Computer Science and Information Systems 15(2):473–485 https://doi.org/10.2298/CSIS170523004B

CpG Islands' Clustering Uncovers Early Development Genes in the Human Genome

Vladimir N. Babenko[1,2], Anton G. Bogomolov[1,2], Roman O. Babenko[2],
Elvira R. Galieva[2] and Yuriy L. Orlov[1,2,3]

[1] Institute of Cytology and Genetics SB RAS, Lavrentyeva 10,
630090 Novosibirsk, Russia
bob@bionet.nsc.ru
[2] Novosibirsk State University, Pirogova 1, 630090 Novosibirsk, Russia
{mantis_anton,galieva,orlov}@bionet.nsc.ru
[3] The A.O. Kovalevsky Institute of Marine Biological Research of RAS, Nakhimov ave. 2,
299011 Sevastopol, Russia

Abstract. We address the problem of the annotation of CpG islands (CGIs) clusters in the human genome. Upon analyzing gene content within CGIs clusters, piRNA, tRNA, and miRNA-encoding genes were found as well as CpG-rich homeobox genes reported previously. Chromosome-wide CGI density is positively correlated with replication timing, confirming that CGIs may serve as open chromatin markers. Early embryonic stage expressed KRAB-ZNF genes abundant at chromosome 19 were found to be interlinked with CGI clusters. We detected that a number of long CGIs and CGI clusters are, in fact, tandem copies with multiple annotated macrosatellites and paralogous genes. This finding implies that tandem expansion of CGIs may serve as a substrate for non-homologous recombination events.

Keywords: CpG islands, bioinformatics, human genome, macrosatellite, genome annotation, genome repeats, DNA methylation

1. Introduction

CpG dinucleotide rich genome regions, also known as CpG islands (CGIs), are important functional elements of vertebrate genomes [1,2]. In particular, in the majority of vertebrate genes, CpG islands coincide with gene promoter areas. In some cases, the transcription from CpG-island containing promoters is bidirectional, this is related to self-complementarity of CG dinucleotides. CGIs are the key contributors to global methylation landscapes. Degenerate content of CGIs (biased CG frequency) assumes a higher probability of tandem repeats and palindromes inside a CGI. To continue our recent work [3], this study identifies tandem duplications within and across human CGIs, representing 400–5000 bp mega-monomers. Moreover, we found intra- and intergene tandem duplications of CGIs. The intergenic CGI duplications were possibly mediated

The materials have been presented first at BELBI conference on bioinformatics at Belgrade University.

by GC-rich subcentromeric and telomeric satellites, as well as by SINE elements. Higher similarity of monomers in tandem repeats indicate a selective pressure and a concerted evolution of such loci. Analysis of the sequence context within intergenic CGI repeats points at their possible role in adjustment of CG-content within a genome segment. Tandem CGIs are transcriptionally active in a wide range of tissues and cell lines. The phenomenon of CGI clustering is manifested most prominently in chromosome 19, known for its abundance of segment duplications and gene expansions. Additionally, CGIs duplications led to the emergence of DXZ4, a unique 45 kb genome segment with variable number tandem repeats (macrosatellite) located on the q arm of chromosome X [4].

CGIs are found in all vertebrate genomes [1]. The formal definition of CGI has several parameters: percentage of CG content > 60 %; ratio of observed and expected GC dinucleotide (cg_obs/cg_exp) > 0.6; island length is >300 bp. This ratio is species-specific, e.g. mouse genome maintains lower CG dinucleotide density than in human [5]. The human genome features about 27-30 thousand CGIs (www.genome.ucsc.edu).

CGIs overlap with promoter regions for 50–70% genes. Such promoters are called CpG promoters [2]. CGIs associated with the promoter of some homeobox genes (HOX, PAX), and about 5% of all other promoters, serve as tissue-specific targets for methyltransferases: after CGI methylation, the promoter becomes repressed (not transcribed). Details of methylation statistics in CGIs are presented in [3]. Here we discuss mathematical approaches and applications for structure analysis of CGI clusters in the human genome, in the context of repeat search and text complexity estimates developed earlier [6,7,8].

2. Materials and Methods

A set of 26 412 CGIs was retrieved from the table cpgIslandExt (www.genome.ucsc.edu; version hg19).

