Saturation analysis for whole-genome bisulfite sequencing data

To the Editor:

Whole-genome bisulfite sequencing (WGBS) has become an integral part of basic and clinical research and has been widely used to generate reference methylomes since 2010 (refs. 1,2). However, because of the initial high cost of a 30× WGBS methylome[3], no saturation analysis has been performed to assess the information that can be harnessed from individual methylome features at different sequencing coverage. Consequently, the International Human Epigenome Consortium (IHEC; http://ihec-epigenomes.org/research/reference-epigenome-standards/) decided to sequence reference methylomes to 30× coverage, which was believed to adequately capture the majority of the methylation signal for subsequent analyses.

Here, we report the first saturation analysis for WGBS. We assessed the effect of coverage on the identification of five features that reveal key aspects of the methylome, including informative CpG sites (iCGs), differentially methylated positions (DMPs), differentially methylated regions (DMRs), blocks of comethylation (COMETs) and differentially methylated COMETs (DMCs). We carried out a downsampling analysis by sequentially removing random WGBS reads—thereby reducing coverage—to assess the loss of information for each of the above features related to coverage, resolution and complexity. Individual CpG methylation states, defined by iCGs, and methylation changes, defined by DMPs, exhibited the highest (single base) level of resolution and lowest level of complexity. In contrast, COMETs and DMCs had the lowest resolution and highest levels of feature complexity, whereas DMRs had medium resolution and complexity. On the basis of this analysis, we show that the current reference methylome coverage (30×) results in ~50% loss of DMPs and is therefore only of limited use for high-resolution feature analysis (e.g., DMPs).

We analyzed 13 WGBS methylomes (M1–13), which are summarized in **Supplementary Table 1** and **Supplementary Methods**[4]. Except for M13, all methylomes were generated by the Roadmap Epigenomics[5] (http://www.roadmapepigenomics.org/) and BLUEPRINT[6] (http://www.blueprint-epigenome.eu/) projects. The same methylomes were also used in a parallel study[4] describing the COMET, DMC and information recovery analyses. To our knowledge, M1–3 are the deepest methylomes reported to date and thus constitute particularly valuable references for future studies.

Downsampling is the method of choice for saturation analysis and assessing coverage-dependent information loss. It requires a static reference methylome against which to downsample a deep-coverage test methylome. Better results are obtained if both methylomes are available in multiple replicates as described below. For the static reference, we evaluated two pre-IHEC (i.e., created before the consortium and its guidelines were established) (M4 (ref. 7), M13 (ref. 8) and four IHEC (M7–10) methylomes (**Fig. 1**) and selected the superior IHEC replicates M7–10 (derived from human embryonic stem cells and generated by the Roadmap Epigenomics Project) against which to downsample deep-coverage test replicates M1–2 (derived from purified human monocytes and generated by the BLUEPRINT Project). For each of the five features described above, the test methylomes (M1–2) were randomly downsampled to different read-coverage levels and assessed for information loss by comparison to the static reference methylomes (M7–10). For the analysis of iCGs, DMPs and DMRs, we used BSmooth[9] and RADmeth[10], and we used COMETgazer[4] and COMETvintage[4] for the analysis of COMETs and DMCs (https://github.com/rifathamoudi/COMETgazer).

Figure 2a shows the saturation analysis of iCGs, DMPs, DMRs, COMETs and DMCs for M1–2 by downsampling from 83× or 91× to 5× sequence coverage. For each coverage and feature, the respective percentages of retained information are plotted on the *y* axis. The total number of M1–2 features

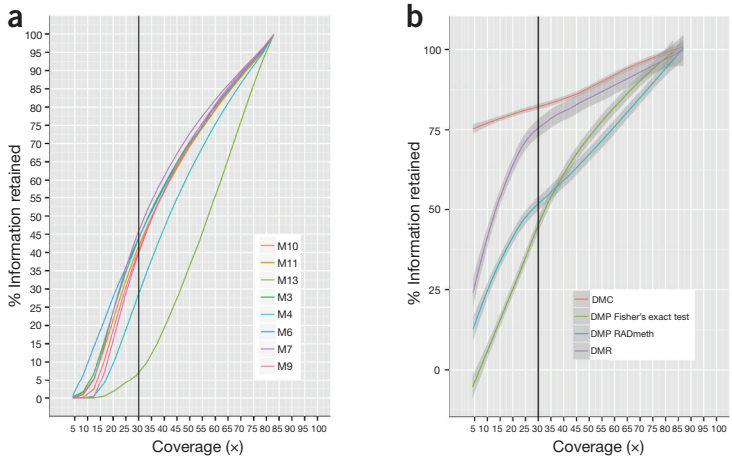

Figure 1 Single-replicate analysis. (**a**) Saturation analysis of DMP calling decay of monocyte methylome (M1) versus pre-IHEC (M4, M13) and IHEC (M3, M6–7, M9–11) methylomes using Fisher's Exact Test. (**b**) Saturation analysis of all differential methylation features using M1–2 and M3, M7–10. Single-replicate DMP calls (M1 versus M3) and replicate RADmeth analysis show a different decay and a crossover pattern. Note that in the single replicate analysis the reference (M3) is at 91×.

