The 1000 Chinese Indigenous Pig Genomes Project provides insights into the genomic architecture of pigs

Supplementary Methods

1. Sex determination

We inferred the gender of 1,011 samples using GATK[1], the coverage of X and Y chromosomes divided by the whole genome coverage, and the coverage of the SRY gene. First, the gender of every sample was inferred based on sequencing data by the DetermineGermlineContigPloidy function of GATK. The second method calculated the coverage of X and Y chromosomes divided by the whole genome coverage using mosdepth (v0.3.3)[2]. The relative coverage of (X, Y) of males is expected to be (0.5, 0.5), and that of females is expected to be (1, 0). From **Fig. 1A**, all samples were clustered as two groups without ambiguous samples. The final method was calculating the mean depth coverage in the SRY gene using Samtools. If the mean coverage of depth was larger than 1×, this sample was considered male, while it was regarded as female. The results of these three methods showed high concordance, and they are listed in **Supplementary Data 1**.

2. The identification of easy- and difficult-to-sequence regions of the genome

We identified the easy- and difficult-to-sequence regions of the reference genome using a previously reported method[3]. Difficult regions included: (1) tandem repeats and homopolymers longer than 6 bp (~67% of difficult regions), (2) segmental duplications (~13% of difficult regions), (iii) low (<25%) and high (>65%) GC content regions (~40% of difficult regions), and (4) regions with low mappability (~31% of difficult regions). Any regions in the Sscrofa11.1 reference genome that did not fall into a difficult region were classified as easy. The number of small variants that fall into the easy- and difficult-to-sequence regions were calculated, respectively.

3. The SNP number and heterozygous/homozygous ratios across different breeds

The distribution of individual SNPs indicated significant breeding and geographical characteristics. Most breeds in southern and southwestern China, such as Wuzhishan pigs, Saba pigs, and Tibetan wild boars, had a higher average number of SNPs per individual than other pig breeds. Additionally, breeds in northeast China (e.g., north of the Qinling-Huaihe line) exhibited an overall decreased level of SNPs (**Supplementary Fig. 7A**). However, the heterozygous/homozygous ratios did not correspond to the individual distributions; the heterozygous/homozygous ratio in the Licha Black pigs was much higher than average (**Supplementary Fig. 7B**), and some breeds in the south and southwest China had lower than average heterozygous/homozygous ratios, such as Wujin and Diannan Small-ear pigs. These results revealed the genetic characteristics, diversity, and complexity of the indigenous pig breeds in large geographical areas across China.

4. The relationship between the detected number of small variants and the count of samples in the dataset

To evaluate the effect of increasing the sample size on variant discovery, we randomly

downsampled the 1KCIGP dataset to different sizes and estimated the total number and variant increase at different sample sizes (**Supplementary Fig. 8**). We found that the number of SNPs and indels continued to increase with increasing sample size. However, the growth rate decreased from an initial average increase of 33,806 and 8,797 per sample to final averages of 2,512 and 1,227 for SNPs and indels, respectively (**Supplementary Fig. 8**).

5. Annotation of genomic variants

The detected small variants and SVs were annotated by ANNOVAR (v2020-06-07)[4] using Ensembl release 107. The deleterious variants annotated by SIFT[5]. The function of SVs was predicted using SnpEff (v4.3t)[6].

6. The imputation accuracy estimation

In this study, we demonstrate the efficacy of our panel, 1KCIGP, by evaluating its imputation accuracy using various pig datasets. The test datasets encompassed 113 pigs from 17 Chinese domestic pig breeds. To enhance the test data's diversity, we incorporated 100 pigs from developed breeds, 30 from European domestic breeds, and 19 pigs from the crossbred population between Chinese domestic pig breeds and European domestic breeds (**Supplementary Data 3**). All pigs from the test datasets not contained in the five panels were used in this study (1KCIGP, Animal-ImputeDB, SWIM, Tong's reference panel, and PHARP). Animal-ImputeDB[7] panel contained 233 samples, SWIM panel[8] contained 2,259 individuals, Tong's reference panel[9] consisted of 1,095 samples, while PHARP web server[10] was constructed from 1,006 individuals in the SRA. The SWIM panel web server only supported several commercial SNP array platforms, and only the phased haplotypes from all publicly available individuals (n = 1,241) could access this server.

The test datasets were phased using SHAPEIT4, and imputation was conducted using Minimac4. We used two indices to evaluate imputation performance: (i) the concordance rate, representing the proportion of correctly imputed genotypes in the total imputed genotypes, and (ii) the squared correlation between the imputed allele dosages and the true typed genotypes. These tests were performed to evaluate genotype imputation performance, each involving ten replicates.

To assess the imputation performance on SNPs detected by low-coverage WGS data, we randomly extracted approximately 1× reads from each test individual and detected SNPs using these reads combined with GATK. An average of 1,038,761 SNPs were detected in these simulated reads. Using the actual reads of the 262 test samples, the detected SNPs in all panels were selected for direct comparison among different panels, and their genotypes were considered benchmarks. A total of 9.4 million SNPs and their genotype were chosen as benchmarks. After imputation these 1,038,761 SNPs, the two indices of imputation performance were calculated with all imputed sites.

To estimate the general imputation accuracy for SNP chips, we simulated three popular commercial array genotyping chips (50K, 60K, and 80K): the Illumina GGP Porcine 50K, Illumina PorcineSNP60, and the Illumina PorcineSNP80. The unsampled sites were used to validate the imputation performance.

To maintain consistency in assembly between 1KCIGP (based on the Sscrofa11.1 genome) and Animal-ImputeDB (based on the Sscrofa10.2 genome), we employed liftOver[11] to align the genome coordinates of Animal-ImputeDB from Sscrofa10.2 to Sscrofa11.1.

7. Test for sex-biased hybridization on chromosome Y

Using the 392 boars, we inferred the autosomal, X chromosomal, and combined chromosomal ancestry with ADMIXTURE. Then, we fix the contribution from Europe to the Asian gene pool equal to the autosomal ADMIXTURE estimates, A (A is the sum of the European-related ancestries from K = 6). Following the equations of the previous study[12], we have $A = (m + f)/2$, where m is the male contribution from Europe, and f is the female contribution from Europe. If European contributions are completely male-biased, then $f = 0$ and $m = 2A$. Conversely, if all European contributions are from females, then $m = 0$. The range of m provides an expectation for the range of European Y chromosome frequencies in the Chinese domestic pig populations. Therefore, using the inferred A for each population, we have the expected Y chromosome frequencies range in the population in $[0, 2A]$. Using a binomial test, we can test if the observed number of Y chromosomes in each Chinese population was higher than expected given the range of European male contributions.

8. The KEGG pathway and GO term enrichment analysis

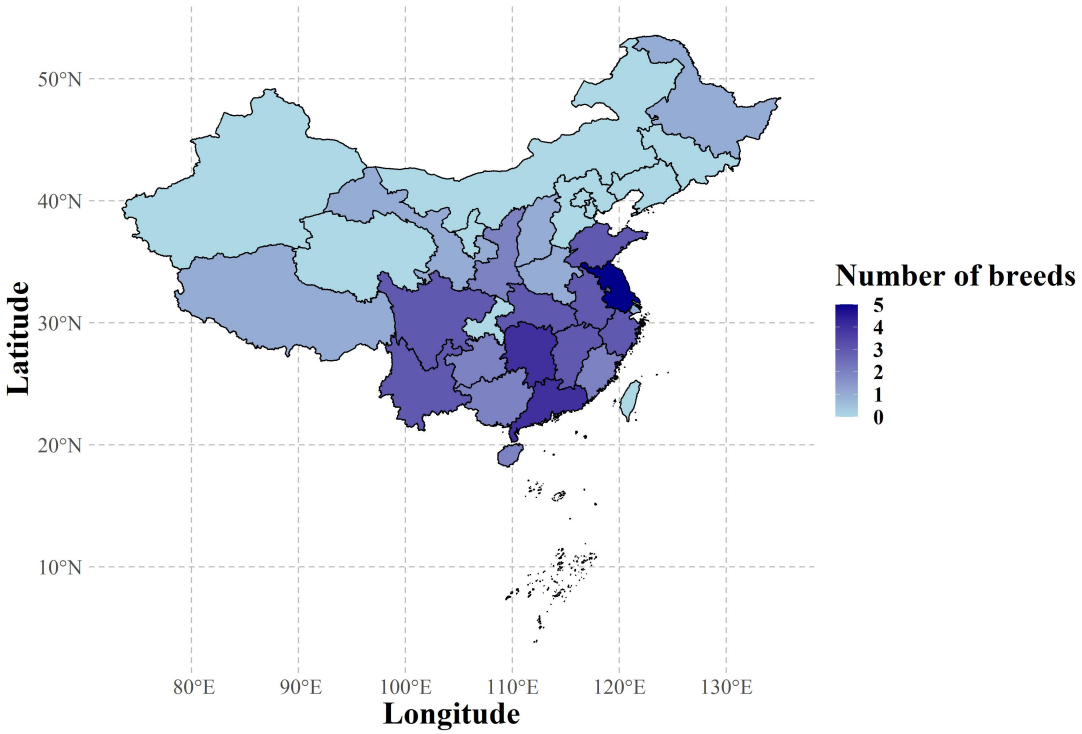
Gene Ontology (GO) terms and KEGG pathway enrichment analyses were carried out through the KOBAS[13].

9. GWAS analyses of biological traits

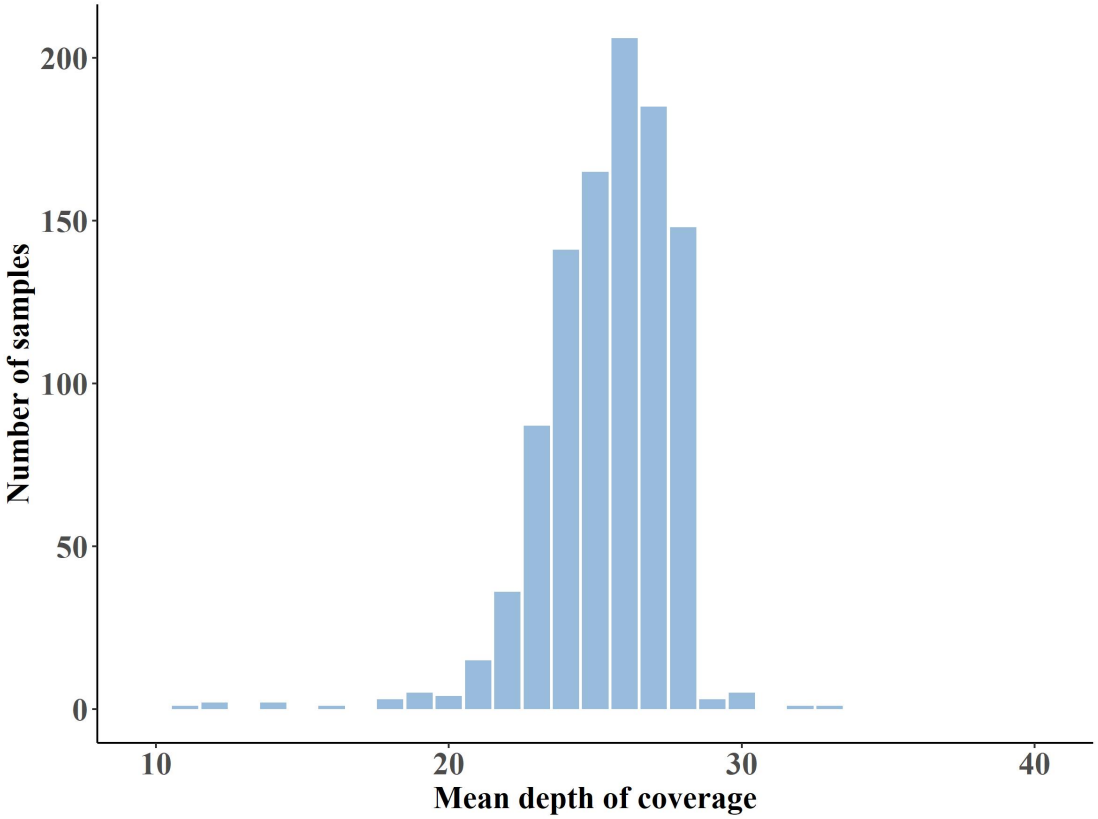
For GWAS analyses focused on two biological traits, we used the mixed linear model to conduct GWAS using the following equation:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Y}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{e},$$

