Gene 498 (2012) 75–80

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Methods paper

nMETR: Technique for facile recovery of hypomethylation genomic tags

Konstantin Baskaev [a], Andrew Garazha [a], Nurshat Gaifullin [b], Maria V. Suntsova [a], Anastassia A. Zabolotneva [a], Anton A. Buzdin [a,*]

[a] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, 16/10 Miklukho-Maklaya, Moscow 117997, Russia
[b] Lomonosov Moscow State University, Lomonosovsky Dr., -5, Moscow 119192, Russia

ARTICLE INFO

Article history:
Accepted 30 January 2012
Available online 13 February 2012

Keywords:
Epigenetics
Methylation
Genome wide analysis
Genomic repeats
Next generation sequencing
Alu retrotransposon

ABSTRACT

Genome-wide methylation studies frequently lack adequate controls to estimate proportions of background reads in the resulting datasets. To generate appropriate control pools, we developed technique termed nMETR (non-methylated tag recovery) based on digestion of genomic DNA with methylation-sensitive restriction enzyme, ligation of adapter oligonucleotide and PCR amplification of non-methylated sites associated with genomic repetitive elements. The protocol takes only two working days to generate amplicons for deep sequencing. We applied nMETR for human DNA using *BspFNI* enzyme and retrotransposon *Alu*-specific primers. 454-sequencing enabled identification of 1113 nMETR tag sites, of them ~65% were parts of CpG islands. Representation of reads inversely correlated with methylation levels, thus confirming nMETR fidelity. We created software that eliminates background reads and enables to map and annotate individual tags on human genome. nMETR tags may serve as the controls for large-scale epigenetic studies and for identifying unmethylated transposable elements located close to genomic CpG islands.

1. Introduction

Methylation of eukaryotic DNA is one of the most important mechanisms governing gene expression and chromatin structure. Assays for DNA methylation are essential for studies of epigenetic mechanisms mediating many aspects of gene expression regulation. Systemic changes of methylation profiles are characteristic for numerous diseases including cancer (Jeronimo et al., 2011) and autoimmune syndromes (Hirst and Marra, 2009). In vertebrate DNA, methylation mostly deals with cytosine residues within the CG dinucleotides, although recent indications suggest that in some mammalian tissues also non-CG cytosine methylation may be functionally significant (Chen et al., 2011).

Methylated or unmethylated state of cytosine residues may attract specific protein complexes mediating their biological functions (Ballestar, 2011). Generally, heavily methylated DNA is associated with gene silencing and chromatin compaction, whereas unmethylated DNA marks active chromatin domains (Ballestar, 2011). Approx. 40% of mammalian genes include in their 5′-terminal parts CG-rich regulatory sequences termed “CpG islands” (Fatemi et al., 2005). The usual formal definition of a CpG island is a region with at least 200 bp and with a GC percentage that is greater than 50% and with an observed/expected CpG ratio that is greater than 60%, where the value of expected CpG is calculated by formula (GC content/2) (Gardiner-Garden and Frommer, 1987). More recently, it has been reported that ~60% of human gene promoters are associated with CpG islands (Bernstein et al., 2007). Moreover, the proportion of human promoters enriched in CG dinucleotides is even higher (72%) (Saxonov et al., 2006). CpG islands themselves may be either associated with known genes or standing alone in genomic sequence (Bernstein et al., 2007), either unique or even incorporated in genomic transposable elements (Bantysh and Buzdin, 2009). In contrast, nonfunctional genomic regions are generally depleted in CG dinucleotides (Bernstein et al., 2007). Changes in methylation states of CpG islands may switch gene activity by modulating their accessibility to transcription factors.

Modern techniques for genome-wide methylation studies may be based on bisulfite conversion of DNA either followed by next generation sequencing (NGS), or by interrogating converted DNA with microarrays (Fazzari and Greally, 2010). Alternatively, methylated DNA may be isolated by using affinity chromatography with reagents binding methylated cytosines (Fisher et al., 2004). Finally, utilizing methylation-sensitive restriction enzymes is considered a method of choice for many applications (Ogoshi et al., 2011). However, thorough analysis of large databanks obtained in such ways raises a question about adequate controls that would permit one to estimate the impacts of false-positive and false-negative sequences in the libraries (Pelizzola and Ecker, 2011; Rauch et al., 2009). An ideal control dataset would meet the following criteria: (*i*) it should be big enough to support genome-wide analyses; (*ii*) it should provide information about many

Abbreviations: BLAT, BLAST-like alignment tool; C, cytidine; G, guanosine; MRE, methylation-sensitive restriction endonuclease; NCBI, National Center for Biotechnology Information; NGS, next generation sequencing; nMETR, non-methylated tag recovery; PCR, polymerase chain reaction; TE, transposable element.

* Corresponding author. Tel./fax: +7 495 727 38 63.
E-mail address: bu3din@mail.ru (A.A. Buzdin).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2012.01.097

independent genomic loci from different chromosomes; (*iii*) it should enable quantification of DNA methylation and (*iv*) should be obtained in an inexpensive, reproducible and easy-to-perform procedure.

