Published online 15 January 2013

Nucleic Acids Research, 2013, Vol. 41, No. 6 ***e72***
doi:10.1093/nar/gks1467

In-solution hybrid capture of bisulfite-converted DNA for targeted bisulfite sequencing of 174 ADME genes

Maxim Ivanov[1], Mart Kals[2], Marina Kacevska[1], Andres Metspalu[2,3,4], Magnus Ingelman-Sundberg[1] and Lili Milani[2,*]

[1]Section of Pharmacogenetics, Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm 17177, Sweden, [2]Estonian Genome Center, University of Tartu, Tartu 51010, Estonia, [3]Estonian Biocentre, Tartu 51010, Estonia and [4]Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu 51010, Estonia

Received July 11, 2012; Revised November 2, 2012; Accepted December 17, 2012

ABSTRACT

DNA methylation is one of the most important epigenetic alterations involved in the control of gene expression. Bisulfite sequencing of genomic DNA is currently the only method to study DNA methylation patterns at single-nucleotide resolution. Hence, next-generation sequencing of bisulfite-converted DNA is the method of choice to investigate DNA methylation profiles at the genome-wide scale. Nevertheless, whole genome sequencing for analysis of human methylomes is expensive, and a method for targeted gene analysis would provide a good alternative in many cases where the primary interest is restricted to a set of genes.

Here, we report the successful use of a custom Agilent SureSelect Target Enrichment system for the hybrid capture of bisulfite-converted DNA. We prepared bisulfite-converted next-generation sequencing libraries, which are enriched for the coding and regulatory regions of 174 ADME genes (i.e. genes involved in the metabolism and distribution of drugs). Sequencing of these libraries on Illumina's HiSeq2000 revealed that the method allows a reliable quantification of methylation levels of CpG sites in the selected genes, and validation of the method using pyrosequencing and the Illumina 450K methylation BeadChips revealed good concordance.

INTRODUCTION

DNA methylation is an important mechanism contributing to the control of gene expression. It is well known that changes in DNA methylation play a role in many human diseases as well as in normal development (1). There are a number of methods developed to assess DNA methylation (2). Currently, bisulfite sequencing is considered the 'gold standard' in DNA methylation analysis, as this method allows the investigation of DNA methylation patterns at a single-nucleotide resolution. Moreover, progress in DNA sequencing technologies has allowed the re-sequencing of whole human genomes within a reasonable time and cost (3). The combination of bisulfite-converted DNA with next-generation sequencing (NGS) allows for a powerful whole epigenome analysis (4).

Coupling target enrichment techniques with bisulfite conversion of DNA allows researchers to focus on genomic regions within cellular or disease-related pathways of interest. It can also dramatically decrease the sequencing cost and time required per sample while maintaining the sequencing depth required for reliable quantification of DNA methylation levels. Currently, many methods for target enrichment of DNA have been reported [reviewed in (5)]. The common feature of all of these methods is to capture the targeted genomic DNA (gDNA) fragments by complementary *in vitro* synthesized oligonucleotide sequences (either baits, primers or probes). Given that bisulfite treatment dramatically decreases the sequence complexity of DNA (as most C residues are converted to Ts), it confers otherwise unrelated sequences into significantly similar ones. Furthermore, bisulfite treatment extensively degrades DNA, which complicates the coupling of enrichment procedures with bisulfite treatment.

Despite these complications, there are some successive examples of combining target enrichment methods with the bisulfite treatment of DNA (6–13). Each of these methods have their own limitations, such as high requirements for the amount of input DNA, a complicated

*To whom correspondence should be addressed. Tel: +372 5304 5400; Fax: +372 7420 286; Email: lili.milani@ut.ee
Correspondence may also be addressed to Magnus Ingelman-Sundberg. Tel: +468 5248 77 35; Fax: +468 3373 27; Email: magnus.ingelman-sundberg@ki.se

in-house protocol for preparation of the capture library, a requirement for special equipment or a restricted number of CpG sites to be captured by primers or probes in a particular region of interest.

We developed a novel protocol for combining DNA bisulfite treatment with the standard in-solution hybrid capture procedure provided by the Agilent SureSelect Target Enrichment System. Using a custom SureSelect library that was modified to capture bisulfite-converted DNA, we were able to enrich bisulfite-converted DNA samples for 3.9 Mb of target genomic non-contiguous intervals. Further sequencing of these target-enriched NGS libraries on Illumina's HiSeq2000 allowed the quantification of methylation states of >40 000 targeted CpG sites at the median depth ranging from 37× to 61× in the human gDNA samples assessed. Herein, we describe this protocol in detail and present results of a pilot study involving the capture of specific genomic regions that encode for enzymes involved in drug metabolism and excretion in four adult hepatic gDNA samples. Moreover, these pilot results serve as insight into novel aspects of gene regulation of drug metabolism and transport enzymes that may potentially explain interindividual differences in drug responses.

MATERIALS AND METHODS

The design of a bisulfite-specific Agilent SureSelect library

A total of 174 genes encoding enzymes for absorption, distribution, metabolism and excretion of drugs (ADME genes) were selected as the genes of interest. Among these genes, 32 encode for the core ADME enzymes and 116 genes encode for enzymes in the extended ADME list as determined by www.pharmaadme.org. In addition, 26 genes encoding transcription factors known to regulate the expression of the aforementioned enzymes were included (see Supplementary File 3 for the complete gene list of interest). The genomic coordinates for each gene were obtained from the UCSC Genome Browser (genome.ucsc.edu) where the genomic region of interest included the gene plus 20 000 bp of both the 5′ and 3′ flanking sequences (Supplementary File 3). In total, our region of interest covered 16.26 Mb of genomic sequences and contained 191 534 CpG sites. These genomic coordinates were uploaded to the Agilent eArray web server (earray.chem.agilent.com/earray), and the SureSelect Target Enrichment library was generated according to the manufacturer's instructions with the following settings: Design Strategy = Centred, Bait Length = 120, Bait Tiling Frequency = 1×, Genome Build = Hg19, Avoid Standard Repeat Masked Regions (RepeatMasker) = ON. As the standard repeat masked regions were avoided during the library generation procedure, the resulting design of the SureSelect library reduced the genomic sequence length of interest to 6.38 Mb, containing 82 184 CpG sites and yeilding 53 152 RNA baits (120 nt each).