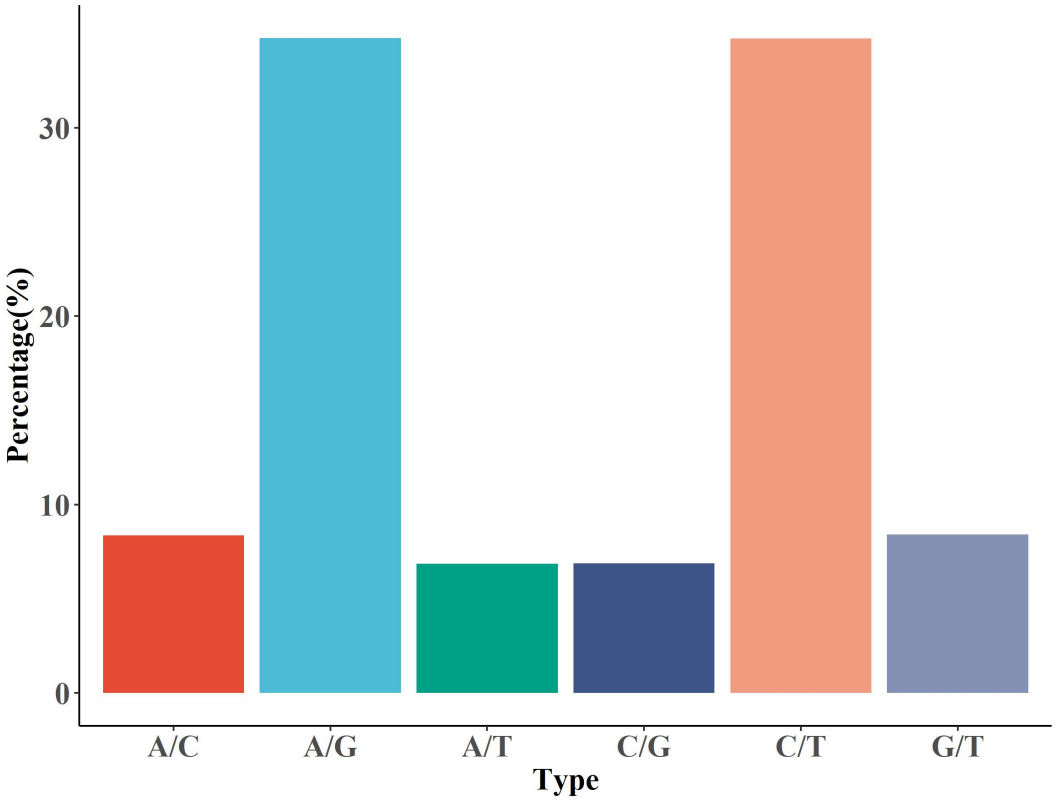
where $\boldsymbol{y}$ referred to the vector of phenotypic values, $\boldsymbol{\alpha}$ represented the vector of SNP effects, which was coded as 0, 1, and 2, corresponding to the three genotypes AA, AB, and BB, with B being the minor allele. $\boldsymbol{\beta}$ represented the vector of fixed effects, including the first five columns of principal components. For body size traits, the fixed effects additionally incorporated geographical distribution. $\boldsymbol{\gamma}$ represented the vector for residual polygenic effects with an assumed distribution $\boldsymbol{\gamma} \sim N(0, \mathbf{G}\sigma_a^2)$, in which σ_a^2 was the additive genetic variance and $\mathbf{G}$ was the marker inferred kinship matrix. $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ were the incidence matrices related to $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. $\boldsymbol{e}$ was the vector for residual residual errors with a putative distribution $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$, in which $\boldsymbol{I}$ was the identity matrix and σ_e^2 was the residual variance.



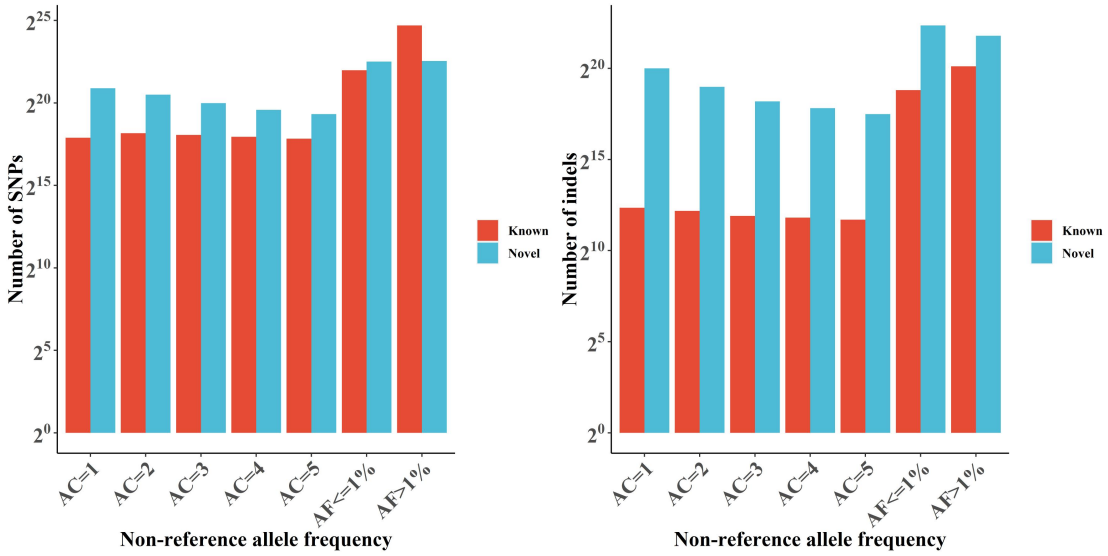
Supplementary Fig. 1. The distribution of collected samples. The shade of the colour in the map corresponds to the number of sampled breeds in each region: darker colours indicate a higher number of breeds sampled from that region, while lighter colours represent regions with fewer sampled breeds. The ggplot2 package invokes the maps package sf to generate this China map. This China map is imported from the National Platform for Common GeoSpatial Information Services (https://www.tianditu.gov.cn/).



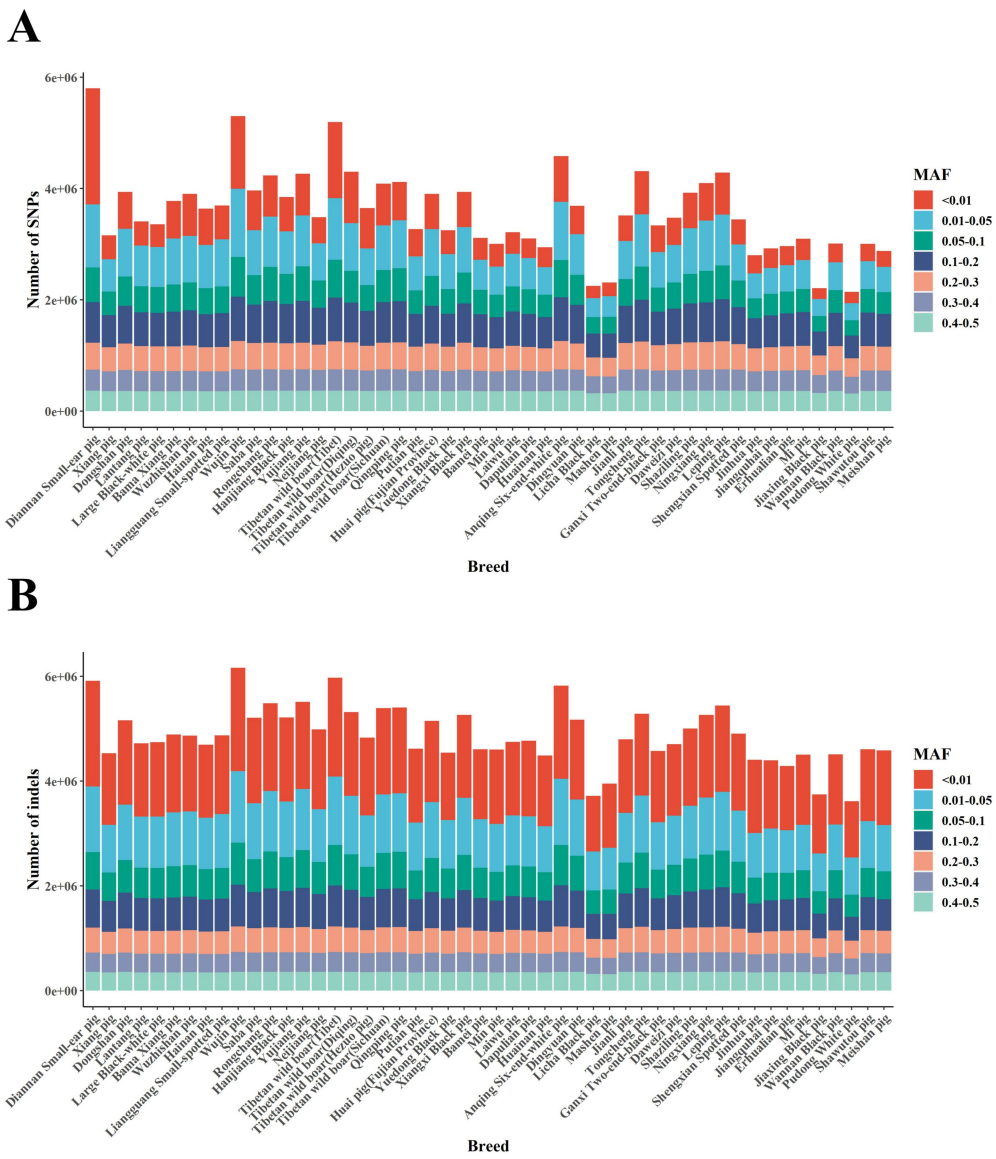
Supplementary Fig. 2. The distribution of the mean sequencing coverage of 1KCIGP individuals.



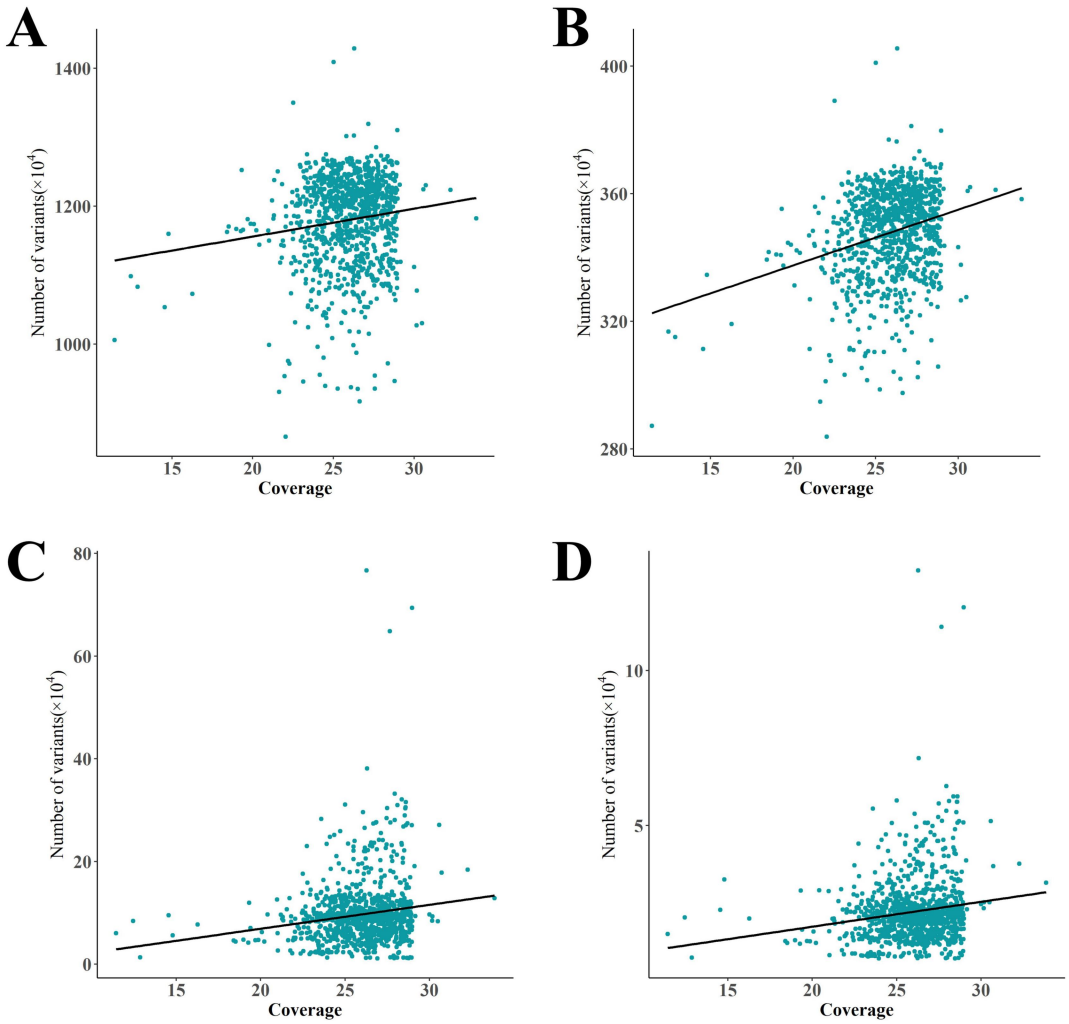
Supplementary Fig. 3. Mutation spectrum of the SNPs.



