PROCEEDINGS OF THE LATVIAN ACADEMY OF SCIENCES. Section B, Vol. 62 (2008), No. 3 (656), pp. 91–102
DOI: 10.2478/v10046-008-0001-6

BIOINFORMATIC ANALYSIS OF EVOLUTIONAL CONSERVATISM AND FUNCTIONAL SIGNIFICANCE OF MICROSATELLITE ALLELES OF HUMAN 14Q13.2 REGION ASSOCIATED WITH TYPE 2 DIABETES MELLITUS

Tatjana Sjakste*, Ilva Poudžiunas*,**, Valdis Pīrāgs**, Māris Lazdiņš***, and Nikolajs Sjakste**

* Institute of Biology, University of Latvia, Miera iela 3, Salaspils, LV-2169, LATVIA; E-mail: tanja@email.lubi.edu.lv

** Faculty of Medicine, University of Latvia, Šarlotes iela 1a, Rīga, LV-1001, LATVIA

*** Faculty of Biology, University of Latvia, Kronvalda bulvāris 4, Rīga, LV-1586, LATVIA

Contributed by Nikolajs Sjakste

The paper deals with bioinformatic and statistical analysis of the possible functional significance of the previously shown association of several microsatellite alleles in intron 6 of the human proteasome core particle PSMA6 *gene (HSMS006) and four other microsatellites localised upstream in human chromosome 14q13.2 (HSMS801, HSMS702, HSMS701, HSMS602) with type 2* diabetes mellitus *in Latvia and Botnia, Finland. Genotype analysis revealed that (CAA)8/(CAA)8 homozygotes of the HSMS602 marker were never found in Type 2 diabetes patients, although 6.56% of the individuals from the control groups were the (CAA)8/(CAA)8 homozygotes. For the HSMS801 marker the (AC)21/(AC)23 genotype was never found in the case group and in the control group it was detected with a frequency 4.40%; these differences were statistically significant (P < 0.05). In contrast to the Latvian population, the distribution of genotype frequencies in cases and controls taken from the Botnian dataset was almost similar. Haplotype analysis showed that in the Latvian population besides haplotypes including alleles differently represented in case and control groups, a combination of some alleles almost equally represented in both groups formed combinations that were more characteristic of either the case group or the control group. This indicates probable independent functional significance of these haplotypes that warrants further investigation. In the Botnian population, more allele combinations were observed, and the distribution of haplotypes in case and control groups differed from that observed in Latvia. The observed haplotype distributions might reflect differences between the studied populations: a homogenous and isolated Botnian vis-à-vis a mixed Latvian population. Linkage disequilibrium (LD) analysis of data on the Latvian population revealed nine of ten two-allele combinations manifesting a high LD. HSMS006 and HSMS602 combination had a low LD; among the analysed markers these were situated at the largest distance from one another. Data on the Botnian population showed that haplotypes in eight of ten combinations had a high LD, including the HSMS006 and HSMS602 combinations. It appears that the two populations differ also in linkage disequilibrium of two-loci haplotypes. Theoretical analysis of a potential functional role of the polymorphisms indicated the significance of the microsatellite length of HSMS602 and HSMS006 for the formation of DNA hairpins. The whole genomic region appears to be conservative in mammals.*

Key words: *type 2* diabetes mellitus*, microsatellites, Chromosome 14, proteasomes,* PSMA6*, polymorphism, evolution, bioinformatics.*

INTRODUCTION

Microsatellites are short di- or trinucleotide repeats interspersed throughout genomes and manifesting high levels of polymorphism in different organisms. For many years, microsatellites have been intensively used as molecular markers. Gradually the interest of researchers in microsatellites decreased as their attention switched to the SNPs. The current avalanche of polymorphism research data, however, makes it essential to re-evaluate the earlier investigations. Most of the results of the association studies on

multifactorial diseases gave negative results or were not reproduced in other populations. We suggest that analysis of numerous polymorphisms on thousands of specimens should be replaced by a more thorough evaluation of the functional significance of any given prospective polymorphism. We start with microsatellites, since it has been shown that these are able to modify the level of gene expression by silencing/enhancing transcription and modulating splicing events (Gebhard *et al*., 1999; Gabellini, 2001). Some authors speculate that microsatellites may be involved in the etiology and pathogenesis of multifactorial diseases (Epplen *et al*., 1997).

For this study we chose a human 14q13 locus, which contains several (TG)n and (AC)n type microsatellites, as well as one (CAA)n repeat. The region contains the proteasomal *PSMA6* gene (14q13.2), as part of a cluster of proteasomal genes of Chromosome 14. Besides the above-mentioned gene, the cluster contains *PSMB5, PSME1, PSME2* (all in 14q11.2), *PSMC6* (14q22) and *PSMA3* (14q23) genes. We have described recently that several microsatellite markers and SNPs in the region are associated with type 2 *diabetes mellitus* (Sjakste *et al*., 2007a; 2007b).

The proteasomal PSMA6 protein, encoded in the genomic site under study, has been known for a long time to be one of the most evolutionary conservative representatives of the alpha-family, a monoclonal antibody raised against this human protein cross-reacts with its homologues in many mammals, birds, amphibia and even plants (Grossi de Sa *et al.,* 1988). mRNAs of the gene in mammals are very similar in structure (Coux *et al.,* 1994). Abundant data on the structure of mammalian genomes that have become recently accessible encouraged us to determine whether evolutionary conservatism of the structural gene is supported by evolutional stability of the genomic region.

The goals of the study were: to evaluate the evolutional conservatism of the 14q13 region with special reference to the microsatellite sequences; to analyse genotype and haplotype distribution of the microsatellite alleles in groups of type 2 *diabetes mellitus* and healthy subjects from two European populations using previously published results (Sjakste *et al*., 2007a) with the aim to evaluate the potential functional significance of the microsatellite sequences by analysing the possible transcription factor binding sites and to investigate changes in the secondary structure of microsatellite sites coupled to changes in repeat numbers.

MATERIALS AND METHODS

The subjects and genotyping were described previously (Sjakste *et al*., 2007a). Briefly the polymorphisms were genotyped in 104 type 2 diabetes patients and 129 control subjects from Latvia and 99 type 2 diabetes patients and 88 control subjects of Finnish origin (Botnia). Age-matched persons without type 2 *diabetes mellitus* diagnosis formed the control groups.

Sequence information on the *Homo sapiens* 14q13.2 genome region (NT_026437) and homologous regions from *Pan troglodytes*14q (NW_115873), *Canis familiaris* chromosome 8 (NW_876327), and *Mus musculus* chromosome 12 (NT_039551) genomes were obtained from the NCBI data base (www.ncbi.nih.gov).

Bioinformatic analysis. A search for homologous genes was performed using the Homologene Programme (www.ncbi.nlm.nih.gov/HomoloGene); sequence alignments were generated by BLAST 2 (www.ncbi.nlm.nih.gov/blast/bl2seq) or the multiple alignment service ClustalW (http://clustalw.genome.jp/). Results on positional synteny of microsatellites and their flanking sequences in all genomes analysed were produced by BLAST search of the particular human marker followed by BLAST 2 alignment of both full marker sequence and primer sequences with corresponding homologene sequences in *Pan troglodytes*, *Canis familiaris*, and *Mus musculus*. Identification of transcription factors binding sites (TFBSs) was performed using Genomatix software (DiAlign TF, Release 3.1, and MatInspector, Release 7.4 tools, at http://www.genomatix.de/). Prediction of possible changes in DNA secondary structure depending on the microsatellite repeat number was carried out with the aid of the MFOLD programme (www.bioinfo.rpi.edu/~zukerm).

Data analysis. The significance of differences between the groups was estimated using the χ^2 test ($Pa < 0.05$). Corrected significance levels (*Pb*) were calculated by means of permutation analysis using PAST (PAlaentological STatistics, ver. 1.63) programme. Linkage disequilibrium was calculated using the Arlequin version 3.01 programme (http://cmpg.unibe.ch/software/arlequin3/; Excoffier, 2005). The programmes were applied as previously described (Rana *et al*., 2004; Tsunoda *et al*., 2004).