In this communication, we report a new method aimed at the generation of control libraries for large-scale methylation studies. We developed technique termed nMETR (non-methylated tag recovery) based on digestion of genomic DNA with methylation-sensitive restriction enzyme, ligation of adapter oligonucleotide and further PCR amplification of non-methylated sites located close to genomic repetitive elements. nMETR procedure is cheap, and its protocol takes only two working days to generate amplicons for deep sequencing. We applied nMETR for human DNA using *BspFNI* enzyme and retrotransposon *Alu*-specific primers. 454-sequencing enabled identification of 1113 loci spread through all of the human chromosomes and harboring *BspFNI* sites adjacent to Alu, of them ~65% were parts of annotated CpG islands. Representation of reads in the library was inversely correlated with methylation levels found for the corresponding loci by bisulfite sequencing, thus confirming efficiency of the method. For the obtained datasets, we created software that eliminates background reads and enables to map and annotate individual tags on human genomic sequence. nMETR tags may serve as the controls for large scale epigenetic studies and for identifying hypomethylated transposable elements located close to genomic CpG islands.

2. Material and methods

2.1. In silico sequence analysis

The consensus sequences of the human repetitive elements were taken from the Repbase Update database (http://www.girinst.org/repbase/update/index.html). Oligonucleotide primers were designed using GeneRunner and Primer 3 software. Homology searches against GenBank were done using the BLAST web server at NCBI (http://www.ncbi.nlm.nih.gov/BLAST). For multiple alignments, BLAST pairwise search, Vector NTI and Clustal W programs (Thompson et al., 1994) were used.

2.2. Oligonucleotides

Oligonucleotides were purchased from Evrogen (Russia) and their sequences are listed in Table 1.

2.3. DNA samples

Human brain genomic DNA sample was kindly provided by Dr. Tatyana Azhikina (Shemyakin-Ovchnnikov Institute of Bioorganic Chemistry, Moscow, Russia). The tissue specimen was sampled *post mortem* from one adult male donor. The biosampling manipulations were done according to E.U. ethical guidelines and approved by the local institutional ethical committees.

2.4. METR procedure

1 μg of human genomic DNA was digested with 5 units of the methylation sensitive restriction endonucleases BspFNI (SibEnzyme, Russia) (recognition site CG^CG). Restriction was carried out for 16 h at 37 °C in 50 μl. Digested DNA was further ligated with the pseudo-doublestranded adapter (A1A2/A3; A1A2, 5′-*TGTAGCGTGAAGACGACAGAAAGGGCGTGGTGCGGAGGGCGGT*-3′; A3, 5′-*AGGGCGTGGTGCGGAGGGCGGT*-3′) annealed as described in (Buzdin et al., 2002), using highly active T4 DNA ligase (SibEnzyme, Russia), for 16 h at 14 °C. 1 μl of the ligation mixture was then PCR amplified with primers A1 + R1 (A1, 5′-*TGTAGCGTGAAGACGACAGAA*-3′; R1, 5′-*AGGTCGAGGCTGCAGTGAGCCGT*-3′), each 0.5 μM. Cycling conditions were the following: initial denaturation for 5 min at 95 °C, followed by a three-step profile: denaturation for 20 s at 95 °C, annealing for 20 s at 60 °C, and extension for 1 min 30 s at 72 °C, for 15 PCR cycles. PCR product was then 10-fold diluted and 1 μl was taken for nested amplification with primers A2 + R2 (A2, 5′-*AGGGCGTGGTGCGGAGGGCGGTCG*-3′, R2, 5′-*CGAGGTTCCAGTGAGCCGTGA*-3′), each 0.5 μM. The primer A2 in addition to adapter sequence had at the 3′ end CG dinucleotide added to improve the selectivity of amplification of the DNA fragments having remnants of the BspFNI restriction site. PCR cycling conditions were as follows: initial denaturation for 5 min at 95 °C, followed by a three-step profile: denaturation for 20 s at 95 °C, annealing for 20 s at 60 °C, and extension for 1 min 30 s at 72 °C, for 10, 15, 20 and 25 cycles. The PCR products were analyzed on 1% agarose gels. DNA fragments lower than 250 bp long were further gel-purified using Wizard Gel and PCR Clean-Up System (Promega). The purified DNA was further either cloned in *E.coli* for Sanger-sequencing of individual colonies, or additionally gel-purified and sequenced using Roche 454 GS FLX apparatus.

Table 1
Oligonucleotide primers used for amplification of bisulfite-treated genomic DNA.

