Structural analysis of group of possible DNA-target sites for the RAG1/2 proteins found in the mouse genome *in silico* and their identification in known types of repetitive elements

A. Yu. Gubsky[1] and V. G. Zinkovsky[2]

[1]Department of Biology, I.I. Mechnikov Odessa National University,
2 Dvoryanskaya str., Odessa 65026, Ukraine

[2]Department of Biotechnology and Molecular Biology, University of Opole,
Kominka 4, PL 45-035 Opole, Poland

gubsky2003@mail.ru

Using mathematical methods and specially developed program algorithms, we have found that the number of previously detected possible target sites of RAG1/2 proteins (cRSS) in the mouse genome is 5.4-fold higher than theoretically expected value. In 71 % of cases examined, cRSS are structural elements of 390 unique types of repetitive elements. We think that some types of repetitive elements can participate in spreading and accumulation of cRSS in the mouse genome. The structure of 5 % motives includes nucleotides typical for majority of recombination signal sequences of functional V, D and J segments of the mouse Ig, Tcr genes (fRSS). The existence of 25 % of such cRSS in the mouse genome may be considered as consequence of random nucleotide combinations. Most often the spacers of 12cRSS and 23cRSS have 58 – 67 % and 30 – 47 % of homology to similar structures of 12fRSS and 23fRSS, respectively.

Keywords: cRSS, V(D)J recombination, RAG1, RAG2.

Introduction. Normally V(D)J recombination system rearranges genes of immunoglobulins (Ig) and T-cell receptors (TCR) in B- and T-lymphocytes precursors [1]. In such cases proteins RAG1 and RAG2 (further referred to as proteins RAG1/2) in complex with proteins HMG1, HMG2, Ku70, Ku80, TdT, XRCC4 initiate double-stranded breaks on the borders of intragenic V, D and J segments by recognizing 28±1 bp and 39±1 bp recombination signal sequences (RSS) [2,3]. However, when V(D)J recombination system gets out of control the proteins RAG1/2 can initiate DNA breaks outside of Ig and TCR loci. As a result of such illegitimate V(D)J recombinase activity some human (HPRT, SCL, SIL, etc.) and mouse (Notch1, etc.) genes can be damaged by translocations or intragenic deletions [4 – 6]. In such cases, the target sites of RAG1/2 proteins may be cryptic recombination signal sequences (cRSS), which are significantly different from the RSS of Ig and TCR genes. These data as well as variability of RSS [7, 8] indicate that RAG1/2 can interact with large variety of target sequences.

At present cRSS are considered as factors that potentially can mediate instability of mammalian genome, if V(D)J recombination system gets out of control. Therefore, it is very important to quantify and to specify locations of hypothetically possible target sites of RAG1/2. By using endonuclease cleavage of plasmid extrachromosomal DNA substrates by RAG1/2 *in vivo*, it has been suggested for the first time that in mammalian genomes (6 billion bp) there are at least 10^7 cRSS (frequency is 1.7×10^{-3}) [9]. In turn, by using RIC scores, Cowell et al. identified 4746 12cRSS and 16439 23cRSS in some mouse and human cDNA and genomic DNA (total size of analyzed sequences was more than 10.5 Mb). They have proposed that in mammalian genome the possible target sites of RAG1/2 may be found with frequency of 5×10^{-4} [10].

Our DNA analysis of 21 mouse chromosomes (Build 35.1) *in silico* indicated that outside of *Ig, Tcr* loci there are 6724 cRSS with theoretically high recombination potential. As 12cRSS and 23cRSS (cRSS with 12 bp and 23 bp spacers, respectively) we considered 28-bp and 39-bp DNA regions of the mouse genome, where heptamer/nonamer sequences (435 unique types of sequences) fully corresponded to CACAGTG/ACAAAAACC structure or had only 14 – 15 common nucleotides [11]. In all cases, the first three nucleotides of heptamers were CAC. Nonamers always had adenine at the fifth, sixth and seventh positions. We found that 2887, 7 and 20 of these cRSS are located in 2373 protein-coding genes, 7 pseudogenes and 20 known duplications of *Ig* and *Tcr* gene fragments. Some cRSS were found in repetitive elements such as B1_Mus1, B1_Mus2, MLT1A1, L1_Mus3, etc. In the case of protein-coding genes, 97 % of found cRSS are located in introns. 12cRSS and 23cRSS (in accordance to the “12/23" rule) theoretically may mediate formation of deletions and inversions of the whole exons in 87 and 100 protein-coding genes (*Large, Rpl23, Trhde, Suz12, Ptprn2, Bach2*, etc.), respectively.

We suppose that cRSS located in the protein-coding genes may cause their damages, if V(D)J recombination system gets out of control. Therefore, it is very important to understand mechanisms of cRSS appearance in the mouse genome as well as to characterize their nucleotide composition. In this study, we have compared structures of all 6724 cRSS with RSS of functional V, D and J segments of mouse *Ig*, *Tcr* genes (fRSS). We have estimated a number of cRSS located in repetitive elements. Also by using developed mathematical and computational approaches, we have estimated a number of cRSS accumulated in the mouse genome as a possible consequence of evolutionary processes.

Materials and methods

The DNA sequence of 21 mouse chromosomes (Build 35.1) was taken from NCBI (ftp://ftp.ncbi.nih.gov/genomes/M_musculus/ARCHIVE/BUILD.35.1/) as 21 files of «gbk» type. To determine number of nucleotides A, T, G, C in DNA of each chromosome, we developed our own program algorithms.

Theoretically expected number (*M* value) of 28-bp and 39-bp regions in mouse DNA correspond to 12cRSS and 23cRSS with a unique type of heptamer/nonamer sequence (435 unique types of sequences) was determined by formula:

$$M \quad (N \quad 1 \qquad)P_A^{\;n_A} \quad P_T^{\;n_T} \quad P_G^{\;n_G} \quad P_C^{\;n_C}, \tag{1}$$

where N is the number of nucleotides in the DNA sequence of analyzed chromosome;   is total number of nucleotides in analyzed 12cRSS or 23cRSS; P_A, P_T, P_G, P_C is a portion of nucleotides A, T, G, C in DNA of the analyzed chromosome; N_A, N_T, N_G, N_C is total number of nucleotides A, T, G, C in the analyzed type of heptamer/nonamer sequence. To determine a theoretically expected number of analogous structures located in a complementary chain of DNA, we have used the same formula.

