A high-throughput screening method for selecting feature SNPs to evaluate breed diversity and infer ancestry

Meilin Zhang, Heng Du, Yu Zhang, et al.

Genome Res. 2025 35: 1875-1886 originally published online July 15, 2025
Access the most recent version at doi:10.1101/gr.280176.124

Method

A high-throughput screening method for selecting feature SNPs to evaluate breed diversity and infer ancestry

Meilin Zhang,[1,3] Heng Du,[1,3] Yu Zhang,[1,3] Yue Zhuo,[1] Zhen Liu,[1] Yahui Xue,[1] Lei Zhou,[1] Sixuan Zhou,[2] Wanying Li,[1] and Jian-Feng Liu[1]

[1]*National Engineering Laboratory for Animal Breeding, Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture; State Key Laboratory of Animal Biotech Breeding; Frontiers Science Center for Molecular Design Breeding (MOE); College of Animal Science and Technology, China Agricultural University, Beijing 100193, China;* [2]*Institute of Animal Husbandry and Veterinary Sciences, Guizhou Academy of Agricultural Sciences, Guiyang, Guizhou 550005, China*

As the scale of deep whole-genome sequencing (WGS) data has grown exponentially, hundreds of millions of single nucleotide polymorphisms (SNPs) have been identified in livestock. Utilizing these massive SNP data in population stratification analysis, ancestry prediction, and breed diversity assessments leads to overfitting issues in computational models and creates computational bottlenecks. Therefore, selecting genetic variants that express high amounts of information for use in population diversity studies and ancestry inference becomes critically important. Here, we develop a method, HITSNP, that combines feature selection and machine learning algorithms to select high-representative SNPs that can effectively estimate breed diversity and infer ancestry. HITSNP outperforms existing feature selection methods in estimating accuracy and computational stability. Furthermore, HITSNP offers a new algorithm to predict the number and composition of ancestral populations using a small number of SNPs, and avoiding calculating the number of clusters. Taken together, HITSNP facilitates the research of population structure, animal breeding, and animal resource protection.

[Supplemental material is available for this article.]

Of the estimated 2 million to 100 million species on Earth, approximately 40 species have been domesticated for agricultural purposes. Over the past 12,000 years, these species have evolved into approximately 6000 to 7000 distinct animal breeds, adapted to specific local environments and production systems, resulting in abundant breed diversity (Scherf 2000). Especially after an initial pulse during the early Holocene, animal domestication became increasingly frequent, leading to the integration of numerous species into human environments and economies (Larson and Fuller 2014). Nowadays, domestic animals play a crucial role in the daily lives of millions of people, supporting the livelihoods of about 2 billion individuals worldwide, which accounts for one-third of the global population (Eusebi et al. 2020). However, numerous domestic animal breeds are endangered under the pressure of purebreds or crossbreds that meet the demand of changing farming systems and breeding objectives (Reist-Marti et al. 2003). To adapt to those challenges, many countries joined efforts to increase the diverse range of animal genetic resources. Therefore, tools for efficient and accurate discovery of the breed diversities of species and ancestries of populations are needed.

Breed diversity can be evaluated through genetic or phenotypic analyses, each offering unique insights (Reist-Marti et al. 2003). Traditionally, genetic diversity assessments relied on pedigree records (Caballero and Toro 2000). However, genealogical records, although valuable, have limitations that hinder their effectiveness in genetic diversity analysis, including incomplete pedigrees for many breeds, idealized assumptions about founders being unrelated and carrying two different alleles, and variations in genetic sharing among full siblings deviate from the theoretical 50% (Eusebi et al. 2020). In addition, confirming ancestries of populations becomes challenging when pedigree data are missing, incomplete, or inaccurate. Hence, with advancements in molecular and sequencing technologies, the evaluation of the breed diversity and ancestries of the population directly based on genomics has gradually been used instead of statistical inferences based on pedigree information.

Various marker-based techniques are available for analyzing the breed diversity and ancestries of populations (Yaro et al. 2017; Eusebi et al. 2020). Many studies compare and utilize different types of marker techniques, including mitochondrial DNA barcoding (mtDNA) (Guo et al. 2006), the Y-Chromosome technique (Bruford et al. 2003), minisatellite and microsatellite markers (Ćurković et al. 2016), and single-nucleotide polymorphism (SNP) (Coll et al. 2014; Yan et al. 2020). Especially for SNPs, common molecular markers in animals with uniform density are high detection accuracy, stable inheritance, and the potential to target functional regions, which usually are used for species genetic analysis, including population structure analyses, genetic diversity evaluation, and ancestry estimation (Kennedy et al. 2003; Zimmerman et al. 2020). Simultaneously, the latest advances in SNP high-throughput arrays and WGS facilitate the SNPs applied in the usual genetic analyses. However, compared with WGS, SNP arrays were primarily designed for based on commercial

[3]**These authors contributed equally to this work.**
Corresponding author: liujf@cau.edu.cn
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.280176.124. Freely available online through the *Genome Research* Open Access option.

breeds, resulting in potential ascertainment bias concerning local breeds (Pérez-Enciso et al. 2015). Moreover, WGS contains information not only on common mutations but also on rare variants of distinct breeds, and its usage opens new possibilities for studies on breed diversity and ancestry estimation (Toro et al. 2009).

At present, deep WGS facilitates SNP detection and indicates that there are millions of SNPs in different species; for example, the Human 1000 Genomes Project (1kGP) identified 111.05 million SNPs in the 3202 human cohort (Byrska-Bishop et al. 2022). However, in assessments of breed diversity and inference of population ancestry, these vast amounts of variations, which either lack predictive power for breeds or are redundant with each other, lead to the curse of dimensionality similar to that seen in genome-wide association studies (GWAS). Using these variations directly for assessments significantly increases detection time and computational complexity costs and may result in overfitting models on real data (Pudjihartono et al. 2022). Hence, many approaches were developed to select a limited number of SNPs with high informational value from a large SNP cohort. These selected SNPs, referred to as feature SNPs, effectively capture the breed diversity and serve as useful tools for breed differentiation and ancestry inference. The widely used selection methods in previous studies involve identifying the most distinguishing features based on typical filtration, such as the fixation index (F_{ST}) (Weir and Cockerham 1984), informativeness (I_n) (Kosoy et al. 2009; Nassir et al. 2009), selective sweep (SS) (Yan et al. 2020; Yang et al. 2022), and principal component analysis (PCA) (Paschou et al. 2007, 2008). In addition, with the development of machine learning, many different methods have emerged in these genetic analyses. For instance, max-relevance and min redundancy (MRMR) (Peng et al. 2005) based on a filter algorithm and sequential forward selection (SFS) (Pudil et al. 1994) based on a wrapper algorithm have been employed to facilitate the construction of a feature library. These methods help to efficiently and rapidly evaluate breed diversity and estimate ancestry. However, these feature selection methods were primarily established to select feature SNPs from SNP genotyping array data (Zhao et al. 2023), and developing an effective method for screening feature SNPs from high-coverage WGS is still imperative.

Furthermore, the traditional ancestry inference methods are mainly based on clustering algorithms, such as STRUCTURE (Pritchard et al. 2000) and ADMIXTURE (Kosoy et al. 2009), and conduct ancestry inference through an unsupervised analysis using a fixed number of clusters, *K*. To help researchers determine the optimal *K* value, ADMIXTURE provides cross-validation error for different *K* values (Alexander and Lange 2011). In contrast, fastSTRUCTURE (Raj et al. 2014) utilizes two additional metrics, the optimal model complexity and the number of nonempty model components, to obtain a reasonable range for *K*. Hence, multiple values of *K* needed to be calculated when inferring the ancestries of populations using these methods. This increases the consumption of computational resources, especially with the large SNP data sets. Therefore, it is worthwhile to exploit the ancestor estimation approach based on the prior information from the reference population, which avoids calculating multiple *K* values.

Here, we propose an algorithmic framework, HITSNP, to screen feature SNPs from deep WGS and evaluate its ability in breed diversity estimation and ancestry inference. Utilizing feature selection and machine learning methods, HITSNP effectively and reliably screens feature SNPs representing breed diversity from high-throughput data. Furthermore, we have outlined a pipeline for predicting the number and composition of ancestors based on a reference population and machine learning classifiers. Finally, we used simulated and real data to demonstrate the performance of HITSNP on feature SNP selection and ancestral prediction analyses.

Results

Methodological overview of HITSNP and summary of results files

We have developed HITSNP, a software based on our feature selection framework (Methods) and machine learning techniques. The software comprises three core modules: "feature SNP screening," "ancestry estimation," and "minimum subset selection" (Fig. 1). The "feature SNP screening" module screens and evaluates feature SNPs using feature selection methods, which comprises three methods: relief + reduce redundant (HITSNP-ReliefRR), max-relevance and min-redundancy (HITSNP-MRMR) (Peng et al. 2005), and cumulative classification ability (HITSNP-CCA) (Fig. 1A; Zhao et al. 2019). The "ancestry estimation" module is mainly used to train classifiers for ancestry prediction (Fig. 1B). Finally, the "minimum subset selection" module searches for the minimum SNP subset capable of distinguishing the breed of the purebred population (Fig. 1C).

In the Results section, we first evaluated the performance of the three modules, followed by a demonstration of HITSNP's practical applications.

