Correspondence to Nature Biotechnology

Resolution-dependent methylome feature analysis of whole-genome bisulfite sequencing data

Emanuele Libertini[1*], Simon C. Heath[2], Rifat A. Hamoudi[3], Marta Gut[2], Michael J. Ziller[4], Javier Herrero[5], Agata Czyz[6], Victor Ruotti[6], Hendrik G. Stunnenberg[7], Mattia Frontini[8,9,10], Willem H. Ouwehand[8,9,10,11], Alexander Meissner[4], Ivo G. Gut[2], Stephan Beck[1*]

[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, MA 02142, USA; Harvard Stem Cell Institute, Cambridge, MA 02138, USA; Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA

[5]Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, London WC1E 6BT, UK

[6] Illumina Inc., San Diego, California 92121, USA

[7]Department of Molecular Biology, Radboud University Nijmegen, Nijmegen 6525 GA, Netherlands

[8]Department of Haematology, University of Cambridge

[9]National Health Service Blood and Transplant, Cambridge Biomedical Campus, Cambridge CB2 0XY, UK;

[10]British Heart Foundation Centre of Excellence, University of Cambridge, Cambridge, United Kingdom

[11]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[*]Correspondence to EL (emanuele.libertini@ucl.ac.uk) and SB (s.beck@ucl.ac.uk)

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[1,2]. However, because of the initial high cost of a 30X WGBS methylome[3], no saturation analysis was carried out to assess the information that can be harnessed from individual methylome features at different sequencing coverage. Consequently, the International Human Epigenome Consortium (IHEC)[4] decided to sequence reference methylomes to 30X 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 inform on 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). Using downsampling by sequentially removing random WGBS reads and thereby reducing coverage, we were able to assess loss of information for each of the above features in a coverage-, resolution- and complexity-dependant manner. Individual CpG methylation states defined by iCGs and methylation changes defined by DMPs represented the highest (single base) level of resolution and lowest level of complexity. Vice versa, COMETs and DMCs had the lowest resolution but represented the highest levels of feature complexity while DMRs represented medium resolution and complexity. Based on this analysis, we showed that the current reference methylome coverage (30X) results in ~50% loss of DMPs and thus is only of limited use for high resolution feature analysis such as DMPs.

We analyzed a total of 13 WGBS methylomes (M1-13) which are summarized in **Supplementary Table 1 and Methods** and shared with *Libertini et al., Nature Communications***.** Except for M13, all methylomes were generated by the Roadmap Epigenomics[11] (www.roadmapepigenomics.org/) and BLUEPRINT[12] (www.blueprint-epigenome.eu/) projects. The same methylomes were also used in a parallel study[9] (*Libertini et al., Nature Communications*) 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 of choice and superior results are achieved if both methylomes are available in multiple replicates as described below. For the static reference, we evaluated two pre-IHEC (M4[5], M13[6]) and four IHEC (M7-10) methylomes (**Figure 1A**) 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[7] and RADmeth[8] and COMETgazer[9] and COMETvintage[9] (*Libertini et al., Nature Communications;* https://github.com/rifathamoudi/COMETgazer) for the analysis of COMETs and DMCs.

Figure 2A shows the saturation analysis of iCGs, DMPs, DMRs, COMETs and DMCs for M1-2 by downsampling from 83X or 91X to 5X 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 called at highest coverage against M7-10 is set to 100%. While 95% of iCGs are retained at the current reference methylome coverage of 30X, only 50% of the 757,623 DMPs called at maximum coverage are called in double replicate analysis using RADmeth (**Figure 2A**) and 45% in single replicate analysis using Fisher's Exact Test (**Figure 1B**) (χ^2, $p < 0.0001$). A 45-50% DMP loss is confirmed using other reference methylomes (M7-10 or M11-12, **Figure 2B**). This loss of information has not previously been reported for methylome analyses at 30X coverage. In comparison, the higher complexity (but lower resolution) DMRs, COMETs and DMCs retain between 85-95% of the information. At 10X coverage ~77% and ~85% of DMC and COMET information, respectively, is retained compared to only ~40% for DMRs. Notably, using first derivatives, the information loss starts at ~85X for DMPs and ~8X for DMCs (Mann-Whitney, $p < 0.0001$) (**Supplementary Information, Statistical Analysis**).

