Genome Biology

METHOD Open Access

epiG: statistical inference and profiling of DNA methylation from whole-genome bisulfite sequencing data

Martin Vincent[1], Kamilla Mundbjerg[2], Jakob Skou Pedersen[3], Gangning Liang[4], Peter A. Jones[5], Torben Falck Ørntoft[3], Karina Dalsgaard Sørensen[3] and Carsten Wiuf[1]*

Abstract

The study of epigenetic heterogeneity at the level of individual cells and in whole populations is the key to understanding cellular differentiation, organismal development, and the evolution of cancer. We develop a statistical method, epiG, to infer and differentiate between different epi-allelic haplotypes, annotated with CpG methylation status and DNA polymorphisms, from whole-genome bisulfite sequencing data, and nucleosome occupancy from NOMe-seq data. We demonstrate the capabilities of the method by inferring allele-specific methylation and nucleosome occupancy in cell lines, and colon and tumor samples, and by benchmarking the method against independent experimental data.

Keywords: CpG methylation, GpC methylation, NOMe-seq, Epi-allelic haplotype, Epi-allele, Epigenetic state

Background

The epigenetic state of a particular genomic region is the combined configuration of the epigenetic modifications in the region, including DNA (CpG) methylation, nucleosome positioning, and histone modifications. Such states are important regulators of many biological processes in health and disease, and at many different biological levels, for example gene regulation, cellular differentiation, and organismal development [1–3]. To understand the biological consequences of CpG methylation and nucleosome positioning, it is, therefore, important to be able to extract reliable epigenetic information from genome-wide sequencing data using statistical methods.

A particularly challenging problem for statistical epigenetic inference is the possible heterogeneous nature of the epigenetic states [4, 5]. Heterogeneity occurs at the population level, for example, in relation to cellular development and differentiation [6], and in the evolution of cancer [7, 8]. It also occurs at the level of individual cells in the form of allele-specific methylation (ASM) [9–13]. ASM might stretch over many kilobases, and is associated with genomic imprinting [9, 14] and generally with genetic variants underlying phenotypic differences and complex diseases [13, 15]. Statistical methods that are able to call regions with ASM, and to infer and differentiate between different stable common epi-alleles in a biological sample, are therefore desirable. We have developed a stand-alone package, epiG, for this purpose.

The method epiG takes as input whole-genome bisulfite sequencing (WGBS) data [4, 16, 17]. It is a likelihood-based method that clusters reads into epi-allelic haplotypes based on sequence similarity, while taking into account experimental errors and biological noise. It outputs the dominating epi-allelic haplotypes of a genomic region of interest, annotated with an inferred CpG methylation and single nucleotide polymorphism (SNP)/somatic nucleotide variant (SNV) profile for each epi-allelic haplotype. The method makes use of prior information on base-calling quality, bisulfite conversion efficiency, and a reference SNP database. Recently, a protocol (NOMe-seq) for simultaneously assessing nucleosome positioning and CpG methylation genome-wide has been developed [18]. epiG has a NOMe-seq track that additionally performs inference on nucleosome occupancy.

*Correspondence: wiuf@math.ku.dk
[1] Department of Mathematical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark
Full list of author information is available at the end of the article

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

State-of-the-art methods for WGBS methylation analysis infer CpG methylation position-wise, reporting a degree of methylation for each CpG site, e.g. [19–21]. Consequently, these methods do not allow for profiling of epigenetic haplotypes. Several methods exist for detecting differentially methylated regions, in particular with the aim of identifying ASM genes [9, 22, 23]. These methods, too, do not profile epigenetic haplotypes directly.

We demonstrate the applicability of epiG by inferring ASM and dominating epi-alleles in tumor and non-tumor WGBS samples. Also, we infer nucleosome positioning in two NOMe-seq samples from cell lines. By benchmarking with independent experimental data, we demonstrate that epiG calls CpG methylation states and genotypes correctly.

Results and discussion

It is not straightforward to extract precise DNA methylation profiles from raw WGBS data due to the complexity of the bisulfite conversion process [17].

Figure 1 illustrates the conversion of genomic DNA into sequenced reads via bisulfite conversion, PCR amplification, and genomic mapping. For a single DNA fragment, the strand and the methylation status (top of Fig. 1) can be recovered from observed error-free reads (bottom of Fig. 1), as shown in Table 1. This is possible because an unmethylated C in a C-G site, with C on the forward strand, is converted into U by bisulfite conversion. After PCR amplification and genomic mapping, this results in an observed T-C site, that is, some reads with T and some reads with C in the matching position. Similarly, a G-C site with an unmethylated C becomes a G-A site after conversion, amplification, and mapping. Thus, the reads with T (G) unambiguously come from the forward strand, whereas the reads with C (A) unambiguously come from the reverse strand (see Table 1).

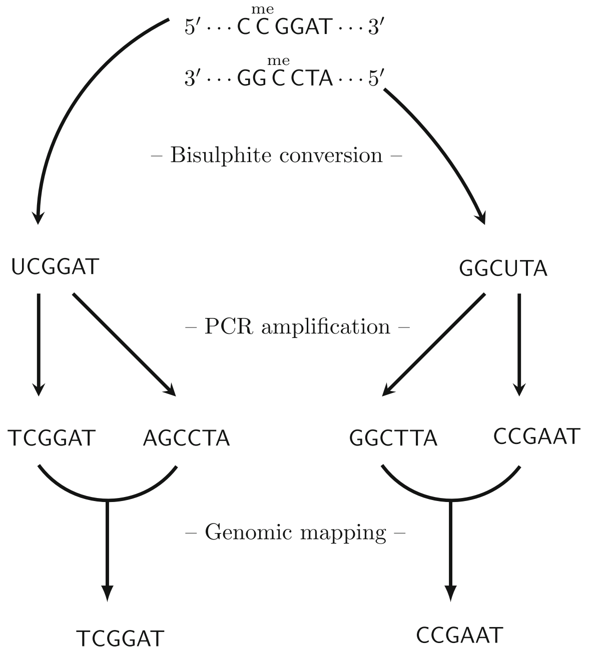
$5' \cdots \mathsf{C}\overset{\mathrm{me}}{\mathsf{C}}\mathsf{GGAT} \cdots 3'$
$3' \cdots \mathsf{GG}\overset{\mathrm{me}}{\mathsf{C}}\mathsf{CTA} \cdots 5'$
– Bisulphite conversion –
UCGGAT
GGCUTA
– PCR amplification –
TCGGAT
AGCCTA
GGCTTA
CCGAAT
– Genomic mapping –
TCGGAT
CCGAAT

Fig. 1 Bisulfite conversion. The DNA fragment is shown at the top as two complementary sequences. There are four possible base pairings of nucleotides and with C possibly methylated, this yields six different pairs in total, all of which are shown at the top of the figure. Bisulfite conversion converts the forward strand into the left sequence and the reverse strand into the right sequence, by changing unmethylated C into U and leaving methylated C as C. Subsequent PCR amplification results in four pairwise complementary types, where U is changed to T, which causes a change to A on the complementary strand. Alignment of the reads to a reference genome results in two types, which are both read in the forward direction. Comparing the two reads yields six possible pairs of nucleotides. If there are no sequencing and PCR errors, then these six (out of 16) are the only possible pairs of nucleotides that might be observed in reads from a single DNA fragment. Each pair in the observed reads, for example A-A, corresponds to a unique pair in the DNA fragment, here A-T

Our method is based on the above insight. The primary goal of epiG is to infer dominant epi-allelic haplotypes (or epi-alleles) by grouping similar reads together to form chains of consecutive reads, called haplotype chains. These will vary in length depending on read length, sequencing depth, and data quality. In addition, the method provides information about genomic and epigenomic variation in the form of

- CpG methylation status
- Nucleosome occupancy (with NOMe-seq data)
- SNPs and SNVs

We use a read mapper (BSMAP [24]) to map reads to the reference genome (hg19). The performance of epiG depends on a number of user-adjustable parameters. These parameters control how noise is treated and how reads are clustered together to form haplotype chains. epiG starts by assigning each read to its own haplotype chain and then, iteratively, it improves the haplotype chains by re-assigning reads to chains one at a time, using an optimization procedure and prior genetic information. Each haplotype chain is characterized by an inferred DNA sequence annotated with CpG methylation and SNPs/SNVs. Noise and low frequency haplotype chains might further be filtered out by removing haplotype chains formed by few reads.

Figure 2 shows an example of three distinct haplotype chains inferred from a region on chromosome 20 from a colon tumor sample (see 'Methods'). One chain is completely unmethylated, while another chain is fully methylated. The last chain is partly methylated. Compared to a consensus method that outputs the average degree of methylation across all reads (irrespective of their

Table 1 Observed reads and the true epigenetic states

A	Read 1	T	C	G	G	A	T
	Read 2	C	C	G	A	A	T
	$5' \rightarrow 3'$	C	$\overset{me}{C}$	G	G	A	T
	$3' \rightarrow 5'$	G	G	$\overset{me}{C}$	C	T	A
B	Read		T	C	G	A	
	$5' \rightarrow 3'$		C/T	$\overset{me}{C}$/C	G	G/A	
	$3' \rightarrow 5'$		G/A	G	$\overset{me}{C}$/C	C/T	

The conditions for epigenetic inference are optimal if there are no sequencing and PCR errors. The table shows inference on the true epigenetic states from reads from a single DNA fragment

A. If error-free reads are available from both strands, one of six possible pairs of nucleotides might be observed at each site by comparing reads, as shown in the two top rows of the table (see also Fig. 1). In each case, the epigenetic state and the strand direction can be inferred unambiguously. For example, if G and A are observed, then the 5′ → 3′ strand has G, while the 3′ → 5′ strand has C unmethylated

B. If reads from only one strand are observed, then inference is in general inconclusive. For each of the four possible nucleotides that might be observed, the true epigenetic state and the strand direction cannot be inferred unambiguously. For example, if T is observed in a read, then the true epigenetic state of the fragment might be C-G or T-A, depending on whether the observed T comes from the 5′ → 3′ strand or the 3′ → 5′ strand. For correct epigenetic inference, it is, therefore, important that reads originating from both strands are observed

epi-allelic haplotype), our method provides information about linked methylated sites on the same epi-allele as well as epi-allelic diversity.