Pig data sets for evaluating the performance of HITSNP

Previous studies screening the feature SNPs with SNP chips and low-coverage sequencing, these methods hardly captured the rare and low-frequency variants that tend to be specific to a population. Our tests showed that although sequencing coverage increased, the number of detected SNPs gradually increased. However, when the coverage of sequencing >20×, the novel-identified SNPs tend to reach trough (Supplemental Fig. S1; Supplemental Methods). Simultaneously, compared with the SNP chip, the deep-coverage sequencing detected hundreds to thousands of times more SNPs, putting more pressure on algorithms in effectively high-throughput screening feature SNPs. To comprehensively evaluate the performance of HITSNP, we used a large cohort in our study. This data set included high-coverage WGS (26.67×) of 1174 samples that were generated from our previous study (Du et al. 2024b,c) and 216 individuals downloaded from the public database (Supplemental Table S1). Using the standard GATK pipeline and after quality control (Methods), we detected 45.50 million SNPs, among which 2.13 million were multiallelic. Annotation of these SNPs revealed a total of 20.27 million variants within pig protein-coding genes (Ensembl Release 112). These included 443,354 exonic, 0.62 million untranslated regions (UTRs), 2418 splice variants, and 19.20 million intronic variants (Supplemental Fig. S2). Particularly focusing on variants within protein-coding exons, we identified 174,162 nonsynonymous SNPs. Across all the SNPs in this data set, 35.78% of the variants were rare variants (minor allele frequency [MAF] < 1%) (Supplemental Fig. S3). Notably, 36.27% of these detected SNPs were identified as novel, not previously reported in the dbSNP database (build 150) (Sherry et al. 2001). We found that each breed possesses a significant number of unique loci, with the highest number observed in the Diannan small-ear pig and the lowest in the Jiaxing Black pig (Supplemental Fig. S4).

To confirm that this data set is suitable for assessing the performance of HITSNP, we inferred the population structure of pigs

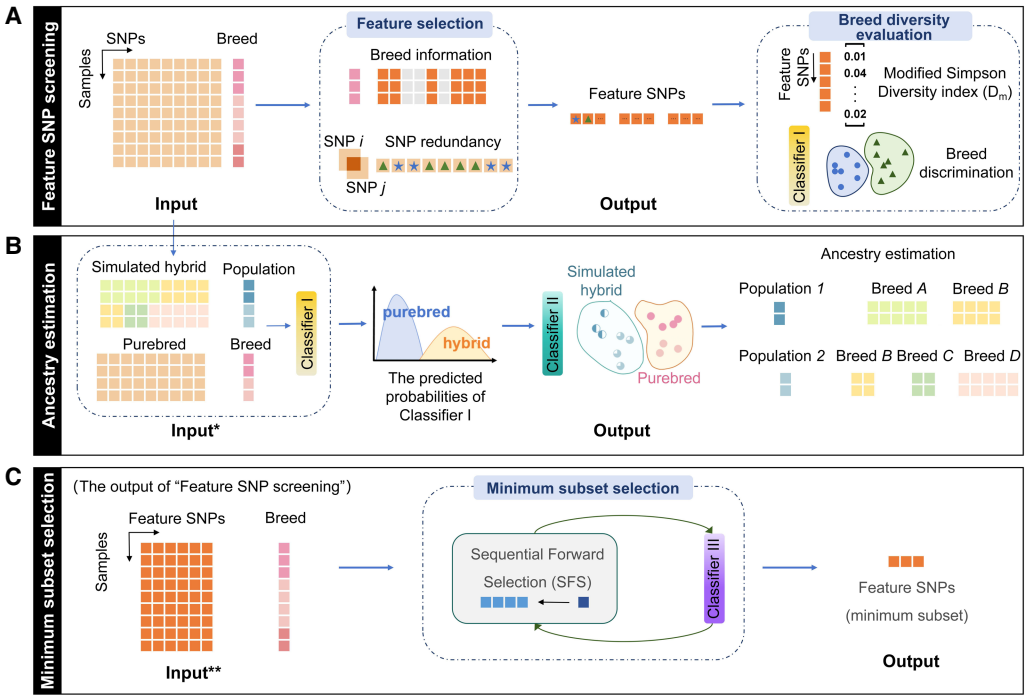

Figure 1. Overview of the HITSNP framework. (*A*) The functionality of the "feature SNP screening" module. This module takes genotype data and corresponding breed information of the samples as input. It outputs the selected feature SNPs and provides evaluation results of breed diversity based on the feature SNPs. (*B*) The "ancestry estimation" module estimates the number of ancestral breeds and the specific ancestral breeds of the hybrid population. Input*: The purebred and classifier I are derived from the "feature SNP screening" output. Simulated hybrids are generated based on the inputs of "feature SNP screening" (purebred data). HITSNP extracts feature SNP information from the genotype data of purebreds and simulated hybrids, which are then used as input for classifier I. (*C*) The "minimum subset selection" module selects the smallest possible feature SNPs set with the capability of breed discrimination. Input**: HITSNP extracts the feature SNPs identified in "feature SNP screening" from the initial input, which are then used as input for the sequential forward selection (SFS) method.

by applying this data set. The t-distributed stochastic neighbor embedding (t-SNE) map indicated that 60 pig populations could be clearly distinguished (Supplemental Fig. S5A). Moreover, the genetic composition showed that these populations could be divided into two large categories, including Asian and European pigs, coinciding with PCA results (Supplemental Fig. S5B). This concurred with the previous report that European and Asian pigs diverged around 1 million years ago (MYA) (Groenen et al. 2012). Focusing on the Asian pigs, the Jeju Black pig diverged from the other Chinese domestic pigs. The population structure of the Chinese domestic pigs indicated that it was consistent with the geographic distribution of different populations (Supplemental Fig. S5C) and could divide these populations into four large groups, including populations mainly resided in the north of China (brown), south of China (purple), the center and east of China (red), and south-west of China (green).

In total, this large-scale and complex data set could satisfy the comprehensive evaluation of HITSNP.

The performance of HITSNP in feature SNP filtration

We compared the utility of the "feature SNP screening" module of HITSNP with F_{ST}, I_n, SS, and PCA methods in feature SNP determination using 9,278,629 variants after linkage disequilibrium (LD) pruning in the above 60 pig populations. We conducted a comparative evaluation of our method and the other four methods from two aspects: one is assessing how well the selected feature SNPs reflect breed diversity, and the other is evaluating the computational stability of each method. We employed a modified Simpson diversity index (D_m) and machine learning classifiers' performance to quantify the ability of selected feature SNPs in reflecting breed diversity. The D_m (Methods) is an assessment metric designed to evaluate the ability of single-feature SNP to reflect breed richness. A higher index indicated greater overall capability in representing breed richness and breed differentiation. The performance of classifiers included accuracy, F1 score, and ROC-AUC. The computational stability of these methods was evaluated using the following criteria: the Jaccard similarity coefficient between cross-validation folds, as well as the standard deviations of the D_m and classifier performance.

We first validated the effectiveness of HITSNP in selecting feature SNPs to reflect the breed diversity of a cohort compared with the unfiltered SNPs. Two application scenarios, including low- and high-density feature SNPs (about 1000 and about 30,000 SNPs), were designed to conduct these comparisons. Because the performance of the classifier can influence the evaluation of breed diversity, we, respectively, selected the highest accuracy parameters for five frequently used classifiers as classifier I of "feature SNP screening" module to ensure the fairness of comparison. The five classifiers include fuzzy *k*-nearest neighbor (FKNN) with PCA applied to the inputs (PCA-FKNN-I), support vector machine (SVM-I), logistic regression (LR-I), bagging ensemble classifiers based on naive Bayes (BagNB-I), and random forest (RF-I) (Supplemental Tables S2, S3). The results showed that the average accuracy of unfiltered SNPs for LR-I, RF-I, and BagNB-I was only 98.60% (Fig. 2A,B). The LR-I and PCA-FKNN-I models were not trained because these

