Single-Cell Genome-Wide Bisulfite Sequencing for Analysis of Epigenetic Heterogeneity

Sébastien A Smallwood[1]*, Heather J Lee[1,5]*, Christof Angermueller[2], Felix Krueger[3], Heba Saadeh[1], Julian Peat[1], Simon R Andrews[3], Oliver Stegle[3], Wolf Reik[1,4,5]#, Gavin Kelsey[1,4]#.

*,# These authors contributed equally

Affiliations:
1: Epigenetics Programme, Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK
2: European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK
3: Bioinformatics Group, Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK
4: Centre for Trophoblast Research, University of Cambridge, Cambridge CB2 3EG, UK
5: Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK

ABSTRACT

We report a single-cell ~~whole-genome~~ genome-wide bisulfite sequencing method (scBS-Seq) capable of accurately measuring DNA methylation at up to 48.4% of CpGs. We observed that ESCs grown in serum/LIF or 2i/LIF both display epigenetic heterogeneity, with "2i-like" cells present in serum cultures. *In silico* integration of 12 individual MII oocyte datasets largely recapitulates the whole DNA methylome, making scBS-Seq a versatile tool to explore DNA methylation in rare cells and heterogeneous populations.

MAIN TEXT

DNA methylation (5mC) is an epigenetic mark with critical roles in regulation and maintenance of cell type specific transcriptional programs[1,2]. Our understanding of 5mC functionality has been revolutionized by the development of bisulfite sequencing (BS-Seq), which offers single cytosine resolution and absolute quantification of 5mC levels genome-wide. Recent technological advances have demonstrated the power of single-cell sequencing analyses for the deconvolution of mixed cell populations[3-5]~~, and.~~ Incorporation of epigenetic information into this single-cell arsenal will transform our understanding of gene regulation, and reveal new insights into the potential biological functions of epigenetic heteogeneity[6]. Here, we report an accurate and reproducible method for single-cell ~~whole-genome~~ genome-wide bisulfite sequencing (scBS-Seq) that allows assessment of DNA methylation heterogeneity within cell populations across the entire genome.

In commonly used BS-Seq protocols, sequencing adapters are first ligated to fragmented DNA and bisulfite conversion is performed, resulting in loss of information due to DNA degradation associated with bisulfite treatment. For scBS-Seq we use a modification of Post-Bisulfite Adaptor Tagging (PBAT)[7], where bisulfite treatment is performed first, resulting in simultaneous DNA fragmentation and conversion of unmethylated cytosines **(Fig. 1a**). Then, complementary strand synthesis is primed using custom oligos containing Illumina adapter sequences and a 3' stretch of 9 random nucleotides. This step is performed 5 times to ensure that maximum numbers of DNA strands are tagged and to generate multiple copies of each fragment. After biotin capture of the tagged strands, the second adapter is similarly integrated, and PCR amplification (12 or 13 cycles) is performed with indexed primers allowing multiple single-cell libraries to be sequenced together.

We performed scBS-Seq on metaphase-II (ovulated) oocytes (MIIs) and mouse embryonic stem cells (ESCs) cultured either in 2i/LIF (2i ESCs) or serum/LIF (serum ESCs) conditions. MIIs are an excellent model for technical assessment as they: i) can be individually handpicked ensuring only one cell

is processed; ii) represent a highly homogeneous population allowing discrimination between technical and biological variability; and iii) present a distinct DNA methylome comprising large-scale hyper- and hypo-methylated domains[8]. ESCs grown in standard serum/LIF conditions exist in a state of dynamic equilibrium that is characterised by transcriptional heterogeneity [9-12], and emerging evidence from immunofluorescence and locus-specific studies has provided the first hints of DNA methylation heterogeneity in ESCs[13]. Recent studies have also demonstrated the remarkable plasticity of the ESC methylome, with genome-wide hypo-methylation being induced by inhibition of FGF signaling using two kinase inhibitors (2i)[13,14]. We therefore use serum and 2i ESCs as a model to determine whether scBS-Seq can reveal DNA methylation heterogeneity at the single-cell level.

12 MII, 12 2i ESC, 20 serum ESC scBS-Seq libraries (and 7 negative controls) and their bulk counterparts (i.e. pools of cells) were sequenced on the Illumina HiSeq platform (100bp paired-end), at a relatively low sequencing depth (average 19.4 million reads). On average, 3.9 million reads were mapped (1.5M-14.3M range), with an average efficiency of 24.6% for single-cell samples (compared to 2.1% average efficiency in negative controls) **(Supplementary Fig. 1 and Supplementary Table 1)**. This relatively low mapping efficiency is mostly due to the presence of low-complexity sequences (poly-T stretches) **(Supplementary Fig. 2)**. We obtained methylation scores on an average of 3.7 million CpG dinucleotides (CpGs; 1.8M-7.7M range) corresponding to 17.7% of all CpGs (8.5-36.2% range) **(Fig. 1b)**. Of importance, more CpGs can ~~realistically~~ be obtained ~~either~~ with deeper sequencing as the limiting duplication plateau has not been reached at this sequencing depth **(Supplementary Fig. 3)**. To validate this, we sequenced two MII scBS-Seq libraries close to saturation and with longer sequencing reads (150bp). Greater sequencing depth resulted in a 1.5- and 1.9-fold increase in the number of CpGs measured **(Supplementary Table 1)**. In addition, because of the broad size distribution of fragments in scBS-Seq libraries **(Supplementary Fig. 1b)**, longer sequencing reads also result in an increase in CpGs covered (9% at saturating sequencing depth, 16% for low sequencing depth). Integrating these additional sequencing results reveals that up to 10.1M CpGs (48.4% of all CpGs) can be obtained by scBS-Seq.

Next, we investigated the reproducibility and accuracy of our scBS-Seq approach. Low levels of non-CpG methylation across all samples revealed a minimum bisulfite conversion efficiency of 97.7% (this was 98.5% by examining the mitochondrial chromosome in ESCs) **(Fig. 1c and Supplementary Table 1)**. CpG sites in MIIs were overwhelmingly called methylated or unmethylated, consistent with a highly digitized output from single cells **(Supplementary Fig.4)**. As expected, global methylation of MIIs is highly homogeneous (33.1 ±0.8%) and 2i ESCs are hypomethylated compared to serum ESCs[13]. Yet strikingly, both 2i and serum ESCs exhibit DNA methylation heterogeneity (serum: 63.9 ±12.4%, 2i: 31.3 ±12.6%) **(Fig. 1c)**. Global methylation levels measured in individual MIIs are slightly lower than bulk (39.0%), but merging all MII datasets result in 38.8% global methylation. To assess scBS-Seq accuracy at CpG resolution, we calculated the pairwise concordance across single oocyte libraries and found an average of 87.6% genome-wide (85.3-88.9% range) and 95.7% in unmethylated CpG islands (CGIs), a highly homogeneous genomic feature, hence demonstrating the technical reproducibly of scBS-Seq **(Fig. 1d)**. Of note, CpG concordance in ESCs is lower (serum: 72.7%, 2i: 69.8%), reflecting the heterogeneity of these cells **(Fig. 1d and Supplementary Fig. 5)**. At lower genomic resolution (2kb windows), we observed high correlation between individual MIIs (on average R=0.92), and between individual MIIs and the bulk sample (on average R=0.95) **(Fig. 1e)**. In addition, for each MII sample, we obtained methylation information on an average of 61.5% of all CGIs (46.3-82.7% range); of the 1,615 CGIs identified as methylated from bulk and informative in individual MIIs, at least 92% were found to be methylated by scBS-Seq, with less than 0.3% being incorrectly called as unmethylated **(Supplementary Fig. 6)**.

