Comprehensive DNA methylation profiling in a human cancer genome identifies novel epigenetic targets


Orion Genomics, St Louis, MO, 2Department of Agronomy, Lilly Hall of LifeSciences 915 West State Street, Purdue University, West Lafayette, IN, 3Department of Statistics, Purdue University, 150 North University Street, West Lafayette, IN, 4Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX and 5Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

*To whom correspondence should be addressed at Science and Technology, Orion Genomics, LLC, 4041 Forest Park Avenue St. Louis, MO 63108, USA. Tel: +1 314.615.6977; Fax: +1 314.615.6975; Email: jeddolah@oriongenomics.com

Using a unique microarray platform for cytosine methylation profiling, the DNA methylation landscape of the human genome was monitored at more than 21,000 sites, including 79% of the annotated transcriptional start sites (TSS). Analysis of an oligodendroglialoma derived cell line LN-18 revealed more than 4000 methylated TSS. The gene-centric analysis indicated a complex pattern of DNA methylation exists along each autosome, with a trend of increasing density approaching the telomeres. Remarkably, 2% of CpG islands (CGI) were densely methylated, and 17% had significant levels of 5 mC, whether or not they corresponded to a TSS. Substantial independent verification, obtained from 95 loci, suggested that this approach is capable of large scale detection of cytosine methylation with an accuracy approaching 90%. In addition, we detected large genomic domains that are also susceptible to DNA methylation reinforced inactivation, such as the HOX cluster on chromosome 7 (CH7). Extrapolation from the data suggests that more than 2000 genomic loci may be susceptible to methylation and associated inactivation, and most have yet to be identified. Finally, we report six new targets of epigenetic inactivation (IRX3, WNT10A, WNT6, RARα, BMP7 and ZGPAT). These targets displayed cell line and tumor specific differential methylation when compared with normal brain samples, suggesting they may have utility as biomarkers. Uniquely, hypermethylation of the CGI within an IRX3 exon was correlated with over-expression of IRX3 in tumor tissues and cell lines relative to normal brain samples.

Introduction

Identifying molecular differences that distinguish tumor tissue from normal tissue is a current topic area of intense interest. Although, tumor genomes display only a limited number of primary sequence differences from the nearly isogenic normal tissues in proximity to them (1), a large number of molecular differences exist. In particular, the spectrum of sequences that normal and tumor genomes specify and mark as silent chromatin have been used as ‘epigenetic signatures’ to molecularly discriminate these cells (2).

Disruption of normal gene regulation is important for carcinogenesis (3,4) resulting in loss, or gain of genetic function. The molecular events that underlie this altered regulation include point-mutations and macro-mutations, such as deletion, amplification or genomic rearrangement (e.g. translocation), that can result in more complex interactions when regulatory genes are affected. Recently, the importance of epigenetic perturbation of gene regulation in the form of changes in chromatin structure has begun to be more fully appreciated (5). In the context of cancer, inappropriate chromatin packaging of genes can lead to gene silencing, or in some cases, ectopic gene expression (6).

Cytosine methylation is a chemically stable mark that may establish, or follow as a consequence of, the packaging of a particular region into silent chromatin. Therefore identification of aberrant genomic DNA methylation that is associated with carcinogenesis should identify loci that are important for disease progression (2).

Mammalian DNA methylation patterns that transmit cellular silencing signals are mitotically maintained with 96–99% fidelity (7). This lies in stark contrast with the primary sequence which is maintained with fidelity over 99.9999% (8). The scale of the difference suggests that epigenetic deregulation may be under appreciated, and recent studies have highlighted the role of the environment, e.g. through dietary folate metabolism, in maintaining DNA methylation and gene silencing, hinting at a mechanism underlying predisposition for cancer (9,10).

Several cancer therapies targeting the maintenance of cytosine methylation and silencing states are in human clinical trials (11–13). Currently, the therapies target either the DNA methylation machinery, or the histone modification machinery. This machinery works synergistically to maintain gene silencing. While such therapies are very promising, the clinical success rates achieved thus far have been similar to more conventional chemotherapies. Therefore, selecting patients for...
such epigenetic therapies or measuring their success may require an understanding and characterization of the sequences affected by epigenetic perturbation in particular diseases.

Finding and characterizing the genomic loci capable of driving carcinogenesis from the epigenetic perspective, as well as those capable of serving as clinically meaningful disease or therapy-specific markers is a pressing need. We chose to address the need through the adaptation of a microarray-based DNA methylation profiling approach for use in the human genome. We report here the results obtained in the course of characterizing the oligodendroglioma derived cell line LN-18 (CRL-26110). Future applications of this technology will serve not only to directly identify those sequences which undergo epigenetic alteration in disease, but also as an epigenetic ‘Ames’ test for environmental history.

Materials and methods

Cell line, normal and tumor brain tissue and blood nucleic acid samples

Cell lines LN-18 (glioblastoma), A-172 (glioblastoma), LN-229 (glioblastoma), T-98G [glioblastoma multiforme (GBM)]; U-138 MG (astrocytoma) and U-87 MG (astrocytoma) were obtained from American Type Culture Collection (Manassas, VA) and cultured under supplier’s recommendations under the manufacturer’s recommended conditions (http://www.atcc.org/). Histologically normal/non-malignant brain tissue (~0.4 mg ea) from the cerebrum (right temporal lobe) of three different patients (15084A1, 16920A1 and 8387B1) were obtained from Astarand plc (Detroit, MI). Primary tumor samples [two astrocytoma (1FMXPFB, JACCIFNL) and one GBM specimen (UAUFPFJD)] were acquired from Genomics Collaborative Inc (Caimbridge, MA). All the tumor samples exhibited >90% neoplastic cellularity. The age range of the normal samples (45.56,87) was controlled relative to the tumors (48.59-59). Genomic DNA was extracted by the MasterPure DNA Purification kit under manufacturer’s recommended conditions (Epicentre Biotechnologies, Madison, WI). Male whole blood genomic DNA, representing a pool of peripheral blood samples obtained from approximately six individuals, was obtained from Novagen (Madison, WI).