To identify significant CGI clustering, the human genome was split into 10Kb non overlapping segments (bins) (243 785 bins total). The number of CGIs per bin (CGI density) was assessed as a total number of CGIs divided by the number of bins λ = 26412/243785 = 0.1. The expected number of CGIs per segment was approximated using a Poisson distribution

$$P(X = k) = \exp(-\lambda)\frac{\lambda^{-k}}{k!} \tag{1}$$

The probability P was computed by the formulae

$$P(X > k) = 1 - \sum_{n=0}^{\mathrm{k}} P(X = n) \tag{2}$$

For the expected number of CGIs per bin given λ = 0.1, the integral probabilities (P-values) are the following: P(X > 0) = 4.7E–3; P(X > 1) = 1.6E–4; P(X > 2) = 3.8E–6; P(X > 3) = 7.7E–8.

False Discovery Rate (FDR) is 1E-3 for bins with four and more CGIs, taking into account adjustment for multiple comparisons (10E-5). Thus, 4 CGIs per bin could be considered a significant threshold of CGI density.

Tandem repeat clusters within the long CGIs were elucidated using TRF [6].

3. CGI Annotation

The median CGI length in the human genome is about 1 Kb. It is therefore reasonable to propose that long CGIs are tandem extensions of monomer islands. In particular, we found 16 CGIs with lengths longer than 10Kb.

For all other CGIs, we searched for their clusterisation within chromosome segments. In total, 25 clusters containing 141 CGI were found located in chromosomal clusters, with 4 or more CGIs per 10Kb (Table 1). The largest CGI clusters were found on chromosomes 19, 4, 2, 7 and 10 (Table 1).

Table 1. Number of clusters and CGI contained by chromosomes (for CpG islands less than 10Kb and with monomer number greater than 3)

Number	**Chromosome**													
	2	**4**	**5**	**6**	**7**	**9**	**10**	**11**	**12**	**14**	**15**	**17**	**19**	**Total**
CGI	13	20	9	4	13	8	12	8	4	4	6	8	32	141
Clusters	3	2	2	1	1	2	3	2	1	1	1	2	4	25

Tandem clusters of CGIs on chromosomes 4 and 19 (2 clusters) mediated by dispersed repeats are related to macrosatellites and include 64 islands. Their features are shown in the next section. Gene-associated CpG islands (141-64 = 63) are assembled into 22 clusters.

The clusters were annotated manually (Table 2). The majority of CpG-island clusters were found within homeobox genes (related to development), in agreement with previous observations [9]. The majority of clusters containing 141 CpG-islands are related to homeobox genes (HOXa,c) and neural development genes (PAX2–5). The number of islands in these genes' loci is 45. Note, that some homeobox genes in single, but long (more than 2Kb) CpG-islands were not considered here. For example, NKX6-2 gene on chromosome 10 resides within a CpG-island more than 3Kb long, which is methylated in majority of tissues. A set of figures underlining CGIs tandem repeat genome structure from Table 2 are based on UCSC browser (www.genome.ucsc.edu) plotting facilities, which is presented in the Supplementary figures file.

In addition to 5 genes encoding for ubiquitously expressed housekeeping genes containing CpG-island clusters in promoter regions, such as splicing factor *PTB1*, myosin *MYO1C*, heparansulphate (membrane protein) *HS3ST3B1*, transcription factor *FAM89B*, histocompatibility gene *HMHA1*, cytochrome P *Cyp26A1_C1* (Table 2), we may note protocadherine gene (*PCDHGA*) containing CG-rich exon cassettes and mediating specific neuron adhesion.

Table 2. Genes associated with CGI clusters

Annotation (Gene class / localization/mediator)	Gene	Chromosome	CGI / length*
Homeobox	*ALX4*	11	4
Promoter	*Cyp26A1_C1*	10	4
Homeobox	*DLX1*	2	5
Homeobox	*EN1*	2	4
Promoter	*FAM89* (AS*)	11	4
Intragenic	*HMHA1*	19	4
Homeobox	*HOXa*	7	13
Homeobox	*HOXc12*	12	4
Membrane peptide	*HS3ST3B1*	17	4
Homeobox	*MSX1*	4	4
Myosin	*MYO1C*	17	4
Homeobox	*NKX2-1*	14	4
Homeobox	*NKX2-3*	10	4
Homeobox	*NKX2-5*	5	5
Homeobox	*NRN1*	6	4
Homeobox (optical.)	*PAX2*	10	4
Homeobox (fetus)	*PAX3*	2	4
Homeobox (CNS)	*PAX5*	9	4
Neural gene	*PCDHGA1*	5	4
Intragenic	*PTB1*	19	4
L1_&_LTR	*tRNA* cluster	1	4(32 361)
SINE	*miRNA* cluster	1	13(40058)
Chr. inversion	*ZNF* (AS*)	10	5(14 472)
Promoter	*EBF3*	10	1(10 527)
Homeobox Telomere (LSAU)	*DUX* cluster	10	6(20 640)
Highly diverged	*piRNA* cluster	15	1?(12 438)
3’ end	*TBX* (AS*)	17	1(10 206)
Homeobox	NKX2-2	20	1(10 782)
Subtelomeric	Low complexity	4	1(13 414)
Homeobox Telomere (LSAU)	*DUX* cluster (D4Z4)	4	9(27 227)
Embryonic TF	*FOXC1*	6	1(11 260)

