Information recovery from low coverage whole-genome bisulfite sequencing

Emanuele Libertini¹, Simon C. Heath², Rifat A. Hamoudi³, Marta Gut², Michael J. Ziller⁴,⁵,⁶, Agata Czyz⁷, Victor Ruotti⁷, Hendrik G. Stunnenberg⁸, Mattia Frontini⁹,¹⁰,¹¹, Willem H. Ouwehand⁹,¹⁰,¹², Alexander Meissner⁴,⁵,⁶, Ivo G. Gut² & Stephan Beck¹

The cost of whole-genome bisulfite sequencing (WGBS) remains a bottleneck for many studies and it is therefore imperative to extract as much information as possible from a given dataset. This is particularly important because even at the recommend 30X coverage for reference methylomes, up to 50% of high-resolution features such as differentially methylated positions (DMPs) cannot be called with current methods as determined by saturation analysis. To address this limitation, we have developed a tool that dynamically segments WGBS methylomes into blocks of comethylation (COMETs) from which lost information can be recovered in the form of differentially methylated COMETs (DMCs). Using this tool, we demonstrate recovery of ~30% of the lost DMP information content as DMCs even at very low (5X) coverage. This constitutes twice the amount that can be recovered using an existing method based on differentially methylated regions (DMRs). In addition, we explored the relationship between COMETs and haplotypes in lymphoblastoid cell lines of African and European origin. Using best fit analysis, we show COMETs to be correlated in a population-specific manner, suggesting that this type of dynamic segmentation may be useful for integrated (epi)genome-wide association studies in the future.

¹ Medical Genomics, UCL Cancer Institute, University College London, London WC1E 6BT, UK. ² Centro Nacional de Análisis Genómico (CNAG), Parc Científic de Barcelona, Torre I, 08028 Barcelona, Spain. ³ Division of Surgery and Interventional Science, University College London, London W1W 7EJ, UK. ⁴ Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. ⁵ Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA. ⁶ Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ⁷ Illumina Inc., San Diego, California 92121, USA. ⁸ Department of Molecular Biology, Radboud University Nijmegen, Nijmegen 6525 GA, Netherlands. ⁹ Department of Haematology, University of Cambridge, Cambridge, CB2 0XY, UK. ¹⁰ National Health Service Blood and Transplant, Cambridge Biomedical Campus, Cambridge, CB2 0XY, UK. ¹¹ British Heart Foundation Centre of Excellence, University of Cambridge, Cambridge, CB2 0QQ, UK. ¹² Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. Correspondence and requests for materials should be addressed to E.L. (email: emanuele.libertini@ucl.ac.uk) or to S.B. (email: s.beck@ucl.ac.uk).
Whole-genome bisulfite sequencing (WGBS) is the method of choice for the generation of reference methylomes and increasingly being used in basic and clinical research as well. To facilitate the complex analysis of such WGBS methylomes, a wide range of pipelines and algorithms has been developed with respect to cost, scale, resolution and biological questions. Informed decisions on resource allocation need to be made to tailor the data analysis to the experimental design while taking into account the advantages and disadvantages of single CpG resolution profiling with WGBS, where methylation estimates are derived from a cell population or a single cell assay. In a separate study, we conducted an assessment of sequencing coverage required for quantitative detection of resolution-dependent methylome features such as differentially methylated positions (DMPs e.g., dynamic CpG sites), differentially methylated regions (DMRs, e.g., tissue or disease specific DMRs) and blocks of comethylation (COMETs), described here. As part of this study, we found that the majority of publicly available methylomes are single replicate, restricting the statistical analysis to e.g., tissue or disease specific DMRs and blocks of comethylation (COMETs), described here. As part of this study, we found that the majority of publicly available methylomes are single replicate, restricting the statistical analysis to e.g., tissue or disease specific DMRs and blocks of comethylation (COMETs), described here.

