Published online 13 October 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 20* ***e136***
doi:10.1093/nar/gkl551

MMASS: an optimized array-based method for assessing CpG island methylation

Ashraf E. K. Ibrahim[1,2,*], Natalie P. Thorne[3,4], Katie Baird[1], Nuno L. Barbosa-Morais[2,5], Simon Tavaré[3,4], V. Peter Collins[1], Andrew H. Wyllie[1], Mark J. Arends[1] and James D. Brenton[2]

[1]Department of Pathology, Division of Molecular Histopathology, Addenbrooke's Hospital, [2]Cancer Genomics Program, Department of Oncology, Hutchison/MRC Research Centre, [3]Department of Oncology and [4]Department of Applied Mathematics & Theoretical Physics, University of Cambridge, Hutchison/MRC Research Centre, Hills Road, Cambridge CB2 2XZ, UK and [5]Institute of Molecular Medicine, Faculty of Medicine, University of Lisbon, Avenue Prof. Egas Moniz, 1649–028 Lisboa, Portugal

Received January 17, 2006; Revised May 10, 2006; Accepted July 14, 2006

ABSTRACT

We describe an optimized microarray method for identifying genome-wide CpG island methylation called microarray-based methylation assessment of single samples (MMASS) which directly compares methylated to unmethylated sequences within a single sample. To improve previous methods we used bioinformatic analysis to predict an optimized combination of methylation-sensitive enzymes that had the highest utility for CpG-island probes and different methods to produce unmethylated representations of test DNA for more sensitive detection of differential methylation by hybridization. Subtraction or methylation-dependent digestion with *Mcr*BC was used with optimized (MMASS-v2) or previously described (MMASS-v1, MMASS-sub) methylation-sensitive enzyme combinations and compared with a published *Mcr*BC method. Comparison was performed using DNA from the cell line HCT116. We show that the distribution of methylation microarray data is inherently skewed and requires exogenous spiked controls for normalization and that analysis of digestion of methylated and unmethylated control sequences together with linear fit models of replicate data showed superior statistical power for the MMASS-v2 method. Comparison with previous methylation data for HCT116 and validation of CpG islands from *PXMP4*, *SFRP2*, *DCC*, *RARB* and *TSEN2* confirmed the accuracy of MMASS-v2 results. The MMASS-v2 method offers improved sensitivity and statistical power for high-throughput microarray identification of differential methylation.

INTRODUCTION

Epigenetic changes are heritable changes that include reversible covalent modifications of histone proteins and methylation of DNA. The vast majority of mammalian DNA methylation is located at the cytosine of CpG dinucleotides which are particularly frequent within CpG islands. The definition of a CpG island continues to evolve but the following criteria are currently accepted (1): a length ⩾500 bp, G + C content ⩾50% and CpG dinucleotides at an observed-to-expected ratio ⩾0.60. Approximately 70% of mammalian genomic CpG dinucleotides are methylated and commonly occur within repetitive elements (2). In contrast, most unmethylated CpG islands span the promoter regions of house-keeping genes and tumour suppressor genes and are critical in gene expression regulation and cell differentiation (3).

The number of cancer-related genes inactivated by epigenetic modifications may equal or exceed the number inactivated by genetic mutations or allele loss (4–10). Therefore, the development of high-throughput methods to characterize methylated and unmethylated CpG islands in normal and neoplastic tissues is vital to enable discovery of methylation markers for cancer predisposition as well as understanding the role of DNA methylation in neoplastic progression and drug resistance (9–11).

Differential methylation hybridization (DMH) is an array-based method for comparing the methylation status of CpG islands between test samples and a common reference (12–17). The two DNAs are first digested with MseI to reduce the size of genomic fragments followed by a combination of methylation-sensitive enzymes that only restrict unmethylated recognition sequences. The MseI recognition sequence (TTAA) is found frequently within bulk DNA, but is rarely found within CpG islands which remain intact after digestion (18). Subsequent linker-mediated PCR results in amplicons

*To whom correspondence should be addressed. Tel: +44 1223 256295; Fax: +44 1223 586670; Email: aeki2@cam.ac.uk that are enriched for methylated sequences. The labelled amplicons are competitively hybridized and the ratio of test to reference signal intensities at each probe on the array reflects methylation differences between the two samples. Nouzova *et al*. (19) modified this method by using digestion with a methylation-dependent enzyme, the homing endonuclease *Mcr*BC. This enzyme has a degenerate methylation recognition sequence that only cleaves methylated DNA and is very frequent in CpG islands. Amplicons from digested DNA therefore represent unmethylated sequences, and competitive hybridization of amplicons from *Mcr*BC digested and undigested DNA from the same sample was used to identify methylated sequences by within-sample comparison. This avoided the need for a common reference design which is advantageous for profiling clinical samples where no appropriate reference tissue may be available or where the available reference sample may not have a 'normal' methylation pattern. However, a potential disadvantage of the Nouzova *et al*. (19) method is that there is unequal representation of methylated and unmethylated sequences in a single hybridization and this may reduce sensitivity to detect differential methylation.

Previous DMH profiling studies used microarrays for which the full sequences of the probes, and consequently their restriction map sites, were unknown (12–17). This prevented rational design of the digestion steps and rigorous analysis of probe performance to exclude artefactual errors (16). For example, if a probe sequence lacks the restriction site for a methylation-sensitive enzyme that digests unmethylated target, the signal from this probe will be falsely assigned as methylated. In this work, we used bioinformatic tools to provide detailed annotation of all probes on a publicly available CpG island array and used this information to develop and validate a high-throughput method called microarray methylation assessment of a single sample (MMASS). We show that MMASS offers improved sensitivity to profile methylated as well as unmethylated CpG islands from single samples.

MATERIALS AND METHODS

Cell line

HCT116 colon cancer cells were cultured in McCoy's 5A modified medium supplemented with 10% foetal bovine serum and 1% penicillin/streptomycin. High-molecular weight DNA was isolated using standard proteinase K and phenol extraction methods.

Derivation of probe sequences and preparation of spike control DNA

Human CpG island arrays containing 13 056 features (HCGI12K) were obtained from the Microarray Centre, University Health Network, Toronto, Canada (http://www.microarrays.ca/products/types.html#HCGI12K). End sequences for the CpG island probes were obtained from the Sanger Centre (available from http://www.sanger.ac.uk/HGP/cgi.shtml) and aligned by BLAST (20) against the NCBI v.35 human genome assembly. Each probe sequence was predicted from contiguous sequence tag alignments containing two MseI recognition sites as MseI digestion was used to create the CpG island library (18). Sequences were further annotated with PERL scripts using BioPerl libraries (21) together with data and libraries from Ensemble (22) (Supplementary Table 1 and Supplementary Perl scripts 1–3). Repetitive sequences were identified using `repeatmasker` (http://www.repeatmasker.org).

Spike control amplicons were prepared by PCR from DNA extracted from normal blood. Methylated spikes were methylated *in vitro* using SssI (New England Biolabs) following the manufacturer's instructions and methylation was confirmed by digestion with appropriate methylation-sensitive enzymes and gel electrophoresis. Methylated and unmethylated spikes were added to the samples before MseI digestion at concentrations corresponding to 1–1000 copies (Supplementary Table 2).

Preparation of genomic DNA

Genomic representation of methylated and unmethylated sequences by enzyme digestion. The MMASS-v1 and MMASS-v2 methods used methylation-sensitive and methylation-dependent enzyme digestion for within-sample comparison (Figure 1). Genomic DNA (2 μg for MMASS-v1 and 1.2 μg for MMASS-v2 methods) was digested overnight in a 30 μl volume using 20 U MseI at 37°C. Digested DNA was then ligated to the linkers H-14 5′-tactccctcggata-3′ and H-24 5′-aggcaactgtgctatccgagggag-3′ which prevented reconstitution of the MseI site. Ligation was carried out in a mixture comprising 30 μl MseI digested DNA, 16 μM annealed linkers, 10× ligase buffer, 1.5 μl of 10 mM ATP, 6 μl PEG 6000, 400 U T4 DNA ligase and 10 U MseI in a total volume of 60 μl at 20°C for 4 h. The ligated DNA fragments were purified using the Qiaquick PCR purification kit (Qiagen), eluted in 100 μl water and vacuum dried. For representation of unmethylated sequences, half the sample was restricted with McrBC, after resuspension in 40 μl water with 10× NEB buffer 2, 10× GTP, 10× BSA and 20 U *Mcr*BC at 37°C for 4 h. For representation of methylated sequences, the other half of the sample was restricted by either the combination of BstUI, HhaI and HpaII (MMASS-v1) in a volume of 30 μl (17) or the combination of AciI, HinP1I, HpyCH4IV and HpaII (MMASS-v2) in a volume of 70 μl with 10× NEB buffer 1, 10× BSA and 20 U of each of the enzymes. A further 10 U of each enzyme was added after 4 h for the MMASS-v2 method and the reaction was allowed to continue for a further 2 h.