RNA was purified from the tumor tissues and cell lines using the RNasey Kit from Qiagen (Valencia, CA). cDNA from normal brain and astrocytoma tissues was purchased from Invitrogen (Carlsbad, CA); cDNA from the cell lines and tumor samples was produced using the oligo dT primers provided in the Superscript III first strand system from Invitrogen (Carlsbad, CA) under the manufacturer’s recommended conditions.

Microarray experimental and analytical approach

Methylation profiling was performed as described previously (14,15) but with adaptations to apply the method to the human genome. The adaptations included the use of target mass normalization before labeling, the employment of long oligonucleotide probes (60mer) on the NimbleGen systems CGH platform, and the use of sequences within the human genome that are incapable of digestion by McrBC as controls.

Methylation-dependent DNA fractionation

LN-18 DNA was mechanically sheared into a uniform molecular weight distribution using a GeneMachines HydroShear apparatus. Sheared DNA was split into four equal portions of 100 μg each. Two portions (treated technical replicates) were digested with McrBC (NEB) in 300 μl total volume including 1× NEB buffer 2.0, 0.1 mg/ml BSA, 2 mM GTP and 200 U McrBC. The remaining two portions were mock-treated under identical conditions except that 20 μl water was added instead of McrBC. Treated and mock-treated reactions were incubated at 37°C overnight. All reactions were treated with 5 μl proteinase K (50 mg/ml) for 1 h at 50°C, and precipitated with EtOH under standard conditions. Pellets were washed twice, dried and resuspended in 50 μl H2O. A total of 50 μg of each fraction was resolved on a 1% low melting point SeaPlaque GTG agarose gel (Cambridge Bio Sciences, Rockland, ME). Untreated and treated replicates were resolved side-by-side. DNA (1 kb) sizing ladder was run adjacent to each untreated/treated pair to guide accurate gel slice excision. Gels were visualized with long-wave ultraviolet (UV), and gel slices including DNA within the modal size range of the untreated fraction (~1–4 kb) were excised with a clean razor blade. DNA was extracted from gel slices using the GenElute agarose spin columns (Sigma–Aldrich).

Human array (OGHA v1.0) design

To identify epigenetic changes in gene regulation, our efforts were focused largely upon the transcription start sites of genes. Since promoter annotation and TSS identification are generally less reliable than gene annotation, we tested the ENSEMBL gene predictions using a highly characterized set of human promoters from the Eukaryotic Promoter Database (EPD, http://www.epd.isb-sib.ch). Substantial similarity was observed between ENSEMBL genepanel SeaPlaque GTG Agarose gel (Cambridge Bio Sciences, Rockland, ME). Untreated and treated replicates were resolved side-by-side. DNA (1 kb) sizing ladder was run adjacent to each untreated/treated pair to guide accurate gel slice excision. Gels were visualized with long-wave ultraviolet (UV), and gel slices including DNA within the modal size range of the untreated fraction (~1–4 kb) were excised with a clean razor blade. DNA was extracted from gel slices using the GenElute agarose spin columns (Sigma–Aldrich).

Human array (OGHA v1.0) design

To identify epigenetic changes in gene regulation, our efforts were focused largely upon the transcription start sites of genes. Since promoter annotation and TSS identification are generally less reliable than gene annotation, we tested the ENSEMBL gene predictions using a highly characterized set of human promoters from the Eukaryotic Promoter Database (EPD, http://www.epd.isb-sib.ch). Substantial similarity was observed between ENSEMBL genepanel SeaPlaque GTG Agarose gel (Cambridge Bio Sciences, Rockland, ME). Untreated and treated replicates were resolved side-by-side. DNA (1 kb) sizing ladder was run adjacent to each untreated/treated pair to guide accurate gel slice excision. Gels were visualized with long-wave ultraviolet (UV), and gel slices including DNA within the modal size range of the untreated fraction (~1–4 kb) were excised with a clean razor blade. DNA was extracted from gel slices using the GenElute agarose spin columns (Sigma–Aldrich).

Dye labeling and microarray hybridization

Each fraction was labeled with Cy5 and separately with Cy5 by in vitro synthesis reactions. A total of 200 ng of fractionated template DNA and 20 μl 2.5× random hexamer cocktail (Invitrogen) were combined in a total volume of 25.5 μl and incubated in a boiling water bath for 5 min. Reactions were placed on ice for 3 min. Reaction volumes were adjusted to 50 μl including 140 μM each dATP, dGTP and dTTP (Invitrogen). 1.5 μl 25 mmol Cyamine 3-3CTP or Cyanine 5-3CTP (NEL), and 100 U 3′-exo’ Klenow (NEB). Reactions were incubated at 37°C for 45 min then inactivated by addition of 50 μl TE. Labeled DNA was purified using MinElute spin columns (Qiagen) and eluted in 35 μl EB buffer. Synthesis yields and Cy-dye specific activities were determined by Nanodrop spectrophotometry. One-half of each labeled reaction was used per microarray hybridization. A duplicated dye-swap experimental design was employed and four microarray hybridizations were performed. Microarray hybridization, washing and scanning were performed by NimbleGen Systems of Iceland under their standard CGH operating protocols.

