Kinship analysis and pedigree reconstruction by RAD sequencing in cattle

Yiming Xu[1,*], Wanqiu Wang[2,*], Minjie Xu[3,*], Binhu Wang[2], Jiefeng Huang[4], Yingsong Wu[4], Yongzhong Xie[1], Jianbo Jian[2,5,6#]

[1] Animal Husbandry and Aquatic Affairs Center, Lianyuan City, Hunan Province, China.

[2] BGI Genomics, Shenzhen, China.

[3] People's Government of Shexian County, Hebei Province, China.

[4] Loudi Municipal Bureau of Agriculture and Rural Affairs, Loudi City, Hunan Province, China.

[5]Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark.

[6]Marine Biology Institute, Shantou University, Shantou, China.

#Correspondence: Jianbo Jian (jbjian@126.com);

*These authors contributed equally: Yiming Xu, Wanqiu Wang (ORCID: 0000-0002-2748-5867), Minjie Xu.

Abstract: Kinship and pedigree information, used for estimating inbreeding, heritability, selection, and gene flow, is useful for breeding and animal conservation. However, as the size of the crossbred population increases, inaccurate generation and parentage recoding in livestock farms increases. Restriction-site-associated DNA sequencing (RAD-Seq) is a cost-effective platform for single nucleotide polymorphism (SNP) discovery and genotyping. Here, we performed a kinship analysis and pedigree reconstruction for Angus and Xiangxi yellow cattle, which benefit from good meat quality and yields, providing a basis for livestock management. A total of 975 cattle, including 923 offspring with 24 known sires and 28 known dams, were sampled and subjected to SNP discovery and genotyping. The identified SNPs panel included

7305 SNPs capturing the maximum difference between paternal and maternal genome information allowing us to distinguish between the F1 and F2 generation with 90% accuracy. In addition, parentage assignment software based on different strategies verified that the cross-assignments. In conclusion, we provided a low-cost and efficient SNP panel for kinship analyses and the improvement of local genetic resources, which are valuable for breed improvement, local resource utilization, and conservation.

Keywords: crossbreeding, RAD-seq, generation classification, parentage assignment

Introduction

The selection of superior individuals as parents in crossbreeding systems is a general tool for genetic improvement (Gregory et al. 1980). Genetic selection is a new approach to selective breeding based on high-density markers covering the entire genome; it shortens the breeding cycle and reduces breeding costs. Predicting progeny from parental lines and selecting the best crosses rely on the accuracy of the pedigree; however, parentage assignment is expensive and laborious in normal livestock production systems. Therefore, the development of reliable genetic markers for kinship analyses and pedigree reconstruction is necessary to meet industry demand and improve livestock management systems.

SNPs are abundant, widespread in the genome, and follow simple models of evolution (Morin et al. 2004). There are various genotyping approaches, including bovine SNP chip, restriction-site-associated DNA sequencing (RAD-seq) (Baird et al. 2008), multiplexed shotgun genotyping (Andolfatto et al. 2011), exome sequencing (Ng et al. 2009), and whole-genome resequencing (WGS). Traditional SNP genotyping methods, which necessitate specific probes or primers tailored to each SNP of interest, can incur substantial costs, particularly for large-scale genotyping efforts. Moreover, these processes are often labor-intensive and time-

consuming, especially in the absence of high-throughput methodologies. Additionally, they may necessitate sophisticated bioinformatics for data analysis, further augmenting the associated costs. In contrast, RAD sequencing generally employs a restricted set of restriction enzymes and barcodes, which can significantly reduce reagent costs compared to individual SNP assays. Despite the complexity of next-generation sequencing (NGS) data, a plethora of tools and pipelines, some open-source, are available to manage RAD data, thereby mitigating the costs of data analysis. Consequently, RAD sequencing is adept at handling large-scale projects and proves more cost-effective for the genotyping of numerous samples or a comprehensive set of SNPs. RAD-seq's appeal extends beyond its economic and temporal advantages, as it is also adept at analyzing complex genomic sequences, unconstrained by the size of the genome. This technique has been effectively leveraged in the construction of linkage maps, comparative genomics, and population genetics studies across a diverse array of organisms

Parentage analysis can generally be divided into two categories: exclusion and likelihood (Flanagan et al. 2019). The former is designed to identify incompatibilities between pairs of individuals to prove that one cannot be the parent of the other (Chakraborty et al. 1974). Parentage assignment using likelihood is well-developed approach (Kalinowski et al. 2007; Marshall et al. 1998). However, different from the rapid development of parentage assignment methods and related software over the past decade (Flanagan et al. 2019), little work has focused on generation classification, an important first step.

Due to its high carcass yield, rapid growth rate, and marbled meat quality, the Angus breed has become the predominant cattle breed in numerous countries. Xiangxi yellow cattle is an indigenous Chinese breed and was included in the *National Protection List of Livestock and*

Poultry Genetic Resources of China in 2006 (Li et al. 2014). This breed can feed on low-quality roughage, is good at climbing hills, and is well-adapted to high temperatures. To improve meat quality and quantity, Angus was crossbred with Xiangxi yellow cattle, creating a new hybrid known as Xiangzhong black cattle, which captures the beneficial traits of each breed. The F1 generation backcrossed with sires to produce the F2 generation. To reduce breeding costs, it is common to breed F1 and F2 generations in one barn after crossbreeding. However, as the size of the crossbred population increases, inaccurate recording of the generation and parentage are more frequent due to error. Therefore, assigning individuals to each generation of the crossbreeding program is necessary, especially for genetic improvement and local resource conservation.

