ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

doi:10.1111/j.1365-2052.2011.02198.x

Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing

A. M. Ramos*, H. J. Megens*, R. P. M. A. Crooijmans*, L. B. Schook† and M. A. M. Groenen*

*Animal Breeding and Genomics Centre, Wageningen University, Marijkeweg 40, 6709 PG Wageningen, The Netherlands. †Department of Animal Sciences, University of Illinois, 374 Edward R. Madigan Lab, 1201 West Gregory Drive, Urbana, IL 61801, USA

Summary

The objectives of this study were to develop breed-specific single nucleotide polymorphisms (SNPs) in five pig breeds sequenced with Illumina's Genome Analyzer and to investigate their usefulness for breed assignment purposes. DNA pools were prepared for Duroc, Landrace, Large White, Pietrain and Wild Boar. The total number of animals used for sequencing was 153. SNP discovery was performed by aligning the filtered reads against Build 7 of the pig genome. A total of 313 964 high confidence SNPs were identified and analysed for the presence of breed-specific SNPs (defined in this context as SNPs for which one of the alleles was detected in only one breed). There were 29 146 putative breed-specific SNPs identified, of which 4441 were included in the PorcineSNP60 beadchip. Upon re-examining the genotypes obtained using the beadchip, 193 SNPs were confirmed as being breed specific. These 193 SNPs were subsequently used to assign an additional 490 individuals from the same breeds, using the sequenced individuals as reference populations. In total, four breed assignment tests were performed. Results showed that for all methods tested 99% of the animals were correctly assigned, with an average probability of assignment of at least 99.2%, indicating the high utility of breed-specific markers for breed assignment and traceability. This study provides a blueprint for the way next-generation sequencing technologies can be used for the identification of breed-specific SNPs, as well as evidence that these SNPs may be a powerful tool for breed assignment and traceability of animal products to their breeds of origin.

Keywords assignment test, breed-specific single nucleotide polymorphisms, next-generation sequencing, pig, traceability.

Introduction

In the last few years, we have witnessed a very rapid development in the field of next-generation sequencing technologies. Presently, several technologies are able to generate large volumes of sequence data in a fast, accurate and inexpensive way, including Illumina's Genome Analyzer (Bennett 2004), Roche's 454 (Margulies *et al.* 2005), ABI's SOLiD (Shendure *et al.* 2005), Helicos (Milos 2008), and Pacific Biosciences' real-time sequencing (Eid *et al.* 2009). The high-throughput nature of these technologies has greatly accelerated the pace of scientific discovery in recent years.

Address for correspondence

A. M. Ramos, Animal Breeding and Genomics Centre, Wageningen University, Marijkeweg 40, 6709 PG Wageningen, The Netherlands.
E-mail: marcos.ramos23@gmail.com

Accepted for publication 9 January 2011

Single nucleotide polymorphisms (SNPs) have become the marker of choice for many studies in animal genetics and genomics, and SNP identification has benefited greatly from the rapid development in next-generation sequencing technologies. To date, several studies have led to the detection of thousands of SNPs in different species (Wiedmann *et al.* 2008; Kerstens *et al.* 2009; Ramos *et al.* 2009; Sánchez *et al.* 2009; van Bers *et al.* 2010). Currently, high-density SNP genotyping assays that allow thousands of genotypes to be obtained simultaneously are already available for several livestock species, including pigs (Ramos *et al.* 2009) and cattle (Matukumalli *et al.* 2009). These tools and developments will play a central role in future studies in animal genetics and genomics.

The potential of several DNA-based methods for the identification of animals at different levels, from individual to breeds and species, has also been investigated. This trend has followed the increasing attention paid by the general public to the origins of the food it consumes. Therefore,

there is a need for systems that can increase consumer confidence in product safety by reliably assuring traceability of food, including products of animal origin such as meat and dairy products. To date, several types of DNA markers have been tested for their potential use in traceability schemes, including microsatellites (Dalvit *et al.* 2008a,b), AFLPs (Negrini *et al.* 2007) and SNPs (Negrini *et al.* 2008a,b). These studies explored the assignment of individuals to their breeds of origin, using likelihood, frequency and Bayesian-based methods, with results showing some promise, but also some limitations. Next-generation sequencing technologies now offer new and unprecedented possibilities for the development of tools that will enhance the progress in the application of DNA tests for the traceability of animals and animal products.