Alternatively, to estimate theoretically expected number of possible target sites of RAG1/2 in the mouse genome, we have performed a search of 28-bp and 39-bp analogues of cRSS (tRSS) in set of sequences with random nucleotide combinations (SRNC). All 5 used sets of SRNC consisted of 21 unique sequences. Each such a sequence had a number of nucleotides A, T, G, C identical to the number of nucleotides of the corresponding mouse chromosome. All these sequences were created by using a generator of random numbers. A total size of sequences in each set of SRNC corresponded to the size of mouse genomic DNA and was equal to 2.57 billions bp (749228240, 750000080, 536296870, 536258140 nucleotides A, T, G, C, respectively). *In silico* search for tRSS in SRNC was performed with the help of the same program algorithms that were used earlier to detect cRSS in mouse genomic DNA. Mean value of tRSS found for these 5 sets of SRNC was used as an expected value for cRSS found in the mouse genome. Both methods of analysis used in this study were developed by paper's authors.

To characterize a nucleotide composition of cRSS, we used frequency matrices [12, 13]. To create frequency matrices, we analyzed 199 and 105 sequences of 12fRSS and 23fRSS of mouse *Ig*, *Tcr* genes (fRSS with 12-bp and 23-bp spacers, respectively). Used fRSS are a part of fRSS that were analyzed in study [10] and were taken from the http://www.duke.edu/-Igcowell. Frequency matrices contained information about relative frequency of nucleotides A, T, G, C in aligned sequences of 12fRSS or 23fRSS. A weighting coefficient (*W*) of each 6724 analyzed sequences cRSS was calculated by formula [14]:

$$W = \frac{1}{m}\sum_{i=1}^{i=m}\left(\frac{n_i}{a_i}\right), \qquad (2)$$

where n_i is nucleotide frequency at position *i* taken from the frequency matrix; a_i is frequency of the most frequent nucleotide at position *i* taken from the frequency matrix; *m* is a total number of nucleotides in the analyzed sequence. *W* equal 1 means that a sequence of 12cRSS or 23cRSS fully corresponds to the consensus sequence of 12fRSS or 23fRSS. In turn, *W* equal 0 shows that there are no common nucleotides among analyzed cRSS and fRSS sequences. The same frequency matrices were used to characterize nucleotide composition of sequences of 12tRSS, 23tRSS as well as 12fRSS, 23fRSS.

To analyze 100-bp fragments adjacent to heptamers of cRSS, we have used BLASTN facility (http://www.ncbi.nlm.nih.gov/igblast/) [15]. It allowed comparing nucleotide composition of analyzed sequences with more than 68 thousand different sequences of V segments of mouse *Ig*, *Tcr* genes and their known duplications (igSeqNt database).

To determine which cRSS are localized in repetitive elements, we have developed algorithms that juxtaposed coordinates of cRSS to coordinates of known in the mouse genome repetitive elements. Coordinates of the first nucleotide of a heptamer and the ninth nucleotide of a nonamer were used as coordinates of the start and the end of cRSS. Coordinates of 9262721 of known repeats were taken from NCBI (ftp://ftp.ncbi.nih.gov/gnomes/Mmusculus/ARCHIVE/BUILD.35.1/masking_coordinates.gz). Repeats classification was performed using Repbase database [16]. Also we used annotations of repeats that were found by programme RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) in some fragments of mouse genomic DNA.

Results and Discussion. In previous work [11] we considered 28-bp and 39-bp DNA regions of mouse genome, where heptamer/nonamer sequences (435 unique types of sequences) fully corresponded to CACAGTG/ACAAAAACC structure or had only 14 – 15 common nucleotides as 12cRSS and 23cRSS. Their spacer regions had random nucleotide composition. Therefore, having made necessary calculations (see materials and methods formula 1), we determined a theoretically expected number of such types of 12cRSS and 23cRSS in DNA of each mouse chromosome. By summarizing *M* values obtained, we have determined that total number of cRSS in both DNA chains of all mouse chromosomes (total size is 2.57 billion nucleotides) must be 1238±35 (619±25 12cRSS and 23cRSS). Similar data have been got when analyzing 5 sets of SRNC. On average 1241 tRSS (606 12tRSS and 635 23tRSS) have been found in sequences of random nucleotide combinations. Thus, theoretically expected number of cRSS is 5.4-fold smaller than amount of cRSS really found in the mouse genome *in silico*. Therefore, we can suppose that about 18 % of cRSS in the mouse genome are a product of random nucleotide combinations. In turn, existence of 82 % cRSS might be a consequence of evolution processes. This group includes also 17 12cRSS and 3 23cRSS found in fragments of duplications of V segments of *Igk* (immunoglobulin k-chain) and *Tra* (T-cell receptor alpha chain) genes. Such motives should be considered as native RSS of *Ig*, *Tcr* genes (further referred to as dRSS).