The next stage involved accommodating the generated custom SureSelect library to capture bisulfite-converted gDNA fragments. To this end, we developed a Python3 script that enabled the conversion of the generated SureSelect bait library to capture bisulfite-converted DNA whereby the threshold number of C-T mismatches was set to 8 (Supplementary File 2). This new output file (containing sequences of bisulfite-converted baits) was uploaded back to the eArray web server as a custom-designed SureSelect library, and the manufactured bisulfite-specific SureSelect library was used in the protocol for the preparation of target-enriched Illumina NGS libraries from bisulfite-converted human gDNA.

Preparation of target-enriched NGS libraries

gDNA from anonymous human liver tissue was isolated using QIAgen DNA Mini kit according to the manufacturer's protocol (Qiagen Cat. #51306). gDNA concentrations were measured with Quant-iT PicoGreen dsDNA assay kit (Invitrogen Cat. #P7589) using SpectraMax Gemini XPS/EM microplate reader (MolecularDevices), and gDNA purity was assessed using Nanodrop 1000 (ThermoScientific). Three micrograms of high-quality gDNA ($A_{260/280}$ = 1.8–2.0) were diluted with 120 µl of TE buffer, transferred to Covaris microTUBEs and subjected to shearing on the Covaris S2 sonicator (Covaris Inc.). Sheared gDNA was then purified with Agencourt AMPure XP beads (BeckmanCoulter Genomics Cat. # A63881) according to the manufacturer's instructions. The DNA was eluted from the beads with nuclease-free water, and 1 µl from each sample was assessed by Agilent 2100 Bioanalyzer (DNA 1000 assay). Following successful shearing, the gDNA was subjected to end blunting, dA-tailing and the ligation with methylated adapters using the TruSeq DNA Sample Prep kit v2 (Illumina Cat. # FC-121-2001). The four gDNA samples were ligated to TruSeq adapters containing different index sequences.

Adapter-ligated DNA was purified with Agencourt AMPure XP beads and then bisulfite converted using the EZ DNA Methylation kit (ZymoResearch, Cat. #D5001) before pre-capture polymerase chain reaction (PCR). Details including components and conditions for the pre-capture PCR, as well as additional information on the comparison of four commercially available bisulfite conversion kits, can be found in Supplementary File 1. Amplified and purified samples were assessed for quality and quantity on the Agilent 2100 Bioanalyzer (DNA 1000 assay).

The PCR amplified DNA was concentrated to ~147 ng/µl using a vacuum concentrator and used for hybridization with the custom SureSelect Target Enrichment library, strictly following the original Agilent instruction manual ['SureSelect Target Enrichment System for Illumina Paired-End Sequencing Library' (G3360-90020), pages 35–47]. Captured DNA fractions were cleaned up and used for the post-capture PCR. For post-capture PCR details, see Supplementary File 1.

Purified post-capture PCR products, which successfully passed the quality check on the Agilent 2100 Bioanalyzer, High Sensitivity DNA assay were precisely quantified with Agilent QPCR NGS Library Quantification Kit for Illumina Genome Analyzer (Agilent Technologies Cat.

#G4880A). The four NGS libraries were then pooled together and sequenced on a single lane of an Illumina HiSeq2000 v3 flowcell, using paired-end sequencing of 100 bp, with 0.5% PhiX spiked into the reaction. For more experimental details, see Supplementary File 1, 'The complete protocol for library preparation'.

Infinium HumanMethylation450 BeadChip assay

From each sample, 500 ng of gDNA was bisulfite modified using the EZ DNA Methylation kit (Zymo Research. Cat. No. D5004) according to the manufacturer's recommendations for the Illumina Infinium assay. The conversion reaction was incubated at 16 cycles of 95°C for 30 s and 50°C for 60 min, followed by a final holding step at 4°C. After purification, 4 µl of bisulfite-converted DNA from each sample was used for hybridization on Infinium HumanMethylation450 (450K) BeadChips, according to the Illumina Infinium HD Methylation protocol. The signal intensities were extracted using the GenomeStudio software. The methylation level of each CpG site was calculated as a beta value according to the fluorescent intensity ratio from the two alleles.

The free software R and the Bioconductor package 'minfi' were used to pre-process the data and for quality control. The original IDAT files from the HiScanSQ scanner were used as input for the minfi package. 'Raw' pre-processing was used to convert the intensities from the red and the green channels into methylated and unmethylated signals. Beta values were computed using Illumina's formula [beta = M/(M + U + 100)]. To combine the data from the Infinium type I and type II probes, peak-based correction was implemented (14). The beta values of all CpG sites with detection *P*-values (calculated by the GenomeStudio software) >0.01 were discarded.