The objectives of this study were to develop specific SNPs in five pig breeds by applying a next-generation sequencing strategy and to validate their utility for breed assignment and traceability purposes.

Materials and methods

Animals and DNA samples

Porcine DNA used for sequencing was obtained from four commercial breeds, namely Duroc (DU), Landrace (LR), Large White (LW) and Pietrain (PI), and the Wild Boar (WB). The numbers of pigs were 32, 27, 35, 22 and 37 for DU, LR, LW, PI and WB, respectively, making a total of 153 animals. DNA pools were made for each breed and contained equal amounts of DNA from each individual. The samples for the commercial breeds originated from the USA, Netherlands and Denmark and are representative of germplasm for pork production, while WB was sampled mainly in Europe, with five samples being collected in Japan.

In the second phase, additional individual pigs from the same breeds were selected for genotyping with the PorcineSNP60 beadchip (Ramos *et al.* 2009). Specifically, the numbers of samples were 57, 74, 110, 82 and 167 for DU, LR, LW, PI and WB, respectively, for a total of 490 samples. These additional samples originated from the same regions as the 153 samples that were sequenced.

Identification of breed-specific SNPs

Details regarding library construction, sequencing and filtering of the Genome Analyzer reads, as well as the procedures adopted for SNP detection and filtering, have been described previously (Ramos *et al.* 2009). For each SNP that passed the applied filtering criteria, all reads were analysed according to the breed information, using the unique identifier with which they had been labelled. A SNP was labelled as breed specific when the allele was only present in one of the five breeds and not detected in any of the other four.

Genotyping with the PorcineSNP60 beadchip and validation of the breed-specific SNPs

To perform validation of the putative breed-specific SNPs, the 153 DNA samples used for sequencing were all genotyped with the PorcineSNP60 beadchip. In addition, a total of 490 individuals from the same breeds used for sequencing were also genotyped with the beadchip, in order to evaluate the usefulness of the breed-specific SNPs for breed assignment purposes. Genotyping was performed at Illumina and ServiceXS following the manufacturer's recommendations. Genotypes that displayed a GenCall score lower than 0.7 were removed from the data set.

A total of 43 582 SNPs identified using the same data set had previously been included in the PorcineSNP60 beadchip, including 4441 putative breed-specific SNPs, which allowed validation of that subset of SNPs. For each SNP, breed specificity was evaluated by comparing the alleles detected with sequencing with the alleles present after genotyping the same DNA samples with the PorcineSNP60 beadchip.

Breed assignment tests

All assignment tests were carried out by defining the 153 individuals that were sequenced as the reference populations, to which the additional 490 animals from the same breeds were assigned.

The genotype data that were tested derived from three sets of SNPs. The first set included all the SNPs that had confirmed their breed specificity after checking the genotypes obtained with the PorcineSNP60 beadchip for the 153 sequenced individuals, for a total of 193 SNPs (hereafter referred to as ALLBSS).

A second set of 100 SNPs was selected from the set of putative breed-specific SNPs that did not confirm breed specificity when their genotypes were checked. This set of 100 SNPs was selected by taking 20 SNPs per breed that displayed the most extreme differences in allele frequency from the other four breeds (hereafter referred to as FREQ). These frequency differences were computed by initially calculating the difference in allele frequency between one breed and the other four, followed by determination of the sum of the four differences previously calculated. This second set of SNPs was selected to simulate a situation where a sequencing effort would be unable to generate any breed-specific SNPs, after checking the genotypes for those SNPs.

Finally, a third set was formed by randomly selecting a set of 100 SNPs from the 43 582 SNPs identified with the same data set that had been included in the Beadchip (hereafter referred to as RANDOM). This was done with the objective of simulating a situation where information on each SNP allele frequency is not considered when selecting a set of SNPs. The SNPs available for selecting the random set also included the set of 193 SNPs that confirmed breed

specificity, but none of these SNPs was included in the set of randomly chosen SNPs.

