High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing

Emily Hodges[1,2*], Andrew D. Smith[1,3*], Jude Kendall[1], Zhenyu Xuan[1], Kandasamy Ravi[1], Michelle Rooks[1,2], Michael Q. Zhang[1], Kenny Ye[4], Arindam Bhattacharjee[5], Leonardo Brizuela[5], W. Richard McCombie[1], Michael Wigler[1], Gregory J. Hannon[1,2†], and James B. Hicks[1†]

[1]Watson School of Biological Sciences
[2]Howard Hughes Medical Institute
Cold Spring Harbor Laboratory
1 Bungtown Road, Cold Spring Harbor, New York 11724, USA

[3]Molecular and Computational Biology
University of Southern California
Los Angeles, California 90089, USA

[4]Department of Epidemiology and Population Health
Albert Einstein College of Medicine
Bronx, New York 10461

[5]Agilent Technologies, Inc.
5301 Stevens Creek Boulevard
Santa Clara, California 95051

*These authors contributed equally to this work.

† To whom correspondence should be addressed:

James B. Hicks
Cold Spring Harbor Laboratory
1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
Phone: +1 (516) 367-8382
Fax: +1 (516) 367-8381
Email: hicks@cshl.edu

Gregory J. Hannon
Phone: +1 (516) 367-8889
Fax: +1 (516) 367-8874
Email: hannon@cshl.edu

Abstract
Regulated changes in DNA methylation occur during normal development and contribute to the stability of epigenetic states. Aberrant methylation is associated with disease progression and is a common feature of cancer genomes. Presently, few methods enable quantitative, large-scale, single-base resolution mapping of DNA methylation states in desired regions of a complex mammalian genome. Here, we present an approach that combines array-based hybrid selection and massively parallel bisulfite sequencing to profile DNA methylation in genomic regions spanning hundreds of thousands of bases. This single molecule strategy enables methylation variable positions to be quantitatively examined with high sampling precision. Using bisulfite capture, we assessed methylation patterns across 324 randomly selected CpG islands (CGI) representing more than 25,000 CpG sites. Using a single lane of Illumina sequencing, methylation states could be definitively called for >90% of target sties. The accuracy of the hybrid-selection approach was verified by spot checking using conventional capillary sequencing of PCR products from bisulfite treated DNA from the same specimens. This confirmed that even partially methylated states could be called successfully. A comparison of human primary and cancer cells revealed multiple differentially methylated regions. More than 25% of islands showed complex methylation patterns either with partial methylation states defining the entire CGI or with contrasting methylation states appearing in specific regional blocks within the island.

Introduction

It has long been known that changes in cellular and organismal characteristics can be inherited without accompanying alterations in genomic sequence (Waddington 1942). This phenomenon, known as epigenetic inheritance, has been proposed to occur through a number of mechanisms, including histone modification and DNA methylation (Holliday and Pugh 1975).

In mammals, DNA methylation is observed mainly at CpG dinucleotides. This modification is propagated via a maintenance methyltransferase, Dnmt1 (Bestor, Laudano et al. 1988), which preferentially recognizes and modifies hemi-methylated CpGs (Bestor 1992). While the vast majority of CpGs are methylated in differentiated mammalian cells (Bird and Taggart 1980), most methylation undergoes waves of erasure and reestablishment during gametogenesis and preimplantation development (Chaillet, Vogt et al. 1991; Monk, Boubelik et al. 1987; Sanford, Clark et al. 1987). The re-establishment of methylation is carried out by *de novo* methyltransferases, Dnmt3a and Dnmt3b (Okano, Xie et al. 1998).

Although CpG dinucleotides are significantly underrepresented in mammalian genomes, certain regions are relatively rich in CpGs, called CpG islands (CGIs; (Bird 1986)). While CGIs are found throughout the genome, they are often associated with promoter regions, with over 70% of annotated genes having CGI-related promoters (Saxonov, Berg et al. 2006). Hypermethylation of promoters is correlated with heterochromatin formation and silenced transcription (Keshet, Lieman-Hurwitz et al. 1986).

Studies of *dnmt1*- and *dnmt3*-mutant mice indicate an essential role for methylation in normal development (Li, Bestor et al. 1992; Okano, Bell et al. 1999). Current models suggest that the regulated and mitotically inherited methylation of specific genomic regions, through the developmental history of a cell, functions to restrict potency and guide cell fate (Reik 2007; Shen, Kondo et al. 2007). Aberrant DNA methylation is associated with disease development and progression.

Despite its importance, mechanisms that guide DNA methylation and the biological impact of global modification patterns remain poorly understood, due in part to the limitations of current methylation profiling technologies. Current profiling methods can be classified roughly into two categories, those that measure methylation at high nucleotide resolution for a modest number of genomic intervals and those capable of surveying the whole genome at low to moderate resolution.

Existing genome-wide approaches typically involve comparative microarray hybridization following fractionation of the genome based upon methyl-cytosine specific antibodies/protein complexes (MeDIP, MIRA) or methylation responsive

enzymes (e.g., MspI/HpaI or McrBc) with sites in CpG-rich regions (Khulan, Thompson et al. 2006; Lippman, Gendrel et al. 2004; Rauch, Wu et al. 2009; Shen, Kondo, Guo, Zhang, Zhang, Ahmed, Shu, Chen, Waterland and Issa 2007; Weber, Davies et al. 2005). The sensitivity of the enzymatic approach is limited by the sequence context of the digestion site and by the number of sites available. Moreover, microarray-based approaches produce an average snapshot of methylation across genomic windows. As a result, resolution of methylation states at individual sites is generally imprecise and can be strongly influenced by CpG density and fragment size (Irizarry, Ladd-Acosta et al. 2008). This drawback hampers the use of such methods for the analysis of imprinted loci and regions with complex methylation patterns.

High-resolution strategies can distinguish methylation states in a semi-quantitative, allele-specific manner at individual CpGs within a defined region. Established protocols that positively identify 5-methylcytosine residues in single strands of genomic DNA exploit the sodium bisulfite-induced deamination of cytosine to uracil. Under denaturing conditions, only methylated cytosines are protected from conversion. To measure methylation levels, bisulfite conversion has been combined with restriction analysis (COBRA) (Xiong and Laird 1997), base-specific cleavage and mass spectrometry (Ehrich, Nelson et al. 2005; Ehrich, Turner et al. 2008), real-time PCR (MethyLight) (Eads, Danenberg et al. 2000) and pyrosequencing (Dupont, Tost et al. 2004). However, these methods are generally limited by their scalability and cost.

Bisulfite sequencing represents the most comprehensive, high-resolution method for determining DNA methylation states. Like SNP detection, the accurate quantification of variable methylation frequencies requires high sampling of individual molecules. High-throughput, single-molecule sequencing instruments have facilitated the genome-wide application of this approach. For example, direct shotgun bisulfite sequencing provided adequate coverage depth and proved cost-effective for a small genome like *Arabidopsis* (119 Mbp) (Cokus, Feng et al. 2008). However, these approaches are currently impractical for routine application in complex mammalian genomes, and simplification of DNA fragment populations (genome partitioning) is still required to boost sampling depth of individual CpG sites (Meissner, Mikkelsen et al. 2008; Taylor, Kramer et al. 2007). This problem becomes compounded as one considers that, within a multicellular organism, there are probably at least as many epigenomic states as there are cell types. Therefore, to understand the impact of epigenetic variation will require both detailed reference maps and the ability to interrogate regions of those reference maps in many samples and cell types at high resolution. Recent strategies for addressing methylation in large genomes have included enzyme directed reduced genomic representation (Brunner, Johnson et al. 2009; Meissner, Mikkelsen, Gu, Wernig, Hanna, Sivachenko, Zhang, Bernstein, Nusbaum, Jaffe et al. 2008) and padlock probe assisted multiplex amplification (Ball, Li et al. 2009; Deng, Shoemaker et al. 2009) followed by massively parallel sequencing.

To this end, we have developed bisulfite capture, a technology platform that combines bisulfite conversion with hybrid selection techniques and deep sequencing. Our profiling method is capable of achieving single nucleotide resolution while simultaneously examining methylation frequencies in tens of thousands of CpGs. Bisulfite capture directs focus to specified CpG regions in a highly parallelized process designed to selectively enhance sequence information content by deeper sampling of targeted bases. Unlike other reduced representation schemes, the selection process is independent of methylation status and the substrate may be tailored to include any unambiguous genomic interval. Here, we describe the application of this approach to determine DNA methylation frequencies in CGIs sampled from a variety of genomic settings including promoters, exons, introns, and intergenic loci. To discern the sensitivity of our approach to detect differential methylation patterns, bisulfite capture was carried out on two model cell lines, a primary skin cell line and a breast cancer cell line. For our study, 324 randomly selected CpG islands encompassing nearly 300kb of genomic space and 25,000 CpG sites were examined in parallel. While global comparison of the two cell lines recapitulates previously described trends, detailed analysis reveals many examples of unexpected complexity in methylation states and instances where sharp transitions from methylated to unmethylated intervals could be finely mapped. Our results demonstrate the unique capacity of the bisulfite capture system to detect site-specific switches in methylation on a readily scalable, cost effective platform.

Results

Experimental Design

Recently, others and we have described the use of custom microarrays as substrates for hybrid selection of high interest regions from complex genomes (Albert, Molla et al. 2007; Hodges, Xuan et al. 2007; Okou, Steinberg et al. 2007). This massively parallel focused resequencing method permits identification of sequence variants within selected genomic intervals spanning thousands to millions of bases. Here, we sought to adapt the same approach for the determination of DNA methylation states. To accomplish this, we integrated bisulfite conversion of genomic DNA into our capture scheme (Fig. 1).