As mentioned above, we have found that some of the examined target sites of RAG1/2 belonged to some repetitive elements. Therefore, we assumed that cRSS of the studied group can be specific for repetitive elements. Thus, in order to support or disprove this hypothesis, we should have to estimate the amount of cRSS located in repeats. Using specially developed algorithms, we have found that 4772 (71 %) cRSS (2077 12cRSS and 2695 23cRSS) are structural elements of 4752 repeats. 2779 (58 %), 1975 (41 %), 18 (<1 %) of such motifs are located in intergenic space, introns and exons of 1698 protein-coding genes. In 140 protein-coding genes (*Adk*, *Ankrd12*, *Abcc4*, *Cacna2d1*, *Ches1*, *Dock1*, *Dpp6*, *Dnahc8*, *Eif2c3*, *Fut8*, *Hk1*, *Il1rapl1*, *Katnal1*, *Map3k5*, *Mapk8ip3*, *Pkhd1l1*, *Prkce*, *Prkg1*, *Ptprg*, *Ptprk*, *Tasp1*, *Tcf12*, *Ulk2*, *Zfp262, etc.*) 12cRSS and 23cRSS (in accordance to the "12/23" rule) can theoretically mediate formation of deletions and inversions of whole exons. Usually, only one sequence of 12cRSS or 23cRSS can be found in repetitive elements. 20 repeats (three repeats B1_Mus1, two repeats B1_Mur2, B1_Mus2, B4 and one repeat B1_Mur1, B1_Mur4, PB1D10, L1_Rod, L1MCa, L1VL4, L1MEb, ORR1A2, ORR1D1-i, MTD, A-rich) are exceptions, where 23cRSS and 12cRSS are overlapped. In this case, the first 28 nucleotides of 23cRSS are a sequence of 12cRSS. Only in 2199 (46 %) cases the cRSS is completely located inside of repeats (further referred to as cRSS of the group "A"). 1561 (71 %), 582 (26 %), 25 (1 %), 15 (0,7 %) and 16 (0,7 %) of them were found in non-LTR retrotransposons (NRT), endogenous retroviruses and LTR retrotransposons (ER/LTR-RT), DNA transposons (DTR), simple repeats (SR), and non-classified by us types of repeats (NTR), respectively. In turn, 2573 (54 %) cRSS are located partially inside of repeats (further referred to as cRSS of the group "B"). In such cases, one part of the cRSS is a fragment of the 3'- or the 5'-end of the repeat, while the other part is a DNA segment flanking this repeat. 2279 (89 %), 151 (6 %), 109 (4 %), 9 (0.3 %) and 25 (1 %) cRSS of this group were found by us in NRT, ER/LTR-RT, SR, DTR, and NTR, respectively. Most often the members of the group "B" (322 (64 %) 12cRSS and 1761 (85 %) 23cRSS) are the structural elements of B1 family repeats. Detailed location data of 12cRSS and 23cRSS group "B" in different repeats are presented in table 1. Thus, the only 1 – 5 or 26 – 27 nucleotides of the 12cRSS sequence are most often detectable in the structure of ER/LTR-RT. In 51 % of cases 3'- and 5'-ends of L1 family repeats are presented by 6 – 20 nucleotide fragments of 23cRSS.

When considering cRSS, we have found them in the structure of 390 unique types of repeats. 345 types belong to the following 18 families: AcHobo, B1, B2, B4, CR1, ERV1, ERVK, ERVL, ID, L1, L2, MaLR, Mariner, MER1, MER2, MIR, MuDR and Tip100. The rest 32 types belong to the simple repeats (types (A)n, (CA)n, (CAAA)n, (T)n, (TG)n, (TTTG)n, (CATG)n, (CTGTG)n, (TCTG)n, (GAA)n, etc.) and 13 to the NTR group (types A-rich, T-rich, GA-rich, RMER1B, MMSAT4, YREP_Mm, etc.). The quantitative estimation shows that most often cRSS can be found in B1_Mus1 and B1_Mus2 repeats of B1 family. 22 % and 21 % of all cRSS were present in these repeats, respectively. In total, we have got 3111 cRSS (65 % from the total amount of motifs found in the repeats) including in the structure of 30 types of repeats related to B1 family (types B1_Mus1, B1_Mus2, B1_Mur1, B1_Mur1, B1_Mur1, B1_Mur1, B1F, PB1, PB1D1, PB1D10, PB1D10B, PB1D10I, PB1D10M, PB1D9, etc.). In turn, in the sequence of MaLR family repeats (74 types of repeats: MTD, MTC, MTEa, ORR1A1, ORR1A2, ORR1B1, ORR1B2, MLT1C, MLT1B, MLT1A1, etc.), L1 family (102 types of repeats: Lx9, Lx8, Lx7, Lx6, Lx5, L1VL4, L1Md_F, L1_Rod, L1_Mur3, L1_Mur2, L1_Mur1, etc.) and ERVK family (46 types of repeats: ETnERV, ETnERV2, RMER6B, RMER6A, RMER17B, RLTR8, RLTR27, RLTR25B, RLTR25A, IAP-d, etc.) we found 12 %, 9 %, and 2 % cRSS, respectively.

Data obtained show that 33 % and 38 % of cRSS studied here are fully or partially located inside of repeats, respectively. Therefore, we consider that these motifs are structures that are more specific to repetitive elements. In 97 % of cases possible DNA breaks induced by cRSS – RAG1/2 interactions would occur inside of repeats. In the last 3 % of cases these breaks would occur in DNA regions flanking repeats, because namely in such regions 165 cRSS of group "B" have a heptamer or at least the first nucleotide of their heptamers. At present we cannot clearly point out in which types of repeats appearance of the cRSS is mediated by evolution processes. However, one can surely state that existence of most cRSS in repeats is not random. This statement is based on the fact that the total amount of cRSS detected in repeats is 3.8-fold higher than theoretically expected number of cRSS in

Table 1.
Analysis of location of 12cRSS and 23cRSS of group "B" found in structure of known classes of repeats and repeats of B1, L1, MaLR, ERVK families

