Genomics Data 5 (2015) 89–93

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Genome-wide mapping of hot spots of DNA double-strand breaks in human cells as a tool for epigenetic studies and cancer genomics

N.A. Tchurikov *, O.V. Kretova, D.M. Fedoseeva, V.R. Chechetkin, M.A. Gorbacheva, A.V. Snezhkina, I.R. Alembekov, G.I. Kravatskaya, Y.V. Kravatsky

Engelhardt Institute of Molecular Biology, Moscow, Russia

ARTICLE INFO

Article history:
Received 15 May 2015
Accepted 24 May 2015
Available online 30 May 2015

Keywords:
Double-strand breaks
Fragile sites
H3K4me3 marks
Bioinformatics
HEK293T

ABSTRACT

Hot spots of DNA double-strand breaks (DSBs) are associated with coordinated expression of genes in chromosomal domains (Tchurikov et al., 2011 [1]; 2013). These 50–150-kb DNA domains (denoted "forum domains") can be visualized by separation of undigested chromosomal DNA in pulsed-field agarose gels (Tchurikov et al., 1988; 1992) and used for genome-wide mapping of the DSBs that produce them. Recently, we described nine hot spots of DSBs in human rDNA genes and observed that, in rDNA units, the hot spots coincide with CTCF binding sites and H3K4me3 marks (Tchurikov et al., 2014), suggesting a role for DSBs in active transcription. Here we have used Illumina sequencing to map DSBs in chromosomes of human HEK293T cells, and describe in detail the experimental design and bioinformatics analysis of the data deposited in the Gene Expression Omnibus with accession number GSE53811 and associated with the study published in DNA Research (Kravatsky et al., 2015). Our data indicate that H3K4me3 marks often coincide with hot spots of DSBs in HEK293T cells and that the mapping of these hot spots is important for cancer genomic studies.

Specifications	
Organism/cell line/tissue	*Homo sapiens*/HEK293T cells
Sex	Female
Sequencer or array type	Illumina Genome Analyzer IIx, Illumina MiSeq
Data format	Raw and processed. Raw data: FASTQ reads. Processed data: BED, WIG, SGR and text table files. Metadata in SOFT and MINiML formats are supplied by GEO for automated processing.
Experimental factors	HEK293T cells were seeded in 10-cm culture plates 1–2 days before experiments in DMEM containing 10% FBS, and were used at approximately 60–80% confluency.
Experimental features	DNA domains, migrating in 0.8% agarose mini-gels from the DNA-agarose plugs, were electroeluted. Biotinylated oligonucleotides were ligated to DNA sequences at DSB sites.
Consent	Level of consent allowed for reuse if applicable (typically for human samples).
Sample source location	Moscow 119334, Russia

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53811.

2. Experimental design, materials and methods

2.1. DNA preparation

The steps of the procedure are shown schematically in Fig. 1A. To reduce non-specific hydrodynamic breakage, DNA samples were isolated from cells embedded in 0.5% low-melt agarose as described previously [3,4,7–9]. About 6 million HEK293T cells in 2 mL of culture medium were pelleted by centrifugation at 2000 rpm in a Minispin centrifuge (Eppendorf), resuspended in 0.3 mL of the same medium, gently mixed at 42 °C with an equal volume of a solution of 1% low-melt agarose L (LKB) in PBS, and distributed on a mold containing 100-μL wells. The mold was covered with Parafilm and placed on ice for 2–5 min. The agarose plugs were then placed in Petri dishes with 5 mL of a solution containing 0.5 M EDTA (pH 9.5), 1% sodium lauroylsarcosine, and 1–2 mg of proteinase K per mL for 40–48 h at 50 °C, and stored at 4 °C in the same solution for 3 months. Each DNA–agarose plug usually contained about 15 μg of DNA, corresponding to about 1 million cells.

