# A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis

Thomas A Down[1,8], Vardhman K Rakyan[2,8], Daniel J Turner[3], Paul Flicek[4], Heng Li[3], Eugene Kulesha[4], Stefan Gräf[4], Nathan Johnson[4], Javier Herrero[4], Eleni M Tomazou[3], Natalie P Thorne[5], Liselotte Bäckdahl[6], Marlis Herberth[7], Kevin L Howe[5], David K Jackson[3], Marcos M Miretti[3], John C Marioni[5], Ewan Birney[4], Tim J P Hubbard[3], Richard Durbin[3], Simon Tavaré[5] & Stephan Beck[6]

**DNA methylation is an indispensible epigenetic modification required for regulating the expression of mammalian genomes. Immunoprecipitation-based methods for DNA methylome analysis are rapidly shifting the bottleneck in this field from data generation to data analysis, necessitating the development of better analytical tools. In particular, an inability to estimate absolute methylation levels remains a major analytical difficulty associated with immunoprecipitation-based DNA methylation profiling. To address this issue, we developed a cross-platform algorithm—Bayesian tool for methylation analysis (Batman)—for analyzing methylated DNA immunoprecipitation (MeDIP) profiles generated using oligonucleotide arrays (MeDIP-chip) or next-generation sequencing (MeDIP-seq). We developed the latter approach to provide a high-resolution whole-genome DNA methylation profile (DNA methylome) of a mammalian genome. Strong correlation of our data, obtained using mature human spermatozoa, with those obtained using bisulfite sequencing suggest that combining MeDIP-seq or MeDIP-chip with Batman provides a robust, quantitative and cost-effective functional genomic strategy for elucidating the function of DNA methylation.**

Modulation of the epigenome—the combination of DNA- and chromatin-associated epigenetic modifications that exist within a cell—is one of the key mechanisms by which cells generate functional diversity from an essentially static genome[1]. The epigenome is a dynamic entity influenced by predetermined genetic programs or external environmental cues. Given the diversity of cell types within complex organisms such as mammals, it is staggering to think of how many epigenomes exist, or are possible, and unraveling this complexity remains an important challenge.

DNA methylation is the only known epigenetic system that modifies the DNA molecule itself. In mammals, it occurs predominantly at CpG dinucleotides and is involved in diverse processes such as development, genomic integrity, X-inactivation and imprinting[2]. Furthermore, perturbed DNA methyation is a hallmark of several human diseases, including cancer. Consequently, there is great interest in experimental and analytical tools for genome-wide (that is, a limited number of genomic sites that are representative of the genome) or whole-genome DNA methylation profiling. In the last few years, a variety of experimental approaches have emerged for genome-wide, and very recently whole-genome, DNA methylation profiling (reviewed in ref. 3). These can be classified into three main categories. The first of these, restriction enzyme–based methods, uses one or more enzymes that restrict DNA only if it is unmethylated (e.g., *Hpa*II or *Not*I) or methylated (e.g., McrBC). These methods, coupled with either microarrays[4–10] or capillary sequencing[11], have been applied to genome-wide DNA methylation profiling of several organisms but are limited to the analysis of CpG sites located within the enzyme recognition site(s). The second group of techniques is based on the reaction between genomic DNA and sodium bisulfite, which converts unmethylated cytosines to uracil (and eventually thymine following amplification), while leaving methylated cytosines unconverted[12]. Bisulfite conversion–based approaches offer single CpG resolution and have been applied to microarrays[13–16], high-throughput PCR sequencing[17,18] and, more recently, to next-generation bisulfite sequencing (BS-seq)[19], resulting in an almost complete DNA methylation profile (DNA methylome) for the ~120-Mb genome of *Arabidopsis thaliana*. However, the reduction of sequence complexity following bisulfite conversion means that it is difficult to design enough unique probes to analyze bisulfite-converted DNA comprehensively on a genome-wide scale on microarrays, whereas the BS-seq approach is currently prohibitively expensive for the routine analysis of large genomes such as human. Methods of the third class use either 5-methylcytosine-specific antibodies (methylated DNA immunoprecipitation[20], MeDIP or mDIP[21]) or methyl-binding domain proteins[22–24] to enrich for the methylated (or unmethylated[25]) fraction of the genome by immunopreciptation. MeDIP/mDIP, combined with microarrays (MeDIP-chip), was used to delineate the first high-resolution whole-genome DNA methylation profile of any

[1]Wellcome Trust Cancer Research UK Gurdon Institute, and Department of Genetics, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK. [2]Institute of Cell and Molecular Science, Barts and The London, 4 Newark Street, London, E1 2AT, UK. [3]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK. [4]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus Hinxton, Cambridge CB10 1SD, UK. [5]Department of Oncology, University of Cambridge, Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [6]UCL Cancer Institute, University College London, 72 Huntley Street, London, WC1E 6BT, UK. [7]Institute of Biotechnology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QT, UK. [8]These authors contributed equally to this work. Correspondence should be addressed to V.K.R. (v.rakyan@qmul.ac.uk), T.A.D. (thomas.down@gurdon.cam.ac.uk) or S.B. (s.beck@ucl.ac.uk).

Published online 8 July 2008; doi:10.1038/nbt1414

genome (*Arabidopsis*[22,26]) and the first high-resolution DNA methylation profile of human promoters[27]. However, until now, it has not been possible to estimate absolute methylation levels from MeDIP, and analysis of regions with low CpG density has been assumed to be problematic[27].

Although no single experimental method offers the 'perfect solution', MeDIP-chip has quickly become a widely used[20–22,26–31] and cost-effective approach for genome-wide and/or whole-genome DNA methylation analysis. Here, we report the development of a cross-platform algorithm—Batman—that can estimate absolute DNA methylation levels, across a wide range of CpG densities, from MeDIP-based experiments. We first demonstrate Batman's performance on MeDIP-chip, and then show it can also be used to analyze MeDIP profiles generated from next-generation sequencing—a technique we called MeDIP-seq, described here. Our MeDIP-seq data represent a high-resolution whole-genome DNA methylation profile of a mammalian genome, which to our knowledge has not been done before. Batman is a cross-platform analytical tool for data generated from microarrays or next-generation sequencing and will aid future studies aiming to understand the role of DNA methylation in the wider context of the epigenome.