From the F1 generation, the phenotype of the offspring is intermediate between those of parent and the female, and it is difficult to distinguish the F1 and F2 generations just by inspection. In the absence of accurate records from local livestock farms, it is nearly impossible to distinguish between the F1 and F2 generations. Therefore, in this study, we evaluated 975 cattle from livestock farms by RAD-seq, including sires (Angus), dams (Xiangxi yellow cattle), the F1 generation from a cross between dams and sires, and the F2 generation from the backcross between F1 and sires and the intercross between F1. Our aims were 1) to distinguish between the F1 and F2 generations according to Mendelian laws of inheritance and 2) to construct a SNP panel with high confidence for kinship analysis and pedigree reconstruction for Xiangzhong black cattle.

Material and Methods

Ethics statement

We declare that all of the experimental procedures involving animals were conducted following the guidelines for experimental animals and were approved by the institutional review board of BGI (approval number BGI-IRB E22013). The samples were collected from a local farm in Loudi, Hunan Province, China.

Sample information

In total, 975 cattle were sampled, including 923 calves, 28 dams, and 24 sires. First, 28 Xiangxi yellow cattle dams were crossed with 24 Aberdeen-angus sires. Semen was collected from Angus bulls and crossed with females of the F1 generation to produce the F2 generation. All F1 and F2 individuals were cultured together in livestock farms. Mating between bulls and cows of the F1 generation was possible. Sires and dams were clearly labeled; however, for offspring cattle, the paternal line was ambiguous.

Library preparation and RAD-seq

For DNA extraction, fresh blood was collected and prepared according to the protocol of solution-based DNA extraction methods (Chacon Cortes et al. 2014). After quality control, DNA samples were subjected to single-digest RAD-seq as described previously, with slight modifications (Baird et al. 2008). Briefly, samples meeting quality standards were digested with TaqI restriction endonuclease for 20 min at 37°C and then randomly fractionated by Covaris Focused-Ultrasonicator. Then, 5 μl of RAD adapter was added to the interruption product. The products were cyclized by rolling circle amplification. The libraries were sequenced on the BGISEQ-500 platform (BGI, Shenzhen, China).

SNP genotyping and selection

The raw sequence data were filtered and trimmed using SOAPnuke (Chen et al. 2018) and separated according to the unambiguous barcodes and the specific enzyme recognition site. Each RAD read was mapped to the *Bos taurus* genome (https://ftp.ncbi.nih.gov/genomes/all/GCF/002/263/795/GCF_002263795.1_ARS-UCD1.2/) using BWA (v0.7.12-r1039) (Li et al. 2010). The bam format was sorted and indexed using Samtools (v1.14) with default parameters except markdup (Li et al. 2009). Then the format was transefered to sam format using Picard (v2.26.10). , The GATK (v4.1.2) were used for SNP and InDel calling with the Variant Call Format (VCF) file with default parameters (McKenna et al. 2010).

Quality control of candidate SNPs was performed using GATK "VariantFiltration." All SNPs that met the criteria for minimizing false positives (QUAL < 250.0, DP < 1500, DP > 6000, MQ < 50.0 || QD < 4.0 || FS > 15.0 || BaseQRankSum < -4.0 || ReadPosRankSum < -3.50 || MQRankSum < -10.0, SOR > 4.0, AN < 975, MQ0 ≥30) were considered potential high-quality SNP markers for subsequent analyses.

For the classification of generations, considering that there were 24 and 28 sires and dams, respectively, alleles with counts of 48 and 56 indicated homozygous loci. To expand the criteria to obtain adequate loci, we allowed for five different alleles, which could reflect sequencing error. Therefore, we selected allele counts of 43–48 and 51–56 in the paternal and maternal genome along with different allelic genotypes for further analysis. Loci with allele depths below 6 were excluded from the SNP set.

For parentage assignment, SNPs with more than two alleles were first excluded. The retained SNPs from paternal data meeting the following criteria were extracted to form an allele set: MAF (minor allele frequency) > 0.25, heterozygosity > 0.2, proportion of missing data < 0.7,

144 *p*-value for Hardy–Weinberg equilibrium > 0.01. Maternal and offspring data were filtered
145 using this allelic set. Finally, sites with an allele depth below 6 and proportion of missing data
146 above 0.8 were excluded.

147 ## Classification of F1 and F2 generations

148 All samples were genotyped based on the SNP panels used for generation classification and
149 loci that were identical to the maternal/paternal allele or were heterozygous were counted. The
150 heterozygous ratio was defined as the number of heterozygous SNPs divided by the total SNPs
151 genotyped. The number of identical maternal/paternal SNPs divided by the total number of
152 SNPs successfully genotyped was defined as the ratio of SNPs inherited from maternal/paternal
153 parents. Samples with a heterozygous ratio above 0.5 were classified as F1 and the remaining
154 samples were assigned to the F2 generation.
155
156 Scatter diagrams of the results were generated using R. For phylogenic tree construction, the *p*-
157 distance matrix was calculated using VCF2Dis-1.44 (https://github.com/BGI-
158 shenzhen/VCF2Dis) and a neighbor-joining tree was generated using PHYLIPNEW with
159 fneighbor (v3.69.650) (Rice et al. 2000). Finally, visualization was performed using Mega
160 (Tamura et al. 2007).
161
162 The population genetic structure was evaluated using the program Admixture. The number of
163 assumed genetic clusters K was set to 5, with 10,000 iterations for each run. The SNP selection
164 was performed based on the following strategies: SNPs were chosen from loci with frequencies
165 ranging from 43 to 48 and 51 to 56 among paternal and maternal data, which exhibited distinct
166 allelic genotypes. The SNP panel comprised a total of 7305 loci. Identity-by-descent (IBD) was
167 analyzed using the PLINK kinship matrix generated based on 7305 SNPs using the kinship

matrix function (-method Centered_IBS) in TASSEL (Bradbury et al. 2007) (v5.0) and visualized using the heatmap package of R.