To test the quality of isolated DNA, fractionation in pulsed-field gels was performed as described previously [1,3,4]. Portions of the original

* Corresponding author at: Engelhardt Institute of Molecular Biology, Vavilov str. 32, Moscow, 119334, Russia. Tel.: +7 499 135 97 53; fax: +7 499 135 14 05.
E-mail address: tchurikov@eimb.ru (N.A. Tchurikov).

http://dx.doi.org/10.1016/j.gdata.2015.05.018

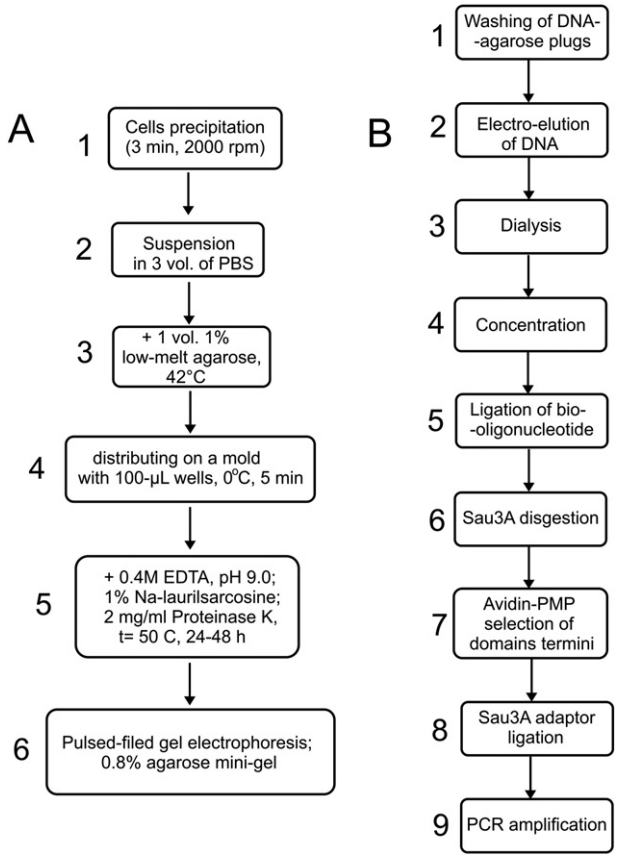

Fig. 1. Schematic representation of the procedures used for isolation of DNA samples inside 0.5% low-melt agarose (A) and the major steps of the RAFT procedure (B).

agarose–DNA plugs (5–50 μL) containing 1–10 μg of DNA were used for electrophoresis without any restriction enzyme digestion. The DNA samples were run in 0.8% agarose gels on a Pulsaphor system (LKB) using a hexagonal electrode and switching times of 25 or 100 s.

For elution of DNA preparations, fractionation in a 0.8% agarose conventional mini-gel was performed. One-half of the DNA–agarose plug was washed in 1× TE three times (for 15 min each), followed by washing (three times) in the same solution containing 17.4 μg/mL phenylmethylsulfonyl fluoride (PMSF) in ethanol. After fractionation in the mini-gel, the ethidium bromide-stained DNA band was excised and electroeluted inside a cellulose-membrane dialysis bag. After overnight dialysis without stirring against 1 L of 0.01× TE at 4 °C, the DNA was concentrated with PEG at 4 °C.