Genomic representation of unmethylated sequences by subtraction. The MMASS-sub method used subtractive hybridization to obtain the unmethylated representation from the starting DNA (Figure 1). Amplicons representing methylated CpG islands were prepared using the MMASS-v1 method as above by digesting 2 μg of DNA and using both halves for methylation-sensitive enzyme digestion. One amplicon was then used as the subtractor DNA from an additional 1 μg of the test DNA digested with MseI. Subtraction was performed using biotin-labelling (BioNick Labeling System; Invitrogen) of the subtracter DNA and recovery with streptavidin-coated magnetic particles (Streptavidin Magnetic Particles; Roche Diagnostics) following the manufacturer's recommendations and as described

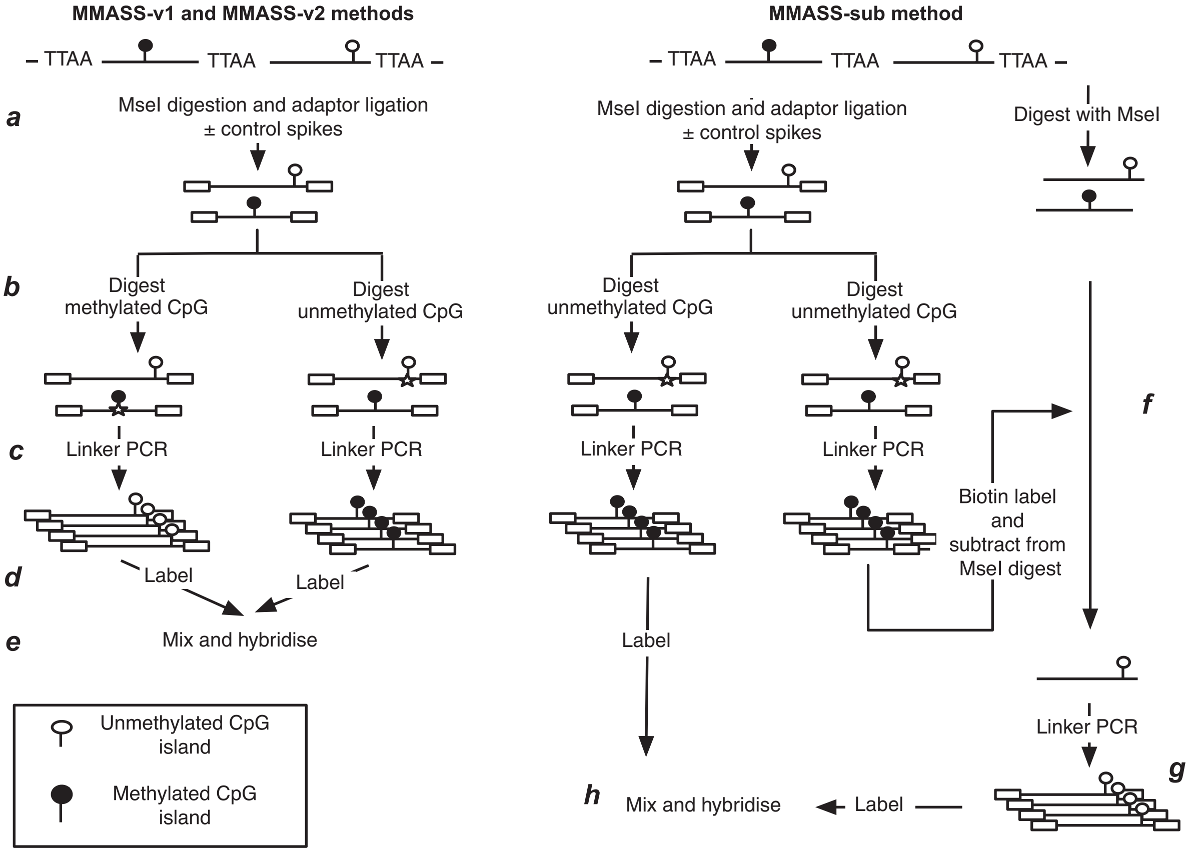

Figure 1. MMASS method. (**a**) DNA from HCT116 was digested with MseI (restriction sequence TTAA) followed by adaptor ligation and addition of control spikes as appropriate. (**b**) One-half of the sample was digested with McrBC (to cut methylated sequences) and the other by a combination of methylation-sensitive enzymes (to cut unmethylated sequences). (**c**) Linker-mediated PCR resulted in two amplicons representing methylated and unmethylated sequences which were (**d**) labeled then (**e**) mixed and hybridised to the CpG island array. For the *MMASS-sub* method, an amplicon representing methylated sequences prepared as described above was subtracted from a MseI digested sample (**f**) resulting in DNA enriched for unmethylated sequences (**g**). The subtracted preparation was subsequently amplified and competitively hybridized against a reciprocal amplicon (**h**).

previously (23). The resulting subtracted DNA (unmethylated representation) was then amplified as below before being hybridized against the remaining methylated amplicon.

Representation using Nouzova method

The Nouzova *et al.* (19) method was carried using both indirect labelling (see below) and as described previously using direct incorporation of Cy3- or Cy5-labelled dCTP and co-hybridization with Cot-1 DNA.

PCR amplification

Each restricted DNA sample was purified using a Qiaquick PCR purification column (Qiagen) and eluted in 100 µl water. PCR amplification was performed in a 300 µl volume comprising 100 µl digested DNA, 10× thermo-start buffer (Applied Biosystems), 100 µM MgCl, 25% DMSO, 200 mM Betaine (Sigma), 0.5 µM H-24 primer, 0.1 µM dNTP mixture and 6 U Deep Vent$_R$ (exo$^-$) DNA polymerase (New England Biolabs). The thermocycling conditions were 5 min at 72°C to fill in the overhanging ends of the ligated DNA fragments, followed by 21 cycles (25 cycles for the MMASS-v2 method) of 1 min at 94°C, 1 min at 65°C and 3 min at 72°C, with a final extension for 10 min at 72°C. Five microlitres of the PCR product was electrophoresed on a 1.5% agarose gel and a diffuse smear pattern between 0.2 and 2kb was taken to indicate successful PCR amplification as described previously (17).

Labelling and hybridization

For each methylated and unmethylated amplicon 300 ng of PCR product was vacuum dried and resuspended in 33 µl of water with 2.5× random primer buffer (BioPrime Labeling Kit; Invitrogen) together with 0.5 ng of control *Arabidopsis thaliana* cDNA (synthesized from pARAB obtained from the Microarray Centre, University Health Network, Toronto, Canada) and denatured at 95°C for 5 min. Each denatured sample was placed on ice with 7.5 µl of 10× dNTP mixture (2 µM each of dATP, dCTP and dGTP, and 0.35 µM dTTP), 1.8 µl of 10 mM aminoallyl-dUTP together with 80 U Klenow Fragment and incubated at 37°C for 2 h then stopped with 5 µl of stop buffer (BioPrime Kit). The total volume was increased to 425 µl with water and unincorporated aminoallyl-dUTP was removed by two centrifugations at 10 000 r.p.m. using a Microcon YM30 concentrator (Millipore). Purified sample was collected by centrifuging the inverted column at 4500 r.p.m. for 5 min and then vacuum dried. Each sample was reconstituted in 4.5 µl of water together with 4.5 µl Cy dye (Amersham-Pharmacia Biotech) in 0.1M sodium bicarbonate titrated with sodium hydroxide to pH 9.0. The mixture was held at room temperature in the dark for 1.5 h and the coupling reaction was stopped by adding 4.5 μl of 4M hydroxylamine and 35 μl of 100 mM sodium acetate (pH 5.2). Labelled DNA was purified using a Qiaquick PCR purification column and then vacuum dried. Both dye-coupled DNAs were then resuspended together in 85 μl of DIG Easy Hyb solution (Roche Diagnostics) together with 5 μl of salmon sperm DNA (10 μg/μl) and denatured at 95°C for 5 min. The hybridization mixture was allowed to cool briefly and 5 μl of yeast tRNA (10 μg/μl) was added and the mixture was held at 65°C for 2 min and allowed to cool to room temperature. Hybridization to the microarray was carried out under a cover slip in a humidified chamber at 37°C for 8 h. The cover slip was floated off in 1× SSC and each slide was washed three times in 1× SSC and 0.1% SDS at 50°C for 15 min followed by removal of SDS at room temperature in 1× SSC and 0.1× SSC for 5 min each. The slides were dried by centrifugation at 500 r.p.m. for 5 min and scanned immediately using the GenePix 4000A scanner (Axon). The settings for PMT gain were adjusted during the initial rapid scan to achieve a balance between the two channels and these settings were used for the high resolution scan. GenePix version 4.1 was used to perform image analysis and feature segmentation.

