Global Advanced Research Journal of Agricultural Science (ISSN: 2315-5094) Vol. 7(2) pp. 034-045, February, 2018 Issue.
Available online http://garj.org/garjas/home

Full Length Research Paper

High Density of Transposable Elements in Sequenced Sequences in Cattle Genomes, Associated With AGC Microsatellites

[1,2]Valery Glazko, [1]Gleb Kosovsky, [1,2]Tatyana Glazko

[1]Research Institute of Fur Farming and Rabbit Breeding Industries n.a. V.A. Afanasyev, Moscow region, 140143, Russia
[2]Russian Timiryazev State Agrarian University — Moscow Agrarian Academy, Moscow, 127550, Russia

Accepted 08 February, 2018

One of the urgent tasks of agricultural biotechnology is the selection of the most informative DNA markers for polylocus genotyping (genome scan) to control the consolidation and dynamics of gene pool of farm animal species, species origin, identification of the animals maximal promising in terms of productivity and adaptive potential. The study discusses some generations of molecular genetic markers that were applied to solve these problems. It presents the results of sequencing of "anonymous" DNA sequences that were used for polylocus genotyping on the basis of DNA fragments flanked by invert repeats of microsatellite loci (ISSR - Inter-Simple Sequence Repeat). The bovine genomic DNA fragments about 500 bp long flanked with inverted repeat $(AGC)_6G$ were sequenced. The regions with identity of more than 80% to SINE, LINE and ERV Bos taurus-specific retrotransposons were found in all the DNA fragments under study. We revealed a trend towards overrepresentation of the regions with identity to retrotransposons in the genomic sequences of bovine leucosis retrovirus-infected cows. The article discusses a link between increased polymorphism of the primer $(AGC)_6G$ amplification spectra and retrotransposon association of this microsatellite.

Keywords: ISSR—PCR—markers, sequencing, SINE, LINE, ERV, LTR, *Bos taurus* genome, polymorphism.

INTRODUCTION

Novel technologies in agricultural farming, especially in dairy cattle breeding, are based mainly on genomic selection. The essence of the methods is as follows: genotypes of hundreds thousands bulls are screened using single nucleotide polymorphism (SNP) to find polylocus genotypes of animals sharing high selection index, which involves scores based on their daughters' lactation performance; this parameter is used further in predication of probable high selection value of bullocks. It is clear that the use of such an approach leads to an essential reduction of a generation length, especially in the cases when commercial DNA-arrays for multiple SNP genotyping are available. Such companies as Illumina (http://www.illumina.com) and Affymetrix (http://wwwaifymetrix. com) have developed various panels of cattle SNP-arrays: for 3,000 DNA fragments 3K (Wiggans GR et al., 2012), 7K (Boichard D et al., 2012), 15K (Khatkar MS et al., 2007), 25K (Raadsma HW et al., 2009), 50K (Matukumalli LK et al., 2009), and more detailed for 800K (Illumina) and 650K and 3 mln DNA fragments (Affymetrix) (Metzker ML, 2010).

The same time, use of high density SNP-arrays often

*Corresponding Author's Email: tglazko@rambler.ru

contradicts the work profitability (Pryce J, Hayes B, 2012, Khatkar MS et al., 2012). Relatively low profitability of SNP application to forecast cattle economic characters can be caused by variety of reasons that can be distributed into 3 groups. The first one is that more than 50% of hereditary variations of the main phenes included into the analysis fall into genomic regions with small phenotypic effects, whose order of values corresponds to the polygenic inheritance (Dekkers ICM, 2012). This can be illustrated by an attempt to map main genes of lactation performance in three French specialized dairy breeds using SNP. This research revealed different genes sharing only one thing: their expression is controlled by a hypothalamic-adrenal axis (Flori L. et al., 2009).

In animals of other species, SNP-genotypes and manifestation of economic characters also can depend on breed characteristics of their genetic structure.

For example, we have analyzed SNP associations within promotors of specialized genes of lipid metabolism (Scd, Lep), myogenesis (Myod, Myf-6) and of pluripotential regulatory protein gene (Opn) with swine pork-and-lard characteristics (Khlopova NS et al., 2012). The revealed associations depended on an animal gender and origin (associations in two-breed crosses differed from those in three-breed). Associations between SNP of Opn gene promoter and pork productivity parameter, fat depth, were the most frequent. These data tell that the examined SNP—genotypes indeed can be used to increase effectiveness of animals' productivity potential forecast but repeatability of these associations is limited to their gender and shared origin.

In addition, heritability estimate of total milk yield, a parameter of milk productivity, is usually low and depends essentially on environmental factors. This can be shown by assessment of bull breeding value on the base of a bull daughters milk productivity in relation with the daughters born in different ecological and geographical regions (Hammami H et al., 2008, 2009a, 2009b).

Techniques used to form SNP-panels comprise the second group of factors causing contradictions. The available techniques are related with inevitable missequencing of cattle genome, problems of differentiation between occasional, structural and rare functionally significant allelic variants of SNP, differences in SNP frequencies within duplicated and single sequences of genomic DNA (Zhan B et al., 2011), along with mistakes in revealing of duplicated genomic sequences (Zimin AV et al., 2012).

Problems of data bulk mathematical processing, associations modeling and their fact-based analysis make the third group of problems in SNP—panels application for control of genome structure of cattle groups and revealing of genotypes associated with phenotypic manifestation of economic traits (Schwarzcnbacher H et al., 2012, Lewis J et al., 2011).