In the following, we applied epiG to publicly available WGBS data sets and to two NOMe-seq data sets from our own lab. See 'Methods' and Table 2 for naming of and a description of the data sets. For all data sets, epiG default parameters were used. See 'Methods,' Table 3, and Additional file 1 for details.

Allele-specific methylation

To illustrate the performance of epiG to detect ASM, we ran epiG on the four WGBS data sets (WA9, colon normal A, colon normal B, and colon tumor B; see Table 2 and 'Methods') near genomic regions that are known to be allele-specific methylated. None of these data sets were used for this or similar purposes in the original publications [15, 25, 26]. We focused particularly on two well-described ASM regions, namely the GNAS locus on chromosome 20 and the *H19* non-coding gene on chromosome 7, both of which are known imprinted regions [9].

In the ASM regions near the GNAS locus, we find either fully methylated or fully unmethylated epi-alleles, strongly suggesting that there are two dominating epi-allelic haplotypes (see Fig. 3).

The paternal copy of the *H19* gene is usually methylated and silent while the maternal copy is hypomethylated or unmethylated [27]. We reach the same conclusion from three of the four data sets, except for the WA9 sample, which shows consistent methylation and a single epi-allelic haplotype throughout the region (see Fig. 4). This observation is in accordance with the previously described timing of *H19* gene activation to the peri-implantation stage of human embryo development [28]; thus, this illustrates the applicability of epiG

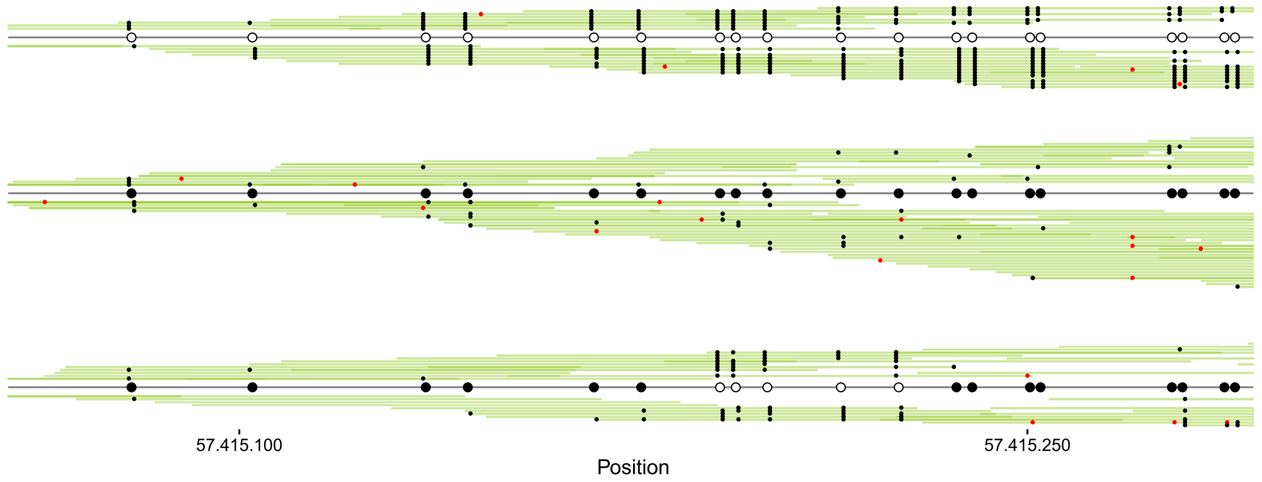

Fig. 2 Example of haplotype chains. Illustration of the inferred haplotype chains in a 250-bp region (position 57,415,050–57,415,300 on chromosome 20) in the GNAS locus from colon tumor B (see 'Methods'). One major and two minor chains are inferred. The major chain is fully methylated while one of the minor chains is completely unmethylated. Each line is one read (*green*), with paired reads on the same line. Converted CpGs are marked in *black*. Bases not matching the reference genome are marked in *red* and are indicative of sequencing errors. The reads in the top part of a chain have an inferred forward direction, while the reads in the bottom part have an inferred reverse direction. After assigning reads to haplotypes chains, the epigenotype of a chain is inferred from all reads in the chain. The epigenotype is shown in the middle of each chain. An *open circle* is an unmethylated CpG site and a *closed circle* is an methylated CpG site. The chains continue beyond the right end point, but are cut for illustrative purposes

Table 2 WGBS and NOMe-seq data sets

Sample name	Description	Method	Layout	Mapped reads	Read/ paired length	GEO	Ref
WA9	Human embryonic stem cell line	WGBS	Paired	1360M	100/186	GSM1521762	[25]
Colon normal A	Sigmoid colon tissue	WGBS	Single	1703M	101/ –	GSM983645	[26]
Colon normal B	Primary colon adjacent to tumor tissue	WGBS	Paired	789M	100/191	GSM1204466	[15]
Colon tumor B	Colon primary adenocarcinoma	WGBS	Paired	807M	100/191	GSM1204465	[15]
PrEC	Prostate epithelial cells	NOMe-seq	Paired	293M	73/128	GSE94361	–
LNCaP	Prostate adenocarcinoma cell line	NOMe-seq	Paired	329M	100/146	GSE94361	–

In the text, we refer to the data sets by their sample names. The data sets colon normal A, PrEC, and LNCaP have lower data quality than the other data sets. The read/paired length is approximate. See Additional file 1: Figure S28 for the position-wise read depth for all samples
WGBS whole-genome bisulfite sequencing

in the study of ASM during embryonic development and cancer.

The region ∼2–4 kb upstream of *H19* is the imprinted control region *H19*-ICR, which is required to establish parent-of-origin imprinting of *H19* [29]. Essentially, we find a single epi-allelic haplotype in this region for each sample. It is methylated in WA9, but has a mixed methylation pattern in the three colon samples with chunks of unmethylated CpGs. If methylated, as in WA9, the upstream *IGF2* gene is activated and *H19* expression is silenced [29].

Figure 5 shows summary statistics for different genomic regions in the four different samples. For allele-specific methylated regions (AMRs), we generally find one or two distinct haplotype chains, while for other regions (promotor, exon, and randomly selected regions), we generally find one haplotype chain, indicating that both DNA copies of a region are methylated in the same way. Additional file 1: Figures S1–S18 show that the number and extent/length of haplotype chains in 18 AMRs [9] in the four samples generally are in high agreement. The position-wise variance of the number of haplotype chains is as low as 0.22 across all positions and AMRs. In contrast, randomly assigning one or two chains to each region would yield a variance of 1.

Table 3 Default parameter values for WGBS and NOMe-seq data

Parameter	WGBS single	WGBS paired	NOMe-seq
α	0.95	0.95	0.95
β	0.05	0.05	0.05
K_0	40	50	40
K_1	1	2	0
K_2	0	0	0
K_3	0	0	2
q	0.9999	0.9999	0.9999

See also Additional file 1: Table S1. K_0 and K_1 are put to different values for single and paired reads, because paired reads are generally longer than single reads (see Table 2). The default values for α and β are conservative estimates of the failed bisulfite conversion rate and the inappropriate bisulfite conversion rate, respectively [32, 36]
WGBS whole-genome bisulfite sequencing

Nucleosome occupancy

The NOMe-seq protocol was developed for simultaneously assessing nucleosome positioning and DNA methylation genome-wide [18]. This protocol uses the ability of the M.CviPI methyltransferase enzyme to methylate the cytosine in a GpC dinucleotide in accessible chromatin. Thus, M.CviPI treatment of intact cell nuclei can demarcate the nucleosome positioning of cells. Following bisulfite conversion and whole-genome sequencing, it is, therefore, possible to infer nucleosome positioning using artificial GpC-methylation as well as endogenous DNA methylation at CpG dinucleotides on the same molecule [18].

Only isolated GpC sites (DGCH sites, or sites not bordering C to the left and G to the right; see ‘Notation and definitions’ in ‘Methods’) can be used for nucleosome inference, as otherwise the signal might be confused with endogenous CpG methylation [18].

We applied epiG to NOMe-seq data from the PrEC and LNCaP samples from our own lab (see Table 2 and ‘Methods’). CTCF binding sites have been shown to have characteristic DNA methylation and nucleosome occupancy patterns. Figures 6 and 7 show an example of nucleosome inference near a CTCF site where different epi-allelic haplotypes can be distinguished. In some of these, the CTCF site is inaccessible due to protein binding. As nucleosome positioning is dynamic and varies across cells, inference on nucleosome length and position is uncertain except in regions where stable nucleosomes are expected [30, 31]. When averaging over many CTCF sites, we see a clear periodic pattern, like what is reported in [18] (see Additional file 1: Figure S19).