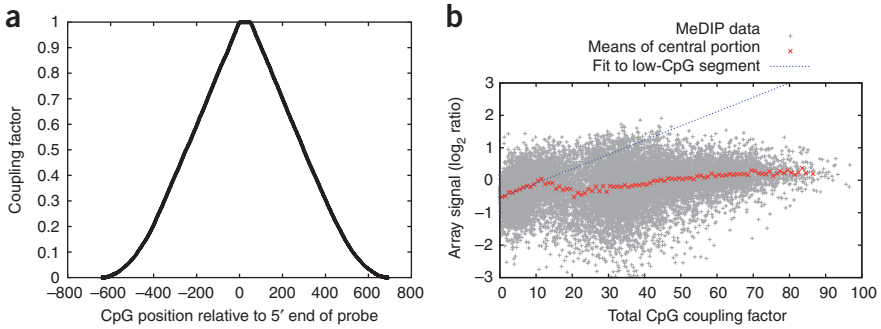


**Figure 1** Calibration of the Batman model against MeDIP-chip data. (**a**) Estimated CpG coupling factors for a MeDIP-chip experiment as a function of the distance between a CpG dinucleotide and a microarray probe. (**b**) Plot of array signal against total CpG coupling factor, showing a linear regression fit to the low-CpG portion, as used in the Batman calibration step. This plot shows all data from one array on chromosome 6.

## RESULTS

### Generation of human genome-wide MeDIP-chip data

MeDIP was performed on three biological replicates of mature spermatozoa from normal human donors (**Supplementary Table 1** online) using a modified version of the original MeDIP protocol[20] (**Supplementary Figs. 1** and **2** online). Human spermatozoa are relatively homogenous, easily obtained and of interest from the point of view of understanding the role of DNA methylation during gametogenesis, fertilization and early embryogenesis. After MeDIP, samples were hybridized to custom high-density oligonucleotide microarrays (Nimblegen Systems) that contained 42,144 regions of interest (ROIs), each typically 500–1,000 bp in length, containing 5–10 unique 50-mer probes. The ROIs overlapped 82% of all known transcriptional start sites (TSSs), 72% of nonpromoter CpG islands and a number of exonic, intronic and intergenic regions in the human genome (Ensembl genome browser[32], *Homo sapiens* release 45.36 g, NCBI36). The correlation coeffecients (Pearson's) ranged from 0.54 to 0.72 among the three biological replicates and 0.82 between a pair of technical replicates (dye swaps), suggesting our MeDIP-chip experiments were reproducible.

### Bayesian tool for methylation analysis (Batman)

The efficiency of immunoprecipitation in MeDIP depends on the density of methylated CpG sites, which vary greatly within any given mammalian genome, making it difficult to distinguish variations in enrichment from confounding CpG density effects[27]. Consequently, until now, it has been impossible to estimate absolute methylation levels from MeDIP experiments, and the analysis of CpG-poor regions, in particular, has been assumed to be difficult[27]. Therefore, to analyze our MeDIP-chip data, we developed a new algorithm that models the effect of varying densities of methylated CpGs on MeDIP enrichment. This transforms normalized MeDIP-chip $\log_2$-ratios into a quantitative measure of DNA methylation across a wide range of CpG densities. Our algorithm, Batman, is implemented as a suite of Java scripts (freely available from http://td-blade.gurdon.cam.ac.uk/software/batman/ under the GNU Lesser General Public License).

Batman relies on the knowledge that almost all DNA methylation in mammals occurs at CpG dinucleotides and, consequently, generates methylation estimates in this context only. We define the coupling factor, $C_{cp}$, between probe $p$ and CpG dinucleotide $c$ as the fraction of DNA molecules hybridizing to probe $p$ that contain the CpG $c$. As we know the approximate range of DNA fragment sizes used in the MeDIP experiment (typically 400–700 bp) and assume that there are no fragment-length biases, this is simply a function of the distance between the probe's genomic location and the CpG dinucleotide. This can be estimated empirically by sampling from the fragment-length distribution and randomly placing each fragment such that it overlaps the probe. The resulting distribution is shown in **Figure 1a**. For a given probe, the sum of coupling factors, which we call $C_{tot}$, gives a measure of local CpG density. Plotting this parameter against the normalized $\log_2$-ratios from a typical MeDIP-chip experiment shows a fairly complex relationship (**Fig. 1b**). However, consistent with the fact that most CpG-poor regions are methylated, whereas the regions richest in CpG motifs (CpG islands) are generally unmethylated, focusing on the low-CpG portion of this plot reveals an approximately linear relationship between the MeDIP-chip output and the density of methylated CpGs as measured by $C_{tot}$. Based on this observation, and assuming that only methylated CpGs contribute to the observed signal, we developed a model whereby the signal observed at each array probe should depend on the methylation states of all nearby CpGs, weighted by the coupling factors between those CpGs and the probe. If we let $m_c$ indicate the methylation state at position $c$, and assume that the errors on the microarray are normally distributed with precision, then we can write a probability distribution for a complete set of array observations, $A$, given a set of methylation states, $m$, as:

$$f(A|m) = \prod_p G(A_p|A_{base} + r\sum_c C_{cp}m_c, v^{-1})$$

where $G(x|\mu, \sigma^2)$ is a Gaussian probability density function. We can now use any standard Bayesian inference approach to find $f(m|A)$, the posterior distribution of the methylation state parameters given the array (MeDIP-chip) data, and thus generate quantitative methylation profile information.