A total of the listed problems returns us to the relatively simpler methods of cattle genome scanning (polylocus genotyping) and development of simpler, more effective and less expensive methods making it possible to solve classical problems of farm animals genetics: exclusion of origin misinterpreting; control of hereditary diseases and pathogen contamination; reconstruction of breed history and genealogy and revealing of their genepool specifics; genomic selection targeted mapping of primary genes of polygenic characters; creating of methods to forecast amount and quality of final products, resistance to housing conditions and infectious agents; creating of genetically based programs of stable use and preservation of local breeds (Kharchenko RN, Glazko VI, 2006, Utsunomiya YT et al., 2014).

In recent years, the listed problems were solved with wide use of microsatellite locus genotyping. Company Applied Biosystems (www.appliedbiosystems.com) upon consultation with Food and Agriculture Organization (FAO) and International Society of Animal Genetics (ISAG) developed a test system for cattle genotyping by 11 microsatellites. However, it was found that a phylogenetic tree of cattle breeds constructed in accordance with microsatellite genotyping differs essentially from a trees reconstructed by genome sequences (Ritz LK et al., 2000, Hassanin A, Ropiquet A, 2004).

The main problem in use of microsatellite based genotyping may be that they involve small numbers of loci (twenty or less), whose polymorphism (allelic variation, spontaneous mutability and rate of allele fixation) varies essentially from one locus to the next. Microsatellites forming inverted repeats within 2,000 base pairs are of special interest for this matter, as they can be used as primers in polymerase chain reaction (PCR) to create polylocus spectra of DNA fragments convenient for genome sequencing (Kharchenko RN, Glazko VI, 2006). The method is named ISSR-PCR (ISSR-Inter-Simple Sequence Repeat).

Availability of structural characteristics of ISSR-marker flanks (looping disposition) is their obvious advantage over, for example, AFLP (Amplified Fragment Length Polymorphism) used in polylocus genotyping; these are anonymous in relation to nucleotide composition of amplified DNA fragments too (Utsunomiya YT et al., 2014).

Dominant character of ISSR-markers manifestation is their essential disadvantage because presence of a DNA fragment of certain length in the spectrum of amplification products resulting PCR, in which a microsatellite fragment was used as a primer, makes it impossible to distinguish between homozygotes and heterozygotes. Moreover, anonymous nucleotide composition of amplified DNA fragments of the same length cannot rule out a possibility of their internal nucleotide heterogeneity at amplification from different genome regions.

One of the objective advantages of ISSR-markers is their definability, that is, a DNA-region to be amplified can be located only in a genome region containing an inverted

microsatellite repeat within 2,000 base pairs; this fact makes it possible to analyse genome distribution of the repeats. Distribution of the inverted AGC repeats is of special interest because it can be found in the cattle genome approximately five time more frequently than in the sequenced genomes of other mammals, and, moreover, in 39% cases these microsatellites are closely associated with a BovA retrotransposon (Elsic CG et al., 2009).

Transposable elements (TEs) or transposons are well-known factors of genomic variability and evolution (Chénais B et al., 2012). For a long time TEs had been recognized as major sources of non-coding or "junk" DNA with uncertain functions in an organism's genome. The current data on TEs abundance in the genomes of different taxa of prokaryotes and eukaryotes show a lack of correlation between the level of organism's evolution and the amount of transposons hosted in its genome. Despite being considered as genetic burden for the host genomes, today exhaustive research projects such as ENCODE (Encyclopedia of DNA Elements) are gathering reliable evidence towards the important role of transposons in environmental adaptation of organisms and disease. Due to their structural elements and ability to multiply, transpose and mutate, TEs become important mutators and providers of regulatory elements that are able to significantly influence the architecture and expression of the host genome (Rebollo R et al., 2011; Walsh et al., 2013).

The aim of our research was the analysis of DNA fragments flanked with an inverted repeat of a microsatellite AGC region, that is, analysis of their nucleotide composition and variability. We sequenced genome DNA fragments from six Holstein cows of Mozhayskoe farm. Length of the fragments was approximately 500 base pairs, and they were flanked with an inverted repeat $(AGC)_6G$. The microsatellite is closely associated with retrotransposons, so we made a comparative analysis of sequences in three cows infected with bovine leucosis retrovirus, and in three infection-free cows, whose contamination was assessed earlier (Kosovsky GYu et al., 2013).

MATERIALS AND METHODS

Total DNA was isolated from whole blood of six Black-and-White "holsteinesed" (that is, originated from the Holsteins) cows. Blood samples were taken in December 2011. Blood was drawn from caudal veins with sterile catheters as sampling device and EDTA as an anticoagulant. Results of our previous study (Kosovsky GYu, et al., 2013) show that 3 cows (№ 306, 333, 456) were infected with bovine leucosis virus, while 3 cows (№ 301, 322, 431) were infection-free.