While scBS-Seq mapped reads are distributed homogeneously across the genome, the enrichment towards exons, promoters and CGIs observed in bulk libraries is exaggerated in scBS-Seq libraries **(Supplementary Fig. 7)**. Thus, scBS-Seq provides information on all genomic contexts, including regulatory regions **(Supplementary Table 2),** with ~~regulatory regions~~ CGIs and promoters being slightly overrepresented. Yet, obtaining ~20% coverage of CpGs per cell means that recurrent information across

samples is dependent on the nature of analytic units; conversely, *in silico* merging of individual datasets rapidly increases the number of CpGs with information **(Supplementary Fig. 8)**. Strikingly, we were able to largely reproduce the entire DNA methylation landscape of MII oocytes ~~at single CpG resolution~~ using only 12 single cells **(Fig. 1e,f and Supplementary Fig. 9)**. This capability is particularly beneficial for homogeneous cell populations, and makes our scBS-Seq approach an important tool to investigate the DNA methylation landscape from very rare material.

We next used the scBS-Seq data to explore DNA methylation heterogeneity in ESCs. A 3kb sliding window was used to estimate the methylation rate across the genome of each ESC, as well as the mean methylation rate and variance across all ESCs **(Fig. 2a)**. Cells were clustered based on the estimated methylation rates, while penalizing uncertainty in estimates due to low read counts. Two distinct clusters could be identified, representing the majority of 2i and serum ESCs **(Fig. 2b)**. Intriguingly, outlier cells from the serum condition clustered with 2i ESCs, implying that serum cultures contain "2i-like" ESCs. This demonstrates the ability of scBS-Seq to identify rare cell types within cell populations. To examine ESC heterogeneity in greater detail, we ranked sites by the estimated cell-to-cell variance, and repeated the cluster analysis for the 300 most variable sites **(Fig. 2c)**. The structure of the resulting clusters was grossly similar to that of the genome-wide analysis, and all 300 variable sites followed the global trend of being more highly methylated in serum than 2i ESCs with high similarity between sites **(Fig. 1c, Fig. 2b,c, Supplementary Figure 10 and Supplementary Fig. 11)**. This observation is consistent with the genome-wide hypo-methylation observed in 2i ESCs[13], and indicates that a major determinant of ESC heterogeneity is the global methylation level. Importantly, detailed analysis by scBS-Seq was also able to identify sites whose methylation varied more than the genome average (as shown in "All" track at the top of the heatmap), including sites with marked heterogeneity even among cells from the same growth condition (e.g. Clusters 5 and 6 in serum ESCs) **(Fig. 2c)**. A comparison of different genomic contexts revealed that regions containing H3K4me1 and H3K27ac, marks associated with active enhancers, have the greatest variance in DNA methylation, whereas CGIs and IAP repeats elements have lower variance than the genome average **(Fig. 2d and Supplementary Fig. 12)**. These findings are consistent with previous observations that distal regulatory elements are differentially methylated between tissues and throughout development[15-17]. ~~Notably, scBS-Seq identified some highly variable sites that overlapped none of the annotated features examined, suggesting that~~ Undoubtedly, further analysis will lead to the discovery of new genomic features with dynamic DNA methylation and regulatory function.

While this manuscript was in preparation, a single-cell reduced representation bisulfite sequencing (scRRBS) method was reported[18], based on the single tube RRBS strategy we previously developed[19]. While scRRBS and scBS-Seq could be seen as complementary, currently our methodology provides, at equivalent sequencing depth, information on ~5 fold more CpGs and ~1.5 fold more CGIs **(Supplementary Fig. 13)**. Future technological developments will undoubtedly allow information to be recovered from most genomic CpGs, the key being the ability to amplify DNA prior to bisulfite conversion. The ability to capture the DNA methylome from individual cells will be critical for a full understanding of early embryonic development, cancer progression and induced pluripotent stem cell (iPSC) generation.

In summary, our work provides a proof-of-principle that large-scale single-cell epigenetic analysis is indeed achievable, and demonstrates that scBS-Seq is a unique and powerful approach to accurately measure DNA methylation across the genome of single cells and to reveal DNA methylation heterogeneity within cell populations.

REFERENCES (main text)

1. Jones, P. A. *Nat. Rev. Genet.* **13**, 484–492 (2012).
2. Smith, Z. D. & Meissner, A. *Nat. Rev. Genet.* **14**, 204–220 (2013).
3. Jaitin, D. A. *et al. Science* **343**, 776–779 (2014).
4. Deng, Q. *et al. Science* **343**, 193–196 (2014).
5. Macaulay, I. C. & Voet, T. *PLoS Genet*. **10**, e1004126 (2014).
6. Lee, H. J. *et al. Cell Stem Cell* **14**, 710–719 (2014)
7. Miura, F. *et al. Nucleic Acids Res*. **40**, e136 (2012).
8. Shirane, K. *et al. PLoS Genet.* **9**, e1003439 (2013).
9. Chambers, I. *et al. Nature* **450**, 1230–1234 (2007).
10. Islam, S. *et al. Nat. Methods* **11**, 163–166 (2013).
11. Hayashi, K., *et al. Cell Stem Cell* **3**, 391–401 (2008).
12. Torres-Padilla, M. E. & Chambers, I. *Development* **141**, 2173–2181 (2014).
13. Ficz, G. *et al. Cell Stem Cell* **13**, 351–359 (2013).
14. Habibi, E. *et al. Cell Stem Cell* **13**, 360–369 (2013).
15. Stadler, M. B. *et al. Nature* **480**, 490-495 (2011).
16. Ziller, M. J. *et al. Nature* **500**, 477–481 (2013).
17. Hon, G. C. *et al. Nat. Genet*. **45**, 1198–1206 (2013).
18. Guo, H. *et al. Genome Research* **23**, 2126–2135 (2013).
19. Smallwood, S. A. *et al. Nat. Genet*. **43**, 811–814 (2011).

METHODS

Sample collection.

MII oocytes were collected from superovulated 4–5-week-old C57BL/6Babr mice, under a stereomicroscope, by mouth pipetting, and stored at -80°C. All mouse studies were done under the guidance issued by the Medical Research Council in "Responsibility in the Use of Animals for Medical Research" (July 1993) and under the authority of Home Office Project Licence 80/2363. Prior to scBS-Seq, 2X oocyte lysis buffer (10mM Tris-Cl pH7.4, 2% SDS) and 0.5µl proteinase K were added (final volume 12µl) followed by incubation at 37°C for 1h. E14 (129/Ola, male) ESCs were cultured in serum/LIF or 2i/LIF conditions as described previously[13]. The 2i ESCs had been maintained in this medium for 24 days and matched serum ESCs were cultured in parallel. Single ESCs were collected by FACS in 12µl of ESC lysis buffer (10mM Tris-Cl pH7.4, 0.6% SDS, 0.5µl proteinase K) using a BD Influx instrument in single cell 1 drop mode. ToPro-3 and Hoechst 33342 staining were used to select for live cells with low DNA content (i.e. in G0/G1). ESCs were incubated at 37°C for 1h and stored at -20°C until required for library preparation. Negative controls were either lysis buffer alone ("empty" tubes) or sorted BD Accudrop Beads, and were prepared and processed concomitantly with all single cell samples.

Single Cell Library Preparation.