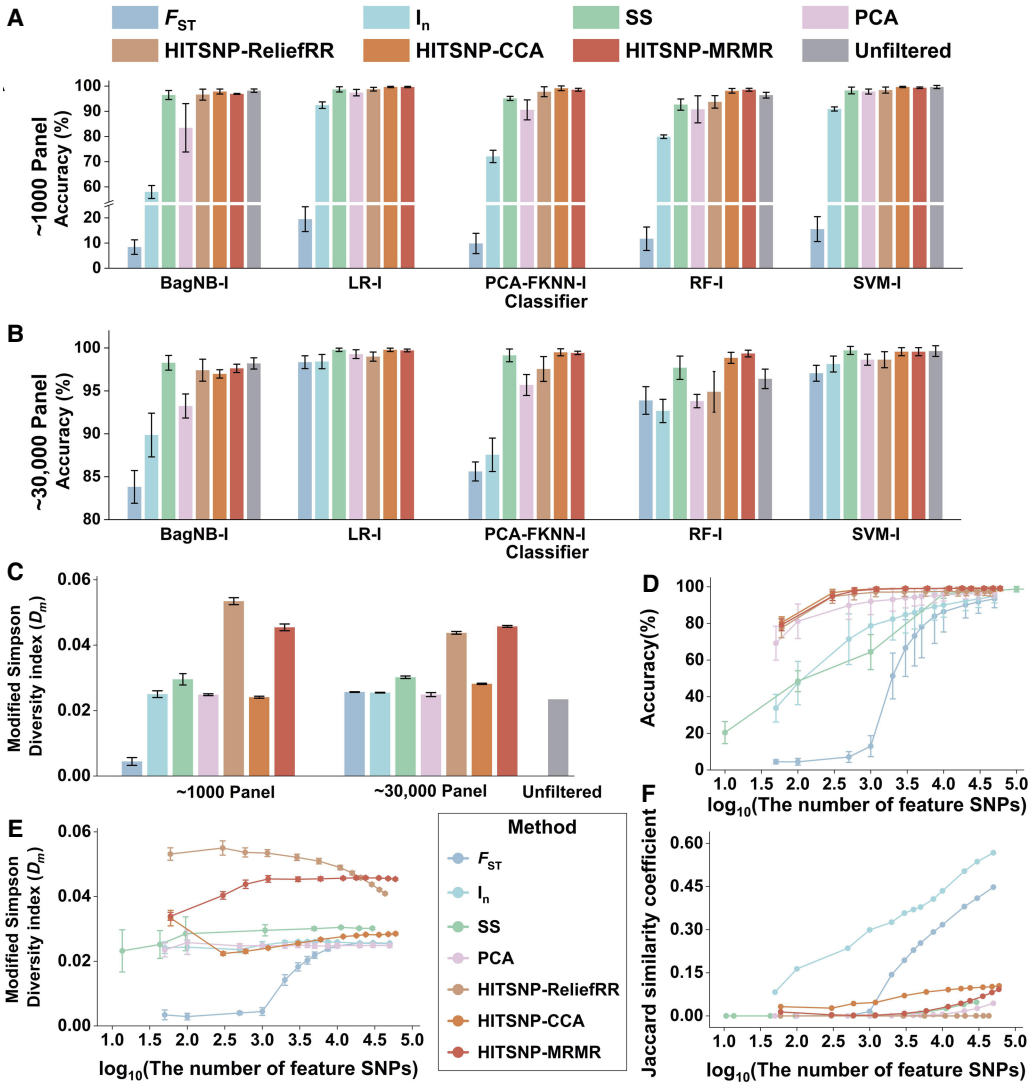

Figure 2. Evaluation of feature SNP selection performance among HITSNP and four other filtration methods. (*A*) Accuracy of five classifiers based on unfiltered SNP set and different feature SNP sets (about 1000 sites) selected by HITSNP and four other filtration methods. (*B*) Accuracy of five classifiers based on unfiltered SNP set and different feature SNP sets (about 30,000 sites) selected by HITSNP and four other filtration methods. (*C*) D_m results of unfiltered SNP data set and feature SNP sets selected by HITSNP and four other filtration methods on about 1000 and about 30,000 SNP sites panels. (*D*) Accuracy of different feature SNP sets selected by HITSNP and four other filtration methods on diverse gradients of SNP sets. (*E*) D_m results of different feature SNP sets selected by HITSNP and four other filtration methods on diverse gradients of SNP sets. (*F*) Jaccard similarity coefficient results of different feature SNP sets selected by HITSNP and four other filtration methods on diverse sizes of SNP sets. (Panels *D*–*F* use the same figure legends; "the number of feature SNPs" represents the average number of feature SNPs across five cross-validation folds for the same gradient of SNP sets.)

models were computationally impractical for this huge SNP data set. The average accuracies of feature SNPs filtered by HITSNP on low- and high-density sizes indicated no significant difference compared with the unfiltered SNPs ($P > 0.05$; *t*-test). In addition, all panels obtained from HITSNP-CCA and the high-density panel filtered by HITSNP-MRMR achieved higher average accuracy than the unfiltered set, which suggested that the magnitude of SNPs in the unfiltered set not only strained computational resources but also negatively impacted classifier accuracy owing to redundant SNPs. We noticed that feature SNP sets of all HITSNP methods achieved a higher mean D_m than the unfiltered set (Fig. 2C). In the unfiltered set, 48.67% of the SNPs had a D_m value of zero, whereas this proportion was significantly decreased in the filtered SNP set of HITSNP. Additionally, HITSNP selects a much more informative feature SNP set than the other four methods on both high- and low-density panels. For five classifiers, three methods of HITSNP and SS showed stably high accuracy on two different density sizes (Fig. 2A,B). Nearly all the indices of filtered feature SNP sets were higher than those of the unfiltered set, which implied that most methods tend to select SNPs with higher values of D_m. Notably, HITSNP-MRMR and HITSNP-ReliefRR showed a much higher D_m value than other methods (Fig. 2C).

To further explore the performance of HITSNP across diverse panel densities, we compared it with four other methods in 12 gradients ranging from about 60 to about 60,000 sites. For HITSNP, we set the target number of feature SNPs per breed. Because of SNP overlaps, the total SNP count is less than or equal to the target multiplied by the number of breeds (Supplemental Methods;

Supplemental Fig. S6). However, the other four methods utilized closely matched total SNP counts for an equivalent performance evaluation (Supplemental Table S4). The classifier accuracy, ROC-AUC, F1 score, and D_m value generally decreased with the number of feature SNPs decreasing, whereas the declining trends varied across these methods (Fig. 2D–F; Supplemental Fig. S7). HITSNP exhibited a relatively lower degree of decline across different methods, maintaining an accuracy of approximately 0.8 even when the number of SNPs was reduced to 60 (select one SNP per breed) (Fig. 2D; Supplemental Table S4). With more than 1000 SNPs selected by these methods, the average accuracy of five classifiers reached stability, ranging from 97.07% to 99.24%. SS achieved high accuracy comparable to HITSNP with more than 5000 SNPs but experienced a rapid decline as the number of SNPs fell below 5000. As the number of SNPs decreased, the performance of the PCA method surpassed SS but remained consistently lower than that of HITSNP. More importantly, for different densities of feature SNP sets, HITSNP demonstrated a superior D_m value compared with the other four methods, with HITSNP-MRMR and HITSNP-ReliefRR being extraordinarily high (Fig. 2E). We noticed that for the HITSNP-ReliefRR, the D_m value increased as the number of SNPs reduced. This pattern was also seen for the HITSNP-CCA, in which the number of SNPs fell from 300 to 60. Moreover, we found that HITSNP demonstrated superior performance in evaluating breed diversity and exhibited excellent stability. It was evident that the three methods under our algorithmic framework generally exhibited lower standard deviations in both the D_m value and classifier performance, particularly at the lower count of feature SNPs. In addition, the Jaccard similarity coefficient for HITSNP-CCA was the highest among the three methods within our algorithmic framework, indicating that HITSNP-CCA maintains relatively high stability among methods in HITSNP (Fig. 2F). The Jaccard similarity coefficient for F_{ST} and I_n showed a significant increase with the density of the SNP panels rising (Fig. 2F). However, their performance in breed diversity significantly lagged behind that of HITSNP and showed larger standard deviations than HITSNP. Overall, HITSNP revealed a stable ability to screen feature SNPs from high-throughput data, accurately representing the breed diversity, especially for lower amounts of feature SNPs.

Ancestry inference of extant individuals

Based on the feature SNPs detected by the "feature SNP screening" module, we further developed the "ancestry estimation" module in the HITSNP, which can accurately predict the ancestral populations of an individual, especially for an admixed individual. To test the performance of the "ancestry estimation" module, we simulated 13 crossbreeding systems (Supplemental Table S5), taking seven distinct breeds (five Asian indigenous pig breeds and two European commercial breeds) as ancestral breeds. We first utilized six simulation populations with ancestral proportions of 1:1 and 1:1:2 to represent hybrid populations, which were analyzed alongside purebred populations as inputs for the module. PCA revealed a clear differentiation between the simulated hybrid populations and the populations of their ancestral breeds across the simulation schemes 1–6 (Supplemental Fig. S8).

Our "ancestry estimation" module is constructed in two steps: evaluating the number of ancestors and estimating the proportions of ancestors. The first step inferred the number of ancestors based on the genotypes of the SNPs selected by the above "feature SNP screening" module. From the five classifiers in classifier I, LR-I, SVM-I, and RF-I were chosen to train classifier II, determining whether an individual has a single ancestor or multiple ancestors. The second step involves selecting the breeds whose predicted probabilities from LR-I are more than 0.05 as the crossbreeds' ancestries. With these predicted ancestries as the reference population, users can further estimate the composition of ancestors through the supervised analysis of ADMIXTURE or other software.

We evaluated the performance of the "ancestry estimation" module based on the classifier I and feature SNPs (about 10,000 SNPs) created by the "feature SNP screening" module. The predicted probability distributions of the three classifiers in classifier I had high Jensen–Shannon distances (JSD) (Supplemental Methods) between purebreds and simulated crossbreeds (Supplemental Table S6), indicating a clear differentiation in their predicted probability distributions. Then, we compared the prediction performance of four machine learning classifiers in classifier II, including decision tree (DT-II), SVM-II, LR-II, and RF-II (Fig. 3A). For the feature SNPs selected from HITSNP-MRMR, the accuracy of all four methods exceeded 99.75%. RF-II performed best among the four methods in evaluation, achieving an accuracy approaching 100.00%. The feature SNPs selected from HITSNP-ReliefRR performed inferior to the other methods, with accuracy in classifier II ranging from 58.23% to 82.03%. This might be because of the lower accuracy of classifier I for HITSNP-RelifRR, which consequently impacted the performance of classifier II.