Target locus[a]	Primer sequence	
	Forward primer	Reverse primer
3q25.2	*TGGGTGGATGTTGATAGGGT*	*CCTAATATCACTACTCCCTAATTTCA*
3q25.2 nested	*TTTTTGAGAGGTTTTTTTGGAGA*	*CAAACCCCTAACCAAAACAACT*
6q21	*GTATTAGTAGTGTTTAAAGTTGTTGGT*	*TTTATTACAACTACCTCACATTCACT*
6q21 nested	*TGAGTGTATAATAAGTTGTTTAGAAGG*	*ACAACTACCTCACATTCACTTTACA*
12q22	*GGTATGTGTTTGGGGTAGTGAT*	*ACCAACACTATAAAACTACCCAAT*
12q22 nested	*TTTGTTTTTTTGGGTATTTTAGTA*	*ATCTTATTTCAATCCTAATCCCA*
12q24.11	*GTTGGTTATGTTTTGGTGGATG*	*ACATCATTCTTCCTTACTTACTTCCT*
12q24.11 nested	*GGTTAGGTTTAGAGTTTATAGGTAGTT*	*ATATCAAACACTATCAAATTACTCCT*
16q22.1	*TTTTAGTTTGGGTAATAGAGTGAGAT*	*CTAACTTCATTTACAATCACTTCCA*
16q22.1 nested	*AGAGAGGGTATATTTTATGTTGAGA*	*AAAATCTACTACATACAACAAAACTACA*
19p13.3	*AGGGGTTGGATATTTTGGTGA*	*CCACACTCAAACCCCACAATAC*
19p13.3 nested	*TAGTTGGTTGGGGGTGTTGA*	*TAAACCCTACCACTAATAACTCACAC*
19q13.11	*TTGTTTTTGGAGATTTATTTGATG*	*ATATCAAACACTACAACAATCCCAT*
19q13.11 nested	*TGGAGATTTATTTGATGGGGA*	*CATACCTATAATCCCAACTACTCCA*
20q11.21	*TTGGAGGTTTTAGGGTGGTTG*	*CCTAACATAACCTCCCTCAATAAAC*
20q11.21 nested	*GTTGGGTTTTGGGAGGGTGT*	*AAAAATCAAACTCTACCCTACAACATC*
22q13.2	*AGATTTGGGTAAGATGGTGAGAT*	*CCTACTAACTCCCAATCCCAAT*
22q13.2 nested	*AAGGATTGTTTGAGTTTAAGAGG*	*CCACTCCCACTAACCTCACC*

[a] Target human genomic locus, primer sets for outer and inner (nested) PCR amplifications.

2.5. 454 DNA sequencing

Deep sequencing was accomplished using Roche 454 GS FLX engine at the Center "Bioengineering" of the Russian Academy of Sciences.

2.6. Analysis of sequencing reads

Individual Sanger-sequenced DNA reads were mapped on the human genome and further analyzed manually using BLAT tool at the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start). To analyze Roche 454 reads, we developed Post-Parser software available through the Web at http://www.postparser.net. The database of mapped and annotated reads is available online at http://nmetr.pparser.net.

2.7. Bisulfite sequencing

Bisulfite treatment was carried out using EpiTect kit (Qiagen) following the manufacturers' recommendations. Prior to bisulfite

conversion, DNA isolated from human tissues was digested with *EcoRI* endonuclease. In all the instances, two or more independent, duplicate bisulfite experiments were performed. Bisulfite-treated DNA was then nested PCR-amplified with the primer sets shown in Table 1. For the first PCR, done with the outer primers, the thermocycling conditions were as follows: first PCR, initial denaturation for 10 min at 95 C, followed by a three-step profile: denaturation for 30 s at 95 °C, annealing for 30 s at 50 °C, and extension for 1 min at 72 °C, for 20 cycles. The nested PCR with the inner primers was carried out under the same conditions, for 30 cycles. The nested PCR products were agarose gel-purified using Wizard gel and PCR clean-up system (Promega) and ligated into the pGEM-T easy vector (Promega) according to the manufacturer protocol, followed by the cloning in *E. coli* and sequencing of the plasmid minipreps from the individual clones. In order to find out the methylation statuses of the individual CG dinucleotides, the sequence data was treated using the BiqAnalyzer software.

2.8. Statistical analysis

Statistical tests were done using the GraphPadPrism software. Graphs and diagrams were built using Microsoft Excel program.

3. Theory

Methylation-sensitive restriction endonucleases (MREs) digest DNA depending on the methylation status of their specific restriction sites. Some MREs cut DNA when these restriction sites are unmethylated, whereas the others cut when recognizing only methylated restriction sites (Bulanenkova et al., 2011). Most commonly used MREs recognize unmethylated restriction sites as the substrates and thus may be used to tag unmethylated genomic loci. Digested DNA may be ligated to oligonucleotide adaptors, which enables PCR amplification of genome-wide pools of hypomethylated DNA tags (Azhikina et al., 2006). Generally, MREs used in epigenetic studies recognize sequences having one or several CG dinucleotides. CG dinucleotides are represented more frequently in the genomic regulatory regions like CpG islands, whereas in the rest of the genome their concentration is far lower due to mutations associated with the cytosine methylation (Cooper et al., 2010). Therefore, using MREs recognizing several rather than one CG dinucleotides makes it possible to enrich for the CpG islands or similar regulatory elements in a pool of target sequences. In this application of nMETR, we used BspFNI MRE that cuts at the unmethylated recognition sequence CGCG.