To reduce the computational cost of analyzing regions with very high CpG density, we took advantage of the fact that CpG methylation state is generally very highly correlated over a scale of hundreds of

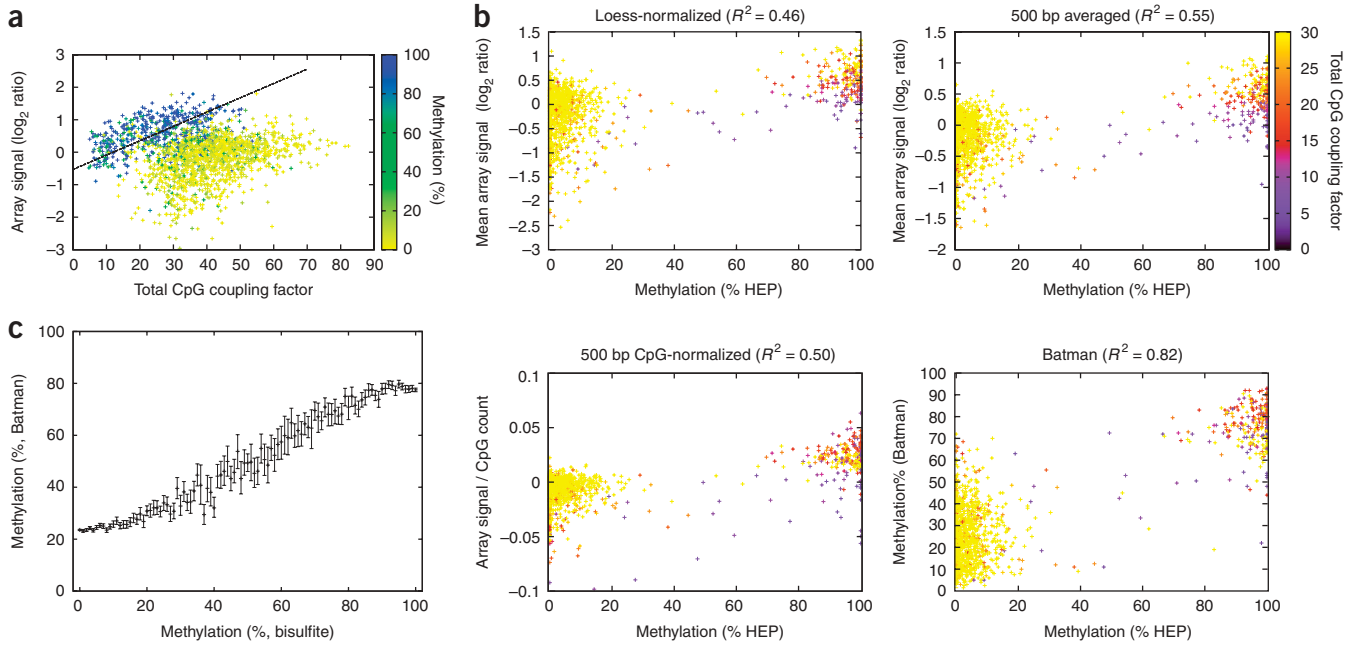


**Figure 2** Comparison of Batman-analyzed MeDIP-chip data with bisulfite-PCR sequencing data from the Human Epigenome Project. (**a**) Plot of MeDIP-chip data against CpG coupling factor, with points colored by methylation values from the HEP bisulfite-sequencing data. All probes that did not overlap with at least one CpG annotated in HEP were excluded. (**b**) Comparisons of MeDIP-chip data with those of HEP using a range of processing strategies: LOESS-normalized $\log_2$-ratios in a 100-bp window centered around a 50-mer probe that overlaps a HEP amplicon (top left), simple averaging of the LOESS-normalized $\log_2$-ratios for all probes within a 500-bp window (top right), averaging of the LOESS-normalized $\log_2$-ratios for all probes within a 500-bp window and then dividing by the observed/expected CpG density (bottom left), Batman analyzed (bottom right). This analysis was derived from 1,481 MeDIP-chip probes that overlapped 667 bisulfite-PCR amplicons from the HEP. HEP methylation values for all CpGs that overlapped any given 100-bp MeDIP-chip window were averaged. Furthermore, to reduce noise in the HEP data set, all 100-bp windows were required to have at least two HEP scores (that is, data from the top and bottom bisulfite-PCR strands for windows containing a single CpG site, or from at least two different CpG sites) that differed by $<50\%$. The purple-yellow (0–30) color bar shows the total CpG coupling factor for each probe. (**c**) Comparison of Batman-quantified MeDIP data with bisulfite data from HEP. Points show the mean Batman output for regions with a given HEP methylation level. Error bars show 95% bootstrap credible intervals.



bases[18]. Instead of modeling every CpG individually, we grouped together all CpGs in 50- or 100-bp windows and assumed that they would have the same methylation state. Inferring the methylation status at each CpG is now a deconvolution problem somewhat analogous to that considered when analyzing chromatin immunoprecipitation data[33]. Standard Bayesian techniques can be used to infer $f(m|A)$, that is, the distribution of likely methylation states given one or more sets of MeDIP-chip outputs. Our implementation of the Batman model uses nested sampling (http://www.inference.phy.cam.ac.uk/bayesys/), a highly robust Monte Carlo technique, to solve this inference problem. For each tiled region of the genome, we used a nested sampler–based approach to generate 100 independent samples from $f(m|A)$. We then summarized the most likely methylation state in 100-bp windows by fitting beta distributions to these samples. The modes of the most likely beta distributions were used as our final methylation calls.