There are, in principle, several ways in which bisulfite conversion could be coupled with hybrid selection. One logical option would be to capture relevant regions of native genomic DNA followed by sodium bisulfite treatment and sequencing. However, this strategy has a major shortcoming in that the hybrid selection step requires large amounts of native, unamplified DNA to be readily available as input (Albert, Molla, Muzny, Nazareth, Wheeler, Song, Richmond, Middle, Rodesch, Packard et al. 2007; Hodges, Xuan, Balija, Kramer, Molla, Smith, Middle, Rodesch, Albert, Hannon et al. 2007; Okou, Steinberg, Middle, Cutler, Albert and Zwick 2007). Substantial amounts of DNA can also be lost

during the harsh process of bisulfite conversion. Because very small amounts of material are generally eluted from the capture arrays, bisulfite conversion post-capture could restrict the number of individually sampled molecules for each variable methylation site. Moreover, for many applications, we desired a method suitable for the analysis of relatively small cell numbers, such as tissue stem cells, or microdissected or laser-captured tumor cells. For these reasons, we developed a platform that permits the use of minimal amounts of starting material, subjecting these samples to bisulfite conversion and amplification prior to hybridization.

We tested our approach using DNA from normal, dermal fibroblast cells (CHP-SKN-1) commonly used as a reference in our microarray studies (Hicks, Krasnitz et al. 2006; Sebat, Lakshmi et al. 2004) and the invasive breast tumor cell line, MDA-MB-231 (ATCC# HTB-26). To prepare samples for sequencing on the Illumina GA2, genomic DNA libraries were generated as previously described with a few important modifications (Fig. 1). First, DNA fragments were ligated to Illumina-compatible adaptors synthesized with 5'-methyl-cytosine instead of cytosine to prevent their conversion by bisulfite treatment. A similar strategy was applied previously for shotgun bisulfite sequencing of the Arabidopsis genome (Cokus, Feng, Zhang, Chen, Merriman, Haudenschild, Pradhan, Nelson, Pellegrini and Jacobsen 2008). Second, following size selection and gel purification, the fragments were denatured and bisulfite converted, so that the status of each CpG site became fixed in the sample. Lastly, the adaptor-ligated fragments were PCR enriched with a polymerase capable of amplifying uracil-rich templates. The amplification process produces ample amounts of input material for hybridization. Ultimately, the library preparation procedure generates four strands (Fig. 1). Two are derived from the original plus and minus strands of the genome. Since these were treated with bisulfite, they are depleted of C, and are designated the T-rich strands. The complements of the converted genomic strands are designated the A-rich strands.

Array Design

There are ~28,000 annotated CGIs in the human genome. CGIs are defined herein as intervals of >200bp with >50% GC content and significant CpG density (Gardiner-Garden and Frommer 1987). As CGIs are relevant targets for DNA methylation, we randomly selected 324 islands between 300-2000 bp representing 258,895 bases of genomic space and 25,000 CpG sites (~0.1% of all CpG sites in the genome). The set was distributed among all autosomes and chromosome X, including 170 islands located within 1500 bp upstream of annotated protein coding genes and 154 islands outside of annotated promoter regions, both intra- and intergenic.

Bisulfite conversion creates a layer of variability between the reference genome and converted template. Therefore, our strategy required an array design that

anticipated the range of possible changes to DNA sequence resulting from cytosine depletion. Standard 60 nucleotide array capture probes are typically designed for one strand of the genomic template (Hodges, Xuan, Balija, Kramer, Molla, Smith, Middle, Rodesch, Albert, Hannon et al. 2007). However, bisulfite conversion and amplification results in four strands comprising two unique double-stranded templates. In principle, it is possible to capture any of the four converted single strands. For symmetric CpG methylation, capture of one of the four products should allow inference of a complete methylation map. However, there have been reports of asymmetric (non-CpG) methylation in some mammalian cell types (Haines, Rodenhiser et al. 2001). Although not the focus of this study, detecting such modifications would require interrogation of products of both genomic strands. Additionally, capturing more than one strand would increase coverage and thus confidence in determining methylation states, but the trade-off would be a reduction in the total genomic area that could be tiled on an array of a given capacity. As a compromise, we chose to capture two strands, the A-rich derivatives of both plus and minus genomic strands (Fig. 1); however, depending upon the biological question, capture of one strand would certainly be sufficient.

For each CpG island, two sets of capture probes were designed, one that assumed full methylation of all CpG residues, and one that assumed full conversion of CpGs to TpGs. Thus, even with a completely random pattern of CpG methylation, only half of the CpG sites within a given probe would contribute a mismatch. Previous studies have quantified the effect of mismatches on hybridization to 60 nucleotide probes printed on Agilent custom arrays (Hughes, Mao et al. 2001), the same selection substrate that we now use in our capture studies (Hodges et al., in press). These reports suggest that up to 6 distributed mismatches are tolerated without a substantial impact on hybridization efficiency. Our previous studies also indicated that the presence of SNPs did not impact the efficiency of capture (Hodges, Xuan, Balija, Kramer, Molla, Smith, Middle, Rodesch, Albert, Hannon et al. 2007). Therefore, we were confident that efficient hybridization could be achieved despite uncertainty in the exact sequence of the A-rich target strands. The mean number of CpGs within probe sequences to the 324 selected CpG islands is 4.68, and the maximum in any probe is 15. Thus, the vast majority of probes are well within the predicted margin of safety for efficient capture (Fig. S1). The designed 60 nucleotide selection probes were tiled every six bases across our contiguous target intervals and synthesized on Agilent 244k arrays.

Mapping bisulfite treated reads

Mapping short sequenced reads requires identifying the genomic locations at which the reference sequence most closely matches that of the read. A small number of mismatches are typically allowed, and when the best match for a given read occurs at two distinct locations, that read is said to map ambiguously.

We infer methylation states only from reads with unambiguous mappings. Bisulfite sequence conversion presents a significant challenge to mapping short reads because the inherent information content of converted DNA is reduced. Since we capture the A-rich strand, and sequence its complement, a T observed in a read may map to a T or a C in the reference genome.

We developed an algorithm for rapidly mapping bisulfite treated reads while accounting for both the C to T conversion at unmethylated cytosines and for the retention of C when a residue is either protected from conversion or unconverted by chance. Our algorithm is based on RMAP (Smith, Xuan et al. 2008) and follows the conventional strategy used in approximate matching. First, we used an "exclusion" stage, requiring candidate mapping locations to have an exact match to the read in a specific subset of positions ("seed" positions). Because the exclusion stage used exact matching, it assumed all Cs in both read and genome sequences have been converted to T. This assumption resulted in a substantial loss of efficiency to the exclusion, and we compensated for this loss by designing tiled seeds. This had the effect of the multiple filtration strategy of Pevzner & Waterman (Pevzner and Waterman 1995) but permitted a highly efficient implementation. In contrast with mapping methods that preprocess the genome, this strategy required relatively little memory and was therefore appropriate for use on nodes of scientific clusters commonly used for analysis of sequencing data.

The algorithm was also designed to take advantage of sequencing quality scores by assigning fractional mismatch penalties based upon the certainty of a base call and by taking into account the fact that a large fraction of Cs are converted to Ts (Figure 2B). For example, in the comparison of site A versus site B in Figure 2, a clear high quality call of G, C or A resulted in a strong penalty for any mismatch. A less high quality call of G, C or A provided an intermediate penalty whose quantitative weight was a function of the individual probabilities of each alternative call (e.g. Figure 2B, site B, position 2). Since we were sequencing bisulfite converted DNA, potential T calls had a nearly equal probability of originating from a genomic T or C. Thus, for cases in which there was a higher probability of a T call than a C call, the lower mismatch penalty for T was also assigned to C (e.g., Figure 2B, site B, position 4). A detailed description of the algorithm, along with a discussion of how to exploit unconverted cytosines without introducing bias, is given in Supplementary Information.

Following bisulfite capture, deep sequencing of the CGI-enriched material generated 20,002,407 raw 36 base reads for MDA-MB-231 and 55,770,254 for CHP-SKN-1 cells (Table 1). Using our mapping algorithm, unique genomic locations were assigned to 7,575,990 and 12,130,697 reads for tumor and normal cells, respectively. We used stringent criteria in mapping, permitting the equivalent, in terms of quality scores, of at most one mismatch per 36-base read. A standard sequencing run on unconverted DNA generally yields 50-60% uniquely mappable reads. In this case, the unsuccessful assignment of more

than half of the reads can be attributed to a combination of highly stringent mapping criteria, reduced sequence complexity following bisulfite conversion, and poor read quality in some Illumina runs. The effect of read quality was also reflected in a comparison of the two samples. The number of sequenced reads and the proportion of those that mapped differed substantially between the samples, which were sequenced on different flow cells. We determined that the total number of mapped reads was sufficient for our experiments, allowing us to investigate the performance of our protocol even when limited data is available. Overall, 6.43 to 11.98% of the reads mapped unambiguously within the targeted CpG islands, corresponding to a substantial enrichment of 711- to 1324-fold for the selected regions from total genomic DNA (Table 1).

Methylation Status of Individual CpGs

An important indication of success for bisulfite capture was that sufficient coverage of the targeted bases was achieved with minimal amounts of sequencing. Using a single Illumina flow cell lane to sequence captured material, 86-91% of the targeted nucleotides were covered by at least 10 reads for each cell line. This is sufficient depth for a confident measure of methylation frequency (see below). It should be noted, however, that both coverage and enrichment rates likely underestimate the performance of the approach, since certain reads from within the target areas cannot be uniquely mapped. For an estimate of the extent of such “dead zones” and their relationship to read length, see Supplementary Table 1.