Genomic repeats occupy most of mammalian DNA (Schumann et al., 2010) and may be either specific to certain chromosome locations like telomere- or centromere-specific repeats, or can be randomly spread through the genomes, like transposable elements (TEs). Different TE families are represented in the host genomes by markedly different copy numbers varying from tens to millions of family members (Gogvadze and Buzdin, 2009; Goodier and Kazazian, 2008).

For example, human genome has $>10^6$ copies of *Alu* retrotransposon. *Alu*s are known to be associated with the GC-rich portion of human genome (Cordaux and Batzer, 2009) and distributed more or less randomly among the different gene clusters (Batzer and Deininger, 2002). For $3*10^9$ nucleotides of human haploid genome, *Alu* retrotransposons are distributed so that there is roughly one copy of *Alu* per every 3 kb of genomic sequence. Accordingly, the estimated distance between the *Alu* and the proximal MRE site is <1.5 kb. After ligating specific oligonucleotide adapters supporting the so-called "PCR suppression" effect (Lukyanov et al., 1997) to BspFNI-digested DNA, one can PCR amplify the resulting fragments tagged by the restriction site at one end and by a fragment of *Alu* on the other end (Fig. 1). The ligated GC-rich "PCR suppression" adapters were chosen because they significantly reduce background amplification by inhibiting PCR with only adapter-specific primers. Simultaneously, when the target-specific primer (designed for *Alu* sequence in this application) anneals

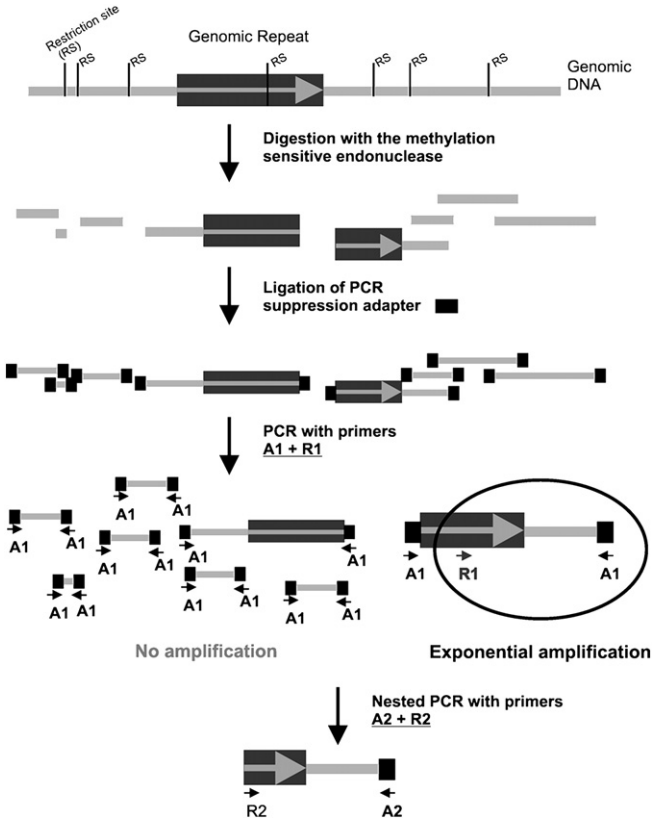

Fig. 1. Schematic representation of nMETR technique. Genomic DNA is digested with methylation sensitive restriction endonucleases. PCR suppression adapters are further ligated, followed by nested PCR amplification with the adapter-specific primers (A1, A2) and genomic repeat-specific primers (R1, R2).

to its complementary site, no PCR-suppression occurs and the fragments of the interest are efficiently amplified. The use of the PCR suppression effect gave rise to numerous experimental techniques many of which are in common use nowadays (Buzdin et al., 2002, 2006; Chalaya et al., 2004; Mamedov et al., 2002; Matz et al., 1997, 2003; Rebrikov et al., 2004).

In the current application, the resulting amplicon represents a set of genomic tags of hypomethylated CGCG sites located close to *Alu* repeats. When sequenced, the proportion of individual nMETR tags is indicative of the overall methylation status of the respective genomic locus. Thus, bioinformatic quantization and mapping of the nMETR tags makes it possible to create characteristic methylation profiles that can be used for the independent experimental validation of larger datasets like whole-genome bisulfitomes and microarray data. Employing other repetitive sequence than *Alu* may modulate representation of nMETR tags in the resulting libraries, according to requirements of the users' research project.

Alternatively, for those interested in the activity of genomic repeats, *e.g.* transposable elements, nMETR provides unique information on the individual repetitive elements located close to non-methylated genomic regions, mostly regulatory CpG islands. This type of mapping might be of significant value for identifying transcriptionally active copies of genomic repetitive elements.