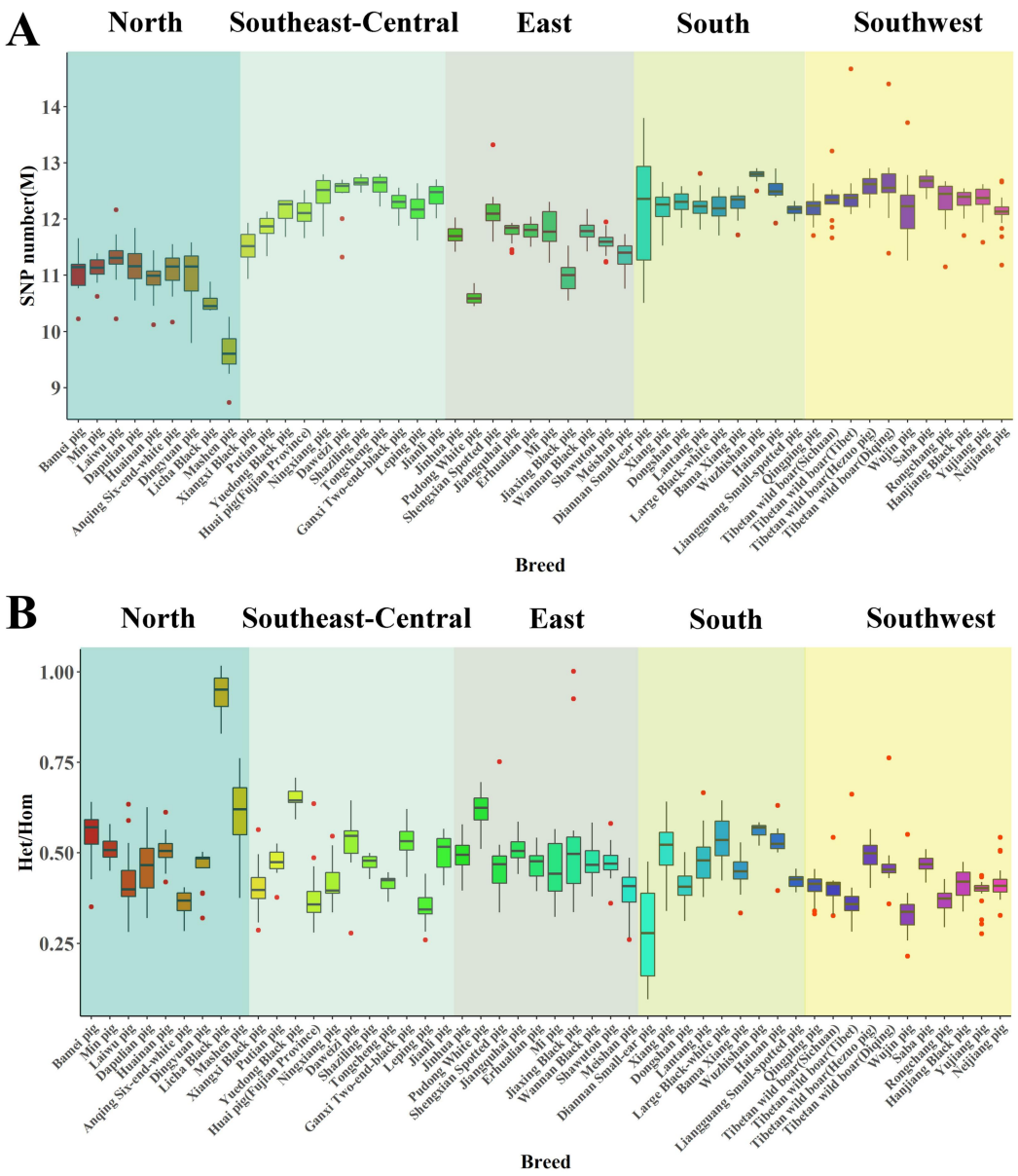
Supplementary Fig. 4. The non-reference allele frequency of SNPs and indels.



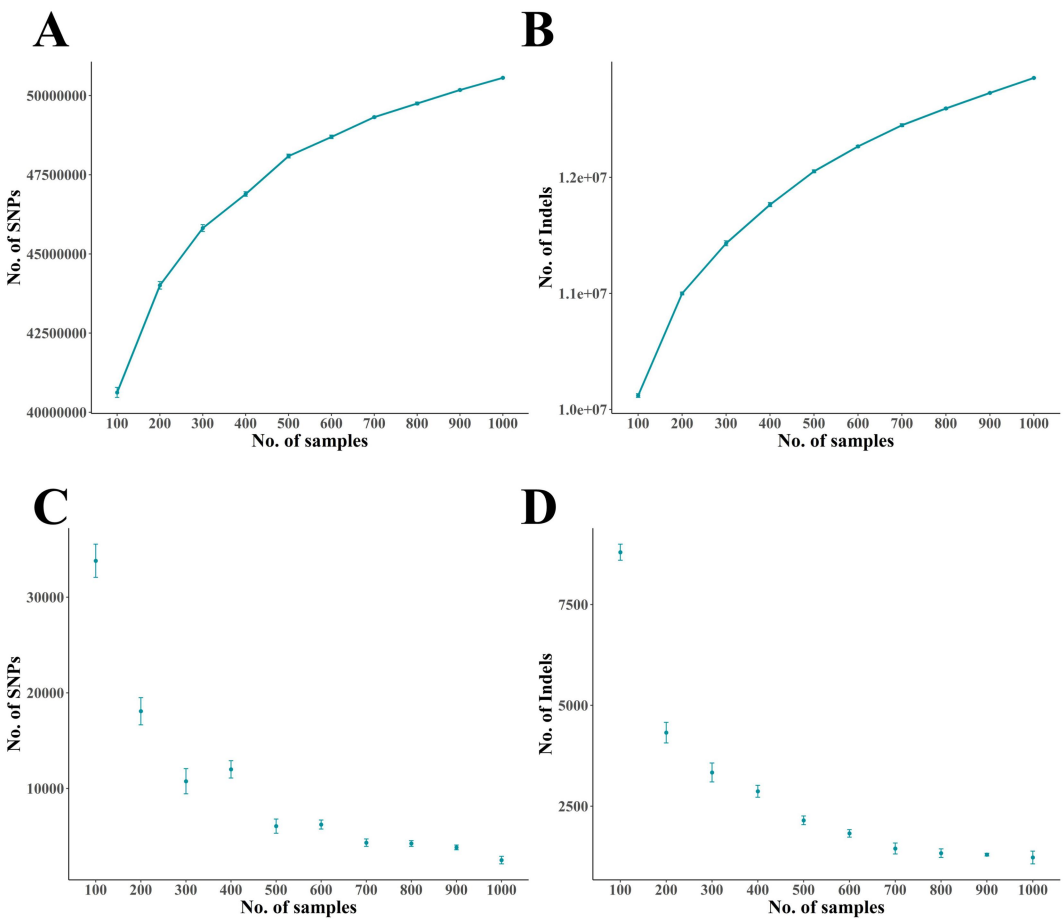
Supplementary Fig. 5. The number of variants detected in each breed for different MAF bins.
A) SNPs, B) indels. Seven MAF bins were divided and represented by different colours.



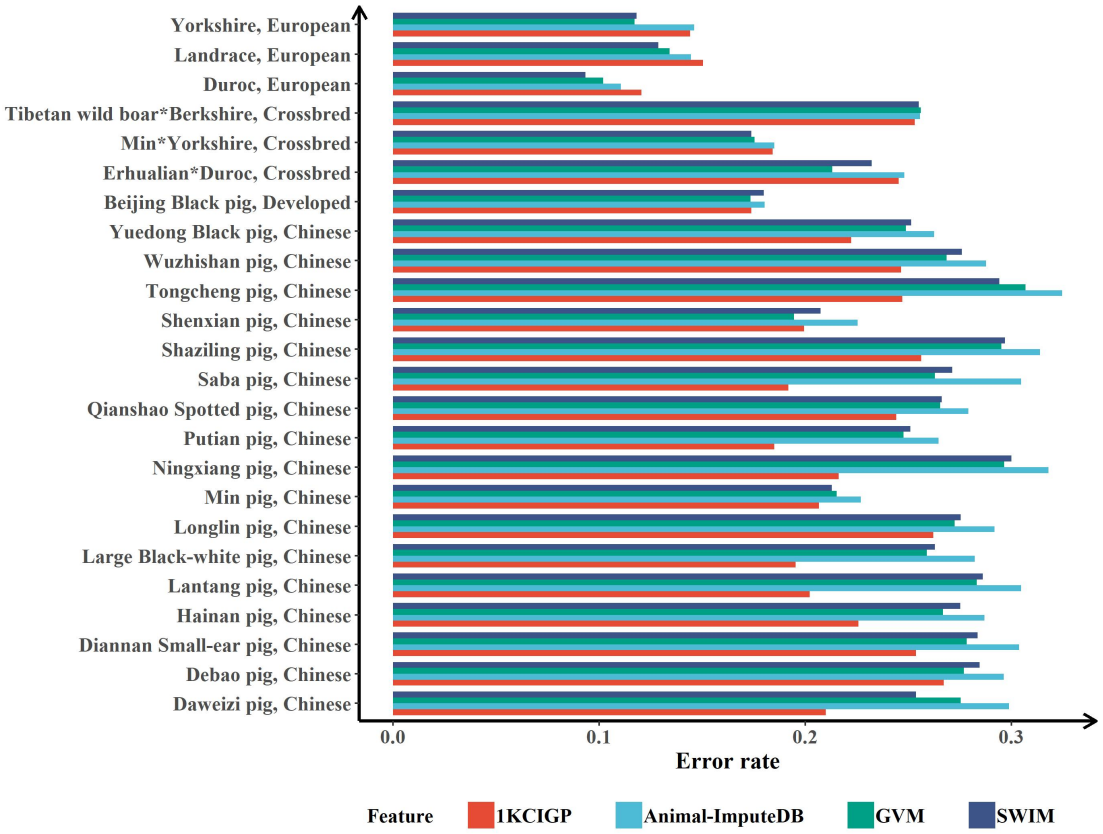
Supplementary Fig. 6. Relationship of genomic coverage (sequencing depth) and number of SNPs or indels. A) and C) indicated the SNPs with MAF > 0.01 or MAF < 0.01 detected in each sample. B) and D) represented the indels with MAF > 0.01 or MAF < 0.01 detected in each sample. Each dot represents the coverage of sequencing depth and genomic variants for each individual.



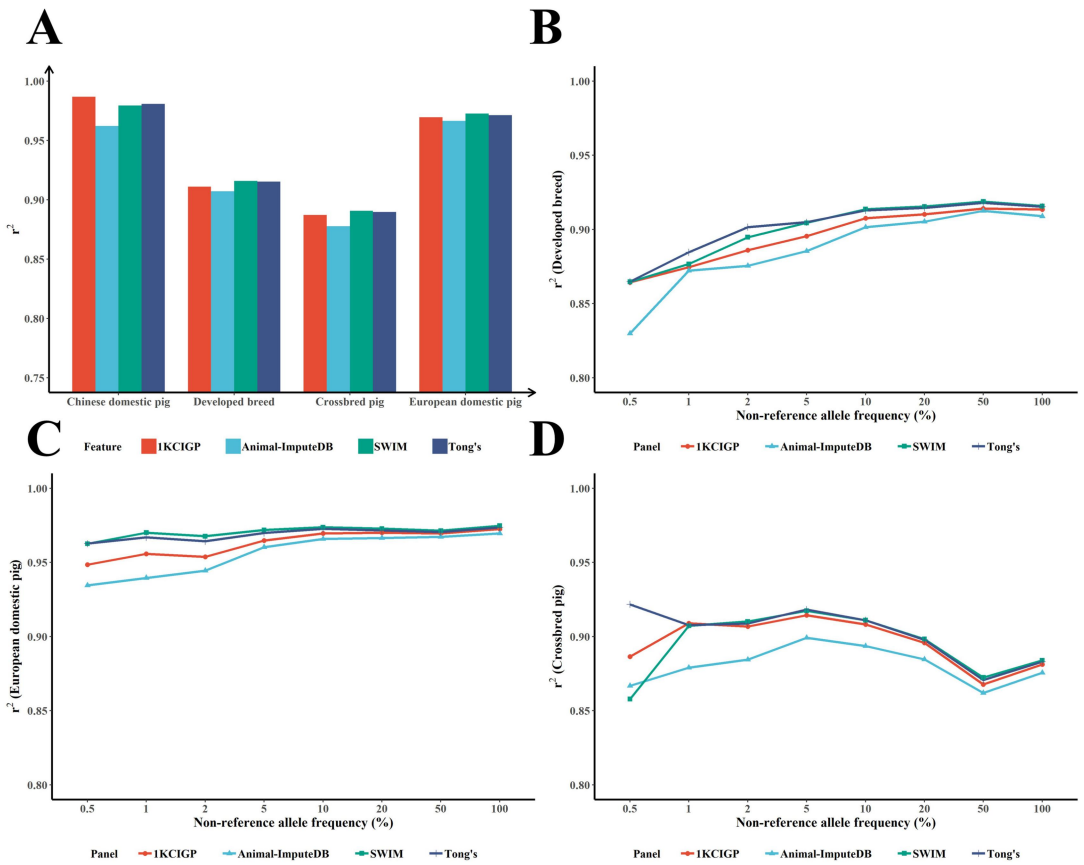
Supplementary Fig. 7. The SNP number and heterozygous SNP ratio in different breeds. A) The SNP number in each breed. The boxplot was generated using the SNP number of each individual in different breeds. B) The heterozygous SNP ratio in each breed. Boxplots show the median, 25th and 75th percentile, the whiskers indicate the minima and maxima, and the points laying outside the whiskers of boxplots represent the outliers.



