Imputation accuracy of bovine spongiform encephalopathy-associated *PRNP* indel polymorphisms from middle-density SNPs arrays

A. Gurgul[1], K. Sieńko[2], K. Żukowski[3], K. Pawlina[1], M. Bugno-Poniewierska[1]

[1]Laboratory of Genomics, National Research Institute of Animal Production, Balice, Poland
[2]Department of Genetics, Wrocław University of Environmental and Life Sciences, Wrocław, Poland
[3]Department of Animal Breeding and Genetics, National Research Institute of Animal Production, Balice, Poland

ABSTRACT: Statistical methods of imputation allow predicting genotypes of markers (which were not genotyped in the whole population) based on known linkage disequilibrium relationships between the flanking polymorphisms and the information obtained from reference datasets used as a pattern. In this study we attempted to predict genotypes of two bovine spongiform encephalopathy (BSE) susceptibility associated indel polymorphisms located in the promoter region of *PRNP* gene relying on the data obtained from middle density SNPs arrays in a sample of the population of Holstein cattle. The two SNPs panels spanning *PRNP* locus were tested in terms of imputation efficiency. Both panels gave satisfactory imputation results showing high accuracy and high probabilities of imputed genotypes. Our results suggest that the approach applied can be used to evaluate the frequency of the disease associated polymorphisms in large populations of animals genotyped with whole-genome SNPs panels based on a limited-size reference population and small financial outlays.

Keywords: BSE; cattle; indel; prion gene

INTRODUCTION

Bovine spongiform encephalopathy (BSE) is one of the most intensively studied prion diseases in animals, mainly because of its infectious nature and capacity for interspecies transmission to humans (Hill et al., 1997; Scott et al., 1999). So far, several prion diseases have been described in humans and other species (Prusiner, 1998). Most of them have their genetic component mainly connected with the polymorphism of prion protein gene (*PRNP*), especially in the region of open reading frame (ORF) encoding prion protein precursor (Shibuya et al., 1998; Parchi et al., 1999; Hilton et al., 2004; Zarranz et al., 2005). Despite many studies, the polymorphism of bovine *PRNP* ORF was not associated with cattle susceptibility to BSE (Hunter et al., 1994; Neibergs et al., 1994). Nevertheless, some BSE-predisposing effect was observed for two insertion/deletion (indel) type polymorphisms located in the region of promoter and intron 1 of *PRNP* (23-bp indel and 12-bp indel, respectively) (Sander et al., 2004). The functional relevance of the polymorphisms is connected with disruption of regulatory element binding sites and alteration of expression level of prion protein gene (Sander et al., 2005).

In recent years, the frequency of BSE-associated deletion alleles of the two indels has been established in a number of cattle populations worldwide to determine their potential genetic susceptibility to BSE (Jeong et al., 2005; Juling et al., 2006; Nakamitsu et al., 2006; Czarnik et al., 2007, 2009; Ün et al., 2008). Taking into consideration the fact that prion gene locus polymorphisms were associated with the level of some production traits (Isler et al., 2006; Strychalski et al., 2011), the indel loci may be added to the selection pressure and their frequency may change over the next generations. The constant monitoring of the allele frequency of these two polymorphisms in the world cattle

population would be reasonable, because it could facilitate the breeding of animals with naturally increased resistance to BSE and could be beneficial for widely comprehended animal health and production of healthy and safe food.

Together with the development of high throughput genotyping methods, especially whole-genome genotyping microarrays, the idea of genomic selection is being introduced to the daily animal breeding (Hayes et al., 2009). The genomic selection requires large-size reference populations, which have been genotyped for whole-genome SNP panels. Genomic SNP panels available on microarrays are designed to capture the largest possible part of genomic variation and can be used for many diverse applications, including imputation of genotypes of markers, which were not genotyped in the whole population using molecular methods (Druet et al., 2010). Thus, in this study we attempted to predict the genotypes of two BSE-associated indel polymorphisms in the locus of *PRNP* gene from middle-density SNPs arrays (Bovine SNP50 BeadChip) and to evaluate the accuracy of such application. The reliable results of imputing indels from SNP data could be used for almost costless monitoring of the genetic susceptibility of Holstein cattle to BSE.

MATERIAL AND METHODS

Samples and DNA preparation. Three hundred and sixteen samples of randomly selected Holstein cattle (both males and females) were studied. DNA was isolated from full blood or semen gathered in the DNA Bank held by the National Research Institute of Animal Production. DNA was evaluated in terms of quantity and quality and normalized to the required concentration of 50 ng/µl.