12cRSS of group "B"		Fragment of 12cRSS sequence located in repeat (bp)						
Location	Number	1–5	6–10	11–15	16–20	21–23	24–25	26–27
NRT	392	$16_{(4)}$	$12_{(3)}$	$20_{(5)}$	$31_{(8)}$	$41_{(11)}$	$35_{(9)}$	$237_{(60)}$
ER/LTR-RT	57	$18_{(31)}$	$2_{(4)}$	$2_{(4)}$	$5_{(9)}$	$8_{(14)}$	$7_{(12)}$	$15_{(26)}$
SR	43	$17_{(39)}$	$5_{(12)}$	$11_{(26)}$	$6_{(14)}$	$3_{(7)}$	$1_{(2)}$	–
DTR	3	$1_{(33)}$	–	–	–	–	–	$2_{(67)}$
NTR	9	$1_{(11)}$	$3_{(34)}$	–	$2_{(22)}$	$2_{(22)}$	–	$1_{(11)}$
B1	322	$2_{(<1)}$	$3_{(1)}$	$5_{(2)}$	$20_{(6)}$	$35_{(11)}$	$30_{(9)}$	$227_{(70)}$
L1	27	$1_{(4)}$	$6_{(22)}$	$9_{(33)}$	$5_{(19)}$	$3_{(11)}$	$1_{(4)}$	$2_{(7)}$
MaLR	42	$11_{(26)}$	$1_{(2)}$	$1_{(2)}$	$4_{(10)}$	$7_{(17)}$	$7_{(17)}$	$11_{(26)}$
ERVK	7	$3_{(43)}$	–	–	$1_{(14)}$	$1_{(14)}$	–	$2_{(29)}$
23cRSS of group "B"		Fragment of 23cRSS sequence located in repeat (bp)						
Location	Number	1–5	6–10	11–20	21–25	26–29	30–31	32–38
NRT	1887	$21_{(1)}$	$18_{(1)}$	$40_{(2)}$	$75_{(4)}$	$105_{(6)}$	$1436_{(76)}$	$192_{(10)}$
ER/LTR-RT	94	$43_{(45)}$	$1_{(1)}$	$5_{(5)}$	$8_{(9)}$	$8_{(9)}$	$15_{(16)}$	$14_{(15)}$
SR	66	$16_{(24)}$	$2_{(3)}$	$25_{(38)}$	$10_{(15)}$	$7_{(11)}$	$2_{(3)}$	$4_{(6)}$
DTR	6	–	$2_{(33)}$	$1_{(17)}$	–	$2_{(33)}$	–	$1_{(17)}$
NTR	16	–	$3_{(19)}$	$4_{(26)}$	$2_{(12)}$	$2_{(12)}$	$3_{(19)}$	$2_{(12)}$
B1	1761	$4_{(<1)}$	$2_{(<1)}$	$22_{(1)}$	$58_{(3)}$	$97_{(6)}$	$1416_{(80)}$	$162_{(9)}$
L1	35	$6_{(17)}$	$10_{(28)}$	$8_{(23)}$	$5_{(14)}$	$3_{(9)}$	–	$3_{(9)}$
MaLR	64	$38_{(59)}$	$1_{(2)}$	$5_{(8)}$	$5_{(8)}$	$3_{(5)}$	$4_{(6)}$	$8_{(12)}$
ERVK	17	$4_{(24)}$	–	–	$1_{(6)}$	$4_{(24)}$	$7_{(41)}$	$1_{(6)}$

Appendix: NRT is non-LTR retrotransposons, ER/LTR-RT is endogenous retroviruses and LTR retrotransposons, DTR is DNA transposons, SR is simple repeats and NTR is types of repeats not classified by us. Percentage is shown in brackets.

all mouse genome. What is the reason for this? We also think that some types of repetitive elements can participate in spreading of full-size cRSS (cRSS of the group "A") in the mouse genome. In other cases, cRSS can appear *de novo* immediately in DNA regions where transpositions of repeats occurred (cRSS of the group "B"). Thereby, certain types of repetitive elements are able to assure evolutionary accumulation of possible target sites of RAG1/2 in mammalian genome.

By using frequency matrices, we found that nucleotide compositions of studied cRSS (both located and not located in repeats) were significantly different from known fRSS (Table 2). Weighting coefficients of 12cRSS and 23cRSS are 0.59 – 0.95 and 0.57 – 0.88, respectively. As expected, the main difference was observed mostly in the spacer regions. We also found that 0.6; 0.2; 0.5; 7; 41; 31; 16; 4; 0.2 % of spacers of 12cRSS had 100; 92; 83; 75; 67; 58; 50; 42; 33 % homology to the spacers of 12fRSS. And 0.1; 0.03; 0.3; 1; 3; 8; 19; 31; 33; 5; 0.03 % spacers of 23cRSS had 100; 91; 65; 61; 57; 52; 48; 43; 39; 35; 30 % homology to the spacers of 23fRSS. In general, the structure of mentioned motives has very high variability. The identified cRSS consist of 57, 146, 392, 1878 and 2623 unique types of heptamers, nonamers, heptamer/nonamer sequences, 12-bp and 23-bp spacers, respectively.

Only 121 (4 %) 12cRSS and 182 (5 %) 23cRSS have weighting coefficients equal to 0.80 – 0.95 and 0.75 – 0.88 (total 303 cRSS). Nucleotides typical for sequences of 140 (70 %) 12fRSS and 80 (76 %) 23fRSS can be found in structure of these motives. Mentioned above 20 dRSS belong indeed to this group of motives. The other 163, 3, 113 and 4 cRSS are located in intergenic space, in pseudogenes *Kif22-ps*, *LOC50777*, *Olfr1452-ps1*, in introns of 78 protein-coding genes (*Astn1*, *Grik3*, *Jak1*, *Kras*, *Mapk4*, *Pcsk5*, *Rb1cc1*, etc.) and in exons of genes *Xkr5*, *LOC545698*, *Slc30a4*, *Neb*, respectively. In genes *Bai3, Large*, *LOC629706*, *Plcb4*, *Tcf4*, *Tnpo3* and *Prkg1*, *Ptprn2*, *Slc24a2*, *Trhde* they theoretically can mediate formation of exon deletions and inversions. Analysis of 100-bp sequences adjacent to heptamers of the described group of cRSS by BLASTN lead to a conclusion that two 12cRSS and one 23cRSS located in intergenic space as well as 23cRSS of pseudogene LOC50777 actually are fragments of undescribed yet duplications of *Igk* and *Tra* genes V segments. In such cases, 79 – 90 bp fragments of analyzed sequences had 84 – 92 % homology to the sequences of known V segments (igSeqNt database). That is why such motives (further referred to as d'RSS) as well as dRSS should be considered as native RSS of mouse *Ig*, *Tcr* genes. In turn, 158 (52 %) cRSS were found in structure of 60 types of repeats (B1_Mur2, B1_Mur3, B1_Mur4, B1_Mus1, B1_Mus2, B3, B4, ETnERV, Lx5, Lx6, Lx7, Lx8, Lx9, MTC, MTD, MTEa, MTEb, PB1, PB1D10, PB1D7, RLTR14, RSINE1, etc.). It is important to note that the same nucleotide compositions are typical for 19 (3 %) 12tRSS and 58 (9 %) 23tRSS found in SRNC. As mentioned above (see materials and methods), total amount of tRSS was used in this study to estimate theoretically expected number of cRSS in the mouse genome (these cRSS are products of random nucleotide combinations). Therefore, making comparisons, we can postulate that the existence of 16 % (19/121) 12cRSS and 32 % (58/182) 23cRSS with weighting coefficients equal to 0.80 – 0.95 and 0.75 – 0.88 are random in the mouse genome.