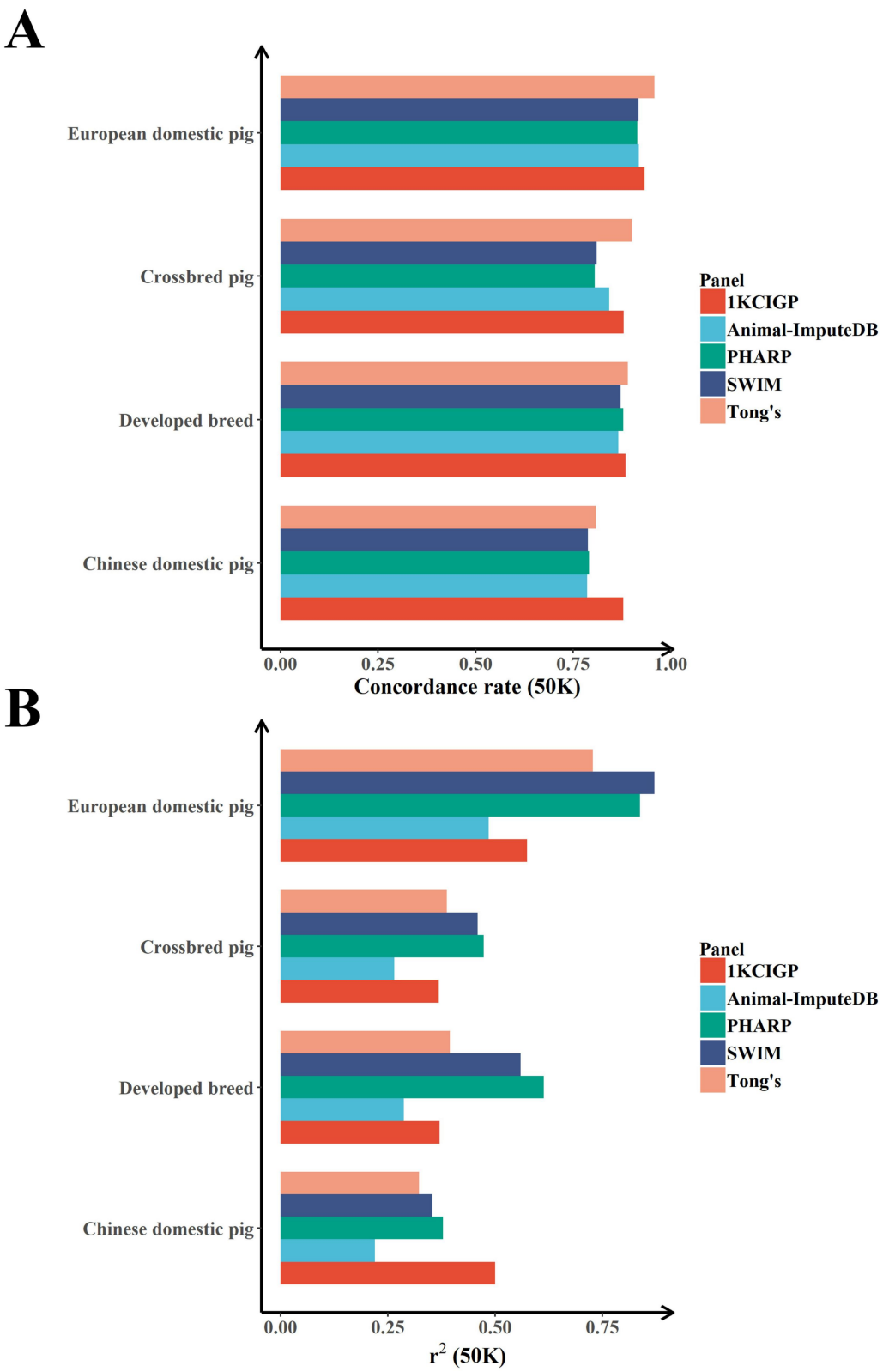
Supplementary Fig. 8. The accumulation of SNPs and indels was estimated by randomly downsampling the 1KCIGP dataset 10 times with intervals of 100 samples. A) Numbers of total SNPs detected related to sample sizes. B) The number of total indels detected is related to sample sizes. C) Numbers of new SNPs were detected by adding new samples for different sample sizes. Points are shown as mean ± SD. D) Numbers of new indels were detected by adding new samples for different sample sizes. Points are shown as mean ± SD.



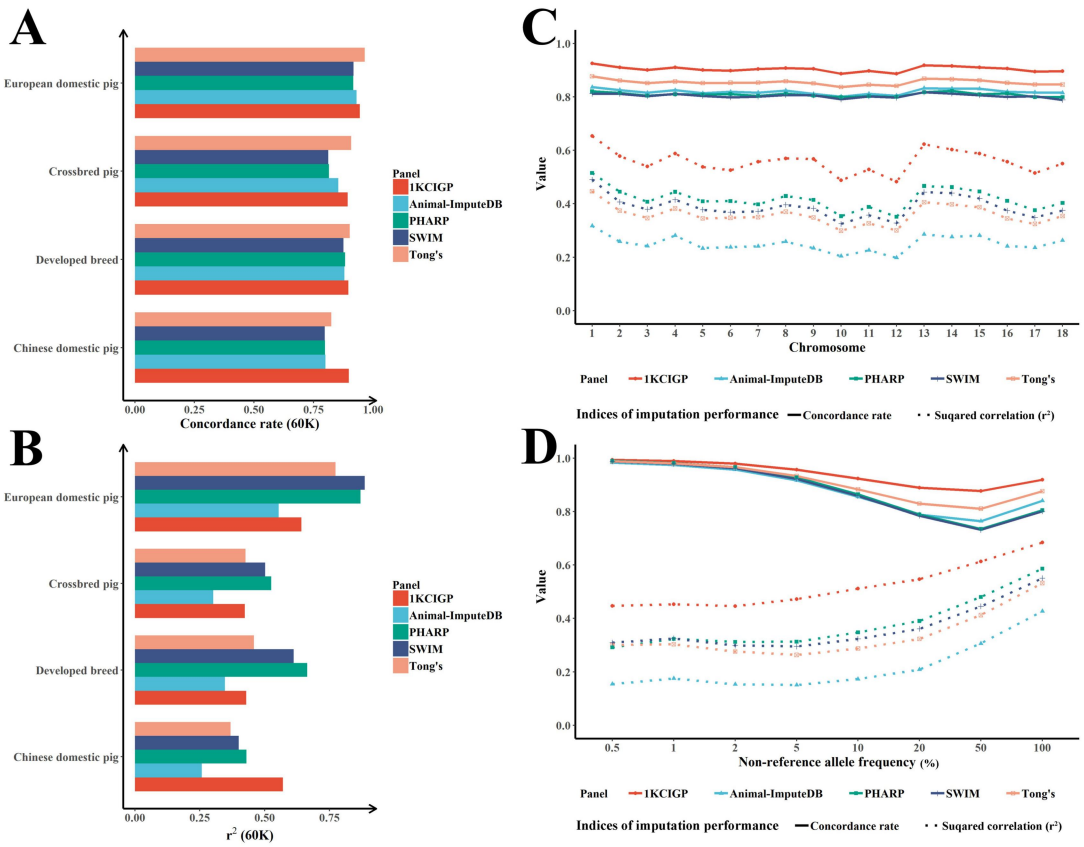
Supplementary Fig. 9. The imputation error rate for four different panels using the 262 test WGS data.



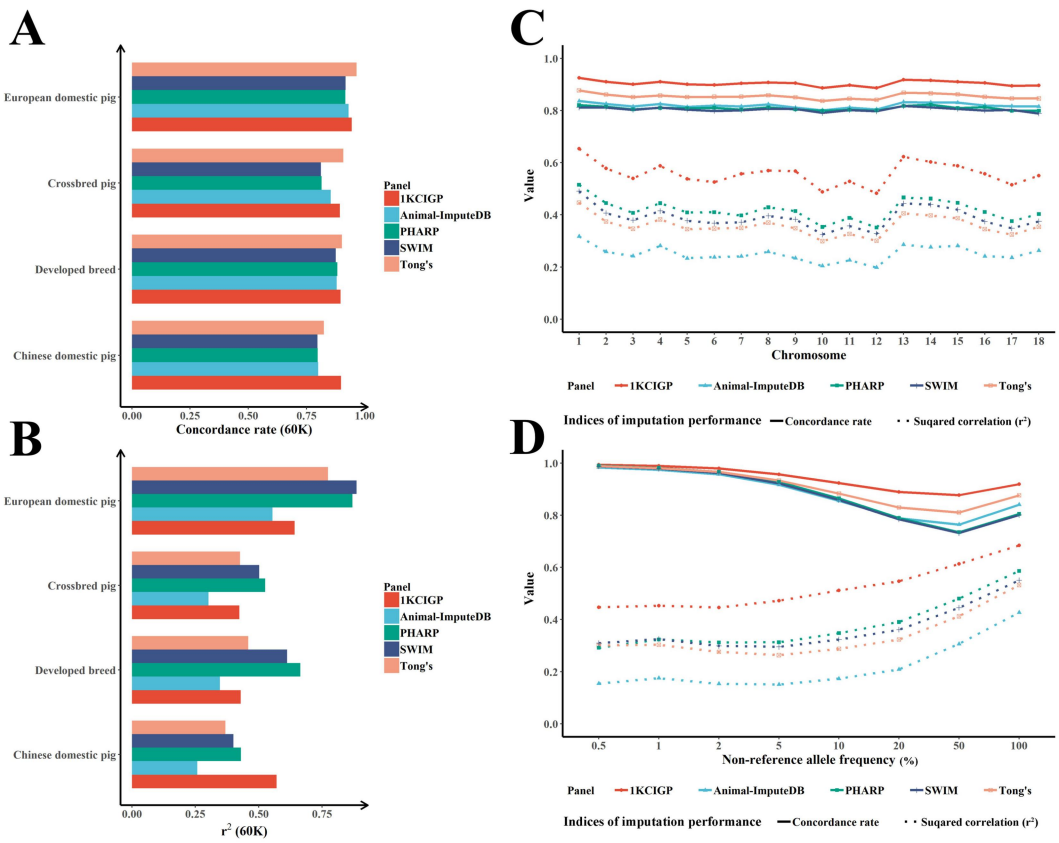
Supplementary Fig. 10. Performance of the 1KCIGP haplotype reference panel in the 262 test WGS data. A) Squared correlation between imputed dosages and true genotypes for different panels. B) Squared correlation between imputed dosages and true genotypes within the stratified non-reference allele frequency bins for developed pig breeds. C) Squared correlation between imputed dosages and true genotypes within the stratified non-reference allele frequency bins for European domestic pigs. D) Squared correlation between imputed dosages and true genotypes within the stratified non-reference allele frequency bins for crossbred pigs.