Bisulfite conversion was performed on cell lysates using the Imprint DNA Modification Kit (Sigma) with the following modifications: all volumes were halved, and chemical denaturation was followed by incubation at 65°C for 90min, 95°C for 3min and 65°C for 20min. Purification was performed as described previously[7], and DNA eluted in 10mM Tris-Cl (pH 8.5) and combined with 0.4mM dNTPs, 0.4µM oligo1 ([Btn]CTACACGACGCTCTTCCGATCTNNNNNNNNN) and 1x Blue Buffer (Sigma) (24µl final) before incubation at 65°C for 3min followed by 4°C pause. 50U of Klenow exo- (Sigma) were added and the samples incubated at 4°C for 5min, +1°C/15s to 37°C, 37°C for 30min. Samples were incubated at

95°C for 1min and transferred immediately to ice before addition of fresh oligo1 (10pmol), Klenow exo- (25U), and dNTPs (1nmol) in 2.5µl total. The samples were incubated at 4°C for 5min, +1°C/15s to 37°C, 37°C for 30min. This random priming and extension was repeated a further 3 times (5 rounds in total). Samples were then incubated with 40U exonuclease I (NEB) for 1h at 37°C before DNA was purified using 0.8x Agencourt Ampure XP beads (Beckman Coulter) according to the manufacturer's guidelines. Samples were eluted in 10mM Tris-Cl (pH 8.5) and incubated with washed M-280 Streptavidin Dynabeads (Life Technologies) for 20min with rotation at room temperature. Beads were washed twice with 0.1N NaOH, and twice with 10mM Tris-Cl (pH 8.5) and re-suspended in 48µl of 0.4mM dNTPs, 0.4µM oligo2 (TGCTGAACCGCTCTTCCGATCTNNNNNNNNN) and 1x Blue Buffer. Samples were incubated at 95°C for 45s and transferred immediately to ice before addition of 100U Klenow exo- (Sigma) and incubation at 4°C for 5min, +1°C/15s to 37°C, 37°C for 90min. Beads were washed with 10mM Tris-Cl (pH 8.5) and resuspended in 50µl of 0.4mM dNTPs, 0.4µM PE1.0 forward primer (AATGATACGGCGACCACCGAGATCTACACTCTTTC-CCTACACGACGCTCTTCCGATCT), 0.4µM indexed iPCRTag reverse primer[20], 1U KAPA HiFi HotStart DNA Polymerase (KAPA Biosystems) in 1x HiFi Fidelity Buffer**.** Libraries were then amplified by PCR as follows: 95°C 2min, 12-13 repeats of (94°C 80s, 65°C 30s, 72°C 30s), 72°C 3min, 4°C hold. Amplified libraries were purified using 0.8x Agencourt Ampure XP beads, according to the manufacturer's guidelines, and were assessed for quality and quantity using High-Sensitivity DNA chips on the Agilent Bioanalyser, and the KAPA Library Quantification Kit for Illumina (KAPA Biosystems). Pools of 12-14 single cell libraries were prepared for 100bp paired-end sequencing on a HiSeq2500 in rapid-run mode (2 lanes/run).

Bulk Sample Library Preparation.
Samples from bulk cell populations were prepared according to the protocol above, with some modifications. For the bulk oocyte sample, 120 MII oocytes were collected and lysed as described above. For ESC bulk cell samples, DNA was purified from cell pellets using the QIAamp micro kit (QIAGEN), according to the manufacturer's instructions, and 50ng of purified DNA was used in the library preparation. One round of first strand synthesis was performed using 0.8mM dNTPs and 4µM oligo1, and second strand synthesis also used 0.8mM dNTPs and 4µM oligo2. Bulk cell libraries were amplified as above with 9-12 cycles of PCR.

Sequencing Data Processing and Data Analysis.
Raw sequence reads were trimmed to remove the first 9 base pairs, adapter contamination and poor quality reads using Trim Galore (v0.3.5, www.bioinformatics.babraham.ac.uk/projects/trim_galore/, parameters: --clip_r1 9 --clip_r2 9 --paired). Due to the multiple rounds of random priming performed with oligo1, scBS-seq libraries are non-directional. Trimmed sequences were first mapped to the human genome (build GRCh37) using Bismark[21] (v0.10.1; parameters: --pe, --bowtie2, --non_directional, --unmapped), resulting in 1.4% mapping efficiency (0.2-13.2% range). Remaining sequences were mapped to the mouse genome (build NCBI37) in single-end mode (Bismark parameters: --bowtie2 --non_directional). Methylation calls were extracted after duplicate sequences had been excluded. For oocyte bulk analysis, our MII bulk dataset was merged *in silico* with previously published datasets[8] (DDBJ/GenBank/EMBL accession number DRA000570). Data visualization and analysis were performed using SeqMonk, custom R and Java scripts. For Figure 1c, CG methylation was calculated as the average of methylation for each CpG position, and non-CpG methylation was extracted from the Bismark reports. Trend line in Figure 1b was calculated using polynomial regression. Percentage of concordance was calculated as the percentage of CpGs presenting the same methylation call at the same genomic position across two cells. For correlation analysis (Pearson's), 2kb windows were defined informative if at least 8 CpGs per window were sequenced. CGI annotation used is from CAP-Seq experiments[22]. Informative CGIs were defined if at least 10 CpGs per CGI were sequenced. Hyper-methylated and hypo-methylated CGIs were defined as ≥80% and ≤20% methylation respectively. Annotation for comparison of genomic contexts **(Fig. 2d, ~~and~~ Supplementary Fig. 12 and Supplementary Table 2)** were extracted from previously published datasets[15,23].

Statistical Analyses.

1) Estimating sample-specific methylation rates

We estimated for each cell j at position i the methylation rate $r_{i,j}$. To increase the coverage across cells, we employed a sliding window approach, which is conceptually similar to approaches that have been used for bulk BS-Seq [24,25]. With window size $w = 3000$bp and step size 600bp, we computed the sum of methylated ($c^+_{i,j}$) and unmethylated ($c^-_{i,j}$) read counts in each window:

$$s^+_{i,j} = \sum_{k=-w/2}^{+w/2} c^+_{i+k,j} \qquad s^-_{i,j} = \sum_{k=-w/2}^{+w/2} c^-_{i+k,j}$$

To estimate methylation rates, we modeled the sum $S^+_{i,j}$ of methylated counts as a Binomial random variable with methylation rate $r_{i,j}$:

$$S^+_{i,j} \sim \mathrm{Bin}(s^+_{i,j} + s^-_{i,j}, r_{i,j})$$

Assuming a Beta (1, 1) prior on $r_{i,j}$, leads to the maximum a posteriori estimator for methylation rates for each window and cell:

$$\hat{r}_{i,j} = \frac{s^+_{i,j} + 1}{s^+_{i,j} + s^-_{i,j} + 2}$$

We approximated the standard error of the rate estimator as follows:

$$\mathrm{SE}[\hat{r}_{i,j}]^2 = \frac{\hat{r}_{i,j}(1 - \hat{r}_{i,j})}{s^+_{i,j} + s^-_{i,j}}$$

2) Estimating mean methylation rates

We used the estimated sample-specific methylation rates $\hat{r}_{i,j}$ to estimate mean methylation rates and cell-to-cell variances. We modeled the mean methylation rate r_i at position i across all cells as a Gaussian random variable with mean $\bar{r}_i$ and variance v_i:

$$r_i \sim N(\bar{r}_i, v_i)$$

To account for differences in the standard errors $\mathrm{SE}[\hat{r}_{i,j}]$, we weighted sample j and position i by $w_{i,j} = \mathrm{SE}[\hat{r}_{i,j}]^{-2}$, and used the weighted maximum likelihood estimator

$$\hat{\bar{r}}_i = \frac{1}{\sum_j w_{i,j}} \sum_j w_{i,j} \hat{r}_{i,j}$$

to estimate $\bar{r}_i$. The corresponding standard error is given by

$$SE[\hat{\bar{r}}_i]^2 = \frac{1}{\sum_j w_{i,j}}.$$

The maximum likelihood estimator of the cell-to-cell methylation variance v_i is

$$\hat{v}_i = \frac{\sum_j w_{i,j}}{(\sum_j w_{i,j})^2 - \sum_j {w_{i,j}}^2} \sum_j w_{i,j} (\hat{r}_{i,j} - \hat{\bar{r}}_i)^2,$$

which is the unbiased weighted sample variance. The chi-squared confidence interval of the variance estimator with confidence level α is

$$\left[\hat{v}^l_i, \hat{v}^u_i\right] = \left[\frac{n_i \hat{v}_i}{\chi^2_{1-\frac{\alpha}{2}, n_i}}, \frac{n_i \hat{v}_i}{\chi^2_{\frac{\alpha}{2}, n_i}}\right].$$

Here, χ^2_{p,n_i} is the p-quantile of the chi-squared distribution with n_i degrees of freedom, where n_i is the sum of sample weights:

$${n_i}^2 = \frac{\sum_j w_{i,j}}{(\sum_j w_{i,j})^2 - \sum_j {w_{i,j}}^2}$$

To determine highly variable methylated sites, we ranked these by the lower bound $\hat{v}^l_i$ of the chi-squared confidence interval and defined the top k sites as the most variable sites. This approach is selecting sites with large estimates of cell to cell variance while penalizing for uncertainty of these estimates, which typically stems from low read counts.

