Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing

Michael J Ziller1–3, Kasper D Hansen4,5, Alexander Meissner1–3 & Martin J Aryee1,6–8

Whole-genome bisulfite sequencing (WGBS) allows genome-wide DNA methylation profiling, but the associated high sequencing costs continue to limit its widespread application. We used several high-coverage reference data sets to experimentally determine minimal sequencing requirements. We present data-derived recommendations for minimum sequencing depth for WGBS libraries, highlight what is gained with increasing coverage and discuss the trade-off between sequencing depth and number of assayed replicates.

DNA methylation is essential for normal development1 and is uniquely distributed in all cell types2–4. WGBS allows for unbiased genome-wide DNA methylation profiling, but little guidance currently exists with regard to the minimal required coverage and other parameters that drive the sensitivity, specificity and costs of this assay. The US National Institutes of Health (NIH) Roadmap Epigenomics Project recommends the use of two replicates with a combined total coverage of 30× (ref. 5). This requires approximately 800 million aligned, high-quality reads (101 bp paired end) for human samples and therefore remains cost prohibitive for large-scale studies. Here, we provide data-driven guidance based on comprehensive simulation experiments using representative high-quality WGBS data sets generated for the Roadmap Epigenomics Project. We focus our analysis on the discovery of differentially methylated regions (DMRs). The findings can inform decisions on the context-specific optimal experimental design strategy for methylation-profiling experiments6.

We explored three experimental scenarios: (i) a comparison of closely related sample types represented by purified CD4 versus CD8 T-cell6, (ii) a more divergent endodermal cell-type comparison represented by embryonic stem cell (ESC)-derived CD184-positive cells versus primary adult liver tissue, and (iii) unrelated brain cortex tissue versus undifferentiated human ESCs (hESCs)2 (Fig. 1a). We used a high coverage level of 30× per sample paired with regional or single CpG–based analysis paradigms to define a set of gold-standard methylation differences. These reference DMRs (refDMRs) were identified with BSsmooth8, an algorithm that uses a smoothing approach to identify regional differences. We also performed secondary analyses using MOABS, a beta-binomial hierarchical model9 approach that analyzes each CpG individually, followed by grouping neighboring differentially methylated cytosines into DMR blocks. As expected, the divergent sample comparisons yielded larger methylation differences (median difference within brain cortex and hESC DMRs = 37.9%; median difference within liver and CD184 DMRs = 39.7%) than the comparison between the closely related cell types (median difference within CD4 and CD8 T-cell DMRs = 21.5%) (Fig. 1b). Using these reference differential methylation sets as benchmarks, we then applied downsampling analysis to ask to what extent our findings would differ had we performed lower-coverage sequencing10,11.

Using the brain cortex–hESC comparison with two biological replicates per group, we observed an initial sharp rise in the fraction of recovered refDMRs as we increased coverage from 1×. The gains in the true positive rate (TPR) fell off rapidly between 8× and 10× and were followed by diminished returns at higher coverage levels (Fig. 1c and Supplementary Fig. 1a). Given the large average differences in methylation levels within DMRs for the brain cortex–hESC comparison (Fig. 1b), it is not surprising that applying a filter for minimum methylation difference in the range of 10–40% had little impact on overall sensitivity (Supplementary Fig. 1b). To investigate the effect of methylation difference magnitude in greater detail, we used two closely related T-cell types (CD4 and CD8) that exhibit considerably smaller between-group methylation differences (Fig. 1b). Interestingly, the sensitivity curve had a similar steep reduction in TPR gains above 10× (Fig. 1c and Supplementary Fig. 1a). As expected, high coverage was most important when the goal was to detect short DMRs with small methylation differences (Fig. 1d and Supplementary Fig. 1c). This observation is particularly relevant in the context of closely related sample types, in which the magnitude of the methylation differences of interest can be used to dictate sequencing depth. For example, our analysis suggests that in order to obtain a target TPR greater than 50%, one may limit consideration to regions with a minimum 20% methylation difference between two conditions at 5× coverage and a minimum 10% methylation difference at 10× (Supplementary Fig. 1b). We also note that

1Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. 2Harvard Stem Cell Institute, Cambridge, Massachusetts, USA. 3Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. 4McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. 5Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA. 6Molecular Pathology Unit, Massachusetts General Hospital, Charlestown, Massachusetts, USA. 7Center for Cancer Research, Massachusetts General Hospital, Charlestown, Massachusetts, USA. 8Department of Pathology, Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to M.J.A. (aryee.martin@mgh.harvard.edu) or A.M. (alex@broadinstitute.org).