2.2. Rapid amplification of forum domains termini (RAFT) procedure

The steps of the procedure are shown schematically in Fig. 1B. About 1.5 μg of isolated DNA (see above) was ligated with 70 ng of double-stranded oligonucleotide (25-bp long 5′-phosphorylated 5′ pCCCCTG CAGTATAAGGAGAATTCGGG 3′ oligonucleotide annealed to a 26-bp long 5′ biotinylated 5′ bio-CCGAATTCTCCTTATACTGCAGGGG 3′ oligonucleotide) in 150 μL of a solution containing 0.1 M NaCl, 50 mM Tris–HCl (pH 7.4), 8 mM $MgCl_2$, 9 mM 2-mercaptoethanol, 7 μM ATP, 7.5% PEG, and 40 units of T4 DNA ligase at 20 °C for 16 h. After heating at 65 °C for 10 min, the DNA preparation was digested with Sau3A enzyme to shorten the forum domain to the positions of the termini attached to the ligated oligonucleotide. The selection of such termini was performed in 0.5-mL Eppendorf tubes using 300 μL of a suspension containing Streptavidin Magnesphere Paramagnetic Particles, (SA-PMP; Promega) according to the manufacturer's recommendations. After extensive washing with 0.5× SSC to remove DNA fragments corresponding to the internal parts of forum domains, the forum termini (FT) DNA preparation was eluted from the SA-PMP using digestion with EcoRI enzyme in a final volume of 50 μL (double-stranded FT). The FT were then ligated with 100× molar excess of double-stranded Sau3A adaptor (5′-phosphorylated 5′ pGATCGTTTGCGGCCGCTTAAGCTTGGG 3′ oligonucleotide annealed to 5′ CCCAAGCTTAAGCGGCCGCAAAC 3′ oligonucleotide). In some experiments, the DNA preparation was eluted from the SA-PMP by incubation at 100 °C for 3 min in 50 μL of 0.01× TE (single-stranded FT). Before heating, the FT preparation was ligated with a 100-fold molar excess of double-stranded Sau3A adaptor in suspension with SA-PMP (see above). Both final DNA samples (double-stranded FT or single-stranded FT) were used for PCR amplifications. PCR amplification (15–20 cycles) in 30 μL of a solution containing 67 mM Tris–HCl (pH 8.4), 6 mM $MgCl_2$, 10 mM 2-mercaptoethanol, 16.6 mM ammonium sulfate, 6.7 μM EDTA, 5 μg/mL BSA, 1 mM dNTPs, 1 μg of primer corresponding to Sau3A adaptor (5′ CCCAAGCTTAAGCGGCCGCAAAC 3′), 1 μg of primer corresponding to biotinylated oligonucleotide (5′ CCGAATTCTCCTTATACTGCAGGGG 3′), and 1 U of Taq polymerase was performed using a Mastercycler Personal thermal cycler (Eppendorf). Amplification conditions were 90 °C for melting, 65 °C for annealing, and 72 °C for extension, for 1 min each. The final DNA sample contained the amplified genome-wide preparation of DNA fragments delimited by a base at a particular DSB and the nearest Sau 3A site.

2.3. Library preparation

Libraries were prepared according to Illumina's instructions accompanying the DNA Sample Kit (Part # 0801-0303). Briefly, DNA was end-repaired using a combination of T4 DNA polymerase, *Escherichia coli* DNA Pol I large fragment (Klenow polymerase), and T4 polynucleotide kinase. The blunt, phosphorylated ends were treated with Klenow fragment and dATP to yield a protruding 3′-A base for ligation of Illumina's adapters, which have a single T-base overhang at the 3′ end. After adapter ligation, DNA was PCR amplified with Illumina primers for 15 cycles. Library fragments of ~200–400 bp and ~400–1200 bp were isolated as bands from an agarose gel, and were sequenced on the Genome Analyzer IIx and MiSeq, respectively, following the manufacturer's protocols.

2.4. Data processing

Fig. 2 shows the bioinformatics pipeline used. The standard Illumina analysis pipeline using their phiX control software was used for base calling. At the first step of processing, quality control was performed using FastQC [10]. Next, reads were trimmed for RAFT primer sequences by use of Cutadapt v. 1.3 [11]. Some options were common for both datasets:

–minimum-length = 30 –trimmed-only –quality-base = 33 –quality-cutoff = 3 -n 2

The option "–trimmed-only" was used to remove from trimmed files all reads that did not have RAFT primers. This option setting ensures that after removal of primers the data set consists of reads of sufficient length to have contained RAFT primers before removal. The following options were applied to remove 5' attached RAFT primers from reads:

–g CCCAAGCTTAAGCGGCCGCAAAC

–g CCGAATTCTCCTTATACTGCAGGGG.

Cutadapt was used in the paired-end mode for the paired-end Illumina GA IIx dataset and in the single-end mode for the single-end MiSeq dataset. At the next step, the trimmed files from both sequencing machines were merged.