Homeobox	*UNCX*	7	2(13 699)
Promoter	*TNRC18*	7	1(10 753)
3’ end	*PLEC*	8	1(11 865)
Promoter	*BCOR*	X	1(15 813)
DXZ4	–	X	54(45712)

Note: AS – antisense transcript within the promoter.

* For merged extra long CGIs the length is denoted.

4. Results

Since the discovery and formal defining of CpG islands [1], the important role of these functional elements in vertebrate genomes is increasingly underlined [2]. In particular, most of the housekeeping genes maintain CGI in promoter regions, and are identified as a target by multiple transcription factors (Sp1, CTCF CCCTC-binding factor, etc.) performing the transcription initiation/repression.

It was revealed that about 30 % of CpG-islands are not related to promoters and reside in intragenic and intergenic regions. Evolutionary conservation of these elements shown in a series of works [2] suggests their functional role has not been defined yet. The hypothesis manifesting CpG islands as one of the major scaffolding factors mediating global genome methylation landscape is actively discussed [2]. CGIs frequently reside in regions of imprinting genes (for example, *PPA2C, DNMT1, TSHZ3, CHST8, ZNF225, ZNF229, DMWD, ZNF331, LILRB4, NLRP2, ZIM2, PEG3, MIMT1, USP29, ZIM3, ZNF264, CHMP2A, MZF1*), not only in promoters, but in gene vicinity [2].

The presence of multiple CpG-islands in homeobox genes was reported earlier [9]. Their post embryonic repression is usually exemplified by PcG complex heterochromatization. Certain *HOX/Hox* regions are methylated. In some aggressive tumors the hypomethylation of homeobox genes was observed. Such etiology (hypomethylation of *HOX*, *PAX* genes) following malignant proliferation arises, in part, due to mutation in methyltransferase gene *DNMT3A* [10]. Notably, the observed skew to the 3’ end of a gene and CGI overlap results in a sharp increase of non-synonymous substitution numbers at the 3’ end of a gene. The dicodon preference of CpG location elucidated earlier [9] underlined that most CG dinucleotides were located at codon junctions. Since, quite often, such CGIs are methylated, the cause of multiple missense mutations is CpG methylation. We observed such missense hotspots at the 3’ end for the genes *TBX, PLEC* (Table 2), as well as for the well known Alzheimer disease related *APOE* gene, which maintains a CGI at its 3’ end [11].

The functional potential of extended genome regions with biased nucleotide content such as CpG-islands in the course of evolution could be based on: a) emergence of an open chromatin state in the genomic region free from nucleosome occupancy [12], and, as a rule, transcription in a wide range of the tissues unless hypermethylated; b) CG-rich content is a subject of duplication due to unequal crosssingover; c) the gene switch-off by promoter methylation and the subsequent heterochromatin formation.

We found the multitude of tandem duplications of CpG-islands in the human genome. CG-rich tandem repeats in non-coding regions were reported previously as VNTR macrosatellites [4, 5]. This property allows using macrosatellites as a population-specific marker. The highest variation of CG-rich tandems is in African populations, the lowest is found in Asians [13], according to overall reported genetic diversity of these populations. CpG-islands are used in population genetics bearing the following names [14] (chromosome arms are in parentheses correspondingly): RS447 (chromosome 4p; non CGI), MSR5p (5p), FLJ40296 (13q), RNU2 (17q) and D4Z4 (4q and 10q), as well as X chromosomal DXZ4 and CT47.