Received 12 December 2013; accepted 24 September 2014; published online 2 November 2014; doi:10.1038/nmeth.3152
whereas altering the minimum methylation threshold affected sensitivity, it had little impact on the coverage level of the sensitivity curve ‘elbow’. In addition, we evaluated a single CpG-based DMR detection method. As expected, this approach had higher coverage requirements to achieve a similar TPR (Supplementary Fig. 1d,e) but provided single-CpG resolution.

We observed that the number of CpGs covered by at least one read dropped rapidly from 90% to 50% as coverage was reduced from 5× to 1× (Supplementary Fig. 1f–h). This uneven coverage directly contributed to a dramatic loss in sensitivity in poorly covered regions. Short DMRs with few CpGs were most susceptible to being missed as coverage decreased below 5× (Fig. 1d and Supplementary Fig. 1c).

In order to examine the sensitivity of our results with respect to the stringency with which the refDMR set is defined, we used a permutation test to compute the familywise error rate (FWER) for all DMRs in the hESC-brain comparison (Online Methods). On the basis of this statistic, we defined a high-stringency refDMR subset, accepting a high false negative rate, and confirmed that shape of the sensitivity-versus-coverage curve remained unchanged (Supplementary Fig. 1i).

We investigated the impact of coverage on DMR discovery specificity and found, as expected, that the false discovery rates (FDRs) in the comparison between the closely related T-cell types were considerably higher than those in the comparison between the brain and hESC samples (Fig. 1e and Supplementary Fig. 1j,k). Similarly to the effect of coverage on the TPR, the greatest rate of FDR improvement occurred at less than 10× coverage for both the divergent and closely related scenarios. However, to satisfy a moderately stringent FDR threshold of 20% in our system, we recommend the use of a higher 15× coverage level for closely related cell types (Fig. 1e).

We next investigated the impact of the number of biological replicate samples on DMR discovery by selecting subsets of hESC and cortex samples to create single-, two- and three-replicate data sets. The relationship between DMR recovery rate and per-sample coverage followed a similar pattern regardless of the number of replicate samples analyzed, with an initial sharp increase in the proportion of DMRs recovered up to an elbow point, followed by more modest subsequent gains (Fig. 2a and Supplementary Fig. 2a). Whereas sensitivity increased with each additional replicate, the diminishing-returns point remained stable in each case between 5× and 10×. We found that decreasing the samples per group from three to two resulted in a modest drop in sensitivity from 77% to 72% at 10× coverage. An experiment with a single replicate per group, however, achieved only 50% sensitivity at the same coverage level. Notably, increasing coverage of the single replicate had limited benefit, resulting in only 60% sensitivity...
Brief Communications

and 18% specificity even when the single replicate was deeply sequenced to 30× (Fig. 2b and Supplementary Fig. 2b). In contrast, a single CpG–based DMR detection method continued to benefit in a more linear fashion from increased per-replicate coverage (Supplementary Fig. 2c).

Finally, we sought to directly address the question of how to balance the number of replicates with the sequencing coverage per replicate. To answer this question we performed in silico experiments, keeping the total sequencing effort across the experiment constant while trading an increase in the number of biological replicates for decreased coverage per sample. We expressed the total sequencing effort across the experiment in terms of single-genome coverage, where 10× total coverage corresponds to 27 Gbp of sequence. We evaluated six scenarios ranging from 10× to 90× (equivalent to 27–243 Gbp) and evaluated sensitivity as a function of the number of biological replicates (Fig. 2c). We identified the number of samples and the corresponding coverage per sample that maximized sensitivity for each scenario. Notably, sensitivity was maximized by keeping the target coverage per sample fairly constant between 5× and 10×, regardless of the total sequencing effort. At a low total sequencing effort of 10×, sensitivity was optimized by dedicating this effort toward covering a single replicate per group at 5×. In contrast, with higher total sequencing resources available, we found minimal benefit in sequencing deeper than 10× per sample and instead obtained improved sensitivity most efficiently by increasing the number of replicates per group.

