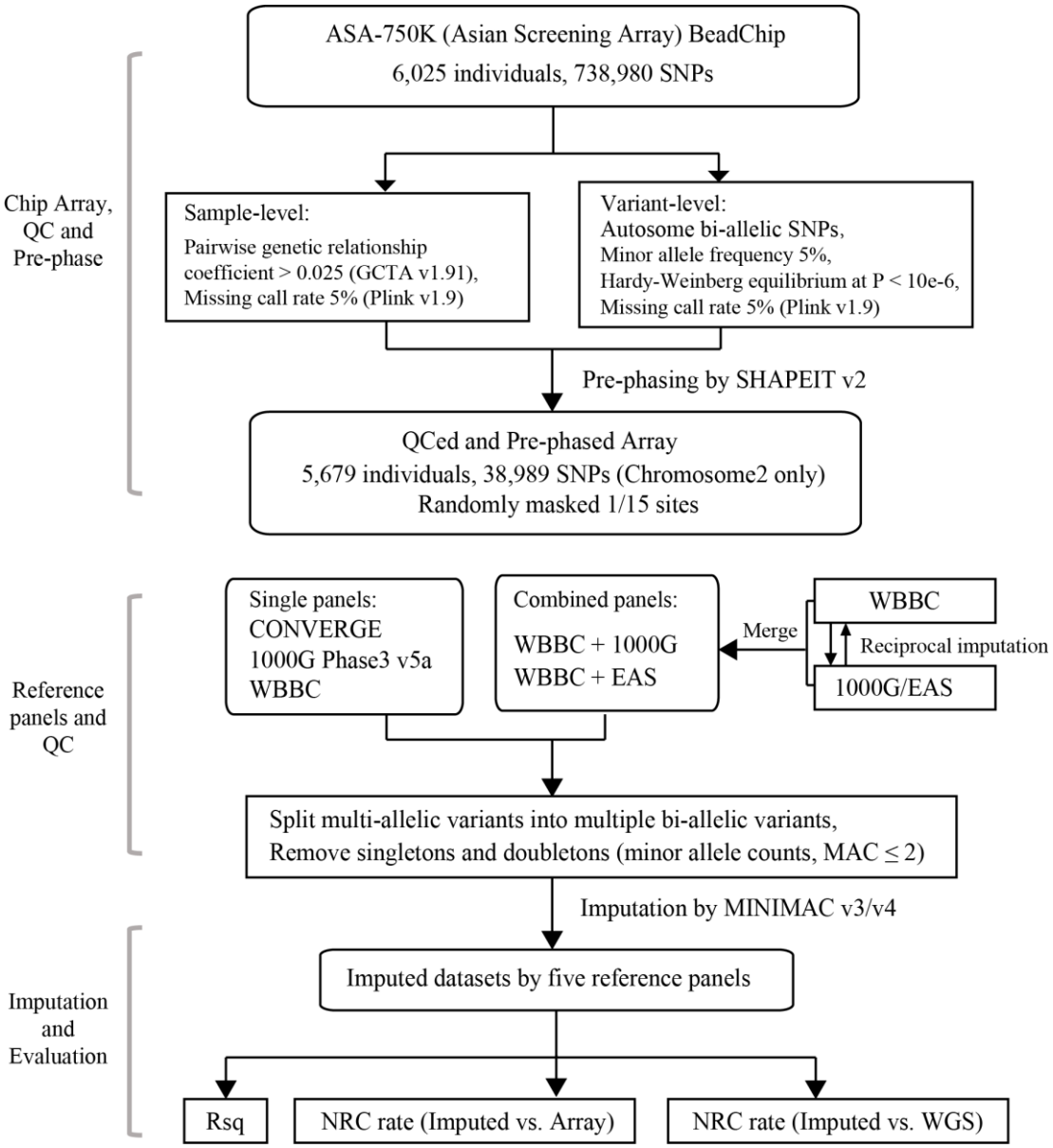


Extended Data Fig. 7 Bootstrap replicates for the tree topology in 27 administrative divisions. The bootstrap replicates were generated by the -bootstrap -k flag of TreeMix software. The plots were y-axis free. The scale bar shows ten times the average standard error of the entries in the sample covariance matrix for the estimated drift parameter.