The assignment tests were performed using the methods implemented in two software packages, namely GeneClass2 (Piry *et al.* 2004) and Structure 2.3.1 (Pritchard *et al.* 2000). The assignment methods available in GeneClass2 included the frequency-based method of Paetkau *et al.* (1995) and the Bayesian-based methods of Rannala & Mountain (1997) (R&M) and Baudouin & Lebrun (2000) (B&L), while Structure implemented the Bayesian-based method developed by Pritchard *et al.* (2000). The runs performed with Structure 2.3.1 used an initial 20 000 burn-in period followed by 100 000 Markov chain iterations, following the recommendations of Falush *et al.* (2007). For the Structure runs, breed information was considered for the individuals that formed the reference populations, while K was set to 5, which was the number of breeds in the reference population set.

Evaluation of the performance of the breed assignment tests was carried out by analysing, for each breed, the number of animals incorrectly assigned, the specificity (calculated as the percentage of animals correctly assigned to their respective breed) and the average probability assignment score.

Finally, we collected several parameters for the 4441 SNPs available for validation, including quality parameters on the GA reads used in these SNPs, MAQ quality parameters, and information regarding the read depth for each of these SNPs. The Welch's unpaired *t*-test was used to investigate whether differences existed between validated and non-validated breed-specific SNPs for these parameters. The Welch *t*-test is a modification of the *t*-test for independent samples that does not assume that the variances for each population are equal, and whose degrees of freedom account for unequal sample sizes, unequal variances and small sample sizes.

Results

SNP discovery

The total number of SNPs discovered, as well as the number of SNPs that passed the stringency criteria and the number of putative breed-specific SNPs identified, are indicated in Table 1. After aligning the 247 million sequences to the reference genome and after application of the stringency criteria used to filter the initial SNP output, the number had been reduced to approximately 314K SNPs. This set of SNPs was then analysed for the presence of breed-specific SNPs.

Among the set of filtered SNPs, total of 29 146 SNPs were identified as putatively breed-specific (Table 1), indicating that one of the alleles was present in only one of the five breeds studied. DU-specific SNPs were the most abundant, while WB displayed the fewest number of breed-specific SNPs. The fact that an AluI short RRL was not created for WB contributed to the smaller number of WB-specific SNPs identified.

Validation of breed-specific SNPs

From the set of 29 146 putative breed-specific SNPs detected, a total of 4441 were included in the PorcineSNP60 beadchip, without prior knowledge about breed specificity. Hence, only approximately 15% of the putative breed-specific SNPs were available for validation. This number was further decreased by 467 SNPs because of failed assays, and 229 SNPs that displayed no variation when their genotypes were analysed (Table 2). A total of 3745 SNPs had suitable genotypes available to proceed with the validation process. For each of these SNPs, the allele variants detected with sequencing and with genotyping were compared, and breed specificity was determined.

A total of 3552 SNPs were not validated as breed specific, because at least one of the other four breeds shared the supposedly specific allele. The number of SNPs that failed to pass the breed specificity test because three or four breeds shared the allele assessed to be specific from sequence data alone was 2738 or 77% of the total number. This was a strong indication that the sequencing strategy adopted was unable to detect these variants, because they were present in those breeds. The remaining 23% of the SNPs, a total of 814, failed the specificity validation because of the fact that one or two of the other breeds carried the same allele. Table 2 includes the details regarding the SNPs that failed to pass the breed specificity validation.

Table 1 Summary of the SNP discovery process, including the number of putative breed-specific SNPs identified in each breed.

			Specific breed					
RRL	SNPs	Breed-specific	DU	LR	LW	PI	WB	SNPs on chip
AluI long	124 568	10 312	4294	633	2211	1910	1264	1811
AluI short	105 290	15 123	5547	4137	3391	2048	0[1]	2158
HaeIII	56 816	2447	483	240	733	512	479	359
MspI	27 290	1264	492	226	257	142	147	113
Total	313 964	29 146	10 816	5236	6592	4612	1890	4441

SNP, single nucleotide polymorphisms.