Variations in coverage depth, the relatively high rate of sequencing error and the fact that individual cytosine residues can be both methylated and unmethylated within a given population of cells necessitated rigorous statistical methods for calling methylation status. We started with two values: the fraction of unconverted cytosines mapping over a CpG and the total number of reads mapping over the CpG. For these studies, we focused on symmetric CpG methylation and therefore collapsed information obtained from both genomic strands. All reads having anything other than a C or T at a given CpG were excluded from analysis. Thus, the “methylated proportion” was defined as the number of reads with a C at a given CpG divided by the number of informative reads. We calculated confidence intervals for the methylated proportion according to Wilson (Wilson 1927) and used these in conjunction with the methylated proportion to call methylation status. Our method assigned methylation states of unmethylated, methylated, partially methylated, or “no call” (to indicate insufficient information). See Methods and Figure S2 for details.

This strategy resulted in confident calls for the vast majority of CpGs in the islands we examined. Increasing sequencing depth would generally increase the number of confidently called CpGs. Of the 25,044 CpG dinucleotides investigated in this analysis, 91.6% in MDA-MB-231 and 92.1% in CHP-SKN-1 could be given

a confident call, either methylated, unmethylated or partially methylated, using the stringent criteria outlined above (Table 2). In both samples, a majority of CpG sites was called either methylated or unmethylated, with only 7% and 12% classified as partially methylated in the normal and tumor cells, respectively. A comparison of methylation frequencies between the two samples showed that the state of many CpG sites closely corresponded in both cell types (Fig. 3A). Of the discordant calls, a higher number were either fully or partially methylated in the tumor sample (Table 3, Fig. 3A). It is notable that among the 22,684 CpGs receiving a confident call in both samples, only 0.2% were called methylated in the normal cell line and unmethylated in the tumor cell line, while 10.3% were unmethylated in CHP-SKN-1 and methylated in MDA-MB-231.

Significant correlation between the methylation states of adjacent CpG sites and a high incidence of short-range comethylation has been reported previously (Eckhardt, Lewin et al. 2006; Irizarry, Ladd-Acosta, Carvalho, Wu, Brandenburg, Jeddeloh, Wen and Feinberg 2008). Therefore, we examined the methylation state of one CpG site as a function of methylation at the subsequent CpG site within our selected CGIs (Fig. 3B, 3C). There was clearly autocorrelation of methylation frequencies through a CGI (0.949 for MDA-MB-231; 0.944 for CHP-SKN-1). Specifically, if a CpG is highly methylated, then the neighboring CpG is more likely to be methylated, and *vice versa* (Fig. 3B, 3C). Furthermore, the concentration of points along the diagonal indicates that partially methylated CpGs are also highly autocorrelated within islands, and will therefore likely reside in a neighborhood of partial methylation.

To validate the accuracy of results obtained with hybrid selection and single molecule sequencing, we selected four CGIs to profile independently with traditional bisulfite cloning and sequencing. The CGIs were specifically selected to validate estimates of intermediate methylation frequency. For each of the four CGIs, multiple overlapping PCR products were generated from the bisulfite converted tumor cell line DNA. Purified amplicons were cloned, and individual colonies were sequenced by traditional capillary sequencing, generating 202 high quality reads. The methylation status of each CpG within each sequenced clone is depicted in Figure 4, along with histograms of CpG methylation frequencies for both traditional bisulfite cloning and bisulfite capture. Excluding the region in Figure 4A, for which too few traditional bisulfite reads were obtained, the methylation frequencies estimated from both methods correspond very closely. We obtained 90% confidence intervals on the methylation proportion at 62 CpGs using traditional bisulfite reads. The confidence intervals overlapped those based on the bisulfite capture at 81% of the CpGs (see Supplementary Table 2). Of the 12 for which the intervals did not overlap, the methylation level estimated using bisulfite capture was closer to 50% on all but two CpG sites. This demonstrates that the hypo/hyper-methylated probe-pair strategy used in bisulfite capture does not bias the capture towards extreme states. In addition, these results also indicate that the higher sampling rates achieved with capture and single molecule sequencing contribute to higher accuracy in calling methylation status.

Patterns of CpG Island Methylation

Changes in DNA methylation patterns have been associated with a number of human diseases, and aberrant DNA methylation contributes causally to tumorigenesis. For example, a significantly elevated proportion of somatic mutations in the tumor suppressor p53 have been found at CpG sites (Rideout, Coetzee et al. 1990). Moreover, tumor genomes are generally hypomethylated, which may contribute to genome instability, perhaps in part by releasing constraints on mobile genetic elements (Lengauer, Kinzler et al. 1997). The global reduction in methylation is accompanied by hypermethylation of individual CGIs, some of which are associated with tumor suppressor genes (Herman and Baylin 2003).

We, therefore, compared patterns of CpG methylation in our normal fibroblast and breast tumor cell lines (Table 4). Consistent with previously observed trends, the distribution of CpG methylation frequencies was largely bimodal (Fig. S3-S4, Fig. 5) with more CGI CpGs in the tumor cell line exhibiting high methylation frequency as compared to the normal fibroblast sample. The aggregate results on individual CpGs in our sampled islands suggest a picture of mostly unmethylated CGIs in the normal cell line and elevated methylation in the tumor cell line. While a little over half of the islands fall into expected categories of fully methylated or fully unmethylated, a surprising number of CGIs displayed more complex methylation profiles. A closer inspection of the individual islands, examples of which are shown in Figure 5 and Figure S5, revealed a rich substructure in many islands that might not be apparent without their examination at the sequence level.

Around 54% of the CGIs showed clearly defined and consistent methylation states across the entire island in both samples. The most common were ‘unmethylated’ islands, with few CGIs assigned as methylated in either the MDA-MB-231 or the CHP-SKN-1 sample (143 cases) (Fig. 5A). A smaller subset (31 cases) showed nearly complete methylation in both samples (Table 5, Fig. 5B). We observed 15 cases that were virtually unmethylated in CHP-SKN-1 but completely methylated in the tumor line, as exemplified by the island at the transcription start site (TSS) of the cell adhesion associated gene FLRT2 (Fig. 5C). We did not observe the converse, where a completely methylated island in CHP-SKN-1 was completely unmethylated in MDA-MB-231; however, there were multiple cases in which methylation was clearly reduced in the tumor line, either in sub-regional blocks or across an entire island (Fig. 5). For about 13% of the islands in this study (41/324), states could not be assigned because all or a large portion of the island overlapped repetitive elements in the genome, and a sufficient proportion of reads could not be uniquely mapped.

Approximately 28% (92 cases) of the investigated CGIs displayed methylation signatures other than simple all or none methylation across the island. These heterogeneous profiles appear in two distinct topologies. First, mixed methylation assignments for individual CpG sites give rise to partial or intermediate methylation, either across the whole island or in sub-regional blocks (Fig. 5D, E, F, and H). This pattern is exemplified by an island located near the 3' end of KCNQ2, a putatively imprinted gene (Luedi, Dietrich et al. 2007), where both samples are partially methylated across the entire island (Fig. 5D). Note that the solid gray bars at the ends of the island represent unassigned CpGs due to their location in the repetitive sequence that often flanks CpG islands. A CGI overlapping the HOXB4 promoter and lying in the HOXB3 transcript appears to be partially methylated in SKN-1, consistent with its prediction to be imprinted (Luedi, Dietrich, Weidman, Bosko, Jirtle and Hartemink 2007), but is fully methylated in MDA-MB-231, consistent with reports that HOX clusters are often methylated in breast cancer (Fig. 4 E, (Rauch, Wang et al. 2007)).

A second topology is defined by sharp transitions from one methylation state to another within an island (Fig 5F, G, H, and I). Figures 5G-I illustrate such structural complexity. Many islands showed some degree of methylation in localized, contiguous blocks. These switches are striking and often define domains of the CGI with respect to methylation. In a number of cases these regions, or blocks of methylation, occupy the same position in both samples. These observations indicate the presence of 'punctuation marks' within CpG islands that likely reflect underlying biological mechanisms.

We noted that islands displaying these transitions often overlap transcription start sites (TSS) and exon junctions. As examples, for islands spanning significant portions of the SSTR4 and RASGRF2 genes, the transitions in methylation occur around or close to the TSS. Likewise, the CGI fully overlapping the GLTPD2 gene locus contains a short domain of mostly methylated CpGs in CHP-SKN-1 that covers the 5'UTR of the gene. Interestingly, the transition to hypomethylation closely corresponds with the first coding exon of the gene. One might speculate that the positions of breakpoints between domains of high methylation and neighboring domains of low methylation may be influenced by gene regulatory mechanisms and local genome structure. The overall biological significance and correlation of such patterns with expression state have yet to be determined.

To complement the comprehensive overview of methylation states in the two cell lines, we also categorized CpG methylation by genomic annotation, examining promoter-associated, genic, and intergenic sites (Figure S6, Supplementary Table 5). As expected, the fibroblast cells displayed a higher number of fully unmethylated CpGs/CGIs in each of the specified regions and the tumor cell line had consistently higher methylation. Notably, for promoter regions the highest proportion of differentially methylated CGIs was heterogeneously methylated in MDA-MB-231. A significant fraction of intragenic CGIs was methylated to some

degree in both cell lines, and nearly half of the intragenic CGIs in MDA-MB-231 were fully methylated.