RESULTS

Synteny conservation of 14q13.2. The upper panel of Figure 1 presents the structure of the genomic site around the human *PSMA6* gene and the localisation of the microsatellite markers used in the study inside introns of three other genes. The chromosome 14q13.2 area analysed spans over 270 kb. The HSMS006 marker ((TG)n motif) is located in the intron 6 of the proteasomal *PSMA6* gene. Four other microsatellite sequences are upstream PSMA6. The markers HSMS701 ((AC)mAT(AC)n motif), HSMS702 ((TG)n), and HSMS801 (AC)n), are localised in introns 3 and 4 of the gene *KIAA0391*, correspondingly, the most upstream HSMS602 ((CAA)n) marker is localized in the first intron of the *C14orf24* gene. The functions of the two latter genes remain unknown, although their mRNAs have been identified (references in Sjakste *et al*., 2004). The NFkappaB gene is localised downstream of the *PSMA6* gene. Figure 1 compares the chromosomal position, exon/intron organisation of the corresponding homologenes in *Pan troglodytes*, *Canis familiaris*, *Mus musculus*, and the positional synteny of microsatellites and/or their flanking sequences. The

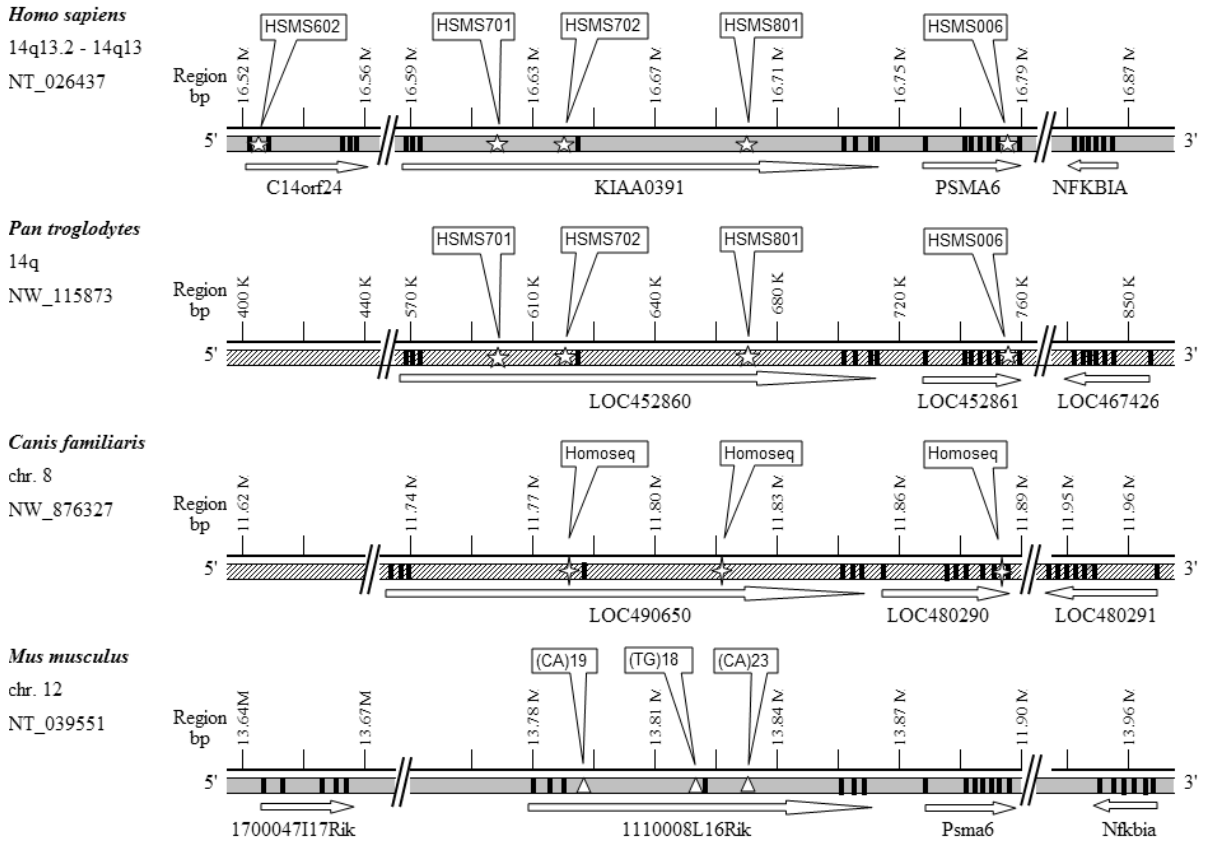

Fig. 1. Structure of the genomic site around the human *PSMA6* gene on Chromosome 14q13.2 and localisation of the microsatellite markers therein, compared to homologous sequences in chimpanzee, dog and mouse genomes. Exons of genes are indicated in black, introns of completely sequenced genomes are presented as grey background, and oblique shading indicates partly sequenced genomes. Human microsatellites and their analogues are indicated by five-ray stars, sequences homologous to regions flanking human microsatellites are indicated by stars. Triangles indicate positions of microsatellites in the mouse genome. Arrows indicate direction of transcription of genes. Homologous sequences are marked with symbol "Homoseq".

synteny of the genes in different mammals is clearly evident. Comparison of the human 14q13.2 sequence with the homologous genomic region of the chimpanzee revealed complete identity of the genomic domains of *KIAA0391*, *PSMA6* and *NFkappaB* genes. Microsatellites corresponding to human HSMS006, HSMS801, HSMS702 and HSMS701 markers are also found in the ape (Fig. 1). Homology cannot be followed upstream in the *KIAA0391* gene due to unfinished sequencing of the site. In dogs, the overall structure of the domain also appears to be similar. Intronic sequences with high similarity to flanking sites of human microsatellites are shown by comparing of human and canine sequences (Fig. 2). Three interspersed TG dinucleotides are observed in the area corresponding to the human HSMS006 site. A similar, but much shorter AC repeat, corresponds to the human HSMS801 microsatellite. An AT-rich sequence corresponds to the human TG-type microsatellite HSMS702. In mouse, flanking regions of human microsatellites were not identified, however; analogues of human TG-type and AC-type microsatellites were found in the domain of the *11100008L16Rik* gene, an analogue of the human *KIAA0391* gene (Fig. 1, lower panel).

Allele analysis. Figure 3 summarises previously published data on the allele frequency distribution in type 2 diabetic and control subjects from Latvia and Botnia. Allele (TG)22 of *HSMS006* is approximately twice as frequent in type 2 diabetic patients as compared to control subjects. Data were reproduced in two populations. Thus, the *HSMS006* (TG)22 allele appeared to be associated with type 2 *diabetes mellitus*.

One of the *HSMS602* alleles ((CAA)8) was less frequent in the case study group compared to controls in Latvia, however, data were not reproduced on Botnian subjects. The (AC)24 allele of HSMS 801 marker was less frequent in the Type 2 diabetes group, but this difference was also observed only in the Latvian population.

Genotypes. In the Latvian population the (CAA)8/(CAA)8 homozygotes of HSMS602 marker were never found in the Type 2 diabetes patients, although 6.56% of the individuals from the control groups were the (CAA)8/(CAA)8 homozygotes. For the HSMS801 marker the (AC)21/(AC)23 genotype was never found in the case group, in the control group it was detected with a frequency 4.4%; these differences were statistically significant ($P < 0.05$). In contrast to the Latvian population, the distribution of genotype frequencies in cases and controls taken from the Botnian dataset was almost similar.