Noise reduction

We are mainly interested in the dominant epi-alleles of a sample. Therefore, we filter out minor epi-alleles and reduce noise by excluding haplotype chains with few reads

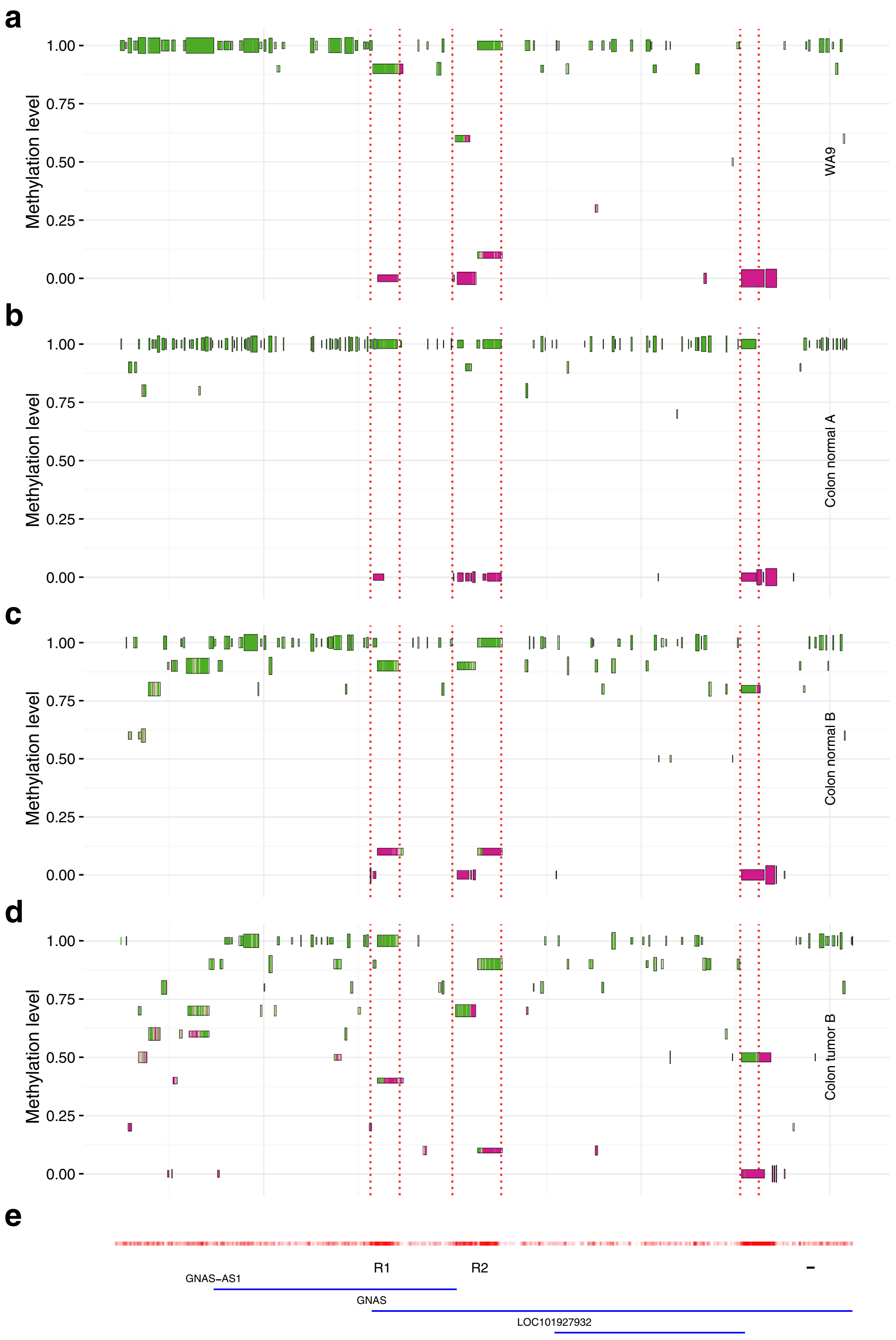

Fig. 3 (See legend on next page.)

Fig. 3 (See figure on previous page.)
Epi-allelic profiling near the GNAS locus. Epi-allelic haplotypes of four samples in a 100-kb region near the GNAS locus with three known regions of ASM (R1, R2, and R3; *vertical dashed lines* [9]). **a** WA9. **b** Colon normal A. **c** Colon normal B. **d** Colon tumor B. **e** CpG site density map (*red*) and known genes (*blue*). In each of **a**–**d**, the *horizontal bars* are inferred epi-allelic haplotypes and the bar thickness represents the average number of reads per position. The degree of methylation is shown in two ways. Firstly, the average methylation level of three consecutive CpGs in a chain is shown in color, where *green* is for fully methylated CpG sites, *purple* for fully unmethylated, and *white* for neither. The haplotype overall methylation level is the average methylation level of all CpGs in the whole chain. In all three regions, we clearly see ASM. The stem cell sample (WA9) shows only one (unmethylated) epi-allele in R3. To the immediate right of R3, all four samples show consistently one unmethylated epi-allele. WA9 and colon normal A are fully methylated outside the ASM regions, whereas a general loss of methylation can be seen in colon tumor B and partly in colon normal B, which is tissue adjacent to the tumor. The density of CpGs is an important determinant of the length of the haplotypes. When the density is low, haplotypes become shorter and more fragmented. *ASM* allele-specific methylation

or low read coverage per position. This is a post-inference step that affects the number of reported haplotype chains and how many haplotypes are filtered out. Thus, the number and the form of the dominating haplotype chains (as those in Figs. 3 and 4) are unaffected by this filtering step.

The depth fraction of a chain is defined as

$$\frac{\text{Total base pairs in chain}}{\text{Total base pairs overlapping the chain}},$$

where the total base pairs in chain is the sum of the lengths of all reads in the chain and the total base pairs overlapping the chain is the sum of the lengths of all reads in the chains overlapping the particular chain.

We use the depth fraction to distinguish noise from signal and do so by only keeping chains for which the inequality

$$a \leq \text{chain length} + b \times \text{depth fraction}$$

holds, for some values of a and b. In general, this means that we keep long chains and chains with a high depth fraction. In all presented analyses, we used $a = 10 \times 10^3$ and $b = 28.5 \times 10^3$ for paired layout WGBS, and $a = 5 \times 10^3$ and $b = 20 \times 10^3$ for single layout WGBS, based on empirical observations (see Fig. 8, Fig. 9, and Additional file 1: Figures S20, S21, and S22). With the chosen values, the main determinant of whether a chain is kept or not is the depth fraction (see the figures). As seen in the figures, a common feature is a trail of haplotype chains with very low depth fraction, corresponding to haplotype chains consisting of very few reads. There is not a gold standard for setting a and b.

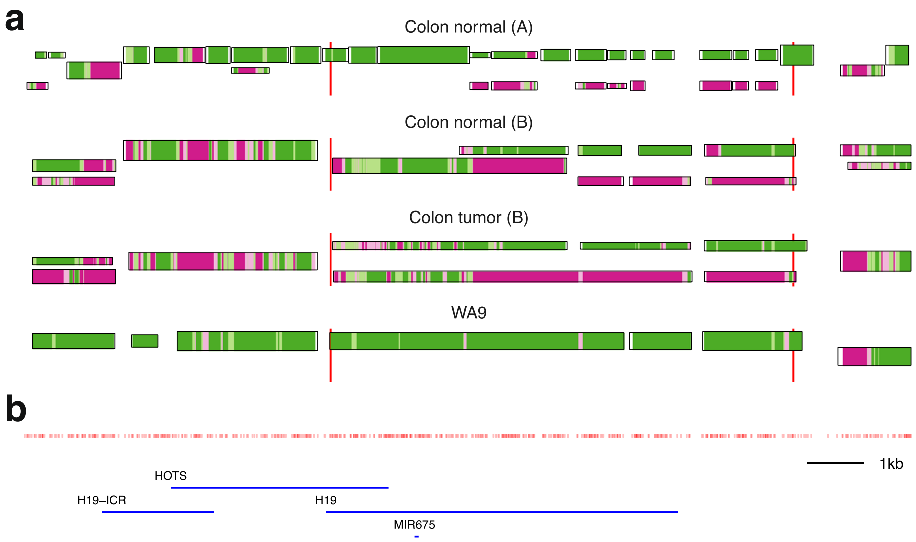

Fig. 4 Epi-allelic profiling near the *H19* gene. Epi-allelic haplotypes of the four samples in a 20-kb region around the *H19* gene with one known ASM region (*vertical red lines* [9]). **a** *Horizontal bars* are inferred haplotype chains and the bar thickness represents the average number of reads per position. The average methylation level of three consecutive CpGs in a chain is shown in color, where *green* is for fully methylated CpG sites, *purple* for fully unmethylated, and *white* for neither (the vertical position of a chain represents the average degree of methylation). WA9 has only one (methylated) epi-allele in the entire ASM region, while the other three show one methylated and one unmethylated epi-allele. The unmethylated region does not span the entire ASM region, which is compiled from 22 methylomes from cell lines and tissue samples (not including the colon) [9]. **b** CpG site density map (*red*) and known genes (*blue*). The region ∼2–4 kb upstream of *H19* is the *H19* imprinting control region (*H19*-ICR). *ASM* allele-specific methylation