SNPs and indel genotyping. All studied animals were genotyped for 54 609 SNP loci included in the Bovine SNP50 BeadChip assay, with the use of standard Infinium ultra protocol provided by the manufacturer (Illumina Inc., San Diego, USA). Raw intensities obtained after scanning of the microarrays with HiScanSQ were analyzed and clustered in GenomeStudio software and assessed for quality by the analysis of call rates and gene call scores. The 23-bp and 12-bp indel polymorphisms were genotyped using PCR method with primers described earlier (Gurgul et al., 2012a, b). Standard amplification procedure implemented in Fast Cycling PCR kit (Qiagen Inc., Valencia, USA) was used.

It was previously shown that the 23-bp indel polymorphism is located in the region upstream of exon 1 (~ –1616 bp according to the transcription start site) and the 12-bp indel is located in intron 1 (~ +300 bp) of *PRNP* (Sander et al., 2005). The genomic positions of indels were established relatively to the *PRNP* transcription start site specified in the Ensembl UMD3.1 genome annotation file (http://www.ensembl.org/Bos_taurus/Info/Index).

Selection of markers for the imputation process. The SNPs were selected on the basis of their genomic coordinates, in this case, location inside or near the *PRNP* gene locus on chromosome 13. The SNPs' genomic positions were retrieved from the map of markers provided by Illumina Corp. showing genomic coordinates according to UMD3.1 cattle genome assembly. Only SNPs that passed frequency test (minor allele frequency > 0.1), quality test (> 90% of calls) and were in Hardy-Weinberg equilibrium ($P > 0.05$) in the studied group were considered. The chromosomal region encompassing *PRNP* was scanned for the presence of regions with limited haplotype diversity (haplotype blocks) with the use of Haploview (Version 4.2) (Barrett et al., 2005). The r^2 correlation coefficient as a measure of pairwise linkage disequilibrium (LD) was calculated with the use of PLINK whole genome analysis toolset (Purcell et al., 2007). Based on the results obtained from haplotype block identification, LD analysis or SNP distribution near the *PRNP* locus, two overlapping 6 SNPs-panels were selected for the imputation (Table 1).

Imputation. Two datasets created based on the data obtained from all studied animals were used for the imputation. The first dataset (257 animals) was composed of individuals with known SNPs and indel genotypes, which were used as a reference population. The second dataset consisted of 59 animals with known SNP genotypes, but their indel genotypes were masked and used for validation of imputation accuracy. The accuracy of imputation was defined as the percentage of correctly imputed haplotypes and genotypes in the validation dataset.

Phasing of genotypes and imputation in validation dataset were performed with BEAGLE software (Version 3.3.2) (Browning and Browning, 2009), which employs hidden Markov model (HMM) and performs local clustering of haplotypes at each marker position to define hidden states. The posterior probability of each imputed genotype generated by BEAGLE was also evaluated.

Table 1. Basic genetic parameters for the SNPs on chromosome 13 included in the two imputation panels

No.	Marker name	Position (bp)	Panel	MAF	O(HET)	HWE	LD between markers (r^2)									
							1	2	3	4	5	6	7	8	9	10
1	ARS-BFGL-NGS-107916	47167515	1	0.08	0.14	0.48	–									
2	ARS-BFGL-NGS-103320	47211871	1	0.48	0.47	0.34	0.11	–								
3	ARS-BFGL-NGS-17702	47256972	1	0.34	0.48	0.20	0.02	0.51	–							
4	BTB-00526711	47289669	1, 2	0.49	0.49	0.75	0.12	0.91	0.48	–						
5	Hapmap39323-BTA-32823	47368346	1, 2	0.47	0.48	0.46	0.10	0.92	0.46	0.93	–					
6	ARS-USMARC-Parent-AY853302-no-rs	47397987	1, 2	0.49	0.49	0.83	0.12	0.90	0.48	0.99	0.92	–				
7	23-bp indel	47398813	1, 2	0.43	0.52	0.25	0.17	0.74	0.62	0.75	0.69	0.75	–			
8	12-bp indel	47400713	1, 2	0.49	0.50	0.91	0.12	0.94	0.51	0.94	0.88	0.94	0.78	–		
9	BTB-01997512	47415727	2	0.19	0.31	0.73	0.41	0.18	0.08	0.19	0.24	0.20	0.03	0.16	–	
10	BTB-00526221	47450332	2	0.26	0.42	0.13	0.29	0.29	0.04	0.27	0.32	0.27	0.29	0.26	0.26	–
11	ARS-BFGL-NGS-80072	47485051	2	0.28	0.40	0.79	0.24	0.37	0.03	0.38	0.36	0.38	0.21	0.38	0.34	0.64