4.1. Structural Properties of Macrosatellites 19SST11 and 19SST12 Regions and Gene Cluster Between Them

The largest non-coding tandem GCI clusters are macrosatellites; 19SST11 and 19SST12 found in chromosome 19 (Table 3). Their total length is approximately 1 Mb. The monomer consists of three subsequent segments: the first segment is an interspersed annotated centromeric CG-rich repeat SST1 (67% CG content; 1.2Kb) [14, 15], the next segment is a Simple Repeat (RepeatMasker classification; 700bp) and the last one is CGI (500bp) (See Supplem. Figures S1, S2). The overall length of the monomer is 2.4Kb. Further on we denoted the monomer as SST1_SR_CGI.

KRAB-ZNF gene clusters [16] are located between 19SST11 and 19SST12 and immediately after them (Table 3; Fig. 1) maintaining the monomer to one of flanking satellites as a core promoter.

The putative emergence scheme of the doublet implies non-equal recombination based on these two tandem inverted clusters 19SST11–19SST12 in chromosome region 19q13.12, which had an inverse segment of 1 Mb between the repeats [14].

High similarity of tandems in opposite orientation (95%) endorses this hypothesis. This is also supported by local mosaic synteny in vertebrates described in [15].

Analysis of genome-wide transcriptome data have shown that the SST1 repeat is transcriptionally active (in majority of tissues) at all chromosomes except an obligatory repressed segment of chromosome Y, where short remnants of the repeat are located. The genome loci containing SST1 in most cases have no CpG-islands. The exceptions are three tandem clusters on chromosomes 19 and 4, as well as SST1_CpG promoters of *KRAB-ZNF* genes located between the tandems on chromosome 19 (Fig.1; Table 3).

The evolutionary mode of CpG-island propagation by duplication was not underscored earlier. There are facts confirming its adaptive value for genes. In particular, chromosome 19 has much more gene and segment duplications in comparison to other chromosomes [15]. As we see from our previous example with the 19SST11/KRAB_ZNF/19SST12 trio, the genes containing active CpG-promoters preserved them after duplications. The monomers evolve in a highly conserved manner both in promoters and flanking macrosatellites. In addition to promoter CGIs, chromosome 19 contains a significant number of 3’ end genic CGIs which are methylated in most cases indicating the increase in gene expression.

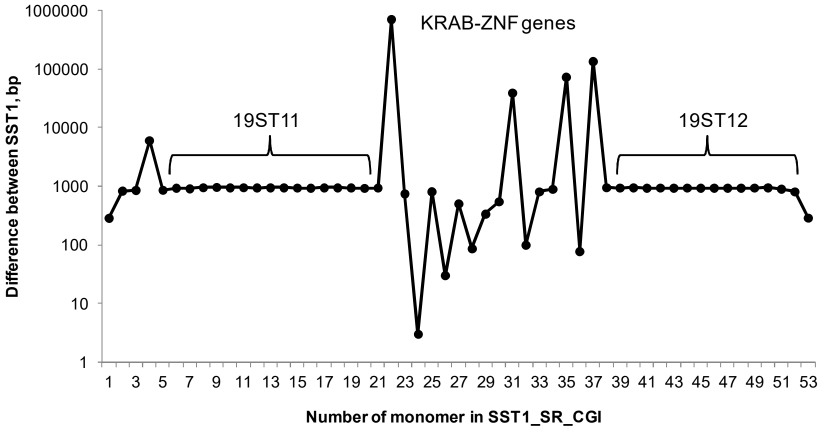

Fig. 1. Distribution of inter-monomer distances SST1_SR_CGI repeats in chromosome 19 macrosatellite doublet 19SST1-19SS12 [4]. Intersatellite region comprises 17 KRAB-ZNF genes with SST1_SR_CGI monomers performing as promoters.

Table 3. List of genes and corresponding CpG islands in promoters located between macrosattellites 19SST11 and 19SST12, and after the tandem. List is sorted by chromosome positioning order.

Cpg_id	CpG-island	Position	Gene
		Macrosatellite 19SST11	
9361	CpG: 56	36 869 564	ZFP14
9362	CpG: 66	36 909 281	ZFP82
9363	CpG: 74	36 912 260	LOC644189
9364	CpG: 44	36 980 190	ZNF566
9365	CpG: 66	37 018 919	ZNF260
9366	CpG: 69	37 063 892	ZNF529
9367	CpG: 86	37 095 680	ZNF382
9367	CpG: 86	37 095 680	ZNF529
9368	CpG: 39	37 157 632	ZNF461
9370	CpG: 53	37 263 381	ZNF850
9372	CpG: 40	37 288 342	LOC284408
9373	CpG: 49	37 328 896	ZNF790
9374	CpG: 26	37 340 918	ZNF345
9374	CpG: 26	37 340 918	ZNF790
9375	CpG: 48	37 406 931	ZNF568
9375	CpG: 48	37 406 931	ZNF829