Statistical analysis of microarray data

A two-stage ANOVA model was employed (16). In the first stage the model fit was: \( y_{ijkl} = \mu + T_i + A_j + D_k + e_{ijkl} \), where \( y = \text{log}(\text{intensity}), \mu = \text{global mean}, T = \text{fixed treatment effect}, A = \text{fixed array effect}, j = 1, 2, 3, 4; D = \text{fixed dye effect}, k = 1, 2; g = 1, 2, \ldots, 21,294 (\text{gene}), f = 1, \ldots, 48; (\text{some genes have replicates for combinations of } T, A \text{ and } D), e_{ijkl} = \text{error, iid } \sim N(0, \sigma^2). \)

Let \( T_1 \) and \( T_2 \) correspond to untreated samples (UT) and treated methylation-depleted (MD) samples, respectively. In the second stage of the ANOVA the residual for each gene \( g \), from the first stage analysis, \( e_{ijkl} \), was used as the response variable in the model: \( e_{ijkl} = G_g + (GT)_g + (GA)_g + (GD)_g + (GTD)_{ijkl} \). Where the error terms, \( e_{ijkl} \), are assumed to be independent \( N(0, \sigma_g^2) \).

Normalization to RCG zero features

Subsequent analysis of the LN-18 data utilized those features having no McrBC half-sites within 1000 bp of genomic sequence spanning the probe [i.e. RCG/kb frequency values of zero (RCG 0)] as controls. One would not expect the RCG0 features to be affected by McrBC due to their lack of McrBC half-sites. As such, any variability observed from these features’ intensities should be solely attributed to the overall treatment, array and dye effects, plus chance variability. Using the RCG0 features to estimate these effects allows for a clearer picture of the methylation status of the other features. For this reason, the first-stage ANOVA model was fit using only the RCG0 features. The estimates of \( \mu, T_1, A_j \) and \( D_k \) were obtained

J.M.Ordway et al.
Comprehensive DNA methylation identifies novel epigenetic targets

For each analyzed genomic DNA sample, 4 µg of genomic DNA was digested with McrBC (NEB) in 100 µl total volume including 1X NEB2 buffer, 0.1 mg/µl BSA, 2 mM GTP and 50 U McrBC overnight at 37°C. In parallel, 4 µg of each genomic DNA sample was mock-treated under identical conditions with the exception that water was substituted for McrBC. Following digestion, enzyme activity was heat inactivated at 65°C for 20 min. A total of 40 ng McrBC-digested and 40 ng mock-treated DNA were analyzed by quantitative real-time PCR in all cases, digested and mock-treated templates were analyzed in adjacent wells. Reactions included 1 µl of each genomic DNA sample and CtMSRE is the cycle-threshold of the Mock (untreated) sample

\[ R = 2 - W, \]

where R is the proportion of fragments that are refractory to treatment.

The amount of DNA remaining after each treatment can be calculated using

\[ \text{DM} = 2^{-\text{(CtMSRE-CtM)}}, \]

where DM is the proportion of densely methylated template DNA. U is the proportion of Unmethylated DNA, CtMSRE is the Cycle-threshold of MSRE sample and CtM is the cycle-threshold for the Mock (untreated) sample.

Confidence in the results obtained using the MSRE restriction one would expect them to be destroyed in a DD. Conversely, if the unmethylated gene copies are in fact present following MSRE restriction one would expect them to be destroyed by the DD. Confidence in the DM and U calls made in this way is assessed by the size of Ct difference between DD and the MSRE/MRDRE. Large Ct differences elicit higher confidence. Measurements with a ΔCt less than 0.5 cannot be resolved apart from machine error.

\[ IM = 1 - \text{DM} - U, \]

where IM is the proportion of Intermediately methylated (IM) DNA. Some molecules can be methylated such that both enzymes are capable of cutting them. That is, they contain 5 mC but are not completely methylated. They exist as part of each of the above calculations. Typically the fraction of IM molecules are identified and calculated only if the ΔCt of the UT-MSRE and UT-MDRE are each above or near 1.0 (e.g. >0.7). An important note about this method is that the IM copy calculation contains both the error functions associated with the MD and U calculations; typically adjusting this calculation for the amount refractory will not affect the calculation.

All ΔCt calculations yield a unit-less parameter reflecting relative proportion of template DNA between various treatments. The fold treatment effects can be applied to the total calculation so that each of the molecular proportions (DM, U and IM) can be converted back into a unit of copies.

RT-PCR expression analysis

A five point (near, 1.5, 1.25, 1.125, 1.625) 5-fold dilution curve of cDNA into water was used as template for limited cycle end-point PCR. No-template (water) and genomic DNA (positive) controls were run in parallel. Primers designed to exonic sequences which span the 3' most intron in either the IRX3 gene or the GAPDH gene were used to monitor expression. The IRX3 primers (Supplementary Table) would amplify a 463 bp product if genomic DNA were the template, while cDNA should yield a product of 293 bp. Similarly, the GAPDH control primers (Supplementary Table)
RESULTS

Methylation profiling was performed using the methylation-dependent restriction enzyme McrBC to separate methylated and unmethylated DNA fragments for labeling and hybridization to microarrays (14). The profiling scheme is depicted in Figure 1. First, purified genomic DNA from a sample of interest is randomly sheared into a uniform size range. The sheared DNA is then split into four portions: two mock treated and two McrBC-digested. In order to restrict DNA, McrBC requires that two methylated half sites (RmC, where R = A or G) lie within 40–3000 bp of each other (optimal spacing 55–103 bp); cutting has been demonstrated to occur between the two half sites in the proximity of one half-site (20–27). Each matched treatment pair was then size fractionated by electrophoresis and the high-molecular weight DNA that was resistant to treatment was purified from excised agarose gel slices. We have employed independent digests so that any variance during digestion can be assessed across technical replicates. The purified DNA represents the whole genome in the mock-treated sample and the unmethylated fraction of the genome in the McrBC-treated sample.