We assessed Batman's quantitative performance by comparing the Batman-analyzed MeDIP-chip data with bisulfite-PCR sequencing, a technique that allows DNA methylation measurements at individual CpG sites. We considered 667 bisulifte-PCR amplicons (spanning a wide range of CpG densities) from the Human Epigenome Project (HEP)[18] that overlapped 1,481 50-mer probes in our microarray. The HEP bisulifte-PCR amplicons were generated from sperm samples different from those used in our MeDIP-chip experiments. **Figure 2a** shows a different version of the Batman calibration plot in which the MeDIP-chip $\log_2$-ratios have been colored according to HEP methylation levels, confirming that the calibration system we use provides a very good fit to the methylated section of the data. **Figure 2b** shows how Batman transforms LOESS-normalized $\log_2$-ratios into more quantitative results ($R^2 = 0.82$, Pearson's) by increasing the dynamic range in low-CpG regions (although some noise still remains in this region). This is a significant improvement over using (i) LOESS-normalized $\log_2$-ratios (in a 100-bp window centered around a 50-mer) probe that overlaps a HEP amplicon, ($R^2 = 0.46$, Pearson's correlation coefficient)), (ii) simple averaging of the LOESS-normalized $\log_2$-ratios for all probes within a 500-bp window ($R^2 = 0.55$, Pearson's correlation coefficient) or (iii) averaging of the LOESS-normalized $\log_2$-ratios for all probes within a 500-bp window and then dividing by the number of CpG sites within that window ($R^2 = 0.50$, Pearson's correlation coefficient). There are two likely explanations for the poor performance of the last method: firstly, as it isn't a Bayesian method, there is no propagation of uncertainty (consequently noise in low-CpG regions is amplified), and secondly, the CpG influence is not necessarily the same for all probes in a 500-bp window. Batman addresses both of these issues. It is important to note that, in addition to estimating methylation levels in CpG-poor regions, Batman also effectively estimates methylation levels in CpG-dense methylated regions. Of the 667 bisulfite-PCR amplicons mentioned above, 15 are classified as CpG islands in the Ensembl genome browser and display $>80\%$ methylation in the HEP. Batman

identified all 15 as being heavily methylated (81–100% methylation, **Supplementary Table 2** online). We further validated the Batman analysis by bisulfite-PCR sequencing of the same sperm samples used for MeDIP-chip. We selected 29 ROIs spanning a range of CpG densities and again observed a very good correlation ($R^2 = 0.85$, **Supplementary Fig. 3** and **Supplementary Table 3** online).

We also tested Batman's performance on an independently generated MeDIP-chip data set[27]. Weber *et al.* (2007) analyzed MeDIP profiles of ~16,000 promoters in human WI38 primary lung fibroblasts using high-density oligonucleotide arrays. We applied Batman to their MeDIP-chip data and analyzed promoters for which they also generated bisulfite-sequencing data (**Supplementary Fig. 4** online). Batman was able to estimate absolute methylation levels over a wide range of CpG densities including low CpG density promoters (or LCPs[27], $CpG_{o/e}$ ~0.2).

As there is still a degree of noise in the Batman results, we also show the mean Batman score for all regions with a given bisulfite methylation state (**Fig. 2c**), demonstrating Batman's output correlates almost linearly with the bisulfite results. It should be noted that Batman rarely outputs very extreme values (close to 0% or 100%) from MeDIP-chip data. This is a consequence of the Bayesian approach taken by Batman: each methylation call is associated with some degree of uncertainty, as represented by a credible interval. As methylation levels <0% or >100% are meaningless, the entire credible interval must fit within a 0–100% scale. This means that the most credible estimates of methylation state are displaced away from the extremes. In principle, it would be possible to correct for this 'compression' artifact by reading values off a curve (**Fig. 2c**). However, this transformation would complicate any consideration of the uncertainties attached to each methylation estimate. As we do not find the compression to be a major problem when working with MeDIP-chip data, we report the output of the Bayesian model directly.

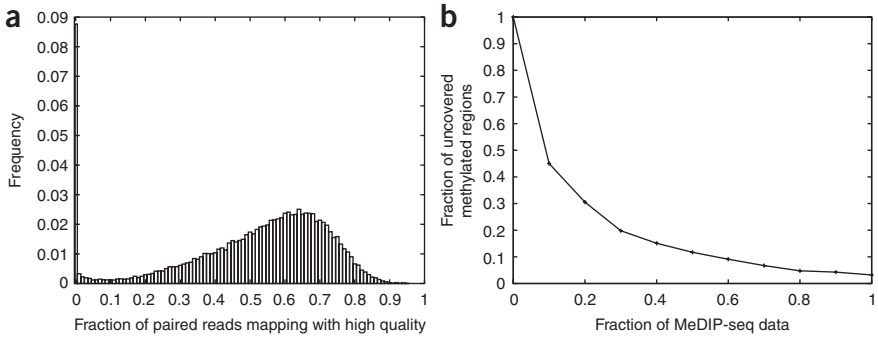


**Figure 3** Mapping quality and genomic coverage of the MeDIP-seq data. (**a**) Histogram showing the fractions of high-quality paired-end read mappings in 50-kb windows across the genome. (**b**) Fraction of methylated regions (>60% methylation) that are not covered by reads in our MeDIP-seq data set. As with all the MeDIP-seq analyses, the reads are extended to a length of 500 bp.

## A human methylome generated using MeDIP-seq

Recently, next-generation sequencing technologies have emerged as powerful tools for whole-genome profiling of epigenetic modifications. They have been combined with chromatin immunoprecipitation (ChIP-Seq)[34,35] for the analysis of histone modifications in human and mouse and with bisulfite sequencing (BS-seq[19]) to elucidate the DNA methylation profile of the 120-Mb *Arabidopsis* genome. Inspired by these approaches, we combined MeDIP with next-generation sequencing—an approach we term MeDIP-seq—to generate a high-resolution whole-genome DNA methylation profile (DNA methylome) of a mammalian genome and show that Batman can also be used to estimate absolute DNA methylation levels from MeDIP-seq DNA methylome data.