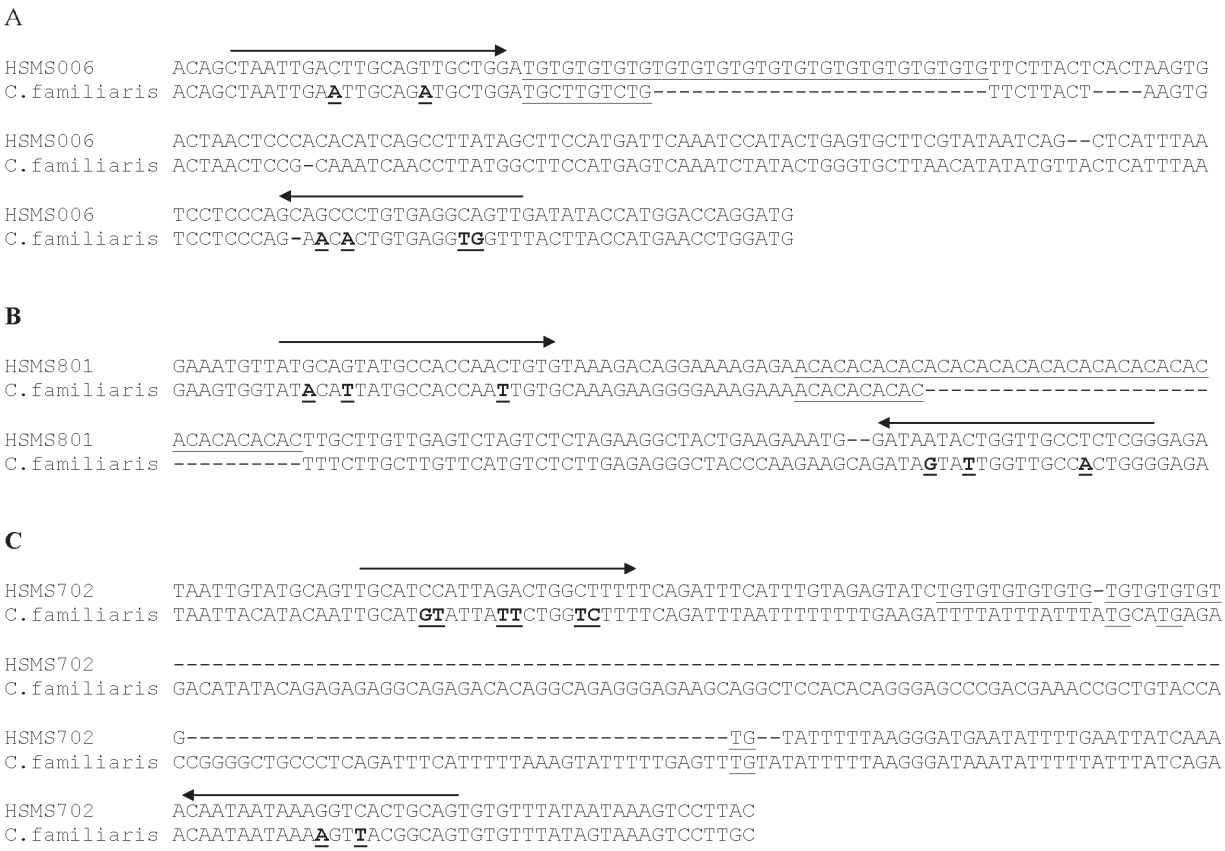

Fig. 2. Alignment of DNA sequences around HSMS006 (A), HSMS801 (B) and HSMS702 (C) markers with homologous canine genome sites. Sequences of microsatellites are underlined. Arrows indicate positions of primers used for amplification of human microsatellite markers. Substitutions constituting differences between human and canine genes inside primer sequences are given in bold and underlined.

Haplotypes. On the basis of allele frequencies of HSMS markers in both case and control groups of the Latvian population, we calculated all of the possible two loci haplotypes and their expected frequencies. We observed a lesser number of haplotypes than would be predicted by random combinations. Haplotypes observed less than four times were excluded from statistical analysis. Data on combination frequencies in case and control groups for the Latvian population are presented in Figure 4; haplotype frequencies not differing significantly in the two groups were omitted. It is interesting that, besides haplotypes including alleles differently represented in case and control groups (A6 or E2), a combination of some alleles almost equally represented in both groups forms combinations that are more characteristic of either the case group (B8/C2, B4/D1C2/E1, D1/E1, for example) or the control group (B7/C3, B7/C5). This suggests independent functional significance of these haplotypes, which warrants further investigation.

In the Botnian population there were more numerous allele combinations (Fig. 5). However, several two-marker combinations were equally distributed between the case and control groups. Combination frequencies that are differently distributed in the case and control groups are shown in Fig. 4. Surprisingly, the distributions differ in the Latvian population. Differences between the groups are represented mostly by other allele combinations. The A6/D1, A6/D6, A6/E1 and A6/E3 are the only haplotypes significantly differing in distribution between case and control groups both in the Latvian and Botnian datasets, however, these differences are opposite. All of these haplotypes are more frequent in Latvian type 2 diabetics and in Botnian controls. The observed haplotype distributions might reflect significant differences between the studied populations: a homogenous and isolated Botnian vis-à-vis a mixed Latvian population.

We calculated the linkage disequilibrium (LD) of the two loci haplotypes. Two microsatellite combinations with sufficiently high LD (LD 0.9) were selected from the pooled case and control groups. Then we calculated LD of the chosen haplotypes in the case and control groups separately. Analysis of data on the Latvian population revealed that nine of ten two allele combinations showed a high LD. The HSMS006 and HSMS602 combination had a low LD; among the analysed markers these were situated at the largest distance from one another. Among the nine other combinations, 28 haplotypes had a high LD, indicating that alleles in these haplotypes are inherited together. In the type 2 diabetes group, 25 of these haplotypes were identified, only 19 of them were probable for the control group. The highest χ^2 and r^2 values were found for the B2/D12 haplotype of the HSMS801/HSMS701 combination. C5/D3 combinations appear to have the strongest linkage (Fig. 6). Data on the Botnian population showed that haplotypes in eight of ten combination had a high LD, including the HSMS006 and

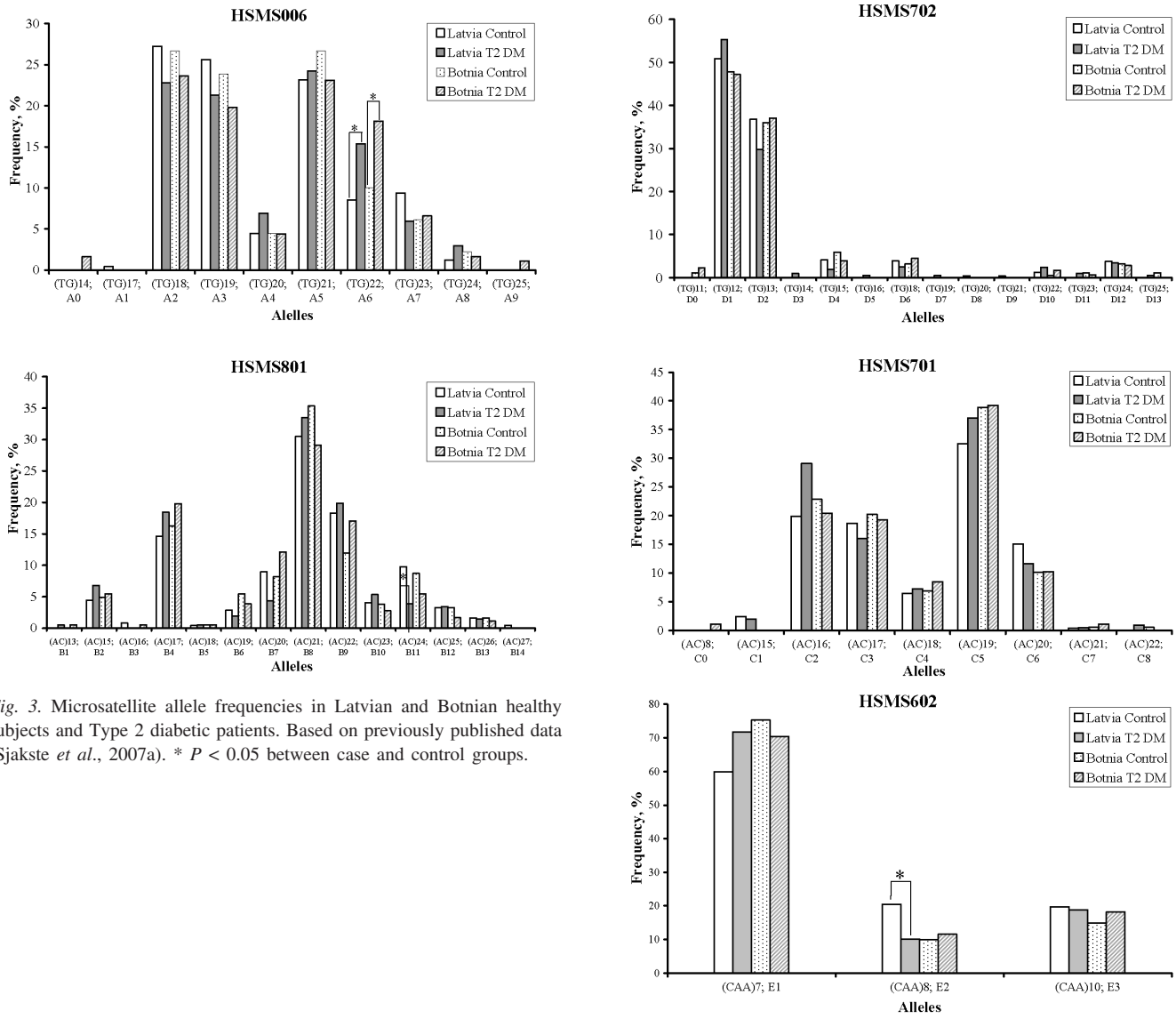

Fig. 3. Microsatellite allele frequencies in Latvian and Botnian healthy subjects and Type 2 diabetic patients. Based on previously published data (Sjakste *et al*., 2007a). * $P < 0.05$ between case and control groups.

HSMS602 combinations. Twenty-five such haplotypes were identified in total. Nineteen haplotypes with a high LD score were characteristic of the case group and 16 of the control group. Combinations A3/C0, A2/D13, B2/D13, B9/D2 and C5/D10 had much higher association than other combinations (Fig. 7). Thus, the two populations differ also in linkage disequilibrium of two-loci haplotypes.