We examined the relationship between dinucleotide frequencies and overall methylation in CGIs. Consistent with earlier reports, a strong negative correlation (-0.39 in CHP-SKN-1 and -0.32 in MDA-MB-231) between CpG density and total CGI methylation was observed (Zhang, Rohde et al. 2009). However, we also observed a strong positive correlation (0.69 in CHP-SKN-1 and 0.54 in MDA-MB-231) between CA/TG frequency and total methylation of the CGIs. Furthermore, sharp cutoffs for frequencies of these dinucleotides can accurately distinguish hypomethylated islands from those showing partial or full methylation, with both strong sensitivity and specificity (see Supplementary Tables 3-4 and Supp. Methods). This suggests existing definitions may not accurately capture the relationship between CpG density and protection from CpG depletion over evolutionary time scales. It is likely that more sophisticated definitions, which may account for characteristics beyond base composition, will be required to define the underlying evolutionary phenomena that produce CGIs.

Discussion

Existing methods for profiling DNA methylation are largely CGI centric and fail to examine methylation in regions beyond those defined as canonical islands (or islands significantly enriched in CpGs). However, the bisulfite capture method is readily programmable, and with the sensitivity and scale achieved here, this approach could be extended to any non-repeat, CpG-containing region in the genome, regardless of CpG density. Bisulfite sequencing of cloned DNA fragments is a well-established gold standard for mapping methylation at high resolution, as exemplified by a recent study of DNA methylation across gene promoter regions on human chromosome 21 (Zhang, Rohde, Tierling, Jurkowski, Bock, Santacruz, Ragozin, Reinhardt, Groth, Walter et al. 2009). This study highlights many of the same features of DNA methylation discussed here. In fact, our method is designed to provide a similar high level of resolution for hundreds of genomic regions without the need for creating individual PCR amplicons and sequencing individual clones. It is currently very costly to perform clone sequencing on the scale necessary to sample thousands of sites in multiple individual samples. Bisulfite capture provided both qualitative and quantitative methylation measurements that were nearly identical to bisulfite sequencing while permitting the highly parallel analysis necessary to understand the biological impact of changes across the epigenome in many cell types and/or individual specimens.

Our approach requires no *a priori* knowledge of the methylation state of target loci. By designing probes corresponding to extreme states, with all CpGs in the target region either fully methylated or unmethylated, we created a probe set that would sufficiently hybridize the selected regions, even if CpG dinucleotides in target fragments were methylated randomly. Since most studies find local

correlation between the methylation states of neighboring CpGs, the overall extent of the mismatch problem is likely to be much lower than the theoretical maximum we anticipated. Nevertheless, recovery of fragments containing both methylated and unmethylated residues provided clear evidence for the unbiased capture of molecules with mixed methylation states. Independent validation using conventional bisulfite sequencing of regions with partial methylation frequencies verified that our approach did not significantly bias the determination of methylation patterns toward local uniformity in CpG status.

Despite its initial success, our current protocol does have room for improvement in enrichment, completeness and uniformity of coverage. While longer reads and increased sequencing depth will improve CpG calling to some extent, the largest gains will likely be made in probe design and array structure. Presently, we capture two genomic strands. However, it is clear that the number of target CpG can be doubled simply by assaying only one strand. Moreover, array densities continue to increase. Recently, the number of probes on the array platform we most commonly use has quadrupled. Finally, we have covered the target CGIs at relatively high tiling density, and many improvements in probe design/selection are possible. Without significant changes to our protocols, it is likely that a 10-fold increase in covered sites can easily be achieved. Besides allowing larger target regions to be examined at greater coverage, more efficient capture arrays, when combined with sample indexing for multiplex captures, will enable targeted profiling of DNA methylation in large numbers of samples, opening the door to potential clinical applications (Laird 2003).

We previously found that genomic repeats could confound efficient capture. To combat this, we eliminate multicopy capture probes based upon average representation of their constituent 15mers in the genome. Because of the reduction in complexity following bisulfite treatment, the same rules could not be directly applied and repeats were not suppressed in these initial studies. Moreover, inclusion of C_0t-1 DNA in hybridizations improves enrichment in conventional captures. Though we did use C_0t-1 in these studies, it was unconverted and thus might not compete effectively with the repeat sequences present in our samples.

Here, we examined clonal cell lines, whose methylation patterns are relatively homogeneous. Tissue-derived samples likely contain multiple methylation states at a given locus, in part because of imprinting and X-inactivation, but mainly because of cell-type heterogeneity in even the most purified populations. Thus, variations in methylation patterns could represent a mixture of several distinct “epitypes”, each of which is a signature of the cell type from which it was derived. The depth of coverage achieved in bisulfite capture, combined with increases in read length, may permit assembly of such epitypes - a procedure analogous to metagenomic assembly. Ultimately, approaches that deeply sample the epigenome at single-nucleotide resolution and at the single molecule level may allow us to detect the presence of rare stem cell populations and to track the

epigenetic reprogramming that correlates with the commitment and fate specification of such multipotent cells to differentiated cell fates.

Materials and Methods

DNA Library Preparation and Bisulfite Conversion

Genomic DNA libraries were generated as previously described with a few important modifications. Briefly, purified cell-line DNA was randomly fragmented by sonication and subsequently treated with a mixture of T4 DNA Polymerase, E. coli DNA polymerase I Klenow fragment, and T4 polynucleotide kinase to repair, blunt and phosphorylate ends according to the manufacturer's instructions (Illumina). The repaired DNA fragments were subsequently 3' adenylated using Klenow exo- fragment (Illumina). After each step, the DNA was recovered using the QIAquick PCR Purification kit (Qiagen). Adenylated fragments were ligated to Illumina-compatible paired-end adaptors synthesized with 5'-methyl-cytosine instead of cytosine (Illumina) and fragments ranging from 150-300 bp were extracted by gel purification using the QIAquick gel extraction kit (Qiagen) followed by elution in 30ul elution buffer. Following size selection and gel purification, the adapter-ligated DNA was divided into two separate reactions to ensure optimal DNA concentration for subsequent cytosine conversion reactions. Fragments were denatured and treated with sodium bisulfite using the EZ DNA methylation gold kit according to the manufacturer's instructions (Zymo). Lastly, the sample was desulfonated and the converted, adaptor-ligated fragments were PCR enriched using paired-end adaptor-compatible primers 1.0 and 2.0 (Illumina) and Expand high fidelity plus PCR system (Roche), a specialized polymerase capable of amplifying the highly denatured, uracil-rich templates. Following amplification, the samples were hybridized to both arrays and captured fragments were recovered and sequenced.

CpG Island Array Capture

20mg of bisulfite-treated DNA was hybridized to custom Agilent 244K microarrays according to the Agilent aCGH protocol with several modifications. Firstly, in addition to 20mg sample DNA, 50mg human c_0t-1 DNA (Invitrogen) and Agilent blocking agent, Agilent aCGH/ChIP Hi-RPM hybridization buffer was supplemented with approximately 1 nmol each of four blocking oligonucleotides (IDT; see Supplementary Table 6) before denaturing at 95ºC. The samples were hybridized at 65ºC for 65h in a rotating microarray oven (SciGene). After hybridization, the arrays were washed at room temperature for 10 min with aCGH wash buffer 1 (Agilent) and washed with aCGH wash buffer 2 (Agilent) at 37ºC for 5 min. Slides were briefly dried at low speed in a slide rack using a centrifuge with a microplate adaptor. Captured bilsulfite-treated DNA fragments hybridized to the arrays were immediately eluted with 490ul nuclease-free water at 95ºC for 5 min in the rotating microarray oven. The fragments were removed from the chamber assembly using a $18^{1/2}$G syringe (BD). Samples were subsequently lyophilized and resuspended for amplification. Five 18-cycle PCR amplifications were performed in parallel for each eluate using Phusion HF PCR master mix

(Finnzymes). Following amplification, the PCR reactions were pooled and purified on Qiagen purification columns.

Single Molecule Sequencing

The DNA was quantified using the Nanodrop 7500 and diluted to a working concentration of 10 nM. Cluster generation was performed for samples representing each array capture in individual lanes of the Illumina GA2 flow cell. An adapter-compatible sequencing primer (Illumina) was hybridized to the prepared flow cell and 36 cycles of base incorporation were carried out on the Illumina GA2 genome analyzer.

Conventional Bisulfite Cloning and Sanger Sequencing

Specific regions of bisulfite treated CHP-SKN-1 and MDA-MB-231 DNA were PCR amplified and their products cloned and sequenced using conventional Sanger sequencing. Briefly, CHP-SKN-1 and MDA-MB-231 genomic DNA was bisulfite converted using the QIAGEN Epitect bisulfite kit according to manufacturer's instructions. The forward and reverse primers were designed for the forward strand using the online primer design tool Methprimer (Li and Dahiya 2002) followed by manual selection of primer sets to satisfy Tm and other requirements. Primer sequences are provided in Supplementary materials (Supplementary Table 7). Thermal cycling was performed as follows: 40 cycles each of denaturation at 92ºC for 50sec, annealing at 52ºC for 1 minute and extension at 72ºC for 1 minute followed by 10 minutes at 72ºC. The PCR products were analyzed on a 2% agarose gel and the reaction mixtures were purified using a PCR purification kit (Qiagen). Purified PCR products were subcloned into the pCR®2.1-TOPO® vector using the TOPO TA cloning kit (Invitrogen) according to the manufacturer's recommendations. Clones were transformed into Top10 competent cells and subsequent colonies were isolated, cultured overnight, and bacterial DNA was purified using the DirectPrep®96 Miniprep kit (Qiagen) according to the provided instructions. The sequencing reaction was performed directly on the purified clones using the M13 Forward and Reverse primers and BigDye version 3.1 DyeDeoxy terminator reaction mixture (Applied Biosystems). Sequences were analyzed on a 3100 genetic Analyzer (Applied Biosystems).