4. Results

We applied a version of nMETR using BspFNI restriction endonuclease as MRE and *Alu* retrotransposon as the repetitive sequence platform, to create a library of hypomethylation tags of genomic DNA isolated from whole human brain. The method is schematically illustrated in Fig. 1. The DNA was digested with BspFNI enzyme and

ligated to double stranded oligonucleotide suppression adapter A1A2/A′. Following phenol-purification, the ligate was PCR-amplified with primers specific to adapter (A1) and to *Alu* sequence (R1). After nested PCR with the respective primers A2 and R2, the amplicon was ligated to TA-cloning vector, cloned in *E. coli* and Sanger-sequenced. Among 200 randomly picked clones, only 8 (4%) were the target sequences having both (*i*) 3′ terminal part of *Alu* and (*ii*) adapter sequence attached to CG dinucleotide left from the BspFNI restriction site. 100% of these nMETR tags have been mapped to certain genomic locations using BLAT software at the UCSC genome browser, and we have found at the appropriate genomic locations complete CGCG motifs recognized by BspFNI enzyme (Fig. 2).

The remaining pool of sequences mostly represented background amplification products with the *Alu*-specific primer only. These amplified background loci had two *Alu* elements directed in a tail-to-tail orientation (Fig. 3). Most of the background reads corresponded to several *Alu–Alu* fragments longer than ~290 bp. To increase the proportion of true nMETR tags in the libraries, we purified the amplified nMETR products shorter than 250 bp from agarose gel, cloned in *E. coli* and sequenced. As expected, at this time among the 200 Sanger-sequenced clones there were 52 true nMETR tags (~26% of the whole library).

By using the aforementioned protocol including gel-purification step, we prepared DNA library for 454 sequencing using Genome Sequencer FLX (Roche). For further analyses, we used only high-quality full-length reads including both (*i*) 3′ terminal part of *Alu* and (ii) adapter sequence attached to CG dinucleotide left from the BspFNI restriction site. Next generation sequencing methods, including Roche 454 pyrosequencing, offer significantly higher performance compared to Sanger sequencing, but yield shorter sequence reads and offer higher error rate. Of 68.729 total reads, ~48% (32.990) were full-length sequences, of them 6.589 were true nMETR tags. Therefore, the abundance of true nMETR tags among the full-length reads was ~20%, and ~10% among all of the acquired 454 reads. The rest was represented by the *Alu–Alu* fragments and by the products of improper adapter ligation and/or of the off-target restriction enzyme activity.

To analyze the reads, we developed a software tool termed "PostParser" that enables finding adapter sequence, *Alu* sequence, mapping of the reads to human genome, and filtering the mapped reads for the presence of CGCG restriction site. The software also annotates the data by providing information on genomic coordinates of the mapped reads, by quantifying number of reads matching to certain genomic loci and by calculating distances between the mapped reads and known structural features like CpG islands and mapped genes and/or RNAs.

PostParser software is built on a modular approach and carried out as a local web-server which allows getting program installed on one powerful computer and run from any network computer. Web-interface was developed using PHP and JavaScript programming languages. MySQL is used as main database in which already obtained

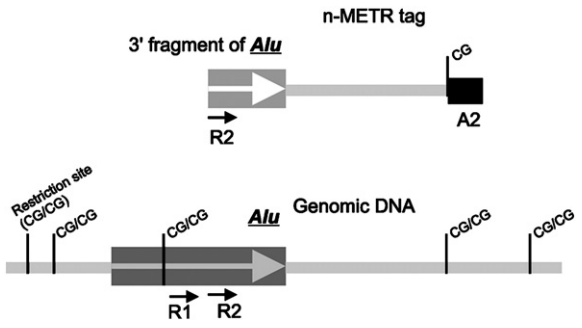

Fig. 2. Human (*Alu*, BspFNI) nMETR tag includes 3′ terminal fragment of Alu repeat, 3′ flanking genomic DNA, CG dinucleotide residual of BspFNI restriction site, and A2 adapter sequence.