9377	CpG: 56	37 568 952	ZNF420
Macrosatellite 19SST12*			
9394	CpG: 45	37 825 101	HKR1
9395	CpG: 57	37 861 691	ZNF527
9397	CpG: 57	37 957 726	ZNF569
9398	CpG: 64	37 959 852	ZNF570
9399	CpG: 26	37 997 790	ZNF793
9400	CpG: 66	38 039 561	LOC100507433
9401	CpG: 56	38 085 148	ZNF540
9401	CpG: 56	38 085 148	ZNF571
9402	CpG: 119	38 145 826	ZFP30
9403	CpG: 47	38 182 793	ZNF781
9404	CpG: 50	38 210 107	ZNF607
9405	CpG: 26	38 270 279	ZNF573

*Note: there is set of imperfect tandem gene duplications after the macrosatellite 19SST12.

4.2. Chromosome 19 Early Replication and Transposable Elements Defense Machinery

Chromosome 19 is the most CGI-rich chromosome in the human genome: the CGI density at chr19 is more than twofold higher than the genome average (Fig.2) [15].

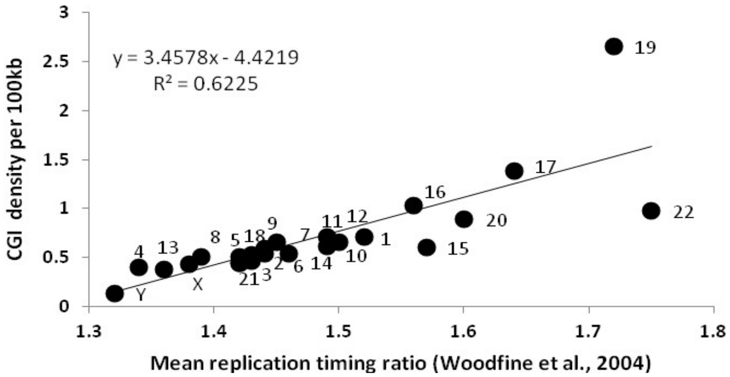

Fig. 2. Mean replication ratio significantly depends on CGI density (P<1.44E-6) in chromosome-wise linear regression analysis.

We analyzed the average replication timing ratio according to data from [17]. We observed strong dependence between CGI density and replication timing (Fig. 2;

P<1.4E-6). Note that chromosome 19 is the earliest to replicate along with the shortest one (chr22). It corroborates the previously reported CGIs to be the open chromatin markers [3], since the majority of origin of replication (ORI) sites reside at transcriptional start sites [18].

One of the reasons for early replication timing is the abundance of KRAB-ZNF genes at this chromosome [17]. It comprises more than two hundred KRAB-ZNF genes organized in clusters [16] and represents the "defense" system against hypomethylated transposable elements early in embryogenesis [19]. It is responsible for identification of non-methylated CpG elements located at retrotransposons. Upon recognition they recruit TET protein, which, in turn, recruits heterochromatin modifiers to repress them, and subsequently changes methylation status [19,20].

4.3. Structural properties of macrosatellites 19SST11 and 19SST12 regions and gene cluster between them

It is worth mentioning that, overall, chromosome 19 is outstanding in maintaining CpG hypomethylated in fetal brain tissues [21]: Hypermethylated/hypomethylated CpG ratio was 0.38 (P<5.9E-23) for chromosome 19, while the closest "over hypomethylated" chromosome 17 maintains it as 0.6 (P<7.5E-14), the whole genome ratio for brain tissues is 0.77 (P<6.4E-54; [21]; Table 1 therein). It implies that intense interplay of hypo/hypermethylated CpG sites during epigenetic reprogramming also makes them a particular target for age related methylation mediated deregulation [21]. The fact underlines that age specific change of the methylation status reported elsewhere [11, 22] may alter/damage the brain function in an ApoE gene harboring large methylated CGI at its 3' end.

The following figures show association between the total number of CGIs and genes across human chromosomes (Fig.3), and between methyltated CGIs and number of correspondent genes by chromosomes (Fig. 4).