LD = linkage disequilibrium, MAF = minor allele frequency, O(HET) = observed heterozygosity, HWE = *P*-value for Hardy-Weinberg equilibrium. Positions are given according to UMD3.1 cattle genome assembly

RESULTS

Indel genotyping results. The analysis of the distribution of indel genotypes showed that both polymorphisms were in Hardy-Weinberg equilibrium at the significance level of 0.05. The allele and genotype frequencies of the indels were shown in Table 2. At both polymorphic sites the major allele was BSE susceptibility-associated deletion (del) variant. The two indel genotypes were phased into three common haplotypes (frequency > 0.01), of which the most common was a haplotype including both deletions (51.2%). The polymorphisms were in strong LD expressed by r^2 correlation coefficient of 0.78.

Selection of marker panels for the imputation procedure. The scan of the genomic region spanning *PRNP* gene revealed the presence of a consistent haplotype block, encompassing promoter region of *PRNP* (Figure 1). Therefore, the first imputation panel was formed by the haplotype block spanning both the indel polymorphisms and six upstream SNPs, which had passed quality tests. The size of the haplotype block was estimated to be 233 kb and the mean r^2 between the markers inside the block was estimated to be 0.76 (±0.19). Only five common (frequency > 0.01) haplotypes were identified within the block along with 13 low-frequency haplotypes. The MAF (minor allele frequency) for the selected SNPs ranged 0.081–0.492 with the mean value of 0.411.

The second marker set was selected on the basis of genomic coordinates and encompassed six informative SNPs – three in upstream region (covered by the haplotype block) and three in 3' downstream region in respect to the positions of the indels. The region spanned 195 kb of a genomic sequence including 109 kb of upstream bases and about 85 kb of a downstream sequence in relation to the indels. The r^2 coefficient between markers included in the region ranged 0.21–0.99 with the mean value of 0.55 (±0.29). The marker genotypes were phased in 34 haplotypes of which 7 occurred with frequency higher than 1%. The mean MAF for the second marker panel was slightly lower than for SNPs in the haplotype block and ranged 0.188–0.492 with the mean value of 0.388. The basic genetic parameters, along with SNP names and positions for both panels, are shown in Table 1.

Imputation accuracy. The imputation accuracy was measured by comparing the imputed haplotypes in the validation dataset of 59 animals with haplotypes obtained from phasing the same individuals with "visible" indel genotypes. Also,

Table 2. Frequency of alleles, genotypes, and haplotypes of the *PRNP* indel polymorphisms in the studied group of animals

Variant	*n*	Allele frequency		Genotype frequency		
		ins	del	ins/ins	ins/del	del/del
23-bp Indel	316	0.429	0.571	0.168	0.522	0.310
12-bp Indel		0.487	0.513	0.237	0.500	0.263
		frequency of haplotypes				
Haplotypes	316	23del–12del		23ins–12ins		23del–12ins
		0.5127		0.429		0.058

ins = insertion, del = deletion

the imputed genotypes at indel loci were directly compared with results of PCR genotyping in the validation dataset. For both approaches and both SNP panels we found full consistency of the results, involving 100% of correctly assigned haplotypes and 100% of correctly imputed indel genotypes. The posterior probability of imputed genotypes ranged 0.939–1 with mean value of 0.990 (±0.014) for the first SNP panel and 0.862–1 with a mean value of 0.971 (±0.036) for the second panel.