To further explore the performance of the "ancestry estimation" module on lower density panels, we assessed the classifier II using about 1000 feature SNPs resulting from the "feature SNP screening" module. With the reduced density of feature SNPs, the accuracy of the four classifiers was relatively diminished compared to those based on about 10,000 SNPs (Fig. 3B). The accuracy for HITSNP-CCA and HITSNP-MRMR could remain >96.00%. We also noticed that RF-II outperformed the other three classifiers in classifier II on both the about 1000 and about 10,000 feature SNP sets filtered by HITSNP.

To further assess the ability of the "ancestry estimation" module to recognize the complex hybrid populations, we conducted other seven simulation schemes with varying ancestry proportions and multibreed hybrids. The results (Supplemental Fig. S9) revealed that when the ancestry proportion of a breed exceeds 80%, classifier II tends to classify hybrid samples as purebred. This indicated that populations with an extremely high proportion of one ancestral breed are not suitable for determining the number of ancestral populations using classifier II. However, for hybrid populations derived from multiple breeds, HITSNP-MRMR performed well across two panels, particularly with the LR-II. This demonstrated that LR-II in classifier II is well suited for determining the number of ancestral populations in scenarios with multiple ancestral origins. By incorporating these additional simulation scenarios, we have demonstrated the applicability of classifier II in ancestry inference. Although the method performs well in scenarios with balanced ancestry proportions, its predictive accuracy decreases in cases with an extremely high ancestry proportion of one breed or complex mixtures of multiple ancestral breeds.

After identifying the purebred and crossbred, we also developed a pipeline to predict the ancestral breeds for the populations with multiple ancestries. For the admixed populations recognized by classifier II, the predictive results generated by classifier I were utilized to estimate their ancestries. For the two different density panels (about 1000 and about 10,000 SNPs), we compared the

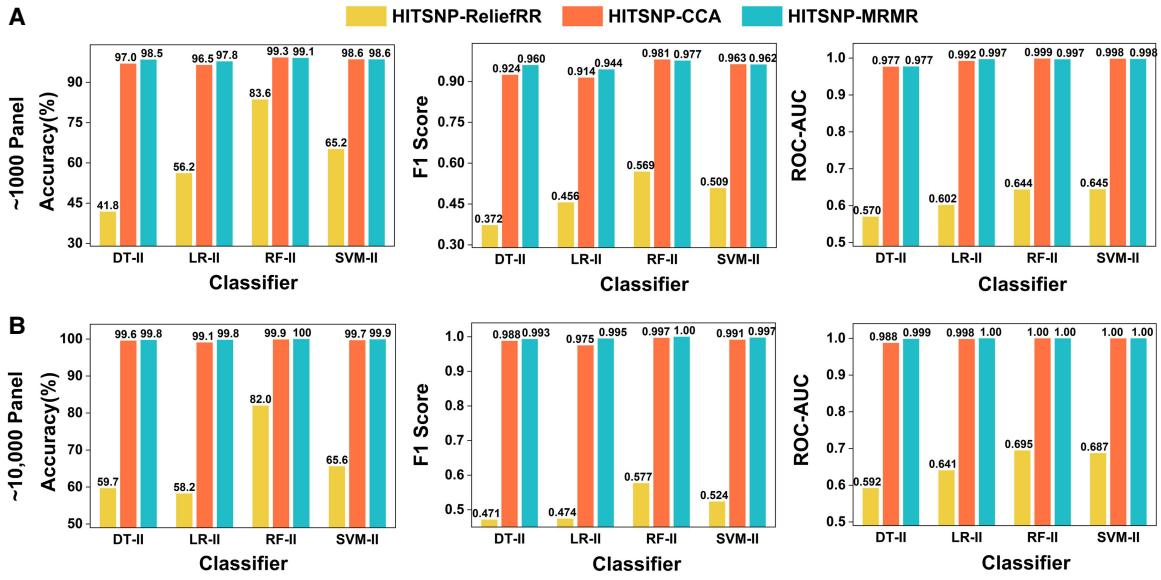

Figure 3. Classifier II performance is based on the feature SNPs selected by HITSNP. (*A*) Accuracy, F1 score, and ROC-AUC of four classifiers in classifier II based on feature SNPs (about 1000 sites) selected by HITSNP. (*B*) Accuracy, F1 score, and ROC-AUC of four classifiers in classifier II based on feature SNPs (about 10,000 sites) selected by HITSNP.

custom "accuracy" (CMA) and "precision" (CMP) (Supplemental Table S7; Supplemental Methods) of five classifiers in classifier I for 0.01 and 0.05 filtering criteria. The CMA and CMP of LR-I were both >97.00%, indicating its outstanding performance. Under the LR-I classifier, the 0.05 criterion exhibited lower CMA but higher CMP than the 0.01 criterion. To ensure accurate ancestry inference for the population, we aimed to minimize the occurrence of false ancestors in the predicted results. Therefore, given that both the 0.01 and 0.05 thresholds exhibit high CMA and CMP across two different densities, we employed the LR-I method with a 0.05 filtering criterion for predicting ancestral breeds.

Searching for the minimum feature SNP data set to accurately predict purebred

To explore the smallest informative SNP combinations that still maintained high effectiveness in purebred prediction, we compared the performance of two algorithms, SFS and recursive feature elimination (RFE). These two methods were wrapper methods that screen SNPs to meet the requirements for fewer SNPs. SFS utilizes a greedy algorithm to iteratively select SNPs from an empty set until the number of SNPs meets the predefined size. The RFE method obtains the desired subset by removing unimportant SNPs. Notably, RFE would be stopped when removing SNPs if the performance of the classifier no longer improves.

To assess the effectiveness of RFE in identifying a minimum feature SNP data set, we conducted minimum subset selection under RFE strategy combined with LR-III, SVM-III, and RF-III of classifier III. Because the feature SNP sets selected by PCA and three methods in HITSNP exhibited high accuracy on the around 1000 site panel, these sets were used as the initial set for minimum subset selection. We established four gradients for the number of feature SNPs (100, 200, 400, and 600, referred to as the set number) to explore the minimum subset. The RFE results indicated that with a lower set number (Fig. 4A; Supplemental Fig. S10), it struggled to reduce the number of feature SNPs to meet our target. Therefore, we compared the true number of SNPs in the resulting subset obtained by each combination, evaluating them based on test set accuracy (Supplemental Fig. S10) and cross-validation accuracy (Supplemental Table S8). Among the three classifiers, LR-III and SVM-III were able to reduce the number of SNPs closer to the set number while maintaining high accuracy Supplemental Fig. S10). In contrast, RFE combined with RF-III failed to further reduce the SNP count or achieve a smaller subset with relatively lower accuracy. The results showed that with the feature SNPs selected by HITSNP-CCA, RFE combined with the SVM-III method could reduce the number of SNPs to the lowest count (370 SNPs), achieving a test set accuracy of 98.92% and a cross-validation accuracy of 99.64%.

Similarly, we evaluated the performance of SFS using the same approach. We conducted a comparative analysis of 25 combinations (by pairing five feature selection methods with five classifiers) employing the SFS strategy. The classifiers used in classifier III included *k*-nearest neighbor (KNN-III), BagNB-III, SVM-III, LR-III, and RF-III. SFS consistently achieved the target number of feature SNPs across all gradients. Except for the initial set derived from PCA, the accuracy of other results increased with the size of the minimum subset increasing (Fig. 4B). At higher set numbers (400 or 600), LR-III and SVM-III selected the minimum subset with higher accuracy, whereas at lower set numbers, BagNB-III demonstrated stable and higher accuracy across three HITSNP methods. Among the lowest set numbers, the most effective combination was HITSNP-CCA and BagNB-III, achieving a test set accuracy of 93.53%. When the set number was 200, the combination of HITSNP-CCA and BagNB-III exhibited the highest generalization performance, and the test set accuracy reached 98.56%. Overall, when aiming to minimize the limited number of SNPs, the SFS strategy remains a suitable choice for balancing accuracy and SNP count, as RFE cannot guarantee a reduction to a predefined target number. Regarding the evaluation of D_m, the minimal subsets selected by SFS generally exhibit higher D_m values compared with RFE, further demonstrating that SFS is more suitable for identifying minimum subsets of a predefined size (Supplemental Fig. S11; Supplemental Table S9).

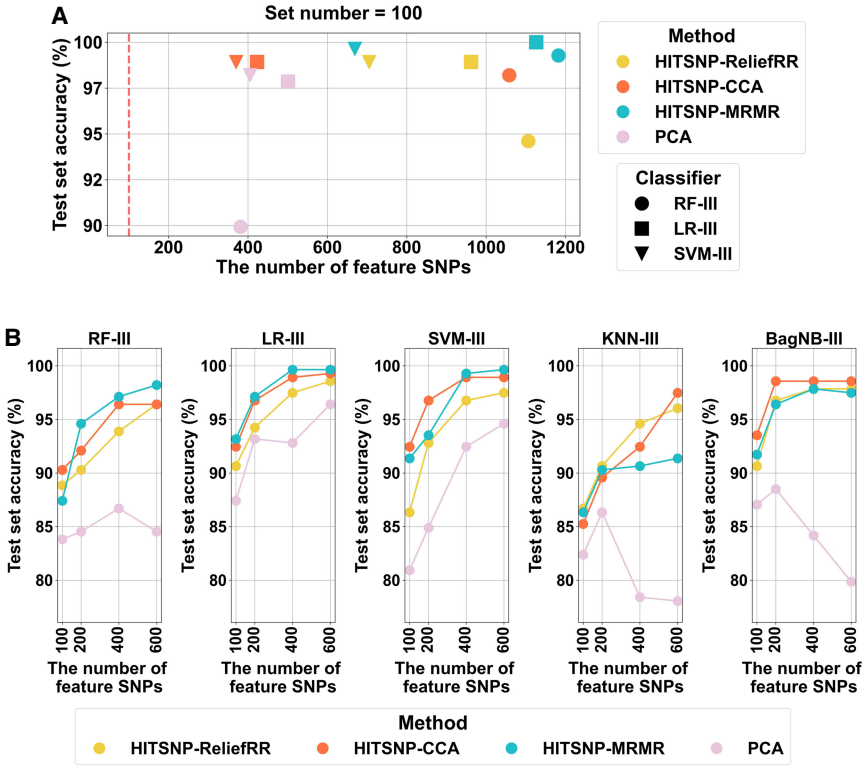

Figure 4. Performance of minimum subset selection using RFE and SFS strategy. (*A*) Test set accuracy and the corresponding number of minimum subsets selected by RFE strategy combined with three classifiers in classifier III when set number = 100. (*B*) Test set accuracy and the corresponding set number of minimum subsets selected by SFS strategy combined with five classifiers in classifier III. For different combinations of SFS strategy, the number of feature SNPs in the minimum subset reached the set number.