COBRA

Sodium bisulphite conversion of HCT116 DNA was performed as described previously (24,25). PCR was then performed on bisulphite-modified DNA samples using primers designed to amplify both methylated and unmethylated DNA (Supplementary Table 3). This was followed by restriction digestion using appropriate enzymes that contain CpG within their recognition sequence as these will change in the DNA samples if the original cytosine bases were unmethylated and followed by quantification using electrophoresis on a 2.5% agarose gel.

Microarray analysis

The limma (26) package within the R environment (27) was used to background-correct, normalize and analyse the data. Where the background exceeded the foreground intensity, the minimum background value for the array was subtracted rather than the local background measurement of the spot. We combined replicate dye-swap arrays for each method using the linear model and empirical Bayes smoothing procedures available in the limma package. A full transcript of all statistical code and the results of computations are provided in the Supplementary Sweave document which allows the analysis to be examined and repeated exactly (28–30). The raw data from the array experiments is available from the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo) under the series accession number GSE5326.

Calculation of spike statistics

For each spiked probe we obtained the spike, digestion and non-digestion effect statistics which represented the spike amount (compared to background level) and the amount of the spike that was digested and undigested, respectively. The spike effect was estimated from the difference in log-intensities between the spiked and unspiked experiments. For the comparisons between labelling methods, intensities were obtained from the channel in which the spike was not expected to be digested and averaged between arrays. The digestion effect was defined as the difference in log-ratios between the spiked (M') and non-spiked (M) arrays for methylated ($D_{m_i} = \bar{M}_i' - \bar{M}_i$) and unmethylated ($D_{u_i} = \bar{M}_i - \bar{M}_i'$) spiked probes ($i$), respectively.

RESULTS

Analysis of probe sequences

We first used bioinformatic methods to predict the complete sequence for all probes on the CpG island arrays, as at the start of the project only end-sequence tags were available (18). The majority of the library was subsequently fully sequenced by the University Health Network Microarray Centre, Toronto (sequences available at http://derlab.med.utoronto.ca/CpGIslands/). After BLAST comparison to the human genome 5435 out of 13 056 (41.6%) probes were selected that had a percentage identity of >97% and <30% masked repeat elements and these were annotated as single copy sequences. A further 1190 probes (9.1%) contained 100% repeat sequences and the remainder was either not identifiable or had an intermediate percentage of repetitive sequences. The restriction sites for all commercially available methylation-sensitive enzymes were identified for unique probes together with the distance to the nearest neighbouring genes and the percentage and type of included repetitive sequences (Supplementary Table 1).

From these analyses, we found that 4160 out of 5435 (76.5%) of the probes on the CpG array would be informative when using the previously described combination of BstUI, HpaII and HhaI enzymes to generated representations of methylated target DNA (17). We predicted that using a novel combination of four enzymes (AciI, HpaII, HinP1I and HpyCH4IV) would utilize 4403 out of 5435 (81%) of the array probes and therefore improve utility. In addition this combination of enzymes was more convenient as all four enzymes could digest efficiently in the same buffer. In contrast, digestion with BstUI, HpaII and HhaI required a two-step digestion protocol with an additional purification step.

We hypothesized that the sensitivity of array-based methylation detection could be improved if greater contrast could be achieved between methylated and unmethylated signal. We therefore evaluated two different methods for generating representations of unmethylated sequences. First, we used McrBC to digest methylated DNA in one-half of the sample for comparison against digestion with the combinations of methylation-sensitive enzymes above (MMASS-v1 and MMASS-v2; Figure 1). Second, we used subtractive hybridization using a subtractor DNA digested with BstUI, HpaII and HhaI (MMASS-sub; Figure 1).

Exploratory data analysis

For each of the methods we obtained four microarray hybridizations, using replicate biological preparations in a balanced dye-swap design and compared the results to the method of

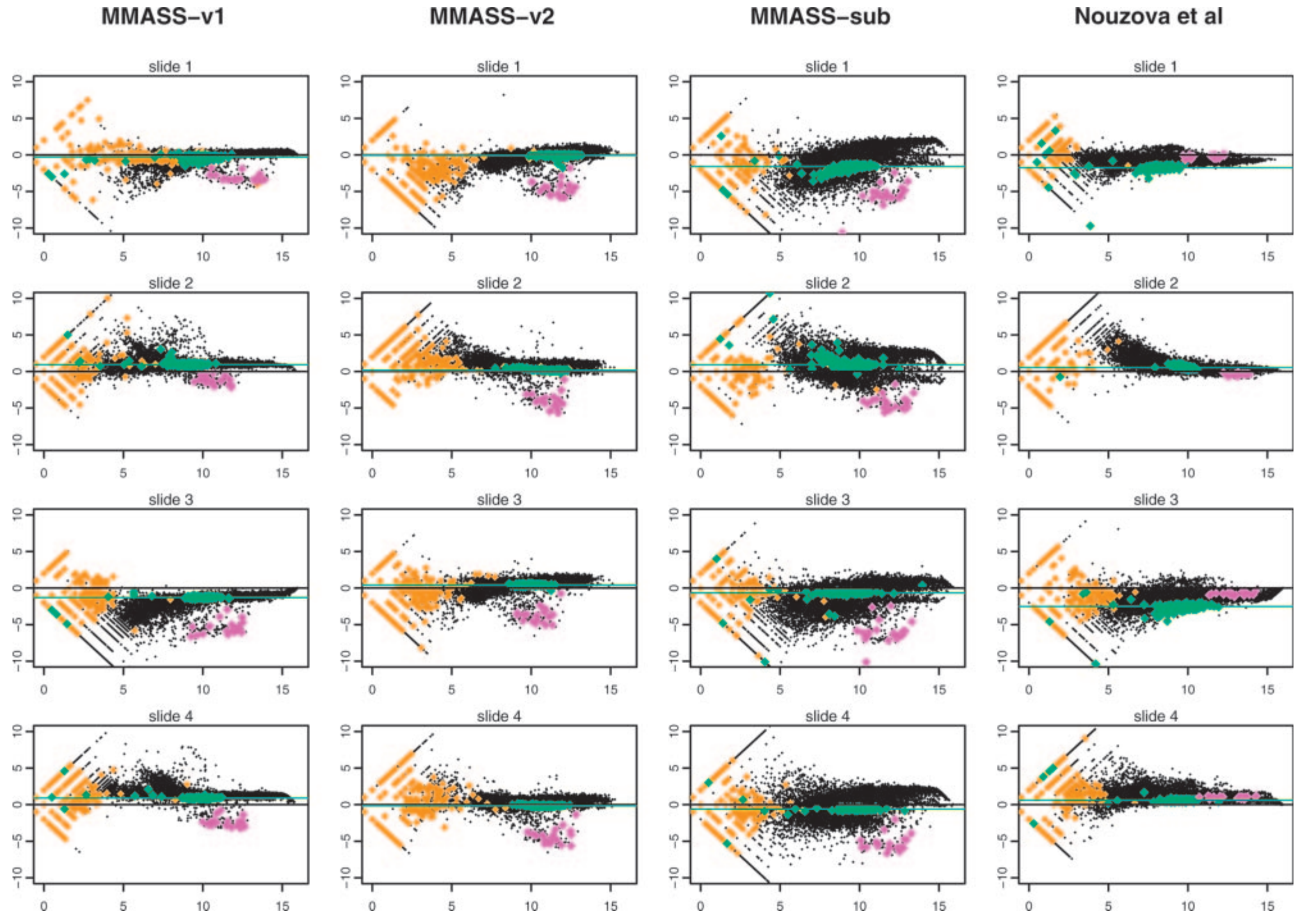

Figure 2. *MA* plots of methylation profiling experiments. Columns show replicate arrays for the MMASS-v1, MMASS-v2, MMASS-sub and Nouzova *et al*. (19) methods. Positive *M* values represent increased ratio of methylated sequences and negative *M* values increased ratio of unmethylated sequences (except for the Nouzova *et al*. method where the negative *M* values are artefactual as no differential signal should be obtained from unmethylated sequences). *A* values represent average fluorescence intensity. Coloured points are *A.thaliana*, green; mitochondrial DNA, purple; empty points, orange. Linear lines at low *A* values are artefacts from background subtraction.