[1]The AluI short RRL was not sequenced in Wild Boar, which was the reason why no putative breed-specific SNPs were detected.

Table 2 Validation of the putative breed-specific SNPs.

SNPs included in the Beadchip	4441
Non-working SNP assays	467
Monomorphic SNPs	229
Number of failed breed-specific SNPs	
One breed	365
Two breeds	449
Three breeds	809
Four breeds	1929
Total	3552
Number of validated breed-specific SNPs	
Duroc	99
Landrace	16
Large White	24
Pietrain	19
Wild Boar	35
Total	193

SNP, single nucleotide polymorphisms.

A total of 193 SNPs were confirmed to be breed-specific after the genotypes for these SNPs were analysed (Table 2). The breed that presented the highest number of specific SNPs was DU, with 99 SNPs, while LW, PI and LR had 24, 19 and 16 specific SNPs, respectively. A total of 35 WB–specific SNPs were also detected. This set of 193 SNPs formed the first of the SNP sets used in the breed assignment tests. The average frequency at which the specific allele was found varied, ranging from 19% in LW to 47.8% in DU (Table 3), and frequencies for the specific allele reached values of 90% in DU and 79% in PI and WB. When the genotypes for the additional 490 individuals were checked, an additional 106 SNPs lost their breed specificity, while 87 SNPs still displayed one allele unique to only one breed (Table 3). WB retained 80% of the specific SNPs, while for LW, only 17% of the SNPs maintained breed specificity after the additional individuals were analysed. The list containing the names of the 193 breed-specific SNPs and their flanking sequence is contained in Table S1. All SNPs have been submitted to NCBI's pig dbSNP database (Submitter Handle: WU_ABGC; Submitter Batch ID: RRL_batch2).

Assignment tests

All assignment methods tested performed extremely well when the SNP sets that included all identified breed-specific SNPs (ALLBSS), and the SNPs with extreme allele frequency differences (FREQ) were used, while assignments were less robust when the set of SNPs chosen randomly (RANDOM) was employed. The results regarding all assignment methods and SNP sets used are summarized in Table 4.

When the ALLBSS SNP set was used, 486 of 490 animals were correctly assigned to their breeds of origin. The four individuals assigned to a wrong breed were LR (3) and LW (1) pigs. These results were identical for the four assignment methods used, and hence the specificity for all methods was 99%, while the average probability of assignment ranged from 99.2% to 99.9% for the Pritchard and R&M methods, respectively. Similar values were obtained for the FREQ SNP set. All methods showed a slightly higher correct assignment percentage when this SNP set was used, relative to the ALLBSS set. For the R&M, B&L and Pritchard methods, only one LR individual was incorrectly assigned, while for the Paetkau method, two LR pigs were assigned to the wrong breed. The specificity for this allele frequency-based SNP set exceeded 99.5% in all methods, while the average probability of assignment surpassed 99.8%. For these SNP sets, no significant differences between the different assignment methods were registered. The number of animals assigned to the wrong breed was higher when the RANDOM SNP set was used, varying from 17 with the B&L method to 47 when the Pritchard *et al.* (2000) method was used. Consequently, the specificity for this SNP set was smaller when compared with the two other SNP sets, varying from 90.2% to 96.4% with the Pritchard and B&L methods, respectively.

Table 3 Breed distribution of the validated breed-specific SNPs when the genotypes for the sequenced and additional individuals were analysed. The frequency values indicated are for the specific allele identified for each SNP.

	Sequenced individuals (n = 153)				Additional individuals (n = 490)				
	Number of specific SNPs	Minimum frequency	Maximum frequency	Average frequency	Number of specific SNPs	Minimum frequency	Maximum frequency	Average frequency	Retention percentage[1]
Duroc	99	0.03	0.9	0.48	42	0.02	0.94	0.4	0.42
Landrace	16	0.05	0.66	0.29	7	0.01	0.3	0.12	0.44
Large White	24	0.05	0.41	0.19	4	0.02	0.19	0.1	0.17
Pietrain	19	0.06	0.79	0.38	6	0.02	0.43	0.22	0.32
Wild Boar	35	0.01	0.79	0.27	28	0.01	0.8	0.21	0.8

SNP, single nucleotide polymorphism.