Pyrosequencing

Specific genomic regions (with read depth $\geq 100\times$) were randomly selected for validation using pyrosequencing of bisulfite-treated DNA. Primer sets, forward, reverse and sequencing primers for 3 amplicons were designed using PyroMark Assay Design 2.0.1.15 software (Qiagen). Methylation states of CpG's for validation were amplified using 20 ng of bisulfite-converted genomic DNA of all four samples investigated and 0.2 µM of forward and reverse primers, one of which was biotinylated. PCR reactions were performed using the PyroMark PCR Kit (Qiagen) optimized for bisulfite-treated DNA. Reaction conditions and PCR cycling were conducted as recommended by the kit instructions, adjusting only for optimized primer annealing temperatures, which were between 53–56°C. A total of 10 µl of PCR product and 0.3 µM of the respective sequencing primer were used for analysis. Quantitative DNA methylation analysis was carried out on the PyroMark Q24 instrument using the recommended PyroMark equipment and solutions (Q24 vacuum workstation, Q24 plates, binding buffer, denaturing solution, wash and annealing buffer) (Qiagen) and streptavidin sepharose high performance beads (34 µM, GE Healthcare). Results were analysed using the PyroMark Q24 Software in the CpG analysis mode, and only methylation values with high quality assessment were considered.

Bioinformatics

The 3′ ends of NGS reads tend to have poor quality and thus may lead to mis-mapping and incorrect methylation calls. Moreover, contamination of reads with adapter sequences also complicates mapping and methylaton calling. To avoid these complications, we performed thorough quality control and trimming of the sequence reads using Trim Galore! wrapper script (version 0.1.4, www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the following settings: –quality 20 –phred64 –fastqc –adapter AGATCGGAAGAGC –stringency 1 –length 0. Finally, sequence pairs were discarded if became not longer that 40 bp after trimming. The quality of the paired-end sequences was controlled before and after the trimming process using FastQC (version 0.10.1, www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Bisulfite-treated reads were aligned to the reference human genome (June 2010, GRCh37/hg19) using Bismark (version 0.7.3, www.bioinformatics.babraham.ac.uk/projects/bismark/) with the following settings: –fastq –phred64-quals –non_directional (15). Bismark served as a wrapper script for short read aligner Bowtie 1 (16). To exclude duplicate reads generated during the PCR amplification, alignments that mapped to the same position in the genome were removed using deduplicate_bismark_alignment_output.pl script, which is included into Bismark distribution. Then, DNA methylation calls were extracted from deduplicated Bismark output SAM files using methylation_extractor script (included into Bismark distribution). As our capture library only targeted the top strand of bisulfite-converted genome, only reads that aligned to the original top strand were considered for calling cytosine methylation.

All subsequent steps of NGS data analysis were done using custom Python3 scripts, which are available on request. First, CpG sites having read depth $<10\times$ were discarded. Among the remaining CpG sites, we selected those CpGs, which were analysed in all four gDNA samples simultaneously. These common CpG sites were further divided into 'on-target' and 'out-of-target' CpGs. 'On-target' CpG sites were defined as those overlapping with the coordinates of baits in our custom Agilent SureSelect Target Enrichment library.

DNA methylation values with their 95% confidence intervals for each CpG site were calculated from the experimental binomial data according to Wilson method (17). CpG sites manifesting variable methylation among four samples were found using pair-wise Fisher's exact test ($\alpha = 0.01$). Visualization of DNA methylation data corresponding to genes of interest was done using Matplotlib library (matplotlib.sourceforge.net). Correlations between NGS data and 450 K data were assessed using GraphPad Prism v5.01 (www.graphpad.com). Coordinates of known SNPs and CpG islands (CGI) were downloaded from the UCSC Table Browser (snp135Common and cpgIslandsExt primary tables, respectively). CGI shores were defined as regions within 2 Kb, but not inside CGIs. The CpG density

of a sequence surrounding a certain genomic position was defined as the number of CpG sites within a 200-bp window centred to the given position. The nucleotide coverage of a genomic interval was calculated as a sum of all nucleotides mapped inside of given interval in all four samples.

RESULTS

The development of an algorithm to design a bisulfite-specific Agilent SureSelect library

The typical way to accommodate a custom Agilent SureSelect library for the hybrid capture of bisulfite-converted gDNA is to simply subject the sequence of each bait to *in silico* bisulfite conversion. However, the methylation state of each particular cytosine residue in a given DNA sample can differ from the state assumed during *in silico* bisulfite conversion of baits, thus leading to a potential mismatch between the bait and the corresponding DNA fragment. A high number of mismatches between a certain bisulfite-converted bait and the corresponding bisulfite-converted gDNA fragment will result in impaired efficiency of hybrid capture.

To avoid this inconsistency, we developed an algorithm converting SureSelect baits into their bisulfite-specific counterparts by taking into account the number of possible mismatches for each bait. Previous studies have determined that, at least for 60 nt baits, as much as six mismatches do not significantly impair the efficiency of hybrid capture (18). Based on this observation, we selected 8 as the threshold of tolerated mismatches between our 120 nt baits and the corresponding DNA fragments. Thus, those baits in the original SureSelect library that cover less than eight CpGs (CpG-poor baits) are expected to have less than eight mismatches with the corresponding bisulfite-converted gDNA fragments at any possible pattern of their methylation. These CpG-poor baits resulted in only one bisulfite-converted bait (assuming that all cytosine residues are unmethylated, and thus all Cs are converted to Ts) (Figure 1).

In contrast, those baits in the original library that cover eight or more CpGs yielded two bisulfite-converted baits: one converted from the original bait assuming that all cytosines are unmethylated, and another obtained assuming that all cytosines in the CpG context are methylated and thereby protected from conversion (see Figure 1). Thus, under any possible pattern of CpG methylation in the gDNA, not more than half of the CpG sites within a given bait would contribute to a mismatch with bisulfite-converted gDNA. At the same time, the original SureSelect baits that did not cover any CpGs were not expected to capture CpG-containing gDNA fragments. Hence, these baits were excluded from the final bisulfite-converted library (see Figure 1).