Our results highlight the minimum sequencing depth required to obtain satisfactory sensitivity and specificity in DMR discovery using WGBS in experimental settings with a limited (3) number of replicate samples. We also note the important trade-off between the two main choices for increasing power: adding additional biological replicates and adding additional per-sample coverage. For DMR identification using the sample types analyzed in this study, we recommend per-sample coverage in the range of 5–15×, depending on the magnitude of methylation differences between the groups and whether a smoothing or single CpG–based DMR identification strategy is used. Sequencing at levels higher than this range leads to wasted resources that would be better spent on an increased number of biological replicates. If the goal, however, is primarily to identify long DMRs with large methylation differences, we find that reducing coverage down to 1× or 2× per sample is acceptable (Supplementary Fig. 1c).

Our findings emphasize that biological replicates should be analyzed separately to increase power, as opposed to being pooled together for analysis. In addition, the findings strongly argue for the use of at least two separate biological replicates for DMR analysis (Supplementary Fig. 2d). Although we provide guidance based on the best currently available sample set, we caution that choosing an appropriate number of biological replicates is a complex issue influenced by the degree of within-group heterogeneity, the magnitude of between-group differences and the presence of confounding factors such as batch effects.

Methods

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

Acknowledgments

This work was supported by the NIH Common Fund (U01ES017155), NIGMS (PO1GM099117) and the New York Stem Cell Foundation. A.M. receives support as a New York Stem Cell Foundation Robertson Investigator.

Author Contributions

M.J.A. and A.M. conceived of the study. M.J.Z., M.J.A. and K.D.H. performed analysis and interpreted results. M.J.Z., M.J.A., K.D.H. and A.M. wrote the paper.

Competing Financial Interests

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Okano, M., Bell, D.W., Haber, D.A. & Li, E. Cell 99, 247–257 (1999).
2. Bird, A. Genes Dev. 16, 6–21 (2002).
3. Reik, W. Nature 447, 425–432 (2007).
4. Ziller, M.J. et al. Nature 500, 477–481 (2013).
5. NIH Roadmap Epigenomics Mapping Consortium. Standards and guidelines for whole genome shotgun bisulfite sequencing http://www.roadmapepigenomics.org/protocol (2011).
6. Mehta, T., Tanik, M. & Allison, D.B. Nat. Genet. 36, 943–947 (2004).
7. Gifford, C.A. et al. Cell 153, 1149–1163 (2013).
8. Hansen, K.D., Langmead, B. & Irizarry, R.A. Genome Biol. 13, R83 (2012).
9. Sun, D. et al. Genome Biol. 15, R38 (2014).
10. Meynert, A.M., Blicknell, L.S., Hurles, M.E., Jackson, A.P. & Taylor, M.S. BMC Bioinformatics 14, 195 (2013).
11. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Genome Res. 21, 2213–2223 (2011).
12. Hansen, K.D. et al. Genome Res. 24, 177–184 (2014).
ONLINE METHODS

Analysis was performed using Python (http://www.python.org/), R/Bioconductor13 and Snakemake14. The analysis code is available at http://aryee.mgh.harvard.edu/publications/bsdepth.

Data sets. The hESC group was composed of two biological replicates of HUES64 taken from GSE46130 and GSE46444, and one replicate of hESC line H9 from GSE62193. The HUES64 samples were collected at different passages almost 1 year apart. The H9 sample has a different genetic background. The cortex samples are primary tissue and were collected from different individuals of similar age (>80) and were obtained from GSE46444. The CD4 and CD8 primary samples were FACS sorted from blood of two different human donors. The CD4, CD8, liver and human cortex data were obtained from GSE46444 and are part of the Roadmap Epigenomics Project and are available through the epigenome roadmap data portal (http://www.roadmapepigenomics.org/). The in vitro–derived endoderm samples (CD184) were also published as part of the Roadmap Epigenomics Project and are available in GEO under GSE46130. All libraries were sequenced with at least 800 million paired-end 100-bp reads. All sex chromosomes were excluded from the analysis.