DISCUSSION

The *PRNP* indel polymorphisms, in the simplest manner, can be genotyped by a PCR method and agarose gel electrophoresis, which is laborious, especially when large numbers of animals have to be analyzed. Nevertheless, the number of genotyped animals needed for reliable determination of the allele frequency in a population can be minimized by application of statistical methods of imputation, which allow for prediction of unobserved genotypes based on flanking markers, whose genotypes are already known. Together with the growing datasets comprising multilocus genotypes of genome-wide SNPs used in genomic selection of Holstein cattle, the statistical methods of imputation of missing observations are being developed. These methods allow for reconstruction of multilocus haplotypes within the regions of interest and prediction of untyped markers genotypes based on known LD relationships between flanking polymorphisms. The

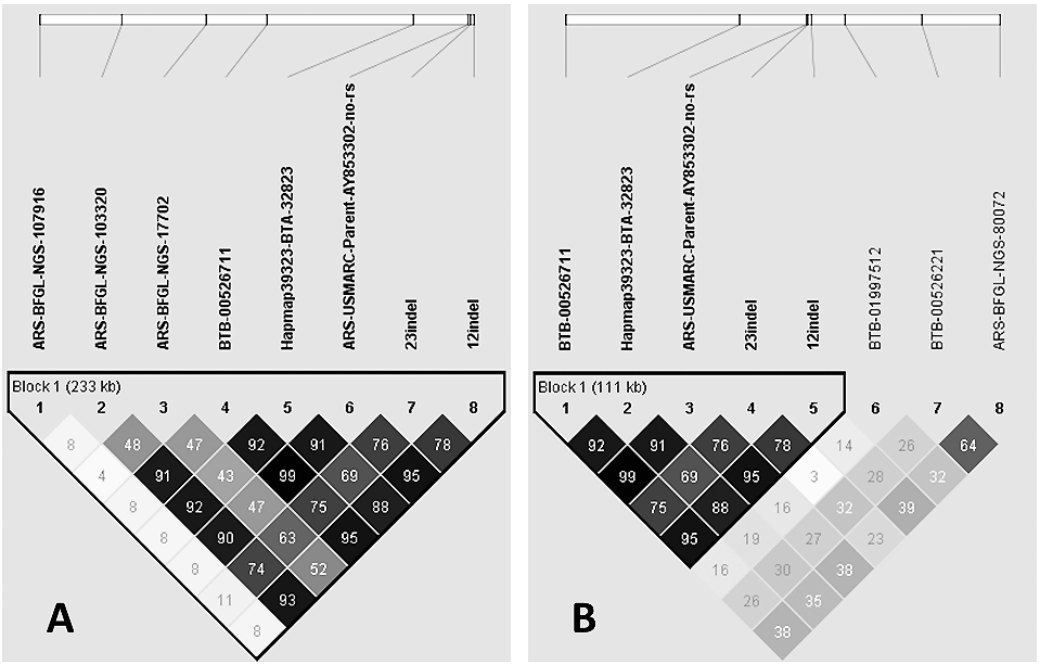

Figure 1. Heat map of linkage disequilibrium for SNPs and indels included in both imputation panels

(**A**) imputation panel formed by a haplotype block (block 1)

(**B**) imputation panel encompassing 6 SNPs surrounding the studied indels

The r^2 correlation coefficient values are given in squares for each polymorphism pair

larger the reference population and the higher LD in the region of missing SNPs, the better the method performance and the more accurate are the results of imputation (Druet et al., 2010; Mulder et al., 2012).

In this study we attempted to use imputation to predict genotypes of two BSE-associated indel polymorphisms in a small sample of population of Polish Holstein cattle and to validate the results obtained by direct PCR genotyping. The two SNP subsets were selected for the imputation on the basis of LD and haplotype structure within a region or simply genomic coordinates encompassing the positions of the indels. Both panels yielded satisfactory imputation results, but the panel formed by the haplotype block near the *PRNP* locus is presumed to be more efficient because of higher LD between the included SNPs and lower haplotype diversity within the block. Also, the recombination rate is significantly lower within the haplotype blocks than between other randomly selected markers, because the haplotype blocks are actually inherited as a single genetic unit (Zhang and Niu, 2010). On the other hand, considerable LD in this region may question the need for imputation as the presence of certain SNPs combination should be a good indicator of the indel genotypes. However, most of SNPs databases used in genomic selection process employ imputation as standard procedure to complete the data, so there is no contraindication to apply this approach also for indel genotypes. In our validation dataset we obtained 100% accuracy of genotype imputation and high genotype probabilities, which suggests that the method would be sufficiently accurate to estimate the frequency of indels in a large population genotyped by SNP arrays. Because the BEAGLE software used is one of the best tools for imputation of markers being in strong LD and is computationally efficient, no other imputation approaches were tested in this study. The size of the reference population used here was sufficient to achieve high accuracy of imputation in the validation dataset and it is presumed to be large enough to impute markers in larger datasets, considering that in the previous studies, reference datasets as small as 100 individuals were sufficient for accurate imputation of multiple SNPs located throughout the genome in hundreds of animals (Pausch et al., 2013). Also, the allele and genotype frequencies of the two studied indels were highly similar to those previously published for the Polish Holstein-Friesian cattle population, which is why no significant result bias caused by reference population structure can be assumed (Czarnik et al., 2011; Gurgul et al., 2012a,b).