The workflow depicted in Figure 1 was implemented in a Python3 script allowing for rapid and easy conversion of the original input Agilent SureSelect library (generated by the Agilent eArray software) to the corresponding bisulfite-specific SureSelect library. This script can be found in Supplementary File 2. Using the approach explained in Figure 1, we generated our bisulfite-specific SureSelect library, which covered 3.9 Mb of target genomic sequences in 174 ADME genes (containing 82 184 CpG sites). Among these CpG sites, 15 432 were located in 262 CGIs and 10 126 in CGI shores (i.e. regions within 2 Kb, but not inside CGIs, manifesting intermediate CpG density).

Efficiency of target enrichment of bisulfite-converted DNA

The main quality metrics characterizing the efficiency of the target enrichment and the performance of the NGS of the four DNA samples are shown in Supplementary Table S1 (see Supplementary File 1). The observed number of NGS reads mapped with read depth $\geq 10\times$ allowed us to reveal the methylation states for >500 000 CpGs for each of the four gDNA samples analysed. Among them, 303 404 CpGs were detected in all four samples, suggesting that the gDNA fragments containing these CpGs are reproducibly captured by our bisulfite-specific SureSelect library. Owing to the inherent effect of bisulfite treatment, we experienced decreased specificity of target enrichment, where 41 922 of the reproducibly captured CpGs are found in the target 3.9 Mb region. Thus, we were able to analyse 51.1% of the 82 184 CpG sites located in the target region at sufficient depth.

The distribution of the read depth for all CpGs in the target region that were analysed in the four samples is shown in Supplementary Figure S2 (see Supplementary File 1). In agreement with these data, the median read depth for CpGs in the target region ranges between $36\times$ and $77\times$ across the four samples (see Supplementary Table S1).

Both *in vitro* bisulfite conversion of gDNA and *in silico* C-to-T conversion of SureSelect baits leads to a strong decrease of GC content (e.g. the median GC content of our SureSelect baits decreases from 49 to 23% on *in silico* bisulfite conversion). We found that GC content of bisulfite-specific baits can serve as a good predictor of both the nucleotide coverage (Supplementary Figure S3A) and the percentage of CpG sites analysed with sufficient read depth (Supplementary Figure S3B) at the corresponding genomic intervals. At that, extremely AT-rich baits (with GC content $\leq$20%) are almost non-functional and can be removed from the layout of the capture library without any significant loss of its performance. For example, 26.9% of our bisulfite-specific SureSelect library was composed of such AT-rich baits (covering 20.8% of the targeted CpGs), but in total they cover as little as 1.4% of all analysed CpG sites.

Moreover, the GC content of the baits (both before and after *in silico* bisulfite conversion) correlates with the number of CpG sites covered by the given bait, i.e. with the CpG density. Accordingly, for CpG-rich baits, a higher percentage of CpG sites could be analysed with sufficient read depth compared with CpG-poor baits (Supplementary Figure S3C).

The variability of DNA methylation

Methylation levels (as well as their 95% confidence intervals) were calculated for the CpGs in the target region, which manifested a read depth $\geq 10\times$ (see Materials and

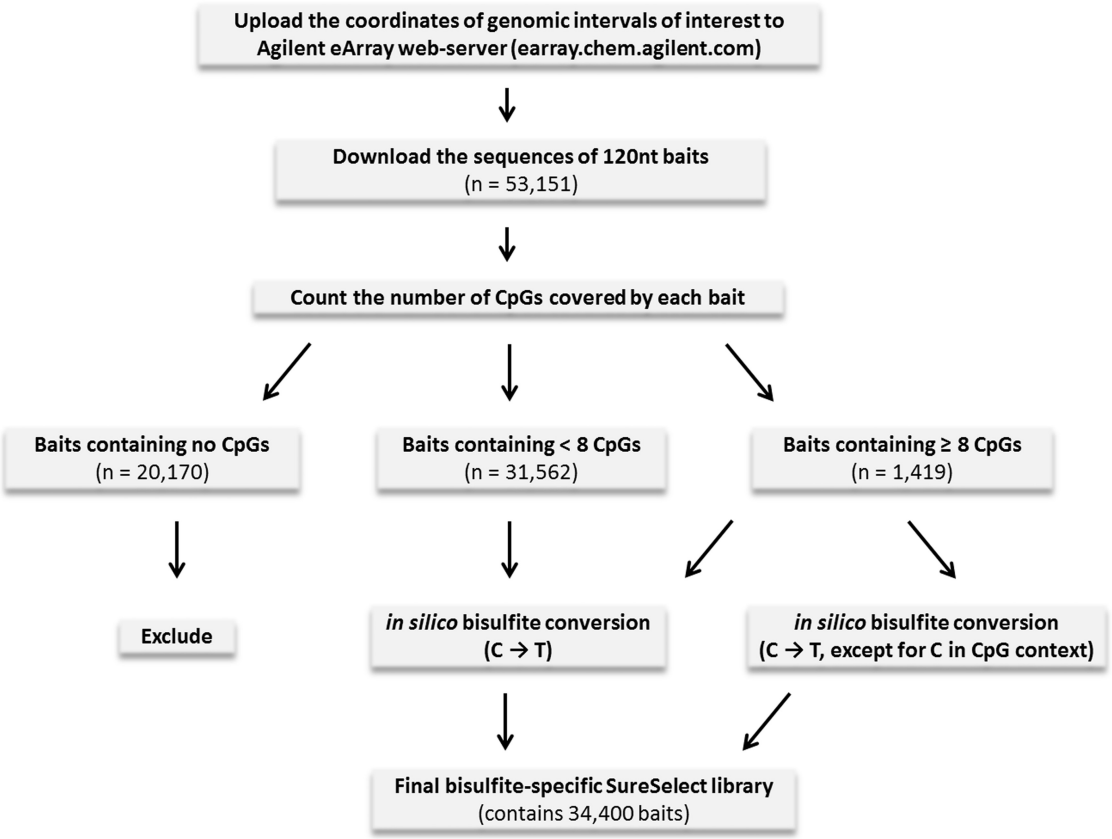

Figure 1. The workflow for designing the SureSelect library specific for bisulfite-converted DNA.