DNA was isolated with Magna^{TM} DNA Prep 200 kit (Isogen Laboratory, Russia). Twenty microliters of incubation mixture contained 2 mcl of 10—fold buffer and 1 mcl (5 units) of Taq-polymerase (Syntol, Russia), 2 mcl of dNTP solutions (10 mM each), 1 mcl of $(AGC)_6G$ primer (20 pmol), 2 pmol (0.5—1 mcg) of DNA, 12 mcl of deionized water. PCR was carried out as follows: initial denaturation: 1 min at 94°C; 35 cycles (30 sec at 94°C, 30 sec at 55°C, and 2 min at 72°C); final elongation: 10 min at 72°C. Products of amplification were electrophoresed in 1% agarose gel. Two ladders, M25 DNA Ladder 1 and M11 DNA Ladder (Sib Enzyme, Russia), were used as markers.

DNA fragments, 400—550 base pairs long, separated from agarose gels were sequenced. DNA-fragments library was produced according to the protocol of the DNA Library Quick Preparation, clonal emulsion PCR and sequencing were carried out according to the manufacturer (Roche) recommendations and with their reagent kits. Nucleotide sequences of amplicons were determined with a genome analyzer GS Junior (Roche). Preparation of amplicon libraries and sequencing were carried out twice. Approximately 22 mln base pairs were determined for the first instrument run.

Seventy thousand and two hundred and twenty three readings were identified as suitable for the further analysis, an average reading length was approximately 315 base pairs. Sequencing of the second amplicon library resulted in determination of approximately 36 mln base pairs and in 87,351 readings with an average length of 415 base pairs.

The sequenced data were organized using the appropriate methods of analysis (Ewing B, Green P, 1998; Ewing B et al., 1998; Gordon D, Abajian C, Green P, 1998). The sequences were clustered according to their identity.

Alignment of each cluster sequences with the Bos taurus chromosome sequences was carried out using algorithms of the BLAST n program (https://blast.ncbi.nlm.nih.gov/). Presence of regions identical to microsatellites and interspersed repeats in the sequences was analyzed with the computer programs Repeat Masker (http://www.repeatmasker.org/) and Giri (http://www.girinst.org).

RESULTS AND DISCUSSION

To analyze nucleotide composition of cattle DNA fragments flanked by an inverted repeat of the microsatellite region AGC, we scanned the genomes of "Holsteinized" cows from Mozhayskoye farm. The cows genomes were sequenced by ISSR-PCR with $(AGC)_6G$ primer; from the spectra of six cows, DNA fragments 450 -500 base pairs long were isolated and their base sequences were determined by pyrosequencing. The sequences were clustered according to their identity that should be ≥95%. Repeatability of the results was controlled by sequencing of libraries of amplicons found in both determinations. The

Table 1: Numbers of clusters included in the analysis of two independent procedures of genome DNA fragments sequencing with approximately 500 base pairs length resulting PCR with $(AGC)_6$ G primer

Cow №	Number of clusters at the first sequencing	Number of clusters at the second sequencing
301	142	316
306	111	126
322	87	120
333	80	38
431	41	105
456	122	128
Total	583	833

amounts of the clusters analyzed after the first and second sequencings are shown in Table 1.

At the first stage of sequence clusters analysis, we examined only those containing not less than 40 sequences (readings). It appeared that each examined cow had its own pattern of base sequences of various lengths (from 200 to 500 base pairs) highly identical (≥97%) to the base sequences of the Bos taurus breed Hereford Btau_4.6.1 and the Bos taurus UMD 3.1 genomes. Sequences of these clusters contain regions of a structural gene encoding the 6th isoform of the CREB5 transcription factor (cyclic AMP-responsive element–binding protein 5 isoform X8), localized in chromosome 4; of an intergenic region 61,581–114,071 base pairs of chromosome 14, the 5'end of which contains a gene encoding a brain-specific angiogenesis inhibitor 1 isoform X, and the 3'end – a gene encoding the ubiquitin carboxyl–terminal hydrolase 24 localized at chromosome 3.

It is significant that at sequencing of a freshly prepared library of the same replicons, these, relatively richer in number of readings, clusters were reproduced again. The same time, a resequencing revealed the clusters, in which numbers of readings excessed 40, that was different from the results of the first sequencing. A list of these regions is presented in Table 2.

Some regions identical more than by 95% to the sequences found in course of the first sequencing are unique, that is, they have no regions over 20-40 base pairs identical to the sequences resulting the second sequencing. These highly identical regions include: genes of FAM35A and calcipressin-2, region localized within an intergenic region 33,872 – 40,724 base pairs of chromosome 10 flanked with the gene of TMCO5B (transmembrane and coiled-coil domain-containing protein SB) at the 5' end and formin-l gene at the 3'end (Table 2). It is worth noting that the cows differed in regard to these sequences, too: the cow № 431 lacked all three of them, and an intergenic fragment 33,872 – 40,724 base pairs of chromosome 10 was found only in one cow, № 322 (Table. 2). These differences may be the results of individual mutations, the more so since a relatively increased density of potential G4-quadruplexes, markers of genomic instability, can be observed in this region (Figure. 1) (Zybailov RL et al., 2013).