Parentage assignment

Parentage assignment was performed using CERVUS (v3.0.7) with the paternity selection function based on the likelihood approach. Based on a simulation, 959 loci showed 91% successful assignment at the 95% confidence level. The R package APIS (v1.0.1) (Griot et al. 2020) was used for parentage assignment based on the distribution of Mendelian transmission probabilities. The assignment error rate was set to 0.05. The first candidate sire given by the software was considered the most probable sire for the offspring and was used for further analyses. Exclusion-based parentage assignment was performed using the hiphop package (v0.0.1) in R (Cockburn et al. 2021). None of the dams and sires were defined as social parents. The first-ranked candidate parent pair was selected as the most likely parent and subjected to subsequent analyses.

Results

Discovery of SNPs by RAD-seq

To enhance the quality of beef and maintain local genetic resources, local farms introduced Aberdeen Angus as sires and crossed them with a local breed, Xiangxi yellow cattle (Figure 1A). Female F1 cattle were backcrossed with sires to produce the F2 generation (Figure 1A). Theoretically, if a sire and dam have different genotypes (AA and aa), the F1 generation will be heterozygous (Aa) and the F2 generation will contain two genotypes (AA, Aa). The F2

generation can be used for further crossbreeding, genetic improvement, and resource conservation. To reduce costs, farmers usually do not breed F1 and F2 generations in separate barns. Furthermore, mating is possible between bulls and cows of the F1 generation. Angus cattle and Xiangzhong Yellow cattle differ mainly in body color and size, i.e., Angus cattle are black and Xiangzhong Yellow cattle are brown. However, their offspring are mainly black and it is difficult to distinguish between F2 progeny based on appearance, without genotyping. The phenotype of F1 hybrids is similar to that of F2 carrying AA and Aa genotypes. In this study, we evaluated 975 cattle by RAD-seq, including 923 offspring with 24 known sires and 28 known dams. All samples were collected from Hunan province, China.

A total of 4.65 Tb of clean data were obtained from the samples, with a mean of 35.59 million reads, 1.59× depth, and 31.06% coverage per individual. An average of 35.96 million reads were mapped to the *B. taurus* reference genome ARS-UCD1.2, with an average mapping rate of 99.8%.

Genetic variants were initially detected by GATK and then underwent preliminary filtering. A total of 8,155,410 SNPs and 1,100,880 InDel variants that satisfied the criteria were retained (Supplementary Tables 1,2). A density plot revealed that SNPs and InDels were preferentially distributed in the proximal telomeric regions (Supplementary Figures 1,2). NC_037357.1, the X chromosome of *B. taurus,* had relatively few SNPs and InDels. The X-chromosome is present in a single copy in males. Compared with those of autosomes, the X-chromosome has a lower recombination rate, lower mutation rate, and smaller effective population size (Schaffner, 2004).

Construction of a SNP panel retaining parental information

Our first aim was to assign cattle lacking clear identifying information to each generation of the crossbreeding program. We had accurate generation information of 28 offspring from the farm record. According to Mendel's laws of inheritance, in a dominant-recessive inheritance system, parents with different homozygous genotypes (AA and aa) will produce heterozygous F1 offspring (Aa). In the backcross of F1 to sires, all F2 progeny show the dominant trait, 50% of individuals will be homozygous and 50% will be heterozygous, on average. If the F1 are intercrossed, on average, the F2 generation is expected to be 25% homozygous with the dominant trait, 50% heterozygous showing the dominant trait (genetic carriers), and 25% homozygous with the recessive trait. The genotypic ratio is 1 (AA):2 (Aa):1 (aa), and the phenotypic ratio is 3:1. Accordingly, homozygous loci from parents with different genotypes may provide pivotal information for identifying the generation of individuals (Figure 1B).

Retained SNPs were further filtered to ensure the maximum difference between paternal and maternal genome information. The workflow used to generate a high confidence SNP panel consisting of a series of the homozygous loci is provided in Figure 1C. First, we chose SNPs that were homozygous in sires (AA) and dams (aa) for alternative alleles. However, only 45 loci were kept from 24 paternal and 28 maternal SNPs. Considering that the sequencing depth of some samples was not sufficient for genotyping, the criteria were expanded slightly to obtain adequate loci. The loci with frequencies of 43–48 and 51–56 among paternal and maternal data and exhibiting different allelic genotypes were selected as candidates for classification. Finally, the SNP panel included 7305 loci.

Subsequently, we constructed a phylogenetic tree based on the panel with data for 24 sires, 28 dams, 15 offspring of the F1 generation, and 13 offspring of the F2 generation. All sires and dams were clearly assigned to distant groups in the phylogenetic tree. In addition, the F1 and

F2 generations were generally separated (Figure 2A). However, it was still difficult to distinguish between F1 and F2 based on these data. Accordingly, it is necessary to consider both the percentage of heterozygous loci and the percentage of loci inherited from dams and sires.