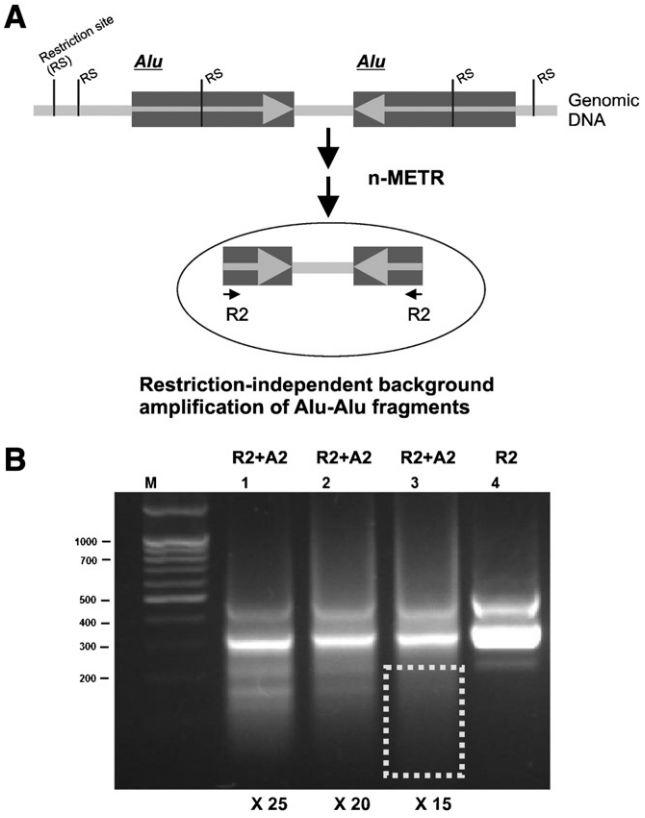

Fig. 3. (A), background PCR products are formed due amplification of Alu-Alu genomic loci with the single Alu-specific primer R1 or R2. (B), representative photograph of (Alu, BspFNI) nMETR products separated in agarose gel. Lane M, DNA ladder; lanes 1–3, nMETR amplification products with the pair of Alu-specific primer R2 and of adapter-specific primer A2, for 25, 20 and 15 PCR cycles, respectively. Lane 4, amplification products with the single Alu-specific primer R2, that correspond to Alu–Alu background amplification. Zone of DNA fragments for gel-purification and further sequencing is boxed.

sequences are stored as well as all related information. External executable programs get connected for resource-intensive process of mapping and annotating. In kernel of the mapping program, well-known BLAT or BLAST is used, whereas in the annotating program we use our original algorithm. Detailed description and links to all mentioned technologies are available through the Web site http://www.postparser.net. PostParser tool enabled us to automatically extract full-length reads and to identify true nMETR tags. The 6589 identified tags represented 1113 human genomic loci, 711 (64%) of them located close to annotated human CpG islands (closer than 200 bp from the tag sequence end). The full nMETR dataset is available from the Web through the link http://nmetr.pparser.net.

Different genomic loci were represented by the different numbers of nMETR tags (Fig. 4). We next tried to assess whether there is a correlation between methylation level of genomic locus and its representation in nMETR tags. To measure DNA methylation levels of particular genomic loci, we used bisulfite sequencing assay (BSA) that enables direct identification of methylated cytosines. The BSA data were processed using BiqAnalyzer software. We analyzed nine genomic loci, of which three were highly represented by nMETR tags (25 reads or more), three had medium representation (5–8 reads), and three were represented by unique tags (Fig. 5). We found that those six genomic loci that had high or medium representations were completely or mostly unmethylated, whereas the ones represented by the unique tags were, in contrast, heavily methylated. We also noticed that there exists an overall correlation between the methylation levels of the restriction site we used (CGCG) and of the enclosing genomic region (Fig. 5). These results evidence in favor of

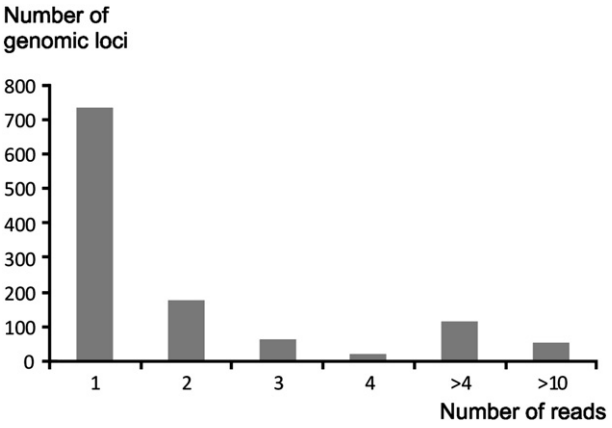

Fig. 4. Distribution of reads among the 6.589 sequenced true nMETR tags.

nMETR applicability and adequacy to the task of large-scale recovery of hypomethylated tags.

We identified a total of 171 genomic loci with five or more reads per locus out of 1113 genomic loci, which gives an efficiency of ~15% in finding confident hypomethylation tags.

5. Discussion

nMETR allows quickly generating pools of hypomethylated genomic sequence tags that can be used as the control datasets for the large-scale DNA methylation studies. Alternatively, nMETR data may be valuable *per se* in the studies of epigenetic regulation of genomic repeats. For example, identification of transposable elements (TEs) located in close vicinity of the unmethylated (or differentially methylated) functional genomic regulatory regions like CpG islands or gene promoters may be helpful for finding highly transcribed individual TE copies. nMETR may be also a method of choice for these applications where the investigators aim to get subsets of hypomethylated sequence tags rather than genome-wide methylation data. nMETR libraries may also vary in the content of the particular tags representing depleted or amplified genomic loci of the source DNAs. This peculiarity of nMETR enabled us to identify several cases of aneuploidy associated with bladder cancer. This has been done by the direct comparison of nMETR tags obtained for the healthy and cancerous tissues. Differences in nMETR tag representation that were not connected with the methylation levels indicated on the regions of aneuploidy (Zabolotneva et al., unpublished data).