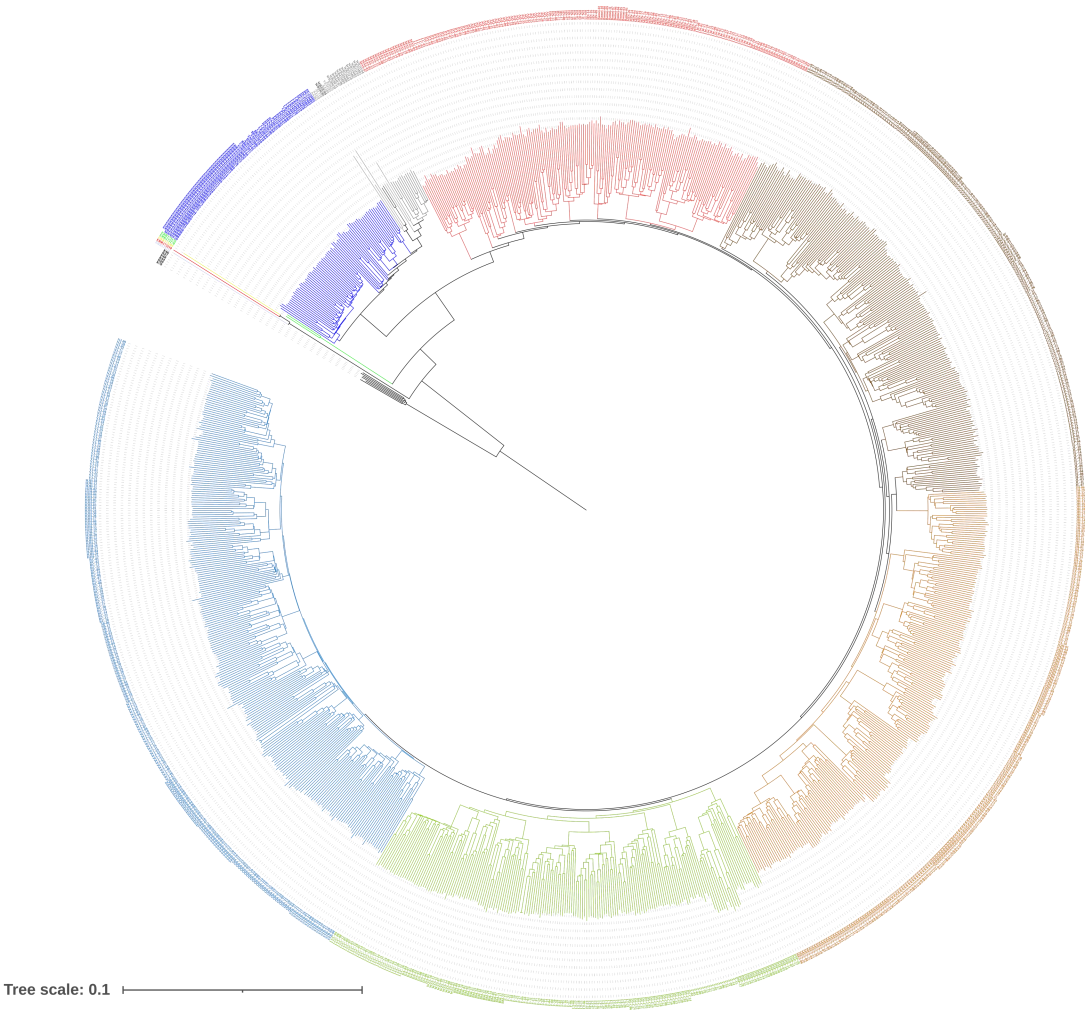
Supplementary Fig. 11. Performance of the 1KCIGP haplotype reference panel in the simulated commercial 50K SNP chip. A) Imputation concordance rate. B) Squared correlation between imputed dosages and true genotypes.



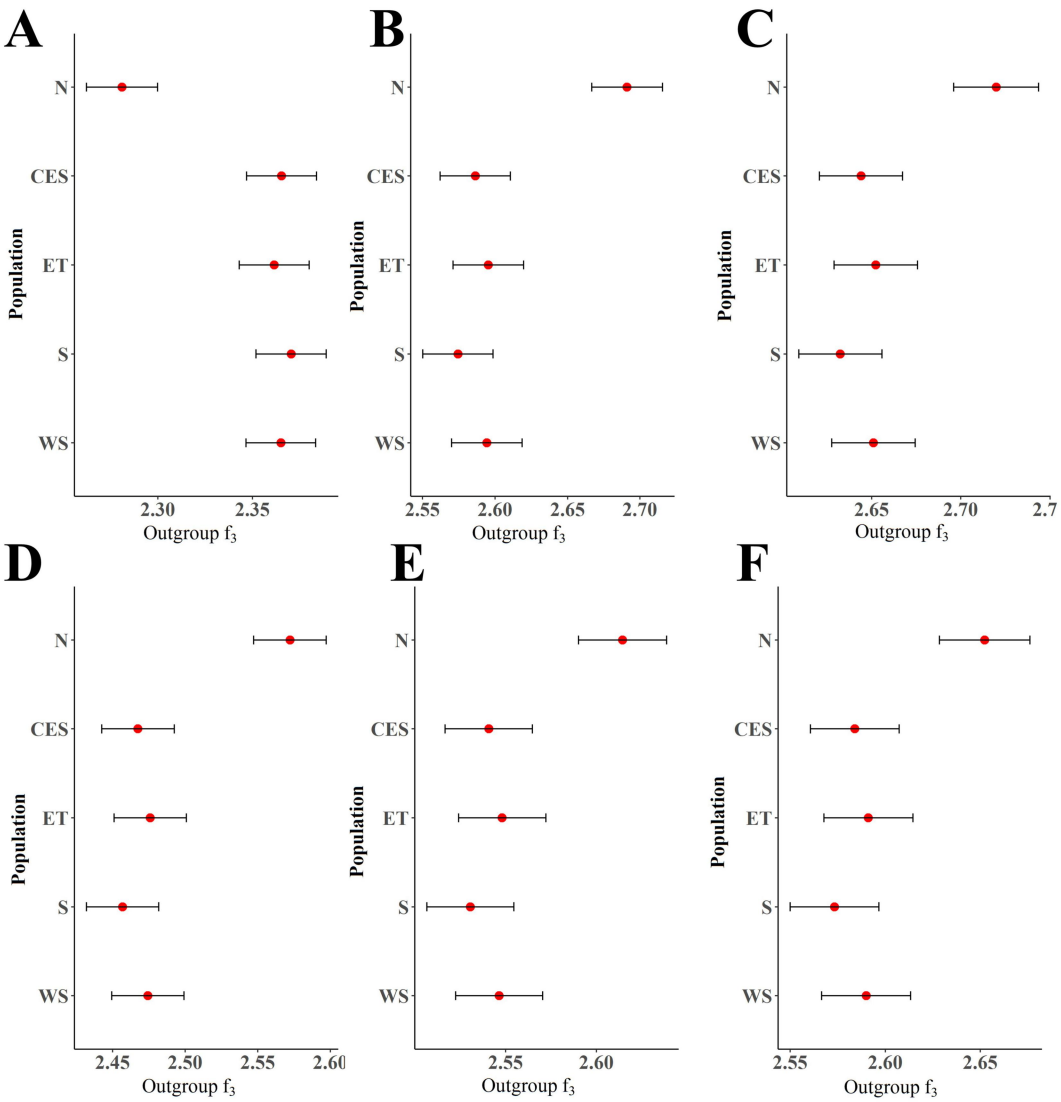
Supplementary Fig. 12. Performance of the 1KCIGP haplotype reference panel in the simulated commercial 60K SNP chip. A) Imputation concordance rate. B) Squared correlation between imputed dosages and true genotypes. C) Imputation accuracy of each chromosome. D) Imputation accuracy within stratified non-reference allele frequency bins.



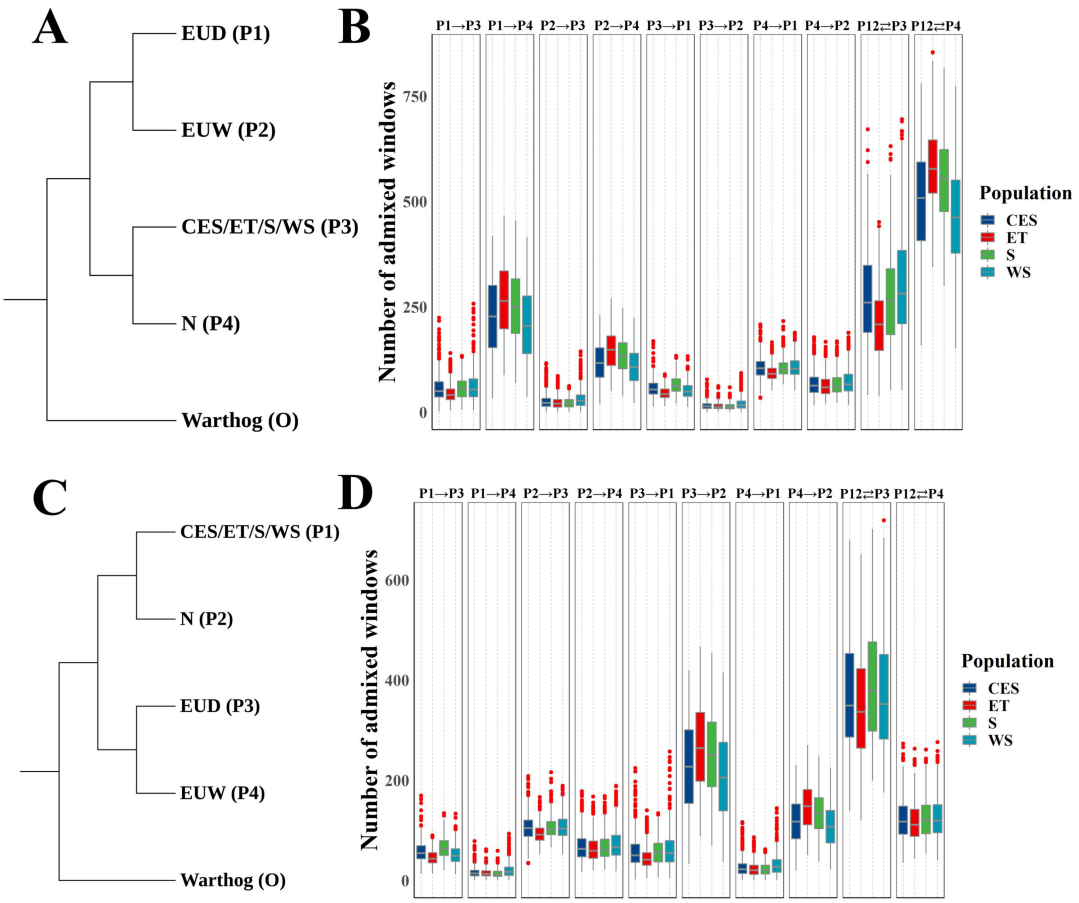
Supplementary Fig. 13. Performance of the 1KCIGP haplotype reference panel in the simulated commercial 80K SNP chip. A) Imputation concordance rate. B) Squared correlation between imputed dosages and true genotypes. C) Imputation accuracy of each chromosome. D) Imputation accuracy within stratified non-reference allele frequency bins.



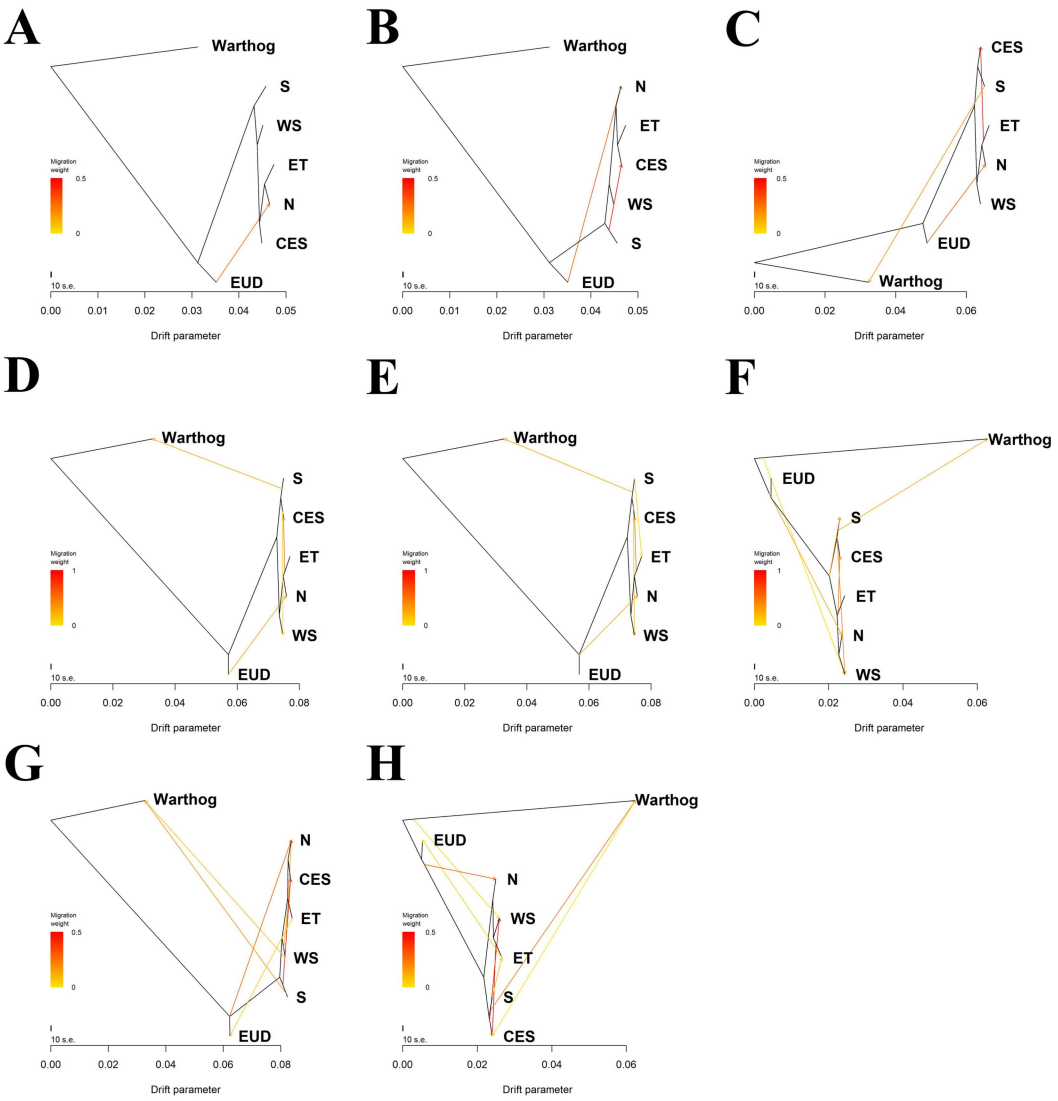
Supplementary Fig. 14. The phylogenetic tree of 1,011 individuals.