CONCLUSION

Summarizing, our study shows that BSE-associated *PRNP* indel polymorphisms can be easily monitored in the population of animals genotyped with whole-genome SNP panels by using standard imputation procedures and only a limited number of reference animals. The approach is cost-efficient and can be easily applied by computational centres which have the ability and resources to perform complex calculations on large datasets.

REFERENCES

Barrett J.C., Fry B., Maller J., Daly M.J. (2005): Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, 21, 263–265.

Browning B.L., Browning S.R. (2009): A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. American Journal of Human Genetics, 84, 210–223.

Czarnik U., Zabolewicz T., Strychalski J., Grzybowski G., Bogusz M., Walawski K. (2007): Deletion/insertion polymorphism of the prion protein gene (*PRNP*) in Polish Holstein-Friesian cattle. Journal of Applied Genetics, 48, 69–71.

Czarnik U., Grzybowski G., Zabolewicz T., Strychalski J., Kaminski S. (2009): Deletion/insertion polymorphism of the prion protein gene (*PRNP*) in Polish Red cattle, Polish White-backed cattle and European bison (*Bison bonasus* L., 1758). Russian Journal of Genetics, 45, 453–459.

Czarnik U., Strychalski J., Zabolewicz T., Pareek C.S. (2011): Populationwide investigation of two indel polymorphisms at the prion protein gene in Polish Holstein-Friesian cattle. Biochemical Genetics, 49, 303–312.

Druet T., Schrooten C., de Roos A.P. (2010): Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. Journal of Dairy Science, 93, 5443–5454.

Gurgul A., Czarnik U., Larska M., Polak M.P., Strychalski J., Słota E. (2012a): Polymorphism of the prion protein gene (*PRNP*) in Polish cattle affected by classical bovine spongiform encephalopathy. Molecular Biology Reports, 39, 5211–5217.

Gurgul A., Polak M.P., Larska M., Słota E. (2012b): *PRNP* and *SPRN* genes polymorphism in atypical bovine spongiform encephalopathy cases diagnosed in Polish cattle. Journal of Applied Genetics, 53, 337–342.

Hayes B.J., Bowman P.J., Chamberlain A.J., Goddard M.E. (2009): Invited review: Genomic selection in dairy cattle: progress and challenges. Journal of Dairy Science, 92, 433–443.

Hill A.F., Desbruslais M., Joiner S., Sidle K.C.L., Gowland I., Collinge J., Doey L.J., Lantos P. (1997): The same prion strain causes vCJD and BSE. Nature, 389, 448–450.

Hilton D.A., Ghani A.C., Conyers L., Edwards P., McCardle L., Ritchie D., Penney M., Hegazy D., Ironside J.W. (2004): Prevalence of lymphoreticular prion protein accumulation in UK tissue samples. Journal of Pathology, 203, 733–739.

Hunter N., Goldmann W., Smith G., Hope J. (1994): Frequencies of PrP gene variants in healthy cattle and cattle with BSE in Scotland. Veterinary Records, 135, 400–403.

Isler B.J., Freking B.A., Thallman R.M., Heaton M.P., Leymaster K.A. (2006): Evaluation of associations between prion haplotypes and growth, carcass, and meat quality traits in a Dorset × Romanov sheep population. Journal of Animal Science, 84, 783–788.

Jeong B.H., Sohn H.J., Lee J.O., Kim N.H., Kim J.I., Lee S.Y., Cho I.S., Joo Y.S., Carp R.I., Kim Y.S. (2005): Polymorphisms of the prion protein gene (*PRNP*) in Hanwoo (*Bos taurus coreanae*) and Holstein cattle. Genes and Genetic Systems, 80, 303–308.

Juling K., Schwarzenbacher H., Williams J.L., Fries R. (2006): A major genetic component of BSE susceptibility. BMC Biology, 4, 33.

Mulder H.A., Calus M.P., Druet T., Schrooten C. (2012): Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. Journal of Dairy Science, 95, 876–889.