Assigning individuals to F1 and F2

From data for the F1 and F2 generation, we extracted the genotypes of all samples using the high confidence SNP panel described above and compared them with the genotypes of sires and dams. The numbers of SNP loci in offspring with the maternal/paternal genotype (paternal/paternal genotype locus, MGL/PGL) were used to calculate the maternal/paternal ratio, and heterozygous loci (H-genotype) were used to calculate the heterozygous ratio. The maternal/paternal ratio was defined as the ratio of MGL/PGL to all genotypes. As mentioned above, all samples belonging to the F1 generation should have heterozygous alleles and these alleles should segregate independently in the F2 generation. However, since we extended the criteria for the SNP panel, it is possible to obtain MGL or PGL in the F1 generation. Consequently, the offspring samples with heterozygous locus percentages above 0.5 were assigned to the F1 generation and the rest were classified as F2. In the F2 generation, offspring with a maternal/paternal ratio above 0.5 were defined as having the most maternal/paternal characteristics.

As expected, 21 out of 28 offspring were correctly assigned based on farm records. A scatter diagram of all samples revealed four groups corresponding to sires (uppermost group in Figure 2B), dams (lower), F1 (left corner), and F2 (near dams) (Figure 2B). Five samples were assigned to the F1 generation but belonged to the F2 generation. It is possible that the F2 individuals were produced from F1 inter-crossing, which influenced the final classification

results (Figure 1B). However, considering that the selection of superior individuals with respect to performance traits from sires and dams is the ultimate goal of this crossbreeding system, it is not necessary to distinguish between these individuals.

The newly developed SNP panel was used to genotype a total of 975 samples. Based on SNP information, almost all offspring were categorized successfully (Supplementary Table 3). Samples were clearly divided into five groups: the F1 generation (heterozygous ratio > 0.5, 616 samples), the F2 generation with strong paternal characteristics (MSF2, heterozygous ratio > 0.5 and paternal ratio > 0.5, 227 samples), the F2 generation with strong maternal characteristics (MDF2, heterozygous ratio > 0.5 and maternal ratio > 0.5, 73 samples), the F2 generation with slight paternal characteristics (LSF2, heterozygous ratio > 0.5 and paternal ratio > maternal ratio, 37 samples), and the F2 generation with slight maternal characteristics (LDF2, heterozygous ratio > 0.5, maternal ratio > paternal ratio, 26 samples). The scatter diagram supported the classification of samples into five groups (Figure 2C). Subsequently, we generated a phylogenetic tree based on the information for all samples (Figure 2D). Samples classified as F2 were distributed across the phylogenetic tree, with one group clustered near sires (MSF2) and another group clustered near dams (MDF2). The samples located in an intermediate position were considered F1 or F2 with slight parental characteristics (LSF2 and LDF2) (Supplementary Table 3).

Validation of classification results

To confirm the correctness of classification, we performed a population structure analysis using all SNPs to observe the systematic differences in allele frequencies among subpopulations (Figure 3A). In total, 975 samples were divided into five subpopulations, consistent with the results for the classification of generations.

Gene IBD is a fundamental concept that explains genetically mediated similarities among relatives (Thompson, 2013). The PI_{HAT} value $(\hat{\pi}, proportion\ IBD)$ obtained using PLINK provides an estimate of IBD. This method is based on the hidden Markov model (HMM) and calculates the probability of IBD = 1, 2, or 0 by method-of-moments estimation. The value of PI_HAT is between 0 and 1, and a higher value indicates a closer relationship. We performed an IBD analysis to evaluate relationships among individuals. Usually, pairs with PI_HAT near 1 are considered identical twins, those with values of approximately 0.5 are considered parent/child or parental identical twins, and those with values of around 0.25 are considered half-siblings. We used the homozygous sites from sires and dams to calculate PI_HAT. All pairs with a PI_HAT value above 0.5 were pairs between sires or dams. Most pairs with PI_HAT values around 0.5 (0.3–0.5) were pairs between sires/dams and MSF2/MDF2 (Supplementary Table 7). These results were in accordance with our previous assignments to provide a basis for selecting MSF2 for breeding and other F2 for conservation.

Finally, we created kinship matrices for individuals containing sires, dams, and the F1 generation. These were multiplied by two to indicate expected covariances between samples and were used for parentage assignment (Figure 3B, Supplementary Table 4). This approach proved to be effective for distinguishing between heterozygous offspring and candidate homozygous offspring with parental traits.

Parentage assignment of the F1 generation

For parentage assignment with the F1 generation, we chose three tools with different underlying algorithms and compared the results. CERVUS is widely and utilizes a likelihood-based

approach, APIS uses the observed distribution of Mendelian transmission probabilities, and HIPHOP extends exclusion approaches with SNP markers.