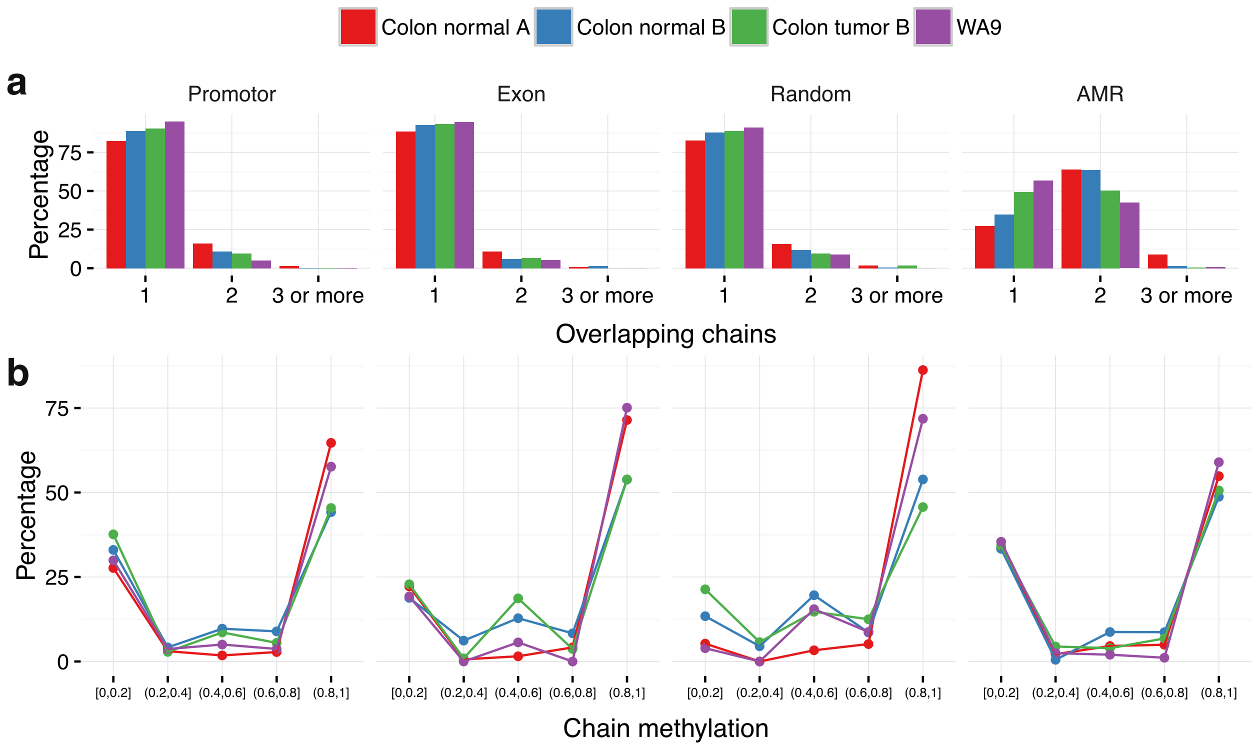

Fig. 5 Summary statistics of the estimated haplotype chains. Collected statistics of the estimated haplotype chains over four different genomic regions: promoters, exons, randomly selected regions, and known AMRs. For each type, except AMR, 100 regions where randomly chosen. For AMR, 18 known AMRs were used. See Additional file 1: Table S2 and [9]. For each region, haplotypes were estimated using the default configuration and the noise was reduced as described in 'Noise reduction' in 'Results and discussion.' **a** The percentage of regions with only one chain, two overlapping chains, and three or more overlapping chains for each region. The majority of regions, except AMRs, were covered by only one chain. **b** The distribution of the average methylation level over the haplotype chains. Chains tend to be either fully methylated or completely unmethylated. *AMR* allele-specific methylated region

Validation and benchmarking

In addition to the qualitative validation presented above, we carried out quantitative validations. Specifically, we measured the ability of epiG to infer position-wise methylation and SNP genotypes by comparing epiG predictions to measurements obtained from Infinium BeadChip and SNP6.0 arrays, with details given below. The results were also compared to results obtained using Bis-SNP [20], a state-of-the-art method for inferring position-wise methylation and genotypes from WGBS and NOMe-seq data. Further, we investigated the robustness of epiG by varying the default parameters.

Methylation validation

To evaluate the inferred methylation states, we used ∼18k isolated CpG sites (HCGD sites; see 'Notation and definitions' in 'Methods') with high and low methylation levels on the Infinium BeadChip from the LNCaP sample. We used isolated sites to distinguish endogenous methylation from methylation caused by the GpC methyltransferase [18].

We first preselected the 5% of CpG sites with the highest β value (high) and the 5% with the lowest β value (low) on the Infinium BeadChip from the LNCaP sample (see 'Methods' for details). Subsequently, we removed all non-isolated CpG sites and ended up with approximately 18k isolated CpG sites for validation of which ∼8k have low β values and ∼10k have high β values. We took high and low as being the true methylation states of the sites. For each of the 18k CpG sites, we then predicted the states high and low from the methylation levels inferred by epiG and Bis-SNP.

For epiG, the inferred level is the percentage of reads with inferred CpG methylation. Bis-SNP outputs automatically an inferred percentage for each site. We calculated the receiver operating characteristic (ROC) curves for epiG and Bis-SNP (see Additional file 1: Figure S23), and chose the thresholds that gave the best classifiers for epiG and Bis-SNP. Specifically, we chose the thresholds that minimized the squared sum $\mathrm{FP}^2 + (1 - \mathrm{TP})^2$, where FP is the false positive rate and TP the true positive rate. The inference failed for 1% (10%) of the sites for

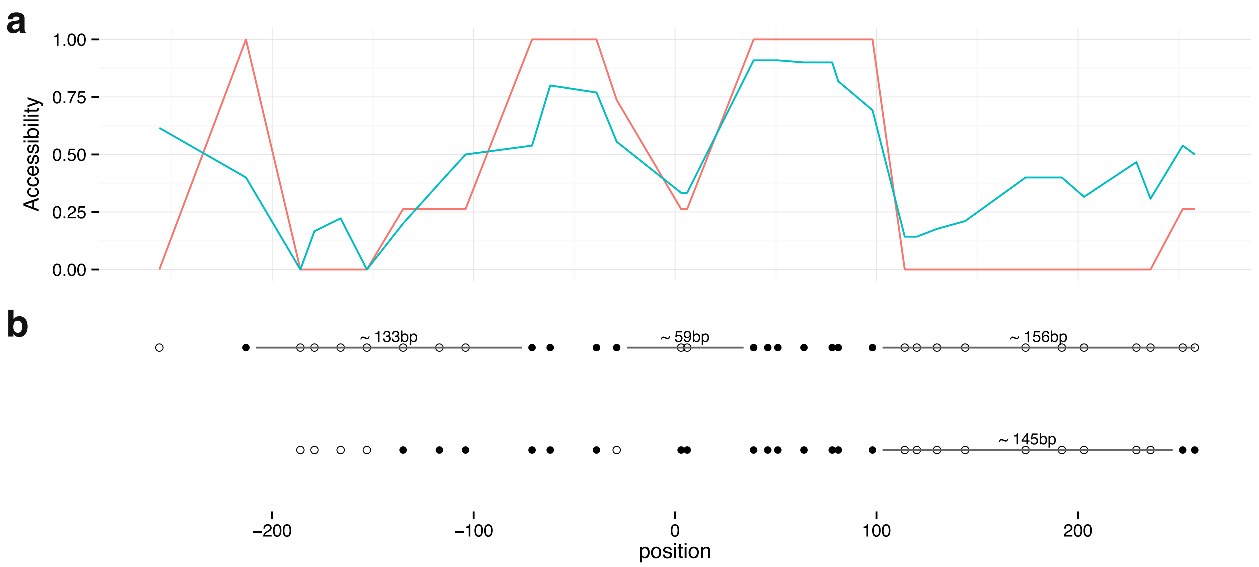

Fig. 6 DNA accessibility near a CTCF site. **a** Consensus DNA accessibility near a representative CTCF site at position 31,109,691 on chromosome 20 in the LNCaP sample. The center of the CTCF site is at position 0 of the plot. *Red line*: consensus based on the two chains shown in (**b**); *blue line*: consensus based on all reads. **b** The two dominating epi-alleles. Each point is an isolated GpC position. *Black*: methylated GpC position (DNA accessible), *white*: unmethylated GpC position (DNA inaccessible). The approximate lengths of the inaccessible intervals are computed (*black lines*), showing that the CTCF site appears to be flanked by nucleosomes (∼149 bp). In one epi-allele, the CTCF site appears to be inaccessible, potentially due to binding of a protein. As nucleosome positioning is dynamic and the read coverage is low for these samples, the inferred intervals are uncertain

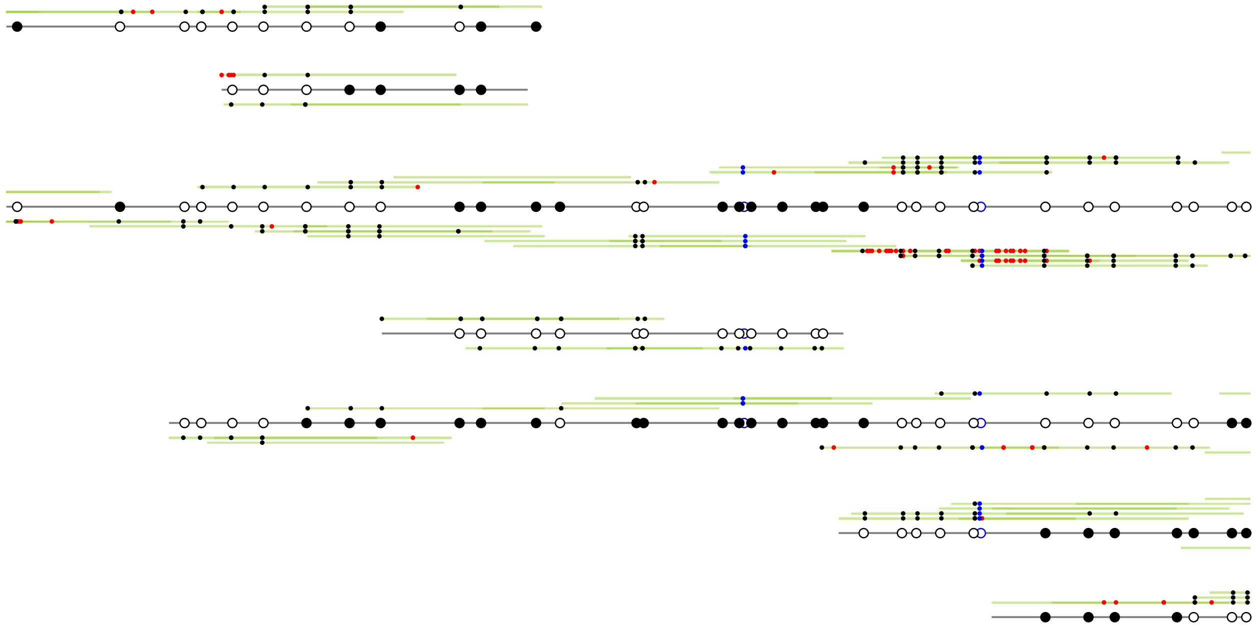

Fig. 7 Haplotype chains near a CTCF site. The inferred haplotype chains in a small 500-bp region near the CTCF site at position 31,109,691 on chromosome 20 (same as in Fig. 6) in the LNCaP sample. Seven chains are found (*black lines with open and closed circles*). The two major chains are also shown in Fig. 6. The reads (*green lines*) assigned to a haplotype chain are shown above (inferred forward direction) and below (inferred reverse direction) the chain, with paired reads on the same line. Converted isolated CpGs are marked in *blue* and converted isolated GpCs in *black*. Bases not matching the reference genome are marked in *red* and indicative of errors (sequencing or bisulfite conversion). After assigning reads to haplotypes chains, the epigenotype of the chain is inferred from the reads: converted GpCs (*solid black circles*), unconverted GpCs (*open black*), and converted CpGs (*open blue*; there are only two of these)