We performed a second MeDIP on one of the sperm samples used in our MeDIP-chip experiments (sample SP3, **Supplementary Table 1**). The immunoprecipitated fraction was then subjected to next-generation sequencing using an Illumina Genome Analyzer. We obtained ~34.2 million single- and ~12 million paired-end reads that were mapped to the human genome using the Maq software (http://maq.sf.net/ and Li *et al.*, data not shown). Only high-quality read placements (Maq quality ≥ 10) were used, resulting in a total of ~26.5 million reads meeting this criterion. To maximize coverage, given the relatively short reads generated by the Illumina Genome Analyzer, we performed a smoothing step on the data by extending each paired-end read to a constant length of 500 bp and representing each singleton read as a 500-bp block centered around the single read's mapping position. We do not expect this step to be necessary if longer fragments are selected.

Assessment of the mapping quality revealed a degree of nonuniformity. For instance, there is a secondary peak of windows with extremely low mapping quality (<10% of reads map with q ≥ 10) (**Fig. 3a**). Many of these windows occur in large (megabase-scale) blocks. Investigation of representative examples suggests that they correspond with known duplications/structural variations in the human genome[32] (data not shown). We chose to mask out these

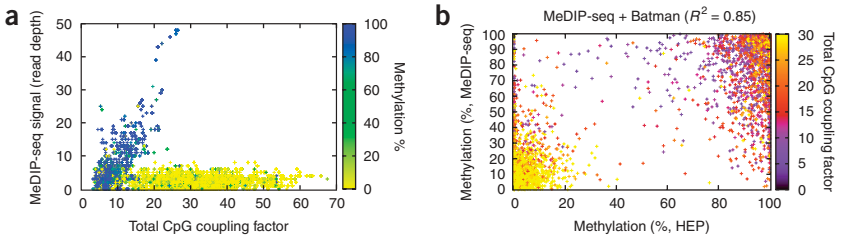


**Figure 4** Comparison of Batman-analyzed MeDIP-seq data with bisulfite-PCR sequencing data from the Human Epigenome Project. (**a**) MeDIP-seq read depth (that is, the number of confidently placed reads overlapping a given point in the genome) for points overlapping HEP amplicons, plotted against total CpG coupling factor. Points are colored according to sperm DNA methylation (yellow-blue represents 0–100% methylation), as measured in HEP[16]. (**b**) MeDIP-seq versus sperm bisulfite-PCR sequencing data from HEP[16]. 100 bp MeDIP-seq tiles are plotted against 1,322 overlapping HEP bisulfite-PCR amplicons. As in **Figure 2b**, HEP methylation values for all CpGs that overlapped any given 100-bp MeDIP-seq tile were averaged, and all 100-bp windows were required to have at least two HEP scores (that is, either data from the top and bottom strand for a single CpG site, or at least two CpG sites) that differed by <50%. The purple-yellow (0–30) color bar on the right of each figure shows the total CpG coupling factor for each 100-bp tile. The same data stratified by CpG density is displayed in **Supplementary Fig. 4** online.

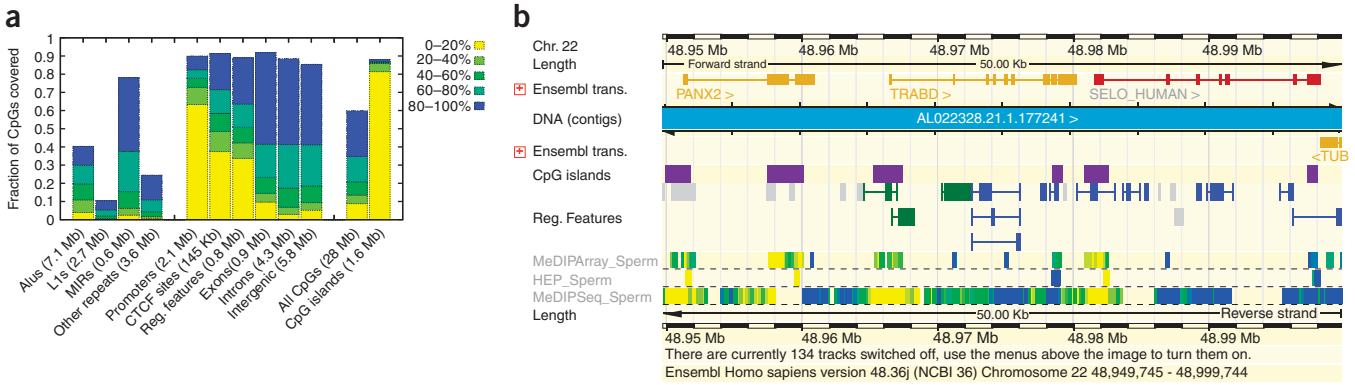


**Figure 5** Genomic coverage and web display of the MeDIP-seq data. (**a**) Genomic coverage of MeDIP-seq (measured as fraction of CpGs). Genomic features are from the Ensembl genome database (release 45). The first ten bars are mutually exclusive, that is, repeats are not included when considering subsequent features. Numbers in parentheses indicate the total number of CpGs within the human genome in that category. Promoters are defined as 2-kb regions centered on annotated transcriptional start sites, and Reg. features represent nonpromoter Regulatory Features in Ensembl. The colors represent the range of DNA methylation levels. (**b**) MeDIP-seq data integrated into Ensembl along with MeDIP-chip data of the same sperm DNA sample, and sperm bisulfite-PCR data from the Human Epigenome Project[16]. The yellow-green-blue color gradient represents 0–100% methylation.

regions, representing ~75 Mb of the genome. These regions are not included in the MeDIP-seq web display or any of our subsequent analyses. Such regions are also likely to be difficult to handle using other DNA methylation profiling strategies.

We then assessed whether sufficient read-depth had been obtained, as an insufficient number of reads would result in some parts of the genome being incorrectly called unmethylated. We therefore considered regions of the genome scored by MeDIP-chip as methylated and calculated the fraction of these regions that were covered by either the complete set of MeDIP-seq reads or by randomly chosen subsets of various sizes (**Fig. 3b**). This shows that our MeDIP-seq data set covers >97% of methylated regions. We consider this coverage to be good, supported by subsequent comparisons of our MeDIP-seq results with bisulfite-PCR sequencing data from the HEP (described below). A further increase in read depth would possibly yield slightly more accurate results, but any such improvement would be subject to rapidly diminishing returns.