In total, 960 SNPs that met the quality criteria were further filtered to determine loci that could provide sufficient statistical power for parentage assignment. CERVUS requires allele frequency estimation and a simulation before formal assignment. Simulation can be used to examine the feasibility of parentage analysis and calculate critical values of likelihood ratios to detemined the confidence of parentage assignments for real data. The following parameters were used for simulation: 30% loci typed, 0.01% genotyping error rate, 10,000 offspring, 50 candidate fathers, 90% of candidate fathers sampled, and the minimum number of typed loci set to 10. The output of strict confidence value of simulation was 95%. Finally, by input the simulation and SNP files, the set containing 28 sires and the panel of 960 SNP markers provided 99.5% successful parentage assignment for 615 offspring (Supplementary Table 5). Among 612 F1–sire relationships, 552 had confidence levels of >80%, including 451 with confidence ≥95%, which indicated a highly significant relationship. The accepted assignment error rate was 0.5 using APIS (Supplementary Figure 3). HIPHOP required a known social parent and year of birth for each cattle at the time of parentage assignment. If the individual is the social parent of the brood, then the social parent parameter was set to 1; otherwise, the parameter was set to 0. All cattle were bred in livestock farms and none were classified as a social parent. We then compared the results obtained by the three algorithms and drew a Venn diagram to visually evaluate shared assignments (Figure 3B). The results obtained using CERVUS and APIS overlapped by more than 75% (450 identical assignments), while HIPHOP assignments agreed with those of the other algorithms in <50% of cases (Figure 3C, Supplementary Table 5).

As mentioned before, HIPHOP relies on exclusion methods. This method identifies incompatibilities between pairs of individuals according to Mendel's laws, and its accuracy depends on the accuracy of the marker data. When the power of the marker set is low, the results become unauthentic. However, it is nearly impossible to obtain error-free data as the sample size and number of markers grow. Consequently, the overlapping assignments obtained by CERVUS and APIS were considered the likely parentage assignments with high confidence (Supplementary Table 6).

Discusstion

RAD-Seq to genotype SNPs is a practical and cost-effective approach compared to traditional methods such as microarrays. Choosing informative SNPs reduces the computational burden of downstream analyses, and allows us to obtain the desired level of accuracy without generating excessive amounts of data or compromising quality. Overall, RAD-Seq remains an efficient tool for genotyping and has wide applications in population genetics research. Our SNP panel offers a cost-effective tool for livestock management by enabling breeders to make informed decisions on the selection of animals for breeding programs. The detailed genetic information provided by RAD-Seq can lead to improved health, productivity, and overall welfare of the livestock. The SNP markers identified in our study can be used to trace and preserve valuable genetic traits within local breeds. This can facilitate targeted breeding strategies to enhance specific characteristics, such as disease resistance, adaptability, and production efficiency. This SNP panel also can be a valuable tool for conservation efforts by providing a means to monitor and manage the genetic diversity of endangered species. It can also aid in the detection of hybridization events and inform conservation breeding programs. We propose that future studies could focus on validating the SNP panel across a broader range of breeds and

environments. The SNP panel could be integrated into existing breeding programs, starting with a pilot study to assess its effectiveness and economic benefits. This would help to establish the panel's versatility and robustness in different contexts.

To assign individuals to the F1/F2 generation, a panel of SNPs can be used to determine the number of loci that are identical to the maternal and paternal alleles or are heterozygous. This allows for the calculation of maternal, paternal, and heterozygous ratios for each offspring. Theoretically, the heterozygous ratio for the F1 generation will be around 1.0, while the maximum ratio was 0.83 in our analysis. Furthermore, the F2 generation was divided into four subclasses according to the maternal ratio and paternal ratio. This likely reflects the law of segregation and the law of independent assortment for non-allelic genes on non-homologous chromosomes, whereas the law of linkage and crossing-over would lead to deviations from the expected ratio. Unexpectedly, the F2 generation contained the MDF2 and LDF2 subgroups. Considering that the F1 generation originated from the cross between dams and sires and the F2 generation from the backcrosses between the F1 generation and sires theoretically, this observation can likely be explained by crossing within the F1 generation, since the proportions of MDF2 and LDF2 in the whole F2 generation were low. In addition, during the SNP selection stage, we allowed five different alleles, which could be the result of sequencing error. This is another reason why the heterozygous ratio of the F1 generation was less than 1.

Although several methods for parentage analysis have been developed, accurate parentage assignment critically depends on the establishment of a molecular marker panel. Accurate pedigree information has been determined in a Mexican registered Holstein population (García-Ruiz et al. 2019) and in Chinese Simmental cattle based on a high-density SNP array (Zhang et al. 2018). Most parentage assignments rely on a microsatellite chip or SNP chip. In taxa without

accurate SNP chips for parentage assignment, it is difficult to achieve this goal. RAD-seq and WGS provide abundant variant information, which enables parentage and population analyses at the same time. Of note, the SNP panel used for parentage assignment should be optimized based on several factors, including linkage disequilibrium, minor allele frequencies, deviation from Hardy-Weinberg equilibrium, genotyping errors, and the frequency of null alleles. In this study, we achieved parentage assignment by extracting the intersection between CERVUS and APIS results. However, 165 offspring could not be assigned to accurate sires. This could probably be attributed to an inherent defect of RAD-seq resulting in a high frequency of null alleles. Null alleles cannot provide accurate predictions, thus weakening the confidence of the prediction results.

Another feasible approach to achieving parentage assignment is based on the genotypes of sex chromosomes and mitochondrial chromosomes. Typically, the X chromosome and Y chromosome do not undergo standard recombination in males, and mitochondrial chromosomes are inherited directionally from the mother (Ballard et al. 2004; Birky Jr 2001; Schaffner 2004). Therefore, alleles on the sex chromosome and mitochondria will be transferred to offspring directly. The Y chromosome is difficult to assemble because it contains many ampliconic and palindromic sequences (Colaco et al. 2018). It is also obvious that SNPs and InDels were sparsely distributed on the X chromosome (Supplementary Figure 1,2) except for the tip region. In conclusion, genotyping by RAD-seq is an efficient method for the classification of generations and parentage assignment. Individuals from different generations identified through the analysis can be used to accelerate the subsequent breeding process and breeding conservation.