Computational data analysis

Reads were mapped with the RMAPBS program, freely available from the authors as Open Source software under the GNU Public License. A suite of software tools was implemented (also available from the authors) to estimate methylation frequencies of individual CpGs, tabulate statistics about methylation

in each CpG island, and compile diagnostic statistics about bisulfite capture experiments. Details are provided in Supplementary Information.

Enrichment was computed as (reads mapped to genome/reads overlapping target regions) / (size of target regions/size of mappable genome). The bisulfite conversion rate was estimated as the ratio of thymines over the sum of cytosines and thymines mapping over genomic non-CpG cytosines. Bisulfite conversion rate was determined using reads mapping anywhere in the genome. Coverage was determined by counting the number of reads mapping over each base in the target regions.

Assigning CpG methylation status

Methylation status of individual CpGs were called using the frequency of methylated reads mapping over each CpG and the total number of reads mapping over the CpG, making use of a Binomial confidence interval. If the upper 0.95 confidence bound was less than 0.25, then we called that CpG unmethylated in the sample. If the lower 0.95 confidence bound was at least 0.75, then we called that CpG methylated in the sample. For the remaining CpGs, if the difference between the upper and lower 0.95 confidence bounds was less than or equal to 0.25, then we called the CpG “partially methylated” in that sample. Regardless of the observed frequency of Cs and Ts mapping over a CpG, if the difference between the upper and lower confidence bounds was greater than 0.25, we concluded that a confident call could not be made. Additional details are given along with graphical description in Supplementary Methods and Figure S1.

Acknowledgments

We thank Danea Rebolini, Laura Cardone and Melissa Kramer for help with Illumina sequencing and Jeremy Hicks and Patty Bird for help in preparing the manuscript and illustrations. We also thank Stephanie Muller for Sanger sequencing of bisulfite PCR clones. EH is supported by training grant T32 CA00917631. This work was supported by grants from the Department of the Army W81XWH04-10477, the DOD Breast Cancer Research Program (GJH), The Breast Cancer Research Foundation (JH, MW), by grants from the NIH (GJH, MQZ, WRM), and by a kind gift from Kathryn W. Davis (GJH). M. W. is an American Cancer Research Professor and GJH is an Investigator of the Howard Hughes Medical Institute.

Figure Legends

Figure 1. Bisulfite capture procedure. Genomic DNA was randomly fragmented according to the standard Illumina protocol and ligated to custom-synthesized adapters in which each C was replaced by 5-meC. The ligation was size fractionated to select material from 150-300 bases in length. The gel-eluted material was treated with sodium bisulfite (see Methods) and then PCR enriched using Illumina Paired-End PCR primers. The resulting products were hybridized to custom-synthesized Agilent 244K arrays containing probes complementary to the A-rich strands. Hybridizations were carried out with Agilent array CGH buffers under standard conditions. After washing, captured fragments were eluted in water at 95^{o}C and amplified again prior to quantification and sequencing on the Illumina GA2 platform.

Figure 2. Mapping bisulfite treated reads. (A) Reads were mapped to the reference genome by minimizing the number of potential mismatches. Any T in a read incurred no penalty for aligning with a C in the genome, and any C in a read was penalized for aligning with a T in the genome. (B) Quality scores were converted to mismatch penalties by assigning a penalty of 0 to the consensus base, and penalizing non-consensus bases proportionately to the difference between their quality score and the consensus base score. A difference of 80 (representing the maximum possible range at a single position) was equated with a penalty of 1.

Figure 3. Distribution of CpG methylation frequencies. A pairwise comparison of methylation at individual CpG sites between the two samples is shown (A). For each sample, scatter plots of the proportion methylated for each CpG (x-axis) and the subsequent neighboring CpG within an island (y-axis, CpG+1) is displayed (B, C). This analysis was restricted to those CpGs with at least 40 reads in both samples.

Figure 4. Methylation status of bisulfite sequenced clones. Four independent CGI loci are shown. Two histograms plot methylation frequencies at individual CpG sites for both the bisulfite capture data (upper) and the conventional bisulfite cloning data (lower) for all four loci (A-D). The block diagrams illustrate methylation state at each CpG site for each individually analyzed clone.

Figure 5. Patterns of methylation in CpG islands. Graphical representation of methylation patterns in 9 CpG islands. A pair of graphics represents each CpG island, one graphic for each sample (CHP-SKN-1 on top, MDA-MB-231 below). Each graphic shows a pair of plots, both with bars indicating the amount of methylated (yellow) and unmethylated (blue) reads mapping over each CpG. The upper plot shows the absolute numbers of reads and spacing between CpGs. The lower plot shows the proportions of methylated and unmethylated reads. Confidence intervals are indicated in grey, and the yellow bar inside the

confidence interval indicates the exact methylation frequency. Similar plots for the remaining CGIs are given in Figure S5.

Supplementary Figure S1. Method of calling CpG methylation status. Calls were determined by considering both methylation rates of reads mapping over the CpG and the width of the 95% confidence interval for the estimate. (A) CpGs for which the confidence interval was contained below 0.25 were called unmethylated; (B) CpGs for which the confidence interval was entirely above 0.75 were called methylated. Partial methylation was called confidently if the confidence interval had width smaller than 0.25 (C) and no call was made if the interval was wider than 0.25 (D).

Supplementary Figure S2. Distribution of maximum probe distances. The numbers of capture array probe pairs (y-axis) as a function of the number of the maximum number of possible mismatches relative to the sequence the probe pairs were designed to detect. This number is equal to half the number of CpGs in the probe.

Supplementary Figure S3. Distribution of CpG methylation frequencies. Histogram showing frequencies of methylation proportions at individual CpGs. Black bars represent data from the normal skin cell line (CHP-SKN-1). Gray bars represent data from breast tumor line MDA-MB-231. Only CpGs covered by sufficient reads to make a confident call at a frequency of 0.5 (at least 41 reads) were included (see Supplementary Methods for details).

Supplementary Figure S4. Distribution of CGI methylation frequencies. The histogram shows frequencies of methylation proportions in CGIs. Black bars represent data from the normal skin cell line (CHP-SKN-1). Gray bars represent data from breast tumor line MDA-MB-231. Only CGIs for which 90% of the CpGs were covered by at least one read were included (see Supplementary Methods for details on assigning methylation frequencies to CGIs).

Supplementary Figure S5. CGI methylation profile plots. Methylation profile plots for all 324 CGIs examined, identical to those presented for selected examples in Figure 5.

Supplementary Figure S6. Comparative changes in CGI methylation states based on genomic location. The promoter class was defined as a CpG located within 1Kbp upstream of the TSS. Likewise, the CGIs overlapping the 1Kbp promoter were also counted. The intragenic class was defined as anything that overlaps the transcript, which is defined to start 1Kbp 3' of the actual start, so not including the promoter as defined above. For the intergenic category, all bases not covered by a transcript were included, with transcripts expanded by 1Kbp. The numbers shown are counts of CpGs that are contained within these 3 kinds of regions, and the numbers for CGIs are for those that overlap the above 3 kinds of regions. Because a CGI can overlap both a promoter and an intergenic region,

the combined categories will sum to more than 324 for each particular methylation state.

Table Legends

Table 1. Bisulfite capture statistics.
Statistics describing data from various stages of the bisulfite capture experiment. See Methods section for definitions of each value.

Table 2. CpG methylation call frequencies.
Summary of methylation states determined for individual CpGs in the CHP-SKN-1 and MDA-MB-231 samples. See Methods section for criteria used to assign these calls.

Table 3. Comparison of CpG calls.
Numbers of CpGs having each combination of calls in the two samples.

Table 4: CGI methylation call frequencies.
Summary of methylation states determined for CpG islands in the CHP-SKN-1 and MDA-MB-231 samples. See Supplementary Methods for criteria used to assign these calls.

Table 5: Comparison of CGI calls.
Numbers of CGIs having each combination of calls in the two samples.

Supplementary Table 1. Bisulfite dead zones.
This table summarizes the portion of the genome and target regions covered by "bisulfite dead zones" for 36-base reads, assuming full bisulfite conversion, and for assumptions of full methylation and no methylation.

Supplementary Table 2. Methylation frequencies inside validation regions
The number of methylated and unmethylated reads mapping over each CpG inside the validation regions, for both traditional bisulfite sequencing and bisulfite capture**.**

Supplementary Table 3-4. Correlation of CGI dinucleotide frequency with methylation frequency in CHP-SKN-1 and MDA-MB-231.

Supplementary Table 5. Comparative changes in CGI methylation states based on genomic location. This table shows the numbers of CGIs having each combination of calls in the two samples with respect to their genomic location either in promoters, intragenic and intergenic regions. See Figure S6 for details regarding how these regions are defined.

Supplementary Table 6. Blocking sequences
This table lists the sequences of the oligonucleotides used to block adaptor self-ligation in the hybridization mixture.

Supplementary Table 7. Bisulfite Sequencing Primers

This table lists the amplified loci and corresponding forward and reverse primers chosen for conventional bisulfite PCR and cloning.

References

Albert, T.J., M.N. Molla, D.M. Muzny, L. Nazareth, D. Wheeler, X. Song, T.A. Richmond, C.M. Middle, M.J. Rodesch, C.J. Packard et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4:** 903-905.