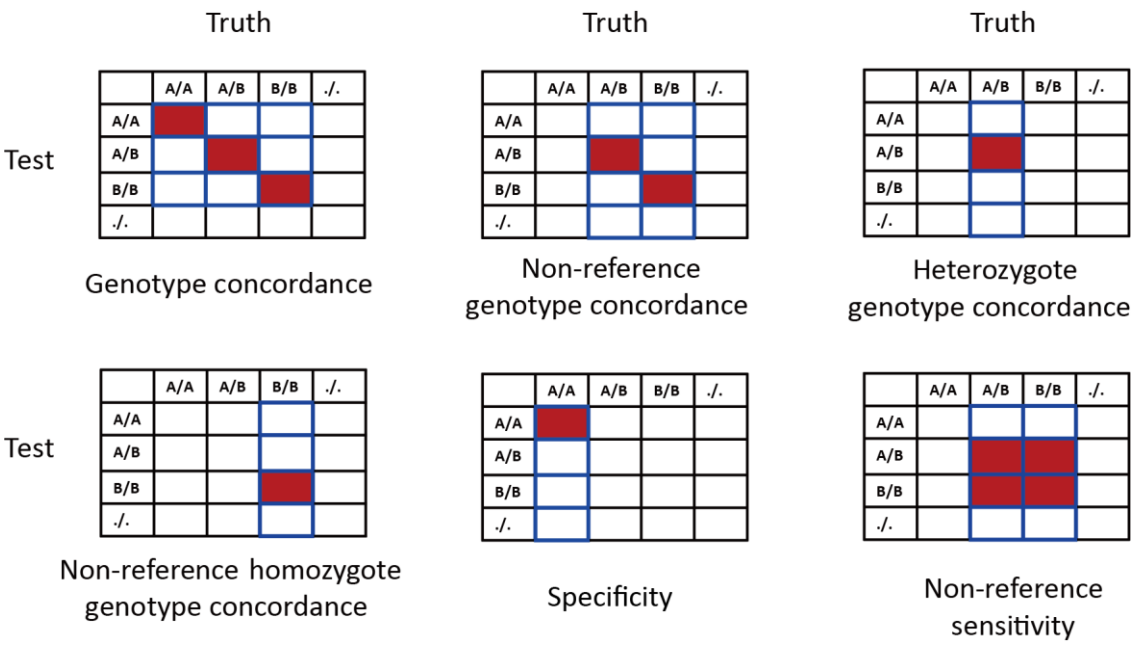
Extended Data Fig. 8 The DAF of SNVs and number of genomic regions with significant selection signatures.

a, The derived allele frequency (DAF) of SNVs with significant selection signatures by SDS in each gene for different populations. The WBBC-Han is all the Han Chinese individuals sequenced by whole-genome sequencing (WGS) in the WBBC cohort. North, Central, South, and Lingnan are the four Han subgroups. EAS, SAS, EUR, AMR and AFR come from the 1000 Genome Project (1KG). **b,** Sharing of genomic regions with higher iHS signals among four subgroups. The number listed within the Venn Diagram indicates the numbers of 200 kb non-overlapping genomic windows in the top 1% of the fraction of SNVs with |iHS| >2.



Extended Data Fig. 9 Imputation evaluation workflow.

Rsq: R-square value calculated by Minimac4. NRC rate: non-reference allele genotype concordance rate. 1000G included 3,284,591 variants and 5,008 haplotypes; CONVERGE included 1,115,342 variants and 23,340 haplotypes. The WBBC included 2,089,508 variants and 8,610 haplotypes. The WBBC+1000G combined panel consisted of 13,618 haplotypes with 4,450,989 variants. The WBBC+EAS combined panel consisted of 9,618 haplotypes with 2,411,382 variants



Extended Data Fig. 10 Definition of genotype concordance, specificity, and sensitivity

Letter A represents reference allele while B represents non-reference (NR) allele. The dot means missing called allele. For each metric, the value equals to the corresponding red rectangles divided by all rectangles with a blue border.

Supplementary Table Legends

Supplementary Table 1. The geographic distribution of all the WBBC cohort samples. The individuals were recruited from 29 of the 34 administrative divisions of the People's Republic of China (PRC).

Supplementary Table 2. The summary of autosomal variants in each individual in the WBBC cohort.

Supplementary Table 3. The statistics of variants in 4,489 individuals.

Supplementary Table 4. The average number of autosomal variants in each genome of four regions.

Supplementary Table 5. The pathogenic and likely pathogenic variants were recorded by Clinvar in the 1,151 healthy individuals.

Supplementary Table 6. Pair-wise Fst values for 27 provinces of the China in WBBC

Supplementary Table 7. Pair-wise Fst values for 26 populations in the 1KG Phase3 and 4 regions of the China in WBBC

Supplementary Table 8. Pair-wise Fst values for 27 provinces of the China in WBBC and 4

continent groups in the 1KG Phase3

Supplementary Table 9. SNVs with significant selection signatures by SDS analysis in the Han Chinese population.

Supplementary Table 10. The top 1% of non-overlapping genomic windows was identified for positive selection in the North Han Chinese population using the iHS statistic. The adjacent regions, including the same genes, were merged. The clusters were sorted by the fraction of SNVs with |iHS| > 2.

Supplementary Table 11. The top 1% of non-overlapping genomic windows was identified for positive selection in the Central Han Chinese population using the iHS statistic. The adjacent regions, including the same genes, were merged. The clusters were sorted by the fraction of SNVs with |iHS| > 2.

Supplementary Table 12. The top 1% of non-overlapping genomic windows was identified for positive selection in the South Han Chinese population using the iHS statistic. The adjacent regions, including the same genes, were merged. The clusters were sorted by the fraction of SNVs with |iHS| > 2.

Supplementary Table 13. The top 1% of non-overlapping genomic windows was identified for positive selection in the Lingnan Han Chinese population using the iHS statistic. The

adjacent regions, including the same genes, were merged. The clusters were sorted by the fraction of SNVs with |iHS| > 2.