3) Clustering

To cluster cells and sites, we considered a complete linkage clustering, and employed the weighted

Euclidean norm as distance measure for comparing sample j with sample j':

$$d(j,j') = \sqrt{\sum_{i=1}^{d} w_i^{j,j'} (\hat{r}_{i,j} - \hat{r}_{i,j'})^2}$$

We defined the weight $w_i^{j,j'}$ at position i as

$$w_i^{j,j'} \propto \sqrt{w_{i,j} w_{i,j'}},$$

and normalized weights to sum up to the total number of positions d. This distance measure places most emphasis on sites that are well covered in both samples.

REFERENCES (Methods)

20. Quail, M. A. et al. *Nat. Methods* **9**, 10–11 (2012).
21. Krueger, F. & Andrews, S. R. *Bioinformatics* **27**, 1571–1572 (2011).
22. Illingworth, R. S. et al. *PLoS Genet.* **6**, e1001134 (2010).
23. Creyghton, M. P. et al. *P.N.A.S.* **107**, 21931-21936 (2010).
24. Li, Y. *et al. PLoS Biol* **8,** e1000533 (2010).
25. Bock, C. *et al. Molecular Cell* **47,** 633–647 (2012).

ACCESSION CODES

Gene Expression Omnibus (GEO): GSE56879.

ACKNOWLEDGEMENTS

We thank Kristina Tabbada and the Welcome Trust Sanger Institute sequencing pipeline team for assistance with Illumina sequencing, Rachael Walker for assistance with FACS and Tim Hore for provision of ESCs maintained in 2i/LIF and serum/LIF conditions. We thank Tim Hore, Jiahao Huang, Iain Macaulay, Stephan Lorenz, Michael Quail, Thierry Voet and Harold Swerdlow for helpful discussions. This work was supported by the Biotechnology and Biological Sciences Research Council grant BB/J004499/1, Medical Research Council grant MR/K011332/1, Wellcome Trust 095645/Z/11/Z, EU FP7 EpiGeneSys and BLUEPRINT.

CONTRIBUTIONS

S.A.S. and H.J.L. designed the study, prepared scBS-Seq libraries, analysed data and wrote the manuscript. F.K., H.S., S.A. performed sequence mapping and analysed data. J.P. contributed to technical developments. C.A. and O.S. analysed data. O.S. provided advice on statistical analyses. W.R. and G.K. supervised the study and wrote the manuscript.

FIGURES LEGENDS.

Figure 1: scBS-Seq is an accurate and reproducible method for genome-wide methylation analysis.

(a) scBS-Seq library preparation is performed in 3 stages: (1) single cells are isolated and lysed before bisulfite conversion is performed; (2) 5 rounds of random priming and extension are performed using oligo1 (which carries the first sequencing adaptor) and newly synthesized fragments are purified; (3) a second random priming and extension step is performed using oligo2 (which carries the second

sequencing adaptor) and the resulting fragments are amplified by PCR. **(b)** Number of CpGs obtained by scBS-Seq correlates with the number of mapped sequences. **(c)** Global level of DNA methylation in a CpG and non-CpG context for single cells, *in silico* merged, and bulk samples. **(d)** Boxplot representation of the pairwise analysis of CpG concordance genome-wide and in unmethylated CGIs. **(e)** Pairwise correlation matrix (Pearson's; 2kb windows) for MII bulk, individual MIIs, and *in silico* merged MII scBS-Seq datasets. **(f)** Screenshots showing CpG methylation (%) quantified over 2kb windows, with red indicating high methylation and blue low methylation. Data are displayed for 4 single MII libraries and the *in silico* merged dataset from all 12 MIIs (MII merged), which closely resemble the methylation landscape of the bulk MII sample. The inset shows the correlation between MII bulk and MII merged.

Figure 2: scBS-Seq reveals DNA methylation heterogeneity in ESCs.
(a) DNA methylation rates were estimated for each ESC using a sliding window across the genome (colored dots in bottom panel, size is inverse of estimation error). The mean methylation rate across cells (black line in bottom panel) and the cell-to-cell variance (blue line in middle panel, 95% confidence interval shaded in light blue) were also estimated. The methylation rates for Bulk serum (green line) and Bulk 2i (orange line) are superimposed in the bottom panel. The region shown as an example includes the *Nanog* locus with some annotated features. **(b)** Genome-wide cluster dendrogram and distance matrix for all ESCs and Bulk samples based on the estimated methylation rates. **(c)** Heatmap for methylation rates of the top 300 most variable sites among single-cell ESC samples. Cluster dendrograms for ~~cells~~ samples (top) and sites (left) are shown. The genome-wide average methylation rate is displayed in the top track ('All'). The main clusters of variable sites are indicated on the right. **(d)** Variance of sites located in different genomic contexts. The shaded gray region indicates the interquartile range for all genome-wide sites.