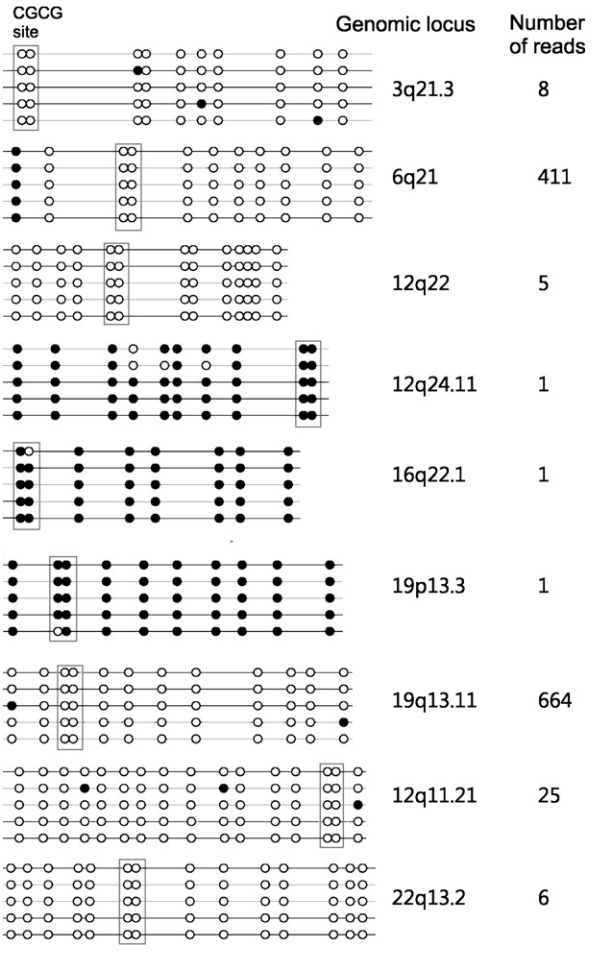

Fig. 5. Representation of bisulfite sequencing results of nine human genomic loci. BspFNI recognition site (CGCG) is boxed.

This technique is applicable to all eukaryotic DNAs having CG methylation and genomic repeats. Genomic repeats may vary from tens to millions in copy number (Kapitonov and Jurka, 2008). In order to get the representation of nMETR tags that fits the best to the individual research project, it is possible to choose among the genomic repetitive sequences those that are characterized by the best features in distribution in genomic DNA. Another possibility is the use of alternative MREs to adjust the number and the quality of nMETR tags. For the amplification of evolutionary old, diverged TE families, like older Alu subfamilies lacking binding sites for the primers used in this study, degenerated PCR primers may be used. The technique may be adopted for any of the currently used next generation sequencing platforms. The only limitation here is that sequencing reads should cover non-repetitive portions of nMETR tags. For the platforms with small sequencing read lengths, pairwise sequencing option may be used to recover both Alu- and MRE site-flanking DNA.

The present nMETR protocol includes a stage of PCR amplification that may bias representation of the nMETR tags, primarily by underrepresenting the sequences with the high GC-content, due to well known "PCR bias effect" (Moskalev et al., 2011).

However, decreasing number of PCR cycles and the use of more processive DNA polymerases during amplification stage may help to avoid this unwanted effect, provided that the amount of DNA required for deep sequencing tends to dramatically decrease over time (Schadt et al., 2010).

For the version of nMETR technology communicated in this report, based on BspFNI enzyme and genomic repeat *Alu*, we anticipate ~60-70.000 of reads to be enough for the characterization of ~1100 human genomic loci, 64% of them located close to annotated CpG islands. Our tests revealed that for this sampling, genomic sequences represented in nMETR libraries by single reads most likely correspond to highly methylated loci, whereas those represented by five or more reads — correspond to mostly unmethylated loci.

Finally, we show that bioinformatic support makes it possible to efficiently analyze the raw nMETR sequence data and to annotate them by sorting individual sequences, quantifying them and mapping on the genome sequence. nMETR tags can be further filtered for the presence or absence of functional genomic features like CpG islands or annotated genes. Overall, we hope that nMETR will be a method of choice for many applications due to its simplicity, robustness and compatibility with the deep sequencing platforms, supplied by a user-friendly bioinformatic interface.

6. Conclusions

We developed an experimental technique termed nMETR that is applicable to generating genome-wide pools of hypomethylated sequence tags. These tags can be used as the controls for large-scale methylome assays or for establishing epigenetic markers. Alternatively, nMETR

tags may serve for identifying unmethylated transposable elements located close to genomic CpG islands. The experimental protocol for this technique is easy to perform and takes only two working days to generate amplicons for deep sequencing.

Acknowledgments

This work was sponsored by the Molecular and Cellular Biology Program of the Presidium of the Russian Academy of Sciences, by the President of the Russian Federation grant 480.2010.4 and by the grant 10-04-00593-a from the Russian Foundation for Basic Research.

References

Azhikina, T., Gainetdinov, I., Skvortsova, Y., Sverdlov, E., 2006. Methylation-free site patterns along a 1-Mb locus on Chr19 in cancerous and normal cells are similar. A new fast approach for analyzing unmethylated CCGG sites distribution. Mol. Genet. Genomics 275, 615–622.