Methods section). Owing to the relatively high read depth observed in the target region, for 90% of the CpG sites, their true methylation levels are expected to differ by not more than 15% from the measured methylation levels. All CpG sites in the target region were checked for possible variability in methylation levels between the four gDNA samples, and 1787 CpGs (4.3% of 41 922 CpGs analysed on target) were found to be differentially methylated. Theoretically, SNPs overlapping with cytosines in CpG context can influence methylation calling, thus providing the basis for false-positive methylation variability. However, we found that only 85 CpGs of 1787 overlap with known common SNPs. Hence, the remaining 1702 CpGs were judged to be differentially methylated among the four gDNA samples analysed. The distribution of these differentially methylated CpGs among the ADME genes of interest is shown in Supplementary File 4 (the corresponding legend can be found in Supplementary File 1). A few examples of the distribution of DNA methylation values along target genomic intervals are shown in Figure 2.

The percentage of CpG sites with variable methylation correlates with the CpG density of the surrounding DNA sequence. The highest percentage of variably methylated CpGs is found in genomic regions with intermediate CpG density (Supplementary Figure S4A). Consistent with variable methylation within intermediate CpG density, the percentage of variably methylated CpGs is the highest in CGI shores and the lowest in CGIs (Supplementary Figure S4B).

When comparing median methylation levels, CGIs were generally hypomethylated, CGI shores were highly variable in methylation levels and genomic regions outside of both CGIs and CGI shores were generally hypermethylated (Supplementary Figure S5), consistent with the current knowledge on CpG density and related methylation states.

Validation of methylation data

The validity of the CpG methylation levels produced by the NGS of our target-enriched samples as well as the bioinformatics analysis was confirmed with two individual techniques, both by pyrosequencing of selected DNA fragments, and by comparison with methylation values produced by the Illumina 450 K Methylation BeadChips.

Pyrosequening, which is generally considered to be a very precise method for the quantification of DNA methylation, was used to validate the results retrieved from three genomic regions, which were randomly selected among those analysed with read depth $\geq 100\times$. In total, the methylation levels of 12 CpG sites in four samples from the NGS study and from the pyrosequencing experiments were compared. The correlation between the methylation levels obtained by the two different methods (Spearman $r = 0.88$) is plotted in Figure 3A. Visualization of these methylation levels plotted against the genomic positions of 12 CpG sites validated is shown in Supplementary Figure S6.