called at highest coverage against M7–10 was set to 100%. Whereas 95% of iCGs were retained at the current reference methylome coverage of 30×, only 50% of the 757,623 DMPs called at maximum coverage were called in double replicate analysis using RADmeth (**Fig. 2a**) and 45% in single-replicate analysis using Fisher's exact test (**Fig. 1b**; χ^2, $P < 0.0001$). A 45–50% DMP loss was confirmed using other reference methylomes (M7–10 or M11–12; **Fig. 2b**). This loss of information has not previously been reported for methylome analyses at 30× coverage. In comparison, the higher complexity (but lower resolution) DMRs, COMETs and DMCs retained 85–95% of the information. At 10× coverage ~77% and ~85% of DMC and COMET information, respectively, was retained compared to only ~40% for DMRs. Notably, using first derivatives, the information loss started at ~85× for DMPs and ~8× for DMCs (Mann-Whitney, $P < 0.0001$) (**Supplementary Methods**, statistical analysis).

The main advantage of WGBS over less expensive enrichment-based methods, such as methylated DNA immunoprecipitation sequencing (MeDIP-seq)[11] is the ability to detect DNA methylation at single-base resolution. MeDIP-seq allows detection only of DMRs but not of DMPs. Whereas reduced representation bisulfite sequencing (RRBS)[5] also has single-base resolution and thus allows detection of DMPs, it covers only ~10% of the methylome, mostly in CpG-rich regions, such as CpG islands.

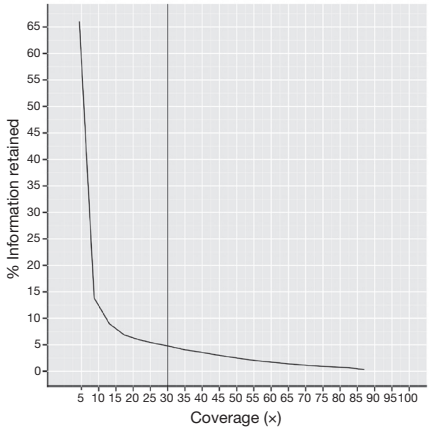

Figure 3 RRBS spike-in simulation. WGBS methylomes (M1–M2) were downsampled and spiked-in with static ~90× RRBS simulated data sets (M14–M15). Replicate DMP analysis of M1–2 versus M7–8 was performed using RADmeth. The percentage (%) information rescued reports the percentage difference in RADmeth DMP calling in the spike-in versus the WGBS alone.

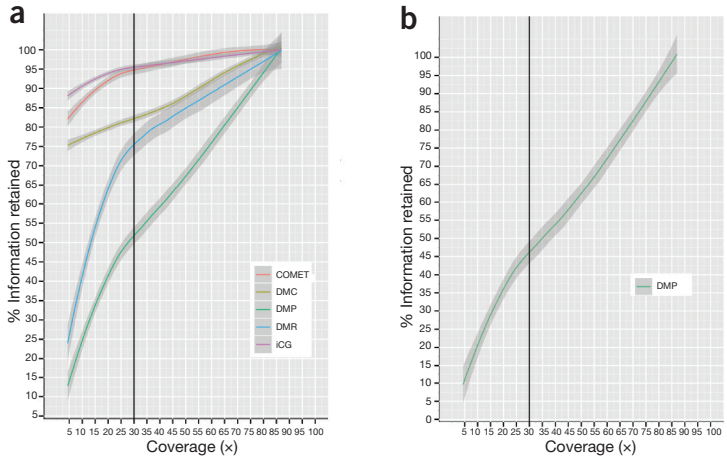

Figure 2 Saturation analysis of deep replicate methylomes. (**a**) Downsampling of methylome features for deep M1–2 against static M7–10. The analysis was conducted with RADmeth for DMPs, BSmooth for DMRs and COMETvintage for DMCs. (**b**) Replicate DMP analysis for deep M1–2 against static M7–10 or M11–12 reference methylomes, as calculated by RADmeth. This represents two independent analyses as combined results showing DMP analysis variation (shaded standard error). Downsampling iterations were run for each of the selected features by shrinking coverage by 5% for each downsampling from 100% to 5% of the data. The absolute deviation from feature calls at 100% is represented as percentage values. Colored Loess curve and shaded standard error provide estimates of information retained at each coverage across all iterations.