[1]The retention percentage refers to the number of SNPs that maintained breed specificity after the genotypes for the additional 490 individuals were analysed.

Table 4 Performance of the breed assignment methods tested for the three SNP sets analysed (ALLBSS, FREQ and RANDOM).

		Rannala & Mountain (1997)			Baudouin & Lebrun (2000)			Paetkau *et al.* (1995)			Pritchard *et al.* (2000)		
Breed	*n*	Assigned to wrong breed	Specificity	Av. Prob.[1]	Assigned to wrong breed	Specificity	Av. Prob.	Assigned to wrong breed	Specificity	Av. Prob.	Assigned to wrong breed	Specificity	Av. Prob.
SNP set 1: all breed-specific SNPs													
Duroc	57	0	1	1	0	1	1	0	1	1	0	1	1
Landrace	74	3	0.959	0.997	3	0.959	0.990	3	0.959	0.982	3	0.959	0.976
Large White	110	1	0.991	0.999	1	0.991	0.999	1	0.991	0.999	1	0.991	0.994
Pietrain	82	0	1	1	0	1	1	0	1	1	0	1	0.999
Wild Boar	167	0	1	0.999	0	1	0.997	0	1	0.999	0	1	0.992
Overall	490	4	0.990	0.999	4	0.990	0.997	4	0.990	0.996	4	0.990	0.992
SNP set 2: SNPs with extreme allele frequency differences													
Duroc	57	0	1	1	0	1	1	0	1	1	0	1	1
Landrace	74	1	0.986	0.996	1	0.986	0.994	2	0.973	0.999	1	0.986	0.991
Large White	110	0	1	1	0	1	0.999	0	1	1	0	1	0.999
Pietrain	82	0	1	1	0	1	1	0	1	1	0	1	0.999
Wild Boar	167	0	1	1	0	1	1	0	1	1	0	1	1
Overall	490	1	0.997	0.999	1	0.997	0.999	2	0.995	0.999	1	0.997	0.998
SNP set 3: randomly chosen SNPs													
Duroc	57	0	1	0.999	0	1.00	0.999	0	1	0.999	0	1	0.999
Landrace	74	12	0.838	0.969	9	0.878	0.957	15	0.797	0.982	27	0.635	0.887
Large White	110	4	0.964	0.999	4	0.964	0.999	2	0.982	0.996	0	1	0.998
Pietrain	82	0	1	0.999	0	1	0.999	0	1	0.999	1	0.988	0.987
Wild Boar	167	4	0.976	0.998	4	0.976	0.992	4	0.976	0.998	19	0.886	1
Overall	490	20	0.956	0.993	17	0.964	0.989	21	0.951	0.995	47	0.902	0.974

SNP, single nucleotide polymorphism.

Specificity was defined as the percentage of animals correctly assigned of the total number of animals assigned.

[1]Average probability of assignment.

Discussion

This study illustrates the usefulness of applying next-generation sequencing technologies in the development of high utility SNP markers, which are, in this context, breed-specific SNPs. The analysed data set was collected with the objective of maximizing the number of SNPs discovered across the five porcine breeds used. However, even though the sequencing strategy was not designed to specifically target this type of SNP, it was still able to identify in excess of 29K putative breed-specific SNPs. To maximize the discovery of these SNPs, the strategy should involve sequencing the RRLs at greater sequence depths to facilitate the identification of the less frequent breed-specific variants and to reduce the number of false positives. Nevertheless, the results obtained with the validated breed-specific SNPs showed that the sequencing strategy adopted in this work was still able to identify SNPs for which the frequency of the specific allele ranged from 1% to 90%. This was a clear indication that even the rare allelic variants of each breed were identified, which confirmed the robustness of using next-generation sequencing for this purpose. Therefore, it is likely that a dedicated sequencing strategy specifically designed to identify breed-specific markers will enhance the discovery rate while reducing the false positive rate of this type of SNP. In addition, the sequencing data used in this work were collected with Illumina's Genome Analyzer I, with reads that were sequenced to a length of 36 nucleotides. Owing to ongoing improvements in sequencing technologies, various alternatives now exist that can deliver even higher volumes of data at a much cheaper price compared to the first-generation Illumina GA that was used in this study. The increase in the availability of sequence data will facilitate SNP discovery and boost the identification of breed-specific variants.