A region more than 95% identical to an intergenic region 12,9097 – 112,110 base pairs of chromosome 21, flanked with the lactadherin precursor gene at the 5'end and abhydrolase domain-containing protein 2 gene at the 3'end, presents in the results of both, the first and the second, sequencings. However, the second sequencing produced several clusters with >40 readings, while after the first sequencing we found them only in two cows, № 333 and 456, and in the clusters with <10 readings. The second sequencing revealed a fragment identical to a chromosome 10 site encoding MEGF11 (multiple EGF-like-domains 11) in all cows, while the first sequencing could find it only in the cow № 322. A region identical to the intergenic fragment 152,949th —88,300th base pairs of chromosome 12 flanked with the gene of E3 ubiquitin-protein ligase RNF146-B at the 5'end and the gene coding a heat shock protein 105 kDa at the 3'end can be found in three animals at the 2nd sequencing and in three cows at the first one: № 301, 306 and 333. A sequence >95% identical to the region of chromosome 11 encoding a not characterized protein, C90rf171, (Table 2, the 2nd sequencing) was found after the 1st sequencing of cow №306, too. However, the results of the first sequencing did not reveal regions identical to the region 15,188th –13,604th base pairs of chromosome 19 flanked with the gene of keratin, type I cytoskeletal 14 at the 5'end and the gene of LOW QUALITY PROTEIN: keratin, type I cytoskeletal 16 at the 3'end that were found in all cows but № 301 after the 2nd sequencing. However, it is worth mentioning that all sequences of this region found in the results of the 2nd sequencing are relatively short (<300 base pairs) and contain guanine repeats predisposed to form alternative DNA structures, G4—quadruplexes, markers of genome instability (Zybailov RL et al., 2013).

Sequences containing regions with high (>97%) identity within different chromosomes in parallel, for example, the 23rd and the 4th at the 1st sequencing, and the 8th, 15th and X at the 2nd one (Table 2), can be found in both series but within the clusters with different numbers of readings. For example, only two cows from six, № 301 and 431, after the

Table 2: Differences between the 1st and the 2nd sequencing of an amplification-produced DNA fragments approximately 500 base pair-long. The fragments of the genome DNAs of the same cows were amplified with the (AGC)6G primer and organized into sequence clusters over 40 identical by >90%

Cow №	№ 301 healthy high yielding	№ 306 infected high yielding	№ 322 healthy low yielding	№ 333 infected high yielding	№431 healthy high yielding	№ 456 infected low yielding
First sequencing	Chromosome 28: the FAM35A gene	<40	<40	<40	absent	<40
	Chromosome 23: the calcipressin-2 gene	Chromosome 23: the calcipressin-2 gene	<40	<40	absent	<40
	absent	absent	Chromosome 10: the intergenic region space, 33,872 base pairs from the 5'end — gene of transmembrane and coiled-coil domain-containing protein 5B and 40,724 base pairs from the 3'end — formin-1 gene	absent	absent	absent
	<40	Region of identity found at 23th and 4th chromosomes	<40	<40	absent	<40
Second sequencing	Chromosome 21: the intergenic region, 12,909 base pairs from the 5'end: lactadherin precursor gene and 112,110 base pairs from the 3'end abhydrolase domain-containing protein 2 gene	Chromosome 21: the intergenic region, 12,909 base pairs from the 5'end: lactadherin precursor gene and 112,110 base pairs from the 3'end abhydrolase domain-containing protein 2 gene	<40	<40	<40	<40
	Chromosome 10: the gene of multiple EGF-like-domains 11	Chromosome 10: the gene of multiple EGF-like-domains 11	<40	<40	<40	<40
	<40	<40	<40	absent	Chromosome 12: the intergenic region, 152,949 base pairs from the 5'end — gene of E3 ubiquitin—protein ligase RNF146—B, 88,300 base pairs from the 3'end — gene of heat shock protein 105 kDa	no

Table 2: Continue

	<40	<40	<40	<40	absent	Chromosome 11: the gene of the uncharacterized protein C90rf171
	absent	<40	<40	<40	<40	Chromosome 19: the intergenic region, 15,188 base pairs from the 5'end = gene of keratin, type I cytoskeletal 14) (1 fragment), 13,604 base pairs from the 3'end — gene of LOW QUALITY PROTEIN: keratin, type I cytoskeletal 16
	Regions of identity at chromosomes 8, 15 and X	Regions of identity at chromosomes 8, 15 and X	Regions of identity at chromosomes 8, 15 and X	Regions of identity at chromosomes 8, 15 and X	Regions of identity at chromosomes 8, 15 and X	Regions of identity at chromosomes 8, 15 and X

Figure 1. Distribution of non-overlapping potential G4—quadruplexes found in the unique fragment identical with an intergenic fragment of 33,872—40,724 base pairs of chromosome 10, 5' end of which includes a gene encoding transmembrane and coiled-coil domain-containing protein 5B, and 3' end — gene encoding formin-1.
Note: a fragment is 487 base pair-long, it contains 3 non overlapping potential G4-quadruplexes (underlined) and 151 overlapping fragments.

1st sequencing, had no regions identical to fragments of all three chromosomes – 8th, 15th and X – together.

The results of two sequences differ rather in quantity than in quality, that is, the presence of different amounts of sequenced DNA fragments (readings) in the clusters identical to the same cattle genome regions. Several factors may explain these differences: automatic exclusion of fragments unfit for sequencing due to some reasons, this could be confirmed by almost twofold decrease in amounts of sequences remaining after deleting of low quality results of sequencing (%Passed Filter); different amounts of the fragments resulting PCR due to low temperature of primer annealing usual for processing of ISSR-PCR markers; context features of sequenced fragments.