Ballestar, E., 2011. An introduction to epigenetics. Adv. Exp. Med. Biol. 711, 1–11.

Bantysh, O.B., Buzdin, A.A., 2009. Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. Biochem. Biokhimiia 74, 1393–1399.

Batzer, M.A., Deininger, P.L., 2002. Alu repeats and human genomic diversity. Nat. Rev. Genet. 3, 370–379.

Bernstein, B.E., Meissner, A., Lander, E.S., 2007. The mammalian epigenome. Cell 128, 669–681.

Bulanenkova, S.S., et al., 2011. Dam methylase accessibility as an instrument for analysis of mammalian chromatin structure. Epigenetics 6.

Buzdin, A., et al., 2002. A technique for genome-wide identification of differences in the interspersed repeats integrations between closely related genomes and its application to detection of human-specific integrations of HERV-K LTRs. Genomics 79, 413–422.

Buzdin, A., Kovalskaya-Alexandrova, E., Gogvadze, E., Sverdlov, E., 2006. GREM, a technique for genome-wide isolation and quantitative analysis of promoter active repeats. Nucleic Acids Res. 34, e67.

Chalaya, T., Gogvadze, E., Buzdin, A., Kovalskaya, E., Sverdlov, E.D., 2004. Improving specificity of DNA hybridization-based methods. Nucleic Acids Res. 32, e130.

Chen, P.Y., Feng, S., Joo, J.W., Jacobsen, S.E., Pellegrini, M., 2011. A comparative analysis of DNA methylation across human embryonic stem cell lines. Genome Biol. 12, R62.

Cooper, D.N., Mort, M., Stenson, P.D., Ball, E.V., Chuzhanova, N.A., 2010. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Hum. Genomics 4, 406–410.

Cordaux, R., Batzer, M.A., 2009. The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10, 691–703.

Fatemi, M., et al., 2005. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucleic Acids Res. 33, e176.

Fazzari, M.J., Greally, J.M., 2010. Introduction to epigenomics and epigenome-wide analysis. Methods Mol. Biol. 620, 243–265.

Fisher, O., Siman-Tov, R., Ankri, S., 2004. Characterization of cytosine methylated regions and 5-cytosine DNA methyltransferase (Ehmeth) in the protozoan parasite Entamoeba histolytica. Nucleic Acids Res. 32, 287–297.

Gardiner-Garden, M., Frommer, M., 1987. CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261–282.

Gogvadze, E., Buzdin, A., 2009. Retroelements and their impact on genome evolution and functioning. Cell Mol. Life Sci. 66, 3727–3742.

Goodier, J.L., Kazazian Jr., H.H., 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135, 23–35.

Hirst, M., Marra, M.A., 2009. Epigenetics and human disease. Int. J. Biochem. Cell Biol. 41, 136–146.

Jeronimo, C., et al., 2011. Epigenetics in prostate cancer: biologic and clinical relevance. Eur. Urol. 60, 753–766.

Kapitonov, V.V., Jurka, J., 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat. Rev. Genet. 9, 411–412 author reply 414.

Lukyanov, K., et al., 1997. Construction of cDNA libraries from small amounts of total RNA using the suppression PCR effect. Biochem. Biophys. Res. Commun. 230, 285–288.

Mamedov, I., Batrak, A., Buzdin, A., Arzumanyan, E., Lebedev, Y., Sverdlov, E.D., 2002. Genome-wide comparison of differences in the integration sites of interspersed repeats between closely related genomes. Nucleic Acids Res. 30, e71.

Matz, M., Usman, N., Shagin, D., Bogdanova, E., Lukyanov, S., 1997. Ordered differential display: a simple method for systematic comparison of gene expression profiles. Nucleic Acids Res. 25, 2541–2542.

Matz, M.V., Alieva, N.O., Chenchik, A., Lukyanov, S., 2003. Amplification of cDNA ends using PCR suppression effect and step-out PCR. Methods Mol. Biol. 221, 41–49.

Moskalev, E.A., et al., 2011. Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression. Nucleic Acids Res. 39, e77.

Ogoshi, K., et al., 2011. Genome-wide profiling of DNA methylation in human cancer cells. Genomics 98, 280–287.

Pelizzola, M., Ecker, J.R., 2011. The DNA methylome. FEBS Lett. 585, 1994–2000.

Rauch, T.A., Wu, X., Zhong, X., Riggs, A.D., Pfeifer, G.P., 2009. A human B cell methylome at 100-base pair resolution. Proc. Natl. Acad. Sci. U. S. A. 106, 671–678.

Rebrikov, D.V., Desai, S.M., Siebert, P.D., Lukyanov, S.A., 2004. Suppression subtractive hybridization. Methods Mol. Biol. 258, 107–134.

Saxonov, S., Berg, P., Brutlag, D.L., 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. U. S. A. 103, 1412–1417.

Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. Hum. Mol. Genet. 19, R227–R240.

Schumann, G.G., et al., 2010. Unique functions of repetitive transcriptomes. Int. Rev. Cell Mol. Biol. 285, 115–188.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.