Next, the same mass of the whole genome sample (untreated, UT) and the methylation-depleted genome sample (treated, T) was independently labeled with different fluorophores. The ratio of mock-treated or ‘untreated’ genome (UT) to methylation-depleted or ‘treated’ genome (T) dye-signal is obtained for each feature (UT/T). Digestion with McrBC removed more than half of the DNA from the T sample, but because the same amount of T and UT DNA was used for each labeling, unmethylated sequences will be enriched in the T sample. This mass normalization step increases the signal from unmethylated sequences, and makes these measurements more reproducible, but requires careful subsequent normalization to internal control probes. The ideal set of control probes are sequences which are devoid of McrBC half-sites for a large stretch of sequence (i.e. >1000 or >2000 bp). The resultant two color sample datasets yield hybridization ratios for each feature, and these are assigned a color: blue (low UT/T ratio = unmethylated), yellow (moderate UT/T ratio) or red (high UT/T ratio = methylated) in Figure 1. Sample analysis utilized a duplicated dye swap on pairs of independently treated and UTs for each genome, requiring four arrays and allowing a balanced Latin-square analysis design (See Materials and methods).

Methylated sequences will have a ratio UT/T that approaches infinity (because of the depletion of methylated DNA in the T sample) while unmethylated sequences will have a low UT/T ratio that should be less than 1 because of the enrichment caused by the mass normalization of the UT and T targets. In practice, because the genomic DNA is sheared to a finite size, probes that are adjacent to methylated sequences are also capable of detecting that methylation (feature Y, Figure 1; yellow and orange probes, Figure 2). The distance at which probes can detect adjacent methylation is related to the degree of shearing, and was termed ‘wingspan’.

DNA methylation profile of LN-18

The stage IV oligodendroglioma derived cell line LN-18 was chosen as the subject genome for this experiment because it was previously demonstrated to be homozygous for a deletion of the p16(CDKN2A) locus on CH9, which acted as a control region for ploidy assessment (see below). Features representing the CDKN2A region were included on the microarray in addition to the CpG islands (CGI) and TSS throughout the annotated genome (above and Materials and methods). The features in Figure 2 are colored according to the representative methylation profile data [Robust Multi-array Average (28) (performed by NG Systems, WI), linear interpretation] obtained from a duplicated dye swap analysis of LN-18. The color scheme is depicted as a gradient from blue (no methylation) to red (highest-methylation). Yellow features detect intermediate levels of methylation or methylation in regions adjacent to the probe due to wingspan. A substantial number of methylated regions (red) are evident, as well as a large number of unmethylated regions (blue). Globally, each autosome displayed a trend of increasing methylation from pericentromeric regions to telomeres (Figures 2 and 5).

The results from the statistical analysis of the LN-18 genome’s cytosine methylation profile are depicted in Figure 3. The signal at each feature was first corrected for technical and biological variation using ANOVA (see Materials and methods), and then normalized to features that represent 1 kb segments of the human genome that contain no McrBC half sites (RCG = 0/kb, denoted hereafter as RCG0). The normalized data were then tested using a
null-hypothesis in which whole genome (UT) signal should be the same as the methyl-depleted (T) signal if the sequences surrounding a given feature were unmethylated. Panel A depicts a graph of the LN-18 methylation ratio ($y$-axis) by feature name ($x$-axis). The methylation data are depicted as the ANOVA corrected signal ratio $[\log_2 (UT/T)]$. Each gene was allowed to have its own variance. In this analysis, methylated sequences are expected to have a positive log ratio, while unmethylated sequences are expected to have a log ratio of zero (corresponding to a UT/T ratio of 1.0). The data points are colored according to the multiple comparison corrected $P$-values based upon an
alpha level of 0.05. Blue data points passed the Holm multiple testing correction and have a P-value <0.001. Pink data points passed the FDR correction and have a P-value <0.05. Ratios that are not significantly different from the control RCG0 features are depicted in gray. Panel B depicts the same data using the respective t-test statistic values (y-axis) with the same x-axis. The t-test data representation allows a visual assessment of the variation on a per feature basis since the t-test values are the log2 (UT/T) signal divided by the standard error for that probe; it corresponds directly with the significance results of panel A.

Table I summarizes the ANOVA results. Approximately 60 loci were used as RCG0 controls for normalization (Table I). Over half of the 21047 measurements detected statistically significant DNA methylation levels relative to the controls, with 4152 significant methylated regions (log ratio > 0) and 6994 significantly unmethylated regions (log ratio < 0) (Table I). The existence of unmethylated regions suggests that the controls could in fact detect some methylation, probably because of ‘wingspan’ from surrounding, RCG containing features.

CpG Islands (CGI) have been characterized as CG rich genomic sequences that do not show CG depletion; they are often in association with genes [(29), for a review see (30)]. CGI are generally thought to be devoid of 5mC, which accounts for the absence of CG depletion which is driven by the mutagenesis of 5mC over evolutionary time, (31–33). Aberrant cytosine methylation of CGI has been associated with inactivation of tumor suppressor genes in human cancers, consequently the DNA methylation status of CGI is of immense interest (5). There are 9786 CGI probes in our design, which reflects roughly 1/3 of the total CGI content in the human genome. The BAC based CGI designed features performed similarly to the UCSC annotated CGI, both were present in the same frequency in both the significant methylated and unmethylated loci (Table I).

As expected, very few (1–2%) of the 9786 CGI features reported dense methylation, whether or not they corresponded to known TSS. However, almost 1 in 3 CG islands had low levels of methylation. In these cases, methylation may be confined to sequences near, but not within the CGI. Features in the region would still detect signal because of ‘wingspan’. Bisulphite sequencing and other validation methods were used to test this idea, as described below.