Nouzova *et al*. (19). DNA from the colorectal cancer cell line HCT116 was used for all experiments as methylation patterns have been well characterized in this cell line (31,32).

The overall quality of individual hybridizations was assessed by inspection of *MA* and spatial plots (33) for each of the arrays (Figure 2 and Supplementary Sweave document). Unsatisfactory array experiments were repeated and 16 high-quality hybridizations were obtained from a total of 19 experiments.

Log ratios (*M*) for control *A.thaliana* probes showed little variation around $M = 0$ indicating high reproducibility (Figure 2). Inspection of blank spots showed uniform low intensities as expected (Figure 2). In contrast to *MA* plots from expression array and array CGH experiments, the distribution of log-ratios from MMASS and Nouzova experiments was not symmetrical. It is important to note that for the Nouzova *et al*. (19) method the log-ratios should theoretically all be positive, as a mixture of methylated and unmethylated sequences was directly compared to unmethylated sequences. Comparison between arrays showed that each method had a characteristic distribution of data points on the *MA* plot that was highly consistent between replicate experiments. For the MMASS-v2 and the MMASS-sub methods there was a bimodal distribution of log-ratios at higher probe intensities (Figures 2 and 3) indicating increased separation between methylated and unmethylated sequences.

We evaluated different strategies for optimum normalization of each method and these are discussed in detail in the Supplementary Sweave document. The *A.thaliana* probes proved unreliable for location normalization except for the MMASS-v2 method where pipetting error was well controlled, allowing the use of median correction (Supplementary Figure S1). Replicate arrays for the MMASS-v1 and Nouzova methods were sufficiently comparable after global loess normalization. For the MMASS-sub method, normalization was performed using a subset of high-intensity methylated clones that demonstrated consistent log-ratios between replicate arrays (Supplementary Figure S2).

Identification of differentially methylated probes

For each of the methods we fitted a linear model followed by empirical Bayes smoothing to obtain *B* statistics (26,34) so that probes could be ranked by the likelihood of differential methylation. Volcano plots summarizing the results showed striking differences in the *B* statistics obtained from

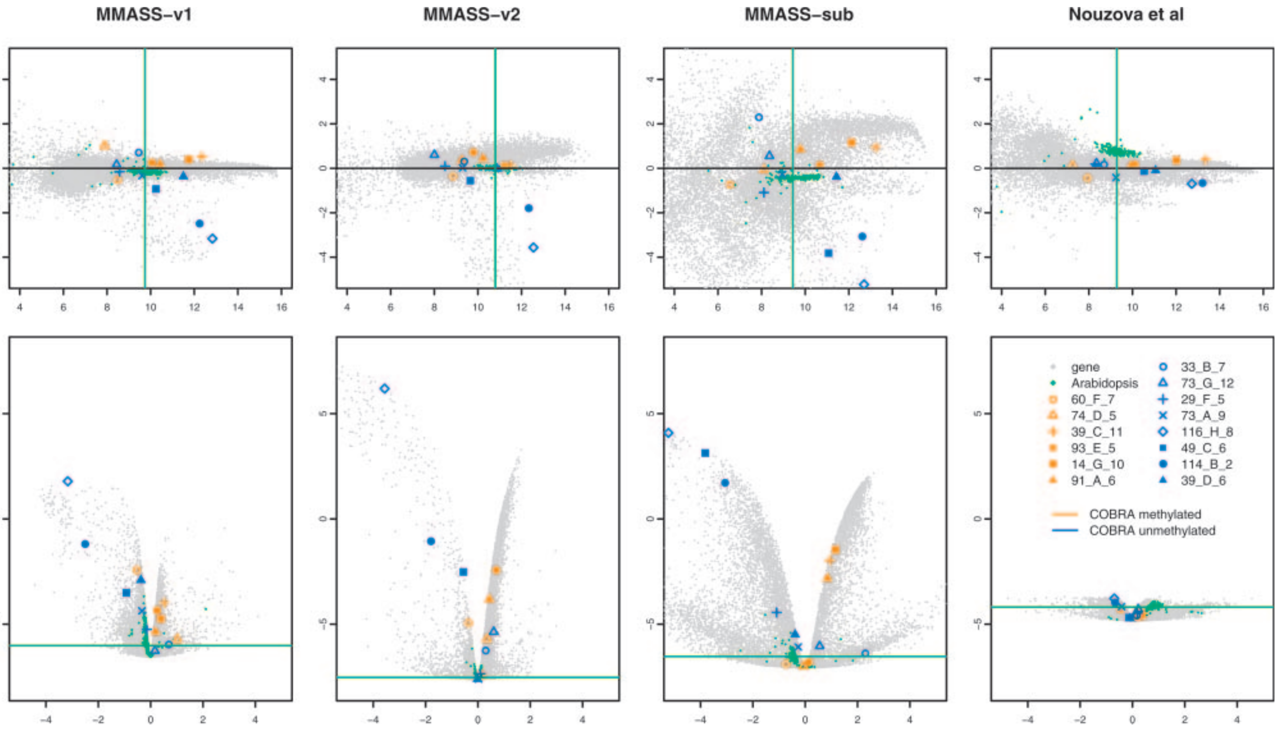

Figure 3. Correlation between COBRA and array methylation results. Plots show data from the linear model fitted to replicate arrays for each of the MMASS-v1, MMASS-v2, MMASS-sub and Nouzova *et al*. (19) methods. *MA* plots (upper panel) and volcano plots (lower panel) are shown. Coloured probes indicate validation results from COBRA analysis. Orange indicates confirmed as methylated and blue indicates unmethylated. Legend shows probe id. See also Supplementary Table 3.

each method (Figure 3). Most notably, the Nouzova method gave a very low and limited range of *B* statistics demonstrating a lack of power to assess methylation. In contrast, the MMASS-v2 and MMASS-sub methods resulted in much higher *B* values indicating better assessment of methylation. However, compared to the MMASS-v2 method, the MMASS-sub method was more variable as shown by the wide spread of points with low *B* values in the MMASS-sub method volcano plot (Figure 3). The MMASS-v2 method resulted in markedly higher *B* values than the MMASS-v1 method. The poor discrimination of the Nouzova data was surprising. To exclude artefact caused by our use of indirect labelling or different hybridization conditions, we repeated the original protocol exactly as described for four additional arrays. No significant increase in performance was obtained by using the unmodified protocol (Supplementary Sweave document).

We next examined results for mitochondrial DNA and repetitive element probes as these are known to be substantially unmethylated and methylated, respectively (35–37). The *MA* plots showed the mitochondrial DNA to be consistently unmethylated in data from the MMASS methods (Figure 2). However the Nouzova *et al*. (19) method had poor sensitivity for distinguishing unmethylated mitochondrial genes. Ranking by *B* statistics showed that the top unmethylated probes were mitochondrial DNA sequences and that the most methylated probes were repeat elements (Figure 4). It is important to note that as mitochondrial and repetitive sequences are present in high copy number in the genome (36) and over-represented on the HCG12K arrays (see earlier), more consistent probe measurements would be expected, making it easier to detect differential methylation for these probes as compared to single copy genes.

We then assessed the effect of spiking *in vitro* methylated and unmethylated target DNAs into the hybridization samples as positive and negative controls for the detection of methylated and unmethylated sequences. We first labelled 32 candidate spikes and hybridized them to two HCGI12K arrays to test the stringency of hybridization (Supplementary Figure S3). Eight spikes which showed correct hybridization and had the largest spike effects (Supplementary Figure S4) were selected for further analysis. Adequacy of *in vitro* methylation was confirmed with BstUI digestion (Supplementary Figure S5) and spikes that were poorly methylated or that had inconsistencies between predicted and actual DNA sequence were excluded from analysis. The spikes were added to two hybridizations for each method leaving the remaining two unspiked so that background measurements for spiked probes could be established.