Considering the results of both SFS and RFE strategies, the SFS approach is more suitable for identifying minimal feature SNP subsets. Specifically, the combination of the HITSNP-CCA feature selection method and BagNB-III classifier suggested more application advantages under the SFS strategy in our data set.

Performance evaluation of screened feature SNPs obtained from HITSNP using real genomes

To evaluate whether HITSNP can effectively select feature SNPs applied for breed inference and breed differentiation, we additionally conducted a test using 65 purebred individuals and 19 hybrid individuals from the public data set not contained in our data set (Supplemental Table S10).

To validate the ability of HITSNP in estimating the number of ancestral breeds, we compared the performance of classifier II on around 1000 and around 10,000 panels using 84 test individuals. With classifier II constructed from around 10,000 feature SNPs, the accuracy of RF-II achieved 92.86%, in which the misclassification rate of hybrids was relatively higher than that of purebreds (Fig. 5A). The accuracies of the other three methods in classifier II were >85.71%. In comparison, the accuracy of classifier II based on the around 1000 feature SNPs was >74.00%, mostly showing a slight decline compared with around 10,000 SNPs (Fig. 5B). Notably, misclassifications of hybrids primarily occurred in the crossbred samples produced by the Min pig and Yorkshire. For these samples, the reason for misclassification was that the breed indicated the highest predicted probability was not their actual ancestral breed.

Breed identification test of 65 purebred samples used five classifiers (PCA-FKNN-I, SVM-I, LR-I, BagNB-I, RF-I) based on different densities of feature SNP subsets from "feature SNP screening" and "minimum subset selection" module (the around 10,000 and around 1000 feature SNP sets from "feature SNP screening" module, 200 SNPs from "minimum subset selection" module). The accuracies of five classifiers were all >95.00% based on the around 10,000 feature SNP set, and the around 1000 feature SNP set was >89.00% (Supplemental Table S11). PCA-FKNN-I and BagNB-I achieved 100.00% accuracy based on the low- and high-density feature SNP set from the "feature SNP screening" module (Supplemental Table S11). The 200 SNPs panel accuracies of SVM-I, LR-I, and BagNB-I were 93.85%, and the majority of misclassifications occurred predominantly in the Tibetan pig breed.

A web-based tool of feature SNPs and their functions in pigs

We have developed a web-based tool, BSP (https://1kcigp.com/BSP), to facilitate ready access to the feature SNPs of our pig data set with 60 breeds. More importantly, the tool integrates our methods and the screened feature SNPs, which allows researchers to

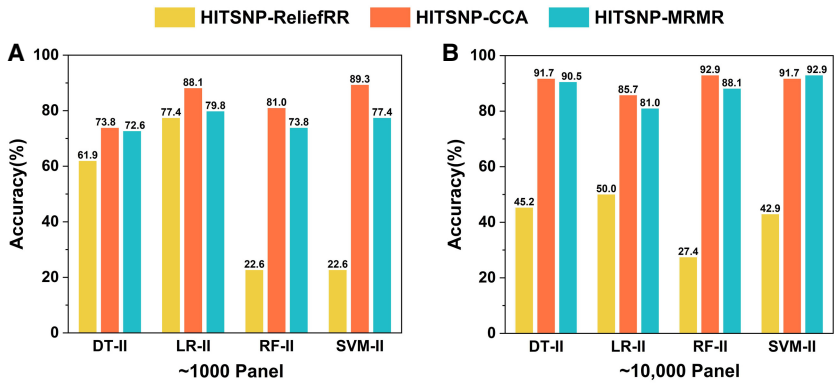

Figure 5. Classifier II accuracy results of the feature SNPs selected by HITSNP for 89 public pig genomes. (*A*) Four classifiers' accuracy in classifier II of the feature SNPs selected by HITSNP on the around 1000 panel. (*B*) Four classifiers' accuracy in classifier II of the feature SNPs selected by HITSNP on the around 10,000 panel.

estimate the ancestries of their sequenced pigs. This web tool could benefit pig research, breeding communities, and breed conservation.

HITSNP applicability to plants

We have included a melon data set (Wang et al. 2021) to validate the effectiveness of HITSNP in selecting feature SNPs for plants. Compared with the pig data set, the accuracy of classifier I on the melon data set was lower but still remained at ~90% (Supplemental Fig. S12). Notably, the PCA-FKNN-I and RF-I outperformed the other classifiers on average in the melon data set, which were not the best classifiers in the pig data set. The successful application of HITSNP on the melon data set demonstrates its broad applicability to both diploid animals and plants with multiple breeds or varieties.

Discussion

In this study, we introduced HITSNP, a stable, effective, and automated tool for screening feature SNPs representing breed diversity from high-throughput data. Compared with four common filter methods, the feature SNPs selected by HITSNP algorithms provided a precise description of breed diversity with high stability, particularly when filtered to low-density feature SNP set. Additionally, we offered a practical pipeline for ancestry inference based on the limited number of feature SNPs. This approach provides prior information on ancestry composition for the supervised analysis, such as ADMIXTURE, thus avoiding the challenges of determining the optimal *K* value.

Traditionally, the methods used in feature SNP screening only consider the breed information or SNP-redundancy; however, the HITSNP framework algorithms balance them well and can efficiently filter feature SNPs, resulting in improved classifier performance and higher D_m values for feature SNP sets across different densities. For instance, compared with the F_{ST}, I_n, and SS methods, which select ranked SNPs relying on breed information, the HITSNP methods consider breed information and evaluate SNP-redundancy through mutual information. Thus, HITSNP can effectively remove the potentially redundant feature SNPs. Additionally, compared with F_{ST} and I_n, HITSNP exhibits a lower Jaccard similarity coefficient across different cross-validations. The Jaccard similarity coefficient for HITSNP is likely influenced primarily by its selection strategy and the degree of sample overlap (Supplemental Methods; Supplemental Fig. S13). Although there are differences in feature SNPs across CVs, the ability of the feature SNP sets screened by HITSNP for population classification remains consistently high across most panel sizes. The PCA method relies on the information of the top principal components (PCs) rather than the breed information of the reference population. Consequently, PCA performs better than F_{ST}, I_n, and SS methods at low density. However, its performance is inferior to that of HITSNP under almost all panel density gradients tested in this study.

In addition to evaluating the entire feature SNPs set in representing breed diversity, we explore the ability of individual SNPs using a customized metric, D_m. Owing to the use of biallelic variants, the individual SNP has an exceedingly low index value when distinguishing among 60 breeds. However, because the comparisons were conducted on the same breed scale, the index can still effectively reflect the breed diversity of individual SNPs. Moreover, HITSNP-ReliefRR and HITSNP-MRMR demonstrated relatively high mean D_m values, further supporting the effectiveness of HITSNP in screening feature SNPs representing breed diversity.

The three algorithms within our framework also differ: HITSNP-ReliefRR and HITSNP-CCA optimize breed information and SNP-redundancy simultaneously by calculating the cumulative value of the product, whereas HITSNP-MRMR calculates subtraction. The accuracies of HITSNP-CCA and HITSNP-MRMR were both high, indicating that the two optimized methods might not significantly affect the feature SNP screening capability. HITSNP-ReliefRR has slightly lower accuracy than the other HITSNP methods, possibly because it selects feature SNPs based on prefiltered SNPs by SS rather than the complete set of unfiltered data.