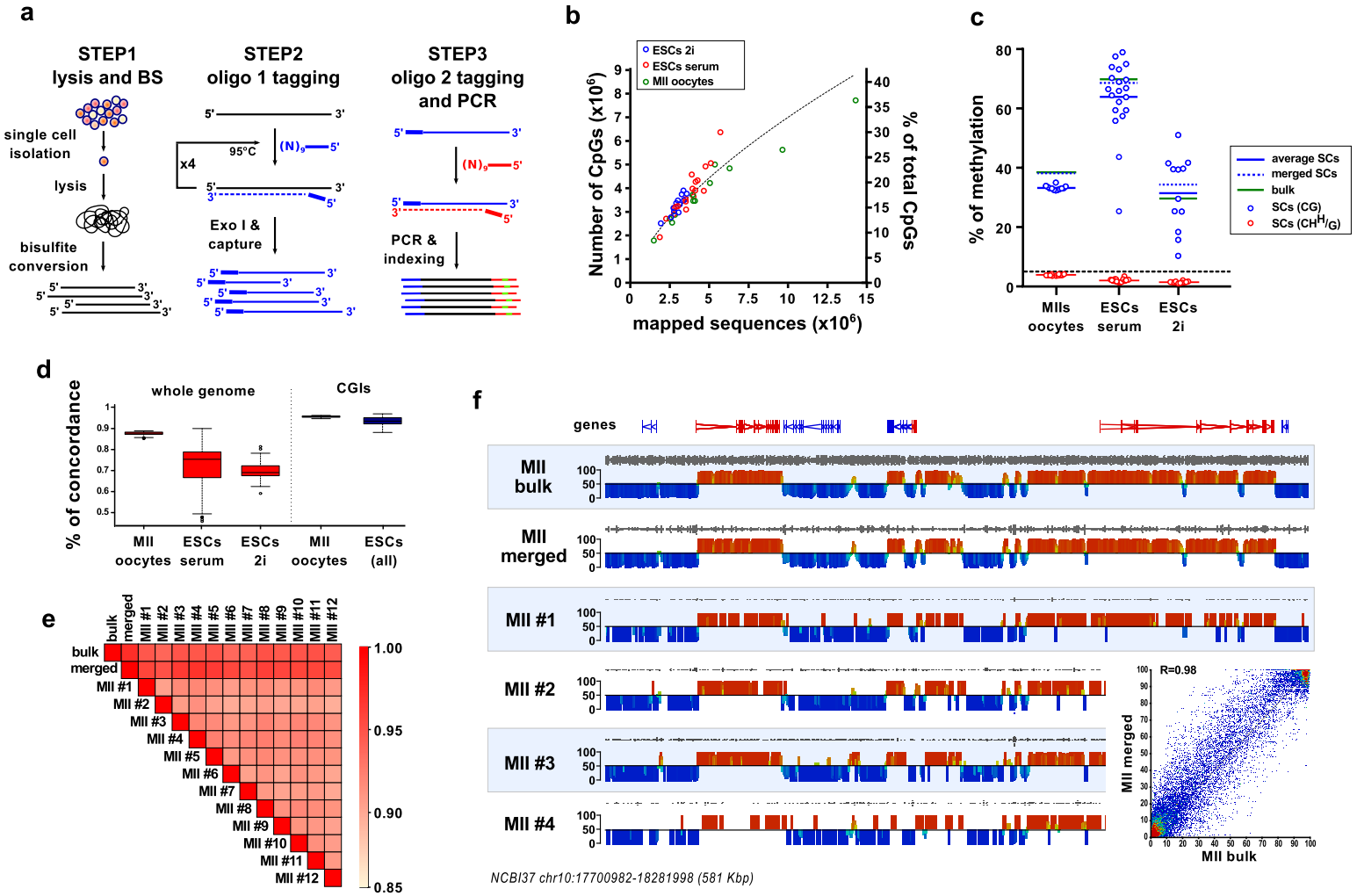

a
STEP1
lysis and BS
STEP2
oligo 1 tagging
STEP3
oligo 2 tagging
and PCR
single cell
isolation
lysis
bisulfite
conversion
95°C
x4
Exo I &
capture
PCR &
indexing
b
ESCs 2i
ESCs serum
MII oocytes
Number of CpGs (x10^6)
% of total CpGs
mapped sequences (x10^6)
c
% of methylation
average SCs
merged SCs
bulk
SCs (CG)
SCs (CH^H/G)
MIIs
oocytes
ESCs
serum
ESCs
2i
d
whole genome
CGIs
% of concordance
MII
oocytes
ESCs
serum
ESCs
2i
MII
oocytes
ESCs
(all)
e
bulk
merged
MII #1
MII #2
MII #3
MII #4
MII #5
MII #6
MII #7
MII #8
MII #9
MII #10
MII #11
MII #12
1.00
0.95
0.90
0.85
f
genes
MII
bulk
MII
merged
MII #1
MII #2
MII #3
MII #4
R=0.98
MII merged
MII bulk
NCBI37 chr10:17700982-18281998 (581 Kbp)

Figure 1

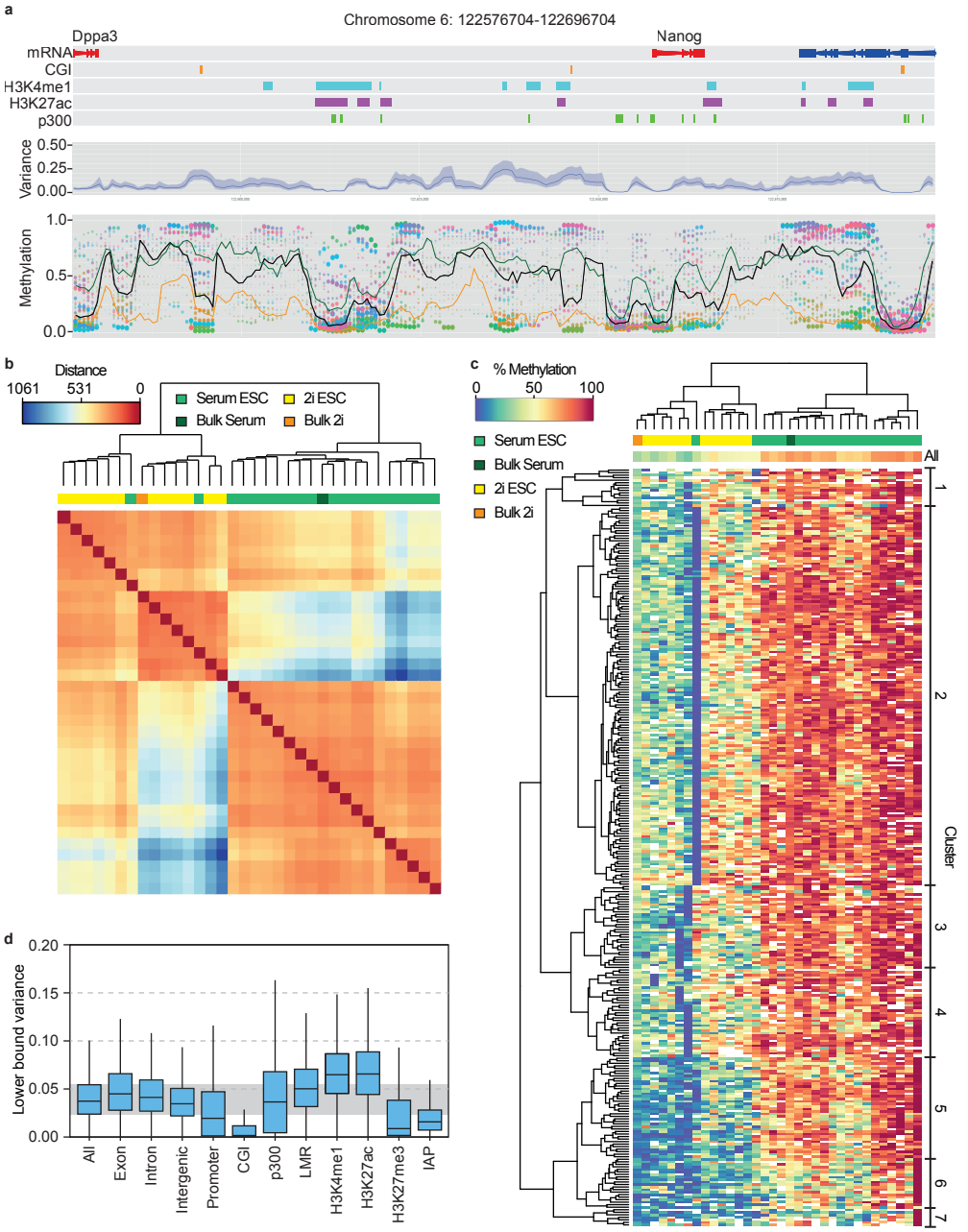

a
Chromosome 6: 122576704-122696704
Dppa3
Nanog
mRNA
CGI
H3K4me1
H3K27ac
p300
Variance
0.50
0.25
0.00
Methylation
1.0
0.5
0.0
b
Distance
1061
531
0
Serum ESC
2i ESC
Bulk Serum
Bulk 2i
c
% Methylation
0
50
100
Serum ESC
Bulk Serum
2i ESC
Bulk 2i
All
Cluster
1
2
3
4
5
6
7
d
Lower bound variance
0.20
0.15
0.10
0.05
0.00
All
Exon
Intron
Intergenic
Promoter
CGI
p300
LMR
H3K4me1
H3K27ac
H3K27me3
IAP

Figure 2

Single-Cell Genome-Wide Bisulfite Sequencing for Analysis of Epigenetic Heterogeneity

Sébastien A Smallwood[1]*, Heather J Lee[1,5]*, Christof Angermueller[2], Felix Krueger[3], Heba Saadeh[1], Julian Peat[1], Simon R Andrews[3], Oliver Stegle[3], Wolf Reik[1,4,5]#, Gavin Kelsey[1,4]#.

*,# These authors contributed equally

SUPPLEMENTARY INFORMATION.