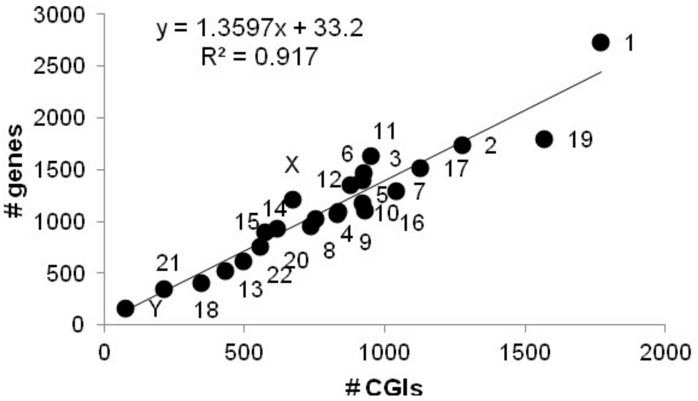

Fig. 3. The total number of CGIs and genes across human chromosomes.

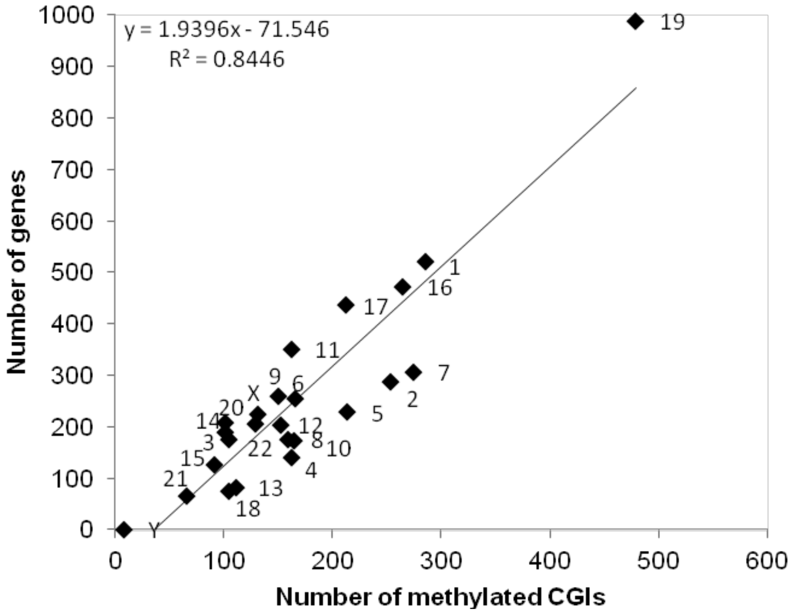

Fig. 4. Correlation between number of methylated CGIs and number of genes across human chromosomes. Chromosome 19 is outstanding in methylated CGIs and corresponding gene numbers.

Notably, *KRAB-ZNF* gene clusters feature unique ambiguous complementation of open chromatin H3K36Me3 histone mark, with repressive H3K9Me3 histone mark attributable to HP1 heterochromatin observed in virtually all adult tissues [23,24], which probably implies its rapid post-embryonic methylation in promoters. Still, we were unable to unequivocally assign the hypermethylated CGIs to *KRAB-ZNF* clusters on chromosome 19, leaving this point as a subject of speculation and further research.

5. Conclusion

Focusing mainly on genome region annotations and identification of general CGIs characteristics, we used a CGI clustering method that is robust relative to the tandem duplication search. Nevertheless the identification of long tandem repetitions with the help of the specialized Tandem Repeats Finder program – TRF [6, 23] shows main macrosatellites/tandem duplications, which we have found taking into account CGI relevance, except for imperfect duplications such as piRNA as a cluster.

We present here several unique cases of multiple duplications of CpG-islands, sometimes merged into superlong CGI (Table 2). It has become clear that the few annotated repetitive elements mediating conservative duplication belong to the SINE types (AluS on sex chromosomes) and to centromere repeats of SST1 on chromosomes

4 and 19. Besides, there are simple CG-rich repeats which have created unique macrosatellite CG island DXZ4 on chromosome X mediating its inactivation, and also macrosatellite D4Z4 (Table 2).

We show that the vast majority of dense CGI clusters are generated by tandem duplications (Table 2). These clusters are connected with genes of early development, including an extended cluster of piRNA (piwiRNA) [24], as well as KRAB-ZNF gene clusters located on chromosome 19.

Recent works discuss the role of methyl-CpG-binding proteins in maintenance and spread of DNA methylation at CpG islands in cancer [25]. Dependence of DNA methylation in human on local topology of CpG sites was shown [26], thus indicating on possible evolutionary role of CGI clusters.