From the total number of putative breed-specific SNPs identified, only approximately 15% were available for validation, because those were the ones that had been included in the PorcineSNP60 beadchip. The remaining number of putative breed-specific SNPs was not the subject of a validation effort, but because the SNPs on the PorcineSNP60 beadchip were optimized for having intermediate allele frequencies over several breeds, it is in fact likely that validation rates of all the putative breed-specific SNPs could be even higher. When the SNPs that failed to confirm breed specificity were analysed, it was clear that the main reason for failure was insufficient sequencing depth. This was demonstrated by the fact that for approximately 77% of the non-validated breed-specific SNPs, the supposedly specific allele was also present in

three or four of the breeds analysed, which is a clear indication that the sequencing strategy was not able to identify all the variants originally present in the breeds sequenced. In the future, this problem will be alleviated by sequencing at greater sequence depths.

Another important detail is the original sampling of the breeds and/or populations to be sequenced. From the 193 validated breed-specific SNPs, 106 lost their specificity when the additional 490 individuals from the same five breeds were analysed. This indicated that the original sampling of the breeds did not capture all of the variation that was present in each of them. Ideally, the sampling should be as representative as possible, which may prove difficult for breeds with a worldwide distribution, such as the porcine breeds used in this study. In any case, a total of 87 SNPs still maintained breed specificity, including an 80% retention rate in WB, indicating that WB is highly differentiated from domesticated *Sus scrofa*. Similarly, the high retention rate of breed-specific SNPs in DU may reflect the high degree of differentiation of this breed from all other European and North American pig breeds. Conversely, the low retention rate of breed-specific SNPs in PI, LW and LR may reflect the low degree of differentiation among these so-called white breeds (Megens *et al.* 2008). In any case, and regardless of the degree of differentiation of each breed, it will still be critical to adopt a sampling strategy that will capture a substantial proportion of the genetic variation of a particular breed. This task will be less demanding in breeds with smaller numbers of individuals, such as local or indigenous breeds, as these will be easier to sample.

The large number of putative breed-specific SNPs identified by sequencing will probably render it impossible to test all of them. Therefore, it would be important to have a way to select the SNPs with the highest probability of being truly specific. The results from the application of the Welch *t*-test to a set of parameters collected in the two SNP groups where breed specificity had been confirmed or discarded revealed significant differences for three parameters, namely total read count for the SNP ($P < 0.0001$), read count for the major allele ($P < 0.0001$) and read count for the minor allele ($P < 0.05$). These were parameters for which the means were higher in the group of SNPs that were confirmed to be breed-specific. These results clearly indicate that from the group of putative breed-specific SNPs identified from the sequence data, SNPs with higher read counts should be prioritized for subsequent validation. This will be important to allow a better selection of SNPs, increase validation rates and decrease the costs associated with the validation process.