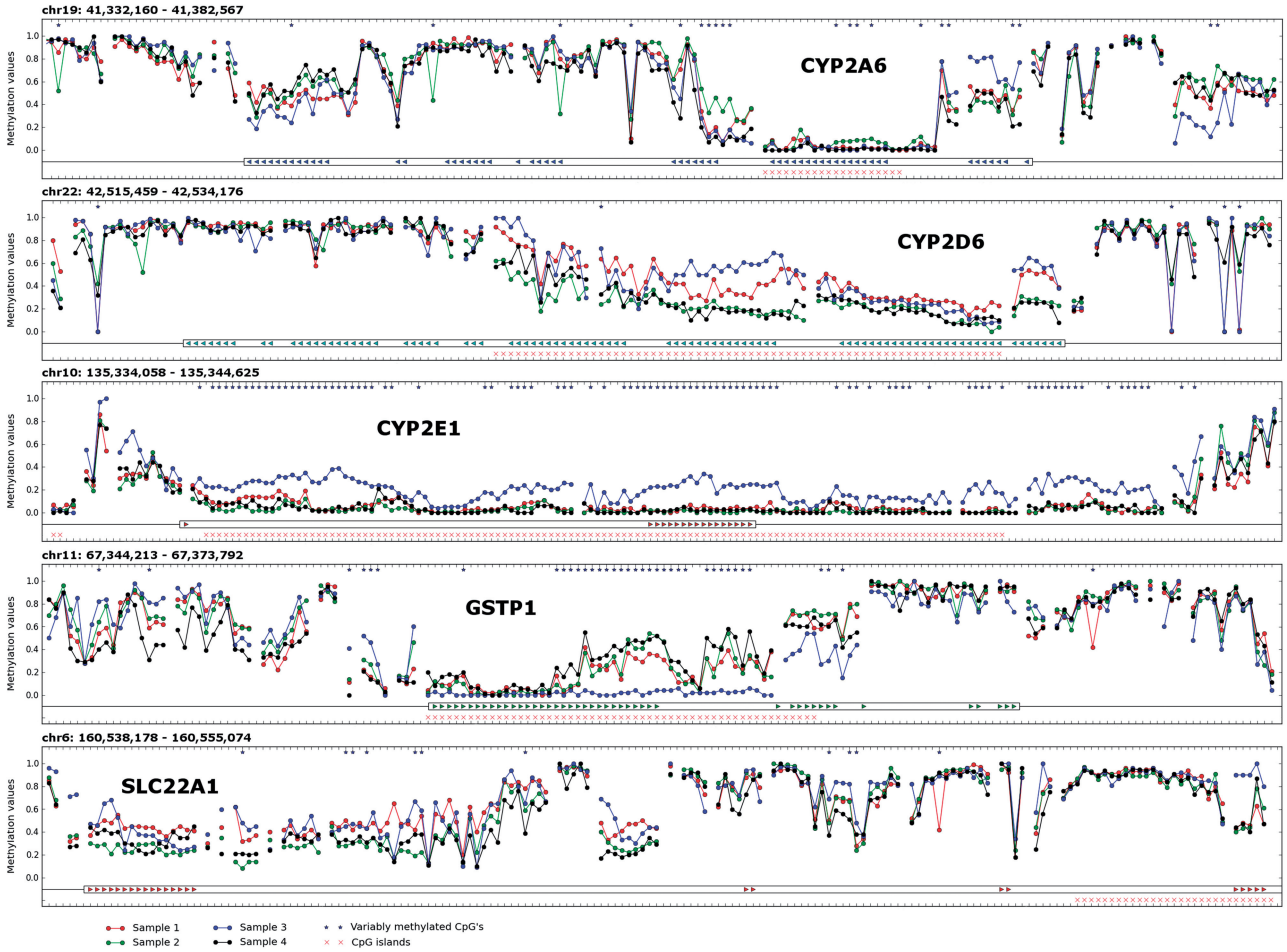

Figure 2. Visualization of DNA methylation values corresponding to the CYP2A6, CYP2D6, CYP2E1, GSTP1 and SLC22A1 genes. The units on the *x*-axis are CpG sites analysed in the given region of interest with read depth ≥10×. The units on the *y*-axis are methylation values (where '1.0' corresponds to fully methylated and '0.0' to unmethylated states). DNA methylation values are plotted as coloured dots (sample 1—red; sample 2—green; sample 3—blue; sample 4—black). For continuous stretches of analysed CpG sites, the corresponding methylation values are connected by coloured lines. CpG sites manifesting statistically significant differences in methylation between four gDNA samples (according to Fisher's exact test) are distinguished by asterisks. CGIs are denoted with pink crosses. Exons of genes of interest are marked with coloured triangles.

The NGS data were also compared with DNA methylation values obtained using the Illumina 450 K BeadChip assays for three of the DNA samples analysed (namely, samples 2, 3 and 4). This assay is expected to interrogate methylation levels of 486 429 CpG sites throughout the whole genome; however, its design is not biased towards ADME genes. This is why only 4933 CpG sites in our 16-Mb region of interest (and, among them, 3650 CpGs in the 3.9-Mb target region) are covered by the design of the BeadChip assay. Among the 348 688 CpG sites, which were detected by the 450 K assay with *P*-values <0.01 in all three samples compared, 1880 CpGs overlapped with those CpGs, which were analysed in the target region of our NGS experiment. A plot showing the correlation (Spearman $r = 0.93$) between the methylation values obtained from the NGS study and the 450 K BeadChips is represented in Figure 3B.

DISCUSSION

The aim of this study was to develop a method for the analysis of DNA methylation patterns in 174 ADME genes (including 20 Kb of their 5′- and 3′-flanking sequences) using bisulfite NGS on the Illumina HiSeq2000 platform. As we did not find existing methods for bisulfite target enrichment to be fully suitable for the purpose, we developed a novel protocol for bisulfite target enrichment, which relies on the hybrid capture of bisulfite-converted gDNA fragments by 120 nt of RNA baits included into a custom Agilent SureSelect library. A brief comparison of published protocols for targeted bisulfite NGS (BS-Seq) is presented in Supplementary File 5.

Essentially, there are two alternative strategies for the integration of a bisulfite treatment step into a hybridization-based target enrichment protocol. The first is target enrichment of native gDNA followed by bisulfite conversion, and the second strategy is to perform the target enrichment on bisulfite-converted gDNA. The advantage of the first strategy is that the specificity of target enrichment remains the same as in the case of the original target enrichment protocol. However, to maintain the DNA methylation state in this scenario, all the required PCR amplification steps have to be omitted, thereby limiting the amount of DNA post capture. Limited

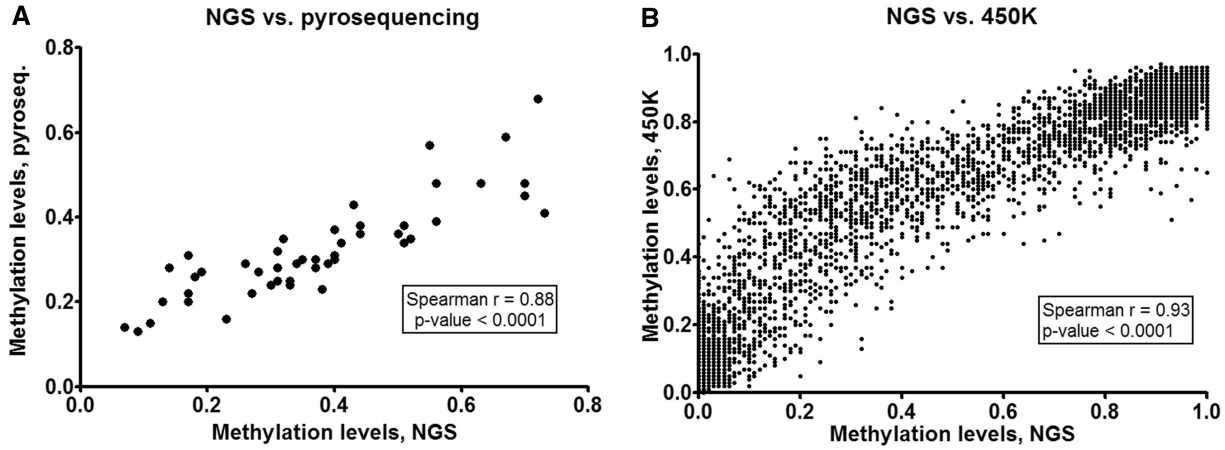

Figure 3. The validation of the NGS data. (**A**) Validation by pyrosequencing. Data points for 12 CpG sites (located in genomic intervals chr9:137 249 931–137 249 946, chr19:16 045 054–16 045 121 and chr16:87 875 316–87 875 361) in samples 1, 2, 3 and 4 are shown on a single plot. (**B**) Validation by Illumina 450 K BeadChip assay. Data points for 1880 CpG sites in samples 2, 3 and 4 are shown on a single plot.

amount of post capture DNA combined with extensive DNA degradation resulting from the proceeding bisulfite treatment reduces the amount of intact and high quality DNA needed for subsequent successful PCR amplification.