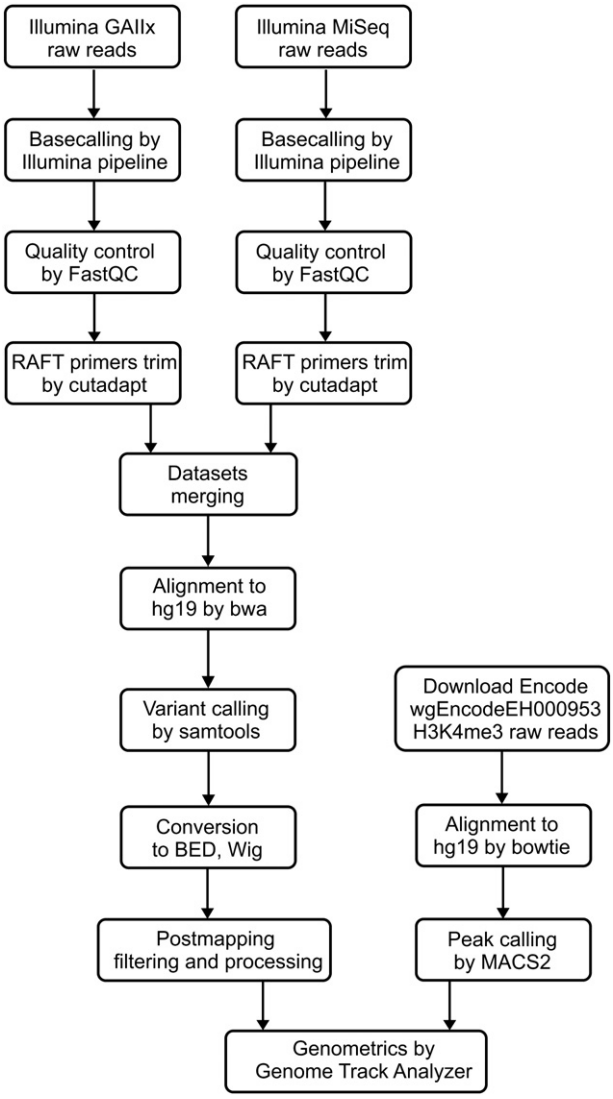

Fig. 2. Bioinformatics pipelines.

Trimmed reads were mapped to hg19/GRCh37p10 in paired-end mode using bwa 0.7.5a [12] and the mem algorithm, and by SAMtools 0.1.19 [13]. Variant calling was also performed using SAMtools. Final mappings were converted for further analysis into tables and formats, including BED and WIG, using ad hoc Perl scripts. Post-mapping filtering was performed as follows. First, all mappings that did not contain a Sau3A recognition sequence (GATC) or contained two or more such sequences (as a result of partial digestion) were removed as erroneous. Second, a coordinate for each DSB at the end opposite the Sau3A sequence was calculated. Next, all mappings that were mapped with coverage below 5% were removed. Finally, all mappings within 1 kb of each other were merged into groups, and the maximum coordinate and coverage value of the group replaced those of the individual mappings. The resulting SGR file contains the DSBs with one-nucleotide resolution and their coverage.

To prepare the H3K4me3 mark dataset, the following steps were performed. The downloaded raw reads for Rep1, Rep2, and Signal from Encode accession wgEncodeEH000953 were aligned to the same genome hg19/GRCh37p10 by use of bowtie v.1 with the following command line options: –best -m 1 –chunkmbs 1024. Peak calling was performed using the MACS2 [14] peak caller with the options callpeak –gsize hs to set the correct genome size. Peak summits obtained from MACS2 were used for further analysis. The genometric analysis of both datasets was performed using Genome Track Analyzer [6].

3. Discussion

The RAFT procedure includes several steps in which very long DNA molecules are manipulated in solution—from the elution of DNA domains to the ligation of biotinylated oligonucleotides (steps 2–5 in Fig. 1B). Although only a gentle mixing of solution was performed, a random fragmentation of forum domains cannot be excluded during these steps. Nevertheless, our previous data strongly demonstrate that the level of this random hydrodynamic fragmentation of DNA molecules in the conditions used is much lower than the non-random fragmentation detected at the hot spots of DSBs [5].