Most often nucleotides which are not typical for fRSS dominate in structure of studied cRSS. 2380 (78 %) 12cRSS and 3249 (88 %) 23cRSS are related to such motives and have lower *W* values (0.65 – 0.74). 30 (15 %) 12fRSS and 20 (19 %) 23fRSS have similar weighting coefficients. Therefore, based on these data we can conclude that cRSS with such nucleotide composition will effectively interact with RAG1/2. In addition, their heptamers and nonamers have all functionally significant nucleotides, whereas changes in spacer regions are not so important to block DNA cleavage by endonucleases [17, 18]. However, we propose that such interactions will be less effective then interactions RAG1/2 to cRSS with higher weighting coefficients. Table 3 shows detailed results of nucleotide compositions of cRSS detected in repeats of known classes. The highest weighting coefficients (0.85 – 0.88 and 0.90) have five 12cRSS, located in MTD, MLT1A1, ORR1B2, MTC and ORR1A4 of MaLR family repeats (LTR retrotransposones). It is important to note that cRSS located in similar repeats can be significantly different from each other. The exception of the rule is a RSS detected in repeats of B1 family, where 51 % 12cRSS and 11 % 23cRSS of such motives have sequences CACAGAGAAACCCTGTC TCAAAAAAACC and CACAGAGAAACCCTGTC TCGAAAAACAAAAACAAAAACC with weighing coefficients 0.71 and 0.73, respectively. Table 4 shows consensus sequences of cRSS found in repeats of various families.

Conclusion. We have found that in the mouse genome there is evolutionary accumulation of possible RAG1/2 proteins target sites. Heptamer/nonamer sequences of such motives match the structure CACAGTG/ACAAAAACC or share 14 – 15 nucleotides of this structure. In 71 % of cases cRSS are structural elements of 390 types of repeats. In a structure of about 5 % motives can be found nucleotides typical for recombination signal sequences of functional V, D and J segments of mouse *Ig* and *Tcr* genes. The existence of 25 % of such cRSS in the mouse genome may be considered as a result of random nucleotide combinations. In most cases, spacers of

Table 2.
Analysis of nucleotide composition of cRSS found in different regions of the mouse genome as well as analysis of nucleotide composition of fRSS and tRSS, using frequency matrices

Analyzed structures			Weighting coefficients						
Location	Type	Number	0,95–0,90	0,89–0,85	0,84–0,80	0,79–0,75	0,74–0,70	0,69–0,65	0,64–0,57
Ig, Tcr-genes	12fRSS	199	$54_{(27)}$	$45_{(23)}$	$41_{(21)}$	$20_{(10)}$	$20_{(10)}$	$10_{(5)}$	$9_{(4)}$
	23fRSS	105	$1_{(1)}$	$34_{(32)}$	$27_{(26)}$	$18_{(17)}$	$16_{(15)}$	$4_{(4)}$	$5_{(5)}$
SRNC	12tRSS	606	–	$1_{(<1)}$	$18_{(3)}$	$127_{(21)}$	$258_{(43)}$	$175_{(29)}$	$27_{(4)}$
	23tRSS	635	–	–	$3_{(<1)}$	$55_{(9)}$	$235_{(37)}$	$280_{(44)}$	$62_{(9)}$
DNA^{1}	12cRSS	3034	$8_{(<1)}$	$14_{(<1)}$	$99_{(3)}$	$474_{(16)}$	$1797_{(59)}$	$583_{(19)}$	$59_{(2)}$
	23cRSS	3690	–	$1_{(<1)}$	$10_{(<1)}$	$171_{(5)}$	$1591_{(43)}$	$1660_{(45)}$	$257_{(7)}$
DNA^{2}	12cRSS	2077	$1_{(<1)}$	$4_{(<1)}$	$52_{(2)}$	$286_{(14)}$	$1372_{(66)}$	$343_{(17)}$	$19_{(1)}$
	23cRSS	2695	–	–	$2_{(<1)}$	$99_{(4)}$	$1230_{(45)}$	$1230_{(45)}$	$134_{(5)}$
DNA^{3}	12cRSS	932	–	–	$39_{(4)}$	$188_{(20)}$	$424_{(46)}$	$240_{(26)}$	$41_{(4)}$
	23cRSS	988	–	–	$4_{(<1)}$	$72_{(7)}$	$358_{(36)}$	$432_{(44)}$	$122_{(12)}$
IR^{A}	12cRSS	498	–	$3_{(<1)}$	$24_{(5)}$	$108_{(22)}$	$223_{(45)}$	$121_{(24)}$	$19_{(4)}$
	23cRSS	533	–	–	$1_{(<1)}$	$35_{(7)}$	$197_{(37)}$	$233_{(44)}$	$67_{(12)}$
IR^{B}	12cRSS	1170	$1_{(<1)}$	$2_{(<1)}$	$32_{(3)}$	$169_{(14)}$	$743_{(64)}$	$214_{(18)}$	$9_{(1)}$
	23cRSS	1609	–	–	$2_{(<1)}$	$63_{(4)}$	$693_{(43)}$	$768_{(48)}$	$83_{(5)}$
IPG^{A}	12cRSS	394	–	–	$17_{(4)}$	$75_{(19)}$	$174_{(44)}$	$109_{(28)}$	$19_{(5)}$
	23cRSS	430	–	–	$4_{(1)}$	$35_{(8)}$	$151_{(35)}$	$189_{(44)}$	$51_{(11)}$
IPG^{B}	12cRSS	904	–	$2_{(<1)}$	$20_{(2)}$	$116_{(13)}$	$628_{(69)}$	$129_{(14)}$	$9_{(1)}$
	23cRSS	1071	–	–	–	$35_{(3)}$	$535_{(50)}$	$451_{(42)}$	$50_{(5)}$
EPG^{A}	12cRSS	44	–	–	$1_{(2)}$	$5_{(11)}$	$24_{(55)}$	$12_{(27)}$	$2_{(5)}$
	23cRSS	26	–	–	–	$2_{(8)}$	$10_{(38)}$	$10_{(38)}$	$4_{(16)}$
EPG^{B}	12cRSS	4	–	–	–	$1_{(25)}$	$3_{(75)}$	–	–
	23cRSS	14	–	–	–	$1_{(7)}$	$3_{(22)}$	$8_{(57)}$	$2_{(14)}$
PSG	12cRSS	3	–	–	$2_{(67)}$	–	–	–	$1_{(33)}$
	23cRSS	4	–	–	$1_{(25)}$	–	$1_{(25)}$	$2_{(50)}$	–
D^{1}	12cRSS	17	$7_{(41)}$	$7_{(41)}$	$3_{(18)}$	–	–	–	–
	23cRSS	3	–	$1_{(33)}$	$2_{(67)}$	–	–	–	–
D^{2}	12cRSS	2	–	$1_{(50)}$	$1_{(50)}$	–	–	–	–
	23cRSS	2	–	–	$2_{(100)}$	–	–	–	–
Del	12cRSS	91	–	$1_{(1)}$	$1_{(1)}$	$11_{(12)}$	$59_{(65)}$	$17_{(19)}$	$2_{(2)}$
	23cRSS	91	–	–	–	$2_{(2)}$	$52_{(57)}$	$30_{(33)}$	$7_{(8)}$
Inv	12cRSS	104	–	–	$3_{(3)}$	$13_{(13)}$	$75_{(72)}$	$12_{(11)}$	$1_{(1)}$
	23cRSS	104	–	–	–	$2_{(2)}$	$60_{(58)}$	$35_{(33)}$	$7_{(7)}$