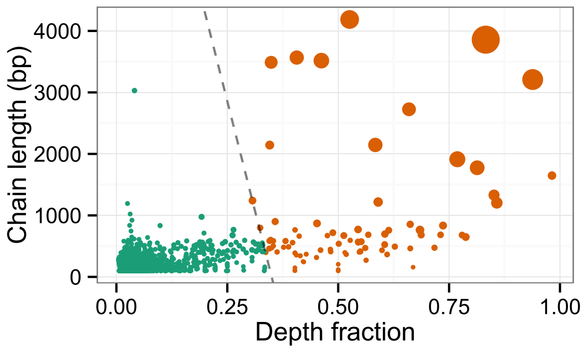

Fig. 8 Depth fraction and length of haplotype chains. Plot showing the lengths of haplotype chains against the depth fraction of the chains in a 100-kbp region near the GNAS locus for the WA9 sample. Each *point* is a chain and the size indicates the number of reads within it. To reduce noise and to extract the dominating haplotype chains, chains falling to the left of the *straight line* are removed as they either have a short length or contain a low fraction of the reads in the region. *Green*: removed chains. *Brown*: kept chains. The exact same cut-off was applied to all samples in all analyses. The chains are plotted against the genomic position in Fig. 9 and in Fig. 3 after the removal of noise

epiG (Bis-SNP). For epiG, these are sites not covered by any reads or sites for which the sequenced nucleotide is uncertain (N).

Using the optimized thresholds, in Fig. 10 we plotted the fraction of true predictions against the read depth. Not surprisingly, and for both methods, the accuracy increases with read depth and was above 90% when the read depth is higher than five reads.

Genotype validation

For SNP genotype validation, we selected all high-confidence genotype calls ($\sim$74k) on chromosomes 1 and 2 on the SNP6.0 array from the LNCaP sample (see ‘Methods’). The genotype calls on the array were taken to be the true genotypes. Of the $\sim$74k SNPs, 170 had zero read depth (no reads available) and, hence, inference failed for both epiG and Bis-SNP for these sites.

For epiG, the combined genotype is computed from the haplotype chains with log-likelihood ratios $F \leq -15$ or $|F_R - F_A| \geq 15$ (see ‘Log-likelihood ratios’ in ‘Methods’ for definitions). This is done to avoid haplotype chains with very few reads affecting the combined genotype. The combined genotype is then the genotype represented by the remaining haplotype chains. For example, if there are two remaining haplotype chains with genotype C and one with T then the combined genotype is CT.

The overall result of the genotype validation can be seen in Fig. 10. The performance differs for the different SNP genotypes (see Additional file 1: Figure S24). It is best for the homozygous genotypes and less good for the genotypes AG and CT, where confusion with methylated cytosines is possible.

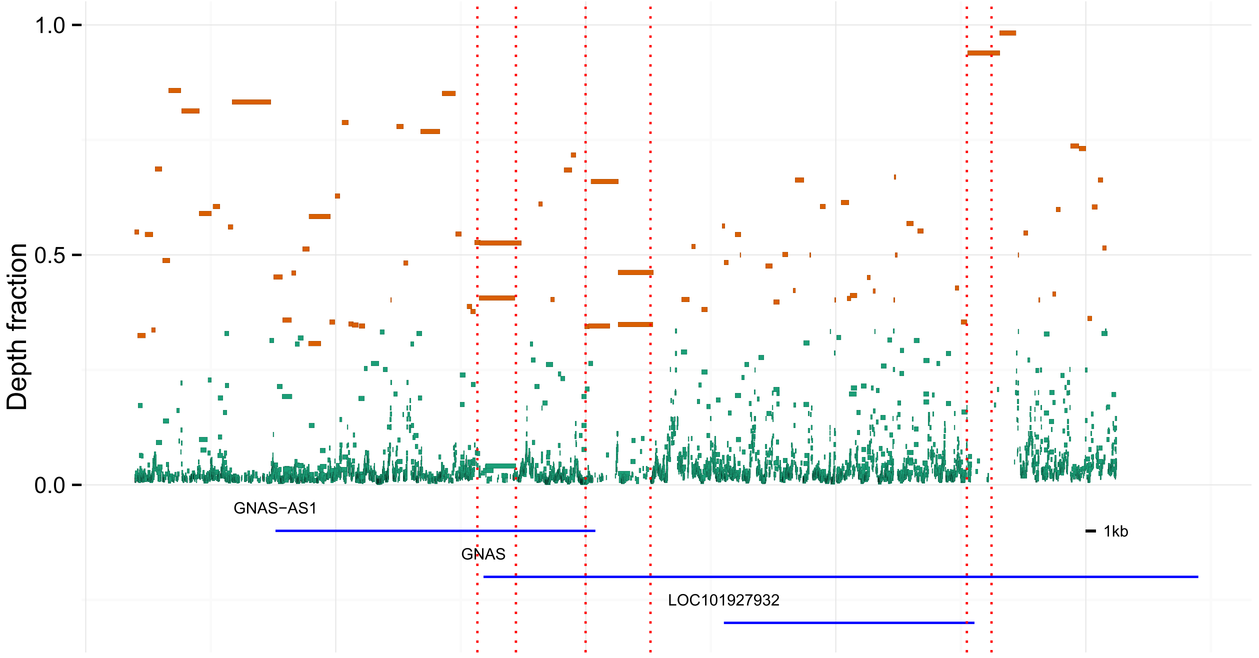

Fig. 9 Inferred haplotype chains. The plot shows all inferred haplotype chains in a 100-kb region near the GNAS locus in the WA9 sample before noise reduction, as Fig. 3. The depth fraction of a chain is plotted against its genomic position. The color indicates whether the chain is kept (*brown*) or interpreted as noise (*green*). See Fig. 8

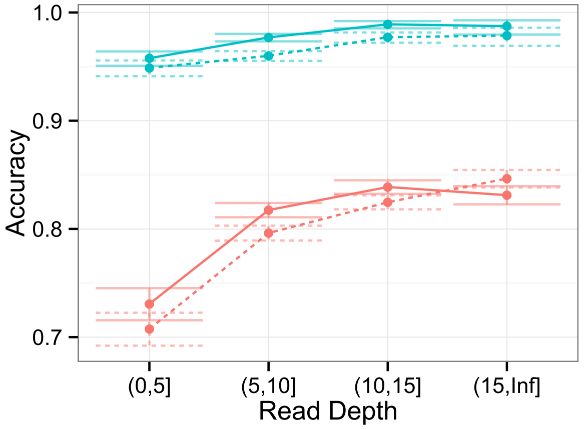

Fig. 10 Position-wise methylation and SNP validation. Accuracy plotted against read depth for epiG (*solid line*) and Bis-SNP (*dotted line*). *Blue lines* are CpG consensus methylation predictions and *red lines* SNP predictions. 95% confidence intervals are shown. The two methods show comparable results. *SNP* single nucleotide polymorphism

Robustness

epiG depends on a number of user-adjustable parameters. To investigate the robustness of the performance of epiG, we varied the default parameters and compared the results for different parameter choices. Of particular interest are the following questions: (1) How does the methylation and genotype validation depend on the parameters? (2) How does the number of dominating haplotype chains vary with the parameters? Different choices of parameters might divide haplotype chains into several smaller chains or create fewer longer chains by merging smaller chains. An important issue is whether the number of chains overlapping a particular site in the genome remains the same for different parameter values. Refer to 'Methods' and Table 3 for a description of the parameters.

Additional file 1: Figure S25 shows the results of varying the failed conversion rate α and the inappropriate conversion rate β on the methylation calls. Specifically, we calculated the area under the curve (AUC) of the ROC curve, as in Additional file 1: Figure S23, for a grid of α and β values (including the default values). The AUC changes only by less than 2%, yielding high robustness in calling methylation status.

Additional file 1: Figure S26 shows the influence of the genotype prior on the genotype calls. The prior depends on a single adjustable parameter q and uses information from hg19 and dbSNP135 (optional), weighted by q; see 'Methods' for details. If dbSNP135 is not used, then epiG assumes there is only one possible genotype for each genomic position, except for private mutations. We repeated the genotype validation study on the same ~74k SNPs using different configurations of the prior. The percentage of correctly called AA and TT genotypes is unaffected by the prior settings (including with and without dbSNP135). For the genotypes CC and GG, the situation is different. Due to the bisulfite conversion process, the true genotypes might be confused with the genotypes CT and GA, respectively, if dbSNP135 is used. In fact, if the read depth is *high*, we see a marked decline in the performance, mainly for low q. This is not the case if dbSNP135 is not used. The calls of the heterozygous genotypes are also affected by changes in the prior setting, again because the bisulfite conversion process creates ambiguities in the interpretation of the data. However, in this case dbSNP135 helps resolve ambiguities unlike for the CC and GG genotypes. The percentage of correctly called SNPs increases by up to ~30 percentage points with the inclusion of dbSNP135, depending on SNP type and read depth. Based on these investigations, we advocate the use of dbSNP135 and a high q.

Note that these results might be extrapolated to non-SNP positions as this corresponds to having a homogeneous genotype and not using dbSNP135. From the discussion above, we conclude that epiG performs well for non-SNP positions. The performance for private mutations is not clear and could be negatively influenced by the prior.