Table 3: Regions with high (>80%) identity with dispersed repeats within clusters of sequences different in 1st and 2nd sequencings

	First sequencing		**Second sequencing**				
	Chromosome 28: a gene of protein FAM35A isoform X1 protein	Chromosome 23: a gene of calcipressin-2	Chromosome 21: an intergenic region, 12,909 base pairs from the 5'end: lactadherin precursor gene and 112,110 base pairs from the 3'end abhydrolase domain-containing protein 2 gene	Chromosome 10 — gene of multiple EGF-like-domains 11	Chromosome 12 — intergenic region, 152,949 base pairs from the 5'end — gene of E3 ubiquitin—protein ligase RNF146—B, 88,300 base pairs from the 3'end — gene of heat shock protein 105 kDa	Chromosome 11 – gene of uncharacterized protein C90rf171	Chromosome 19 – intergenic region, 15,188 base pairs from the 5'end = gene of keratin, type I cytoskeletal 14) (1 fragment), 13,604 base pairs from the 3'end — gene of LOW QUALITY PROTEIN: keratin, type I cytoskeletal 16
Length of a cluster	473	412	476	362	456	336	350
Identity with dispersed repeats (~>80%)	Positions 17-189 CHR-2A SINE Non-LTR Ruminants retrotransposon	Positions 47-408 L1-2_BT-	Positions 28-80 MER87B_BT – long terminal repeat of cattle 1st class endogenous retrovirus (EVR1)	Positions 8-38 BEL9-I_DR LTR – Danio rerio retrotransposon	Positions 167-369 Bov-tA1 SINE subfamily of Ruminants	Positions 5-46 Gypsy3-ZM_I	Positions 115-314 L1
	Positions 190-471 L1MC4 ancestor mammal class specific subfamily LINE-1	-	Positions 240-476 BovB_Oa2951-3187-NonLTR/RTE	Positions 39-125 LTR Gypsy	-	Positions 257-328 L1NE3C_3end-L1	Positions 323-350 Gypsy-139_AA-I
	-	-	-	Positions 226-333 – Gypsy-like LTR	-		

To examine possible context features of the sequenced fragments with different amounts of readings after the 1st and the 2nd sequencing, we screened these sequences for the fragments identical to various repeats that can be a problem at amplification and sequencing; the screening was made with the programs Repeat Masker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) and Giri (http://www.girinst.org). In all the cases, the screening revealed regions identical to SINE, LINE and endogenous retroviruses species, specific mainly for Bos taurus (Table 3). At that, some fragments contained products of retrotransposons recombination, for example, a site of chromosome 28 encoding protein FAM35A isoform X1, 473 base pair-long, where we found a fragment more than 95% identical to SINE within a fragment from the 17th to the 189th base pairs, and to LINE 1 in fragment from the 190th to the 471st base pairs; fragment of chromosome 10 encoding multiple EGF-like-domains 11, 362 base pair-long, containing a fragment, from the 8th to the 38th base pairs, more than 95% identical to the long terminal repeat (LTR) of the endogenous virus, initially described in fishes, and from the 39th to the 125th base pairs to the LTR of another endogenous retrovirus, Gypsy (Table 3). In some cases, almost the whole fragment was more than 95% identical to a retrotransposon, for example, a fragment of chromosome 23,412 base pair-long, encoding calcipressin-2, where a fragment including the 47th–408th base pairs was identical to the L1-2_BT sequence.

Table 4: Regions highly identical to sequences in the 23rd and 4th chromosomes

Cow number						
Fragment length	465	461	476	472	no	476
Positions ERV1-2C-LTR_BT	1—232	1—232	1—232	1—232		1—232
Fragment #1	15—249	15—251	15—251	15—250		15—250
L1-2_BT	1770—1999	1777—1999	1770—1999	1770—1999		1770—1999
Fragment #1	250—465	252—461	252—470	251 —466		251—468
Fragment length	431	404	434	434		386
Positions ERV1-2C-LTR_BT	1-232	1-232	1-232	1-232		1-232
Fragment #2	15—249	15—254	15—250	15—251		15—253
L1-2_BT	1805—1999	1837—1999	1803—1999	1804—1999		1855—1999
Fragment #2	250—431	255—404	251—433	252—434		254—386
Fragment length	447	261				
Positions ERV1-2C-LTR_BT	1—232	46—232				
Fragment #3	15-249	67-258				
L1-2_BT	1792—1999	no				

Figure 2. Regions highly identical (>80%) to the 1 – 232 LTR of the 1 class endogenous virus of Bos Taurus (ERV1-2C-LTR_BT) and region from the 1770th to the 1999th LINE of the same species L1—2_BT, the 1st fragment of the cluster from the 15th to the 249th base, and the 2nd from the 250th to the 465th base (color marked)

We found a similar situation at the fragments identical to regions of different chromosomes in parallel. For example, in all clusters, with a sequence highly identical (>97%) to fragments of both 23rd and 4th chromosomes, the sequence includes fragments, from the 15th to the 249th identical (>80%) to the 1st – 232nd base pairs of LTR of the 1st class endogenous virus of Bos taurus (ERV1-2C-LTR_BT) and to the 250th – 465th base pairs to the region 1,770th – 1,999th base pairs of LINE of the same species, L1-2_BT (Table 4, Figure. 2). The same animals had clusters of various lengths and identities, clusters of animals infected with the bovine leucosis virus differed somewhat more (Table 4).