Appendix: SRNC is sets of sequences with random nucleotide combination. DNA^{1} is all analyzed mouse genomic DNA. DNA^{2} is genomic DNA, which includes repeats only. DNA^{3} is genomic DNA, which does not include repeats and sequences of known and identified in this study duplications of V segments of mouse *Ig*, and *Tcr* genes. IR is intergenic space. IPG is introns of protein-coding genes. EPG is exons of protein-coding genes. PSG is pseudogenes. D^{1} is known duplications of V segments of mouse *Ig*, *TCR* genes. D^{2} is identified in this study duplications of V segments of mouse *Ig*, *TCR* genes. Del and Inv is cRSS, which theoretically can mediate formation of deletions and inversions, respectively. Index ^{A} indicates that cRSS in this regions are located in repeats. Index ^{B} indicates that cRSS in this regions are not located in repeats.

Table 3.
Analysis of nucleotide composition of 12cRSS and 23cRSS found in repeats of different classes, using frequency matrices

cRSS				Weighting coefficients						
Classes of repeats	Group	Type		0,95–0,90	0,89–0,85	0,84–0,80	0,79–0,75	0,74–0,70	0,69–0,65	0,64–0,57
		12/23	Number							
NRT	A	12cRSS	1181	–	–	$22_{(2)}$	$142_{(12)}$	$865_{(73)}$	$141_{(12)}$	$11_{(1)}$
		23cRSS	380	–	–	$1_{(<1)}$	$35_{(9)}$	$143_{(38)}$	$167_{(44)}$	$34_{(9)}$
	B	12cRSS	392	–	–	$7_{(2)}$	$56_{(14)}$	$277_{(71)}$	$49_{(12)}$	$3_{(1)}$
		23cRSS	1887	–	–	–	$32_{(2)}$	$935_{(50)}$	$858_{(45)}$	$62_{(3)}$
ER/LTR-RT	A	12cRSS	360	–	$4_{(1)}$	$14_{(4)}$	$45_{(12)}$	$157_{(44)}$	$135_{(38)}$	$5_{(1)}$
		23cRSS	222	–	–	$1_{(<1)}$	$21_{(9)}$	$91_{(41)}$	$89_{(40)}$	$20_{(9)}$
	B	12cRSS	57	$1_{(2)}$	–	$6_{(10)}$	$14_{(24)}$	$34_{(60)}$	$2_{(4)}$	–
		23cRSS	94	–	–	–	$9_{(10)}$	$24_{(26)}$	$54_{(57)}$	$7_{(7)}$
DTR	A	12cRSS	9	–	–	–	$2_{(22)}$	$3_{(33)}$	$4_{(45)}$	–
		23cRSS	16	–	–	–	–	$4_{(25)}$	$10_{(62)}$	$2_{(13)}$
	B	12cRSS	3	–	–	$2_{(67)}$	–	$1_{(33)}$	–	–
		23cRSS	6	–	–	–	–	$5_{(83)}$	$1_{(17)}$	–
SR	A	12cRSS	9	–	–	–	$6_{(67)}$	$1_{(11)}$	$2_{(22)}$	–
		23cRSS	6	–	–	–	–	$4_{(67)}$	$2_{(33)}$	–
	B	12cRSS	43	–	–	–	$15_{(35)}$	$22_{(51)}$	$6_{(14)}$	–
		23cRSS	66	–	–	–	$2_{(3)}$	$18_{(27)}$	$39_{(59)}$	$7_{(11)}$
NTR	A	12cRSS	14	–	–	$1_{(7)}$	$4_{(29)}$	$6_{(43)}$	$3_{(21)}$	–
		23cRSS	2	–	–	–	–	$2_{(100)}$	–	–
	B	12cRSS	9	–	–	–	$2_{(22)}$	$6_{(67)}$	$1_{(11)}$	–
		23cRSS	16	–	–	–	–	$4_{(25)}$	$10_{(63)}$	$2_{(12)}$

Appendix: NRT is non-LTR retrotransposons, ER/LTR-RT is endogenous retroviruses and LTR retrotransposons, DTR is DNA transposons, SR is simple repeats and NTR is types of repeats not classified by us. Percentage is shown in brackets.

analyzed 12cRSS and 23cRSS had 58 – 67 % and 30 – 47 % homology with spacers of fRSS, respectively.