Additional file 1: Figure S27 shows that the number of haplotype chains overlapping a particular site in the genome essentially remains unchanged when varying the values of $K_0 \geq 10$ (bp) and $K_1 \geq 1$ (CpG sites), which control the required overlap between reads and haplotype chains. This is, in particular, the case for the ASM regions. However, individual haplotype chains might still be split up or merged according to the values of K_0 and K_1.

Conclusions

Various statistical methods exist for drawing inference on epigenetic patterns in general and ASM specifically. Admixture models, e.g. as implemented in armfinder [9], have been used to identify AMRs from WGBS data without inferring the structure of the underlying epi-allelic haplotypes. Other methods, such as Bis-SNP [20] and Bismark [21], infer CpG methylation levels position-wise (and they output the degree of methylation) from WGBS data without distinguishing between epi-alleles.

We have developed a novel method to infer CpG methylation states from WGBS data and epigenetic states, including DNA methylation and nucleosome positioning, from NOMe-seq data. The method groups similar reads into haplotype chains, thereby making it possible to draw inference at the haplotype level rather than at the nucleotide level alone. In particular, we have demonstrated that epiG is able to infer allele-specific methylated epi-alleles in different WGBS samples and to reveal information about nucleosome occupancy in NOMe-seq data. The separation of reads into distinct epi-allelic haplogroups is a first step in providing a statistical method that is able to extract information about epigenetic mixtures of cell populations and their frequencies, and ultimately enable the comparison of epigenetic states between cell types.

We benchmarked epiG against array data from prostate cells grown in culture and showed that epiG performed comparably to Bis-SNP with respect to position-wise methylation. The software epiG is available for download (see 'Software' in 'Methods' for details).

Methods

Statistical model

We first formulate a statistical model for WGBS (and NOMe-seq) data based on the bisulfite sequencing protocol described in [17]. The model takes into account sequencing errors, and failed and inappropriate bisulfite conversions [32], as well as errors in methylation of GpC sites (NOMe-seq data only) [18]. The model consists of a conversion model and a sequencing model. See Fig. 11 for the overall structure.

The true *epigenotype* of a position in a DNA fragment is one of

$$\mathsf{C}, \mathsf{G}, \mathsf{A}, \mathsf{T}, \overset{\text{me}}{\mathsf{C}}, \text{ or } \underset{\text{me}}{\mathsf{G}},$$

representing the nucleotides cytosine, guanine, adenine, and thymine, plus methylated cytosine and guanine complementary to a methylated cytosine on the reverse strand, respectively. (The epigenotype is haploid and not diploid as is normally the case.) We observe the epigenotype subject to two levels of noise: first of all, the DNA fragment is potentially treated with GpC methyltransferase (NOMe-seq), then bisulfite converted and PCR amplified. Secondly, it is passed through a sequencing machine. For NOMe-seq data, we take the methylation state of the enzyme-treated genome to be the true epigenotype.

We are interested in the probability of the observed nucleotide D in a given position in a read as a function of the epigenotype g and the strand r (fwd, rev) from which the read originates. Both g and r are in this context unknown parameters that we aim to estimate. The probability is modeled in two steps, reflecting the two levels of noise.

First, we let Z be the nucleotide at the position in the DNA fragment after conversion and PCR amplification. The probability of Z depends on the parameters g and r, the rate α of failed bisulfite conversions, and the rate β of inappropriate bisulfite conversions. It is given in Table 4 and builds on the observations in Fig. 1.

Second, the probability of D given Z depends on the reliability of the base-calling. We treat the reliability as an additional parameter, ϵ, in the model and obtain this parameter from the Phred sequencing quality scores [33, 34]. Specifically, we put

$$\mathbf{P}(D = d \mid Z = z) = \begin{cases} 1 - \epsilon, & \text{if } d = z, \\ \frac{\epsilon}{3}, & \text{if } d \neq z, \end{cases}$$

for $d, z = \mathsf{C}, \mathsf{G}, \mathsf{A}, \mathsf{T}$. As we treat N as an unknown nucleotide, we have $\mathbf{P}(D = \mathsf{N} \mid Z = z) = 1$. We have

$$\begin{aligned} &\mathbf{P}(D = d \mid g, r) \\ &\quad = \sum_{z \in \{\mathsf{C},\mathsf{G},\mathsf{A},\mathsf{T}\}} \mathbf{P}(D = d \mid Z = z)\mathbf{P}(Z = z \mid g, r) \\ &\quad = \left(1 - \frac{4}{3}\epsilon\right) \mathbf{P}(Z = d \mid g, r) + \frac{\epsilon}{3}, \end{aligned} \tag{1}$$

for $d = \mathsf{C}, \mathsf{G}, \mathsf{A}, \mathsf{T}$ and $\mathbf{P}(D = \mathsf{N} \mid g, r) = 1$.

The model can be made position or context dependent by changing the default settings (see Additional file 1). In particular, the quality score always depends on the position.

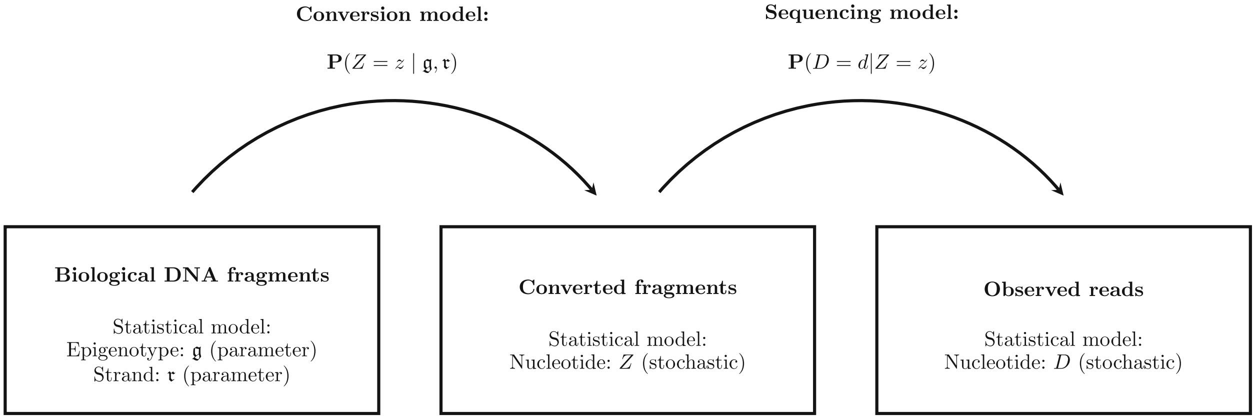
Conversion model:
$\mathbf{P}(Z = z \mid \mathfrak{g}, \mathfrak{r})$
Sequencing model:
$\mathbf{P}(D = d | Z = z)$
Biological DNA fragments
Statistical model:
Epigenotype: $\mathfrak{g}$ (parameter)
Strand: $\mathfrak{r}$ (parameter)
Converted fragments
Statistical model:
Nucleotide: Z (stochastic)
Observed reads
Statistical model:
Nucleotide: D (stochastic)

Fig. 11 Outline of the statistical model

Table 4 The bisulfite conversion model

	Epigenotype g					
Z	C	G	A	T	$\overset{\text{me}}{\mathsf{C}}$	$\underset{\text{me}}{\mathsf{G}}$
C	$\hat{\rho}\alpha$	0	0	0	$1-\hat{\rho}\beta$	0
G	0	$\rho\alpha$	0	0	0	$1-\rho\beta$
A	0	$1-\rho\alpha$	1	0	0	$\rho\beta$
T	$1-\hat{\rho}\alpha$	0	0	1	$\hat{\rho}\beta$	0

The probability $\mathbf{P}(Z = z \mid g, r)$ of Z given the epigenotype g and the strand r. Here α is the rate of failed bisulfite conversion and β is the rate of inappropriately converted nucleotides. The parameter ρ is 0 if $r =$ fwd, 1 if $r =$ rev, and $\hat{\rho} = 1 - \rho$. For example, $\mathbf{P}(Z = \mathsf{C} \mid g = \mathsf{C}, r = \text{fwd}) = \alpha$

Overview of the inference procedure

The sequenced reads come from different cells that might or might not share epigenetic structure. Typically, we find several non-compatible epigenetic haplotypes, as well as reads that support a common, or similar, origin. We call a collection of DNA fragments from the same epigenetic haplotype for an *epigenetic haplotype chain* or just a haplotype chain.

We propose an algorithm that does the following:

(1) Infer haplotype chains, that is, cluster the reads according to their epigenetic haplotype.
(2) Infer the epigenotype of each position and strand of each read in a haplotype chain.

These inference steps are done using a constrained maximum posterior procedure based on the model Eq. 1, as described in the 'Details of the inference procedure.'

Notation and definitions

We use standard nomenclature and let H be any nucleotide different from G, and D be any nucleotide different from C [35]. A CpG site is said to be isolated if the context is HCGD and likewise, a GpC is said to be isolated if the context is DGCH.

Let n denote the number of reads and b_i the haplotype chain of read i. If read i and j come from DNA fragments with the same epigenetic haplotype, then $b_i = b_j$. The strand of read i is called r_i. We refer to $B = (b_1, \ldots, b_n)$ as the *haplotype structure* and to $R = (r_1, \ldots, r_n)$ as a *strand assignment.*

Each position in a haplotype chain has a unique epigenotype. For a haplotype chain b, let

$$g_b = \left(g_{S_b b}, \ldots, g_{(S_b+L_b-1)b}\right) \tag{2}$$

be an assignment of an epigenotype to each position in the chain, where S_b is the start position of the chain (with respect to a reference genome) and L_b the length of the chain. We refer to g_b as an *epigenotype chain.* The collection of epigenotype chains, one for each haplotype chain, is denoted by G and referred to as an *epigenotype assignment.*

An epigenotype chain is said to be *strand compatible*, if the epigenotypes of each CpG site are either CG or $\overset{\text{me}}{\mathsf{C}}\underset{\text{me}}{\mathsf{G}}$, that is, if C is methylated on the forward strand, then the C on the reverse strand complementary to G is also methylated. In NOMe-seq mode, this is required only for isolated CpG sites as well as all isolated GpC sites, for which the epigenotypes should be GC or $\underset{\text{me}}{\mathsf{G}}\overset{\text{me}}{\mathsf{C}}$.