Research on ancestry inference is typically applied in population stratification in GWAS analysis and scientific fields such as population history, medical research, and forensics (Halder et al. 2012; Elhaik et al. 2014; Suarez-Pajes et al. 2021; Alladio et al. 2022). Current ancestry inference methods rarely have the direct functionality to accurately distinguish purebreds and hybrids, which is a critical concern in plant and animal breeding. We have developed a method that directly distinguishes purebreds and hybrids using a machine learning classifier based on their distribution differences in prediction probabilities. This functionality is achieved by classifier II, which is trained based on the prediction results of classifier I in HITSNP. Therefore, the accuracy of classifier I is extremely important in accurately inferring the number of ancestral breeds. This importance was demonstrated by the different performances of classifier II that was trained on results derived from HITSNP-CCA, HITSNP-MRMR, and HITSNP-ReliefRR. In addition to estimating purebreds and hybrids, we also developed a practical ancestry inference pipeline for hybrids based on the results of classifier I, which provides prior information for setting the *K* value. Determining the optimal *K* based on the cross-validation error from ADMIXTURE can be distorted by hierarchical population stratification or uneven sample sizes (Puechmaille 2016; Janes et al. 2017). In contrast, HITSNP infers the ancestral breeds through individual-level classifier predictions, achieving robust ancestry prediction that is unaffected by population size or population stratification. This advantage suggests HITSNP may outperform ADMIXTURE in scenarios in which there is limited or no prior knowledge about the population for which ancestry inference is required (Supplemental Methods; Supplemental Fig. S14). Although some methods (Zhang et al. 2008; Chen et al. 2013; Bansal and Libiger 2015), such as SNPweights and iAdmix, can estimate ancestry without determining an optimal *K* value by relying on the information of the reference population, they are suitable for high-density or whole-genome data. In contrast, the pipeline of HITSNP can achieve ancestry inference using a limited number of SNPs obtained from the "feature SNP screening" module.

In summary, HITSNP is a practical and effective tool for extracting feature SNPs, which also provides prior information for ancestry inference based on reference population data. We consider that this tool will be useful in understanding the role of SNPs in evaluating breed diversity and contributing to different breed formations.

Methods

HITSNP algorithm

HITSNP is mainly based on machine learning techniques and a feature selection framework that combines breed information and

SNP-redundancy. Its inputs include a quality-controlled biallelic SNP data set and the breeding information of its corresponding sample cohort. The tool consists of three core modules. First, the "feature SNP screening" module utilizes the feature selection framework to identify a subset of feature SNPs from the input data. This module finally provides a data set of feature SNPs, combined with the assessment results of feature SNPs on breed diversity. Second, the "ancestry estimation" module is mainly based on the "feature SNP screening" result, which generates simulated data for hybrid populations and trains classifiers for ancestry prediction. It outputs each trained classifier along with its performance. Third, the "minimum subset selection" module also depends on the above feature SNPs and searches the minimum SNP subset that can distinguish the breed of the purebred population while meeting the specified requirements on the number of SNPs. The detailed algorithms underlying these modules are described in the following sections.

Feature SNP selection module

We have developed a feature selection framework for high-throughput screening feature SNP based on the filter method in this module. This framework consists of two main indicators: breed diversity capability of a single-feature SNP and redundancy between two feature SNPs. This module provides three distinct methods to separately select feature SNPs, including a modified relief method (HITSNP-ReliefRR) in this study, HITSNP-CCA, and HITSNP-MRMR method. The latter two methods were constructed based on standard CCA and MRMR algorithms, respectively.

Based on the relief method, the HITSNP-ReliefRR algorithm additionally employs an incremental search strategy to identify the optimal subset of features. This selection process is guided by weights and mutual information derived from the relief algorithm (Urbanowicz et al. 2018). A one-vs-the-rest (OvR) multiclass strategy was adopted, and the top SNPs selected at each iteration were merged into the final SNP panel.

The HITSNP-ReliefRR methodology employed in this study consists of the following two procedures. First, the Relief method was applied to calculate the weight values of each feature SNP. The initial weight value is set to zero. A random selection of target R_i($i\in\{1,2,\ldots,m\}$) is made. By finding the nearest hit (H, the nearest sample of the same class) and the nearest miss (M, the nearest sample of a different class) for the target R_i sample, the weight of each feature is calculated using the following formula:

$$W[A] = W[A] - \frac{\mathrm{diff}(A, R_i, H)}{m} + \frac{\mathrm{diff}(A, R_i, M)}{m}, \quad (1)$$

where W[A] means the weight of feature "A," and diff means differences observed between the target sample R_i and neighboring instances of R_i (H or M) on the feature "A" (Supplemental Methods). m refers to the number of nearest neighbors that are considered when calculating the feature weights. In HITSNP, we default to setting m as six.

Then, we further selected the feature SNPs through incremental learning combined with mutual information. The feature selection process in the HITSNP-ReliefRR method starts with an empty set and incrementally selects features. The feature with the highest weight W[A] will be selected as the first feature SNP, and the classification capability $R(S_1)$ is set to this highest weight. For subset S_{j-1}, which represents a collection of j−1 selected feature SNPs, the jth feature SNP "A_j" is chosen from the remaining set to maximize the classification capability $R(S_j)$ of the subset S_j. The value of $R(S_j)$ is calculated as follows:

$$R(S_j) = R(S_{j-1}) + (1 - \max_{i\in\{1,\ldots,j-1\}} NI_{A_iA_j})W(A_j), \quad (2)$$

where $R(S_j)$ is the feature's classification capability of the subset S_j, $W(A_j)$ means the weight of feature "A_j." $NI_{A_iA_j}$ is the standardized mutual information (NMI) between A_i and A_j (Supplemental Methods).

The HITSNP-CCA and HITSNP-MRMR methods (Supplemental Methods) adopted different criteria to choose the feature SNPs in this module. By maximizing cumulative classification ability, the HITSNP-CCA algorithm could enhance the estimation accuracy of breeds differentiation and determine breed attribution and then minimize the number of SNPs based on normalized mutual information. The HITSNP-MRMR algorithm relies on the mean value of mutual information across all feature SNPs and breed labels for feature selection. Combining the above criterion, it utilizes the average mutual information between pairs of feature SNPs to reduce redundancy.

Ancestry estimation module

We propose an intuitive approach for ancestry prediction based on the prediction probability distribution characteristics of machine learning classifiers derived from the "feature SNP screening" module. Using the same data set employed in the "feature SNP screening" module, HITSNP simulates hybrid populations based on a previously reported method (Pardo-Seco et al. 2014). In brief, HITSNP randomly mixes haplotypes from the hypothetical ancestral populations and their corresponding proportions to generate a new hybrid population. This process is repeated 100 times to create a large enough population that satisfies the following classifier training. The input data set supplied to this module and the resulting simulated hybrid population were divided into two types labeled as "purebreds" data and "simulated crossbreeds" data. Then, it extracts the feature SNPs from "purebreds" and "simulated crossbreeds" data to train the multiple classifiers in classifier II to recognize these two different labels. These data were split into 7:3 train set and test set. Three classifiers in classifier I, including LR-I, SVM-I, and RF-I, take the train set as input and transfer the top six prediction probabilities of each classifier to train classifier II. Finally, the performance of four classifiers II (SVM-II, DT-II, LR-II, RF-II) in purebred and crossbreed prediction will be evaluated. Additionally, for crossbreeds, the certain breed with its probability of classifier I (LR-I) greater than 0.05 will be considered as the ancestor of the crossbred.

Minimum subset selection module

To minimize the number of feature SNPs, we employed a wrapped feature selection method known as SFS in this module, with the accuracy of the classifier serving as the evaluation criterion. The SFS algorithm initiates with an empty set and gradually selects feature SNPs based on the evaluation criterion until it reaches the expected number of feature SNPs or until there is no further improvement in model accuracy. In HITSNP, SFS was conducted combined with the BagNB-III classifier. Subsequently, HITSNP utilizes cross-validation and test set accuracy to evaluate the performance of the minimal subset screening strategy and the generalization performance of the optimal subset derived from these five strategies.

Assessment metrics

To evaluate the performance of HITSNP in screening feature SNPs, which can accurately assess the breed diversity, we adopt two evaluation dimensions: a customized metric (D_m) based on the Simpson diversity index and F_{ST} value, and the performance of the machine learning classifier. The D_m value was applied to evaluate the single-feature SNP's capacity to differentiate breeds and reflect breed richness. The metric calculation is based on the F_{ST}

values calculated between each breed and populations outside that breed.

The SNP is considered inadequate for distinguishing breeds with corresponding F_{ST} values below a predefined low threshold (one). Contrarily, F_{ST} values exceeding a high threshold (h) were regarded as facilitating the distinguishment of those pairwise breeds, and the SNP received a reward coefficient in the D_m. The D_m is represented with the following formula:

$$D_m = \begin{cases} \frac{\sum F_{ST_h}}{n_h}\left(1 - \frac{n_l(n_l-1)}{N(N-1)}\right) & F_{ST} > h \\ h\left(1 - \frac{n_l(n_l-1)}{N(N-1)}\right) & \text{otherwise} \end{cases}, \quad (3)$$

where N was the total number of breeds, and n_l represents the number of F_{ST} values below one. The term $\frac{n_l(n_l-1)}{N(N-1)}$ will assign a lower score to SNP with a higher number of F_{ST} values below one. n_h means the number of the F_{ST} values higher than h; F_{ST_h} means the exact value of F_{ST}, which is higher than h. We compared the D_m with other calculation methods, such as the mean of F_{ST}, demonstrating that D_m more effectively assigns higher scores to SNPs that can accurately distinguish a greater number of breeds from the reference population. This ensures D_m provides a more precise representation of breed richness as reflected by individual SNPs (Supplemental Methods; Supplemental Fig. S15). In addition to independently assessing the performance of each SNP using the D_m, the mean D_m value of the feature SNP data set was calculated as a metric to assess their capability in breed classification.

In addition, we train machine learning classifiers based on the feature SNPs and evaluate the effectiveness of the feature SNPs through the classifier's performance. When evaluating the classifiers, we ensured a high level of accuracy while balancing the F1 score and ROC-AUC. This comprehensive evaluation allows us to assess the effectiveness of the feature SNP subsets.