А. Ю. Губский, В. Г. Зиньковский

Структурный анализ группы возможных ДНК-мишеней белков RAG1/2, обнаруженных в геноме мыши *in silico*, и их идентификация в известных типах повторяющихся элементов

Резюме

С использованием математических методов анализа, а также специально разработанных алгоритмов установлено, что количество ранее обнаруженных в геноме мыши возможных сайтов – мишеней белков RAG1/2 (cRSS) в 5,4 раза превышает теоретически ожидаемое число. В 71 % случаев cRSS являются структурными элементами 390 типов повторов. В структуре около 5 % мотивов обнаружены нуклеотиды, типичные для большинства сигнальных последовательностей рекомбинации функциональных V, D, J-сегментов Ig- и Tcr-генов мыши (fRSS). Существование 25 % из них в ДНК анализируемого вида можно рассматривать как следствие случайных комбинаций нуклеотидов. Структуры спейсерных участков исследуемых 12cRSS и 23cRSS, как правило, имеют 58–67 и 30–47 % гомологии с аналогичными структурами fRSS соответственно.

Ключевые слова: cRSS, V(D)J-рекомбинация, RAG1, RAG2.

А. Ю. Губський, В. Г. Зіньковський

Структурний аналіз групи можливих ДНК-мішеней білків RAG1/2, виявлених у геномі миші *in silico*, та їхня ідентифікація у відомих типах повторюваних елементів

Резюме

З використанням математичних методів аналізу, а також спеціально розроблених алгоритмів встановлено, що кількість виявлених у геномі миші можливих сайтів – мішеней білків RAG1/2 в 5,4 разу перевищує теоретично очікуване число. У 71 % випадків cRSS є структурними елементами 390 типів повторів. У структурі близько 5 % мотивів виявлено нуклеотиди, типові для більшості сигнальних послідовностей рекомбінації функціональних V, D, J-сегментів Ig- і Tcr-генів миші (fRSS). Існування 25 % з них у ДНК досліджуваного виду потрібно розглядати як наслідок випадкових комбінацій нуклеотидів. Структури спейсерних ділянок аналізованих 12cRSS і 23cRSS,

Table 4.
Consensus sequences of 12cRSS and 23cRSS found in repeats of 18 families

FR	TG	12cRSS		23cRSS	
		Number	Consensus sequence	Number	Consensus sequence
AcHobo	A	0	–	1	CACCGTGCAAAGTTCTAAAAAAAAAAAAAAAAAAAAAACC
B1	A	924	CACAGAGAAACCCTGTCTCAAAAAAACC	104	CACAGTGARTTCCAGGMTAGCCAGGAATACACAAAAACC
	B	322	CACAGAGAAACCCTGTCTCAAAAAAACC	1761	CACAGAGAAACCCTGTCTCGAAAAACAAAAACAAAAACC
B2	A	31	CACAGTGTGAGTGCTAGGAACAAAAACC	12	CACAGTGTACTCAAAMAACAAAACAAAAAAACAAAAACC
	B	9	CACAGTGAAACAAAAMAAAACAAAAACC	15	CACAGTGAAATAAAAAAAAAAAAAAHAAAAACAAAAACC
B4	A	48	CACAGTGAAAGAGCSTAACACAAAAACC	63	CACAGTGCCCTGGGTTCCATCCCCAGCACCACAAAAACC
	B	24	CACAGTGAAAAAAAAACAAACAAAAACC	42	CACAGTGAAATCATGTAACAACAAAAAMAAACAAAAACC
CR1	A	1	CACACTGAAATCACAGTCAACTAAAACC	0	–
ERV1	A	5	CACAGTGAGRGWKAAAARMACAAAAACC	7	CACAGTSAGTMTHATMCAGCAAAACCAAAAACAAAAACC
	B	4	CACAGTGATTCAWCWGTACASAAAAAGC	2	CACAGTSYTRRYMCMRARMWWRKSMAGMRGAMARAAACM
ERVK	A	22	CACAGTGTCAAGMAATWWWACAAAAACC	62	CACAGTGGTTGTTGTTGATCTTGTGGAAAAACACAAACC
	B	4	CACAGTGGRNCANRAMAAAACAAAAATC	10	CACAGTGATWAATWWCAAAATACATAAGGTACAAAAACC
ERVL	A	4	CACAGTGNRNACAAAWTNMACMAAAACC	9	CACAGTGAKAMMTWACTGGACCHAAWMGAGACAAAAACC
	B	2	CACAGTGTTWAWMWMRRCKWCAGAAACC	4	CACAKTGNAGTAGCCWGAGTTTGWAAGCNRACRAAAACC
ID	A	0	–	2	CACAGWGCMCYRKGCTKKRTCCCYARMAYYWCAAAAASY
	B	3	CACAGTGAADAAAAHADGAACAAAAACC	6	CACAGTGCCCTGGGTTCAATCCCCARCAYCACAAAAACC
L1	A	163	CACAGTGAAAAAAAAAAAACAAAAACC	189	CACAGTGAAACATGGATACAACAAAAAAAAACAAAAACC
	B	27	CACAGTGGCTYTAWAAGAAACAAAAACC	36	CACAGTGWAAATAWAAAAAAAAAAAAAAAAACAAAAACC
L2	A	6	CACAGTGRGTGMGAMASAVACAAAAACC	1	CACAGACTAAAATTCCAGTGGGGAAGACAGACAAAAACC
	B	0	–	4	CACAGTGCNAAAGCANNTATYGGNNNAWAMACAAAAAMC
MaLR	A	319	CACAGTGTCTGTTCACAGCACTAAAACC	143	CACAGTGAGGAAATACTGGACCAAAGCAAGACAAAAACC
	B	41	CACAGTGATAAAAAACATAACAAAAACC	64	CACAGTAATGTTTAACTTACACCATAACCTACAAAAACA
Mariner	A	0	–	1	CACATTTCTTTCTGCCAATGTCTTTCTGTAACAAAAACC
MER1	A	6	CACAGTGWYAASVCCCKARACAAAAACC	13	CACAGTGDWGAAWAAAGTCAWTAAGRAAAAACAAAAACC
	B	3	CACAGBGADACGTVVCCDAACAAAAACC	5	CACAGTGGCAGAAMACATWGTWYAAMCMATACAAAAACC
MER2	A	2	CACAGTCYKMAAAWAWRAAAMARAAACC	2	CACASYGTRWTWACWKYWKMSWWMKRWWSKASAAAAAYC
MIR	A	1	CACAGTGAAGAGTACTTAGACTAAAACA	7	CACAGAGHYTGTATYAHCTAATTTAAWTAYACAAAAACC
	B	4	CACAGTGTGNAATATCNTAACAAAAACC	7	CACAGTGHTGKYHTCTTTCTAMTTCATCTCACAAAAACC
MuDR	A	0	–	1	CACAGTGATTTCGAAGAAACTAGGCTGCCTACATAAATC
Tip100	A	1	CACAGGTGCCTTTACTTACACAAAAACC	0	–
SR	A	9	CACAGAGAKACACATACASACAAAAACA	3	CACAGTGCAHVCACACACATATGVATACACACACAAACA
	B	43	CACAGTGAAAAAAAAAAAACAAAAACC	64	CACAGTGMAAAACAAAAAAAAAAAAAAAAAACAAAAACC