The data on the distribution of hot spots of DSBs in the human genome could be used for the study of chromosomal breakage associated with regulation of gene expression and different genomic rearrangements (translocations, inversions, and deletions).

We studied the positional and ordering correlations between DSBs and H3K4me3 marks in the chromosomes of human HEK293T cells using Genome Track Analyzer [6]. The H3K4me3 mark is a well-known promoter-specific histone modification that is associated with transcription and active genes. This epigenetic mark selectively directs global TFIID recruitment to active genes, some of which are also p53 targets [15]. The summary of correlations is shown in Table 1 and demonstrates strong positional correlations between DSBs and H3K4me3 peak summits for all chromosomes of H293T cells. Such correlations support the hypothesis regarding the relationships between DSBs and coordinated gene expression [2]. Interesting questions arise from the ordering correlations, which are significant only for chromosomes 2, 3, 18, and 19 and show that in these chromosomes H3K4me3 peak summits often precede DSBs. In future work we plan to analyze the significant correlation pairs for these chromosomes in different genome browsers and automatic annotation systems to reveal the possible biological meaning of these correlations. The strong correlation between H3K4me3 marks and hot spots of DSBs has been described in human rDNA units, suggesting an important role for DNA breaks in actively transcribed genes [5].

Fig. 3 shows one example of DSB mapping important for cancer genomic studies. The BAM file was used in locating hot spots of DSBs inside the TMPRSS2 and ERG genes located on the minus strand of chr21 at a distance about 3 Mb. These genes were selected because recurrent gene fusions between TMPRSS2 and ETS family genes occur at high frequency in prostate cancer [16]. We detected several regions in the TMPRSS2 and ERG genes that are enriched with DSBs. Deletion, rather than translocation, is reported to be the main mechanism for TMPRSS2-ERG gene fusion (81 vs. 19%) [16]. Detected hot spots of DSBs (Fig. 3) could be involved in such genomic rearrangements. It has been shown that the regions possessing hot spots of DSBs in human rDNA genes often form contacts with other genomic regions also possessing hot spots of DSBs, and it has been suggested that this fact could explain the origin of Robertsonian translocations [5]. It is known that regions of the same chromosome make 3D contacts more often than between different chromosomes [17]. TMPRSS2 and ERG genes are located very close to each other on chr21, providing a potential for contacts between their regions possessing DSB hot spots. Currently, we are performing 4C (circular chromosome conformation capture) experiments in order to study genomic contacts between these genes, to uncover the possible mechanism of this and some other cancerogenic gene fusions.

Our data suggest that hot spots of DSBs are associated with various epigenetic mechanisms of gene regulation and with the formation of 3D chromosomal structures, both of which are conserved in different cell types, with dramatic consequences for genomic integrity should they go awry [2,5]. Hence, data on the distribution of DSB hot spots in the human genome provide a new tool for studies of cancer genomics and genomic features associated with the regulation of gene expression.

Table 1
Correlation of the data on mapping of DSBs and H3K4me3 marks in HEK293T cells.

Chromosome	z	Correlation	p	z_p	Ordering	p_z
chr1	11.142	strong corr	0.0000	-1.6450		0.1000
chr2	3.218	strong corr	0.0013	-3.0600	2⇒1	0.0022
chr3	6.001	strong corr	0.0000	-2.0400	2⇒1	0.0414
chr4	3.501	strong corr	0.0005	-1.0700		0.2845
chr5	4.925	strong corr	0.0000	0.6430		0.5205
chr6	5.856	strong corr	0.0000	1.3680		0.1712
chr7	5.461	strong corr	0.0000	-1.7190		0.0857
chr8	3.864	strong corr	0.0001	-0.6400		0.5224
chr9	5.439	strong corr	0.0000	0.2440		0.8075
chr10	3.818	strong corr	0.0001	-0.2390		0.8110
chr11	4.857	strong corr	0.0000	0.2150		0.8295
chr12	4.160	strong corr	0.0000	-0.7700		0.4411
chr13	2.395	corr	0.0166	-0.1390		0.8897
chr14	3.096	strong corr	0.0020	-0.4870		0.6260
chr15	2.989	strong corr	0.0028	-1.7060		0.0881
chr16	5.115	strong corr	0.0000	-0.7840		0.4328
chr17	2.701	strong corr	0.0069	-0.1660		0.8685
chr18	2.222	corr	0.0263	-2.3430	2⇒1	0.0191
chr19	5.534	strong corr	0.0000	-2.8680	2⇒1	0.0041
chr20	3.995	strong corr	0.0001	1.0110		0.3122
chr21	2.748	strong corr	0.0060	-1.2790		0.2008
chr22	2.651	strong corr	0.0080	-0.3760		0.7066
chrX	3.085	strong corr	0.0020	-0.6310		0.5282
Genome-wide	29.000	strong corr	0.0000	-3.8930	2⇒1	0.0001