Another area of recent advancement includes analyses based on patterns of comethylation which were first observed over short (~1 kb) distances by chromosome-wide profiling and subsequently confirmed by WGBS. More recently, similar types of analyses have been developed for the identification of regulatory regions using methylSeekR. GeMes is a method of choice for the generation of reference methylomes and increasingly being used in basic and clinical research as well. To facilitate the complex analysis of such WGBS methylomes, a wide range of pipelines and algorithms has been developed with respect to cost, scale, resolution and biological questions. Informed decisions on resource allocation need to be made to tailor the data analysis to the experimental design while taking into account the advantages and disadvantages of single CpG resolution profiling with WGBS, where methylation estimates are derived from a cell population or a single cell assay. In a separate study, we conducted an assessment of sequencing coverage required for quantitative detection of resolution-dependent methylome features such as differentially methylated positions (DMPs e.g., dynamic CpG sites), differentially methylated regions (DMRs, e.g., tissue or disease specific DMRs) and blocks of comethylation (COMETs), described here. As part of this study, we found that the majority of publicly available methylomes are single replicate, restricting the statistical analysis to e.g., tissue or disease specific DMRs and blocks of comethylation (COMETs), described here.

Another area of recent advancement includes analyses based on patterns of comethylation which were first observed over short (~1 kb) distances by chromosome-wide profiling and subsequently confirmed by WGBS. More recently, similar types of analyses have been developed for the identification of regulatory regions using methylSeekR. GeMes is a method of choice for the generation of reference methylomes and increasingly being used in basic and clinical research as well. To facilitate the complex analysis of such WGBS methylomes, a wide range of pipelines and algorithms has been developed with respect to cost, scale, resolution and biological questions. Informed decisions on resource allocation need to be made to tailor the data analysis to the experimental design while taking into account the advantages and disadvantages of single CpG resolution profiling with WGBS, where methylation estimates are derived from a cell population or a single cell assay. In a separate study, we conducted an assessment of sequencing coverage required for quantitative detection of resolution-dependent methylome features such as differentially methylated positions (DMPs e.g., dynamic CpG sites), differentially methylated regions (DMRs, e.g., tissue or disease specific DMRs) and blocks of comethylation (COMETs), described here. As part of this study, we found that the majority of publicly available methylomes are single replicate, restricting the statistical analysis to e.g., tissue or disease specific DMRs and blocks of comethylation (COMETs), described here.

Another area of recent advancement includes analyses based on patterns of comethylation which were first observed over short (~1 kb) distances by chromosome-wide profiling and subsequently confirmed by WGBS. More recently, similar types of analyses have been developed for the identification of regulatory regions using methylSeekR. GeMes is a method of choice for the generation of reference methylomes and increasingly being used in basic and clinical research as well. To facilitate the complex analysis of such WGBS methylomes, a wide range of pipelines and algorithms has been developed with respect to cost, scale, resolution and biological questions. Informed decisions on resource allocation need to be made to tailor the data analysis to the experimental design while taking into account the advantages and disadvantages of single CpG resolution profiling with WGBS, where methylation estimates are derived from a cell population or a single cell assay. In a separate study, we conducted an assessment of sequencing coverage required for quantitative detection of resolution-dependent methylome features such as differentially methylated positions (DMPs e.g., dynamic CpG sites), differentially methylated regions (DMRs, e.g., tissue or disease specific DMRs) and blocks of comethylation (COMETs), described here. As part of this study, we found that the majority of publicly available methylomes are single replicate, restricting the statistical analysis to e.g., tissue or disease specific DMRs and blocks of comethylation (COMETs), described here. As part of this study, we found that the majority of publicly available methylomes are single replicate, restricting the statistical analysis to e.g., tissue or disease specific DMRs and blocks of comethylation (COMETs), described here.
DMC analysis recovered ~35% of the estimated RADmeth DMPs at maximum coverage, and ~30% at only 5X. In contrast, DMR analysis recovered only ~20% of the DMPs at maximum coverage and ~10% at 5X. The difference between DMR and DMC performance is most likely caused by individual DMPs disregarded by DMR callers but able to break COMETs and thus detected by COMETvintage. Figure 3b shows an example of a DMC between M1–2 and M7–10 created by fragmentation of COMETs. For comparison, the underlying DMPs and DMRs are also shown at maximum and 30X coverage. Calling of DMRs and COMETs at different coverage is highly reproducible (Supplementary Fig. 3).
Relationship between co-methylation and haplotypes. Finally, we explored the relationship between COMETs and haplotypes. As WGBS methylome data become available on a population-wide level, high complexity feature analysis such as COMETs may also offer a way to generate an epigenetic equivalent of the haplotype map (HapMap). To illustrate this potential, we generated a 37X methylome (M5) of an African (YRU) HapMap cell line (GM18507) with known linkage disequilibrium (LD) structure and compared the YRU-derived COMETs with the corresponding YRU haplotype blocks defined by LD (Figure 4, Supplementary Figure 5). A best fit analysis revealed high correlation ($r = 0.86$, $P$-value = 0.00112) for $r^2 = 0.9$ and $OMg = 0.1$ (Supplementary Table 4) which decreased as expected by 0.4% in significance when replacing YRU by a more distant and less fragmented European (CEU) haplotype (Supplementary Fig. 6). Taken together, these findings suggest a possible functional relationship between genetic and epigenetic (DNA methylation) variants in line with recent observations using related analyses.