SNP/microsatellite haplotypes. Our previous data indicated that one of the genotypes formed by SNPs in the *PSMA6* gene (-8CG) was significantly more frequent in type 2 diabetes patients, and haplotype C^{-110}/G^{-8}, compared to C^{-110}/C^{-8}, in association with the disease risk (Sjakste *et al.,* 2007b). As part of samples analysed in the above study was used also for microsatellite analysis in the Latvian population. We performed analysis of two- and three-loci haplotypes formed by the SNPs and HSMS006 microsatellite alleles. Interestingly, several combinations were more often found in the case group (Tables 1 and 2). We have also calculated the potential LD relationship between two types of polymorphisms. However, no statistically significant associations were revealed for any combination (not shown).

Transcription factor binding sites. Data on possible binding sites of transcription factors around the sites of the studied microsatellites are presented in Figure 8. Microsatellite sequences do not contain any transcription factor binding sites; in some cases a microsatellite extremity forms a transcription site together with the flanking sequence (NK1 binding site on 3' extremity of HSMS801, MMEF binding site in HSMS702 and the downstream adjacent sequence). Sequences upstream from several microsatellites contain multiple binding sites for numerous transcription factors. Interestingly, the CMYB binding site is found in the 5' flanking regions of HSMS006, HSMS801 and HSMS702 markers; the NFAT binding site exists in the vicinity of both HSMS801 and HSMS701 markers.

Potential DNA secondary structures. Figure 9 presents models of formation of secondary structure in the regions of two of the studied microsatellites, depending on the number of repeats therein. In the HSMS602 marker, an increase of

Table 1

DISTRIBUTION OF TWO-LOCI HAPLOTYPES FORMED BY SNPS IN THE *PSMA6* GENE AND ALLELES OF HSMS006 MICROSATELLITE MARKER IN CASE AND CONTROL GROUPS FROM LATVIA

Two-loci haplotypes -110A/HSMS006

Collection	Samples containing microsatellite HSMS006	Samples containing HSMS006 and SNP -110/-8	Haplotypes		Haplotypes -110A/HSMS006 (Haplotypes (samples; x – homozygote by MS)							
			Total for HSMS006 and SNP -110	SNP -110 A	TG17	TG18*	TG19	TG20	TG21	TG22	TG23	TG24
Control	123	103	412	13 (11 samples)	0 (0; 0)	2 (1; 1)	4 (3; 0)	1 (1; 0)	4 (4; 0)	6 (5; 0)	7 (5; 1)	2 (1; 0)
T2 DM	101	94	376	13 (13 samples)	0 (0; 0)	9 (6; 3)	5 (4; 1)	1 (1; 0)	6 (4; 2)	2 (2; 0)	2 (2; 0)	1 (1; 0)
P*a*					–	0.024	0.639	0.948	0.439	0.200	0.127	0.618
P*b*					–	0.031	0.743	1	0.534	0.288	0.095	1

Two-loci haplotypes -8G/HSMS006

Collection	Samples containing microsatellite HSMS006	Samples containing HSMS006 and SNP -110/-8	Haplotypes		Haplotypes -8G/HSMS006 (Haplotypes (samples; x – homozygote by MS)							
			T for HSMS006 and SNP -8	SNP -8 G	TG17	TG18	TG19	TG20	TG21	TG22*	TG23	TG24
Control	123	103	412	17 (17 samples)	0 (0; 0)	9 (7; 2)	9 (9; 0)	2 (2; 0)	6 (5; 1)	2 (2; 0)	5 (5; 0)	1 (1; 0)
T2 DM	101	94	376	27 (27 samples)	0 (0; 0)	8 (8; 0)	12 (11; 1)	5 (5; 0)	12 (10; 2)	8 (8; 0)	8 (7; 1)	1 (1; 0)
P*a*					–	0.957	0.393	0.211	0.112	0.042	0.322	0.948
P*b*					–	1	0.505	0.122	0.153	0.020	0.273	1

the CAA repeat number up to ten repeats triggers changes in DNA secondary structure different from that of (CAA)7 and (CAA)8 alleles. In the HSMS006 microsatellite, an increase of TG repeat number from 17 to 18 is followed by a change in the hairpin structure.

DISCUSSION

PSMA6 protein is known to be one of the most evolutionary conservative representatives of the alpha-family of proteasomal proteins (Grossi de Sa *et al.,* 1988; Coux *et al.,* 1994). Our data indicate that conservatism is characteristic, not only of the coding region of the gene, but also of the organisation of the surrounding genomic area, including the repeated sequences. Our observation does not contradict the data of other workers, since it is known that the dinucleotide repetitive sequences (TG/CA)n are widespread and conserved in many loci of higher eukaryotes and on average occur every 30 kb in human, and 18 and 21 kb in mouse and rat genomes, correspondingly. The positions of these repeats have been conserved between closely related species, such as humans and other primates. To a lesser extent, positions of GT repetitive sequences have been conserved between species in such distantly related groups as primates and rodents (Stallings *et al.,* 1991).

Our results indicate that a TG repeat polymorphism in the intron 6 of the human proteasome subunit *PSMA* gene is associated with type 2 diabetes in two different populations. The (TG)22 allele is two times more frequent in patients as compared to controls. The very description of (TG)n microsatellites longer than 22 repeats in an intron of *PSMA6* gene is novel, as it was previously considered that proteasomal genes contain only shorter repeats (Sharma *et al.,* 2005a). Long (TG)n repeats form Z-DNA structures and this could explain the physiological significance of these alleles (Sharma *et al.,* 2005b). Moreover, growing evidence has been accumulated over several years that point to multiple functional roles of microsatellites in various biological processes. Intronic sequences seem to be able to modify the level of gene expression by silencing or enhancing the transcription and splicing events (Horikawa *et al.,* 2000). It was observed that (TG/CA)n microsatellites have a propensity to undergo structural transitions in *in vivo* conditions and could modulate gene expression in several genes of different organisms (see for reference Sharma *et al.,* 2005a) including human housekeeping genes (Sharma *et al.,* 2003).

Table 2

DISTRIBUTION OF THREE-LOCI HAPLOTYPES FORMED BY SNPS IN THE *PSMA6* GENE AND ALLELES OF HSMS006 MICROSATELLITE MARKER IN CASE AND CONTROL GROUPS FROM LATVIA

Three-loci haplotypes - 110C/-8G/HSMS006

Collection	Samples for microsatellite HSMS006	Samples for HSMS006 and SNP -110/-8	Haplotypes		Haplotypes -110C/-8G/HSMS006							
			All for HSMS006 and SNP -110/-8	SNP -110C/-8G	TG17	TG18	TG19	TG20	TG21*	TG22*	TG23	TG24
Control	123	103	824	30 (15 samples)	0	18	16	4	12	2	8	0
T2 DM	101	94	752	54 (27 samples)	0	16	24	10	24	16	16	2
P*a*					–	0.939	0.124	0.077	0.024	0.000508	0.065	0.139
P*b*					–	1	0.147	0.057	0.026	0.0008	0.065	0.231

Three-loci haplotypes - 110A/-8C/HSMS006

Collection	Samples for microsatellite HSMS006	Samples for HSMS006 and SNP -110/-8	Haplotypes		Haplotypes -110A/-8C/HSMS006							
			All for HSMS006 and SNP -110/-8	SNP -110A/-8C	TG17	TG18*	TG19	TG20	TG21	TG22	TG23*	TG24
Control	123	103	824	22 (11 samples)	0	4	8	2	8	12	14	4
T2 DM	101	94	752	26 (13 samples)	0	18	10	2	12	4	4	2
P*a*					–	0.00147	0.507	0.927	0.274	0.070	0.0313	0.481
P*b*					–	0.0009	0.644	1	0.359	0.081	0.0153	0.688

Furthermore, the (TG/CA)n repeats have been observed to act as stimulators in recombination (Majewski and Ott, 2000; Gabellini, 2001) and as a widespread regulatory element of alternative splicing (Hui *et al,* 2003; 2005). Change of the repeat number *per se* alters the local spatial relationships of transcription factor interactions and nucleosome formation pattern (Kashi and King, 2006). Our prediction of changes in the hairpin formation depending on repeat number in microsatellites HSMS006 and HSMS602 also indicates possible alterations in the framework of transcription factors binding the area. Interactions of intermediate filaments with microsatellites may also depend on the repeat number of the latter resulting in alterations of the cell functions (Tolstonog *et al.,* 2001; 2005); changes in secondary structure might be also crucial for these interactions. Thus, prediction of possible hairpins in intronic sequences of *PSMA6* genes indicates the possibility for formation of unpaired regions in the gene. These regions give rise to the so-called S1-hypersensitive sites with altered base pairing. Such sites have been found both in structural (Larsen and Weintraub, 1982) and regulatory (Recillas Targa *et al*., 1994) sites in the chicken alpha-globin domain (reviewed in Sjakste and Sjakste, 2002). The regulatory role of possible hairpins might be also related to formation of the matrix attachment regions (MARs regions within MARs, that possess an intrinsic propensity to unwind under negative superhelical strain (double stranded base unpairing regions (BURs)). Several chromatin-associated proteins specifically recognize BURs (poly (ADP-ribose) polymerase (PARP-1), Ku autoantigen, SAF-A, HMG-I(Y), nucleolin and p53). Interaction between BUR-binding proteins may not only provide an architectural core but also recruit functional multi-molecular complexes at the base of chromatin loops to affect multiple distant genes (Galande, 2002). A specific role of partially unwound DNA sequences in anchoring of DNA-loops to the nuclear matrix and alterations of nuclear matrix attachment sites were revealed in several of our previous studies (see Sjakste and Sjakste, 2007a, 2007b) for review.