### Batman analysis of MeDIP-seq

Two slight modifications to the MeDIP-chip version of Batman were required when handling MeDIP-seq data. First, as the read-out we use for MeDIP-seq is an absolute read density (which we sampled at arbitrary 50-bp intervals along the genome) rather than a $\log_2$-ratio, a different model was required. Based on visual inspection of the MeDIP-seq data (**Fig. 4a**), we used a polynomial model of order 2 instead of the linear model used for MeDIP-chip (**Figs. 1b** and **2a**). Second, the Gaussian error model was no longer appropriate (as the read density can never fall below zero), but a rectified Gaussian model could be used in a closely analogous manner. Following these two modifications, inference was performed as described above. We selected an output resolution of 100 bp as a good compromise between fast computation and high resolution. This resolution is likely to be sufficient for many applications, as the methylation status of CpG sites within <1,000 bp is significantly correlated (e.g., ~75% for ≤100 bp). Initially, this process gave results covering the entire human genome except for assembly gaps and regions containing no CpGs. However, as the short reads from the Illumina Genome Analyzer cannot be unambiguously mapped onto some repetitive parts of the genome, we expect Batman to undercall the methylation levels of the interior parts of large repeats. In recognition of this, we discarded all Batman results overlapping 500-bp genomic tiles with >50% repeat coverage. Comparison of the Batman-analyzed MeDIP-seq results with sperm data from the HEP revealed a strong overall correlation ($R^2 = 0.85$) (**Fig. 4b** and **Supplementary Fig. 5** online).

Our MeDIP-seq data provides high-resolution, quantitative coverage for ~90% of all CpG sites within CpG islands, promoters and other regulatory sequences, exons and introns, and ~60% of all CpGs in the human genome (**Fig. 5a**). This represents a ~20× improvement in coverage over existing methods. The use of paired-end sequencing allowed us to measure DNA methylation levels of some small repeat-element families. Several recent studies on human[36,37] and mouse[38,39] have reported epigenetic variability at repeat elements, and these could have phenotypic consequences. Furthermore, epigenetic silencing of repeat elements is thought to be critical for genomic integrity[40]. MeDIP-seq thus provides a means of analyzing such repeat elements in future DNA methylome studies. Consistent with previous observations from genome-wide studies[18,27], Batman analysis of the MeDIP-seq data reveals that promoters display an inverse correlation between CpG density and methylation (**Supplementary Fig. 6** online); CpG islands are predominantly unmethylated; a significant proportion of sites recognized by CTCF, a DNA binding protein involved in insulator activity, are unmethylated[41]; and most other regions of the genome are methylated. Our Batman-analyzed MeDIP-seq data (and the MeDIP-chip data) are freely accessible via the Ensembl Genome Browser (**Fig. 5b** and http://www.ensembl.org/), representing a useful resource for the scientific community.

## DISCUSSION

Unraveling the complexities of the epigenome is a very important objective, and recent years have seen the development of several strategies for genome-wide analysis of epigenetic marks, including DNA methylation. One of the principal challenges now is to develop more powerful analytical tools to interpret the vast amounts of data that continue to be generated.

Here, we have reported the development and validation of Batman, a cross-platform algorithm for the quantitative analysis of MeDIP data generated using either arrays (MeDIP-chip) or next-generation sequencing technologies (MeDIP-seq, representing a high-resolution DNA methylome of a mammalian genome). Batman, combined with MeDIP-chip or MeDIP-seq, estimates absolute methylation levels over

a wide range of CpG densities. This is a very useful property for DNA methylome analyses, as it will allow more effective profiling, including CpG-poor regions that have traditionally been overlooked in most DNA methylome studies. Furthermore, estimation of absolute DNA methylation levels will facilitate cross-platform comparisons.

Although several strategies now exist for DNA methylation profiling, there are, to the best of our knowledge, only two others that compare with combining Batman with MeDIP-chip or MeDIP-seq in terms of genomic coverage and quantitative performance. The first, comprehensive high-throughput arrays for relative methylation (CHARM), combines a tiling-array design strategy with statistical procedures that average information from neighboring genomic locations[42]. The authors applied CHARM to an assay involving digesting DNA with McrBC, which restricts methylated DNA (recognition sequence $R^mC(N)_{55–103}R^mC$). The enzyme is used on size-selected (1.5–4.0 kb) DNA to fractionate unmethylated DNA after digestion, which is co-hybridized on arrays with DNA similarly processed but not cut with the enzyme. Although CHARM correlates well with bisulfite conversion–based data ($R^2 = 0.76$)[42], it is not a 'stand-alone' algorithm but rather a strategy that requires the use of a particular array design. Moreover, it is unclear whether it can be adapted to next-generation sequencing technologies. It does not estimate absolute DNA methylation levels and suffers to some degree in the ability to discriminate highly methylated from highly unmethylated CpG islands[42]. Interestingly, the authors also tested MeDIP-chip and concluded that it cannot be used to analyze CpG-poor regions[42]. Our results show that combining MeDIP with Batman can provide absolute DNA methylation levels across a range of CpG densities (including CpG-poor regions) from arrays or next-generation sequencing.

Another recently reported approach, BS-seq, was used to delineate a DNA methylome for the ~120-Mb *Arabidopsis* genome[19]. BS-seq has the ability to provide single bp–resolution DNA methylation profiles, which is indeed a very useful property. However, at current sequencing costs, such an approach is still prohibitively expensive to analyze larger genomes such as the human, which is ~25× bigger than the *Arabidopsis* genome. Based on our results, we estimate that ~40 million paired-end reads (less than a single run of an Illumina Genome Analyzer) are sufficient to generate a high-quality mammalian methylome, whereas ~3.8 Gb of sequence (which would equate to >40 million paired-end reads) was required to generate a single-base pair resolution (~20× coverage) methylome for the ~120-Mb *Arabidopsis* genome using BS-seq[19]. Also, even though single-CpG resolution is desirable, the fact that the methylation status of CpG sites within <1,000 bp is significantly correlated[18] (e.g., ~75% for ≤100 bp) means that the ~100 bp resolution is suitable for many applications.