The main advantage of WGBS over less expensive enrichment-based methods, such as methylated DNA immunoprecipitation sequencing (MeDIP-seq)[10] is the ability to detect DNA methylation at single base resolution. MeDIP-seq only allows detection of DMRs but not DMPs. While reduced representation bisulfite sequencing (RRBS) [11] also has single base resolution and thus allows detection of DMPs, it only covers about 10% of the methylome, mostly in CpG-rich regions such as CpG islands. The increased resolution and coverage of WGBS enables the identification of genome-wide DMPs as exemplified for the identification of dynamic CpG sites through analysis of over 40 WGBS data sets[12]. As our saturation analysis revealed that DMP calling at ~30X coverage only captures ~50% of DMPs in a replicate analysis, we assessed if 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 30X and ~12% at 10X (**Figure 3**). This figure 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 Programme[13], BLUEPRINT[14] and other members of the International Human Epigenome Consortium (http://www.ihec-epigenomes.org/).

Our results demonstrate that methylomes generated at 30X coverage and single replicate are 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[9] and COMETvintage[9]; *Libertini et al., Nature Communications*) that allow partial recovery of the lost information even at low (5X) coverage. These methods do require 2 methylome replicates, indicating that replicates are more important than coverage in terms of maximising the accuracy of signal that can be identified from the data. Currently, IHEC standards allow for single replicate methylomes and 60% of current IHEC methylomes are in fact single replicate. Based on the results of this saturation analysis, we recommend multiple replicates for future methylome sequencing.

Acknowledgements

We thank the National Institute for Health Research (NIHR) Cambridge BioResource volunteers for participation, staff for volunteer recruitment, SAB and Management Committee for support and the NIHR Cambridge Biomedical Research Centre for funding. SCH, MG, IGG and HGS were supported by EU-FP7 project BLUEPRINT (282510). MJZ and AM were supported by the NIH Common Fund (U01ES017155). MF was supported by the BHF Cambridge Centre of Excellence [RE/13/6/30180]. WHO was supported by EU-FP7 project BLUEPRINT (282510), the NIHR, the British Heart Foundation (RP-PG-0310-1002, RG/09/12/28096).and NHS Blood and Transplant. JH was supported by The Monument Trust. 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).

Competing Financial Interests

AC and VR 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.

References

1. Bock, C. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol.* **28**, 1106-14 (2010).

2. Harris, R.A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol.* **28**, 1097-105 (2010).

3. Beck, S. Taking the measure of the methylome. *Nat Biotechnol*. **28**, 1026-8 (2010).

4. http://ihec-epigenomes.org/research/reference-epigenome-standards/

5. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. **462**, 315-22 (2009).

6. Li, Y., *et al*. The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol. **8**, e1000533 (2010).

7. Hansen, K.D., Langmead, B., Irizarry, R.A.. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13**, R83 (2012).

8. Dolzhenko, E., & Smith, A. D. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics **15**: 215 (2014).

9. https://github.com/rifathamoudi/COMETgazer **[Nat Comms reference goes here]**

10. Down. *et al*. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnol*. **26**, 779-85 (2008).

11. Gu. *et al*. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protoc*. **6**, 468-81 (2011).

12. Ziller, M.J. *et al*. Charting a dynamic DNA methylation landscape of the human genome. *Nature*. **500**, 477-81 (2013).

13. Satterlee, J.S., Schübeler, D., Ng, H.H. Tackling the epigenome: challenges and opportunities for collaboration. *Nat Biotechnol*. **28**, 1039-44 (2010).