QC and read alignment. WGBS libraries were aligned using BSMap to the hg19/GRCh37 reference assembly15 and were quality controlled for bisulfite conversion rate. All samples had a bisulfite conversion rate greater than 97%.

Downsampled data sets. Each data set was downsampled by randomly selecting reads from the original aligned BAM files to create simulated samples with 1×, 2×, 5×, 10×, 20× and 30× coverage. The number of 100-bp reads needed to achieve these coverage levels was calculated as \((\text{target coverage})(2.7 \times 10^9 \text{ bp})/100 \text{ bp}\). Target coverage was defined as the total coverage across both DNA strands. For each subsampled data set, we exclusively selected reads from the corresponding high-coverage data set to create lower-coverage versions of each data set while maintaining the biological replicate information at each coverage level.

Identifying differentially methylated regions. Regional smoothing analysis. DMRs were identified using the BSsmooth algorithm6. For the two- and three-replicate experiments, we calculated a differential methylation \(t\)-statistic for each CpG covered by at least one read in each sample. These \(t\)-statistics were then smoothed across neighboring CpGs, and regions with the largest smoothed values (top 1% for cortex vs. hESC and liver vs. CD184; top 0.1% for CD4 vs. CD8) were identified as candidate DMRs. In the case of single-replicate experiments, we substituted the methylation difference at each CpG for the \(t\)-statistic. We used a permutation approach to identify a stringently defined subset of gold-standard reference DMRs, where we accept a high false negative rate with the goal of achieving a low false positive rate. Permutation tests were used as recently described12. The output of the permutation test is an estimated familywise error rate (FWER), which is corrected for multiple testing. We included only DMRs with an FWER of 0 in the stringent reference set, representing those DMRs where no putative DMR in any of the permutations is larger with a bigger methylation difference. In addition, we required a minimal methylation difference on the DMR level of 0.1 between conditions.

Single-CpG analysis. We used the mcomp function in the MOABS (model-based analysis of bisulfite sequencing data)9 framework with the following parameters settings “-d 1 --doComp=1 --doDmcScan=1 --doDmrScan=1 --minDmcsInDmr=1.” All other parameters were left at their default values. This procedure first identified differentially methylated cytosines (DMCs) followed by merging DMCs into DMRs. We required DMRs to harbor at least one DMC. We used the results of the algorithm described as 2 in ref. 9 (beta-binomial hierarchical model) to determine DMRs.

Sensitivity. Sensitivity (true positive rate, TPR) was calculated as the number of refDMRs that overlapped the downsampled DMR set by at least 1 bp. refDMRs were defined as DMRs present in the high-coverage (30×) data set using two replicates in each group, with the exception of the analyses shown in Figure 2 and Supplementary Figure 1e, where three replicates were used for the hESC-cortex comparison. Sensitivity was measured at 1×, 2×, 5×, 20× and 30× and then smoothed for plotting purposes using the cubic spline fitting algorithm implemented in the R function smooth.spline in the stats package.

Specificity. In order to assess the impact of sequencing coverage on the rate of false positive DMR discovery, we estimated false discovery rates (FDRs) for the downsampled data sets. Under the assumption that DMRs discoverable at low coverage will not become undetectable as coverage increases, we calculated FDR as the fraction of DMRs discovered in the low-coverage data sets that were not overlapping with DMRs present in the high-coverage data set. Specificity was measured at 1×, 2×, 5×, 20× and 30× and then smoothed for plotting purposes using the cubic spline fitting algorithm implemented in the R function smooth.spline in the stats package.

Replicate analysis. To assess the impact of number of replicates on DMR discovery, we identified brain cortex tissue vs. undifferentiated ESC DMRs using subsets of the six samples used for the reference DMR set. At each coverage level, we generated DMRs from each of the nine possible 1 vs. 1 replicate pairings and each of the nine possible 2 vs. 2 replicate pairings. We calculated the average true and false positive rates across the permutations.

13. Gentleman, R.C. et al. Genome Biol. 5, R80 (2004).
14. Köster, J. & Rahmann, S. Bioinformatics 28, 2520–2522 (2012).
15. Xi, Y. & Li, W. BMC Bioinformatics 10, 232 (2009).