Discussion

A recent saturation analysis of WGBS data revealed a major limitation for calling DMPs in methylomes generated at the recommended reference coverage of 30X (ref. 8). Using a novel approach of segmenting WGBS methylomes into COMETs for subsequent calling of DMCs, we present a solution that is able to recover approximately 30% of the lost DMP information content in the form of DMCs, doubling the recovery achievable to date by DMR analysis. However, our COMET/DMC analysis is not without limitation either. As for DMR analysis, DMP recovery by DMC analysis is not possible at single CpG level. For that, the corresponding DMRs and DMCs need to be subjected to additional targeted BS-seq for which a variety of methods are readily available. However, as biological processes predominantly involve multiple and frequently clustered changes in CpG methylation, DMR/DMC resolution will be adequate for many functional studies. An alternative solution would be to recover lost DMPs by imputation which proved highly successful for the recovery of single nucleotide polymorphisms (SNPs) in low-coverage whole-genome sequencing. Towards this goal, a first method (ChromImpute) was recently developed and shown to be capable of imputing epigenomic maps with as little as 26% of supporting experimental data. While the imputed data were similar to the observed experimental data and even surpassed them in consistency, multiple complementary data were required to impute any particular mark, e.g., it is currently not possible to impute DMPs from WGBS data alone.

In addition to DMP recovery, we show COMET analysis to complement low-resolution functional methylome studies using PMD analysis. The COMETgazer algorithm provides a fine-grained segmentation of the methylome which breaks down variable regions (and detects regions of transitions) with an average block size of $\sim 1,000$ bp for COMETs compared to...
Figure 4 | Correlation between African (YRU) haplotype blocks and YRU COMETs derived from M5. Median haplotype block size defined by $r^2 > 0.9$ versus median COMET size defined by OMg = 0.1. Data was tiled over fixed windows of 100,000 bp and scaled over 0–1 (Methods).

$\sim$25,000 bp for PMDs, facilitating the identification of novel regulatory elements such as promoters and enhancers within PMDs through differential methylation using a negative binominal model. We propose that the oscillatory patterns of DNA methylation and the number of COMETMs (the fragmentation) may be used as an additional metric to characterize epigenomes and are currently pursuing an integrative analysis with other epigenomic datasets including additional modalities (histone medications, RNAseq and HiSeq) from the International Human Epigenome Consortium (IHEC). A more speculative application of COMET analysis may be to harness it in the future for epigenome-wide association studies and the generation of an epigenetic equivalent of the haplotype map. Although only based on two cell lines from African and European descent, our finding that the relationship between COMETMs and haplotype blocks appears to be population-specific is certainly interesting and warrants further investigation once WGBS data become available on a population level.

Methods

The key metrics of the methylomes used here are summarized in Supplementary Table 1 and further details are described below.

Datasets. The M1 and M2 datasets have been deposited together into EGA under accession number EGAD00001001261. The M3 dataset was downloaded from GEO using accession numbers: GSM1112840 (M7), GSM1112841 (M8) GSM916051 (M9), GSM1112848 (M10), belonging to superseries GSE46644. The M4 dataset was downloaded from GEO series GSE17917 described in Lister et al. The M5 dataset has been deposited into GEO under accession GSE66285. M6 was obtained from EGA under accession number EGAD0000100673. M11 and M12 were obtained from GSM1112838 and GSM1112842, respectively. M13 was downloaded from GSE17972.