To quantify the amount by which the spikes were digested, we calculated the spike effect at each of the four methylated and unmethylated spiked probes and compared this to the spike remaining after digestion (Figure 5 and Supplementary Figure S4). For unmethylated spikes, the largest spike effect was seen for probes shown in Figure 5g and h where almost complete digestion by the MMASS-v1 and MMASS-v2 methods was observed. As expected there was minimal digestion by the Nouzova *et al*. (19) method as McrBC does not restrict unmethylated sequences. The spike effect at probes shown in Figure 5 e and f was too small to allow meaningful interpretation. For methylated spikes (Figure 5a–d) the largest digestion effect was observed for the MMASS-sub

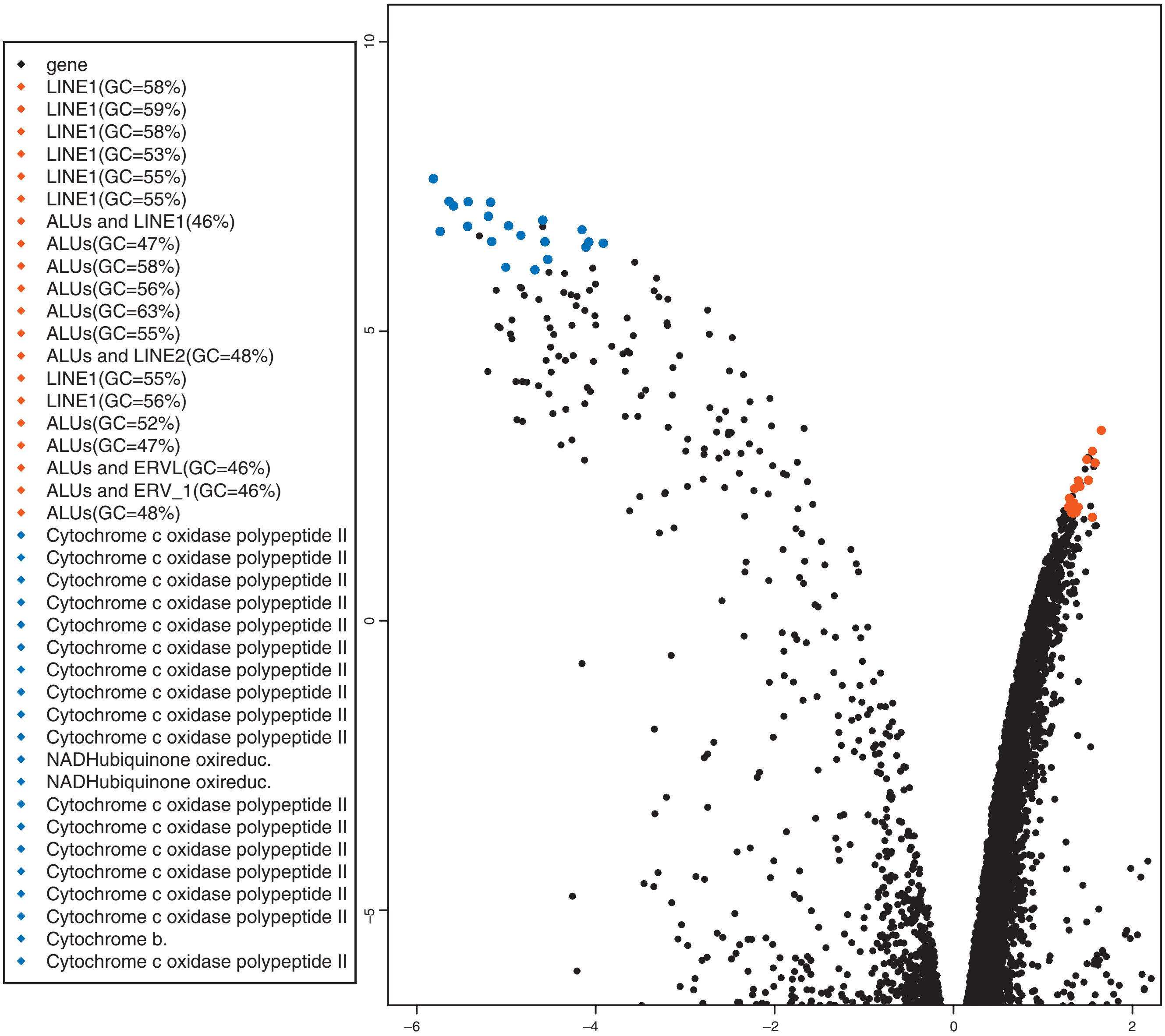

Figure 4. Mitochondrial and repetitive sequence probes show highest differential methylation. Volcano plot of linear fit model combining four experiments using the MMASS-v2 method. The top 20 methylated and unmethylated sequences are plotted as vermilion and blue points respectively. Legend shows probe type.

and MMASS-v2 method. However the subtraction process could also have attenuated the spiked sequences, increasing the apparent effect. There was little digestion effect seen for the Nouzova (19) and MMASS-v1 methods.

To validate the results for single copy genes, we selected 14 probes randomly within low, medium and high average probe intensity (*A*) ranges and compared results from array hybridizations from each MMASS method with independent assessment by COBRA (Figure 3, Supplementary Figure S6 and Supplementary Table 3). Although COBRA only surveys two to four CpGs in an amplicon, our experience in cancer samples is that this gives a good indication of the methylation status of the locus. Results from probes with *A* values higher than the median intensity of the *A.thaliana* control spots were more consistent with the COBRA results and these higher intensity probes were also more consistent across all MMASS methods (Figure 3 and Supplementary Table 3).

The MMASS-sub method resulted in greatest separation between the methylated and unmethylated COBRA validated clones (Figure 3). The ranking of the MMASS probes by degree of methylation was consistent with full and partial methylation results detected by COBRA (Figure 3 and Supplementary Table 3).

Validation of methylation of cancer-related genes

We then examined 325 single copy probes identified by the MMASS-v2 method with values of $B > -3$ as this cut-off was consistent with the COBRA validation experiments

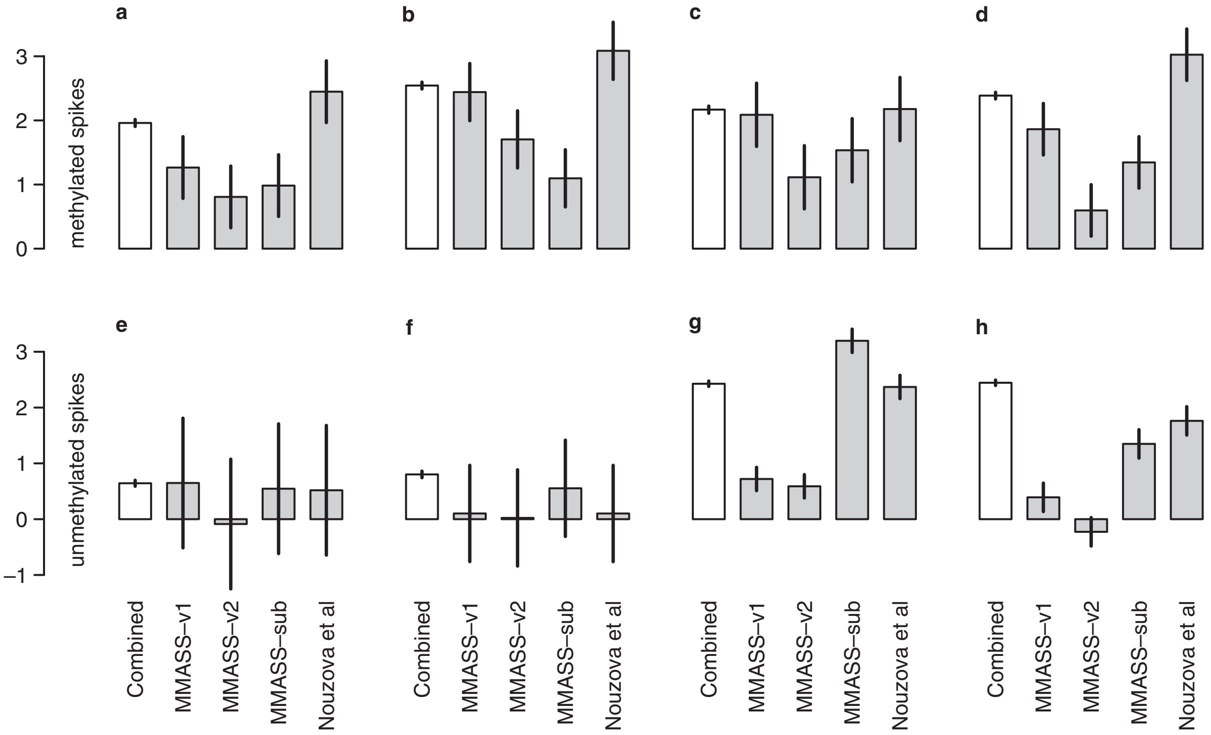

Figure 5. Digestion of methylated and unmethylated spikes. Bar plot shows the spike effect averaged over all the methods (open bars) of methylated (upper panel) and unmethylated (lower panel) spiked probes. Non-digestion effects (grey bars) for each probe are shown by method and should be compared to open bars. Vertical lines over bars indicate 2 SEs around the mean.