14. Adams, D., *et al*. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol*. **30**, 224-6 (2012).

Figure Legends

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. Based on this result, we decided to exclude the pre IHEC methylomes (M4, M13) from the main analysis. The analysis highlights a potential technical issue of pre IHEC methylomes generated on GAII compared to IHEC methylomes generated on HiSeq platforms. **B**, Saturation analysis of all differential methylation features using M1-2 and M3, M7-10. Single replicate DMP calls (M1 vs 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 91X.

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. Coloured loess curve and shaded standard error provide estimates of information retained at each coverage across all iterations.

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

Figure 1

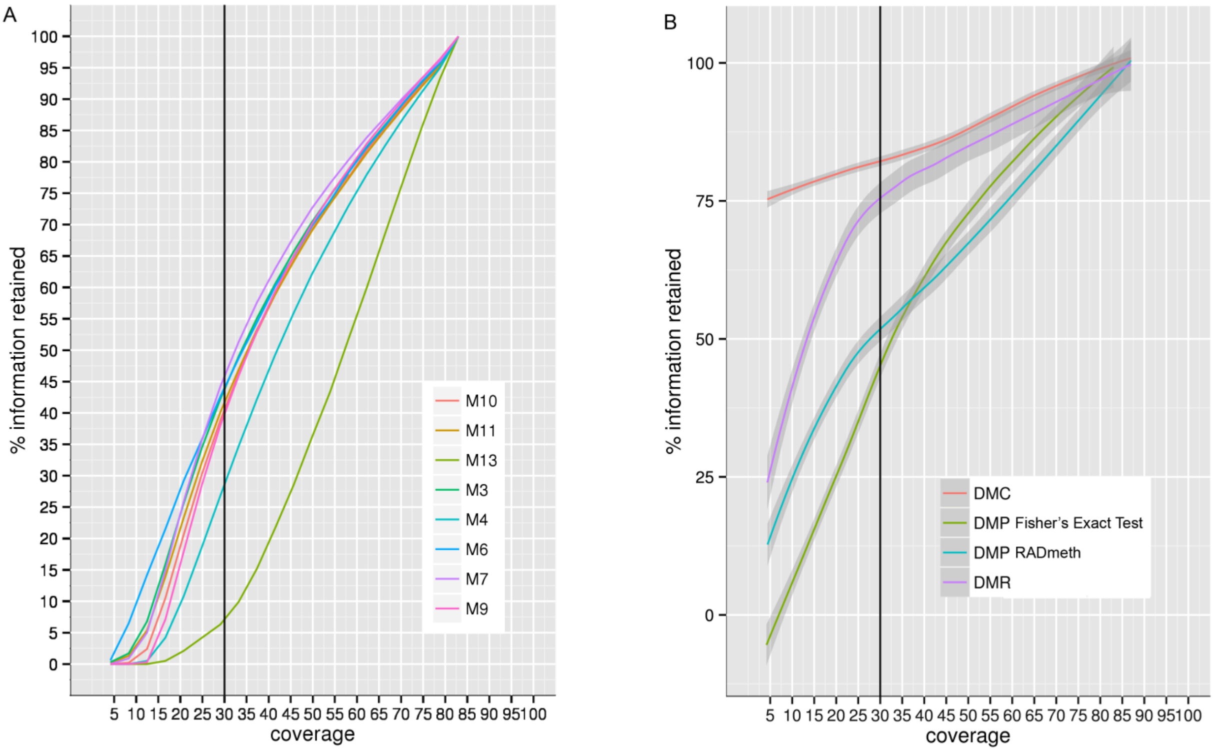

A
% information retained
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
coverage
M10
M11
M13
M3
M4
M6
M7
M9
B
% information retained
100
75
50
25
0
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
coverage
DMC
DMP Fisher's Exact Test
DMP RADmeth
DMR