Supplementary Figure 1: (a) Mapping efficiency of scBS-Seq samples and negative controls. Boxplot representation of the mapping efficiencies for each single cell and negative control (red crosses represent individual cell values). The overall higher mapping efficiency of oocytes versus ESCs can be explained by the amount of DNA in each cells (4n for MII oocytes and 2n for ESCs), resulting in a relatively lower contribution of spurious sequences in MIIs (see Supplementary Fig. 2). All negative controls had less than 3.5% mapping efficiency (the dashed line indicates 5% mapping efficiency). **(b) Visualization of scBS-Seq library fragment size distribution on the Bioanalyser platform**. The Bioanalyser trace of library MII#1 is shown as an example.

Supplementary Figure 2: Contribution of spurious sequences to scBS-Seq mapping efficiency.
(a) The relatively low mapping efficiency of scBS-Seq is associated with a significant fraction of sequences mapping at multiple genomic locations, which are therefore discarded. **(b)** Analysis of the G+C content of the raw sequences (i.e. prior to mapping) of scBS-Seq libraries revealed many with <3% G+C, absent from bulk samples. These correspond to poly-T stretches (poly-Ts) (i.e., $(T)_N$ with N>50). Poly-Ts are present in both actual samples and corresponding negative controls suggesting a contaminant as their main source of origin. **(c,d)** The amount of poly-Ts is higher in ESCs than oocytes, and the percentage of sequences with poly-Ts and sequences with multiple alignments are tightly correlated across samples. **(e)** This suggests that poly-Ts are the major cause for scBS-Seq low mapping efficiency. To test this, we trimmed, from the raw fasq file, sequences containing poly-Ts of at least 50bp in size and repeated the mapping. This resulted in a drastic reduction in the percentage of sequences with multiple alignments and an increase in the percentage of sequences with unique alignments. Poly-Ts are inherent to our current methodology, and while alternative protocols we developed do not generate these artifacts, they still yield significantly fewer measured CpGs.

Supplementary Figure 3: Saturation level of scBS-Seq libraries.
For each individual MII scBS-Seq library and one representative example of bulk BS-Seq (PBAT), the percentage of informative CpGs is plotted for 10% increments of mapped sequences. This demonstrates that in contrast to the bulk BS-Seq example (black line), MIIs scBS-Seq libraries (colored lines) have not reached the plateau of saturating sequencing depth, indicating that further sequencing would yield additional information. MII#2 Deep Seq and MII#5 Deep Seq correspond to the deeper sequencing of these libraries (see main text and Supplementary Table 1).

Supplementary Figure 4: scBS-Seq generates a digital output of DNA methylation.

(a) For each single MII BS-Seq library, and for the bulk MII sample, CpGs were grouped based on their read depth. The proportion of CpGs in each group with a methylation value of either 0% or 100% (digital output) was calculated for each sample. The boxplot represents the results from all 12 single MII libraries. The results from the bulk MII sample are superimposed as solid blue circles. As expected, the proportion of digital CpGs in the scBS-Seq libraries was very high (>90% for read depth 2-5 in all cells, dashed line). In contrast, the bulk sample had fewer digital CpGs (66% at read depth 5) due to cell-to-cell variability within the population. **(b)** Histograms of the distribution of CpG methylation values for MII bulk and MII single cells for CpGs with at least 2 reads.

Supplementary Figure 5: CpG concordance obtained from MIIs and ESCs using scBS-Seq.
(a) CpG concordance was calculated for each cell pair as the proportion of overlapping CpGs with identical methylation state. On average, 1.8M CpGs were measured for each pairwise analysis. Within each cell types, the order from bottom – up is the same than in Supplementary Table1 (For oocytes bottom sample is MII#1 and top sample is MII#12). **(b)** Pearson correlation matrix of MIIs, 2i ESCs and serum ESCs scBS-Seq was calculated using 2kb window methylation values.

Supplementary Figure 6: scBS-Seq accurately determines CpG island (CGI) methylation status in MII oocytes.
(a) Heatmap displaying in individual MII libraries the methylation level of CGIs identified as methylated (>80%) and unmethylated (<20%; random selection) in bulk. The number on top indicates the number of individual MIIs in which CGIs are commonly informative. The discrepancy between the number of methylated and unmethylated CGIs informative across single cells reflects the different CpG density between these 2 groups as previously described[19]. **(b)** Histogram displaying for MII bulk and individual MII libraries the percentage of total CGIs (23,020) found methylated, unmethylated, with an intermediate level of methylation, and the percentage of wrong calls (i.e., CGI methylated in bulk (>80%) and called unmethylated (<20%) in single cells, and *vice versa*). **(c)** Boxplot presenting the methylation level in each individual MII of CGIs found methylated in bulk (>80%). The percentage of these CGIs informative in each MII with a methylation level lower than 80% is shown below the plot. **(d)** Similar to (c) for unmethylated CGIs (<20%).

Supplementary Figure 7: scBS-Seq provides information on all genomic contexts.
(a) Snapshot displaying read distribution across 61Mbp of chromosome 19. Below the annotation tracks are displayed the mapped reads and the quantification (number of reads per 25kb window (log)). **(b)** The representation of different genomic contexts in single cell and bulk libraries is shown as fold enrichment over the expected value (dashed line). The boxplot represents the values for all single cell samples, and the bulk samples are superimposed as blue diamonds (MII), purple crosses (serum ESCs) and red plus signs (2i ESCs).

Supplementary Figure 8: Union and intersect for scBS-Seq libraries.
Number of CpGs **(a)** and CGIs **(b)** for the union and intersect of all possible combinations of the 12 individual MII scBS-Seq libraries. The union shows that pooling data from multiple scBS-Seq samples increases the number of measured sites. The intersect shows that the number of measured sites common to multiple scBS-Seq datasets decreases as the number of datasets increases. Dotted lines show the information obtained in standard BS-Seq experiments as well as the number of CpGs and CGIs in the mouse genome.

Supplementary Figure 9: scBS-Seq snapshot of the imprinted locus Plagl1.
The imprinted *Plagl1* locus (top) and *Plagl1* maternal DMR/CGI (bottom) is shown for all 12 individual MIIs, MIIs merged and MII bulk. Quantification is absolute level of methylation (%), at individual CpG resolution, as indicated on the scale on the left of each sample (0 is 0% methylation, 1 is 100% methylation).

Supplementary Figure 10: Comparison of cluster analyses for ESCs.
Cluster dendrograms are shown for **(a)** genome-wide methylation estimates (equivalent to the dendrogram shown in Figure 2b) and **(b)** the top 300 most variable sites among single ESC samples (equivalent to the dendrogram shown in Figure 2c). The cell IDs are included for direct comparison between dendrograms. **(c)** The distance matrix for the 300 most variable sites is grossly similar to that for all sites (Figure 2b). Cells are presented in the order shown in (b).

Supplementary Figure 11: Cluster dendrogram and distance matrix for the most variable sites in ESCs.
The top 300 ranked most variable sites in ESCs show similar methylation patterns across ESCs, as indicated by the low distance between sites. ~~The clusters indicated below the distance matrix correspond to those in Figure 2c.~~