(Supplementary Table 4). From these, 22 were selected that were proximal to genes reported previously as having cancer-related functions (Figure 6), including DNA replication and repair (*PMS2L4*, *MCM7* and *BRCA1*) and tumour suppressor function in colorectal cancer (*SFRP2*) (38). Validation of the methylation status of five CpG islands (*PXMP4*, *SFRP2*, *DCC*, *RARB* and the unmethylated housekeeping gene *TSEN2*) confirmed correct array results using COBRA or MSP (Supplementary Figure S7). Our array result for *HNRPA2B1* (Figure 6) was not in agreement with previous data that has shown it to be unmethylated in HCT116 (39) but we were unable to obtain a satisfactory MSP result to confirm this (data not shown). The reproducibility of the MMASS-v2 method was also demonstrated by the finding of very similar *B* values for several duplicate probes from single copy genes, including *MCM7* (Figure 6 and Supplementary Figure S8).

We also examined array data from 23 probes representing 9 genes (*SYK*, *ZFP37*, *DIRAS3*, *RARB*, *LMX1A*, *DAPK1*, *SFRP2*, *FAT* and *RASSF1*) that have been reported previously to be methylated in HCT116 (32,40). Inspection of *MA* and volcano plots for each of the four methods (Figure 7) showed that the MMASS-v2 results were most consistent with previous data.

DISCUSSION

Genomic profiling of methylated and unmethylated sequences using methylation-sensitive restriction enzyme digestion and hybridization to microarrays is a potentially powerful and convenient technique. However, in contrast to work carried out on expression microarray data, no detailed assessment of the effects of different protocols or analysis methods has been performed (17,19,31,41,42). We have developed and optimized new restriction enzyme methods to profile both methylated and unmethylated sequences within a single sample.

The three MMASS methods resulted in very consistent data representation between replicate experiments but there were marked differences in sensitivity. The MMASS-sub method increased the power to resolve methylation differences as compared to the previously published Nouzova *et al*. (19) method, but also increased noise (Figure 3). The subtraction steps were time consuming, and there remains a theoretical disadvantage that the subtraction may compound errors caused by partial digestion. For example, an excess of the partially digested sequences in the subtracter DNA amplicon could result in disproportionate removal of target DNA and a skewed representation of methylation. The MMASS-v2 method resulted in better representation of the methylation status of the target DNA (Figure 7) and had less noise, and therefore increased power, as compared to other methods (Figures 3 and 7). This may be in part because of better digestion of unmethylated sequences (Figure 5). As additional fresh enzymes were added in the MMASS-v2 method and digestion was carried out in a single step using one buffer–enzyme combination, minimizing potential loss of sample.

The poor performance of the Nouzova *et al*. (19) method was surprising and cannot be explained simply by technical reasons, such as failure of McrBC digestion, as all experiments were carried out using the same conditions, batch

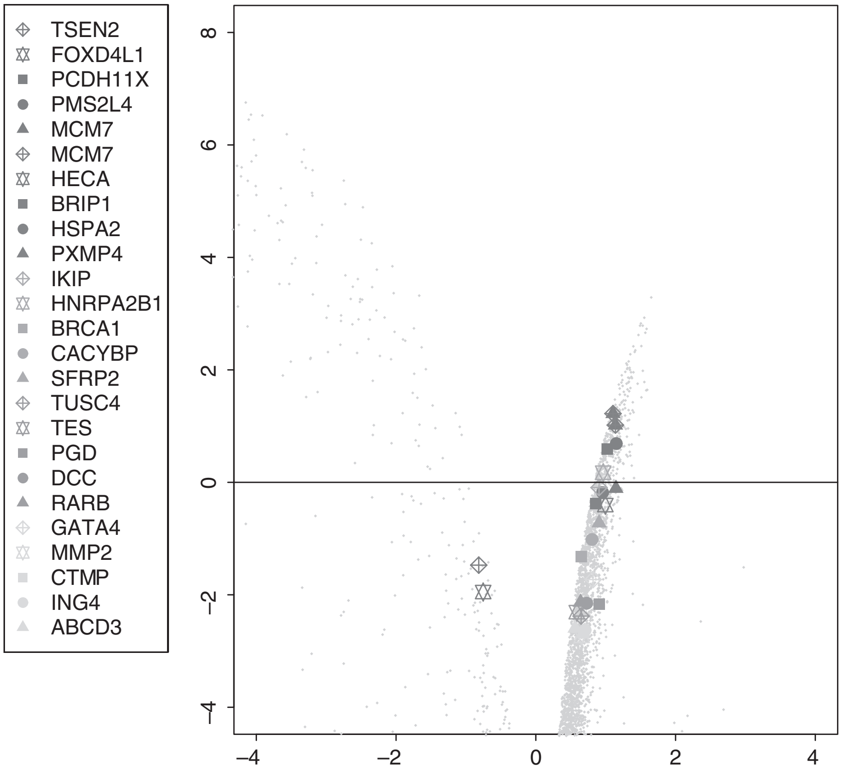

Figure 6. Methylation results for 22 cancer-related probes in HCT116 using the MMASS-v2 method. Figure shows detail from volcano plot of linear fit model. Legend shows gene symbols.

of enzyme and *in vitro* methylated spikes. In addition, the dynamic range for probe data was very similar between our Nouzova experiments and the original publication. It is possible that other effects such as array quality or the higher genomic complexity of the amplicon from the undigested DNA (containing unmethylated and methylated sequences) may have altered spike-probe hybridization results. However in contrast to MMASS, the Nouzova method has very poor sensitivity for detecting hypomethylation (Figure 2) such as mitochondrial spots.

The mixtures being compared by hybridization may have had strong effects on sensitivity. The direct comparison of methylated to unmethylated representations appears more sensitive (larger *M* values) than comparisons to a mixture of methylated and unmethylated sequences as in the Nouzova *et al*. (19) method. The MMASS methods resolved with high precision the methylation status for repetitive and mitochondrial target DNAs as these are represented in high copy number in the genome (Figure 4). They also were able to resolve single copy CpG methylation and identify correctly the methylation status of a number of CpG islands which have been described previously to be methylated or unmethylated in HCT116 (Figures 6 and 7).

The bioinformatic analysis of methylation array data is very different to that of expression data in which symmetrical distribution of log-ratios is assumed and the main aim of normalization is to remove dye bias. We show here that data distributions from different methods are inherently skewed and may be bi-modal at high intensities. It is not possible to estimate how much asymmetry to expect since this will depend upon the method used and global levels of methylation in the samples. We have carefully investigated and applied appropriate methods for these analyses. From these data it is clear that proper normalization is fundamentally reliant on exogenous controls including the spiked *A.thaliana* cDNA used here, but better reagents are needed. Significant collaborative efforts are now underway for designing reproducible control spikes for expression studies (43). It is important to note that use of simplistic location-based normalization in other datasets is likely to have prevented detection of real effects and combining probe-level data between different datasets that have used different methods and comparator DNAs may be impossible to achieve.

We were able to optimize our methods by using bioinformatic tools to identify and annotate the predicted probe sequences on the HCGI12K array and to identify the optimum set of restriction enzyme sites to maximize probe utilization. Optimization of this enzyme set was based on analysis of 5435 CpG island sequences and therefore likely to be of high utility to other CpG island platforms. This may also have contributed to the improved effects seen with the MMASS-v2 method. However, our analysis was limited by the low number of informative probes caused by inclusion of repeat and nonsense sequences from the original library. Improved array platforms with better representation of all CpG islands across the genome, as well as fine mapping within individual CpG islands, are now needed for detailed studies.