The increased resolution and coverage of WGBS enables the identification of genome-wide DMPs, as exemplified by the identification of dynamic CpG sites through analysis of over 40 WGBS data sets[6]. Because our saturation analysis reveals that DMP calling at ~30× coverage captured only ~50% of DMPs in a replicate analysis, we next investigated whether part of the lost information could be recovered through RRBS spike-in. As DMP loss occurs frequently in CpG-rich sequences, we spiked simulated RRBS (M14–15) into WGBS (M1–M2) data, resulting in a quantitative DMP recovery of 5% at 30× and ~12% at 10× (**Fig. 3**). **Figure 3** can be used as a guide to estimate DMP information gain for spiking RRBS into WGBS at different coverage.

We report the first saturation analysis for WGBS-based methylomes that has implications for subsequent feature analyses of the reference methylomes generated by the Roadmap Epigenomics Project[12], BLUEPRINT[13] and other members of the International Human Epigenome Consortium (http://www.ihec-epigenomes.org/). Our results demonstrate that methylomes generated at 30× coverage and single replicates were not adequate for quantitative identification of DMPs, arguably the most desirable feature of WGBS methylome analysis. To improve detection of methylation features from existing data, we have developed two algorithms (COMETgazer[4] and COMETvintage[4]) that enable partial recovery of the lost information, even at low (5×) coverage. These methods require two methylome replicates, indicating that replicates are more important than coverage in terms of maximizing the accuracy of signal that can be identified from the data. Currently, IHEC standards allow single-replicate methylomes and 60% of current IHEC methylomes are in fact single replicates. On the basis of the results of this saturation analysis, we recommend multiple replicates for future methylome sequencing.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nbt.3524).

ACKNOWLEDGMENTS

We thank the National Institute for Health Research (NIHR) Cambridge BioResource volunteers for participation, staff for volunteer recruitment, S.A.B. and Management Committee for support and the NIHR Cambridge Biomedical Research Centre for funding. S.C.H., M.G., I.G.G. and H.G.S. were supported by EU-FP7 project BLUEPRINT (282510). M.J.Z. and A.M. were supported by the US National Institutes of Health Common Fund (U01ES017155). M.F. was supported by the BHF Cambridge Centre of Excellence (RE/13/6/30180). W.H.O. was supported by EU-FP7 project BLUEPRINT (282510), the NIHR, the British Heart Foundation (RP-PG-0310-1002, RG/09/12/28096) and the NHS Blood and Transplant. J.H. was supported by The Monument Trust. E.L. and S.B. 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).

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details are available in the online version of the paper (doi:10.1038/nbt.3524).

Emanuele Libertini[1], Simon C Heath[2], Rifat A Hamoudi[3], Marta Gut[2], Michael J Ziller[4–6], Javier Herrero[7], Agata Czyz[8], Victor Ruotti[8], Hendrik G Stunnenberg[9], Mattia Frontini[10–12], Willem H Ouwehand[10–13], Alexander Meissner[4–6], Ivo G Gut[2] & Stephan Beck[1]

[1]Medical Genomics, UCL Cancer Institute, University College London, London, UK. [2]Centro Nacional de Análisis Genómico (CNAG), Parc Científic de Barcelona, Barcelona, Spain. [3]Division of Surgery and Interventional Science, University College London, London, UK. [4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [5]Harvard Stem Cell Institute, Cambridge, Massachusetts, USA. [6]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [7]Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, London, UK. [8]Illumina Inc., San Diego, California, USA. [9]Department of Molecular Biology, Radboud University Nijmegen, Nijmegen, Netherlands. [10]Department of Haematology, University of Cambridge, Cambridge, United Kingdom. [11]National Health Service Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK. [12]British Heart Foundation Centre of Excellence, University of Cambridge, Cambridge, UK. [13]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.
e-mail: emanuele.libertini@ucl.ac.uk or s.beck@ucl.ac.uk

Published online 27 June 2016;
doi:10.1038/nbt.3524

1. Bock, C. *et al. Nat. Biotechnol.* **28**, 1106–1114 (2010).
2. Harris, R.A. *et al. Nat. Biotechnol.* **28**, 1097–1105 (2010).
3. Beck, S. *Nat. Biotechnol.* **28**, 1026–1028 (2010).
4. Libertini, E. *et al. Nat. Commun.* **7**, 11306 (2016).
5. Gu, H. *et al. Nat. Protoc.* **6**, 468–481 (2011).
6. Ziller, M.J. *et al. Nature* **500**, 477–481 (2013).
7. Lister, R. *et al. Nature* **462**, 315–322 (2009).
8. Li, Y. *et al. PLoS Biol.* **8**, e1000533 (2010).
9. Hansen, K.D., Langmead, B. & Irizarry, R.A. *Genome Biol.* **13**, R83 (2012).
10. Dolzhenko, E. & Smith, A.D. *BMC Bioinformatics* **15**, 215 (2014).
11. Down, T.A. *et al. Nat. Biotechnol.* **26**, 779–785 (2008).
12. Satterlee, J.S., Schübeler, D. & Ng, H.H. *Nat. Biotechnol.* **28**, 1039–1044 (2010).
13. Adams, D. *et al. Nat. Biotechnol.* **30**, 224–226 (2012).