The functional significance of microsatellites is supported by the increasing number of reports on association of different pathologies with certain intronic microsatellite alleles, including the recent finding of an association of type 2 *diabetes mellitus* with polymorphism of the intronic microsatellite of the *TCF7L2* gene (Grant *et al.,* 2006).

Thus, we can hypothesise that the *PSMA6* gene expression is altered in patients with type 2 diabetes. Positioning of the transcription factor binding sites around the microsatellite and changes in the DNA secondary structure, coupled to an increase in the repeat number support this idea. This could influence the efficiency of proteasome-mediated degradation of some key transcription factors or enzymes involved in insulin signalling in muscle or other tissues. Accumula-

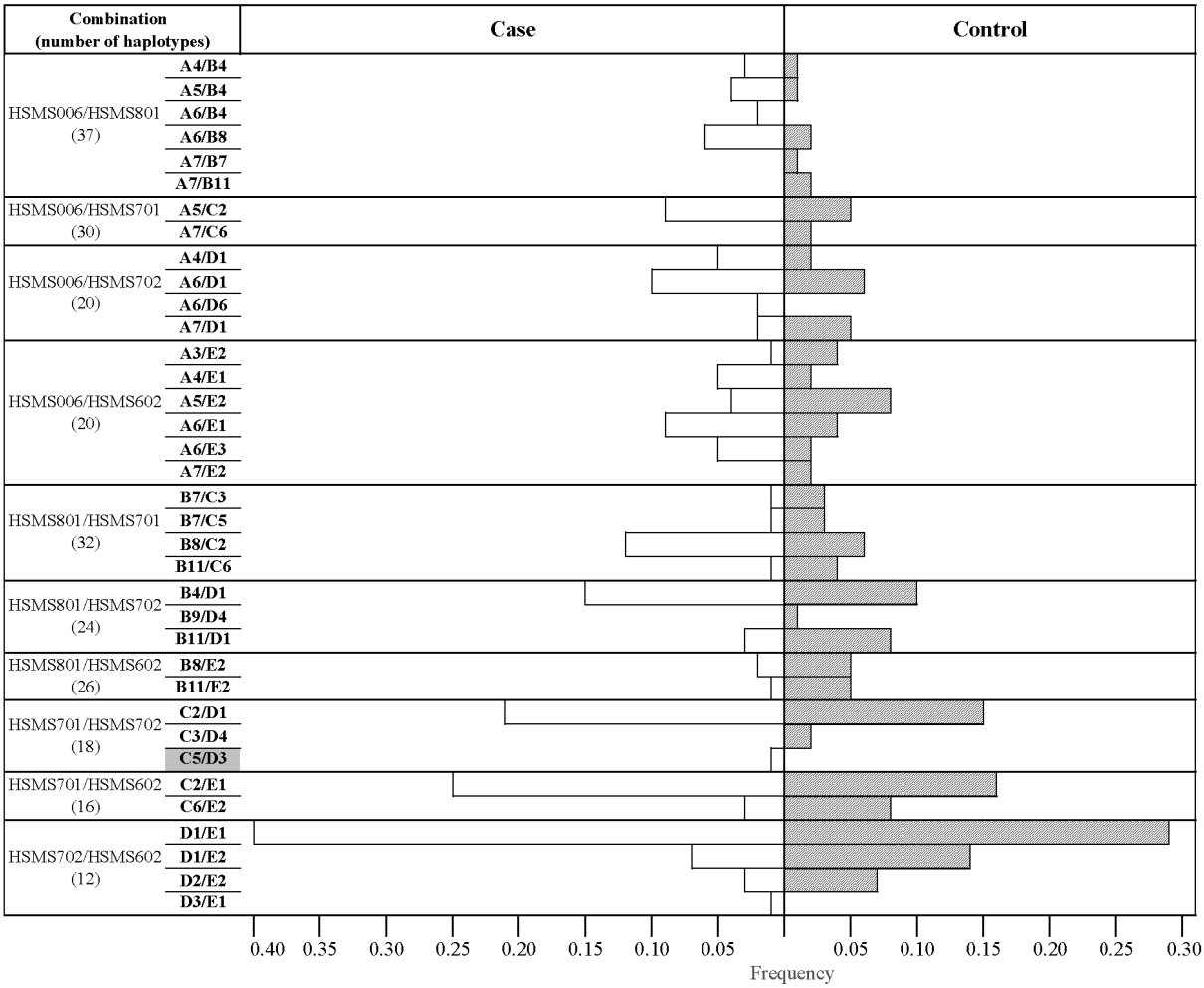

Fig. 4. Frequencies of microsatellite haplotypes manifesting statistically significant differences ($P < 0.05$) in case and control groups of the Latvian population. Designation of alleles is given in Figure 3. Designations of haplotypes with high LD are shadowed.

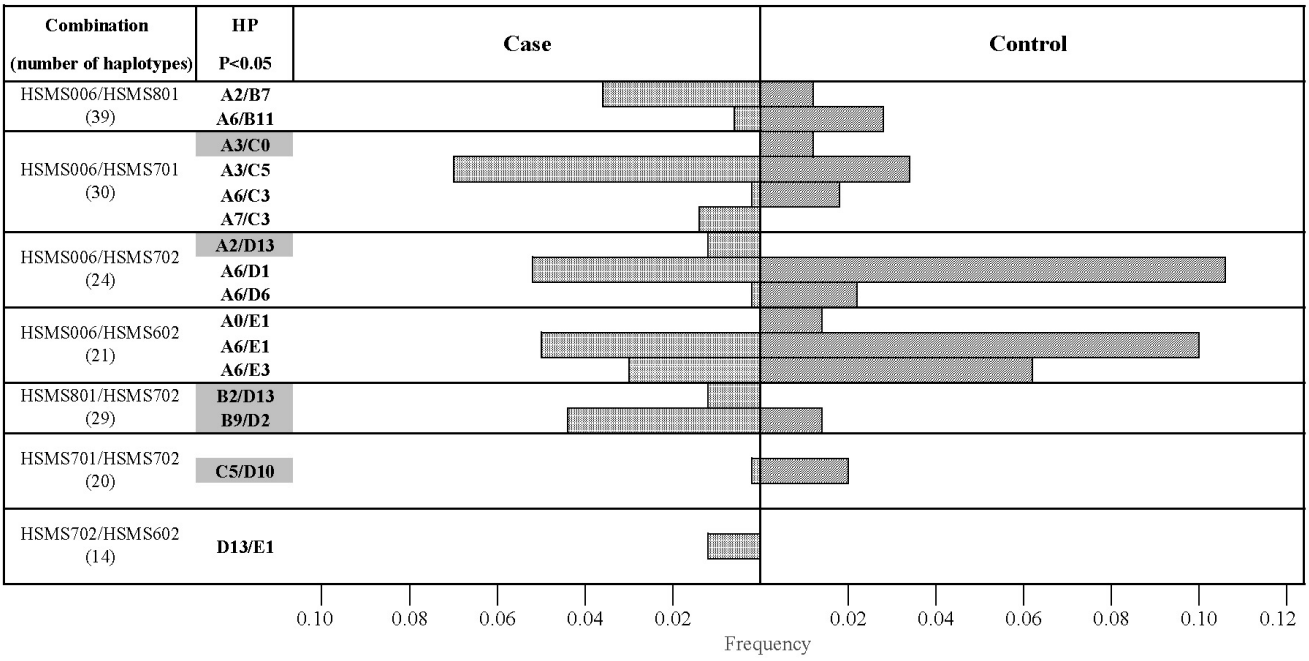

Fig. 5. Frequencies of microsatellite haplotypes manifesting statistically significant differences ($P < 0.05$) in case and control groups of the Botnian population. Designation of alleles is given in Figure 3. Designations of haplotypes with high LD are shadowed.