The results from the breed assignment tests when the SNP set ALLBSS was used were excellent, because 99% of the individuals (486 of 490) were correctly assigned to their breeds of origin with a very high probability of assignment (Table 4). A slight improvement in the results, namely in the specificity, was observed with the FREQ SNP set for all assignment methods used, while the values for the average probability of assignment were comparable. For the FREQ SNP set, only two LR individuals were incorrectly assigned when the method of Paetkau *et al.* (1995) was used, while for all other methods, only one LR individual was assigned to the wrong breed (Table 4). The FREQ SNP set was selected from the group of putative breed-specific SNPs that failed to validate breed specificity. This SNP set was used to simulate the possibility that a sequencing effort would fail to identify any validated breed-specific SNPs. It should be emphasized that very stringent criteria for specificity were used in the validation procedure, because any putative breed-specific SNP was not validated if a single count of the specific allele was observed in a different breed. However, these non-validated breed-specific SNPs still displayed substantial differences in the allele frequencies for each of the breeds, which were sufficient to allow an outstanding power for breed assignment purposes. The results obtained for the ALLBSS and FREQ SNP sets clearly indicated that prior information about the variants present in each breed, obtained with next-generation sequencing, will identify SNP markers that will be extremely useful for breed assignment purposes. In contrast, the RANDOM SNP set was chosen without any prior information on the allelic variants. The breed assignment results obtained with this set were clearly worse when compared with the previous sets. The number of animals assigned to the wrong breed varied from 17, for the B&L method implemented in GeneClass2, to 47 for the method of Pritchard *et al.* (2000) implemented in Structure 2.3.1., and a decrease in the average probability of assignment was observed for all methods. These differences illustrated the benefit that will be gained by having prior information about the breed distribution of SNP allelic variants. All assignment tests were performed using reference populations, which were the individuals that had originally been sequenced, in order to simulate a practical, 'real-world' application of this type of technology. The slightly better performance of the FREQ SNP set, when compared with the ALLBSS set, could possibly be because of the impact of including SNPs with very low minor allele frequency in the ALLBSS set. Nevertheless, the outstanding results obtained with the ALLBSS and FREQ SNP sets indicate that this approach can potentially be translated into applications used in the traceability of animals and their products to their breeds of origin.

Previous studies performed in different cattle breeds have explored the use of either AFLP (Negrini *et al.* 2007) or microsatellite (Dalvit *et al.* 2008a,b) markers for breed assignment purposes, with results that were less convincing that the ones presented here. In addition, the use of these types of markers will likely decrease in the future, mainly because of the higher costs associated with using AFLPs and microsatellites compared to SNPs. Other studies used SNP

markers for breed assignment purposes in cattle, with SNP sets selected in candidate genes for meat quality and production traits (Negrini *et al.* 2008a,b), but without any prior information about the allele frequencies for those markers in the breeds analysed. Overall, the results from those studies did not achieve the levels of specificity and probability assignment that were observed in this work, even though comparable results were obtained for some of the cattle breeds analysed in the studies performed by Negrini *et al.* (2008a,b).

In principle, the approach proposed in this study can be applied to any populations that differ at an unknown number of locations in their genome, including populations with varying degrees of differentiation between them. Some of these differences will be specific to one of the populations and can be used for multiple applications. The most variable factor will be the amount of sequencing needed to identify the target number of breed-specific markers. However, as mentioned previously, the technology has been improving and the costs decreasing, which will make the amount of sequencing needed of smaller concern. In the European Union, there are many animal products to which the Product of Designated Origin (PDO) or Protected Geographic Indication (PGI) labels have been awarded. The production of many of these products is based on the use of a single breed, which is usually also a pre-requisite for the PDO/PGI label to be awarded. The many examples of this situation span different species and countries and include fresh meat, processed meat products, cheeses and other products. Customers generally associate the PDO/PGI products with higher quality, and these products are normally sold at higher prices, making them vulnerable to fraud. Hence, effective and reliable control mechanisms and verification systems are needed for food safety control, to assure consumer confidence and to prevent frauds. The approach proposed in the present study will be suitable to achieve all these goals, at least for the products where a single breed is used, by using state of the art sequencing technologies for the development of high utility SNP markers.

Conclusions

This study provides a blueprint for the use of next-generation sequencing technologies in the identification of breed-specific SNP markers, by offering possible strategies for how to sample and sequence the breeds/populations of interest, select sets of putative breed-specific SNPs displaying higher validation rates, and offering strong evidence of the utility of using breed-specific SNPs for assigning individuals to their breeds of origin. In fact, the number of correct breed allocations surpassed 99% for all assignment methods tested, indicating that this approach will be a powerful tool in the molecular traceability of animal products to their breeds of origin.