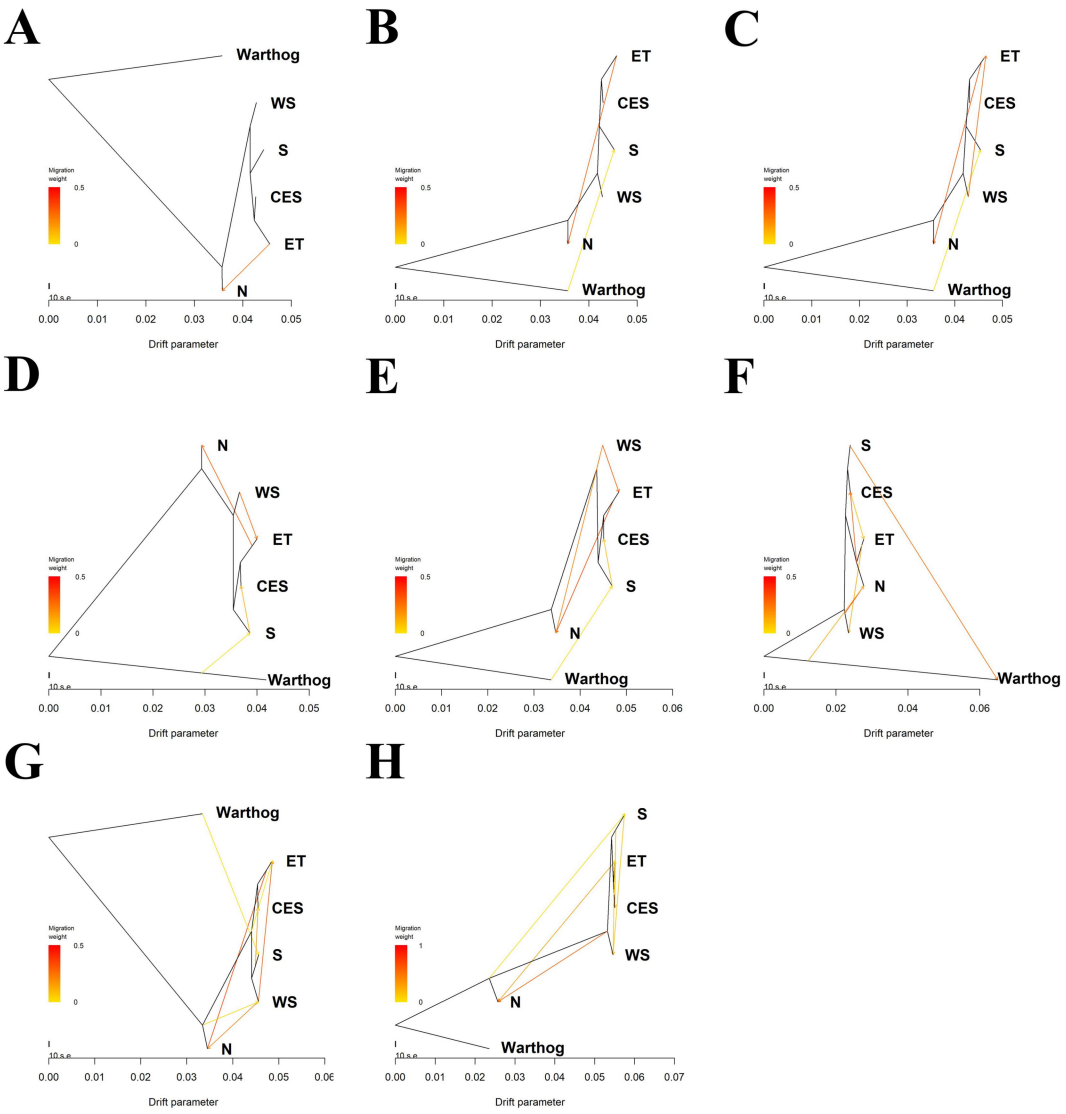
Supplementary Fig. 15. The outgroup f_3 results. Warthog was used as the target. Different plots represented the population excluded from Chinese domestic pigs. A) (*X*, *Sus* from Island Southeast Asia; Warthog). B) (*X*, European wild boar; Warthog). C) (*X*, wild boar from Near East; Warthog). D) (*X*, Ancient European domestic pig; Warthog). E) (*X*, Ancient domestic pig from the Near East; Warthog). F) (*X*, Ancient wild boar from the Near East; Warthog). The individuals of the Chinese domestic pig used in this analysis are the same as in Fig. 3F. The data points are presented as estimated f_3 statistics ± s.e.. The horizontal bars represent ± 1s.e..



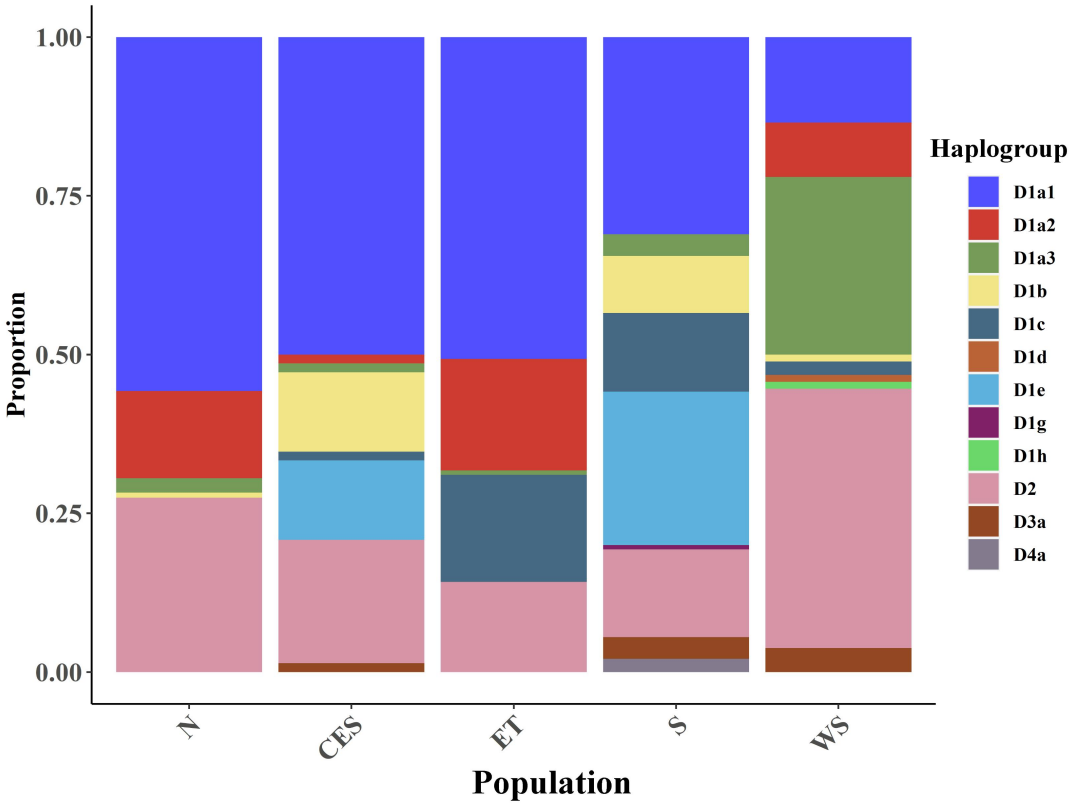
Supplementary Fig. 16. Admixture between Chinese domestic pigs and European pigs estimated using D_{FOIL}. A) Schematic of the topology used for the first configuration of D_{FOIL} analyses applied for population N. B) The D_{FOIL} analyses results of first configuration for population N (n=5×1×11×9 when P3 represents an individual from population CES; n=5×1×10×9 when P3 represents ET; n=5×1×9×9 when P3 represents S; n=5×1×11×9 when P3 represents WS). C) Schematic of the topology used for the second configuration of D_{FOIL} analyses applied for population N. D) The D_{FOIL} analyses results of second configuration for population N (n=5×1×11×9 when P1 represents an individual from population CES; n=5×1×10×9 when P1 represents ET; n=5×1×9×9 when P1 represents S; n=5×1×11×9 when P1 represents WS). Boxplots show the median, 25th and 75th percentile, the whiskers indicate the minima and maxima, and the points laying outside the whiskers of boxplots represent the outliers.



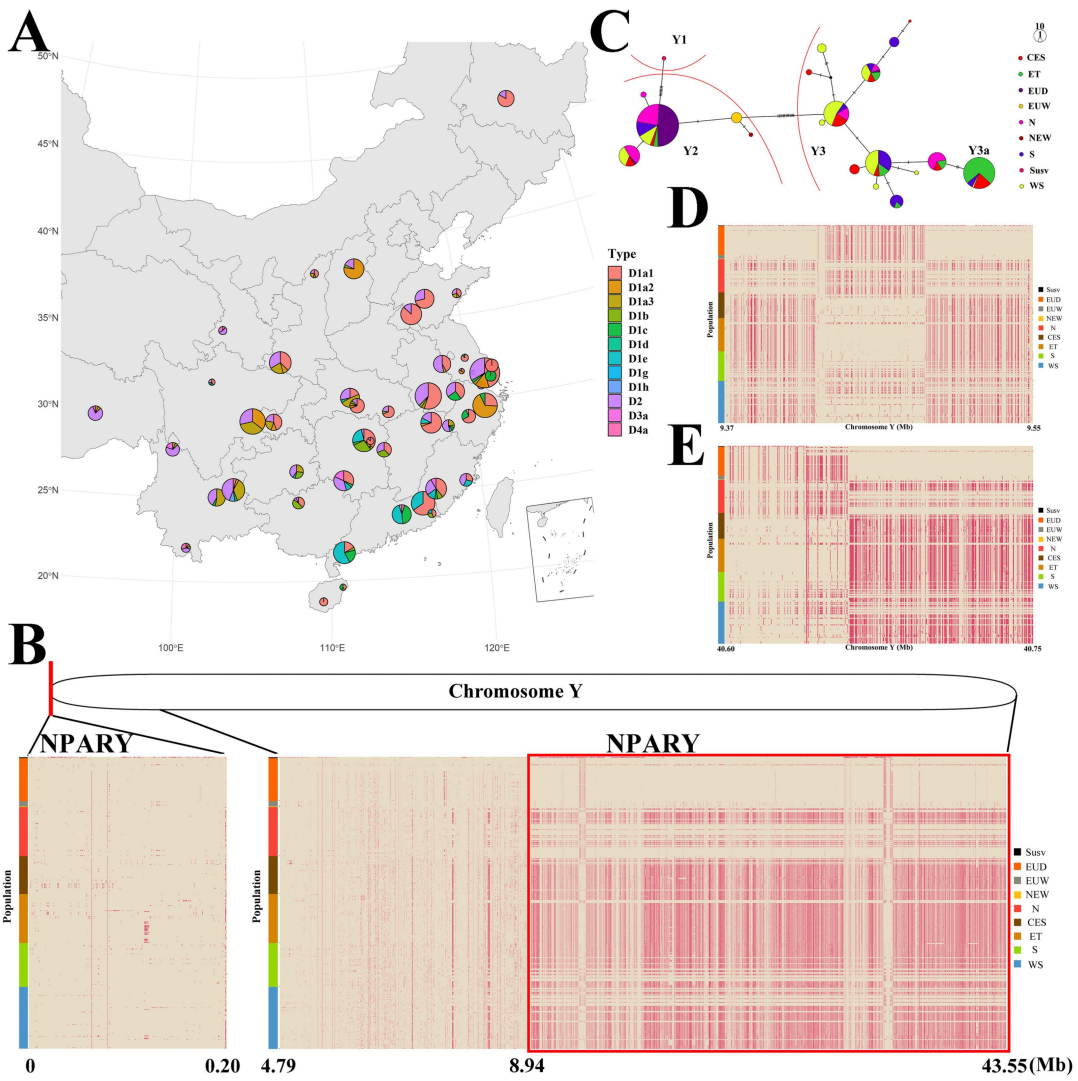
Supplementary Fig. 17. Treemix results of Chinese domestic pig populations and European domestic pigs. Warthog was used as an outgroup. A)-H) represented migrate edges from 1-8.