Appendix: FR is families of repeats; TG is types of cRSS group; SR is simple repeats; Y, R, W, S, K, M, D, V, H, B, N is C/T, A/G, A/T, G/C, T/G, C/A, A/T/G, A/G/C, A/T/C, T/G/C, A/T/G/C, respectively.

як правило, мають 58–67 і 30–47 % гомології з аналогічними структурами fRSS відповідно.

Ключові слова: cRSS, V(D)J-рекомбінація, RAG1, RAG2.

REFERENCES

1. *Tonegawa, S.* Somatic generation of antibody diversity // Nature.–1983.–**302**.–P. 575–581.
2. *Tonegawa S., Brack C., Hozumi N., Pirrotta V.* Organization of immunoglobulin genes // Cold Spring Harbor Symp. Quant. Biol.–1978.–**42**.–P. 921–931.
3. *Oettinger M., Schatz D., Gorka C., Baltimore D*. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination // Science.–1990.–**248**.–P. 1517–1523.
4. *Aplan P. D., Lombardi D. P., Ginsberg A. M., Cossman J., Bertness V. L., Kirsch I. R*. Disruption of the human SCL locus by «illegitimate» V-(D)-J recombinase activity // Science.–1990.–**256**.–P. 1426–1429.
5. *Tsuji H., Ishii-Ohba H., Katsube T., Ukai H., Aizawa S., Doi M., Hioki K., Ogiu T*. Involvement of illegitimate V(D)J recombination or microhomology-mediated nonhomologouse end-joining in the formation of intragenic deletions of the *Notch*1 gene in mouse thymic lymphomas // Cancer Res.–2004.–**15**.–P. 8882–8890.
6. *Scheerer J. B., Xi L., Knapp G. W., Setzer R. W., Bigbee W. L., Fuscoe J. C.* Quantification of illegitimate V(D)J recombinase mediated mutations in lymphocytes of newborns and adults // Mutat. Res.–1999.–**431**.–P. 291–303.
7. Gubsky *A. Yu*. Structural analysis of recombination signal sequences of three types V, D, J segments of human immunoglobulin and T-cell receptors genes // Odessa Medical Journal. – 2005. – **5** – P. 10 – 12.
8. *Matsuda F., Ishi K., Bourvagnet P., Kumma I., Hayashida H., Miyata T., Honjo T.* The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus // J. Exp. Med.–1998.–**188**.–P. 2151–2162.
9. *Lewis S. M., Agard E., Suh S., Czyzyk L*. Cryptic signals and the fidelity of V(D)J joining // Mol. and Cell. Biol.–1997.–**17**.– P. 3125–3136.
10. *Cowell L. G., Davila M., Yang K., Kepler T. B., Kelsoe G.* Prospective estimation of recombination signal efficiency and identification of functional cryptic signals in the genome by statistical modeling // J. Exp. Med.–2003.–**197**.–P. 207–220.
11. Gubsky *A. Yu.*, Zinkovsky V. G. Search of cRSS with supposedly high recombination potential and identification of their location in the mouse genome // Odessa Medical Journal. – 2006. – **98**, № 6. – P. 11 – 14.
12. *Hawley D. K., McClure W. R.* Compilation and analysis of *Escherichia coli* promoter DNA sequences // Nucl. Acids Res.–1983.–**11**.–P. 2237–2255.
13. *Stormo G. D.* DNA binding sites: representation and discovery // Bioinformatics.–2000.–**16**.–P. 16–23.
14. *Harr R., Haggstrom M., Gustafsson P.* Search algorithm for pattern math analysis of nucleic acid sequences // Nucl. Acids Res.–1983.–**11**.–P. 2943–2957.
15. *Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucl. Acids Res.–1997.–**25**.–P. 3389–3402.
16. *Jurka J.* Repbase update: a database and an electronic journal of repetitive elements // Trends Genet.–2000.–**16**.–P. 418–420.
17. *Fanning L., Connor A., Baetz K., Ramsden D., Wu G. E.* Mouse RSS spacer sequences affect the rate of V(D)J recombination // Immunogenetics.–1996.–**40**.–P. 146–150.
18. *Akamatsu Y., Tsurushita N., Nagawa F., Matsuoka M., Okazaki K., Imai M., Sakano H.* Essential residues in V(D)J recombination signals // J. Immunol.–1994.–**153**.–P. 4520–4529.

UDC 575.113 + 577 .2 + 599.89
Received 02.10.07