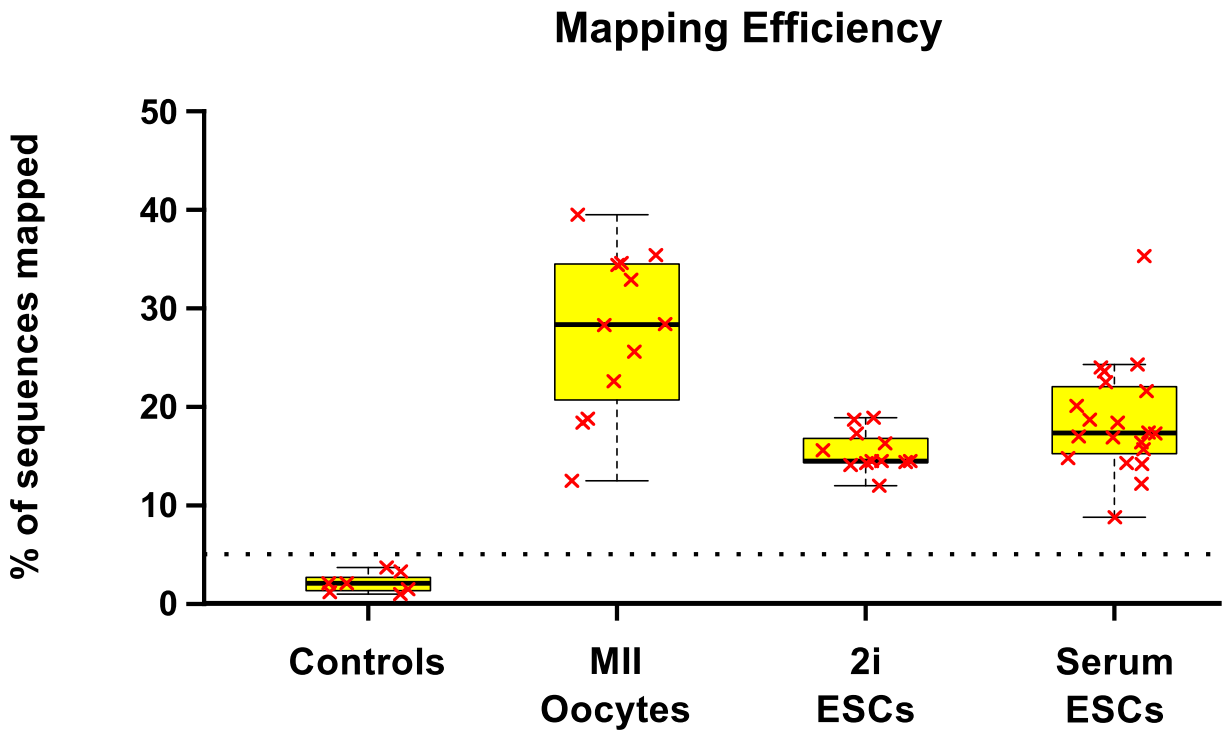
Supplementary Figure 12: Detailed variance analysis for different genomic contexts.
(a) Receiver Operating Characteristic (ROC) curves showing the fraction of annotated sites (sensitivity) versus the fraction of non-annotated sites (1-specificity). Sites with high variance are more likely to belong to a given genomic context if the ROC curve is above the diagonal (e.g.H3K4me1), and less likely to belong to genomic contexts if the ROC curve is below the diagonal (e.g. CGI). **(b)** Different genomic contexts have different mean methylation values. **(c)** For most genomic contexts, variance was greatest for sites with mean methylation rates close to 50%. H3K27ac and H3K4me1 sites were among the most variable, even after accounting for mean methylation rate. CGI and p300 sites with intermediate mean methylation rates were also highly variable.

Supplementary Figure 13: Comparison of scRRBS and scBS-Seq in MII oocytes.
(a) Summary table showing the number of raw sequences, informative CpGs and CGIs. For scRRBS, the number of CpG dinucleotides and the number of informative CGIs were calculated using the methylation calls present in the .bed file of GEO accession number GSE47343 from Guo *et al*.[18]. **(b)** Plots showing the number of raw sequences generated and the corresponding number of CpGs obtained in MII oocytes for both methods.

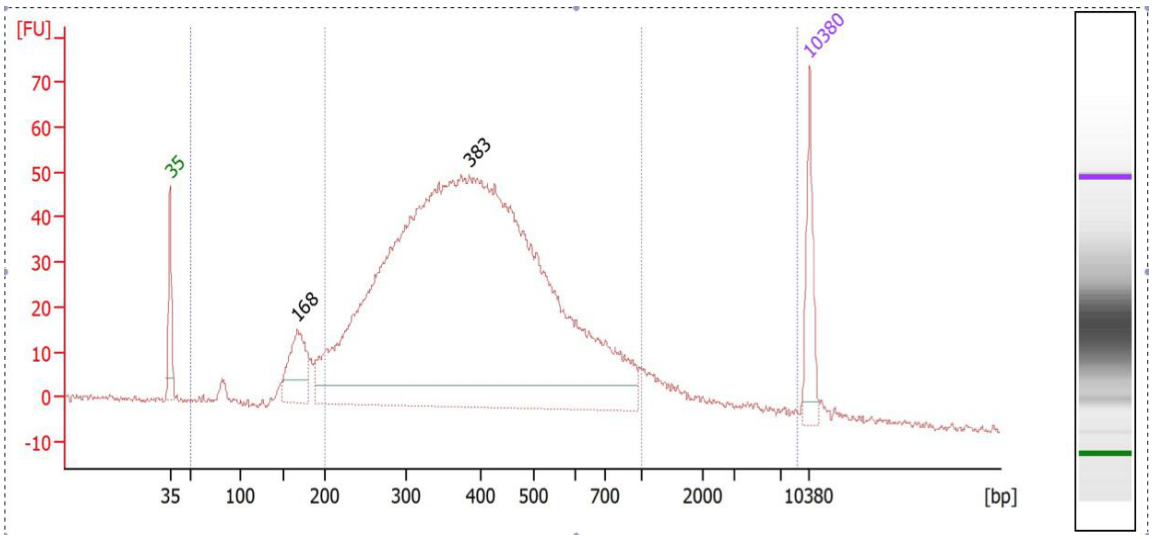
Supplementary Table 1: Information on sequencing, level of methylation and number of informative CpGs for all scBS-Seq samples, negative controls and bulk samples.

Supplementary Table 2: Representation of regulatory regions in ESC scBS-Seq datasets.
The number and proportion of CGIs, promoters, LMRs, H3K4me1, p300, H3K27ac, H3K27me3 and IAPs covered by at least 5 CpG sites is given for each ESC scBS-Seq and Bulk dataset.