**Independent verification suggests high accuracy with a quantitative capacity**

Independent verification of DNA methylation levels was performed using two methods. The first was clone-based bisulphite sequencing. 13 loci were selected from a range of the microarray data, and 1 locus devoid of RCG motifs was selected. More than 10 clones were analyzed per locus (19). The locus devoid of RCG motifs was substantially methylated, and because no McrBC half-sties were in proximity of the probe it yielded a negative log ratio (Supplementary Table). Of the 13 remaining loci, 11 yielded statistically significant measurements: 6 were methylated and 5 were unmethylated according to the microarray experiments. In 9 of the 11 cases, the methylation occupancy information determined by bisulphite sequencing agreed with the array prediction (Figure 5A) (Supplementary Table). Interestingly, the data from the 9 concordant suggested that there may be a relationship between the array’s output ratio value and the average regional 5mC occupancy (Figure 4A–C). In the two discordant cases, the data were consistent with a wingspan effect upon the probes (Figure 4A, two points off line). Panel B contains bar-charts depicting the average methylation density at each locus. For each locale the y-axis depicts the average methylation occupancy at the CGs within the analyzed interval. The CGs are spaced along the x-axis by their relative position.

The second verification technique employed a more high-throughput quantitative PCR approach recently developed by
Fig. 4. Independent accuracy analysis suggests quantitative capacity within array analysis. (A) A scatter plot of the average 5 mC density determined by bisulphite sequencing obtained from the 11 statistically significant loci are plotted versus the normalized log₂ (UT/T) for each feature. The $R^2$ value obtained reflects the result of a linear correlation analysis. Two of the 11 were considered to be discordant with the regression line. (B) Bar charts depicting high resolution cytosine methylation analysis from loci including the points i, ii and iii. Each of the charts indicate the position of each CG analyzed within each window on the x-axis, and the average relative 5 mC occupancy for greater than 10 clones at each site within the locus on the y-axis. The ANOVA corrected signal log ratio for each locus is indicated, along with the multiple testing correction $P$-value intervals the measurements fell within, are the inset. The position of the array probe within the interval is denoted by the black bar in each window. (C) A scatter-plot of methylation values for 84 genomic loci is shown. Signal ratios from MethylScope are plotted against the change in cycle threshold (ΔCt) measured by MethylScreen quantitative PCR (34). The data are categorized into four quadrants: In clockwise order, false negatives, methylated predictions, false positives and unmethylated predictions. Methylated and unmethylated features are those for which both assays agree. A cut-off of 0.5 cycles is used for the MethylScreen data, as explained in the text (34). The data points are colored according to their significance level: FDR=open circle, Holm=filled circle, grey points are not significant.
our group [MethylScreen, see Materials and methods and (34)]. We randomly selected 50 loci without regard to the magnitude of the microarray measurement, and selected an additional 34 loci representing the range of measurements. We performed MethylScreen upon these 84 loci, 4 of which were included in the bisulphite sequencing study. Figure 4C depicts the log ratio of UT/T (x-axis) plotted against the ΔCt obtained by MethylScreen (y-axis). Like Figure 3A, the data points are colored by their statistical significance level, grey data were not significant. The limit of detection of methylation by MethylScreen was set to a ΔCt of 0.5 cycles. This was because under ideal circumstances, quantitative PCR has a platform error of ±0.5 Ct. Using these criteria, the upper left quadrant represents false negatives in which significantly hypermethylated regions were not detected by the array (6 total). The upper right quadrant represents methylated features (25 total) as determined by both assays. The lower right quadrant represents false positives (3 total). Finally, the lower left quadrant represents unmethylated features (25 total). A total of 25 of the randomly selected loci were no more or less methylated than the controls and so met the null hypothesis (gray). Of these, only 5 were substantially methylated according to MethylScreen, indicating that only a small number of features in this class failed to be detected by the array hybridization, with the remaining being unmethylated. Altogether, the qPCR (MethylScreen) and microarray (MethylScope) approaches agreed 85% of the time. These validation experiments, along with the bisulphite sequencing data, indicate that the MethylScreen and MethylScope technologies are highly reproducible, and support the view that the MethylScope array hybridization approach generates an accurate characterization of genome-wide cytosine methylation. In the few cases where they disagree, the level of methylation was almost always very low, and in only three cases (one false negative and two false positives) did the MethylScope and MethylScreen techniques differ markedly. We cannot exclude the possibility that both ‘false positives’ were detecting methylation adjacent to the probe.

Simultaneous methylation profile and copy number assessment

LN-18 is homozygous for a deletion of the cell cycle regulatory gene CDKN2A (35), and so we included 161 probes from this region on the array as controls. Close inspection of Figure 3A appeared to indicate there was a lack of hybridization of either the UT or T fractions to the probe elements (black arrow) alphabetically between the CGI features (red line) and the promoter features (black line). The 161 probes dedicated to CDKN2A/B fall into this region. In an attempt to confirm that the array could detect the absence of the CDKN2A locus, we sorted the t-values (signal/error) such that the features were depicted in chromosome order. CH7, CH8 and CH9 are illustrated in Figure 5. Another region of the genome with a similar feature density to that present within the CDKN2A/B region is the HOX gene cluster on CH7, represented by 41 features. The boxes highlight each region in Figure 5. The HOX cluster is hypermethylated while the CDKN2A region on CH9 has no significant signal. The mean of the t-values corresponding to the 161 features from the CDKN2A region is −1.8, with a standard deviation of 1.3, indicating that for the vast majority of the probes the signal (UT/T) was of the same order as that of its standard error, and therefore not above background (Supplementary Data). We estimated the size of the deletion to be <12 Mb; it likely lies between base pair positions CH9: 19219777 and 31243697.