Let $\bar{g}$ denote the nucleotide of an epigenotype g. For example, if $g = \mathsf{C}$, then also $\bar{g} = \mathsf{C}$ and if $g = \overset{\text{me}}{\mathsf{C}}$, then $\bar{g} = \mathsf{C}$.

Feasible haplotype chains

For a haplotype chain b, let $c(b,j)$ be the read depth of the chain at position j, that is, $c(b,j) \geq 0$ is the number of reads in the chain overlapping position j. The range of the haplotype is the interval $[S_b, S_b + L_b - 1]$ from the start position of the first read of the chain until the end position of the last read of the chain.

A haplotype chain b with epigenotype chain g_b is said to be *feasible* if

1. It has positive read depth throughout its range, that is, if $c(b,j) > 0$ for all $j \in [S_b, S_b + L_b - 1]$.
2. It consists of a single read, or it is obtained from another feasible haplotype chain b' by adding one read i_b, such that
   (a) The overlap between b' and i_b is at least K_0 positions.
   (b) The overlap between b' and i_b contains at least K_1 CpGs, K_2 isolated CpGs, and K_3 isolated GpCs.
3. The epigenotype chain g_b is strand compatible.

A haplotype structure B is said to be feasible if all haplotype chains of B are feasible. The haplotype structure that assigns every read to its own haplotype chain is feasible.

Default values for K_0, K_1, K_2, and K_3 are listed in Table 3. A high overlap between two reads makes it more likely the reads originate from the same epi-allelic haplotype, than a small overlap. However, if we require a large overlap, few reads will cluster together, leading to many haplotype chains with few reads. There is, therefore, a balance to strike. If reads are long, we can in general ask for a larger overlap than if reads are small. If two reads do not overlap in any CpGs, then there is no epigenetic evidence they are from the same epi-allelic haplotype. Hence K_1 should be at least 1.

Details of the inference procedure

The likelihood and priors

The haplotype structure B, the strand R, and epigenotype assignment G are unknown parameters and will be inferred from the data. The parameters α, β, and ϵ will be fixed (assumed known) in the inference procedure and not estimated from the data.

The lengths of the reads are denoted by $l_1, \ldots, l_n$. We observe the nucleotide sequences

$$\left(d_{i1}, \ldots, d_{il_i}\right) \in \{\mathsf{C}, \mathsf{G}, \mathsf{A}, \mathsf{T}, \mathsf{N}\}^{l_i}, \quad i = 1, \ldots, n,$$

of the n reads. In our data analyses, l_i is typically around 75–100.

With this in mind, the likelihood of the observed reads is given by

$$\mathcal{L}(B, G, R) = \prod_{i=1}^{n} \prod_{j=1}^{l_i} \mathbf{P}(D = d_{ij} \mid g_{(s_i+j)b_i}, r_i),$$

where $s_1, \ldots, s_n$ are the start positions of the reads in the genome. Here $g_{(s_i+j)b_i}$ is the epigenotype at the jth position in the ith read of the haplotype chain g_{b_i} containing the read (see Eq. 2). The parameters α, β, and ϵ are suppressed in $\mathbf{P}$ as they are not estimated from the data.

We add priors on B, G, and R. We use a uniform prior on R, such that all strand assignments a priori are equally likely. The prior on B controls the number of haplotype chains, that is, the tendency of reads to be grouped into the same haplotype chain (potentially at the cost of imposing errors). The prior on G is derived from hg19 (reference genome) and dbSNP135. The purpose of this prior is to guide the inference in situations of low read depth, abundant errors, or high methylation levels.

Optimization

The posterior likelihood for the optimization problem is defined as

$$\Lambda(B, G, R) = \pi_0(B)\pi_1(G\,|B)\mathcal{L}(B, G, R), \tag{3}$$

where $\pi_0(B)$ and $\pi_1(G\,|B)$ are the priors on the haplotype structure and the epigenotype assignment, respectively. As the prior on R is uniform, it can be omitted. In other words, $\Lambda(B, G, R)$ is the likelihood of the data weighted by $\pi_0(B)$ and $\pi_1(G\,|B)$.

The inference procedure updates estimates of B, G, and R iteratively, such that the posterior likelihood increases in each step. In the $(k+1)$th step, new estimates $\widehat{B}_{k+1}$ and $\widehat{R}_{k+1}$ are proposed from the estimates $\widehat{B}_k$ and $\widehat{R}_k$ in the previous step. Given $\widehat{B}_{k+1}$ and $\widehat{R}_{k+1}$, the optimal estimate $\widehat{G}_{k+1}$ of G can be computed directly (see Additional file 1). The algorithm is guaranteed to converge in a finite number of steps. However, in rare cases, the final estimates are found to be suboptimal (see Additional file 1). The initial haplotype structure assigns all reads to their own haplotype chain. Note that if we do not use a prior on B, then the algorithm will remain in the initial configuration, as the likelihood does not increase by re-assigning reads to other chains.

For further details and implementation, see Additional file 1.

Priors

Let I_b denote the set of all reads in haplotype chain b, hence $I_b = \{i = 1, \ldots, n \mid b_i = b\}$. The prior on B is defined as

$$\pi_0(B) \propto \prod_{i=1}^{n} \sqrt{L_{b_i}} \left(\sum_{j \in I_{b_i}} l_j \right)^2.$$

The prior gives higher weight to longer chains as well as to chains for which the total length of all reads is high. If all reads have the same length, then

$$\pi_0(B) \propto \prod_{i=1}^{n} n_i^2 \sqrt{L_{b_i}},$$

where n_i is the number of reads in I_{b_i}, and the prior simply weights the number of reads in the chains to their length. The exponents could be chosen in many ways. Here the number of reads in a chain weights higher (exponent 2) than the length of the chain (exponent 0.5); thus, we favor thick chains to long (thin) chains. This is sensible, as we are particularly interested in the dominating epi-alleles.

The general form of the prior is based on practical considerations. The space of all haplotype structures B is a highly complex space and for computational reasons is it essential that the prior weight is relatively straightforward to calculate.

We assume the prior $\pi_1(G\,|B)$ is of the form

$$\pi_1(G\,|B) \propto \prod_{j} \prod_{b \in B(j)} q^{\mathbb{I}\left(\overline{g}_{jb}\right)} (1-q)^{1-\mathbb{I}\left(\overline{g}_{jb}\right)},$$

where the first product is over all positions in the haplotype chains, $B(j)$ is the set of haplotype chains overlapping position j, $\overline{g}_{jb}$ is the nucleotide of position j in chain b (Eq. 2), and $\mathbb{I}(\overline{g}_{jb})$ is 1 if the genotype is found in hg19 or dbSNP135, and it is 0 otherwise (dbSNP135 is optional). If $q = 0.5$, then $\pi(G\,|B)$ is independent of hg19 and dbSNP135 and gives the same prior weight to all observations. The prior serves two purposes: the main purpose is to guide epiG in situations where the data might be interpreted in different ways because the bisulfite conversion process creates ambiguities; secondly, it weights the likelihood of introducing private mutations. We take $q = 0.9999$ and use the same value of q for all data sets, even though it might be reasonable to adjust the parameter according to what is believed about the sample; for example, cancer cells are generally more exposed to private mutations than other cell types.

There is a missing normalization constant obtained by normalizing the distribution $\pi_0(B)\pi_1(G\,|B)$ to 1. It has no influence on the estimation procedure. In practice, we, therefore, ignore the normalization constant.

Log-likelihood ratios

To assess the fitted epigenotypes, we compute three likelihood ratios. For a given haplotype structure and strand assignment, we compute a log-likelihood for each haplotype chain b that assesses the significance of a particular epigenotype x in position j in the chain,

$$\ell(x) = \sum_{b_i=b} \log \mathbf{P}(D = d_{ij} \mid x, r_i),$$

where the sum is over all reads in the chain overlapping the position (see Eq. 1). The dependence on b and j is suppressed in $\ell(x)$.

The log fit-ratio is defined as

$$F = 2\left(\max_{x\notin\{g,\bar{g}\}} \ell(x) - \max_{x\in\{g,\bar{g}\}} \ell(x)\right), \qquad (4)$$

where g is the inferred epigenotype in the position. The log fit-ratio measures how well the inferred epigenotype (or the corresponding unmethylated nucleotide) fits compared to any other epigenotype. It can take any value since $\ell(x)$ is only maximized as part of the full posterior likelihood (see Eq. 3).

Additionally, the fit of the genomic reference nucleotide (or the alternative nucleotide, if relevant) can be similarly assessed:

$$F_R = 2\left(\max_{x\notin\{y,y^{\mathrm{me}}\}} \ell(x) - \max_{x\in\{y,y^{\mathrm{me}}\}} \ell(x)\right) \qquad (5)$$

(similarly F_A for the alternative nucleotide), where y denotes the reference (or alternative) nucleotide and y^{me} denotes the methylated nucleotide when $y = \mathsf{C}, \mathsf{G}$. Again, this quantity can take any value.

WGBS data

We downloaded four publicly available WGBS data sets (see Table 2 and Additional file 1: Figure S28 for details). All data sets were mapped to the hg19 reference genome using BSMAP with the standard configuration [24]. For a paired design, reads not properly paired were removed. The bam file was sorted using Picard tools [37].

LNCaP and PrEC cell culture

LNCaP prostate adenocarcinoma cells and PrEC normal primary prostate epithelial cells were obtained from the American Type Culture Collection (ATCC). LNCaP cells were grown in RPMI 1640 with L-glutamine supplemented with 10% fetal bovine serum, 100 U/ml penicillin, and 100 μg/ml streptomycin (Life Technologies). The authenticity of the LNCaP cell line was confirmed by short tandem repeat analysis (www.identicell.dk). PrEC cells were grown in Prostate Epithelial Cell Basal Medium (ATCC) supplemented with Prostate Epithelial Cell Growth Kit (ATCC). LNCaP and PrEC cells were harvested at 80% confluence using Trypsin-EDTA (Life Technologies) or Trypsin-EDTA for primary cells (ATCC) and Trypsin Neutralizing Solution (ATCC), respectively. For microarray analyses, genomic DNA from PrEC and LNCaP cells was extracted using the Puregene DNA purification kit (Gentra Systems) with proteinase K treatment (100 U, 30 min at 55 °C), as described previously [38].