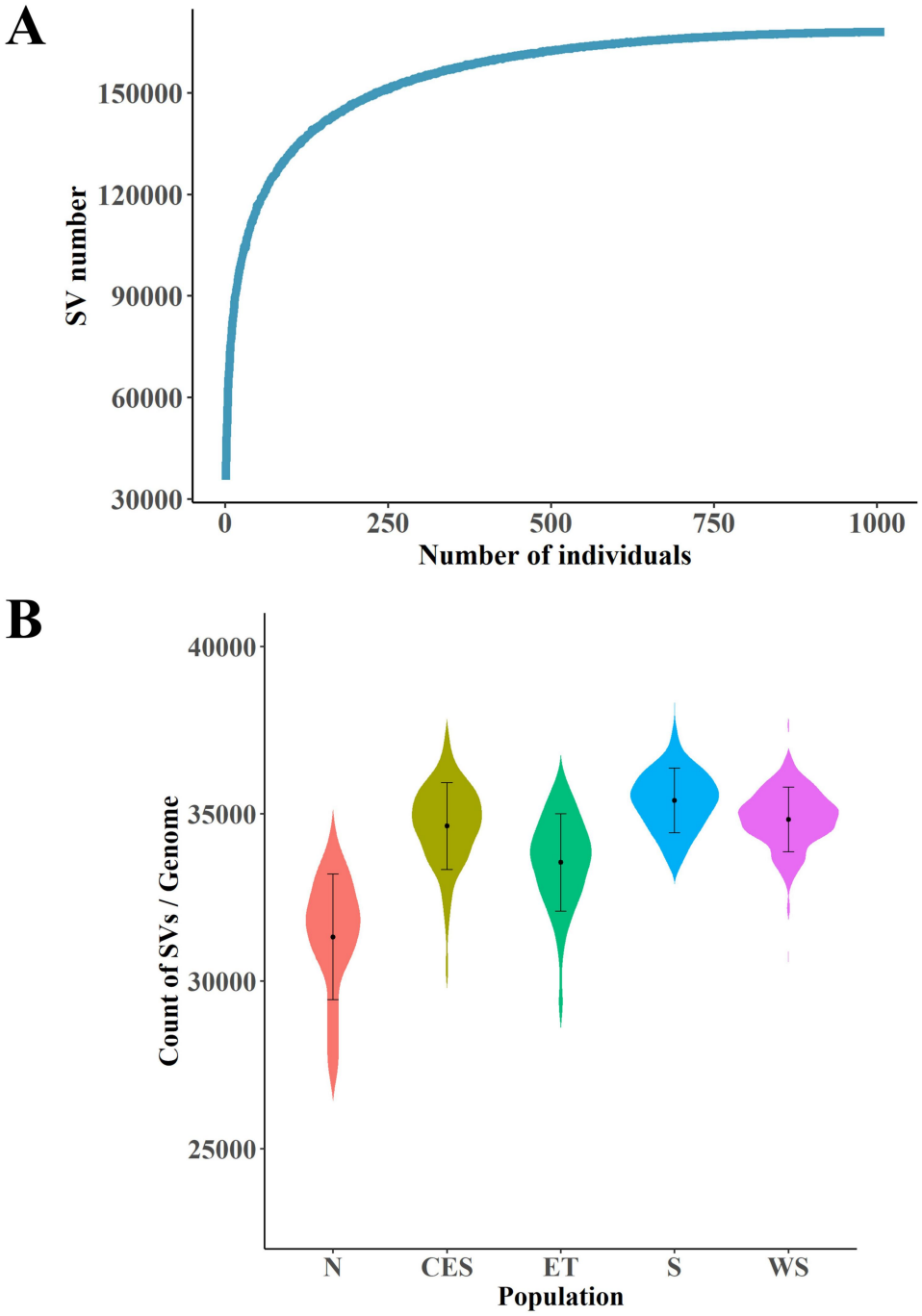
Supplementary Fig. 18. Treemix results of Chinese domestic pig populations. Warthog was used as an outgroup. A)-H) represented migrate edges from 1-8.



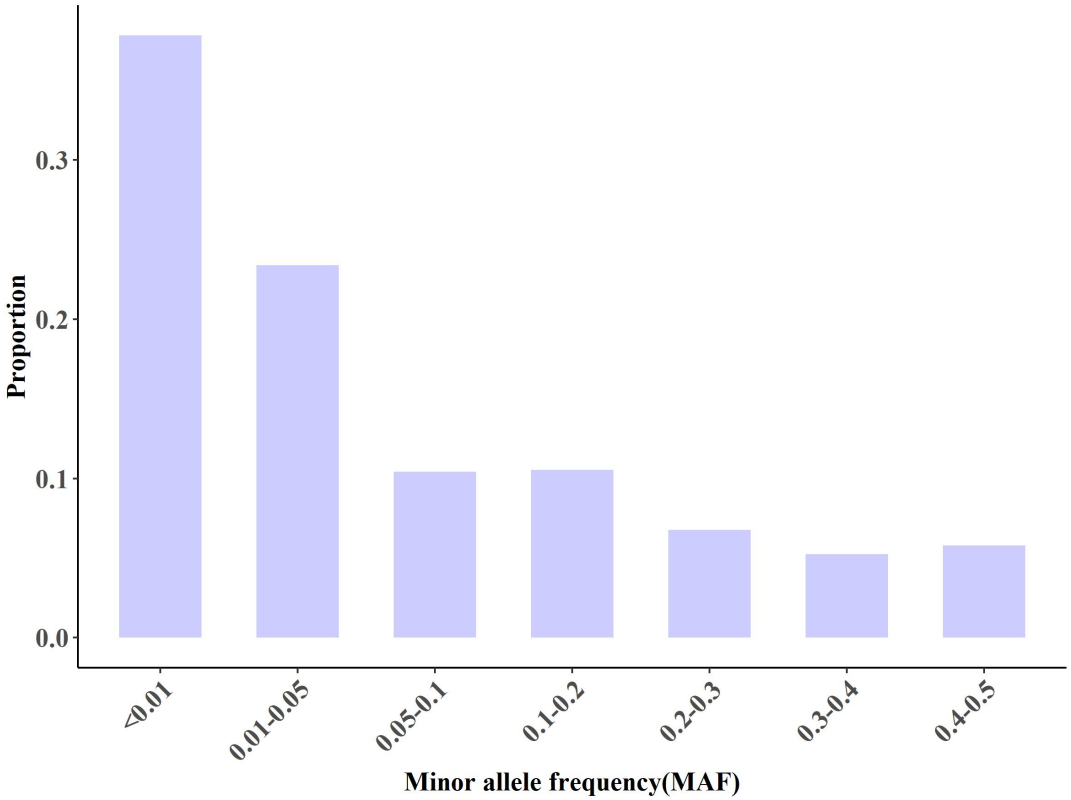
Supplementary Fig. 19. The mitochondrial haplogroups in different Chinese domestic pig populations.



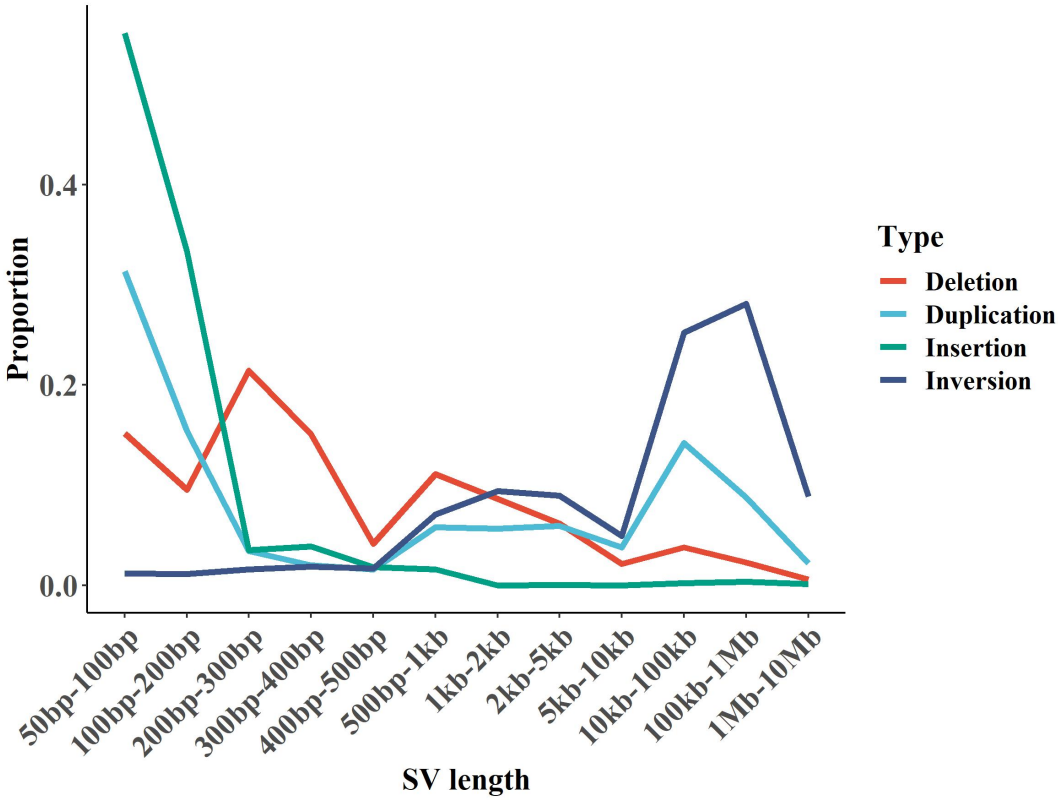
Supplementary Fig. 20. The mitochondrial and Y-chromosome haplogroups in different Chinese domestic pig populations. A) The mitochondrial haplogroups in Chinese domestic pigs. The ggplot2 package invokes the maps package sf to generate a China map. This China map is imported from the National Platform for Common GeoSpatial Information Services (https://www.tianditu.gov.cn/). B) The haplotype patterns on chromosome Y are shared in Eurasian pigs. Alleles that are identical or different from the ones on the reference genome are indicated by creamy white or red, respectively. C) Median-joining (MJ) network of Y-chromosome haplotypes. D) The haplotype patterns on the *LOC100625207* gene. E) The haplotype patterns on the *LOC102159347* gene. In B)-E), Susv: *Sus verrucosus*.



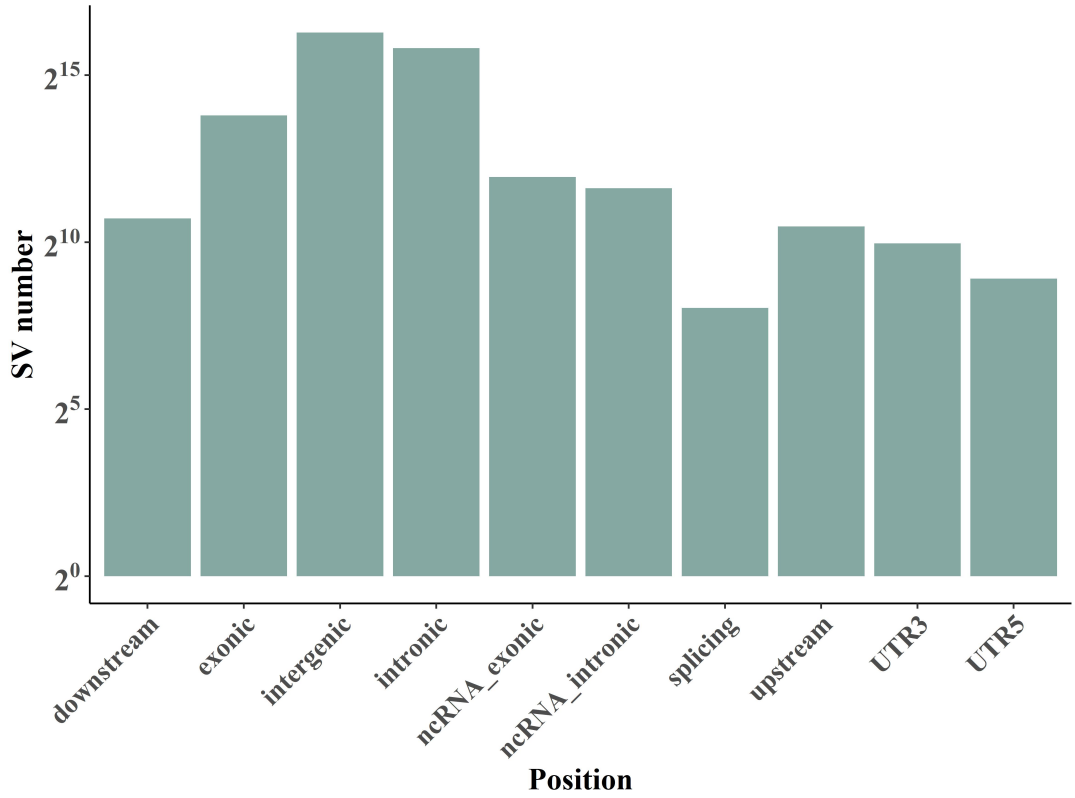
Supplementary Fig. 21. The SVs in 1KCIGP samples. A) The increase in the number of SVs detected with the increase of individuals. B) The SVs numbers in each individual for five distinct Chinese domestic pig sub-groups. The individuals for each sub-group used in this analysis are the same as in Fig. 3F. The points are presented as the mean number of SVs. The horizontal bars represent ± 1SD.