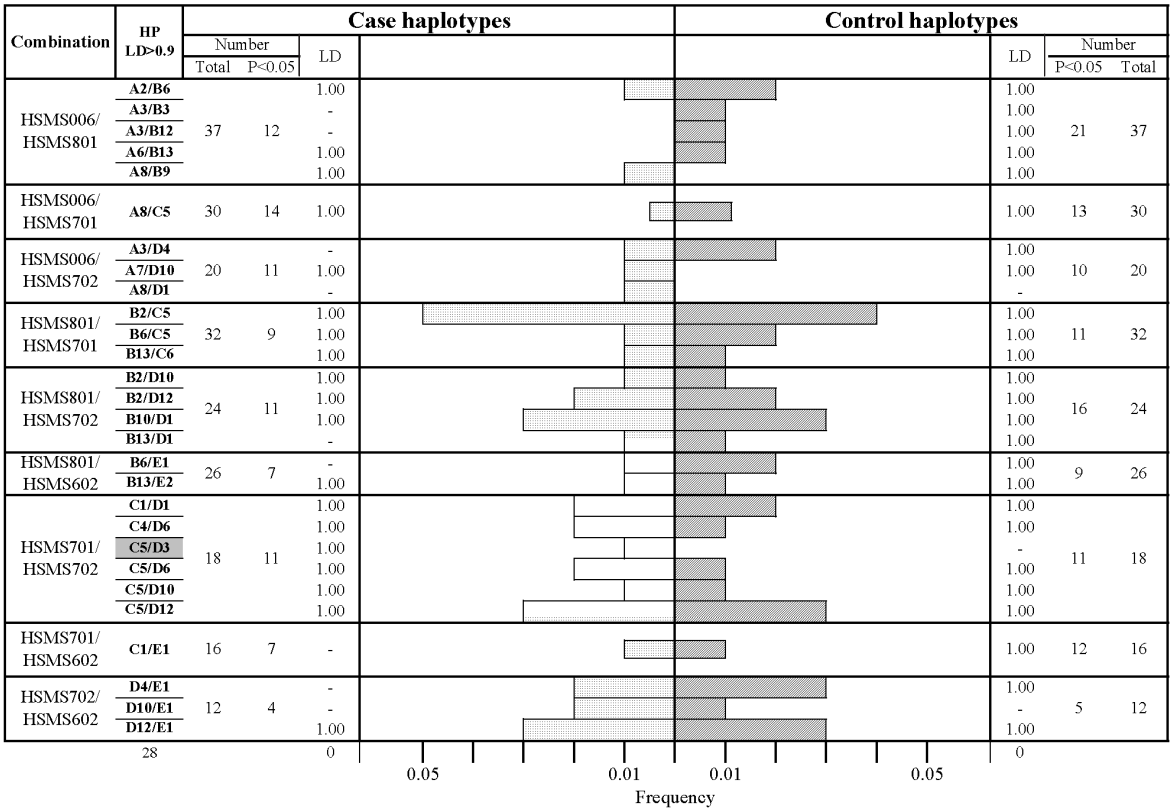

Fig 6. Linkage disequilibrium (LD) of the two loci haplotypes in case and control groups of the Latvian population. Combinations with sufficiently high linkage disequilibrium (LD > 0.9) are presented. Designations of haplotypes found with statistically different frequencies in case and control groups are shadowed.

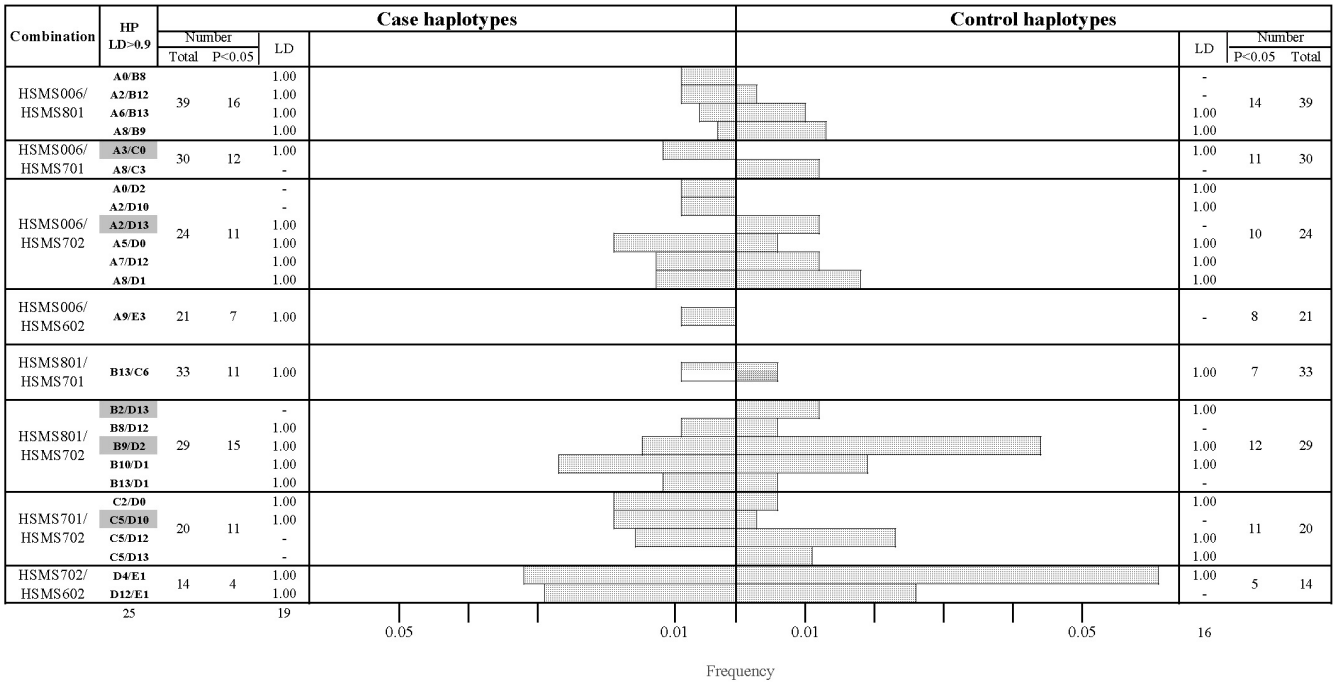

Fig. 7. Linkage disequilibrium (LD) of the two loci haplotypes in case and control groups of the Botnian population. Combinations with sufficiently high linkage disequilibrium (LD > 0.9) are presented. Designations of haplotypes found with statistically different frequencies in case and control groups are shadowed.

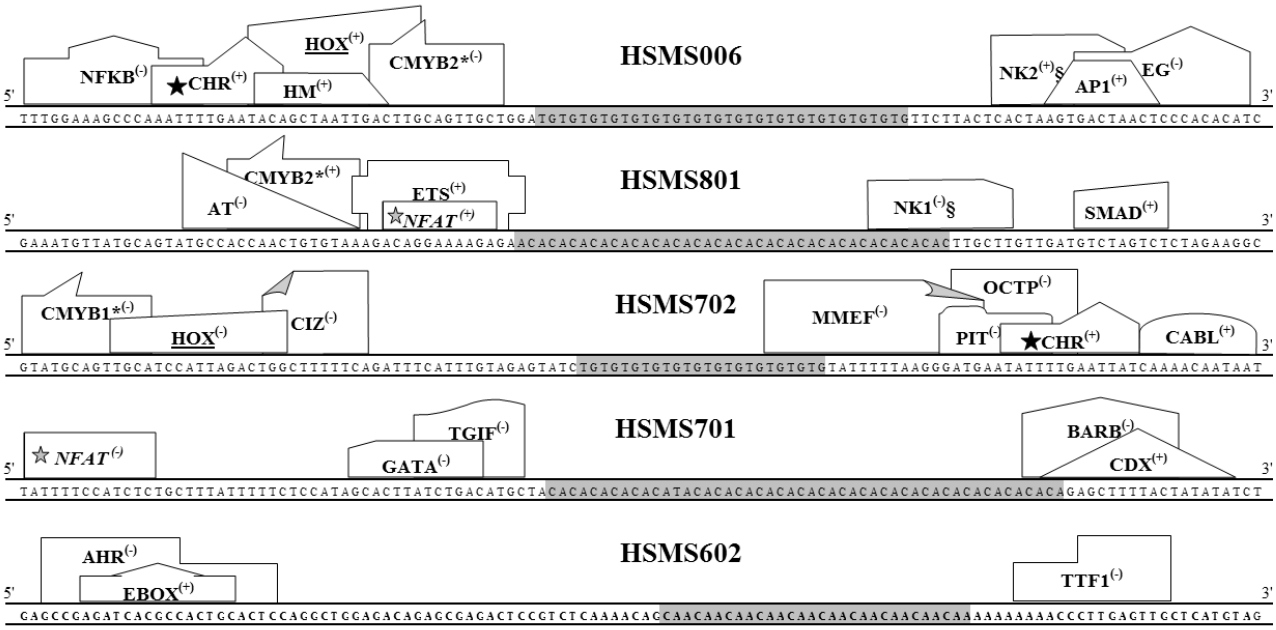

Fig. 8. Possible binding sites of transcription factors in the vicinity of the 14q13 microsatellite markers. Positions of sequences of transcription factor binding sites in NT_026437 are as follows: HSMS006: 16783851 – 16783975; HSMS801: 16699844 – 16699968; HSMS702 – 16649426 – 16649550; HSMS701 – 16620531 – 16620655; HSMS602: 16516649 – 16516773. Names of factors are given in boxes. (+) – binding to positive strand; (-) – binding to negative strand. Additional symbols (§, *, stars) indicate transcription factor binding sites found near two or more markers.