Data processing and analysis. Read mapping. Two reference sequences were prepared based on the hg19 human reference; reference_C2T had the C residues replaced by Ts, and reference_G2A had the Gs replaced by As. The sample preparation protocol followed ensures that reads from end 1 are from either of the original DNA strands, and are therefore generally C deficient (as unmethylated C residues are converted to T), and reads from end 2 are from the complement to the original strands and are therefore generally G deficient. The read data were fully converted prior to alignment, converting the remaining C’s to T’s in end 1, and converting G’s to A’s in end 2. The WGBS data was aligned using the GEM aligner (Kulis et al.) and (Marco-Sola et al.) allowing up to 4 mismatches from the reference. Uniquely mapping reads were selected where both read end1 mapped consistently, and no other consistent set of mappings for a read pair was found with the same number of mismatches. Duplicate read pairs were identified as read pair mapping to the same position at both ends, and such pairs were merged to produce a consensus sequence for downstream analysis. Overlapping read pairs were handled by generating a single long read with the overlapping portion representing the consensus between the two ends. After read mapping, the reference sequence (C2T or G2A) that the read pair mapped to was recorded, and the original read data restored. Prior to further analysis, the 5 base pairs at the start of both read ends were trimmed since methylation estimation from these positions are unreliable due to the end repair procedure during sample preparation.

Inference of genotype and methylation status. Genotype and methylation status were estimated simultaneously using software developed at the Centre National d’Analyse Genomique. A Bayesian model, taking into account the probability of under and over conversion and sequencing error, was used to estimate the joint posterior probability of genotype and methylation at each genomic site covered by at least 2 reads. The marginal posterior genotype probability was estimated by numerical integration of the joint posterior (using Gaussian quadrature). For sites where a single genotype was present >99% of the posterior distribution, the maximum likelihood estimate of the methylation and the standard error of the estimate conditional the most probable genotype were calculated. CpG calls for downstream analysis were produced from pairs of sites called as homozogous C followed by homozoygous G with high confidence (posterior probability > 0.99).

Feature and saturation analysis. The feature and saturation analysis was conducted at University College London. All methylation were analysed on autosomes only. Features of increasing complexity were defined and computed for the subsequent saturation analysis as follows: Informative CpG sites (iCgs) were defined as canonical CpGs of at least 8X read coverage. Coverage was calculated on median counts across all iCgs and shown as iCg saturation curve. Counts were independently and randomly downsampled for every CpG. Differentially methylated positions (DMPs) in replicate analysis were defined as iCgs of genome-wide significance (P < 0.05 after FDR adjustment), 10% methylation difference and computed with RADmeth (Dolzhenko and Smith). Single replicate DMPs were called with Fisher’s Exact Test after Benjamini-Hochberg FDR adjustment (P < 0.05) with minimum 10% methylation difference and computed with custom software. Differentially methylated regions (DMRs) were defined as iCgs with 10% minimum methylation difference and at least 3 DMPs per region and computed using Sboom (Hansen et al.). Blocks of comethylated (COMETMs) were defined and computed using COMETGazer and differentially methylated COMETMs (DMCOMETMs) were defined and computed using COMETvintage. The workflow for COMET analysis is shown in Supplementary Fig. 1 and discussed in the section workflow and feature definitions. Here we describe the definition of COMETMs, and their relative count distributions as follows:

Definition of blocks of co-methylation (COMETMs): the COMETgazer algorithm. We define the stochastic oscillator of methylation (OM) (Fig. 1a) as the one-series percentage change of CpG methylation (estimated on beta values, based on smoothed count data) in a single sample calculated as follows:

\[
\text{OM} = \text{Index CpG} \times \% \ change = \text{single CpG delta} = \frac{[\text{IndexCP} - \text{IndexPP}]}{\text{IndexPP}} \times 100
\]

where IndexCP defines current CpG beta value and IndexPP is previous (upstream) CpG beta value. The harmonics define the segmentation of the COMETMs in a sequential manner across chromosomes treating the data as if it were a time series. This definition is inspired by Ulrich et al. (2013) (ref. 32) and Ryan et al. (2014) (ref. 33), applications for calculating K series % changes for stock variation in financial modelling. The relationship between methylation levels and delta OM values is shown for a representative region in Fig. 1b.