MMASS has several advantages over current high-throughput methods; e.g. MMASS in common with DMH employs a universal primer complementary to the ligated adaptor rather than a complex sequence-specific primer

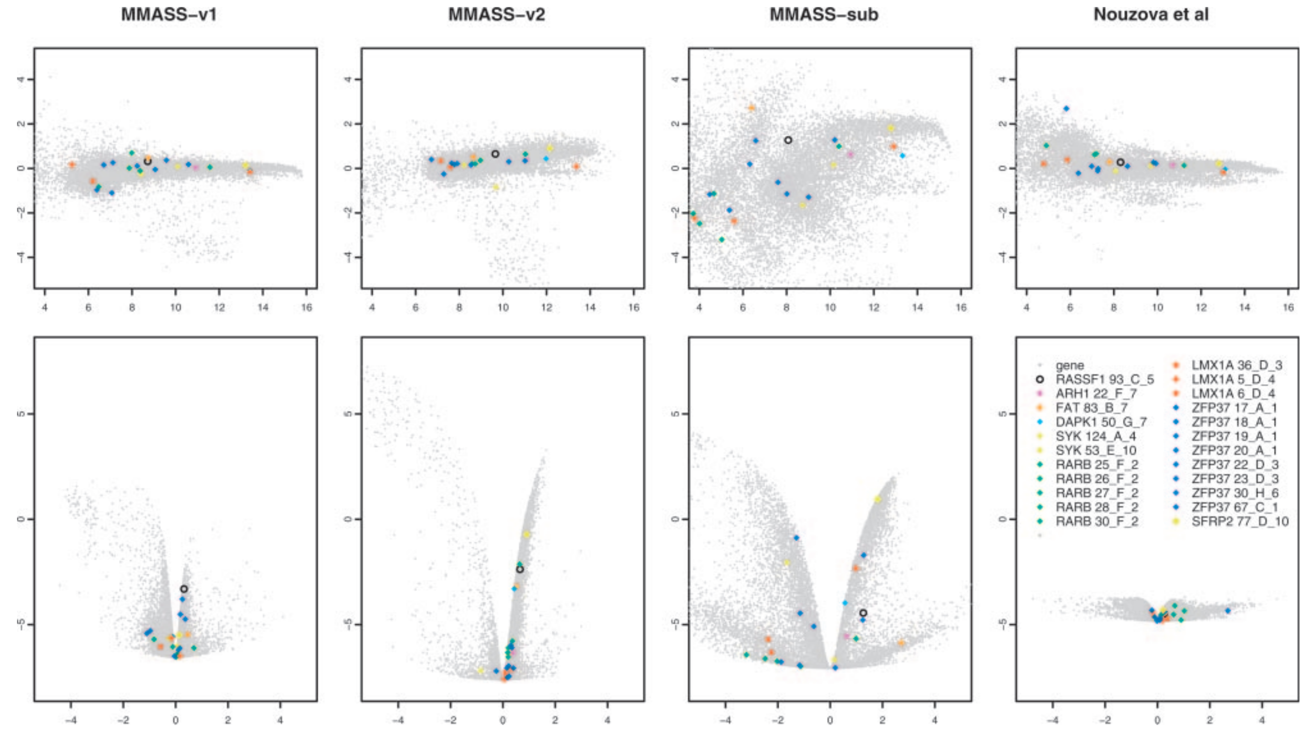

Figure 7. Methylation results for 9 genes (23 independent probes) reported previously to be methylated in HCT116. *MA* plots (upper panel) and volcano plots (lower panel) of linear fit models for each method are shown. Probes representing the same gene are plotted with the same colour. Legend shows gene symbol and probe id.

design such as in methylation-specific oligonucleotide microarray (44) and MALDI mass spectrometry (45). Methylation analysis using BAC microarrays requires the use of rare cutting methylation-sensitive enzymes which limits resolution to a single BAC probe, or only provides an average estimate of methylation across a large genomic region (46,47). Other methods that have used within-sample comparison methylation analysis have either used non-optimized enzyme combinations (48) or complex specific linker/enzyme pairings that have not been shown to improve sensitivity (49). MMASS is able to resolve the overall methylation status of a single copy CpG island probe on a spectrum from mostly unmethylated to mostly methylated. Our results also show that in contrast to DMH and the Nouzova *et al*. (19) method, we were able to detect unmethylated sequences such as housekeeping genes (Figure 6). This will be particularly important in the context of the human epigenome project and for cancer studies where comparison is needed for both methylated and unmethylated sequences (50).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

A.E.K.I. conceived the study, conducted the microarray experiments and drafted the manuscript. N.P.T. participated in the study design as well as manuscript preparation and conducted the analysis. K.B. helped in conducting the microarray and COBRA experiments. N.L.B.M. annotated the array platform sequences and wrote the Perl scripts used for bioinformatic analysis. S.T. participated in data analysis and manuscript preparation. V.P.C. participated in the coordination of the study. A.H.W. participated in the coordination of the study. M.J.A. participated in designing and coordinating the study and manuscript preparation. J.D.B. participated in designing the experiments, coordinating the study, data analysis and manuscript preparation. All authors read and approved the final manuscript. We thank Dr Koichi Ichimura for technical advice on genomic hybridization protocols. This work was funded by grants from Cancer Research UK (CR-UK) and the Department of Pathology, University of Cambridge. A.E.K.I. is a CR-UK Bobby Moore Fellow and J.D.B. is a CR-UK Senior Clinical Research Fellow. N.L.B.-M. is supported by a PhD Fellowship (PRAXIS XXI SFRH/BD/2914/2000) from Fundação para a Ciência e a Tecnologia, Portugal. Funding to pay the Open Access publication charges for this article was provided by grants from Cancer Research-UK.

Conflict of interest statement. None declared.

REFERENCES

1. Wang,Y. and Leung,F.C.C. (2004) An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics*, **20**, 1170–1177.
2. Turker,M.S. and Bestor,T.H. (1997) Formation of methylation patterns in the mammalian genome. *Mutat. Res.*, **386**, 119–130.
3. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes. Dev.*, **16**, 6–21.
4. Jones,P.A. and Laird,P.W. (1999) Cancer epigenetics comes of age. *Nature Genet.*, **21**, 163–167.