LN-18 displays HOX cluster hypermethylation

The HOX interval is ~165 kb, and our array contains approximately 1 feature per 4 kb. The values from our analysis are indicated as a blue line in Figure 6. Methylation profiling data from the same region was previously obtained from colon cancer and fibroblast cell lines, by using immunoprecipitation with antibodies raised against 5 mC and overlapping BAC clones on the array (36), and these data are indicated as a green line in Figure 6. In this particular region, the BACs were larger than average, with more than 80 kb between the midpoints of the BACS. The TSS for the genes in the interval are depicted as arrows pointed in the direction of their transcription (not to scale). Thus, while both analyses indicate methylation, locally unmethylated regions could also be detected using our higher resolution approach, afforded by a higher density and smaller size of probes used. Unmethylated regions included the TSS from only 3 of the HOXA genes surveyed (see Discussion).

Indications of biological relevance

In an effort to discern whether or not the methylation events detected were biologically meaningful, we ran the 25 quantitative PCR assays that detected regional methylation against a panel of genomic DNAs from six brain cancer derived cell lines, as well as pooled male and female peripheral blood which were collected independently from the tumor origin. We employed the CGI in the TH2B gene as a positive control for single copy gene associated 5 mC detection (37). Among these loci, 8 were differentially methylated in the majority of cell lines relative to peripheral blood (Figure 7). Two of these 8 loci were previously demonstrated to be subject to DNA methylation mediated gene silencing; RASSF1A (38) and E-Cadherin (CDH1) (39). The remaining six genes that displayed
Comprehensive DNA methylation identifies novel epigenetic targets

hypermethylation relative to blood, and normal brain tissue have not been reported to be subject to epigenetic silencing. These newly discovered methylated loci were associated with the IRX3, BMP7, WNT10A, WNT6, RARa and ZGPAT genes. The same panel of loci was differentially methylated in a comparison of primary brain tissue from three apparently normal patients in comparison to primary tumor tissues derived from two astrocytomas and one glioma (GBM).

**IRX3 CGI methylation is correlated with IRX3 overexpression**

Interestingly, the methylated CGI detected at the IRX3 locus corresponds to an exon rather than the gene’s promoter (Figure 8). A high resolution analysis of the annotation and DNA methylation data obtained from the MethylScope analysis of the LN-18 genome is depicted in Figure 8A. The promoter of IRX3 was significantly unmethylated, while the CGI in exon 2 was methylated. Bisulphite genomic sequence was obtained from the IRX3 exonic CGI (Figure 8B). The graph displays the average methylation occupancy per CG in the interval when considering the 38 clones sequenced, the X and Y axes are the same as those in Figure 4B. MethylScreen and bisulphite sequencing analyses demonstrated that this region is unmethylated in normal blood and normal brain DNA (Figure 7 and data not shown). The position of the array feature is denoted by the gray bar in Figure 8B. There are six HhaI restriction sites in the analyzed region, five of which appear to be nearly always occupied. The MethylScreen amplicon designed for the qPCR analysis presented in Figure 7 surveys all six of these HhaI sites. This amplicon was then employed in a MethylScreen assay (Figure 9). The kinetic profiles obtained from each of the four conditions of the assay obtained from each genomic sample are displayed. The data depict results from cell lines (upper row), normal brain tissue (middle row) and three brain tumors (lower row); each genome’s results are color coded by their treatment: mock restriction (red), HhaI restriction (green), McrBC restriction (blue) and a HhaI +...
McrBC DD (orange). The pie chart insets display the result of each assay where red = fraction of molecules with all HhaI sites occupied, green = fraction of molecules devoid of RmC methylation between the primers, and the yellow = fraction of molecules susceptible to both treatments (see Materials and methods for calculations). The exon appeared to be hypermethylated in all the tumors and cell lines relative to the three independent normal brain samples, reminiscent of the differentially methylated regions (DMR) in imprinted genes like IGF2 and H19. We therefore

Fig. 8. High resolution analysis of IRX3 hypermethylation reveals an exonic rather than promoter location. (A) A gbrowse view of the NCBIv35 genome build with the ENSEMBL_36 annotation is depicted along with the LN-18 microarray measurements of local DNA methylation. The arrow scale at the top denotes the base pair position along the chromosome. The first data track contains a bar representing the IRX3 locus, the arrow depicts its direction of transcription. The second track is the splicing model of the transcript as determined by ENSEMBL_36. The third data track (RCG) depicts the relative abundance of methylatable McrBC recognition sites within a defined 1 kb genomic window. The fourth track (OGHA) denotes the positions of the two array features (60mer probes). The fifth track depicts a scaled black bar chart of the ANOVA corrected array measurements for the two features Log2 (UT/T); the actual measurements along with their P-values are indicated below each feature. Positive log ratios indicate 5 mC and negative log ratios indicate its relative absence. (B) Bisulphite sequencing results from analysis of the LN-18 methylation pattern surrounding the array feature are depicted as a scatter plot. The x-axis depicts the relative position of each base pair within the interval. The average methylation occupancy at each CG is represented as an open circle. The area under the dashed line between points represents the methylation density. Black arrows denote the position of HhaI restriction sites in the interval (one was outside the sequenced region). The black circles represent methylation occupancies of CGs at HhaI restriction sites within the interval. The gray bar is the position of the array probe.
hypothesized that IRX3 may be over-expressed in tumors and cell lines (due to its intragenic position) rather than silenced. Since the MethylScreen results from normal brain tissues are not consistent with hemizygous methylation, we do not suspect IRX3 as being imprinted (Figure 9, middle row).