Microarray analyses

All microarray analyses were performed by service provider Aros Applied Biotechnology, Aarhus, Denmark. For SNP genotyping of LNCaP and PrEC cells, genomic DNA was labeled and hybridized to the Genome-Wide Human SNP6.0 array (Affymetrix, Santa Clara, CA, USA), as described previously [39]. SNP6.0 data processing and analysis were performed as previously [39].

For methylation profiling, 1 μg of DNA was bisulfite converted, whole-genome amplified, and analyzed on the Infinium Human Methylation450 BeadChip (Illumina, San Diego, CA) according to the protocol provided by the manufacturer. This array interrogates the methylation states of >485,000 CpG sites per sample at single nucleotide resolution, where each investigated CpG site is assigned a β value ranging from 0 (fully unmethylated) to 1 (fully methylated), corresponding to the ratio of the methylated signal divided by the sum of the methylated and unmethylated signals. Normalized peak-corrected β values were used.

Genome-wide nucleosome footprinting assay

NOMe-seq was performed as previously described [18]. Briefly, exponentially growing cells were washed with phosphate buffered saline, trypsinized, and incubated with ice-cold lysis buffer (10 mM Tris, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA, and 0.5% NP-40) for 5 min on ice to isolate intact nuclei. Nuclei were washed with ice-cold wash buffer (10 mM Tris, pH7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA), resuspended in ice-cold 1× GpC buffer (New England BioLabs), and treated with 200 units of M.CviPI enzyme supplemented with 1.5 μL S-adenosylmethionine (SAM) for 7.5 min with a boost of 100 units enzyme and 0.75 μL SAM for an additional 7.5 min. Genomic DNA was isolated by standard phenol-chloroform extraction and ethanol precipitation. WGBS libraries were generated using 2–5 μg of DNA as previously described and sequenced on a HiSeq 2000 performed by the USC Epigenome Center [40]. Sequencing reads were mapped to the hg19 reference genome using BSMAP.

Additional file

Additional file 1: Additional information and figures. Additional information about the statistical method, in particular, the optimization and implementation of epiG, as well as additional figures referenced in the main text. (PDF 1880 kb)

Acknowledgments

We thank Søren Vang for valuable assistance with data handling.

Funding

CW acknowledges support from the Danish Cancer Society and the Danish Research Council. This work was supported by the Danish Strategic Research Council/Innovation Fund Denmark.

Availability of data and materials

An implementation of the model and inference procedure are available in the R package epiG. The package is available on GitHub (github.com/vincent-dk/epiG) under the GNU General Public License version 2 (https://www.gnu.org/licenses). Further details about the implementation of epiG can be found in Additional file 1. Essential scripts used for data analysis and how to use epiG are also available on GitHub (github.com/vincent-dk/using-epiG). The version of the software used for this publication is available and assigned to doi:10.5281/zenodo.229860 (epiG software) and doi:10.5281/zenodo.229861 (primary scripts).

Details on how Bis-SNP is run can be found in Additional file 1.

GEO accession numbers for WGBS data used in this study are GSM1521762, GSM983645, GSM1204466, and GSM1204465, and for NOMe-seq data, GSE94361 (see also Table 2).

Authors' contributions

MV developed the model, assisted by CW. MV developed the optimization procedure and implemented, tested, and analyzed the WGSB and the NOMe-seq data. MV and CW interpreted the results and wrote the paper with input from the other authors. KM, KDS, TFØ, JSP, GL, and PAJ performed laboratory experiments, including on NOMe-seq data, and provided scientific input. We thank two anonymous reviewers for comments and suggestions that improved the presentation of the method. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

No ethics approval was needed for this study.

Author details

[1]Department of Mathematical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark. [2]USC Norris Comprehensive Cancer Center, Keck School of Medicine, 90089-9176 Los Angeles, CA, USA. [3]Department of Molecular Medicine, Aarhus University Hospital, 8200 Aarhus, Denmark. [4]Department of Urology, Keck School of Medicine, University of Southern California, 90089 Los Angeles, CA, USA. [5]Van Andel Research Institute, 49503 Grand Rapids, MI, USA.

Received: 17 June 2016 Accepted: 7 February 2017
Published online: 21 February 2017

References

1. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003;33:245–54.
2. Rakyan V, Down T, Balding D, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2001;12:529–41.
3. Dawson M, Kouzarides T. Cancer epigenetics: from mechanism to therapy. Cell. 2012;150:12–27.
4. Laird P. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet. 2010;11:191–203.
5. Mikeska T, Candiloro I, Dobrovic A. The implications of heterogeneous DNA methylation for the accurate quantification of methylation. Epigenomics. 2010;2:561–73.
6. Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature. 2007;447:415–32.
7. Feinberg A, Irizarry R. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci USA. 2010;170:1757–64.
8. Issa JP. Epigenetic variation and cellular Darwinism. Nat Genet. 43;2011:724–5.
9. Fang F, Hodges E, Molaro A, Dean M, Hannon G, Smith A. Genomic landscape of human allele-specific DNA methylation. Proc Natl Acad Sci USA. 2012;109:7332–7.
10. Tycko B. Allele-specific DNA methylation: beyond imprinting. Human Mol Genet. 2010;19(R2):210–20.
11. Meaburn E, Schalkwyk L, Mill J. Allele-specific methylation in the human genome. Epigenetics. 2010;5(7):578–82.
12. Babak T, DeVeale B, Tsang B, Li X, Smith K, Kukurba K, et al. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. Nat Genet. 2015;47(5):544–9.
13. Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res. 2010;20(7):883–9.
14. Keverne E. Epigenetically regulated imprinted genes and foetal programming. Neurotox Res. 2010;18:386–92.
15. Ziller M, Gu H, Müller F, Donaghey J, Tsai L, Kohlbacher O. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500:477–81.
16. Lister R, Pelizzola M, Dowen R, Hawkins R, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.
17. Krueger F, Kreck B, Franke A, Andrews S. DNA methylome analysis using short bisulfite sequencing data. Nat Methods. 2012;9:145–51.
18. Kelly T, Liu Y, Lay F, Liang G, Berman B, Jones P. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res. 2012;22:2497–506.
19. Hansen K, Langmead B, Irizarry R. Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13:83.
20. Liu Y, Siegmund K, Laird P, Berman B. Bis-SNP: combined DNA methylation and SNP calling for bisulfite-seq data. Genome Biol. 2012;13:61.
21. Krueger F, Andrews S. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics. 2011;27:1571–3.
22. Hebestreit K, Dugas M, Klein HU. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics. 2010;29(13):1647–53.
23. Peng K, Ecker J. Detection of allele-specific methylation through a generalized heterogeneous epigenome model. Bioinformatics. 2012;28(12):163–71.
24. Xi Y, Li W. BSMAP: whole genome bisulfite sequence mapping program. BMC Bioinform. 2009;10:232.
25. Ziller M, Edri R, Yaffe Y, Donaghey J, Pop R, Mallard W, et al. Dissecting neural differentiation regulatory networks through epigenetic footprinting. Nature. 2015;518:355–9.
26. Roadmap Epigenomics Consortium, Kundaje A, Meuleman Q, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
27. Banet G, Bibi O, Matouk I, Ayesh S, Laster M, Kimber K, et al. Characterization of human and mouse h19 regulatory sequences. Mol Biol Rep. 2000;27:157–65.
28. Rugg-Gunn P, Ferguson-Smith A, Pedersen R. Status of genomic imprinting in human embryonic stem cells as revealed by a large cohort of independently derived and maintained lines. Hum Mol Genet. 2007;16: 243–51.
29. Gebert C, Kunkel D, Grinberg A, Pfeifer K. H19 imprinting control region methylation requires an imprinted environment only in the male germ line. Mol Cell Biol. 2010;30(5):1108–15.
30. Taberlay P, Statham A, Kelly T, Clark S, Jones P. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. 2014;24:1421–32.

31. Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol. 2013;20(3):267–73.
32. Genereux D, Johnson W, Burden A, Stöger R, Laird C. Errors in the bisulfite conversion of DNA: modulating inappropriate-and failed-conversion frequencies. Nuc Acids Res. 2008;36:150.
33. Ewing B, Hillier L, Wendl M, Green P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. 1998;8:175–85.
34. Ewing B, Green P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. 1998;8:186–94.
35. Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nuc Acids Res. 1985;13(9):3021–30.
36. Leontiou C, Hadjidaniel M, Mina P, Antoniou P, Ioannides M, Patsalis P. Bisulfite conversion of DNA: performance comparison of different kits and methylation quantitation of epigenetic biomarkers that have the potential to be used in non-invasive prenatal testing. PLoS ONE. 2015;10(8):0135058.
37. Wysoker A, Tibbetts K, Fennell T. The Picard toolkit (icardtools 1.5.3.) http://broadinstitute.github.io/picard.
38. Abildgaard M, Borre M, Mortensen M, Ulhøi B, Tørring N, Wild P, et al. Downregulation of zinc finger protein 132 in prostate cancer is associated with aberrant promoter hypermethylation and poor prognosis. Int J Cancer. 2013;130:885–95.
39. Schepeler T, Lamy P, Hvidberg V, Laurberg J, Fristrup N, Reinert T, et al. A high resolution genomic portrait of bladder cancer: correlation between genomic aberrations and the DNA damage response. Oncogene. 32;2012:3577–86.
40. Blatter A, Yao L, Witt H, Guo Y, Nicolet C, Berman B, et al. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. Genome Biol. 2014;15:469.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central