Conclusion

Crossbreeding is a widely used and effective tool to increase genetic diversity within a breed. However, successful crossbreeding relies on accurate marking and recording of newborn calves, which is laborious and increases in difficulty as the calf population grows. Traditional SNP genotyping may be more suitable for smaller projects or when specific SNPs are the focus of the research. RAD sequencing is generally considered a more cost-effective and scalable solution for large-scale SNP genotyping projects due to its lower per sample costs and ability to multiplex many samples simultaneously. The choice between the two methods should be based on the specific requirements of the project, including the number of samples, the number of SNPs to be genotyped, and the available resources. In this study, we perfomed RAD sequencing and identified the F1 and F2 generation from Angus cattle and Xiangxi yellow cattle crosses according to Mendelian laws of inheritance and selected a SNP panel with high confidence for a kinship analysis and pedigree reconstruction. The F1 generation and MSF2 can be applied for breed selection and the LDF2 and MDF2 generation can be maintained for breed conservation. To the best of our knowledge, this is the first application of a RAD-seq-based approach for simultaneous generation classification and parentage assignment. The combination of the efficiency of RNA-seq and advances in kinship analysis is expected to improve breed management, local resource utilization, and conservation.

Data Availability

The raw sequencing reads were deposited at NCBI under PRJNA1063367, the SNP file was deposited under fighshare (https://doi.org/10.6084/ m9.figshare.25807246).

Figures

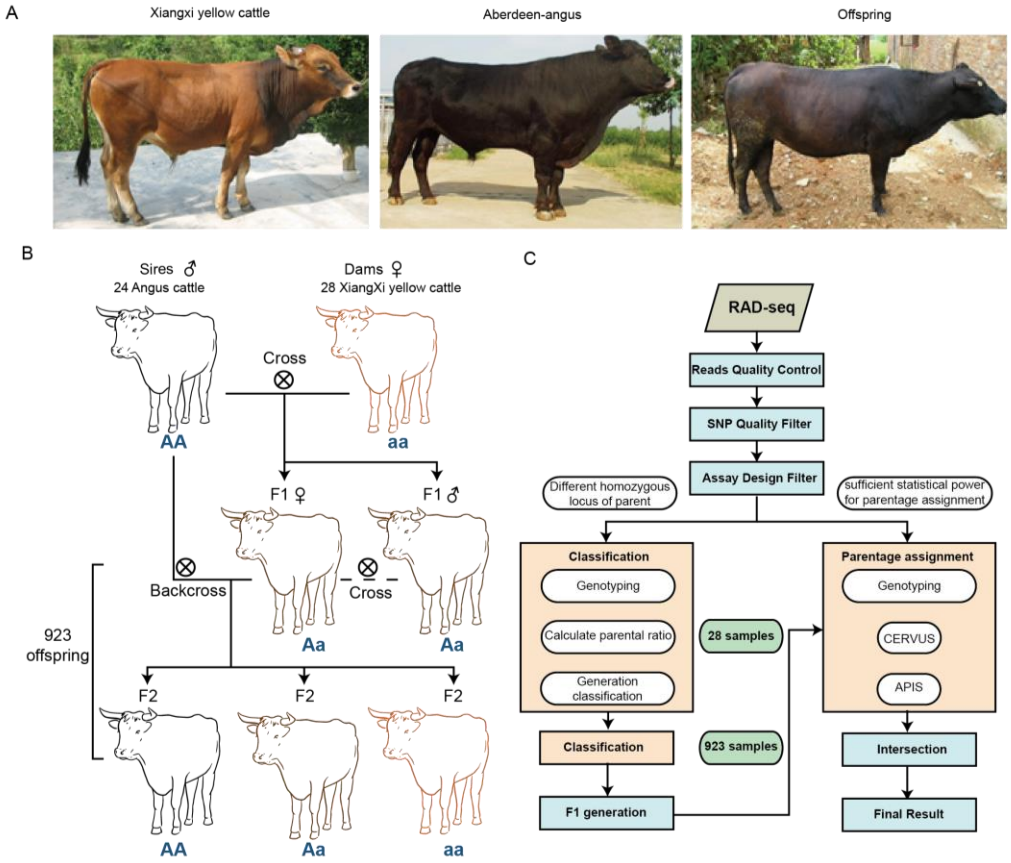
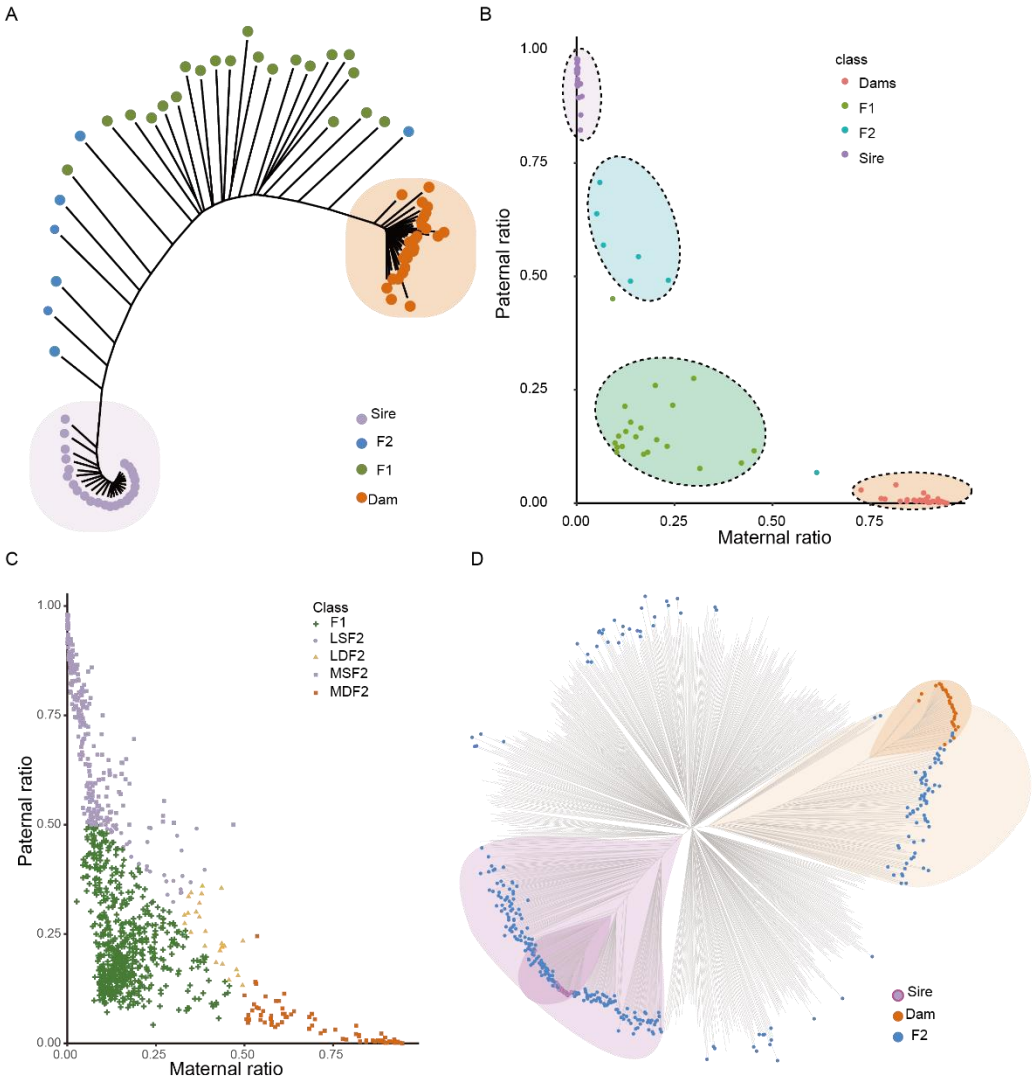