The expression of IRX3 in tumor and normal brain samples was assessed using semi-quantitative PCR amplification of dilutions of oligo-dT primed cDNA libraries. IRX3 expression was measured relative to GAPDH expression as a positive control (Figure 10A). The MethylScreen IRX3 DNA methylation result from each sample is depicted in Figure 10B. The expression amplicon selected for each target spanned an intron of ~100 bp or more to eliminate genomic DNA contamination of the cDNA libraries. In all cases, the level of DNA contamination was negligible (data not shown). Comparison of amplification from each dilution indicated that all three of the tumors expressed between 5- and 8-fold more IRX3 than the normal tissue library (with a SD of 3-fold). Commercially acquired astrocytoma cDNA libraries behaved similarly (data not shown), although matched DNA samples were not available. LN-18 produced more than 20-fold more IRX3 than normal brain, and had the most methylated DNA (Figures 10A and B, 9). The U87MG cell line expressed ~4-fold less IRX3 than LN-18, but ~5-fold more than the normal tissue, making its expression level most similar to the tumor libraries (data not shown). Interestingly, it was also the sample with a methylation pattern most similar to the tumors (Figure 9).

**DISCUSSION**

**Precise, accurate and quantitative whole genome DNA methylation profiles**

DNA methylation is a stable epigenetic signal associated with gene silencing (2), and quantitative measurement of DNA methylation during disease has immense appeal as a diagnostic tool, especially in cancer. Several DNA methylation profiling approaches have been developed (14,36,40–50). Few have demonstrated the ability to identify both methylated and unmethylated sequences with statistical confidence, and fewer yet, have the capacity to monitor the entire genome.

Biologically meaningful changes in DNA methylation typically occur over regions with multiple CG dinucleotides,
which are methylated in concert. Thus, regions of methylated sequence rather than single bases may be the most important unit of information in the epigenome. The unique enzymatic activity of McrBC is well suited for identifying these regions, as it recognizes at least half of the potentially methylated CG-dinucleotides in the human genome (RCG half sites); no other enzymatic system has such a broad capacity. Further, the methods described here monitor methylation of repeated sequence elements as well as unique sequences. Because each array feature has the capacity to detect methylation adjacent to it, unique features may be employed to monitor adjacent repeat methylation (Figure 1). The first reported epigenetic alteration in cancer genomes was a loss of methylation from repeated elements (51), and such changes in methylation may have profound consequences for the genome as a whole (52). Our approach is also unique in its ability to report the relative density of 5 mC around each feature (Figure 4) as well as to identify changes in regional genomic copy number (Figure 5); although a relatively high probe density was required to observe this effect.

The purpose of this study was to demonstrate that methylated and unmethylated sequences could be detected within the human genome, and to understand the precision and accuracy of the cytosine methylation profiling procedure. Because a duplicated dye-swap experiment was performed within each DNA sample, and each feature was replicated on the array, a stringent measure of precision can be gauged through a correlation analysis between the independent technical replicates. The average correlation ($R^2$) was 0.77 for LN-18 and ranged between 0.72 and 0.83. Recent experiments involving nearly 50 arrays have displayed $R^2$ averages of 0.87 and a range of 0.81 to 0.93 (J. M. Ordway, R. M. Kendall and J. A. Jeddeloh, unpublished data). Using two independent validation approaches, we have demonstrated that the MethylScope array hybridization method is robust and accurate. In total, between the two verification approaches, 95 loci were tested, 70 of which were identified as either significantly methylated or unmethylated according to the microarray. The array measurements agreed with the verification measurements 59 times so that the accuracy of the array approach is at least 84% (59/70), assuming that bisulphite sequencing and MethylScreen are almost always accurate. Importantly, the verification of data obtained from measurements which were not statistically significant yielded appropriate results; i.e. these features were predominantly unmethylated as predicted by our analytical hypothesis (20/25). However, future analyses would not be able to identify the true status of these measurements without a second analytical method (i.e. not array only).

These values should be considered a conservative estimate of accuracy because of wingspan effects (Figure 1). The MethylScreen quantitative PCR approach is only sensitive to DNA methylation between the primers while the array itself is capable of reporting methylation of a larger region.

**LN-18 epigenetic landscape**

Our analysis of the LN-18 genome revealed more than 4000 methylated loci, by design the majority of which were associated with the transcription start sites of human genes (Table I). CpG islands (CGI) were disproportionately unmethylated in agreement with previous reports [for review a see (30)]; however far more were hypermethylated than anticipated if both sparse and dense methylation are considered (Table I). A conservative estimate of the CGI hypermethylation is revealed by the CGI features reporting dense methylation, which were 2% of the total CGI features. However, since approximately half of the features with intermediate array ratios were also methylated within the feature, upwards of 19% (2 + 17%) of the CGI in this genome are methylated (Table I). Perhaps this unexpected finding reflects the more extreme epigenomic representation displayed by cancer derived cell lines.

Importantly, the CGI associated with TSS were most often associated with unmethylated and intermediate array ratios. This observation highlights the importance of their unmethylated status, since half of these intermediate measurements will likely represent unmethylated features with adjacent methylated DNA. Further, it suggests that many more TSS may be susceptible to the spread of silencing
from these proximally methylated elements than originally anticipated (5,30). The use of ultra-high density microarrays with much higher feature densities will permit the more accurate mapping of methylation changes in the future.

Among the independently verified hypermethylated loci that the array discovered were E-cadherin and RASSF1A. These loci were important for two reasons. First, they served as an internal positive control since their ability to become hypermethylated had been shown previously (38,39). Second, the methylation pattern present at the RASSF1A promoter region and the high density feature placement suggested wingspan. As predicted, features some distance from the verified center of methylation, detected methylation proximal to them, though with lower confidence (data not shown). Wingspan was confirmed elsewhere by bisulfite sequencing (Figure 4).

Because a given feature seemed to be capable of detecting adjacent methylation, the quantitative value obtained from any probe represents a complex output. Signal from a feature provides not only data regarding the density of methylation surrounding the feature, but also information about its chromosomal context. Future studies are aimed at analytical methods to de-convolute the contribution of both signals.