Nakamitsu S., Miyazawa T., Horiuchi M., Onoe S., Ohoba Y., Kitagawa H., Ishiguro N. (2006): Sequence variation of bovine prion protein gene in Japanese cattle (Holstein and Japanese Black). Journal of Veterinary Medical Science, 68, 27–33.

Neibergs H.L., Ryan A.M., Womack J.E., Spooner R.L., Williams J.L. (1994): Polymorphism analysis of the prion gene in BSE-affected and unaffected cattle. Animal Genetics, 25, 313–317.

Parchi P., Giese A., Capellari S., Brown P., Schulz-Schaeffer W., Windl O., Zerr I., Budka H., Kopp N., Piccardo P., Poser S., Rojiani A., Streichemberger N., Julien J., Vital C., Ghetti B., Gambetti P., Kretzschmar H. (1999): Classification of sporadic Creutzfeldt-Jakob disease based on molecular and phenotypic analysis of 300 subjects. Annals of Neurology, 46, 224–233.

Pausch H., Aigner B., Emmerling R., Edel C., Götz K.U., Fries R. (2013): Imputation of high-density genotypes in the Fleckvieh cattle population. Genetics Selection Evolution, 45, 3.

Prusiner S.B. (1998): Prions. Proceedings of the National Academy of Sciences of the United States of America, 95, 13363–13383.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., Maller J., Sklar P., de Bakker P.I.W., Daly M.J., Sham P.C. (2007): PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81, 559–575.

Sander P., Hamann H., Pfeiffer I., Wemheuer W., Brenig B., Groschup M.H., Ziegler U., Distl O., Leeb T. (2004): Analysis of sequence variability of the bovine prion protein gene (*PRNP*) in German cattle breeds. Neurogenetics, 5, 19–25.

Sander P., Hamann H., Drogemuller C., Kashkevich K., Schiebel K., Leeb T. (2005): Bovine prion protein gene (*PRNP*) promoter polymorphisms modulate *PRNP* expression and may be responsible for differences in bovine spongiform encephalopathy susceptibility. The Journal of Biological Chemistry, 280, 37408–37414.

Scott M.R., Will R., Ironside J., Nguyen H-O.B., Tremblay P., DeArmond S.J., Prusiner S.B. (1999): Compelling transgenic evidence for transmission of bovine spongiform encephalopathy prions to humans. Proceedings of the National Academy of Sciences of the United States of America, 96, 15137–15142.

Shibuya S., Higuchi J., Shin R.W., Tateishi J., Kitamoto T. (1998): Codon 219 Lys allele of *PRNP* is not found in sporadic Creutzfeldt-Jakob disease. Annals of Neurology, 43, 826–828.

Strychalski J., Czarnik U., Pierzchała M., Pareek C.S. (2011): Relationship between the insertion/deletion polymorphism within the promoter and the intron 1 sequence of the *PRNP* gene and milk performance traits in cattle. Czech Journal of Animal Science, 56, 151–156.

Ün C., Oztabak K., Ozdemir N., Tesfaye D., Mengi A., Schellander K. (2008): Detection of bovine spongiform encephalopathy-related prion protein gene promoter polymorphisms in local Turkish cattle. Biochemical Genetics, 46, 820–827.

Zarranz J.J., Digon A., Atares B., Rodriguez-Martinez A.B., Arce A., Carrera N., Fernandez-Manchola I., Fernandez-Martinez M., Fernandez-Maiztegui C., Forcadas I., Galdos L., Gomez-Esteban J.C., Ibanez A., Lezcano E., Lopez de Munain A., Marti-Masso J.F., Mendibe M.M., Urtasun M., Uterga J.M., Saracibar N., Velasco F., De Pancorbo M.M. (2005): Phenotypic variability in familial prion diseases due to the *D178N* mutation. Journal of Neurology, Neurosurgery and Psychiatry, 76, 1491–1496.

Zhang Y., Niu T. (2010): Haplotype structure. In: Lin S., Zhao H. (eds): Handbook on Analyzing Human Genetic Data. Springer-Verlag, Berlin-Heidelberg, 25–79.

Received: 2013–06–05
Accepted after corrections: 2014–01–30

Corresponding Author

Dr. Artur Gurgul, National Research Institute of Animal Production, Laboratory of Genomics, Krakowska 1, 32-083 Balice, Poland

Phone: +48 666 801 309, fax: +48 122 856 550, e-mail: artur.gurgul@izoo.krakow.pl