Although Batman in its present form performs well, we see opportunities for future development of MeDIP–post-processing platforms, especially with regard to the use of sequencing technologies. In particular, when analyzing paired-end MeDIP-seq data, it should be possible to take advantage of the exact mapping positions of each read, rather than summarizing the data as a set of read-depth samples, thereby improving the resolution. Also, it would be interesting to apply Batman to the analysis of *Arabidopsis* MeDIP data. Although both CpG and non-CpG methylation is found in *Arabidopsis*[22,26], gene bodies contain predominantly the former, and therefore, it should be possible to use Batman for the analysis of genic regions.

In the near future, the integration of (epi)genomic and functional approaches is going to be crucial for elucidating the biological role of DNA methylation. The need for such an integrated approach is also evident from the recently announced National Institutes of Health Epigenome Roadmap Initiative calling for mapping of reference DNA methylation profiles on an unprecedented scale (http://nihroadmap.nih.gov/epigenomics/). Combining the Batman algorithm with MeDIP-chip or MeDIP-seq should provide cost-effective strategies for quantitative, high-resolution DNA methylome analysis and will contribute toward elucidating the role of the epigenome in health and disease.

## METHODS

**Sperm samples.** Human mature spermatozoa were obtained as part of the major histocompatibility complex (MHC) Haplotype Project (http://www.sanger.ac.uk/HGP/Chr6/MHC/) under Cambridge Local Research Ethics Committee approvals LREC-03/094 and LREC-04/Q0108/46.

**Methylated DNA immunoprecipitation.** MeDIP was performed using a previously published protocol[20], but we also included a ligation-mediated PCR (LM-PCR) step[43] to amplify the material (the LM-PCR step was not performed for MeDIP-seq). Hybridizations of pre- and post-LM-PCR samples on custom tile-path arrays (2 kb resolution) for the human MHC showed that the LM-PCR did not introduce significant bias (**Supplementary Fig. 1**). A detailed protocol is provided in the **Supplementary Methods** online.

**Custom oligonucleotide array design and pre-Batman processing.** Our microarray consists of 382,178 50-bp probes. Although we aimed to target all annotated TSSs and nonpromoter CpG islands (CGIs), we were unable to design suitable unique probes for 18% of the TSSs and 28% of nonpromoter CGIs. The array also contained 50-mer probes tiled at ~100-bp density across the entire human MHC, and promoters and nonpromoter CpG islands on the X- and Y-chromosomes. Analyses of these regions will be presented elsewhere. The array was originally designed using the NCBI build 35 version of the human genome assembly, but then mapped to NCBI build 36 using Exonerate[44]. To be mapped, probes were required to align full length and without gaps or mismatches. Probes that aligned more than once to the NCBI36 sequence were removed from the analysis. Tiled regions were defined by clustering uniquely mapped probes within 200 bp of one another. Singleton probes were discarded. The tiled regions were then divided into 500-bp ROIs. After hybridization (performed by Nimblegen using their standard conditions), arrays were LOESS-normalized using custom R-scripts before Batman anlaysis of the resulting $\log_2$ ratios.

**Illumina Genome Analyzer sequencing.** Based on the manufacturer's recommended protocol, we nebulized 10 µg sperm DNA (SP3 in **Supplementary Table 1**) with compressed nitrogen for 6 min at 32 p.s.i., giving fragments of <800 bp. We then end repaired, phosphorylated and A-tailed the fragmented DNA and ligated Illumina paired end adapters to fragments. Of this, we used ~1 µg of adaptor-ligated DNA for subsequent MeDIP enrichment (performed as described above but without the LM-PCR step). Because the quantity of DNA obtained after MeDIP was low (~30 ng), we deviated from the standard Illumina protocol and amplified the sample using Illumina paired-end PCR primers before gel electrophoresis and size-selecting libraries. We excised bands from the gel to produce libraries with insert sizes of 85–160 bp and quantified these libraries using an Agilent Bioanalyzer 2100. We prepared paired-end flowcells with 3.2 pM DNA (using two-primer chemistry) using the manufacturer's recommended protocol and sequenced for 36 cycles on an Illumina genome analyzer fitted with a paired-end module. The reads were mapped onto the human genome reference sequence using the high-performance alignment software 'maq' (http://maq.sf.net/) before Batman analysis.

**Statistics.** All correlation coefficients were computed using Pearson's product-moment formula. All credible intervals were estimated by bootstrapping. All other statistical procedures related to Batman are described in the main text.

**Accession codes.** ArrayExpress: MeDIP-chip data have been deposited with number E-TABM-445; MeDIP-seq data have been deposited with number E-TABM-482.

*Note: Supplementary information is available on the Nature Biotechnology website.*

**ACKNOWLEDGMENTS**

T.A.D., E.M.T., L.B., K.L.H., D.K.J., M.M.M., H.L., T.J.P.H., S.B., D.J.T., R.D. were supported by the Wellcome Trust. V.K.R. was supported by the Barts and The London Charitable Trust, and a C.J. Martin Fellowship from the National Health and Medical Research Council, Australia. S.G., N.J. and M.H. were supported by an EU grant (High-throughput Epigenetic Regulatory Organization in Chromatin (HEROIC), LSHG-CT-2005-018883) under the 6th Framework Program to S.B. (M.H.) and E.B. (S.G., N.J.). N.P.T., J.C.M. and S.T. were supported by grant C14303/A8646 from Cancer Research UK.