Highly identical regions, approximately 400 base pairs long, were found at 3 chromosomes in parallel: at the 8th chromosome (intergenic region from the 267,117th to the 107,184th base pairs flanked with a gene encoding protein S100-A11 at the 5'end and gene of LOW QUALITY PROTEIN: ras-related GTP-binding protein A at the 3'end; at the 15th chromosome (intergenic region from the 59,844th to the 649,718th base pairs, the 5'end of which contains a gene encoding the progesterone receptor, and the 3'end contains a gene of contact in 5 precursor; and at the X chromosome (intergenic region from the 234,390th to the 43,179th, 5'end contains a gene of tumor necrosis factor receptor superfamily member 27, and the 3'end contains a gene of ligase E3 ubiquitin-protein ligase SIAH1-like. We found sequences homologous to the LTR Gypsy, ERV1-2-I_BT, SINE BOVA2 in these regions. In addition, we found apparent polymorphism not only with regard to the fragment lengths but to the deletions within the fragments in the clusters of different lengths of the same cow, and in the clusters of different animals. These data suggest increased frequency of spontaneous mutagenesis to be typical for these regions.

These data tell that the sequences located in the examined DNA regions flanked with the inverted $(AGC)_6G$ repeat are rich in regions highly identical to such mobile elements as L1-2_BT, BOVA, ERV1-2-I_BT and to the Gypsy LTR widely distributed in various genomes. It is reasonable to expect that the increased variability of these sequences tells about spontaneous mutagenesis in different cell populations of the same cow and between different animals. It is notable that, in whole, the regions identical to the regions of such Bos taurus retrotransposons as LTR ERV3, Bov B, L2, Bov-tA are

Table 5: Representativity of the regions identical to dispersed repeats and bacterial genomes in the cattle DNA sequences flanked with the inverted $(AGC)_6G$ repeat

Pseudomonas fluorescens						
Number of clusters	11		4			
Number of reading per a cluster	11		3-8			
Fragment length	442		441-470			
Acidovorax ebreus TPSV						
Number of clusters			1			
Number of reading per a cluster			10			
Fragment length			461			
Streptomyces bingchenggensis BCW-1						
Number of clusters		1				
Number of reading per a cluster		2				
Fragment length		372				
Anaplasma marginale						
Number of clusters						3
Number of reading per a cluster						2
Fragment length						11-467
Sanguibacter keddieii						
Number of clusters				1		1
Number of reading per a cluster				2		4
Fragment length				478		480
Propionbacterium acnes						
Number of clusters		3		1		
Number of reading per a cluster		3-11		2		
Fragment length		405-478		310		
R21-1_PINeSL non-LTR retrotransposons of Phytophthora genomes						
Number of clusters		2				
Number of reading per a cluster		5-8				
Fragment length		426-463				
L-1_2BT-	47 - 422	47 - 422				48 - 422
ERV1-2C-I-LTR_BT	423 - 490	425- 492				423 - 490
Number of clusters	1	1				1
Number of reading per a cluster	6	4				4
Fragment length	492	492				496
Copia-7_ES-I-	1 - 42					
Copia-141_SB-LTR	180 - 252					
Number of clusters	1					
Number of reading per a cluster	5					
Fragment length	433					

Table 5: Continue

L1-2_BT	47 - 263	47 - 422	47 - 419	46 - 417		46 - 417
Number of clusters	3	5	2	4		9
Number of reading per a cluster	2-5	2-92	2-4	2-14		2-19
Fragment length	264-302	256-486	419-436	255-476		157-494
Bov-tA3	55 - 254	30 - 238				
Number of clusters	1	1				
Number of reading per a cluster	2	6				
Fragment length	256	243				
BOVA2		30 - 158				21 - 216
Number of clusters		1				1
Number of reading per a cluster		6				2
Fragment length		450				230
LTR2C_BT ERV1-			3 - 60			
BovB_Oa			63 - 370	25 - 335		
Number of clusters			1	1		
Number of reading per a cluster			4	2		
Fragment length			378	374		
Jockey-3_Dgri non-LTR retro			1 - 100			
Number of clusters			1			
Number of reading per a cluster			2			
Fragment length			413			
Harbinger-5_Ami-DNA transposon			33 - 224			
LINE-1-58_SB			371 - 450			
Number of clusters			1			
Number of reading per a cluster			2			
Fragment length			454			
L1-2_VPA-				26 - 74		
BovB-				86 - 267		
L1-3_Vpa				276 - 389		
Number of clusters				1		
Number of reading per a cluster				2		
Fragment length				407		
L1_Carn5_3end-						14 - 124
L1-2_Trt						139 - 357
Number of clusters						1
Number of reading per a cluster						6
Fragment length						377
L1-2_Trt						21 - 220
Number of clusters						2
Number of reading per a cluster						2
Fragment length						220-248

rarer in the sequenced clusters of all cows free of bovine leucosis virus at both sequencings.

The clusters resulting both sequencings contain regions lacking identity to the genomes of Bos Taurus breed Hereford Btau_4.6.1 and Bos_Taurus_UMD_3.1. Usually, these regions contain sequences of bacterial DNA (Table 5). In most such clusters these regions are presented by small numbers of readings, however, showing the sequencing capabilities and presence of the AGC microsatellite in the bacterial genomes, they are of some interest (Table 5).