Acknowledgements

Financial support for Antonio Marcos Ramos was provided by the Marie Curie Intra-European Fellowship PIEF-GA-2008-220390. The authors thank Mr. Zhao Zhen Sun for his assistance with some of the bioinformatics procedures.

References

Baudouin L. & Lebrun P. (2000) An operational Bayesian approach for the identification of sexually reproduced cross-fertilized populations using molecular markers. *Acta Horticulturae* **546,** 81–93.

Bennett S. (2004) Solexa Ltd. *Pharmacogenomics* **5,** 433–8.

van Bers N.E., van Oers K., Kerstens H.H., Dibbits B.W., Crooijmans R.P., Visser M.E. & Groenen M.A. (2010) Genome-wide SNP detection in the great tit *Parus major* using high throughput sequencing. *Molecular Ecology* **19,** 89–99.

Dalvit C., De Marchi M., Targhetta C., Gervaso M. & Cassandro M. (2008a) Genetic traceability of meat using microsatellite markers. *Food Research International* **41,** 301–7.

Dalvit C., De Marchi M., Dal Zotto R., Gervaso M., Meuwissen T. & Cassandro M. (2008b) Breed assignment test in four Italian beef cattle breeds. *Meat Science* **80,** 389–95.

Eid J., Fehr A., Gray J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–8.

Falush D., Stephens M.W. & Pritchard J.K. (2007) Inference of population structure using multilocus genotype data, dominant markers and null alleles. *Molecular Ecology Notes* **7,** 574–8.

Kerstens H.H., Crooijmans R.P., Veenendaal A., Dibbits B.W., Chin-A-Woeng T.F., den Dunnen J.T. & Groenen M.A. (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics* **10,** 479.

Margulies M., Egholm M., Altman W.E. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–80.

Matukumalli L.K., Lawley C.T., Schnabel R.D. *et al.* (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* **4,** e5350.

Megens H.J., Crooijmans R.P.M.A., San Cristobal M., Hui X., Li N. & Groenen M.A.M. (2008) Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. *Genetics, Selection, Evolution* **40,** 103–28.

Milos P. (2008) Helicos bioSciences. *Pharmacogenomics* **9,** 477–80.

Negrini R., Milanesi E., Colli L., Pellecchia M., Nicoloso L., Crepaldi P., Lenstra J.A. & Ajmone-Marsan P. (2007) Breed assignment of Italian cattle using biallelic AFLP markers. *Animal Genetics* **38,** 147–53.

Negrini R., Nicoloso L., Crepaldi P. *et al.* (2008a) Assessing SNP markers for assigning individuals to cattle populations. *Animal Genetics* **40,** 18–26.

Negrini R., Nicoloso L., Crepaldi P. *et al.* (2008b) Traceability of four European Protected Geographic Indication (PGI) beef products using Single Nucleotide Polymorphisms (SNP) and Bayesian statistics. *Meat Science* **80,** 1212–7.

Paetkau D., Calvert W., Stirling I. & Strobeck C. (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* **4,** 347–54.

Piry S., Alapetite A., Cornuet J.M., Paetkau D., Baudouin L. & Estoup A. (2004) GeneClass2: a software for genetic assignment and first-generation migrant detection. *Journal of Heredity* **95,** 536–9.

Pritchard J.K., Stephens M. & Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–59.

Ramos A.M., Crooijmans R.P., Affara N.A. *et al.* (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS ONE* **4,** e6524.

Rannala B. & Mountain J. (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences* **94,** 9197–201.

Sánchez C.C., Smith T.P., Wiedmann R.T., Vallejo R.L., Salem M., Yao J. & Rexroad C.E. (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics* **10,** 559.

Shendure J., Porreca G.J., Reppas N.B., Lin X., McCutcheon J.P., Rosenbaum A.M., Wang M.D., Zhang K., Mitra R.D. & Church G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309,** 1728–32.

Wiedmann R.T., Smith T.P. & Nonneman D.J. (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics* **9,** 81.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Information regarding the SNP name, base substitution and flanking sequence for the 193 validated breed specific SNPs.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.