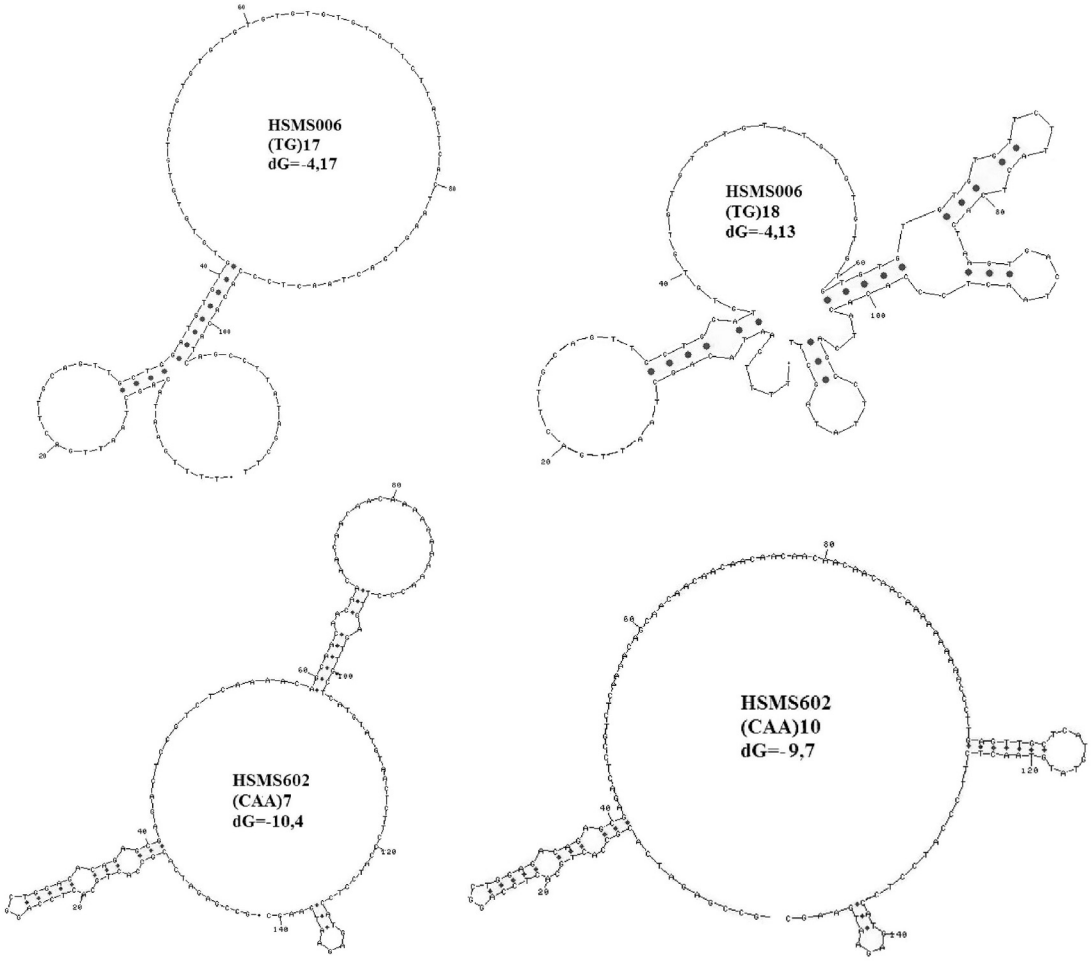

Fig. 9. Computer models of modifications in secondary structures triggered by increase of microsatellite repeat numbers.

tion of the non-degraded proteasome substrates could make the cells less sensitive to the action of insulin. It was reported that an SNP at the -8 position from the transcription starting point of the *PSMA6* gene, causing increase of its expression, is associated with risk of myocardial infarction in Japanese population (Ozaki *et al.,* 2006). This association was not found in European populations (Sjakste *et al*., 2007b). The association of some alleles of HSMS602 and HSMS801 with susceptibility or resistance to type 2 *diabetes mellitus* is difficult to explain, as the functions of the corresponding genes remain unknown. Although this should be taken only as speculation at the moment, the discovery of an association between microsatellite polymorphism in the *PSMA*6 gene area and type 2 *diabetes mellitus* makes the genes interesting candidates for type 2 diabetes susceptibility genes. Modification of transcription or the translation level of a single protein of the proteasomal particle can affect formation of the particle that consists of several subunits. Impaired assembly of proteasomes can affect regulation of the internalisation of the insulin receptor, of the control of the amount of insulin receptor substrates 1 and 2, and of insulin degradation (Rome *et al*., 2004). We hope that the nature of the association of 14q13.2 microsatellite polymorphism with type 2 *diabetes mellitus* described above will be explained on the basis of experimental studies in the very near future.

ACKNOWLEDGEMENTS

This work was supported by grants from the Latvian Council of Science and European Foundation of Social Development (ESF). Support and encouragement of E. Grēns as founder of genomics research in Latvia is greatly acknowledged.

REFERENCES

Coux, O., Nothwang, H.G., Silva Pereira, I., Recillas Targa, F., Bey, F., Scherrer, K. (1994). Phylogenic relationships of the amino acid sequences of prosome (proteasome, MCP) subunits. *Mol. Gen. Genet.,* **245**, 769–780.

Epplen, C., Santos, E.J., Maueler, W., van Helden, P., Epplen, J.T. (1997). On simple repetitive DNA sequences and complex diseases. *Electrophoresis,* **18**, 1577–1585.

Excoffier, L., Laval, G., Schneider, S. (2005). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online,* **1**, 47–50.

Gabellini, N. (2001). A polymorphic GT repeat from the human cardiac $Na^{+}Ca^{2+}$ exchanger intron 2 activates splicing. *Eur. J. Biochem.,* **268**, 1076–1083.

Galande, S. (2002) Chromatic (dis)organization and cancer: BUR-binding proteins as biomarkers for cancer. *Curr.Cancer Drug Targets*, **2**, 157–190.

Gebhard, F., Zänker, K.S., Brandt, B. (1999). Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J. Biol. Chem.,* **274**, 13176–13180.

Grant, S.F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., Styrkarsdottir, U., Magnusson, K.P., Walters, G.B., Palsdottir, E., Jonsdottir, T., Gudmundsdottir, T., Gylfason, A., Saemundsdottir, J., Wilensky, R.L., Reilly, M.P., Rader, D.J., Bagger, Y., Christiansen, C., Gudnason, V., Sigurdsson, G., Thorsteinsdottir, U., Gulcher, J.R., Kong, A., Stefansson, K. (2006). Variant of transcription factor 7 – like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nat. Genet.,* **38**, 320–323.

Grossi de Sa, M.-F., Martins de Sa, C., Harper, F., Olink-Coux, M., Huesca, M., Scherrer K. (1988). The association of prosomes with some of the intermediate filament networks of the animal cell. *J. Cell Biol.,* **107**, 1517–1530.

Horikawa, Y., Oda, N., Cox, N.J. (2000). Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat. Genet.,* **26**, 135–137.

Hui. J., Hung, L.H., Heiner, M., Schreiner, S., Neumuller, N., Reither, G., Haas S.A., Bindereif, A. (2005). Intronic CA-repeat and CA-rich elements: A new class of regulators of mammalian alternative splicing. *EMBO* (European Molecular Biology Organisation) *J.,* **24**, 1988–1998.