5. Herman,J.G. (1999) Hypermethylation of tumor suppressor genes in cancer. *Semin. Cancer Biol.*, **9**, 359–367.
6. Jones,P.A. and Baylin,S.B. (2002) The fundamental role of epigenetic events in cancer. *Nature Rev. Genet.*, **3**, 415–428.
7. Merlo,A., Herman,J.G., Mao,L., Lee,D.J., Gabrielson,E., Burger,P.C., Baylin,S.B. and Sidransky,D. (1995) 5′ CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nature Med.*, **1**, 686–692.
8. Herman,J.G., Latif,F., Weng,Y., Lerman,M.I., Zbar,B., Liu,S., Samid,D., Duan,D.S., Gnarra,J.R. and Linehan,W.M. (1994) Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc. Natl Acad. Sci. USA*, **91**, 9700–9704.
9. Palmisano,W.A., Divine,K.K., Saccomanno,G., Gilliland,F.D., Baylin,S.B., Herman,J.G. and Belinsky,S.A. (2000) Predicting lung cancer by detecting aberrant promoter methylation in sputum. *Cancer Res.*, **60**, 5954–5958.
10. Cui,H., Cruz-Correa,M., Giardiello,F.M., Hutcheon,D.F., Kafonek,D.R., Brandenburg,S., Wu,Y., He,X., Powe,N.R. and Feinberg,A.P. (2003) Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. *Science*, **299**, 1753–1755.
11. Widschwendter,M. and Jones,P.A. (2002) The potential prognostic, predictive, and therapeutic values of DNA methylation in cancer. *Clin. Cancer Res.*, **8**, 17–21.
12. Huang,T.H., Perry,M.R. and Laux,D.E. (1999) Methylation profiling of CpG islands in human breast cancer cells. *Hum. Mol. Genet.*, **8**, 459–470.
13. Wei,S.H., Chen,C.-M., Strathdee,G., Harnsomburana,J., Shyu,C.-R., Rahmatpanah,F., Shi,H., Ng,S.-W., Yan,P.S., Nephew,K.P. *et al.* (2002) Methylation microarray analysis of late-stage ovarian carcinomas distinguishes progression-free survival in patients and identifies candidate epigenetic markers. *Clin. Cancer Res.*, **8**, 2246–2252.
14. Yan,P.S., Chen,C.M., Shi,H., Rahmatpanah,F., Wei,S.H., Caldwell,C.W. and Huang,T.H. (2001) Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. *Cancer Res.*, **61**, 8375–8380.
15. Yan,P.S., Perry,M.R., Laux,D.E., Asare,A.L., Caldwell,C.W. and Huang,T.H. (2000) CpG island arrays: an application toward deciphering epigenetic signatures of breast cancer. *Clin. Cancer Res.*, **6**, 1432–1438.
16. Yan,P.S., Chen,C.-M., Shi,H., Rahmatpanah,F., Wei,S.H. and Huang,T.H.-M. (2002) Applications of CpG island microarrays for high-throughput analysis of DNA methylation. *J. Nutr.*, **132**, S2430–S2434.
17. Yan,P.S., Efferth,T., Chen,H.-L., Lin,J., Rodel,F., Fuzesi,L. and Huang,T.H.-M. (2002) Use of CpG island microarrays to identify colorectal tumors with a high degree of concurrent methylation. *Methods*, **27**, 162–169.
18. Cross,S.H., Charlton,J.A., Nan,X. and Bird,A.P. (1994) Purification of CpG islands using a methylated DNA binding column. *Nature Genet.*, **6**, 236–244.
19. Nouzova,M., Holtan,N., Oshiro,M.M., Isett,R.B., Munoz-Rodriguez,J.L., List,A.F., Narro,M.L., Miller,S.J., Merchant,N.C. and Futscher,B.W. (2004) Epigenomic changes during leukemia cell differentiation: analysis of histone acetylation and cytosine methylation using CpG island microarrays. *J. Pharmacol. Exp. Ther.*, **311**, 968–981.
20. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
21. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al.* (2002) The bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
22. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2001) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
23. Rouquier,S., Trask,B.J., Taviaux,S., van den Engh,G., Diriong,S., Lennon,G.G. and Giorgi,D. (1995) Direct selection of cDNAs using whole chromosomes. *Nucleic Acids Res.*, **23**, 4415–4420.
24. Xiong,Z. and Laird,P.W. (1997) COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic Acids Res.*, **25**, 2532–2534.
25. Sadri,R. and Hornsby,P.J. (1996) Rapid analysis of DNA methylation using new restriction enzyme sites created by bisulfite modification. *Nucleic Acids Res.*, **24**, 5058–5059.
26. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.
27. R Development Core Team (2005) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
28. Leisch,F. and Rossini,A.J. (2003) Reproducible statistical research. *Chance*, **16**, 41–45.
29. Hardle,W. and Ronz,B. (eds) (2002) Sweave: dynamic generation of statistical reports using literature data analysis. In *Proceedings of the conference on Computational Statistics*, Berlin. Physika Verlag, Heidelberg, Germany, pp. 575–580.
30. Gentleman,R. (2004) Reproducible research: a bioinformatics case study. *Stat. Appl. Genet. Mol. Biol.*, **3**.
31. Paz,M.F., Wei,S., Cigudosa,J.C., Rodriguez-Perales,S., Peinado,M.A., Huang,T.H.-M. and Esteller,M. (2003) Genetic unmasking of epigenetically silenced tumor suppressor genes in colon cancer cells deficient in DNA methyltransferases. *Hum. Mol. Genet.*, **12**, 2209–2219.
32. Lind,G.E., Thorstensen,L., Lovig,T., Meling,G.I., Hamelin,R., Rognum,T.O., Esteller,M. and Lothe,R.A. (2004) A CpG island hypermethylation profile of primary colorectal carcinomas and colon cancer cell lines. *Mol. Cancer*, **3**, 28.
33. Dudoit,S., Yang,Y.H., Callow,M.J. and Speed,T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–140.
34. Lonnstedt,I. and Speed,T.P. (2002) Replicated microarray data. *Statistica Sinica*, **12**, 31–46.
35. Groot,G.S. and Kroon,A.M. (1979) Mitochondrial DNA from various organisms does not contain internally methylated cytosine in -CCGG- sequences. *Biochim. Biophys. Acta*, **564**, 355–357.
36. Maekawa,M., Taniguchi,T., Higashi,H., Sugimura,H., Sugano,K. and Kanno,T. (2004) Methylation of mitochondrial DNA is not a useful marker for cancer detection. *Clin. Chem.*, **50**, 1480–1481.
37. Burden,A.F., Manley,N.C., Clark,A.D., Gartler,S.M., Laird,C.D. and Hansen,R.S. (2005) Hemimethylation and non-CpG methylation levels in a promoter region of human LINE-1 (L1) repeated elements. *J. Biol. Chem.*, **280**, 14413–14419.
38. Suzuki,H., Toyota,M., Nojima,M., Mori,M. and Imai,K. (2005) SFRP, a family of new colorectal tumor suppressor candidate genes. *Nippon Rinsho.*, **63**, 707–719.
39. Antoniou,M., Harland,L., Mustoe,T., Williams,S., Holdstock,J., Yague,E., Mulcahy,T., Griffiths,M., Edwards,S., Ioannou,P.A. *et al.* (2003) Transgenes encompassing dual-promoter CpG islands from the human TBP and HNRPA2B1 loci are resistant to heterochromatin-mediated silencing. *Genomics*, **82**, 269–279.
40. Paz,M.F., Fraga,M.F., Avila,S., Guo,M., Pollan,M., Herman,J.G. and Esteller,M. (2003b) A systematic profile of DNA methylation in human cancer cell lines. *Cancer Res.*, **63**, 1114–1121.
41. Leu,Y.-W., Yan,P.S., Fan,M., Jin,V.X., Liu,J.C., Curran,E.M., Welshons,W.V., Wei,S.H., Davuluri,R.V., Plass,C. *et al.* (2004) Loss of estrogen receptor signaling triggers epigenetic silencing of downstream targets in breast cancer. *Cancer Res.*, **64**, 8184–8192.
42. Shi,H., Yan,P.S., Chen,C.-M., Rahmatpanah,F., Lofton-Day,C., Caldwell,C.W. and Huang,T.H.-M. (2002) Expressed CpG island sequence tag microarray for dual screening of DNA hypermethylation and gene silencing in cancer cells. *Cancer Res.*, **62**, 3214–3220.
43. Baker,S.C., Bauer,S.R., Beyer,R.P., Brenton,J.D., Bromley,B., Burrill,J., Causton,H., Conley,M.P., Elespuru,R., Fero,M. *et al.* (2005) The external RNA controls consortium: a progress report. *Nature Methods*, **2**, 731–734.
44. Shi,H., Maier,S., Nimmrich,I., Yan,P.S., Caldwell,C.W., Olek,A. and Huang,T.H.-M. (2003) Oligonucleotide-based microarray for DNA methylation analysis: principles and applications. *J. Cell Biochem.*, **88**, 138–143.
45. Tost,J., Schatz,P., Schuster,M., Berlin,K. and Gut,I.G. (2003) Analysis and accurate quantification of CpG methylation by MALDI mass spectrometry. *Nucleic Acids Res.*, **31**, e50.
46. Weber,M., Davies,J.J., Wittig,D., Oakeley,E.J., Haase,M., Lam,W.L. and Schubeler,D. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genet.*, **37**, 853–862.

47. Ching,T.-T., Maunakea,A.K., Jun,P., Hong,C., Zardo,G., Pinkel,D., Albertson,D.G., Fridlyand,J., Mao,J.-H., Shchors,K. *et al.* (2005) Epigenome analyses using BAC microarrays identify evolutionary conservation of tissue-specific methylation of SHANK3. *Nature Genet.*, **37**, 645–651.
48. Wang,Y., Yu,Q., Cho,A.H., Rondeau,G., Welsh,J., Adamson,E., Mercola,D. and McClelland,M. (2005) Survey of differentially methylated promoters in prostate cancer cell lines. *Neoplasia*, **7**, 748–760.
49. Schumacher,A., Kapranov,P., Kaminsky,Z., Flanagan,J., Assadzadeh,A., Yau,P., Virtanen,C., Winegarden,N., Cheng,J., Gingeras,T. *et al.* (2006) Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res.*, **34**, 528–542.
50. Wu,H., Chen,Y., Liang,J., Shi,B., Wu,G., Zhang,Y., Wang,D., Li,R., Yi,X., Zhang,H. *et al.* (2005) Hypomethylation-linked activation of PAX2 mediates tamoxifen-stimulated endometrial carcinogenesis. *Nature*, **438**, 981–987.