Figure 1: Workflow for generation classification and the selection of a SNP panel retaining parental information

A. Representative pictures of the three breeds

B. Schematic diagram of the crossbreeding system

C. Flowchart detailing the workflow for generation classification and parentage assignment



435

436 **Figure 2: Differentiation between the F1 and F2 generation from a group of cattle**

437 A. Phylogenetic tree of 80 samples containing sires, dams, and offspring with known

438 generation information.

439 B. Scatter diagram of 80 samples according to the paternal ratio and maternal ratio.

440 C. Scatter diagram of 923 samples according to the paternal ratio and maternal ratio. Different

441 colors represent different subclasses.

442 D. Phylogenetic tree of all 975 samples containing sires, dams, and offspring.

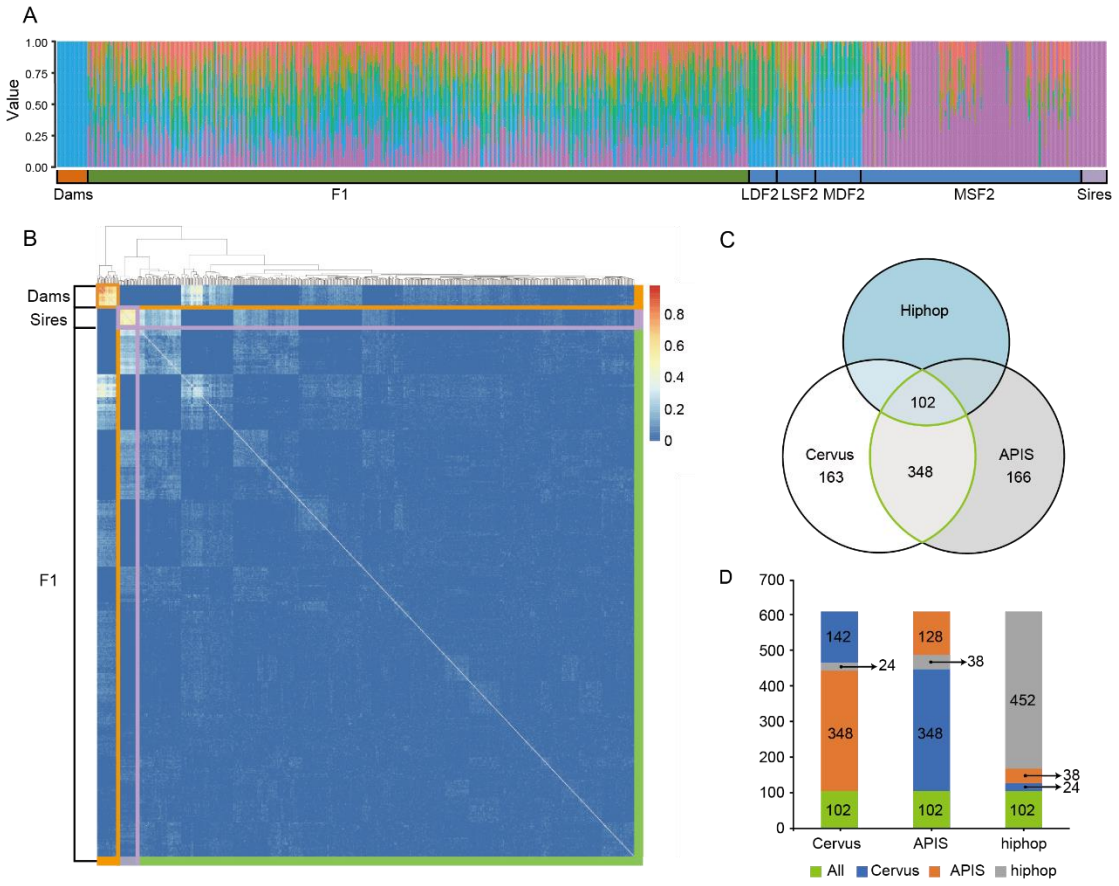


Figure 3: Parentage assignment of the F1 generation

A. Population structure analysis of 975 samples with K = 5.

B. Heatmap representing the correlations between 615 samples using kinship matrices.

C. Venn diagram indicating parentage assignments obtained using three algorithms. The green oval represents the intersection between Cervus and APIS, defined as the final assignments.

D. Stacked columns representing the composition within each parentage assignment results based on three programs.

REFERENCES

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T.T., Mast, J., Sunayama-Morita, T., Stern, D.L., Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome research 2011. 21, 610–617.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A.,. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one 2008. 3, e3376.

Ballard, J.W.O., Whitlock, M.C., The incomplete natural history of mitochondria. Molecular ecology 2004. 13, 729–744.

Birky Jr, C.W., The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annual review of genetics 2001. 35, 125–148.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S., TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 2007. 23, 2633–2635.

Chacon Cortes, D.F., Griffiths, L., Methods for extracting genomic DNA from whole blood samples: current perspectives. Journal of Biorepository Science for Applied Medicine 2014, 1–9.

Chakraborty, R., Shaw, M., Schull, W.J., Exclusion of paternity: the current state of the art. American Journal of Human Genetics 1974. 26, 477.

Chen, Yuxin, Chen, Yongsheng, Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience 2018. 7, gix120.

Cockburn, A., Peñalba, J.V., Jaccoud, D., Kilian, A., Brouwer, L., Double, M.C., Margraf, N., Osmond, H.L., Kruuk, L.E.B., Pol, M., HIPHOP : Improved paternity assignment among close relatives using a simple exclusion method for biallelic markers. Mol Ecol Resour 2021. 21, 1850–1865.