The method where bisulfite treatment is used after hybridization-based DNA capture is best illustrated by the study of Lee *et al*. (11) reporting the successful enrichment of an 8-Mb target region using a custom oligonucleotide library. The authors managed to increase the number of intact DNA molecules post capture and bisulfite treatment by using high amounts of starting gDNA (as much as 20–30 μg) and up to six hybrid capture reactions in parallel for each gDNA sample. Using the Agilent SureSelect Human All Exon Kit where native DNA is also captured and then bisulfite treated, Wang *et al*. (13) demonstrated that with the optimization of the experimental conditions, 2 μg of input gDNA can be successfully used to enrich 38 Mb of genomic sequence (13). Moreover, Agilent recently announced their new SureSelect Human Methyl-SEQ system, claiming to enrich 84 Mb of genomic sequence from 3 μg of DNA with the use of a predesigned SureSelectXT library.

A variation to these hybridization-based target enrichment approaches are methods that use target amplification by capture and ligation. Recently, two independent studies using ligation-based approaches, also performing enrichment of native gDNA followed by bisulfite treatment, showed successful DNA capture with low input gDNA requirements (200–250 ng) (9,10). Therefore, ligation-based protocols can be considered as another alternative to the hybridization-based methods, especially if the amount of starting material is limited.

As previously mentioned, the second possible strategy for coupling target enrichment with bisulfite conversion involves bisulfite treatment of DNA before the hybrid capture. As this strategy uses bisufite-treated DNA and hence does not require omitting PCR amplification steps before capture, limited intact DNA post capture and bisulfite treatment can potentially be avoided. However, the specificity of the hybrid capture itself is expected to be impaired owing to the decreased complexity of bisulfite-converted DNA sequences that can result in a high percentage of NGS reads outside of the target region. Moreover, the sequence of bisulfite-converted DNA can be only partially predicted from the sequence of the corresponding native DNA. This complicates the library design for DNA capture, as cytosines in the CpG context may be either cytosines or thymines after amplification, depending on the methylation state.

Despite these complications, the validity of target enrichment on bisulfite-converted DNA was first demonstrated in two independent studies using molecular inversion probes or padlock probes (7,8). Later, the commercial microdroplet PCR method was successfully applied for bisulfite-converted DNA, yielding the methylation states of >77 000 CpG sites localized in the promoters of 2100 genes (12). Additionally, it was demonstrated that 60-nt probes can also be successfully used for array-based hybrid capture of 258 Kb of bisulfite-converted DNA (6). In agreement with the aforementioned common considerations, the specificity of the hybrid capture was shown to be impaired, with not more than 12% of mapped bisulfite reads being in the target genomic intervals (6). Nevertheless, the validation of the NGS data with traditional Sanger bisulfite sequencing allowed the authors to conclude that the capture of bisulfite-converted DNA was not biased towards particular methylation states of original gDNA fragments (6). Thus, hybrid capture of bisulfite-converted DNA can be used for target enrichment; however, the existing protocols have a low genomic coverage of target-enrichment libraries.

To this end, we developed a protocol for the Agilent SureSelect Target Enrichment System involving the bisulfite treatment step before the hybrid capture (see Materials and Methods). We used this modified SureSelect protocol to examine four different gDNA samples that had four barcoded Illumina libraries for

paired-end sequencing. These libraries were pooled together and sequenced on a single lane of a v3 flowcell on the Illumina HiSeq2000 platform.

As expected, the percentage of reads mapped on target (4.0–7.2%) is significantly lower than is usually observed in non-bisulfite target-enrichment experiments (i.e. 70–80%) (see Supplementary Table S1). The comparison of our protocol with the similar protocol developed by Hodges *et al*. (6) reveals some important improvements, including the enrichment of up to 6 Mb of genomic sequences of interest (versus 258 Kb), a significantly lower required DNA input, and the usage of in-solution hybrid capture, which does not require any special equipment, as opposed to solid-phase oligonucleotide arrays.

The number of CpG sites analysed in the target region (n = 41 922) constitutes 51.1% of the total number of CpG sites (n = 82 184), which are located in non-repetitive sequences in our 16.26-Mb genomic region of interest and are covered by the designed SureSelect baits. This means that certain bisulfite-specific SureSelect baits work less efficiently than others. We found that those baits, which became extremely AT-rich (GC content ≤20%) on *in silico* bisulfite conversion, were unlikely to ensure sufficient read depth at the corresponding CpG sites. At that, GC content of baits (and hence the percentage of CpGs analysed with sufficient read depth) positively correlates with CpG density of targeted genomic regions. Our custom SureSelect library contains a substantial proportion of AT-rich baits, as the selection of genomic regions of interest was based solely on the coordinates of ADME genes, and it was not skewed towards a certain specific CpG density. Otherwise, if only genomic intervals with high and/or intermediate CG density would be used as templates for the design of the bisulfite-specific SureSelect baits, one could expect somewhat better quality metrics of target enrichment.