Hui, J., Reither, G., Bindereif, A. (2003). Novel functional role of CA repeats and hnRNP L in RNA stability. *RNA*, **9**, 931–936.

Kashi, Y., King, D.G. (2006). Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.,* **22**, 253–259.

Larsen, A., Weintraub, H. (1982). An altered DNA conformation detected by S1 nuclease occurs at specific regions in active chick globin chromatin. *Cell,* **29**, 609–622.

Majewski, J., Ott, J. (2000). GT repeats are associated with recombination on human chromosome 22. *Gen. Res.,* **10**, 1108–1114.

Ozaki, K., Sato, H., Iida, A., Mizuno, H., Nakamura, T., Miyamoto, Y., Takahashi, A., Tsunoda, T., Ikegawa, S., Kamatani, N., Hori, M., Nakamura, Y., Tanaka, T. (2006). A functional SNP in *PSMA6* confers risk of myocardial infarction in the Japanese population. *Nat. Genet.*, **38**, 921–925.

Rana, N.A., Ebenezer, N.D., Webster A.R., Linares A.R., Whitehouse D.B., Povey S., Hardcastle A.J. (2004). Recombination hotspots and block structure of linkage disequilibrium in the human genome exemplified by detailed analysis of PGM1 on 1p31. *Hum. Mol. Genet*., **13**, 3089–3102.

Recillas-Targa, F., Razin, S. V., de Moura Gallo, C. V., Scherrer, K. (1994). Excision close to matrix attachment regions of the whole domain of the chicken alpha-globin gene domain by nuclease S1 and characterisation of the framing structures. *Proc. Natl. Acad. Sci. USA*, **91**, 4422–4426.

Rome, S., Meugnier, E., Vidal, H., (2004). The ubiquitin–proteasome pathway is a new partner for the control of insulin signaling. *Curr. Opin. Clin. Nutr. Metab. Care,* **7**, 249–254.

Sharma, V.K., B-Rao, C., Sharma, A., Brahmachari, S.K., Ramachandran, S. (2003). (TG:CA)(n) repeats in human housekeeping genes. *J. Biomol. Struct. Dyn.,* **21**, 303–310.

Sharma, V.K., Brahmachari, S.K., Ramachandar, S. (2005a) $(TG/CA)_n$ repeats in human gene families: Abundance and selective patters of distribution according to function and gene length. *BMC Genomics,* **6**, 83.

Sharma, S., Rajan, U.M., Kumar, A., Soni, A., Ghosh, B. (2005b). A novel (TG)n(GA)m repeat polymorphism 254 bp downstream of the mast cell chymase (CMA1) gene is associated with atopic asthma and total serum IgE levels. *J. Hum. Genet.,* **50**, 276–282.

Sjakste, N.I., Sjakste, T.G. (2002). Structure of globin gene domains in mammals and birds. *Russ. J. Genet.*, **38**(12), 1342–1358.

Sjakste, N.I., Sjakste, T.G. (2007a). Possible involvement of DNA breaks in epigenetic regulationof cell differentiation. *Russ. J. Genet.*, **43**(5), 467–484.

Sjakste, N., Sjakste, T. (2007b). Possible involvement of DNA strand breaks in regulation of cell differentiation**.** *Eur. J. Histochem*, **51**, 81–94.

Sjakste, T., Eglite, J., Sochnevs, A., Marga, M., Pirags, V., Collan, Y., Sjakste N. (2004). Microsatellite genotyping of chromosome 14q13.2–14q13 in the vicinity of proteasomal gene *PSMA6* and association with Graves' disease in the Latvian population. *Immunogenetics,* **56**, 238–243.

Sjakste, T., Kalis, M., Poudziuas, I., Pirags, V., Lazdins, M., Groop, L., Sjakste, N. (2007a). Association of microsatellite polymorphisms of the human 14q13.2 region with type 2 diabetes mellitus in Latvian and Finnish populations. *Ann. Human Genet.,* **71**, 772–776.

Sjakste, T., Poudziunas, I., Ninio, E., Perret, C., Pirags, V., Nicaud, V., Lazdins, M., Evans, A., Morrison, C., Cambien, F., Sjakste N. (2007b). SNPs of *PSMA6* gene—investigation of possible association with myocardial infarction and type 2 diabetes mellitus. *Russ. J. Genet.*, **43**, 442–448.

Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E., Moyzis, R.K. (1991). Evolution and distribution of (GT)n repetitive sequences in mammalian genomes. *Genomics,* **10**, 807–815.

Tolstonog, G.V., Mothes, E., Shoeman, R.L., Traub, P. (2001). Isolation of SDS-stable complexes of the intermediate filament protein vimentin with repetitive, mobile, nuclear matrix attachment region, and mitochondrial DNA sequence elements from cultured mouse and human fibroblasts. *DNA Cell. Biol.,* **20**, 531–554.

Tolstonog, G.V., Li, G., Shoeman, R.L., Traub, P. (2005). Interaction *in vitro* of type III intermediate filament proteins with higher order structures of single-stranded DNA, particularly with G-quadruplex DNA. *DNA Cell Biol.,* **24**, 85–110.

Tsunoda, T., Lathrop, G.M., Sekine, A., Yamada, R., Takahashi, A., Ohnishi, Y., Tanaka, T., Nakamura, Y. (2004). Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome. *Hum. Mol. Genetics*, **13**, 1623–1632.

Received 29 October 2007

AR 2. TIPA CUKURA DIABĒTU SAISTĪTA CILVĒKA 14Q13.2 REĢIONA MIKROSATELĪTU ALĒĻU EVOLUCIONĀRĀ KONSERVATĪVISMA UN FUNKCIONĀLĀS NOZĪMES BIOINFORMĀTISKĀ UN STATISTISKĀ ANALĪZE

Izmantojot bioinformātikas un statistikas metodes, izanalizēta iespējamā funkcionālā nozīme cilvēka proteasomu kodola proteīna *PSMA6* gēna 6. intronā lokalizētā mikrosatelīta (HSMS006) un citu augšup lokalizētu četru mikrosatelītu (HSMS801, HSMS702, HSMS701, HSMS602) polimorfismiem, kas ir asociēti ar 2. tipa cukura diabētu. Genotipu analīze parādīja, ka marķiera HSMS602 genotips (CAA)8/(CAA)8 netika novērots 2. tipa diabēta pacientiem, kaut gan kontroles grupā tas bija 6.56% indivīdiem. HSMS801 marķiera (AC)21/(AC)23 genotips arī netika atrasts slimnieku grupā, bet kontroles grupā tas tika atrasts ar biežumu 4.40%; šīs atšķirības bija statistiski nozīmīgas ($P < 0.05$). Pretstatā Latvijas populācijai, genotipu frekvenču sadalījums slimnieku un kontroles grupās Botnijas (Somija) populācijā bija gandrīz vienāds. Haplotipu analīze parādīja, ka Latvijas populācijā bez haplotipiem, kuros iekļautās alēles ar dažādu biežumu sastopamas slimnieku un kontroles grupās, dažu alēļu kombinācijas ir gandrīz vienādi pārstāvētas abās grupās. Tādējādi veidojās kombinācijas, kas raksturīgas tikai gadījuma vai tikai kontroles grupām. Tas liecina par neatkarīgu šo haplotipu funkcionālo nozīmi, un to būtu vērts pētīt nākotnē. Botnijas populācijā tika novērots lielāks alēļu kombināciju skaits; haplotipu izplatība slimnieku un kontroles grupā atšķiras no Latvijas populācijā konstatētās. Haplotipu sadales atšķirības, iespējams, atspoguļo atšķirību starp pētītajām populācijām: viendabīga un izolēta Botnijas populācija pret jaukto Latvijas populāciju. Latvijas populācijas datu nelīdzsvarotās saistības (LD) analīze pierādīja, ka deviņām no desmit divu alēļu kombinācijām ir augsts LD rādītājs. HSMS006 un HSMS602 kombinācijai tas ir zems; starp analizētajiem marķieriem šie divi mikrosatelīti atrodas vislielākajā attālumā viens no otra. Botnijas populācijas datu analīze parādīja, ka astoņiem no desmit kombināciju haplotipiem ir augsts LD rādītājs, ieskaitot arī HSMS006 un HSMS602 marķieru kombināciju. Atklājās, ka šīs divas populācijas atšķiras arī pēc divu lokusu haplotipu nelīdzsvarotās saistības. Iespējamās polimorfismu funkcionālās nozīmes teorētiskajā analīzē konstatējām, ka HSMS006 un HSMS602 mikrosatelītu garums ietekmē DNS sekundārās struktūras (matadatu) veidošanos. Salīdzinot dažādu zīdītāju genomus, izrādījās, ka analizētais reģions ir evolucionāri konservatīvs.