Colaco, S., Modi, D., Genetics of the human Y chromosome and its association with male infertility. Reproductive Biology and Endocrinology 2018. 16, 14.

Flanagan, S.P., Jones, A.G., The future of parentage analysis: From microsatellites to SNPs and beyond. Mol Ecol 2019. 28, 544–567.

García-Ruiz, A., Wiggans, G.R., Ruiz-López, F.J., Pedigree verification and parentage assignment using genomic information in the Mexican Holstein population. Journal of

Dairy Science 2019. 102, 1806–1810.

Gregory, K.E., Cundiff, L.V., Crossbreeding in Beef Cattle: Evaluation of Systems1. Journal of Animal Science 1980. 51, 1224–1242.

Griot, R., Allal, F., Brard-Fudulea, S., Morvezen, R., Haffray, P., Phocas, F., Vandeputte, M., APIS: An auto-adaptive parentage inference software that tolerates missing parents. Mol Ecol Resour 2020. 20, 579–590.

Kalinowski, S.T., Taper, M.L., Marshall, T.C., Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular ecology 2007. 16, 1099–1106.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., The sequence alignment/map format and SAMtools. Bioinformatics 2009. 25, 2078–2079.

Li, L., Zhu, Y., Wang, X., He, Y., Cao, B., Effects of different dietary energy and protein levels and sex on growth performance, carcass characteristics and meat quality of F1 Angus × Chinese Xiangxi yellow cattle. Journal of Animal Science and Biotechnology 2014. 5, 21.

Marshall, T.C., Slate, J., Kruuk, L.E.B., Pemberton, J.M., Statistical confidence for likelihood-based paternity inference in natural populations. Molecular ecology 1998. 7, 639–655.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 2010. 20, 1297–1303.

Morin, P.A., Luikart, G., Wayne, R.K., the SNP workshop group, SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution 2004. 19, 208–216.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009. 461, 272–276.

Rice, P., Longden, I., Bleasby, A., EMBOSS: the European molecular biology open software suite. Trends in genetics 2000. 16, 276–277.

Schaffner, S.F., The X chromosome in population genetics. Nature Reviews Genetics 2004. 5, 43–51.

519 Tamura, K., Dudley, J., Nei, M., Kumar, S., MEGA4: molecular evolutionary genetics analysis
520 (MEGA) software version 4.0. Molecular biology and evolution 2007. 24, 1596–1599.
521 Thompson, E.A., Identity by Descent: Variation in Meiosis, Across Genomes, and in
522 Populations. Genetics 2013. 194, 301–326.
523 Zhang, T., Guo, L., Shi, M., Xu, L., Chen, Y., Zhang, L., Gao, H., Li, J., Gao, X., Selection and
524 effectiveness of informative SNPs for paternity in Chinese Simmental cattle based on a
525 high-density SNP array. Gene 2018. 673, 211–216.
526