COMETMs are thus calculated using the following COMETgazer algorithm:

1. Define the CpG data points: by definition only canonical iCg are taken into consideration
2. Smooth CpG methylation (beta) scores
3. Calculate OM (Single CpG delta) globally
4. Define COMETMs as regions of contiguous iCg where OM$_i$ and OM$_{i+1}$ oscillate around 0, with the arbitrary parameter (threshold) of dynamic oscillation termed OMg (oscillator of methylation grade) set to be $\pm 0.1$ OM over smoothed beta scores (at least +/− 10% methylation difference), roughly representing 8% of the delta (OM) distribution. This step is illustrated in Fig. 1c,d.
Definition of OORTcloud distributions and the COMETVintage algorithm. Observed Oscillatory Rhythm Transition of COMET Longitudinally Obtained Undulation Domains (OORTcloud) was calculated by binning COMETs over 100,000 bp windows at each methylation level (high: hCOMET; medium: mCOMET; low: lCOMET) as shown in Supplementary Fig. 3. In this manner, we created three distributions of COMET domains.

For the DMC analysis, OORTcloud distributions were built in a count matrix in R. Differential myelocyte structure as defined by sample COMET counts was assessed with a negative binomial model using replicate values for the two samples (M1–M2, i.e., monocytes and the individual hESC replicates M7–10) using the Bioconductor package edgeR (Robinson et al.)\textsuperscript{3}. Statistical significance is taken to be at P<0.05. An example of DMP, DMR and COMET comparison is shown in Fig. 3b and Supplementary Fig. 4.

Workflow and feature definitions. The workflow for COMET analysis (https://github.com/ratifhamoudi/COMETgazer) is shown in Supplementary Fig. 1, and involves the following 3 steps:

**Step 1** (tool: COMETgazer)

Key feature: COMET, i.e., region of co-methylation

Process: For each myelocyte, individual CpG methylation level (beta) distributions were used to compute OM scores and segment samples into COMETs.

**Step 2** (tool: COMETVintage)

Key feature: OORTcloud distributions, i.e., distributions of COMET counts

Process: For each myelocyte, COMETs were binned into count distributions reflecting the COMET domains (OORTcloud).

**Step 3** (tool: COMETVintage)

Key feature: Differentially methylated COMETs (DMC)

Process: For differential methylation analysis, COMET domains (OORTcloud) were assembled into a count matrix to call regions of differential methylated COMET (DMC) counts.

Relationship between COMETs and linkage disequilibrium. Data were normalized to a 0–1 scale in order to compare COMETs with linkage disequilibrium (LD). Haplotype blocks for the sample Coriell NA18507 (HapMap GM18507) representing an African (YRU) haplotype and HapMap data for an European (CEU) haplotype block representing an European (CEU) haplotype. The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol. 8, e1000533 (2010).


Code availability. COMETgazer is available for download at https://github.com/ratifhamoudi/COMETgazer.

References


Acknowledgements
We gratefully acknowledge the participation of all NIHR Cambridge BioResource volunteers. We thank members of the Cambridge BioResource SAB and Management Committee for their support of our study and the National Institute for Health Research Cambridge Biomedical Research Centre for funding. Mattia Frontini was supported by the BHF Cambridge Centre of Excellence [RE/13/6/30180]. Research in the Ouwehand laboratory is supported by EU-FP7 project BLUEPRINT (282510) and by program grants from the National Institute for Health Research (NIHR, http://www.nihr.ac.uk); and the British Heart Foundation under numbers RP-PG-0310-1002 and RG/09/12/28096 (http://www.bhf.org.uk). The laboratory receives funding from the NHS Blood and Transplant for facilities. EL and SB were supported by EU-FP7 projects EpiTrain (316758), EpiGeneSys (257082) and BLUEPRINT (282510), the Wellcome Trust (99148) and a Royal Society Wolfson Research Merit Award (WM100023).

Author contributions

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: A.C. and V.R. are employees of Illumina Inc., a public company that develops and markets systems for genetic analysis. All other authors declare no competing financial interests. All authors declare no competing non-financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Libertini, E. et al. Information recovery from low coverage whole-genome bisulfite sequencing. Nat. Commun. 7:11306 doi: 10.1038/ncomms11306 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/