The revealed fragments of bacterial genomes lacking sequences identical to those in the cattle genome may be distributed into two types (Table 5): those associated with plant and soil microbiota (Pseudomonas fluorescens, Acidovorax ebreus, Streptomyces bingcnenggensis, Phytophtora), and mammal-associated or mammal pathogens (Sanguibacter keddieii, Propionibacterium acnes, Anaplasma marginale).

All other sequences were highly identical to the retrotransposons, mainly to L1-2_BT, SINE Bos Taurus and LTRs of some endogenous retroviruses of ancient origin and specific for Bos Taurus (Table 5). Conspicuous is the fact that the regions identical to the L1-2_BT in the sequences lacking identity with Bos Taurus genomes present in the GenBank are longer than those having the identities (Tables 4, 5), at that products of recombinations between L1-2_BT and ERV1-2C-I-LTR_BT (Table 5) of the first group are ordered differently than in the second one (Table 4).

These data tell that the variants of the retrotransposon found in these fragments and products of their recombination appeared relatively recently: after the separation between specialized Hereford beef cattle, whose genome sequence is presented in the GenBank, and the Holstein-based cattle examined in the study. A number of these non-identical sequences is somewhat higher in animals infected with bovine leucosis virus than in non-infected ones. It suggests that a relatively increased sensitivity to retroviral contamination is associated to some extent with the intracellular mechanisms of retrotransposition protection.

Such an idea needs further studies because in spite of a small number of analysed animals, and complexity of results of double sequencing of DNA fragments flanked with the $(AGC)_6G$ inverted repeat, it shows their participation in high frequency of spontaneous mutating of cattle genome.

After the whole cattle genome had been sequenced in 2009, the international consortium paid special attention to the AGC microsatellite. This attention can be explained by several factors: 1) the microsatellite is found in the cattle genome 90- and 142-fold more frequently than in the human and dog genomes, respectively (Elsic CG, Tellam RL, Worley KC, 2009); 2) in the cattle genome, the AGC microsatellite in 39% cases is associated with the presence of Bov-A2 SINE. The element has the AGC sequence in the tail piece and is a more recent derivative of an evolutionary ancient cattle repeat, Bov-B (LINE). It is worth mentioning that sequences homologous to this long dispersed nuclear element can be met in 65 taxa and can be involved into horizontal transfer of genetic material between various taxa (Gordon D, Abajian C, Green P, 1998). In total, LINE_L1 takes 11.3% of cattle genome, BovA — 10.1%, LTR_ERV1 and LTR_ERV3 — approximately 0.8%, microsatellite AGC — 0.1%, at that according to (Elsic CG, Tellam RL, Worley KC, 2009), localization of the microsatellite AGC correlates with the presence of not only SINE elements, but microsatellites CCG, CG, AGG, ACC, AC, AG and ACG, too. Comparison of these data with the obtained sequences tells that LINE and LTR elements of endogenous retroviruses are found in the fragments flanked with inverted AGC repeat more frequently than the microsatellite sequences (Tables 3 – 5).

The data allows us to make the following conclusion: a sequencing performance depends on a nucleotide context. When the nucleotide context contains homopolymers, regions with predisposition to form alternative DNA structures, and retrotransposon regions, sequencing can result in fragments of various lengths. Moreover, presence of these elements promotes missequencing and spontaneous mutagenesis.

Use of the $(AGC)_6G$ primer for cattle genome scanning makes it possible to obtain polylocus spectra of amplification products, whose polymorphism is closely associated with retrotranspositions and products of recombinations between retrotransposons. We noticed a definite trend towards increased frequency of regions identical to retrotransposons or products of their recombination within sequenced cattle genome infected with bovine leucosis virus. Spectra of amplification products obtained with the use of the $(AGC)_6G$ primer could be especially effective for study of interbreed genetic differentiation, including that associated with the retroviral infections.

REFERENCES

Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, Gietzen KJ, Hayes BJ, Lawley CT, Sonstegard TS, Van Tassell CP, Van Raden PM, Viaud-Martinez KA, Wiggans GR (2012). Bovine LD Consortium. Design of a Bovine Low-Density SNP Array Optimized for Imputation. *PLoS One*, 7(3):e34130. doi:10.1371/journal.pone.0034130

Chénais B, Caruso A, Hiard S, Casse N (2012). The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509(1): 7-15. doi: 10.1016/j.gene.2012.07.042

Dekkers JCM (2012). Application of Genomics Tools to Animal Breeding. *Current Genomics*, 13(3):207—212.

Elsic CG, Tellam RL, Worley KC (2009). The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science*, 324(5926):522—528.

Ewing B, Green P (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3):186—194.

Ewing B, Hillier L, Wendl M C, Green, P (1998). Base—calling of automated sequencer traces using phred. 1. Accuracy assessment. *Genome Research*, 8(3):175 – 185.

Flori L, Fritz S, Jaffrézic F, Boussaha M, Gut I, Heath S, Foulley JL, Gautier M (2009). The Genome Response to Artificial Selection: A Case Study in Dairy Cattle. *PLoS ONE*, 4(8): e6595. doi:10.137l/journal.pone.0006595.

Gordon D, Abajian C, Green P (1998). Consed: a graphical tool for sequence finishing. *Genome Research*, 8(3):195—202.

Hammami H, Rekik B, Bastin C, Soyeurt H, Bormann J, Stoll, Gengler N (2009a). Environmental sensitivity for milk yield in Luxembourg and Tunisian Holsteins by herd management level. *Journal of Dairy Science*, 92(9):4604-4612.