Comprehensive nature of analysis supports epigenetic mapping

Recently, the need for large-scale epigenetic mapping of the human genome (The Human Epigenome Project) has been realized in the context of cancer, neurological and other diseases (53). The American Association for Cancer Research, National Human Genome Research Institute, as well as National Cancer Institute have recently called for a Human Epigenome Project to determine genome-wide patterns of DNA and histone modifications. Such a coordinated effort would run in parallel with the Cancer Genome Anatomy Project which will identify genetic changes in cancer cells. Epigenetic profiles are also important in testing stem cell lines for their suitability for therapeutic cloning (54,55).

Preliminary efforts to generate such epigenomic maps have included the large scale bisulphite sequencing of several human chromosomal regions (56,57); the analysis of modified histones associated with chromosomes 21 and 22 by chromatin immunoprecipitation (ChIP) chip (58), as well as array-based methylation profiling using plasmid and BAC subclones from the human genome (36,59).

The method we describe here, in addition to detecting biomarkers differentially methylated in disease, has potential to rapidly map quantitative methylation differences throughout the genome at very high resolution, and could make a substantial contribution to epigenomic maps of this sort. The analysis presented here demonstrates that any whole genome effort will need better than 4 kb resolution in order to fully capture the profile of the epigenetic landscape and resolve the wingspan effects. Higher probe density would mitigate any negative effects of this possibility and are readily achievable.

Targets of epigenetic modification identified

Several targets of DNA hypermethylation were identified in the course of this characterization. We have independently confirmed that the testis specific histone 2B (TH2B) locus is methylated in at least 7 non-testes tissues (brain, cervix, ovary, lung, colon, breast and lymphocytes), suggesting conservation of epigenetic regulation from rodents to humans (Figure 7 and data not shown) (37). This finding makes the locus useful for a positive control. Ten of 13 HOX cluster transcriptional starts sites were in methylated regions in the chromosome 7 HOX cluster (Figures 5 and 6). It appeared that only three TSS in the region were substantially unmethylated, Q96MZ23, HOX3 and HOX11 (Figure 6 and Supplementary Table). Members of this HOX cluster have previously been found to be concomitantly hypermethylated in cell lines and tumors (60). These findings suggest that domains larger than individual genes may also be targets of epigenetic alteration. HOX hypermethylation may be general to cell lines other than LN-18, since the SW48 (colon tumor cell line) profile reported by Weber et al. (36) has similar characteristics (Figure 6). However, the local highs and lows in the HOX region (Figure 6) are only apparent from our analysis. Interestingly, inappropriate CDKN2A and HOX gene regulation in brain cancer, and altered DNA methyltransferase functions have been mechanistically connected through interactions with the polycomb repressive complex member BMI1 (61).

Finding that 6 of the 8 differentially methylated loci identified were novel hypermethylation targets suggests that the full number of genomic loci susceptible to epigenetic modification is large, and that the majority have yet to be revealed. Four of these genes have been implicated in normal brain development (IRX3 (62), BMP7 (63), WNT6 (64) and WNT10A (65)). Finally, RARz is a member of a gene family that includes other genes subject to hypermethylation in cancer (66).

If this fraction of differentially methylated sequences (8 of 24) is representative of the whole dataset, we would predict that approximately 1500 loci in the human genome might behave similarly (~35% of the 4152 methylated loci). This value is in close agreement with an estimate put forth by Baylin and colleagues (43). However, neither estimate considers the likely existence of thousands of loci with the opposite molecular phenotype (unmethylated in tumors).

This study identified a hypermethylated region in exon 2 of the IRX3 gene and coincident upregulation of IRX3 transcription in glial-derived primary tumors and tumor cell lines, relative to normal brain. Given the reported functions of IRX (iroquois) family proteins, this finding may have implications for the development of gliomas. IRX family genes encode homeobox transcription factors conserved from nematodes to humans (67). In vertebrates, these factors participate in regulation of proneural genes during early neurulation (68) as well as in antero-posterior and dorsoventral subdivision of the neural plate [reviewed in (69)]. In these contexts, regulation of IRX3 activity has been shown to be directly responsive to signaling from both the Wnt and sonic hedgehog (Shh) pathways (70,71). Traditionally, high-grade gliomas (including glioblastoma and astrocytoma) have been presumed to arise from glial cells. However, recent studies support a model in which human gloma arises from neural stem-like cells (72–75). Given that Shh functions in maintaining neural stem cell characteristics (76), and ectopic expression of Irx3 in chick embryos is sufficient to induce Shh responsiveness (77), it is tempting to speculate that overexpression of IRX3 may be functionally involved in glial tumorigenesis. Furthermore, the interesting finding of coordinate differential methylation of HOXA10, WNT10A, WNT6 and IRX3 genes in primary glial tumors and tumor cell lines raises the possibility of an underlying epigenetic
signature that decouples these developmental factors from their normal regulatory mechanisms.

The power of genomic survey to detect biomarkers is best illustrated by the exonic CGI within the IRX3 gene. Current views of the importance of epigenetic regulation in cancer are based upon the belief that the regulatory regions of tumor suppressor genes become hypermethylated and/or that the regulatory regions of oncogenes may become unmethylated, and that these alterations lead to changes in expression. While certainly true, IRX3 serves as a reminder of the danger implicit in strictly interpreting rules before the landscape of epigenetic alteration has been fully characterized.

Clearly, comparing methylation profiles of normal brain tissues and primary tumors with this data set would be the optimal means to discover and characterize the number and types of novel hypermethylation and hypomethylation targets. Such studies using our array assay are presently underway.

**Supplementary data**

Supplementary data are available at *Carcinogenesis* online

**Acknowledgements**

The authors thank Z. Lippman, V. Colot, R. Wilson, W.R. McCombie, E.J. Richards, M. and J. Finney, S. Smith, R. Green and Nimblegen Technologies. R.A. Martienssen, J.D. McPherson, R.W. Doerge, and the authors from Orion Genomics recognize a financial interest in the company.

**References**