z and z_p are calculated by Genome Track Analyzer [6] and characterize the positional and ordering correlations between DSBs and H3K4me3 peak summits. The 1% significance thresholds for $|z|$ and $|z_p|$ in the case of random correlations correspond to 2.58, while 5% significance thresholds correspond to 1.96. The negative values of z_p indicate that H3K4me3 mark peak summits precede DSBs for some chromosomes (2, 3, 18, 19). The corresponding p-values were calculated using Gaussian statistics. All data have number of pairs of the nearest neighbors (NN) exceeding 50 to ensure the applicability of Gaussian statistics.

Color and values legend	
***z* -value**	**Meaning**
$\|z\| < 1.8$	Insignificant, no correlation
$1.8 \le \|z\| < 1.96$	Fuzzy correlation
$1.96 \le \|z\| < 2.58$	Significant correlation, $p < 0.05$
$\|z\| \ge 2.58$	Strong correlation, $p < 0.01$

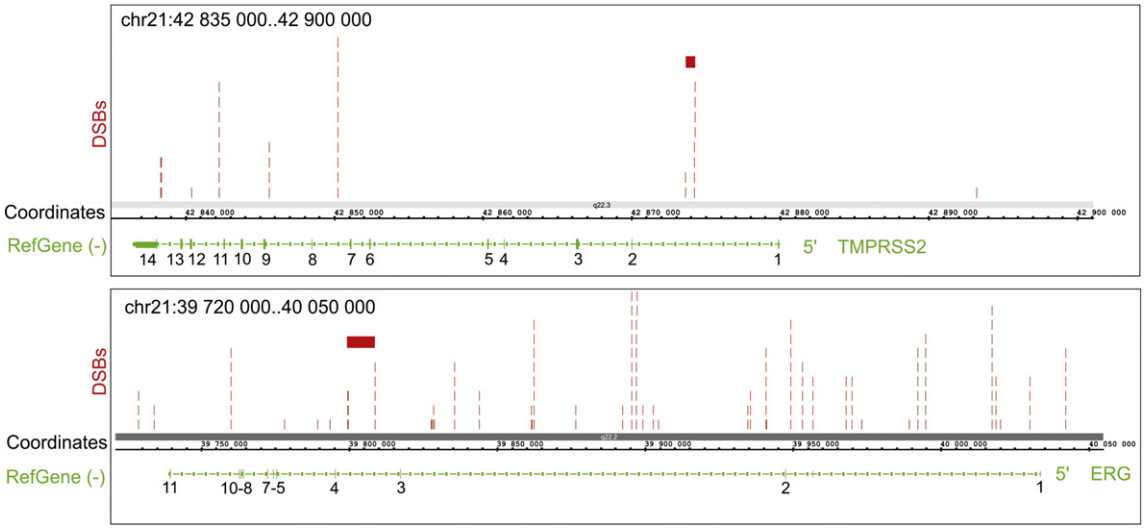

Fig. 3. The mapping of hot spots of DSBs inside TMPRSS2 and ERG genes. The mapping results using the BAM file are shown using IGB Browser on Human Feb. 2009 (GRCh37/hg19) Assembly. The values at genes indicate exons numbers. The red bars indicate the regions that are involved very often in fusion variant possessing exon 1 from TMPRSS2 and exons 4–11 from ERG in prostate cancer [16].