Despite these complications, we were able to assess the methylation levels of 41 922 CpG sites in target regions with sufficient fidelity. The validation of the DNA methylation data obtained from the NGS study with both pyrosequencing and Illumina 450K BeadChip assay shows strong correlations (see Figure 3). Moreover, NGS-derived DNA methylation values do not seem to manifest a systematic shift towards either hyper- or hypomethylated states of analysed DNA fragments, thus suggesting that hybrid capture of bisulfite-converted DNA is apparently not biased towards specific methylation patterns at targeted CpG sites.

In addition, among the targeted CpG sites, 1702 were shown to be differentially methylated among four human liver samples. The percentage of variably methylated CpG sites (from the number of CpG sites analysed in this study) was shown to be higher in the regions with intermediate CpG density, namely, in CGI shores, which is in line with previous observations (19). Hence, CGI shores deserve increased attention when studying interindividual differences in DNA methylation in human livers.

Interestingly, the percentage of variably methylated CpG sites also varies significantly among ADME genes (see Supplementary File 4). Some ADME genes are characterized by relatively high percentage of variably methylated CpGs (e.g. CYP2E1, GSTP1, SLC7A5) compared with others. One can suggest that such genes are more probable to be regulated by DNA methylation than those showing low percentage of variably methylated CpGs. These considerations should however be regarded as preliminary because of the limited number of liver gDNA samples analysed in this study.

Despite the recent progress in the development of novel methods for targeted bisulfite sequencing, protocols with higher efficiency are needed, which will widen the opportunities to analyse DNA methylation patterns in every genomic region of choice and thus contribute to further discoveries in the field of epigenomics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3, Supplementary Figures 1–6 and Supplementary Files 1–5.

ACKNOWLEDGEMENTS

The authors are indebted to Drs Inger Johansson, Inger Jonasson, Emily LeProust, Darren Marjenberg, Silva Kasela, Viljo Soo and Kaarel Krjutshkov for valuable assistance.

FUNDING

European Union through the European Social Fund [MJD71], European Regional Development Fund, in the frame of the Centre of Excellence in Genomics; Estonian Science Foundation [ETF9293]; Swedish Medical Research Council [21384]; EUFP7/Colipa [NOTOX]; IMI-JU MIP-DILI Grant [115336]; targeted financing from the Estonian Government [SF0180142s08]; EU-FP7 grant [OPENGENE]. Funding for open access charge: European Union through the European Social Fund [MJD71].

Conflict of interest statement. None declared.

REFERENCES

1. Gopalakrishnan,S., Van Emburgh,B.O. and Robertson,K.D. (2008) DNA methylation in development and human disease. *Mutat. Res.*, **647**, 30–38.
2. Laird,P.W. (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
3. Metzker,M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
4. Fouse,S.D., Nagarajan,R.O. and Costello,J.F. (2010) Genome-scale DNA methylation analysis. *Epigenomics*, **2**, 105–117.
5. Mamanova,L., Coffey,A.J., Scott,C.E., Kozarewa,I., Turner,E.H., Kumar,A., Howard,E., Shendure,J. and Turner,D.J. (2010) Target-enrichment strategies for next-generation sequencing. *Nat. Methods*, **7**, 111–118.
6. Hodges,E., Smith,A.D., Kendall,J., Xuan,Z., Ravi,K., Rooks,M., Zhang,M.Q., Ye,K., Bhattacharjee,A., Brizuela,L. *et al*. (2009) High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res.*, **19**, 1593–1605.

7. Ball,M.P., Li,J.B., Gao,Y., Lee,J.H., LeProust,E.M., Park,I.H., Xie,B., Daley,G.Q. and Church,G.M. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.*, **27**, 361–368.
8. Deng,J., Shoemaker,R., Xie,B., Gore,A., LeProust,E.M., Antosiewicz-Bourget,J., Egli,D., Maherali,N., Park,I.H., Yu,J. *et al.* (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol.*, **27**, 353–360.
9. Nautiyal,S., Carlton,V.E., Lu,Y., Ireland,J.S., Flaucher,D., Moorhead,M., Gray,J.W., Spellman,P., Mindrinos,M., Berg,P. *et al.* (2010) High-throughput method for analyzing methylation of CpGs in targeted genomic regions. *Proc. Natl Acad. Sci. USA*, **107**, 12587–12592.
10. Varley,K.E. and Mitra,R.D. (2010) Bisulfite Patch PCR enables multiplexed sequencing of promoter methylation across cancer samples. *Genome Res.*, **20**, 1279–1287.
11. Lee,E.J., Pei,L., Srivastava,G., Joshi,T., Kushwaha,G., Choi,J.H., Robertson,K.D., Wang,X., Colbourne,J.K., Zhang,L. *et al.* (2011) Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing. *Nucleic Acids Res.*, **39**, e127.
12. Komori,H.K., LaMere,S.A., Torkamani,A., Hart,G.T., Kotsopoulos,S., Warner,J., Samuels,M.L., Olson,J., Head,S.R., Ordoukhanian,P. *et al.* (2011) Application of microdroplet PCR for large-scale targeted bisulfite sequencing. *Genome Res.*, **21**, 1738–1745.
13. Wang,J., Jiang,H., Ji,G., Gao,F., Wu,M., Sun,J., Luo,H., Wu,J., Wu,R. and Zhang,X. (2011) High resolution profiling of human exon methylation by liquid hybridization capture-based bisulfite sequencing. *BMC Genomics*, **12**, 597.
14. Dedeurwaerder,S., Defrance,M., Calonne,E., Denis,H., Sotiriou,C. and Fuks,F. (2011) Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, **3**, 771–784.
15. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
16. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
17. Newcombe,R.G. (1998) Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat. Med.*, **17**, 857–872.
18. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.
19. Irizarry,R.A., Ladd-Acosta,C., Wen,B., Wu,Z., Montano,C., Onyango,P., Cui,H., Gabo,K., Rongione,M., Webster,M. *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **41**, 178–186.