**AUTHOR CONTRIBUTIONS**

T.A.D. co-conceived the study, wrote the Batman algorithm, co-analyzed data and co-wrote the paper; V.K.R. co-conceived the study, performed the bulk of the experimental work, co-analyzed data, co-wrote the paper and provided overall project management; D.J.T. performed the Illumina Genome Analyzer sequencing; H.L. performed the maq analysis; P.F., E.K., S.G., N.J. and J.H. performed some data analysis and designed the Ensembl web display for the data reported here; E.M.T., L.B. and M.H. performed experimental work; K.L.H. and D.K.J. assisted with array design; N.P.T. and J.C.M. performed preliminary array analysis; M.M.M. supplied materials; E.B., T.J.P.H., R.D. and S.T. provided intellectual input; S.B. co-conceived the study, co-wrote the paper and provided overall project management. T.A.D. and V.K.R. contributed equally to this work.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/



1. Bernstein, B.E., Meissner, A. & Lander, E.S. The mammalian epigenome. *Cell* **128**, 669–681 (2007).
2. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
3. Beck, S. & Rakyan, V.K. The methylome: approaches for global DNA methylation profiling. *Trends Genet.* **24**, 231–237 (2008).
4. Tompa, R. *et al.* Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. *Curr. Biol.* **12**, 65–68 (2002).
5. Lippman, Z. *et al.* Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**, 471–476 (2004).
6. Khulan, B. *et al.* Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res.* **16**, 1046–1055 (2006).
7. Schumacher, A. *et al.* Microarray-based DNA methylation profiling: Technology and applications. *Nucleic Acids Res.* **34**, 528–542 (2006).
8. Ordway, J.M. *et al.* Comprehensive DNA methylation profiling in a human cancer genome identifies novel epigenetic targets. *Carcinogenesis* **27**, 2409–2423 (2006).
9. Ching, T.T. *et al.* Epigenome analyses using BAC microarrays identify evolutionary conservation of tissue-specific methylation of SHANK3. *Nat. Genet.* **37**, 645–651 (2005).
10. Shen, L. *et al.* Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet.* **3**, 2023–2036 (2007).
11. Rollins, R.A. *et al.* Large-scale structure of genomic methylation patterns. *Genome Res.* **16**, 157–163 (2006).
12. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA* **89**, 1827–1831 (1992).
13. Gitan, R.S. *et al.* Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. *Genome Res.* **12**, 158–164 (2002).
14. Adorjan, P. *et al.* Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res.* **30**, e21 (2002).
15. Bibikova, M. *et al.* High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.* **16**, 383–393 (2006).
16. Reinders, J. *et al.* Genome-wide, high-resolution DNA methylation profiling using bisulfite-mediated cytosine conversion. *Genome Res.* **18**, 469–476 (2008).
17. Rakyan, V.K. *et al.* DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol.* **2**, e405 (2004).
18. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* **38**, 1378–1385 (2006).
19. Cokus, S.J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
20. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37**, 853–862 (2005).
21. Keshet, I. *et al.* Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat. Genet.* **38**, 149–153 (2006).
22. Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* **126**, 1189–1201 (2006).
23. Gebhard, C. *et al.* Genome-wide profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia. *Cancer Res.* **66**, 6118–6128 (2006).
24. Rauch, T., Li, H., Wu, X. & Pfeifer, G.P. MIRA-assisted microarray analysis, a new technology for the determination of DNA methylation patterns, identifies frequent methylation of homeodomain-containing genes in lung cancer cells. *Cancer Res.* **66**, 7939–7947 (2006).
25. Illingworth, R. *et al.* A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.* **6**, e22 (2008).
26. Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. & Henikoff, S. Genome-wide analysis of *Arabidopsis* thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.* **39**, 61–69 (2006).
27. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–466 (2007).
28. Yasui, D.H. *et al.* Integrated epigenomic analyses of neuronal MeCP2 reveal a role for long-range interaction with active genes. *Proc. Natl. Acad. Sci. USA* **104**, 19416–19421 (2007).
29. Jacinto, F.V., Ballestar, E., Ropero, S. & Esteller, M. Discovery of epigenetically silenced genes by methylated DNA immunoprecipitation in colon cancer cells. *Cancer Res.* **67**, 11481–11486 (2007).
30. Cheng, A.S. *et al.* Epithelial progeny of estrogen-exposed breast progenitor cells display a cancer-like methylome. *Cancer Res.* **68**, 1786–1796 (2008).
31. Fouse, S.D. *et al.* Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. *Cell Stem Cell* **2**, 160–169 (2008).
32. Flicek, P. *et al.* Ensembl 2008. *Nucleic Acids Res.* **36**, D707–D714 (2008).
33. Qi, Y. *et al.* High-resolution computational models of genome binding events. *Nat. Biotechnol.* **24**, 963–970 (2006).
34. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
35. Mikkelsen, T.S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
36. Sandovici, I. *et al.* Interindividual variability and parent of origin DNA methylation differences at specific human Alu elements. *Hum. Mol. Genet.* **14**, 2135–2143 (2005).
37. Flanagan, J.M. *et al.* Intra- and interindividual epigenetic variation in human germ cells. *Am. J. Hum. Genet.* **79**, 67–84 (2006).
38. Morgan, H.D., Sutherland, H.G., Martin, D.I. & Whitelaw, E. Epigenetic inheritance at the agouti locus in the mouse. *Nat. Genet.* **23**, 314–318 (1999).
39. Rakyan, V.K. *et al.* Transgenerational inheritance of epigenetic states at the murine *Axin*$^{Fu}$ allele occurs after maternal and paternal transmission. *Proc. Natl. Acad. Sci. USA* **100**, 2538–2543 (2003).
40. Bestor, T.H. The host defence function of genomic methylation patterns. *Novartis Found. Symp.* **214**, 187–195 (1999).
41. Irizarry, R.A. *et al.* Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.* **18**, 780–790 (2008).
42. Mukhopadhyay, R. *et al.* The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res.* **14**, 1594–1602 (2004).
43. Oberley, M.J. & Farnham, P.J. Probing chromatin immunoprecipitates with CpG-island microarrays to identify genomic sites occupied by DNA-binding proteins. *Methods Enzymol.* **371**, 577–596 (2003).
44. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31–34 (2005).