Ball, M.P., J.B. Li, Y. Gao, J.H. Lee, E.M. LeProust, I.H. Park, B. Xie, G.Q. Daley, and G.M. Church. 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **27:** 361-368.

Bestor, T., A. Laudano, R. Mattaliano, and V. Ingram. 1988. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J Mol Biol* **203:** 971-983.

Bestor, T.H. 1992. Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. *EMBO J* **11:** 2611-2617.

Bird, A.P. 1986. CpG-rich islands and the function of DNA methylation. *Nature* **321:** 209-213.

Bird, A.P. and M.H. Taggart. 1980. Variable patterns of total DNA and rDNA methylation in animals. *Nucleic Acids Res* **8:** 1485-1497.

Brunner, A.L., D.S. Johnson, S.W. Kim, A. Valouev, T.E. Reddy, N.F. Neff, E. Anton, C. Medina, L. Nguyen, E. Chiao et al. 2009. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res*.

Chaillet, J.R., T.F. Vogt, D.R. Beier, and P. Leder. 1991. Parental-specific methylation of an imprinted transgene is established during gametogenesis and progressively changes during embryogenesis. *Cell* **66:** 77-83.

Cokus, S.J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C.D. Haudenschild, S. Pradhan, S.F. Nelson, M. Pellegrini, and S.E. Jacobsen. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452:** 215-219.

Deng, J., R. Shoemaker, B. Xie, A. Gore, E.M. LeProust, J. Antosiewicz-Bourget, D. Egli, N. Maherali, I.H. Park, J. Yu et al. 2009. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* **27:** 353-360.

Dupont, J.M., J. Tost, H. Jammes, and I.G. Gut. 2004. De novo quantitative bisulfite sequencing using the pyrosequencing technology. *Anal Biochem* **333:** 119-127.

Eads, C.A., K.D. Danenberg, K. Kawakami, L.B. Saltz, C. Blake, D. Shibata, P.V. Danenberg, and P.W. Laird. 2000. MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Res* **28:** E32.

Eckhardt, F., J. Lewin, R. Cortese, V.K. Rakyan, J. Attwood, M. Burger, J. Burton, T.V. Cox, R. Davies, T.A. Down et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38:** 1378-1385.

Ehrich, M., M.R. Nelson, P. Stanssens, M. Zabeau, T. Liloglou, G. Xinarianos, C.R. Cantor, J.K. Field, and D. van den Boom. 2005. Quantitative high-

throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proc Natl Acad Sci U S A* **102:** 15785-15790.

Ehrich, M., J. Turner, P. Gibbs, L. Lipton, M. Giovanneti, C. Cantor, and D. van den Boom. 2008. Cytosine methylation profiling of cancer cell lines. *Proc Natl Acad Sci U S A* **105:** 4844-4849.

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

Haines, T.R., D.I. Rodenhiser, and P.J. Ainsworth. 2001. Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. *Dev Biol* **240:** 585-598.

Herman, J.G. and S.B. Baylin. 2003. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* **349:** 2042-2054.

Hicks, J., A. Krasnitz, B. Lakshmi, N.E. Navin, M. Riggs, E. Leibu, D. Esposito, J. Alexander, J. Troge, V. Grubor et al. 2006. Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res* **16:** 1465-1479.

Hodges, E., Z. Xuan, V. Balija, M. Kramer, M.N. Molla, S.W. Smith, C.M. Middle, M.J. Rodesch, T.J. Albert, G.J. Hannon et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39:** 1522-1527.

Holliday, R. and J.E. Pugh. 1975. DNA modification mechanisms and gene activity during development. *Science* **187:** 226-232.

Hughes, T.R., M. Mao, A.R. Jones, J. Burchard, M.J. Marton, K.W. Shannon, S.M. Lefkowitz, M. Ziman, J.M. Schelter, M.R. Meyer et al. 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19:** 342-347.

Irizarry, R.A., C. Ladd-Acosta, B. Carvalho, H. Wu, S.A. Brandenburg, J.A. Jeddeloh, B. Wen, and A.P. Feinberg. 2008. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res* **18:** 780-790.

Keshet, I., J. Lieman-Hurwitz, and H. Cedar. 1986. DNA methylation affects the formation of active chromatin. *Cell* **44:** 535-543.

Khulan, B., R.F. Thompson, K. Ye, M.J. Fazzari, M. Suzuki, E. Stasiek, M.E. Figueroa, J.L. Glass, Q. Chen, C. Montagna et al. 2006. Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res* **16:** 1046-1055.

Laird, P.W. 2003. The power and the promise of DNA methylation markers. *Nat Rev Cancer* **3:** 253-266.

Lengauer, C., K.W. Kinzler, and B. Vogelstein. 1997. DNA methylation and genetic instability in colorectal cancer cells. *Proc Natl Acad Sci U S A* **94:** 2545-2550.

Li, E., T.H. Bestor, and R. Jaenisch. 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69:** 915-926.

Li, L.C. and R. Dahiya. 2002. MethPrimer: designing primers for methylation PCRs. *Bioinformatics* **18:** 1427-1431.

Lippman, Z., A.V. Gendrel, M. Black, M.W. Vaughn, N. Dedhia, W.R. McCombie, K. Lavine, V. Mittal, B. May, K.D. Kasschau et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430:** 471-476.

Luedi, P.P., F.S. Dietrich, J.R. Weidman, J.M. Bosko, R.L. Jirtle, and A.J. Hartemink. 2007. Computational and experimental identification of novel human imprinted genes. *Genome Res* **17:** 1723-1730.

Meissner, A., T.S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B.E. Bernstein, C. Nusbaum, D.B. Jaffe et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454:** 766-770.

Monk, M., M. Boubelik, and S. Lehnert. 1987. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. *Development* **99:** 371-382.

Okano, M., D.W. Bell, D.A. Haber, and E. Li. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99:** 247-257.

Okano, M., S. Xie, and E. Li. 1998. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* **19:** 219-220.

Okou, D.T., K.M. Steinberg, C. Middle, D.J. Cutler, T.J. Albert, and M.E. Zwick. 2007. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* **4:** 907-909.

Pevzner, P.A. and M.S. Waterman. 1995. Multiple filtration and approximate pattern matching. *Algorithmica* **13:** 135-154.

Rauch, T., Z. Wang, X. Zhang, X. Zhong, X. Wu, S.K. Lau, K.H. Kernstine, A.D. Riggs, and G.P. Pfeifer. 2007. Homeobox gene methylation in lung cancer studied by genome-wide analysis with a microarray-based methylated CpG island recovery assay. *Proc Natl Acad Sci U S A* **104:** 5527-5532.

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

Reik, W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447:** 425-432.

Rideout, W.M., 3rd, G.A. Coetzee, A.F. Olumi, and P.A. Jones. 1990. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science* **249:** 1288-1290.

Sanford, J.P., H.J. Clark, V.M. Chapman, and J. Rossant. 1987. Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse. *Genes Dev* **1:** 1039-1046.

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

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525-528.

Shen, L., Y. Kondo, Y. Guo, J. Zhang, L. Zhang, S. Ahmed, J. Shu, X. Chen, R.A. Waterland, and J.P. Issa. 2007. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet* **3:** 2023-2036.

Smith, A., Z. Xuan, and M. Zhang. 2008. Using quality scores and longer reads improves accuracy of Solexa read mapping *BMC Bioinformatics* **9:** 128.

Taylor, K.H., R.S. Kramer, J.W. Davis, J. Guo, D.J. Duff, D. Xu, C.W. Caldwell, and H. Shi. 2007. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res* **67:** 8511-8518.

Waddington, C. 1942. The epigenotype. *Endeavour* **1:** 18-20.

Weber, M., J.J. Davies, D. Wittig, E.J. Oakeley, M. Haase, W.L. Lam, and D. Schubeler. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37:** 853-862.

Wilson, E.B. 1927. Probable inference , the law of succession, and statistical inference. . *Journal of the American Statistical Association* **22:** 209-212.

Xiong, Z. and P.W. Laird. 1997. COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic Acids Res* **25:** 2532-2534.

Zhang, Y., C. Rohde, S. Tierling, T.P. Jurkowski, C. Bock, D. Santacruz, S. Ragozin, R. Reinhardt, M. Groth, J. Walter et al. 2009. DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS Genet* **5:** e1000438.

Table 1. Bisulfite Capture Statistics.

Sample	**CHP-SKN-1**	**MBA-MB-231**
*Reads sequenced**	55,770,254	20,002,207
Reads mapped (unambiguous)	12,130,697	7,575,990
Reads in target region	780,471	907,592
Percent mapped reads in target	6.43%	11.98%
Enrichment	711.14	1324.14
Target region coverage (at least one read)	94.23%	93.56%
Target region coverage (at least 10 reads)	92.97%	92.50%
Median read depth at target CpGs	95	105
*Bisulfite conversion rate***	98.85%	98.75%
Target region size	258,571	
Genome size†	2,858,008,658	
Expected % mapped reads in target (i.e. uncaptured)	0.009%	

**Numbers represent sequenced data combined from multiple lanes (4 lanes for CHP-SKN-1 and 2 lanes for MDA-MB-231).*

***Includes reads mapping outside target regions*

†Excludes unassembled regions larger than 1000 bases

Table 2. CpG methylation call frequencies.

