The use of next generation sequencing technology in the identification of specific SNPs for breed assignment and traceability of animal products

A.M. Ramos[*], H.J. Megens[*], R.P.M.A. Crooijmans[*], L.B. Schook[§], M.A.M. Groenen[*]

Introduction

The advent of next-generation sequencing will likely revolutionize the field of animal genetics and genomics. Today, technologies such as Illumina's Genome Analyzer (Bennett 2004), Roche's 454 (Margulies et al. 2005), Helicos (Milos et al. 2008) and Pacific Biosciences' real-time sequencing (Eid et al. 2009) allow the generation of large volumes of sequencing data in a fast, accurate and inexpensive way.

One of the fields where these technologies have been widely applied is SNP discovery. Studies performed in several species of domestic animals led to the identification of thousands of SNPs and to the development of high density SNP genotyping beadchips.

Several DNA-based methods have also been investigated for their potential use in the identification of animals at different levels, from individuals, to breeds and species. 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 sequenced with Illumina's Genome Analyzer and investigate their utility for breed assignment purposes.

Material and methods

Animals and DNA samples. Porcine DNA was collected in five pig breeds, namely Duroc (DU), Landrace (LR), Large White (LW), Pietrain (PI) and Wild Boar (WB). For all breeds, the samples were as representative as possible of their worldwide distribution. A total of 153 animals were used, including 32, 27, 35, 22 and 37 samples collected in DU, LR, LW, PI and WB, respectively. DNA pools were formed for each breed and contained equal amount of DNA from all the individuals sampled. Besides the animals used for sequencing additional individuals were gathered and genotyped with the PorcineSNP60 beadchip (Ramos et al. 2009). The number of additional samples was 57, 74, 110, 82 and 167 for DU, LR, LW, PI and WB, respectively, for a total of 490 samples.

Sequencing, SNP detection and filtering. The DNA pools were digested with three restriction enzymes (*Alu*I, *Hae*III and *Msp*I) and sequenced to a length of 36 nucleotides on a 1G Genome Analyzer (GA). The criteria used to filter the GA reads included the presence or

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

absence of each restriction enzyme motif, the presence of poly-A,C,G,T regions, the average quality score for each read (minimum threshold set at 12) and the number of times each read was detected (all reads present in excess of five times the sequence depth were removed). Each of the reads that passed these criteria was subsequently labeled with a unique identifier that included breed information.

The reads were aligned with MAQ (Li et al. 2008) to Build 7 of the pig genome. SNPs were filtered for several quality criteria that included MAQ mapping quality parameters, the number of reads for the minor allele and the total number of reads. SNPs were labeled as breed specific when an allele was present in one of the five breeds, but not in the other four.

Validation of breed specific SNPs. The 153 samples used for sequencing, as well as the additional 490 samples, were genotyped with the PorcineSNP60 beadchip. A total of 4,441 SNPs identified as breed specific at the bioinformatics level were included in the Beadchip, which allowed validation of this subset of SNPs by comparing the alleles detected with sequencing with the alleles present after genotyping the same DNA samples.

Breed assignment tests. The assignment tests were performed using the 153 sequenced individuals as reference populations to which the additional 490 animals were assigned. The genotypes used were derived from the set of SNPs that had confirmed breed specificity after checking the genotypes obtained with the PorcineSNP60 beadchip. The assignment tests were conducted using the methods implemented in the software packages GENECLASS2 (Piry et al., 2004), which included the frequency based method of Paetkau and colleagues (1995) and the Bayesian based methods of Rannala & Mountain (1997) and Baudouin & Lebrun (2001), and also the Bayesian method implemented in Structure 2.3.1 (Pritchard et al. 2000). The performance of each assignment test was evaluated by analyzing the number of animals assigned to the wrong breed, the specificity (defined as the number of animals correctly assigned) and the average probability assignment score.

Results and discussion

SNP discovery. The total number of unfiltered SNPs was 9,048,038, but most were not real SNPs because MAQ reports as a SNP all variation detected in the GA reads and between them and the reference genome. After filtering the raw output a total of 313,964 SNPs remained in the dataset, which was then analyzed for the presence of breed specific SNPs. At the bioinformatics level, a total of 29, 146 SNPs were identified as breed specific.

Validation of breed specific SNPs. From the set of breed specific SNPs detected at the bioinformatics level, a total of 4,441 had been included in the PorcineSNP60 beadchip, even though breed specificity was not considered as a criteria in the selection of which SNPs to include on the beadchip. Hence, only approximately 15% of the breed specific SNPs were available for validation. This number was further decreased because 467 SNPs had nonworking assays and 229 SNPs displayed no variation when their genotypes were analyzed. In the end, a total of 3,745 SNPs were available to proceed with the validation process. This information is summarized in Table 1.