6. Supplementary Material

The manuscript includes supplementary materials (separate file with figures is available at the journal web-site and http://lcg.nsu.ru/belbi).

Acknowledgments. The work was supported by Novosibirsk State University (for ERG, ROB) and ICG SB RAS budget project (0324-2018-0017) (for VNB, AGB, YLO). Authors are grateful to A.O.Bragin, I.V.Chadaeva, G.V.Orlova and N. Beka for the help in manuscript preparation and science discussion.

References

1. Gardiner-Garden, M., Frommer, M.: CpG islands in vertebrate genomes. J. Mol. Biol., Vol. 196, 261-282. (1987)
2. Deaton, A.M., Bird, A.: CpG islands and the regulation of transcription. Gen. Dev., Vol. 25, No. 10, 1010-1022. (2011)
3. Babenko, V., Chadaeva, I., Orlov, Y.: Genomic landscape of CpG rich elements in human genome. BMC evolutionary biology, Vol. 17(Suppl 1), 19. (2017)
4. Chadwick B.P.: DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. Gen. Res., Vol. 18, 1259-1269. (2008)
5. Illingworth, R.S., Bird, A.P.: CpG islands – 'a rough guide'. FEBS Lett., Vol. 583, No. 11, 1713-20. (2009)
6. Gelfand, Y, Rodriguez, A, Benson, G.: TRDB–the Tandem Repeats Database. Nucleic Acids Res., Vol. 35(Database issue), D80–87. (2007)
7. Babenko, V.N., Kosarev, P.S., Vishnevsky, O.V., Levitsky, V.G., Basin, V.V., Frolov, A.S.: Investigating extended regulatory regions of genomic DNA sequences. Bioinformatics, Vol. 15, No. 7-8, 644-53. (1999)
8. Orlov, Y.L., Te Boekhorst, R., Abnizova, I.I.: Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. J Bioinform Comput Biol., Vol. 4, 523-36. (2006)
9. Branciamore, S., Chen, Z.X., Riggs, A.D., Rodin, S.N.: CpG island clusters and pro-epigenetic selection for CpGs in protein-coding exons of HOX and other transcription factors. Proc. Natl Acad. Sci. USA, Vol. 107, No. 35, 15485-15490. (2010)

10. Qu, Y., Lennartsson, A., Gaidzik, V.I., Deneberg, S., Karimi, M., Bengtzén, S., Höglund, M., Bullinger, L., Döhner, K., Lehmann, S. Differential methylation in CN-AML preferentially targets non-CGI regions and is dictated by DNMT3A mutational status and associated with predominant hypomethylation of HOX genes. Epigenetics, Vol. 9, No. 8, 1108-1119. (2014)
11. Ma, Y., Smith, C.E., Lai, C.Q., Irvin, M.R,, Parnell, L.D., Lee, Y.C., Pham, L., Aslibekyan S., Claas, S.A., Tsai, M.Y., Borecki, I.B., Kabagambe, E.K., Berciano, S., Ordovás, J.M., Absher D.M., Arnett, D.K.: Genetic variants modify the effect of age on APOE methylation in the Genetics of Lipid Lowering Drugs and Diet Network study. Aging Cell, Vol. 14, No.1, 49-59. (2015)
12. Goh, W.S., Orlov, Y., Li, J., Clarke, N.D.: Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. PLoS Comput Biol., Vol. 6, No. 1, e1000649. (2010)
13. Schaap, M., Lemmers, R.J., Maassen, R., van der Vliet, P.J., Hoogerheide, L.F., van Dijk, H.K., Baştürk, N., de Knijff, P., van der Maarel, S.M.: Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. BMC Genomics, Vol. 4, No. 14, 143. (2013)
14. Tremblay, D.C., Alexander, G. Jr., Moseley, S., Chadwick, B.P.: Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. BMC Genomics, Vol. 15, No. 11, 632. (2010)
15. Grimwood, J., Gordon, L.A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., … Stubbs, L., Rokhsar, D.S., Myers, R.M., Rubin, E.M., Lucas, S.M.: The DNA sequence and biology of human chromosome 19. Nature, Vol. 428, No. 6982, 529-535. (2004)
16. Vogel, M.J., Guelen, L., de Wit, E., Peric-Hupkes, D., Lodén, M. et al.: Human heterochromatin proteins form large domains containing KRAB-ZNF genes. Genome Res., Vol. 16, No. 12, 1493-1504. (2006)
17. Woodfine, K., Fiegler, H., Beare, D.M., Collins, J.E., McCann, O.T., Young, B.D., Debernardi, S., Mott, R., Dunham, I., Carter, N.P.: Replication timing of the human genome. Hum Mol Genet., Vol. 13. No. 2, 191-202. (2004)
18. Langley, A.R., Gräf, S., Smith, J.C., Krude, T.: Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). Nucleic Acids Res. Vol. 44. No. 21, 10230-10247. (2016)
19. Long, H.K., Blackledge, N.P., Klose, R.J.: ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. Biochem Soc Trans., Vol. 41, No. 3, 727-740. (2013)
20. Imbeault, M., Trono, D.: As time goes by: KRABs evolve to KAP endogenous retroelements. Dev Cell., Vol. 31, No. 3, 257-258. (2014)
21. Spiers, H., Hannon, E., Schalkwyk, L.C., Smith, R., Wong, C.C. et al.: Methylomic trajectories across human fetal brain development. Genome Res., Vol. 25, No. 3, 338-352. (2015)
22. Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., Thurman, R.E., Kaul, R., Myers, R.M., Stamatoyannopoulos, J.A.: Widespread plasticity in CTCF occupancy linked to DNA methylation. Gen. Res., Vol. 22, No. 9, 1680-1688. (2012)
23. Warburton, P.E., Hasson, D., Guillem, F., Lescale, C., Jin, X., Abrusan, G.: Analysis of the largest tandemly repeated DNA families in the human genome. BMC Genomics, Vol. 9, No. 533. (2008)
24. Williams, Z., Morozov, P., Mihailovic, A., Lin, C., Puvvula, P.K., Juranek, S., Rosenwaks, Z., Tuschl, T.: Discovery and characterization of piRNAs in the human fetal ovary. Cell Rep., Vol. 13, No. 4, 854-863. (2015)
25. Stirzaker, C., Song, J.Z., Ng, W., Du, Q., Armstrong, N.J., Locke, W.J., Statham, A.L., French, H., Pidsley, R., Valdes-Mora, F., Zotenko, E., Clark, S.J.:Methyl-CpG-binding protein MBD2 plays a key role in maintenance and spread of DNA methylation at CpG islands and shores in cancer. Oncogene, Vol. 36, No. 10, :1328-1338. (2017)