Sample	**CHP-SKN-1**		**MDA-MB-231**	
Unmethylated	18398	73.46%	13456	53.73%
Partially Methylated	2018	8.06%	3681	14.70%
Methylated	2660	10.62%	5791	23.12%
No Call	1968	7.86%	2116	8.45%
	25044		25044	

Total called in CHP-SKN-1	23076	92.14%
Total called in MDA-MB-231	22928	91.55%

Table 3. Comparison of CpG calls.

	MDA-MB-231				
CHP-SKN-1	*Unmethylated*	*Partially Methylated*	*Methylated*	*No Call*	*Total*
Unmethylated	13162	2588	2342	306	18398
Partially Methylated	172	612	1198	36	2018
Methylated	46	416	2148	50	2660
No Call	76	65	103	1724	1968
Total	13456	3681	5791	2116	25044

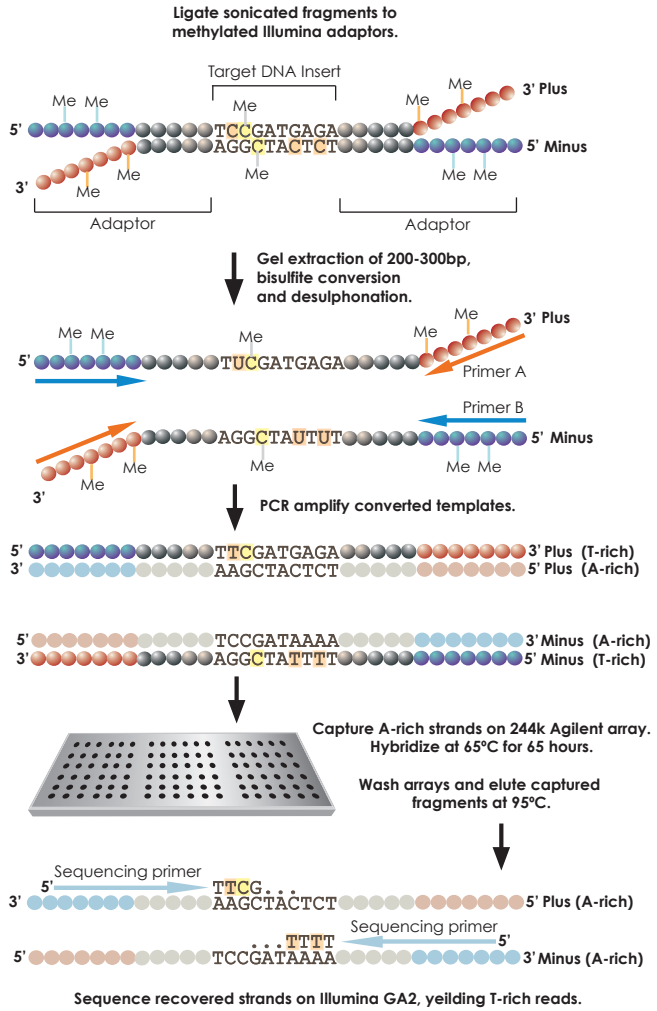
Total called in both samples	22684

Table 4. CGI call frequencies.

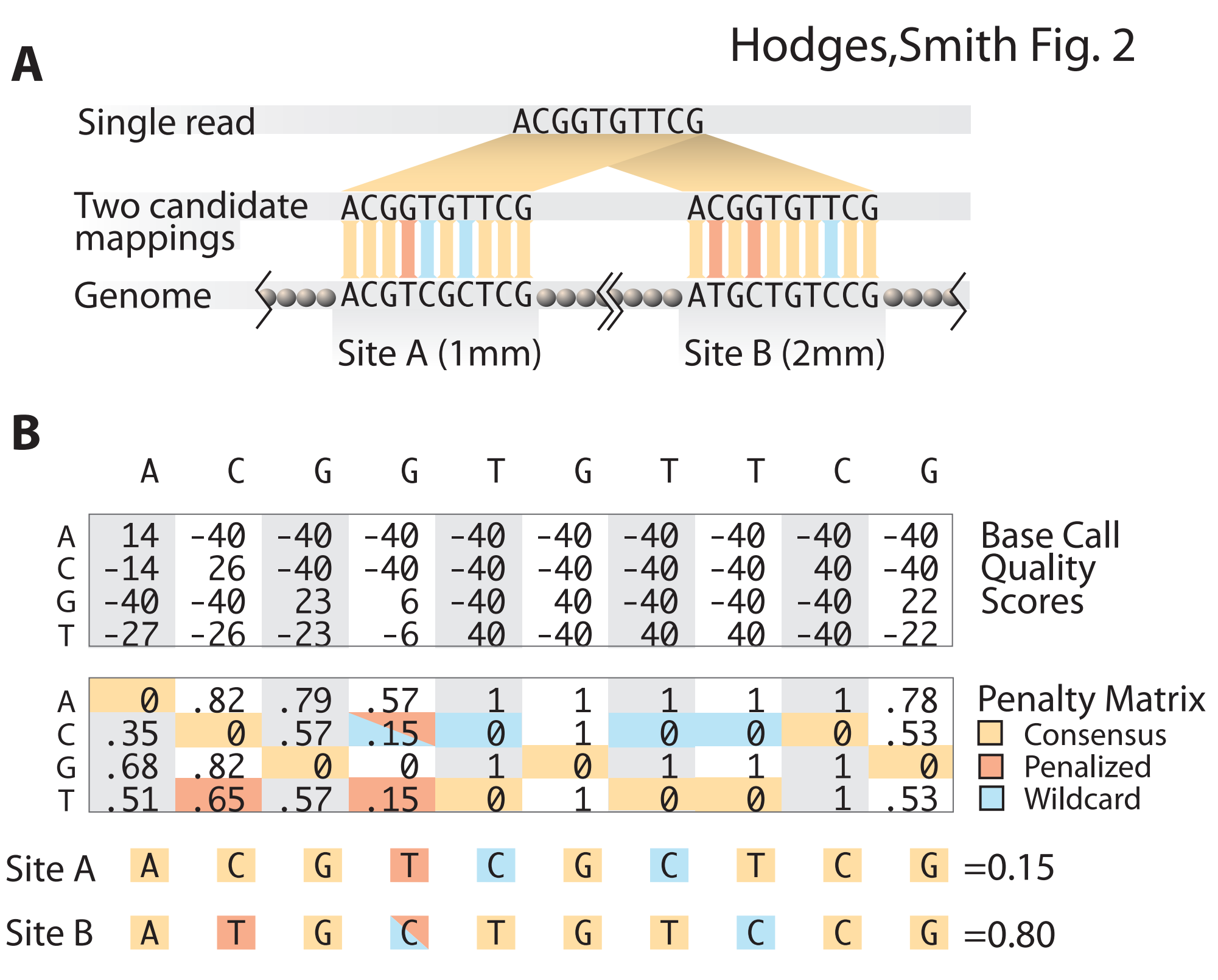
Sample	**CHP-SKN-1**		**MDA-MB-231**	
Unmethylated	210	64.81%	145	44.75%
Partially Methylated	31	9.57%	71	21.91%
Methylated	42	12.96%	64	19.75%
No Call	41	12.65%	44	13.58%
	324		324	

Table 5. Comparison of CGI calls.

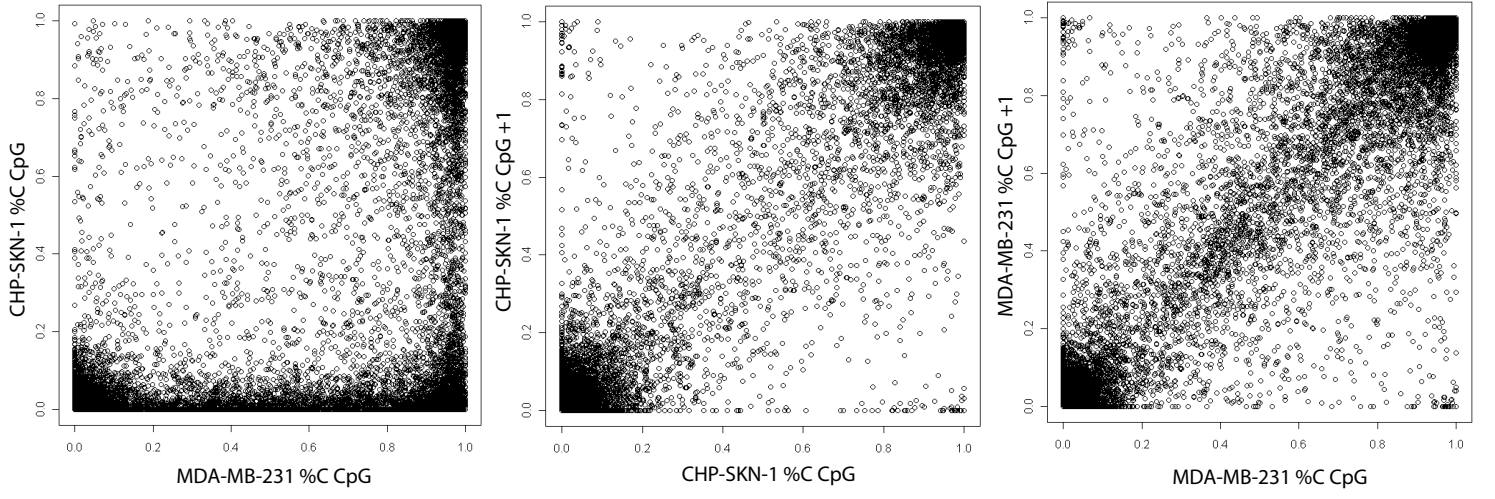
	MDA-MB-231				
CHP-SKN-1	*Unmethylated*	*Partially Methylated*	*Methylated*	*No Call*	*Total*
Unmethylated	143	51	15	1	210
Partially Methylated	2	10	18	1	31
Methylated	0	10	31	1	42
No Call	0	0	0	41	41
Total	145	71	64	44	324