Figure 2

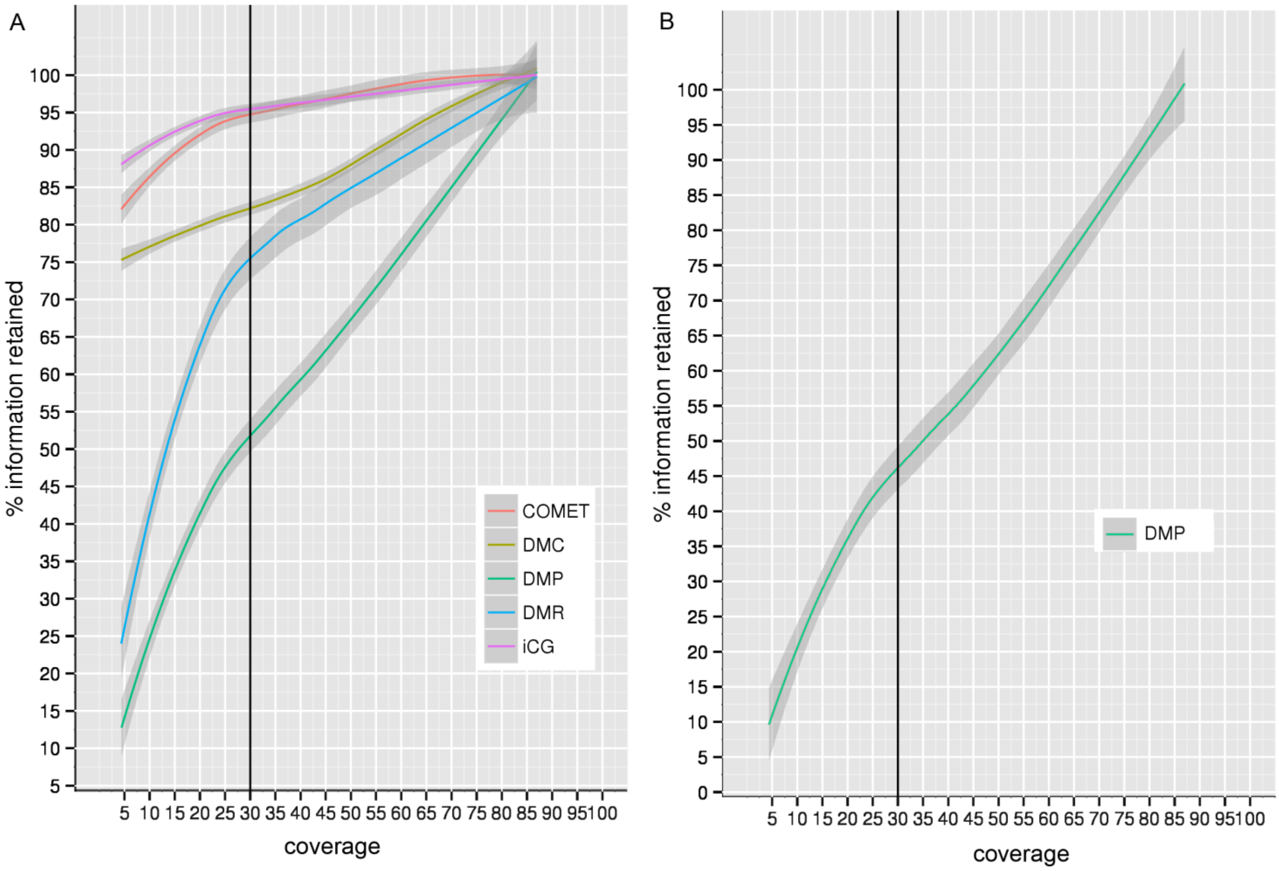
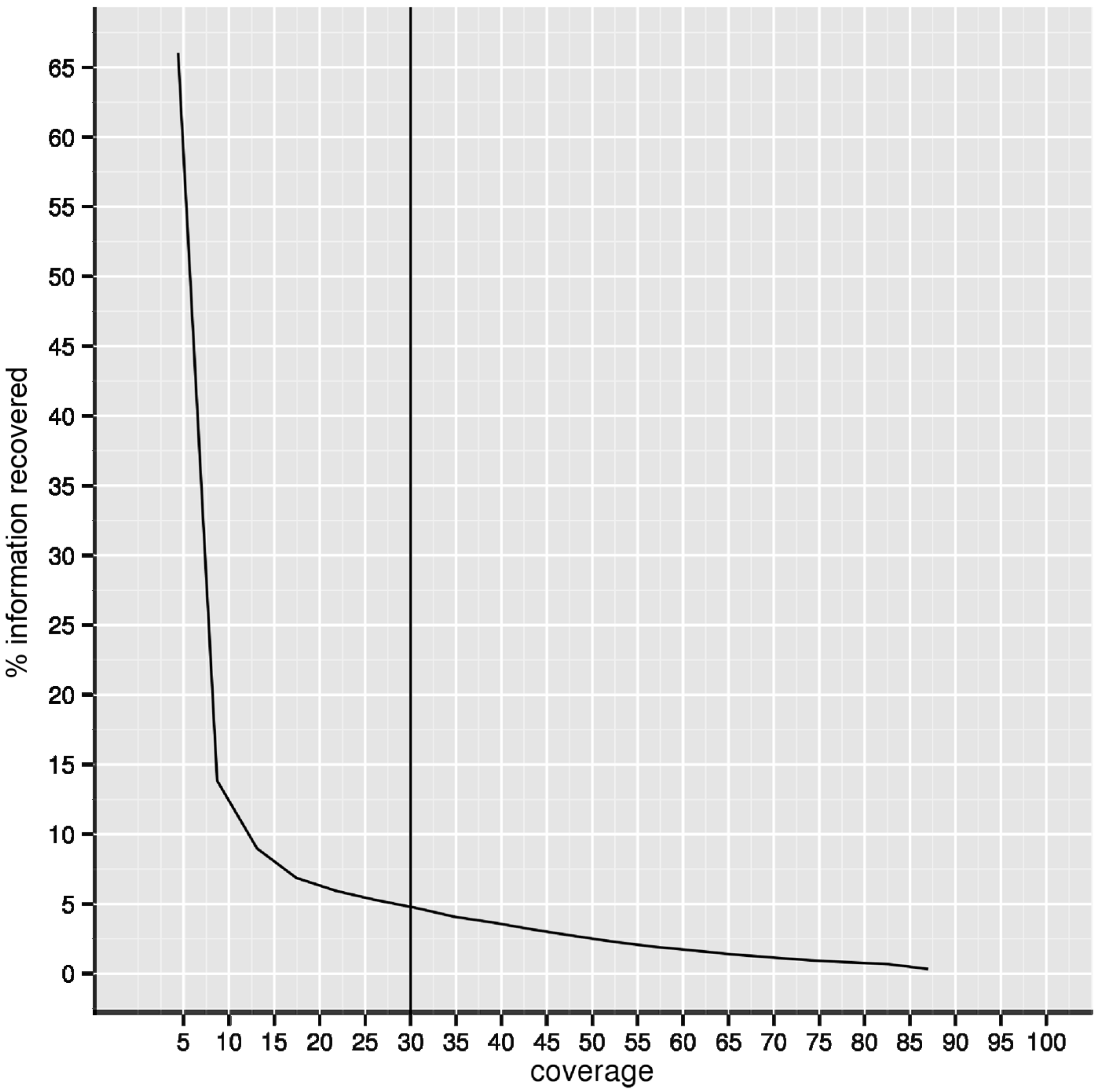

A
% information retained
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
coverage
COMET
DMC
DMP
DMR
iCG
B
% information retained
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
coverage
DMP

Figure 3


% information recovered
65
60
55
50
45
40
35
30
25
20
15
10
5
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
coverage