The pig SNP data set used for training and testing the HITSNP

A diverse panel of 1174 pigs from 47 different breeds was sequenced in our previous study (Supplemental Table S1; Du et al. 2024a,b). These samples were sequenced by the MGISEQ-2000 platform (MGI) with 150 bp paired-end reads. The sequencing samples in this study had a mean coverage of ~26.67×. Additionally, the genome data of 216 individuals was downloaded from the public database, comprising 13 nonrepetitive distinct breeds (Supplemental Table S1).

The raw reads of all samples were initially filtered and trimmed using Trim Galore! (v0.6.1) (Martin 2011). Subsequently, all cleaned reads from each individual were aligned to the Sscrofa11.1 reference genome using the Burrows–Wheeler aligner (BWA; v0.7.17-r1188) (Li and Durbin 2009). Duplicated reads were removed, and the alignment results were sorted using the genome analysis toolkit (GATK; v4.0.12.0) (DePristo et al. 2011) and SAMtools (v1.9) (Li et al. 2009). SNPs were identified and filtered using GATK with the following criteria: (1) variant confidence/unfiltered depth of nonreference samples (QD) > 2.0; (2) RMS mapping quality (MQ) > 40.0; (3) Phred-scaled *P*-value using Fisher's exact test to detect strand bias in the reads (FS) < 60.0; (4) strand bias estimated by the symmetric odds ratio test (SOR) < 3.0; (5) the μ-based Z-approximation from the Mann–Whitney *U* test for mapping qualities (MQRankSum) > −12.5; (6) the μ-based Z-approximation from the Mann–Whitney *U* test for the distance from the end of the read for reads with the alternate allele (ReadPosRankSum) > −8.0; (7) and no more than three SNPs clustered in a 10 bp window.

These high-quality 45.50 million SNPs were further filtered, and the SNPs with genotyping rates smaller than 0.99 and MAF smaller than 0.01 were removed by PLINK (v1.9) (Purcell et al. 2007). The SNPs were pruned using PLINK with the parameters "--indep-pairwise 50 10 0.9". After that, the remaining SNPs were used to evaluate HITSNP and other breed diversity and ancestry estimation methods. The PCA and t-SNE (van der Maaten and Hinton 2008) were used to visualize the genetic distance of indigenous and foreign pig breeds.

Evaluation of HITSNP using pig SNP data

After LD pruning, the remaining biallelic variants in the pig SNP data set were formatted into zero, one, and two representations. These data were further used to evaluate the performance of three modules in HITSNP and compare them with other widely used methods. Because of the computational complexity and considerable time associated with Relief in managing a vast number of SNPs, a preliminary prefiltering step is undertaken before applying HITSNP-ReliefRR. Based on the F_{ST} and nucleotide diversity (θ_π) ratio, the prefiltering process selected approximately 380,000 feature SNPs as input for the HITSNP-ReliefRR. θ_π values were calculated using a 10 kb sliding window. The θ_π ratio was determined as the $\log_2$ of the quotient between the θ_π values of two breeds. The top 0.5% of windows from both tails of the θ_π ratio distribution were selected across all breed pairs. Additionally, the top 0.5% of the F_{ST} values within the selected windows would be the result of prefiltering process, with F_{ST} values calculated using VCFtools (Danecek et al. 2011) for each breed pair.

To evaluate the performance of our feature selection framework, we also used four popular feature selection methods (F_{ST}, I_n, SS, and PCA methods) compared with our method. The F_{ST} method selected feature SNPs by ranking the mean F_{ST} values calculated by VCFtools for 60 breeds. Informativeness for the I_n method was calculated using a previously described method (Rosenberg et al. 2003). The PCA method ranked and selected feature SNPs by assigning weights based on their contribution to the explained variance from the top 100 PCs. The SS method applied the same procedure as the filtering process described above. We conducted a comparison of the feature SNPs yielded by HITSNP (utilizing three different algorithms) and the other four filtration methods across two distinct SNP subset sizes: around 1000 and around 30,000. We also compared their performance to the unfiltered SNP data set. Subsequently, we compared the breed diversity of feature SNPs and the stability of the above filtration methods across different panel densities. The comparison was mainly based on the following aspects: the ability of feature SNPs to represent breed diversity and the stability of filtration methods. D_m and the machine learning classifiers' performance were used to quantify the breed diversity of feature SNPs. Additionally, the standard deviations of these metrics and the Jaccard similarity coefficient between cross-validation folds were utilized to evaluate the stability of filtering methods.

For ancestry inference, we designed a series of simulations to evaluate the performance of HITSNP. The selection of simulated ancestors primarily considered three factors: the number of ancestors, genetic distance, and expected ancestry proportion. Thirteen hybrid scenarios (Supplemental Table S5) were simulated, with the Bama Xiang pig, Jiaxing Black pig, Dapulian pig, Yorkshire, Landrace, and Duroc selected as ancestral breeds, each generating 2000 offspring per program. We employed two indexes, CMA and CMP, to examine whether ancestral breeds could be identified based on the classifier's estimated prediction probability (Supplemental Methods).

We divided the "purebreds" data and the "simulated crossbreeds" data into 70% training sets and 30% test sets to train

suitable machine learning classifiers (classifier II) and compared their performance. We used the results from the "feature SNP screening" module, based on the around 1000 and around 10,000 feature SNPs, as inputs for training classifier II.

To investigate the minimal informative SNP combinations that are highly effective in predicting purebreds, we scanned optimal feature subsets using wrapped feature selection methods based on around 1000 SNPs, with the classifier's accuracy serving as the evaluation criterion. A random cross-validation set was chosen for testing, combining two wrapper feature search strategies, SFS and RFE. Two strategies were evaluated under the same target SNP set sizes (100, 200, 400, 600). SFS was combined with five machine-learning classifiers (KNN-III, BagNB-III, SVM-III, LR-III, RF-III). Because RFE requires the weight of input features, only SVM-III, LR-III, and RF-III classifiers were utilized to assess the performance of the subsets. The accuracy of combined classifiers and the D_m value were used to evaluate the performance of the minimum subset.

Real-data test for the screened feature SNPs from the pig data set

Real-data testing involved sequencing data from 65 purebred samples representing seven breeds and 19 hybrid samples derived from five crossbreeding combinations. The sequencing data of these 84 individuals were downloaded from the public data set (Supplemental Table S9). This assessment aimed to evaluate the efficiency of feature SNPs and classifiers in breed diversity and ancestry inference (Supplemental Methods). In this study, we employed the Beagle (v5.2) (Browning et al. 2018) to impute the test samples using our population as the reference group to reduce the impact of missing SNP. Feature SNPs were then extracted from the imputed data, and their genotypes were unified with the minor alleles across the reference population before being converted to zero, one, and two. Subsequently, the data were fed into the machine learning classifier, and we assessed the discriminatory power of these models in distinguishing between purebred and hybrid populations (classifier II) and in the breed inference of purebred (classifier I).

Evaluation of HITSNP using melon SNP data

We utilized a genotyping-by-sequencing (GBS) data set (Wang et al. 2021) comprising 2081 melon samples from two major subspecies: thick-skinned melon (*ssp. melo*) and thin-skinned melon (*ssp. agrestis*). The GBS data were downloaded from Cucurbit Genomics (http://cucurbitgenomics.org/ftp/GBS_SNP/melon/raw_GBS_data/). The SNPs were filtered using PLINK based on the following criteria: (1) MAF > 0.01 and (2) missing rate < 0.05. After filtering, 34,787 SNPs were retained and used for feature SNP selection with HITSNP to distinguish melon subspecies. We evaluated the screening results using five classifiers in classifier I. Additionally, we visualized the genetic distribution of the two subspecies using PCA and t-SNE plots.

Data access

The SNP data generated in this study have been submitted to Zenodo (https://doi.org/10.5281/zenodo.13785208). HITSNP is available at GitHub (https://github.com/CAU-TeamLiuJF/HITSNP) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was financially supported by a National Science and Technology major project (2022ZD0115704), the National Natural Science Foundations of China (3227200469 and 32302708), Chinese Universities Scientific Fund (2023TC196), Science and Technology Program of Guizhou Province (Qian Kehe Support [2022] Key 032), the Earmarked Fund for China Agriculture Research System (no. CARS-pig-35), and the 2115 Talent Development Program of China Agricultural University. We acknowledge the computational support provided by the High-Performance Computing Platform of China Agricultural University.

Author contributions: J-F.L. conceived and designed the study, directed the project, provided all data and computational resources, supervised bioinformatic and statistical analyses, and revised the paper. M.Z., H.D., and Y.Zhang designed the analytical strategy and performed analysis processes. Y.Zhuo and Z.L. wrote the software. Y.X. and L.Z. conducted the software validation. S.Z. and W.L. revised the manuscript. All authors read and approved the final version.