26. Lövkvist, C., Dodd, I.B., Sneppen, K., Haerter, J.O.: DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic Acids Res., Vol. 44, No. 11, 5123-32. (2016)

Vladimir N Babenko, PhD, is a senior scientist at the Institute of Cytology and Genetics SB RAS and Novosibirsk State University, Novosibirsk, Russia. Professional interests: bioinformatics, genomics, statistical genetics, human genetics, programming. He is the author and co-author of numerous journal and conference papers, participated in several national and international research projects. Authored and co-authored about 100 papers.

Anton G Bogomolov is a junior scientist working in bioinformatics and bioimaging analysis at the Institute of Cytology and Genetics SB RAS and Novosibirsk State University, Novosibirsk, Russia. He is the author and co-author of numerous journal and conference papers, participated in several national research projects in Russia.

Roman O Babenko is a master student of mathematical department at Novosibirsk State University, Russia. Professional interests: bioinformatics, web development, programming.

Elvira R Galieva, PhD, is a senior scientist at Novosibirsk State University, Russia, working also as a teaching assistant with undergraduate students. Professional interests: human genetics, microscopy, DNA analysis. She is the author and co-author of numerous journal and conference papers.

Yuriy L Orlov, Professor of the Russian Academy of Sciences, is a senior scientist at the Institute of Cytology and Genetics SB RAS, and Head of the Computer Genomics Laboratory, Life Sciences Department, Novosibirsk State University, Novosibirsk, Russia lecturing in the field of mathematical methods in biology. He is Leading Scientist, at the A.O. Kovalevsky Institute of Marine Biological Research of the RAS, Sevastopol, Russia. He authored and co-authored over 200 works presented and published at international and national conferences, symposia and journals in the field of bioinformatics. He has organized a number of scientific and professional projects and studies supported by Russian and international grant agencies. He is guest editor and editorial board member in several journals series (BioMed Central, JBCB, PeerJ, Frontiers in Genetics), as well as a member of program committees of scientific conferences (BGRS - Bioinformatics of Genome Regulation and Structure \ Systems Biology conference series). His primary areas of interest are: bioinformatics, genomics, neuroinformatics, and databases.

Received: May 23, 2017; Accepted: April 30, 2018.