A total of 3,552 SNPs did not confirm breed specificity because at least one of the other four breeds had displayed the supposedly specific allele. We further investigated the number of breeds that were causing each SNP to lose its specificity (Table 1). The number of SNPs that failed to pass the breed specificity test because three or four breeds also contained the allele thought to be specific was 2,738, which accounted for approximately 77% of the total

number. This indicated that the sequencing strategy adopted was unable to detect these variants, since they were present in those breeds. This was not surprising because the strategy had originally been designed to identify SNPs with high minor allele frequency across breeds. In the future, studies targeting the identification of breed specific SNPs should prioritize sequencing at greater read depths, which will facilitate the discovery of the rarer breed specific variants.

Table 1: Validation of breed specific SNPs

SNPs included in the Beadchip		4,441
Nonworking SNP assays		467
Monomorphic SNPs		229
Number of failed breed specific SNPs	**One breed**	365
	Two breeds	449
	Three breeds	809
	Four breeds	1,929
	Total	3,552
Number of validated breed specific SNPs	**Duroc**	99
	Landrace	16
	Large White	24
	Pietrain	19
	Wild Boar	35
	Total	193

Despite the limitations of the sequencing strategy, a total of 193 SNPs were confirmed to be breed specific after their genotypes were analyzed. The breed which presented the highest number of specific SNPs was Duroc, with 99 SNPs, followed by Wild Boar, Large White, Pietrain and Landrace (Table 1). The average frequency at which the specific allele was found was highest in Duroc (0.478) followed by Pietrain, Landrace, Wild Boar and Large White, that displayed frequencies of 0.381, 0.286, 0.27 and 0.19, respectively. Even though the validation rate of the breed specific SNPs was low, our study showed that next generation sequencing technologies allow unprecedented power to detect the unique features that define the genetic architecture of porcine breeds.

Assignment tests. The results of the assignment tests performed using the set of validated breed specific SNPs are indicated in Table 2. All the methods tested for assigning the additional 490 individuals to their breeds of origin performed extremely well. A total of 486 animals were correctly assigned to their breeds of origin. The four individuals assigned to a wrong breed derived from the Landrace (3) and Large White (1) breeds. These results were identical for the four assignment methods used, hence the specificity for all methods was 99.2%. The values for the average probability of assignment were extremely high and ranged from 99.2% to 99.9% for the Pritchard et al. (2000) and Rannala & Mountain (1997) methods, respectively. These results clearly indicated the usefulness of using the set of breed specific SNPs in the assignment of individuals to their original breeds. The application of this type of markers will allow the molecular traceability of several animal products awarded with the PGI/PDO labels that require the use of a single breed in their production.

Conclusion

This study provides a blueprint on how next generation sequencing technologies can be utilized in the identification of breed specific SNPs. The results presented clearly indicate that the set of breed specific SNPs displayed very high power for breed assignment. The number of correct allocations surpassed 99% for all assignment methods tested and, thus, may be a powerful tool in the traceability of animal products to their breeds of origin.

Table 2: Performance of the breed assignment methods tested

Method		Duroc	Landrace	Large white	Pietrain	Wild Boar	Overall
Rannala & Mountain	**Incorrect**[§]	0	3	1	0	0	4
	Specificity	1	0.959	0.991	1	1	0.992
	Av. Prob.	1	0.997	0.999	1	0.999	0.999
Baudouin & Lebrun	**Incorrect**	0	3	1	0	0	4
	Specificity	1	0.959	0.991	1	1	0.992
	Av. Prob.	1	0.99	0.999	1	0.997	0.997
Paetkau et al.	**Incorrect**	0	3	1	0	0	4
	Specificity	1	0.959	0.991	1	1	0.992
	Av. Prob.	1	0.982	0.999	1	0.999	0.996
Pritchard et al.	**Incorrect**	0	3	1	0	0	4
	Specificity	1	0.959	0.991	1	1	0.992
	Av. Prob.	1	0.976	0.994	0.999	0.992	0.992

§ Incorrect refers to the number of individuals assigned to the wrong breed

References

Paetkau, D., Calvert, W., Stirling, I. *et al.* (1995). *Mol. Ecol.* 4(3): 347-354.

Rannala, B., and Mountain, J.L. (1997). *Proc. Natl. Acad. Sci. USA* 94(17): 9197-9201.

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). *Genetics* 155(2): 945-959.

Baudouin, L., and Lebrun, P. (2001). *Proc. Int. Symp. Molecular Markers* 546: 81-94.

Bennett, S. (2004). *Pharmacogenomics* 5(4): 433-438.

Piry, S., Alapetite, A., Cornuet, J.M. *et al.* (2004). *J. Hered*. 95(6): 536-539.

Margulies, M., Egholm, M., Altman, W.E. *et al.* (2005). *Nature* 437(7057): 376-380.

Li, H., Ruan, J., and Durbin, R. (2008). *Genome Res.* 18(11): 1851-1858.

Milos, P. (2008). *Pharmacogenomics* 9(4): 477-480.

Eid, J., Fehr, A., Gray, J. *et al.* (2009). *Science* 323(5910): 133-138.

Ramos, A.M., Crooijmans, R.P., Affara, N.A. *et al.* (2009). *PLoS One* 4(8): e6524.