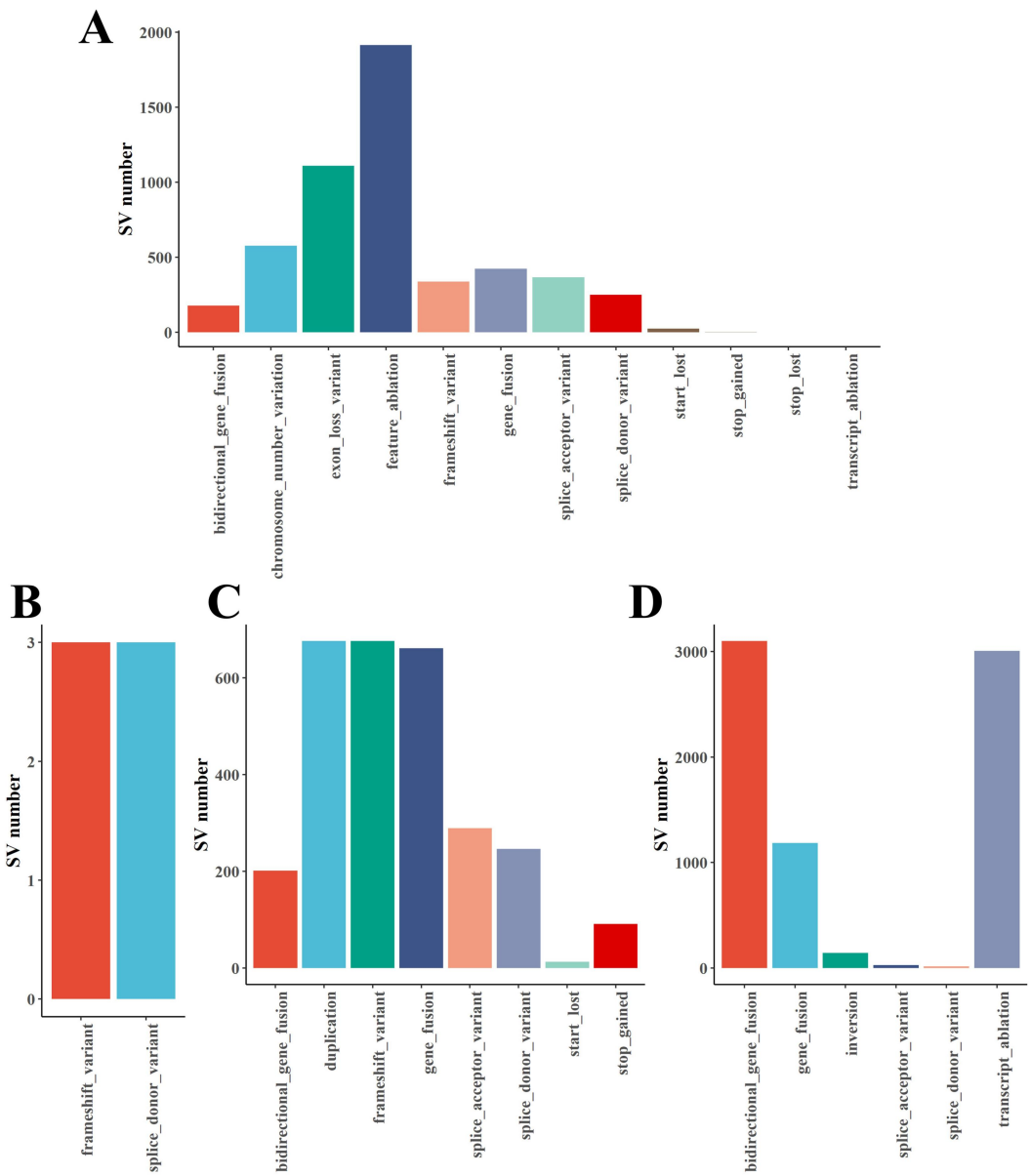
Supplementary Fig. 22. The frequency of SVs.



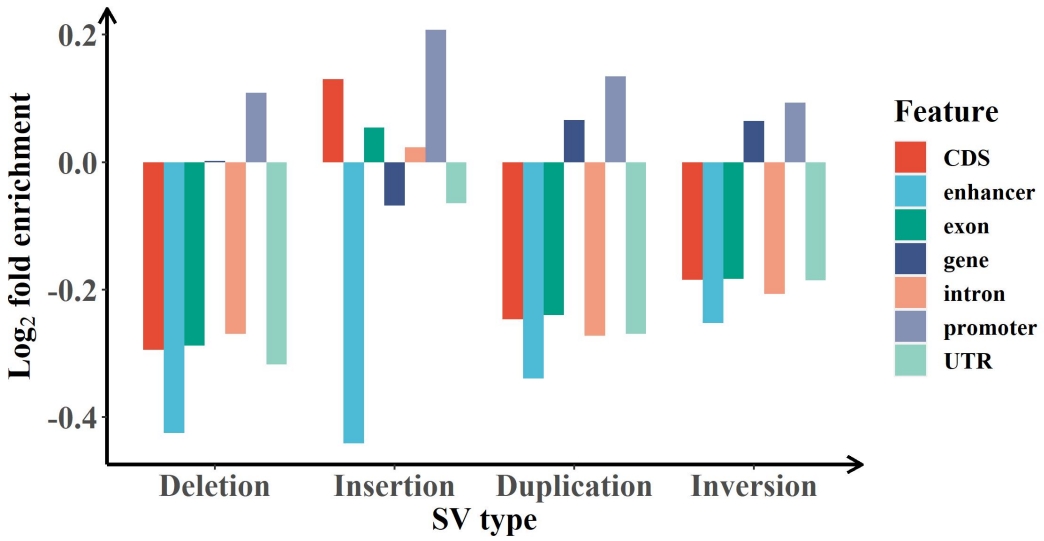
Supplementary Fig. 23. SV density of different sizes for different SV types.



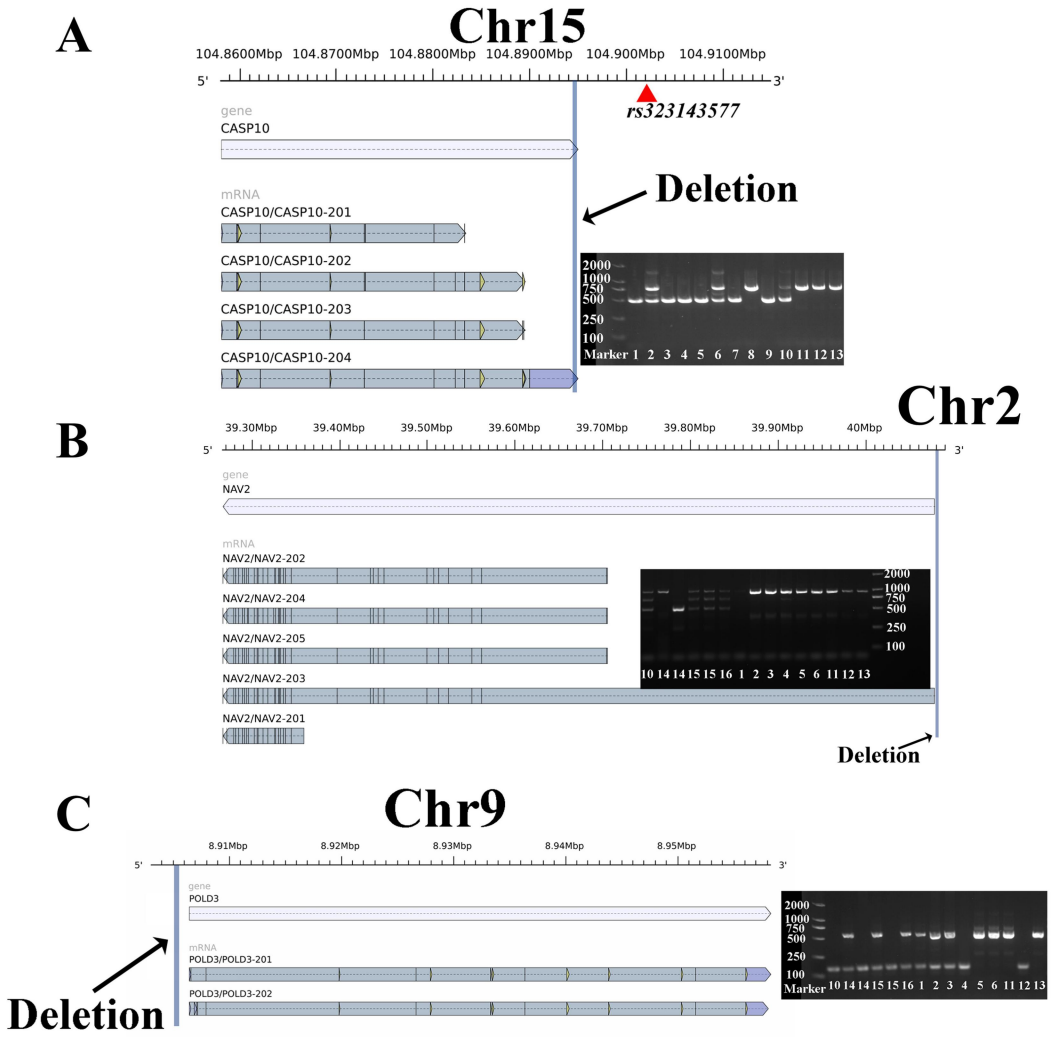
Supplementary Fig. 24. The annotation results of SVs.



Supplementary Fig. 25. Number of high impact annotations per SnpEff effect for SVs. A) deletions, B) insertions, C) duplications, and D) inversions.



Supplementary Fig. 26. Enrichment/depletion of different SV types within various functional genomic features.



Supplementary Fig. 27. The SVs in genes. A) The deletion in the 3'UTR of the *CASP10* gene. B) The deletion in the promoter of the *NAV2* gene. C) The deletion in the promoter of the *POLD3* gene. The right plot was the PCR results of the above SVs. The numbers from 1 to 16 represented the different breeds: 1.Wuzhishan pig, 2.Tongcheng pig, 3.Bama Xiang pig, 4.Rongchang pig, 5.Ningxiang pig, 6.Mieshan pig, 7.Diannan Small-ear pig, 8.Bamei pig, 9.Jinhua pig, 10.Tibetan wild boar (Hezuo pig), 11.Yorkshire, 12.Landrace, 13.Duroc, 14.Tibetan wild boar (Sichuan), 15.Tibetan wild boar (Tibet), 16. Tibetan wild boar (Diqing).

Supplementary References

1. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-+ (2011).
2. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
3. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).
4. Wang, K., Li, M. Y. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, (2010).
5. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat Protoc* **11**, 1–9 (2016).
6. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
7. Yang, W. *et al.* Animal-ImputeDB: a comprehensive database with multiple animal reference panels for genotype imputation. *Nucleic Acids Research* **48**, D659–D667 (2020).
8. Ding, R. *et al.* The SWine IMputation (SWIM) haplotype reference panel enables nucleotide resolution genetic mapping in pigs. *Commun Biol* **6**, 1–10 (2023).
9. Tong, X. *et al.* Accurate haplotype construction and detection of selection signatures enabled by high quality pig genome sequences. *Nat Commun* **14**, 5126 (2023).
10. Wang, Z. *et al.* PHARP: a pig haplotype reference panel for genotype imputation. *Sci Rep* **12**, 12645 (2022).
11. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Research* **32**, D493–D496 (2004).
12. Goldberg, A., Verdu, P. & Rosenberg, N. A. Autosomal Admixture Levels Are Informative About Sex Bias in Admixed Populations. *Genetics* **198**, 1209–1229 (2014).
13. Bu, D. C. *et al.* KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Research* **49**, W317–W325 (2021).