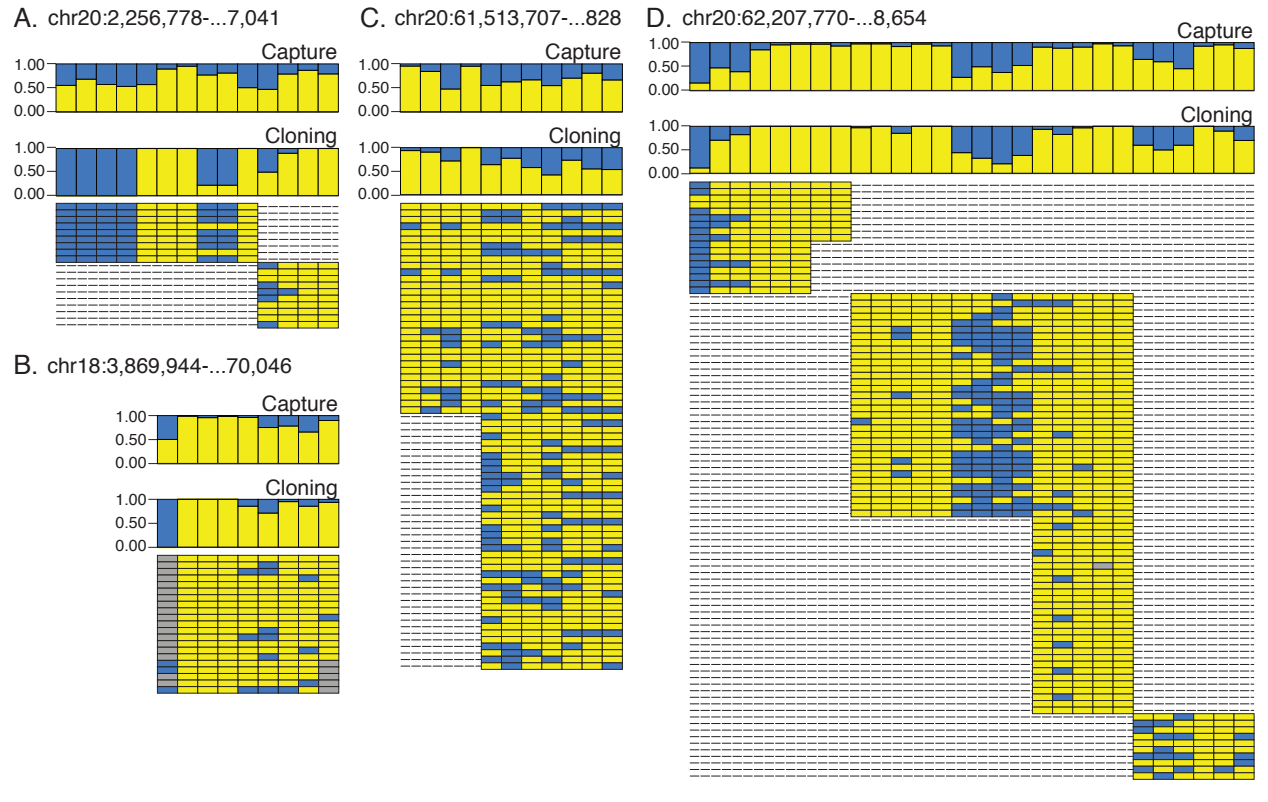

Ligate sonicated fragments to methylated Illumina adaptors.
Target DNA Insert
Me
5'
TCCGATGAGA
AGGCTACTCT
3' Plus
5' Minus
3'
Adaptor
Adaptor
Gel extraction of 200-300bp, bisulfite conversion and desulphonation.
TUCGATGAGA
Primer A
Primer B
AGGCTAUTUT
PCR amplify converted templates.
TTCGATGAGA
3' Plus (T-rich)
AAGCTACTCT
5' Plus (A-rich)
TCCGATAAAA
3' Minus (A-rich)
AGGCTATTTT
5' Minus (T-rich)
Capture A-rich strands on 244k Agilent array. Hybridize at 65ºC for 65 hours.
Wash arrays and elute captured fragments at 95ºC.
Sequencing primer
TTCG...
5' Plus (A-rich)
...TTTT
Sequencing primer
TCCGATAAAA
3' Minus (A-rich)
Sequence recovered strands on Illumina GA2, yeilding T-rich reads.

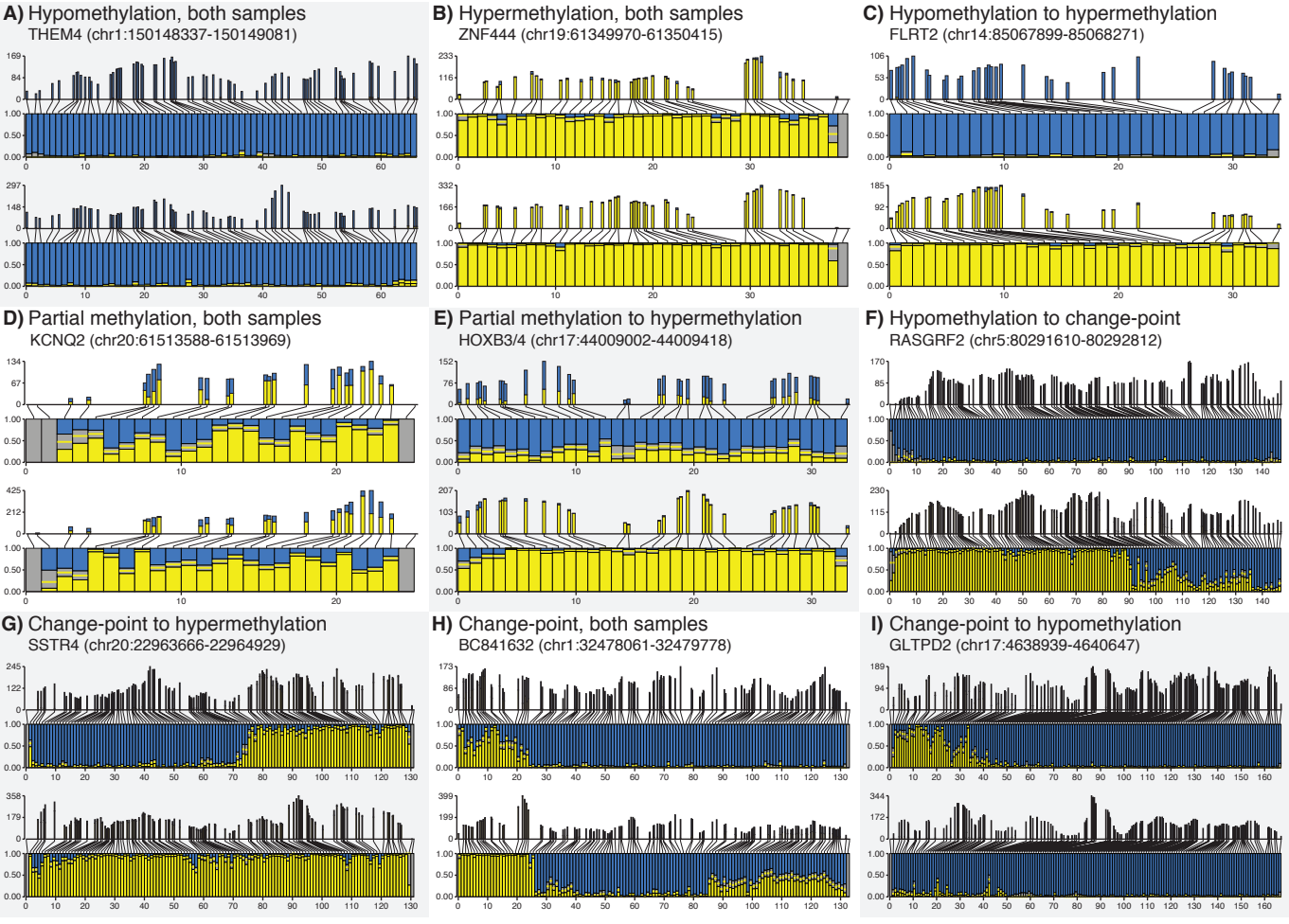
Hodges,Smith Fig. 2



Hodges,Smith Fig. 3




A. chr20:2,256,778-...7,041
Capture
Cloning
1.00
0.50
0.00
B. chr18:3,869,944-...70,046
Capture
Cloning
1.00
0.50
0.00
C. chr20:61,513,707-...828
Capture
Cloning
1.00
0.50
0.00
D. chr20:62,207,770-...8,654
Capture
Cloning
1.00
0.50
0.00


A) Hypomethylation, both samples
THEM4 (chr1:150148337-150149081)
B) Hypermethylation, both samples
ZNF444 (chr19:61349970-61350415)
C) Hypomethylation to hypermethylation
FLRT2 (chr14:85067899-85068271)
D) Partial methylation, both samples
KCNQ2 (chr20:61513588-61513969)
E) Partial methylation to hypermethylation
HOXB3/4 (chr17:44009002-44009418)
F) Hypomethylation to change-point
RASGRF2 (chr5:80291610-80292812)
G) Change-point to hypermethylation
SSTR4 (chr20:22963666-22964929)
H) Change-point, both samples
BC841632 (chr1:32478061-32479778)
I) Change-point to hypomethylation
GLTPD2 (chr17:4638939-4640647)