Acknowledgments

This work was supported by a grant from the Russian Science Foundation (project no. 15-14-00005).

References

[1] N.A. Tchurikov, O.V. Kretova, D.V. Sosin, I.A. Zykov, I.F. Zhimulev, Y.V. Kravatsky, Genome-wide profiling of forum domains in *Drosophila melanogaster*. Nucleic Acids Res. 39 (2011) 3667–3685.
[2] N.A. Tchurikov, O.V. Kretova, D.M. Fedoseeva, D.V. Sosin, S.A. Grachev, M.V. Serebraykova, S.A. Romanenko, N.V. Vorobieva, Y.V. Kravatsky, DNA double-strand breaks coupled with PARP1 and HNRNPA2B1 binding sites flank coordinately expressed domains in human chromosomes. PLoS Genet. 9 (4) (2013) e1003429.
[3] N.A. Tchurikov, N.A. Ponomarenko, L.G. Airich, Isolation of forum DNA—a specific fraction in human DNA. Dokl. Akad. Nauk SSSR 303 (1988) 491–497.
[4] N.A. Tchurikov, N.A. Ponomarenko, Detection of DNA domains in *Drosophila*, human and plant chromosomes possessing mainly 50- to 150-kilobase stretches of DNA. Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 6751–6755.
[5] N.A. Tchurikov, D.M. Fedoseeva, D.V. Sosin, A.V. Snezhkina, N.V. Melnikova, A.V. Kudryavtseva, Y.V. Kravatsky, O.V. Kretova, Hot spots of DNA double-strand breaks and genomic contacts of human rDNA units are involved in epigenetic regulation. J. Mol. Cell Biol. (Oct 3 2014) (pii: mju038. [Epub ahead of print] PMID: 25280477).
[6] Y.V. Kravatsky, V.R. Chechetkin, N.A. Tchurikov, G.I. Kravatskaya, Genome-wide study of correlations between genomic features and their relationship with the regulation of gene expression. DNA Res. 22 (2015) 109–119.
[7] N.A. Tchurikov, A.N. Krasnov, N.A. Ponomarenko, Y.B. Golova, B.K. Chernov, Forum domain in *Drosophila melanogaster* cut locus possesses looped domains inside. Nucleic Acids Res. 26 (1998) 3221–3227.
[8] N.A. Tchurikov, O.V. Kretova, B.K. Chernov, Y.B. Golova, I.F. Zhimulev, I.A. Zykov, SuUR protein binds to the boundary regions separating forum domains in *Drosophila melanogaster*. J. Biol. Chem. 279 (2004) 11705–11710.
[9] N.A. Tchurikov, O.V. Kretova, D.M. Fedoseeva, V.R. Chechetkin, M.A. Gorbacheva, A.A. Karnaukhov, G.I. Kravatskaya, Y.V. Kravatsky, Mapping of genomic double-strand breaks by ligation of biotinylated oligonucleotides to forum domains: analysis of the data obtained for human rDNA units. Genomics Data 3 (2015) 15–18.
[10] S. Andrews, FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
[11] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, no. 1, http://dx.doi.org/10.14806/ej.17.1.200.
[12] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26 (2010) 589–595.
[13] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map (SAM) format and SAMtools. Bioinformatics 25 (2009) 2078–2209.
[14] Zhang, et al., Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9 (2008) R137.
[15] S.M. Lauberth, T. Nakayama, X. Wu, et al., H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. Cell 152 (2013) 1021–1036.
[16] J.J. Tu, S. Rohan, J. Kao, N. Kitabayashi, S. Mathew, Y.T. Chen, Gene fusions between TMPRSS2 and ETS family genes in prostate cancer: frequency and transcript variant analysis by RT-PCR and FISH on paraffin-embedded tissues. Mod. Pathol. 20 (2007) 921–928.
[17] J. Dekker, K. Rippe, M. Dekker, et al., Capturing chromosome conformation. Science 95 (2002) 1306–1311.