Hammami H, Rekik B, Soyeurt H, Bastin C, Bay B, Stoll J, Gengler N (2009b). Accessing genotype by environment interaction using within- and across-country test-day random regression sire models. *Journal of Animal Breeding and Genetics*, 126(5):366-377.

Hammami H, Rekik B, Soyeurt H, Bastin C, Stoll J, Gengler N (2008). Genotype x environment interaction for milk yield in Holsteins using Luxembourg and Tunisian populations. *Journal of Dairy Science*, 91(9):3661—3671.

Hassanin A, Ropiquet A (2004). Molecular phylogeny of the tribe Bovini (Bovidae, Bovinae) and the taxonomic status of the Kouprey, Bos sauveli Urbain 1937. *Molecular Phylogenetics Evolution*, 33(3):896—907.

Kharchenko RN, Glazko VI (2006). *DNA-technologies in Development of Agrobiolog*. Moscow: Voskresenie.

Khatkar MS, Moser G, Hayes BI, Raadsma HW (2012). Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. *BMC Genomics*, 13:538. doi:10.1186/14717-2164-13-538

Khatkar MS, Zenger KR, Hobbs M, Hawken RJ, Cavanagh JA, Barris W, McClintock AE, McClintock S, Thomson RC, Tier B, Nicholas FW, Raadsma HW (2007). A primary assembly of a bovine haplotype block map based on a 15,036-singleenucleotide polymorphism panel genotyped in Holstein-Friesian cattle. *Genetics*, 176(2):763-772.

Khlopova NS, Stefanon B, Guatti D, Glazko TT, Glazko VI (2012). Mononukleotidnyj polimorfizm genovekandidatov kontrolya pokazatelej produktivnosti svinej. *Doklady Rossiiskoi Academii Selskokhazyaistvennykh Nauk*, no 4:39-45.

Kosovsky GYU, Sotnikova EA, Mudrik NN, Cuong Vu Chi, Toan Tran Xuan, Hoan Tran Xuan, Glazko VI (2013). Diagnostic enzootic bovine leukosis infection using primers of genes gag and pol. *Journal of Veterinariya*, no 8:58—61.

Lewis J, Abas Z, Dadousis C, Lykidis D, Paschou P, Drineas P (2011). Tracing Cattle Breeds with Principal Components Analysis Ancestry Informative SNPs. *PLoS ONE* 6(4): e18007. doi: 10.1371/journal.pone.0018007.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS, VanTassell GP (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One*, 4(4): e5350. doi: 10.1371/journal.pone.0005350.

Metzker ML (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31—46.

Pryce J, Hayes B (2012). A review of how dairy farmers can use and profit from genomic technologies. *Animal Production Science*, 52(2—3):180—184.

Raadsma HW, Khatkar MS, Moser G, Hobbs M, Crump RE, Cavanagh JA, Tier B (2009). Genome wide association studies in dairy cattle using high density SNP scans. *Prac Assoc Advmt Anim Breed Genet*, 18:151-154.

Rebollo R, Romanish MT, Mager DL (2011). Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics*, 46(1):21-42

Ritz LK, Glowatzki-Mullis ML, Mac Hugh DE, Gaillard C (2000). Phylogenetic analysis of the tribe Bovini using microsatellites. *Journal of Animal Genetic*, 31(3):178—185.

Schwarzcnbacher H, Dolezal M, Flisikowski K, Seefried F, Wurmser C, Schlotterer C, Fries R (2012). Combining evidence of selection with association analysis increases power to detect regions influencing complex traits in dairy cattle. *BMC Genomics*, 13:48. doi:10.1186/14717216443748.

Utsunomiya YT, Bomba L, Lucente G, Colli L, Negrini R, Lenstra JA, Erhardt G, Garcia JF, Ajmone-Marsan P (2014). European Cattle Genetic Diversity Consortium. Revisiting AFLP fingerprinting for an unbiased assessment of genetic structure and differentiation of taurine and zebu cattle. *BMC Genetics*, 15:47. doi:10.1186/1471—2156-15-47

Walsh AM, Kortschak RD, Gardnerb MG, Bertozzia T, Adelsona DL (2013). Widespread horizontal transfer of retrotransposons. *Proc Natl Acad Sci USA*, 110(3):1012 – 1016.

Wiggans GR, Cooper TA, VanRaden RM., Olson KM, Tooker ME (2012). Use of the Illumina Bovine3K Bead Chip in dairy genomic evaluation. *Journal of Dairy Science*, 95(3):1552 – 1558.

Zhan B, Fadista J, Thomson B, Hedegaard J, Panitz E, Bendixen C (2011). Global assessment of genomic variation in cattle by genome resequencing and high-through put genotyping. *BMC Genomics*, 12:557. doi: 10.1186/1471-2164-12-557.

Zimin AV, Kelley DR, Roberts M, Marcais G, Salzberg SL, Yorke IA (2012). Mis-Assembled «Segmental Duplications» in Two Versions of the Bos Taurus Genome. *PLoS ONE*, 7(8): e42680. doi: 10.1371/ journal.pone.0042680.

Zybailov RL, Sherpa MD, Glazko GV, Raney KD, Glazko VI (2013). G4-quadruplexes and genome instability. *Molecular Biology*, 47(2):197-204.