References

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12:** 246. doi:10.1186/1471-2105-12-246

Alladio E, Poggiali B, Cosenza G, Pilli E. 2022. Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field. *Sci Rep* **12:** 8974. doi:10.1038/s41598-022-12903-0

Bansal V, Libiger O. 2015. Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. *BMC Bioinformatics* **16:** 4. doi:10.1186/s12859-014-0418-7

Browning BL, Zhou Y, Browning SR. 2018. A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet* **103:** 338–348. doi:10.1016/j.ajhg.2018.07.015

Bruford MW, Bradley DG, Luikart G. 2003. DNA markers reveal the complexity of livestock domestication. *Nat Rev Genet* **4:** 900–910. doi:10.1038/nrg1203

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185:** 3426–3440.e19. doi:10.1016/j.cell.2022.08.004

Caballero A, Toro MA. 2000. Interrelations between effective population size and other pedigree tools for the management of conserved populations. *Genet Res* **75:** 331–343. doi:10.1017/S0016672399004449

Chen C-Y, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. 2013. Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29:** 1399–1406. doi:10.1093/bioinformatics/btt144

Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. 2014. A robust SNP barcode for typing mycobacterium tuberculosis complex strains. *Nat Commun* **5:** 4812. doi:10.1038/ncomms5812

Ćurković M, Ramljak J, Ivanković S, Mioč B, Ivanković A, Pavić V, Brka M, Veit-Kensch C, Medugorac I. 2016. The genetic diversity and structure of 18 sheep breeds exposed to isolation and selection. *J Anim Breed Genet* **133:** 71–80. doi:10.1111/jbg.12160

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27:** 2156–2158. doi:10.1093/bioinformatics/btr330

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491–498. doi:10.1038/ng.806

Du H, Diao C, Zhuo Y, Zheng X, Hu Z, Lu S, Jin W, Zhou L, Liu J-F. 2024a. Assembly of novel sequences for Chinese domestic pigs reveals new genes and regulatory variants providing new insights into their diversity. *Genomics* **116:** 110782. doi:10.1016/j.ygeno.2024.110782

Du H, Zhou L, Liu Z, Zhuo Y, Zhang M, Huang Q, Lu S, Xing K, Jiang L, Liu J-F. 2024b. The 1000 Chinese Indigenous Pig Genomes Project provides insights into the genomic architecture of pigs. *Nat Commun* **15:** 10137. doi:10.1038/s41467-024-54471-z

Du H, Zhuo Y, Lu S, Li W, Zhou L, Sun F, Liu G, Liu J-F. 2024c. Pangenome reveals gene content variations and structural variants contributing to Pig characteristics. *Genomics Proteomics Bioinformatics* **22:** qzae081. doi:10.1093/gpbjnl/qzae081

Elhaik E, Tatarinova T, Chebotarev D, Piras IS, Maria Calò C, De Montis A, Atzori M, Marini M, Tofanelli S, Francalacci P, et al. 2014. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun* **5:** 3513. doi:10.1038/ncomms4513

Eusebi PG, Martinez A, Cortes O. 2020. Genomic tools for effective conservation of livestock breed diversity. *Diversity (Basel)* **12:** 8. doi:10.3390/d12010008

Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491:** 393–398. doi:10.1038/nature11622

Guo SC, Savolainen P, Su JP, Zhang Q, Qi DL, Zhou J, Zhong Y, Zhao XQ, Liu JQ. 2006. Origin of mitochondrial DNA diversity of domestic yaks. *BMC Evol Biol* **6:** 73. doi:10.1186/1471-2148-6-73

Halder I, Kip KE, Mulukutla SR, Aiyer AN, Marroquin OC, Huggins GS, Reis SE. 2012. Biogeographic ancestry, self-identified race, and admixture-phenotype associations in the heart SCORE study. *Am J Epidemiol* **176:** 146–155. doi:10.1093/aje/kwr518

Janes JK, Miller JM, Dupuis JR, Malenfant RM, Gorrell JC, Cullingham CI, Andrew RL. 2017. The $K=2$ conundrum. *Mol Ecol* **26:** 3594–3602. doi:10.1111/mec.14187

Kennedy GC, Matsuzaki H, Dong S, Liu W, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, et al. 2003. Large-scale genotyping of complex DNA. *Nat Biotechnol* **21:** 1233–1237. doi:10.1038/nbt869

Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, et al. 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat* **30:** 69–78. doi:10.1002/humu.20822

Larson G, Fuller DQ. 2014. The evolution of animal domestication. *Annu Rev Ecol Evol Syst* **45:** 115–136. doi:10.1146/annurev-ecolsys-110512-135813

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760. doi:10.1093/bioinformatics/btp324

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17:** 10–12. doi:10.14806/ej.17.1.200

Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, et al. 2009. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genet* **10:** 39. doi:10.1186/1471-2156-10-39

Pardo-Seco J, Martinón-Torres F, Salas A. 2014. Evaluating the accuracy of AIM panels at quantifying genome ancestry. *BMC Genomics* **15:** 543. doi:10.1186/1471-2164-15-543

Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, Drineas P. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet* **3:** e160. doi:10.1371/journal.pgen.0030160

Paschou P, Drineas P, Lewis J, Nievergelt CM, Nickerson DA, Smith JD, Ridker PM, Chasman DI, Krauss RM, Ziv E. 2008. Tracing sub-structure in the European American population with PCA-informative markers. *PLoS Genet* **4:** e1000114. doi:10.1371/journal.pgen.1000114

Peng H, Long F, Ding C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* **27:** 1226–1238. doi:10.1109/TPAMI.2005.159

Pérez-Enciso M, Rincón JC, Legarra A. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genet Sel Evol* **47:** 43. doi:10.1186/s12711-015-0117-5

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959. doi:10.1093/genetics/155.2.945

Pudil P, Novovičová J, Kittler J. 1994. Floating search methods in feature selection. *Pattern Recognit Lett* **15:** 1119–1125. doi:10.1016/0167-8655(94)90127-9

Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. 2022. A review of feature selection methods for machine learning-based disease risk prediction. *Front Bioinform* **2:** 927312. doi:10.3389/fbinf.2022.927312

Puechmaille SJ. 2016. The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Mol Ecol Resour* **16:** 608–627. doi:10.1111/1755-0998.12512

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575. doi:10.1086/519795

Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197:** 573–589. doi:10.1534/genetics.114.164350

Reist-Marti SB, Simianer H, Gibson J, Hanotte O, Rege JEO. 2003. Weitzman's approach and conservation of breed diversity: an application to African cattle breeds. *Conserv Biol* **17:** 1299–1311. doi:10.1046/j.1523-1739.2003.01587.x

Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003. Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* **73:** 1402–1422. doi:10.1086/380416

Scherf BD. 2000. *World watch list for domestic animal diversity*, 3rd ed. Food and Agriculture Organization of the United Nations, Rome.

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29:** 308–311. doi:10.1093/nar/29.1.308

Suarez-Pajes E, Díaz-de Usera A, Marcelino-Rodríguez I, Guillen-Guio B, Flores C. 2021. Genetic ancestry inference and its application for the genetic mapping of human diseases. *Int J Mol Sci* **22:** 6962. doi:10.3390/ijms22136962

Toro MA, Fernández J, Caballero A. 2009. Molecular characterization of breeds and its use in conservation. *Livest Sci* **120:** 174–195. doi:10.1016/j.livsci.2008.07.003

Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. 2018. Relief-based feature selection: introduction and review. *J Biomed Inform* **85:** 189–203. doi:10.1016/j.jbi.2018.07.014

van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J Mach Learn Res* **9:** 2579–2605.

Wang X, Ando K, Wu S, Reddy UK, Tamang P, Bao K, Hammar SA, Grumet R, McCreight JD, Fei Z. 2021. Genetic characterization of melon accessions in the U.S. national plant germplasm system and construction of a melon core collection. *Mol Hortic* **1:** 11. doi:10.1186/s43897-021-00014-9

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution (N Y)* **38:** 1358–1370. doi:10.2307/2408641

Yan J, Zou D, Li C, Zhang Z, Song S, Wang X. 2020. SR4R: an integrative SNP resource for genomic breeding and population research in rice. *Genomics Proteomics Bioinformatics* **18:** 173–185. doi:10.1016/j.gpb.2020.03.002

Yang C, Yan J, Jiang S, Li X, Min H, Wang X, Hao D. 2022. Resequencing 250 soybean accessions: new insights into genes associated with agronomic traits and genetic networks. *Genomics Proteomics Bioinformatics* **20:** 29–41. doi:10.1016/j.gpb.2021.02.009

Yaro M, Munyard KA, Stear MJ, Groth DM. 2017. Molecular identification of livestock breeds: a tool for modern conservation biology. *Biol Rev Camb Philos Soc* **92:** 993–1010. doi:10.1111/brv.12265

Zhang W, Zou S, Song J. 2008. Term-tissue specific models for prediction of gene ontology biological processes using transcriptional profiles of aging in *Drosophila melanogaster*. *BMC Bioinformatics* **9:** 129. doi:10.1186/1471-2105-9-129

Zhao S, Shi C-M, Ma L, Liu Q, Liu Y, Wu F, Chi L, Chen H. 2019. AIM-SNPtag: a computationally efficient approach for developing ancestry-informative SNP panels. *Forensic Sci Int: Genet* **38:** 245–253. doi:10.1016/j.fsigen.2018.10.015

Zhao C, Wang D, Teng J, Yang C, Zhang X, Wei X, Zhang Q. 2023. Breed identification using breed-informative SNPs and machine learning based on whole genome sequence data and SNP chip data. *J Anim Sci Biotechnol* **14:** 85. doi:10.1186/s40104-023-00880-x

Zimmerman SJ, Aldridge CL, Oyler-McCance SJ. 2020. An empirical comparison of population genetic analyses using microsatellite and SNP data for a species of conservation concern. *BMC Genomics* **21:** 382. doi:10.1186/s12864-020-06783-9

Received October 30, 2024; accepted in revised form May 22, 2025.