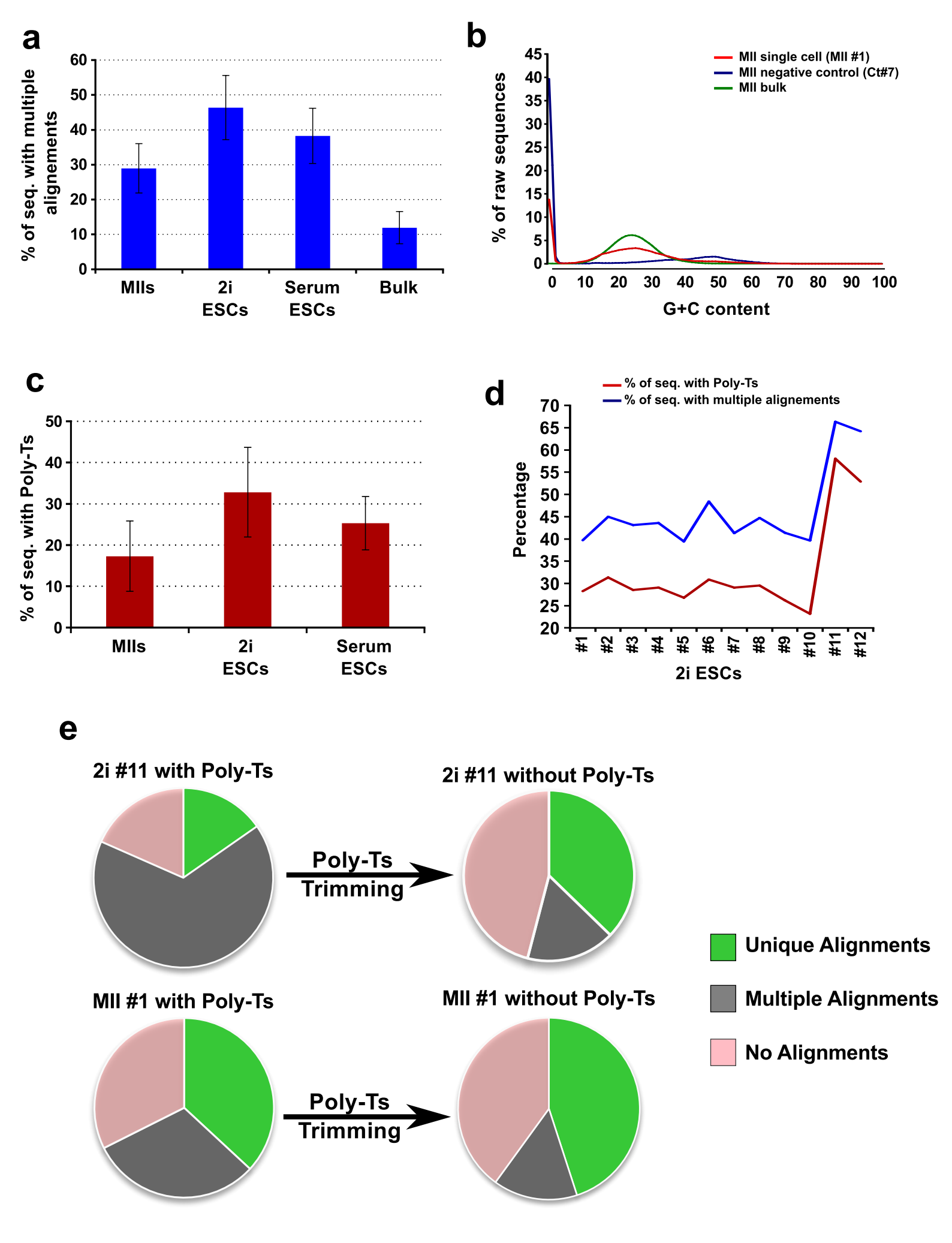
a



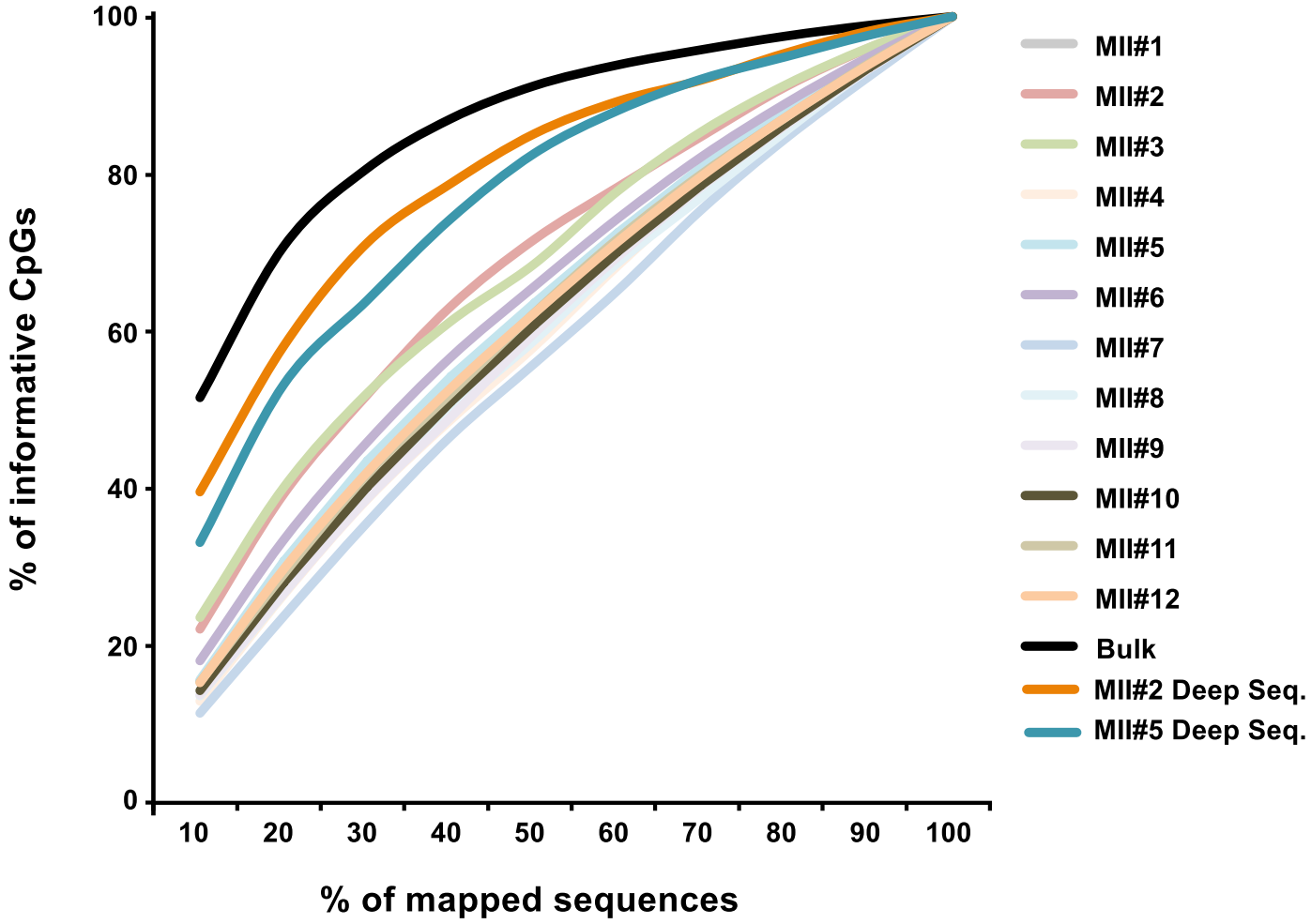
b



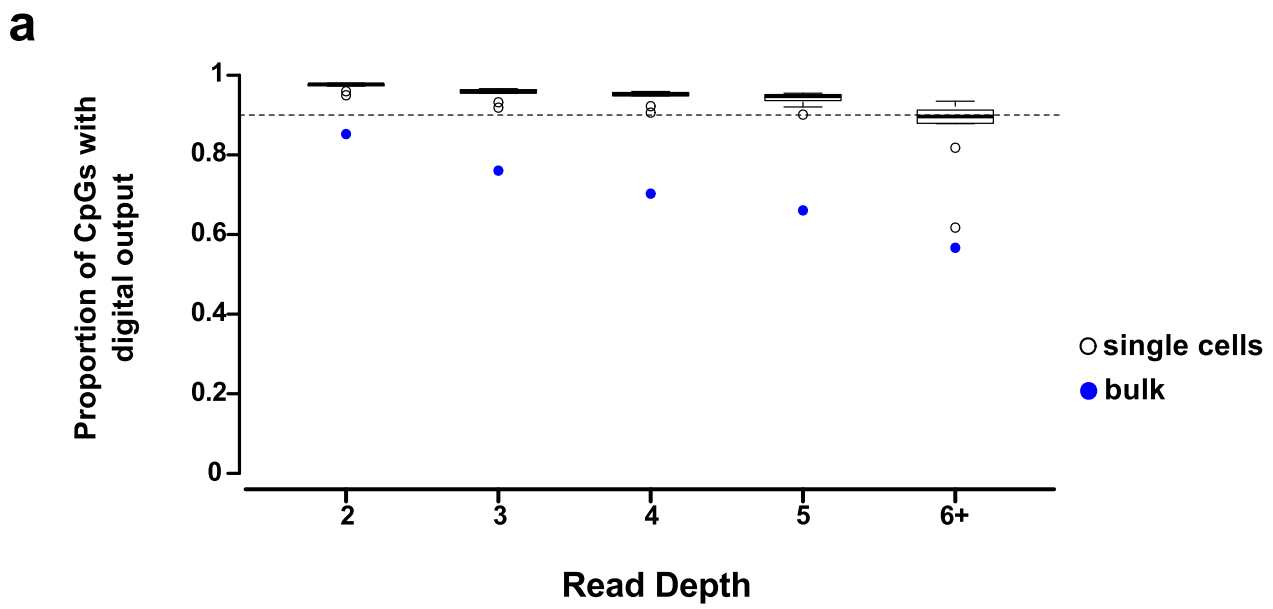
Supplementary Figure 1

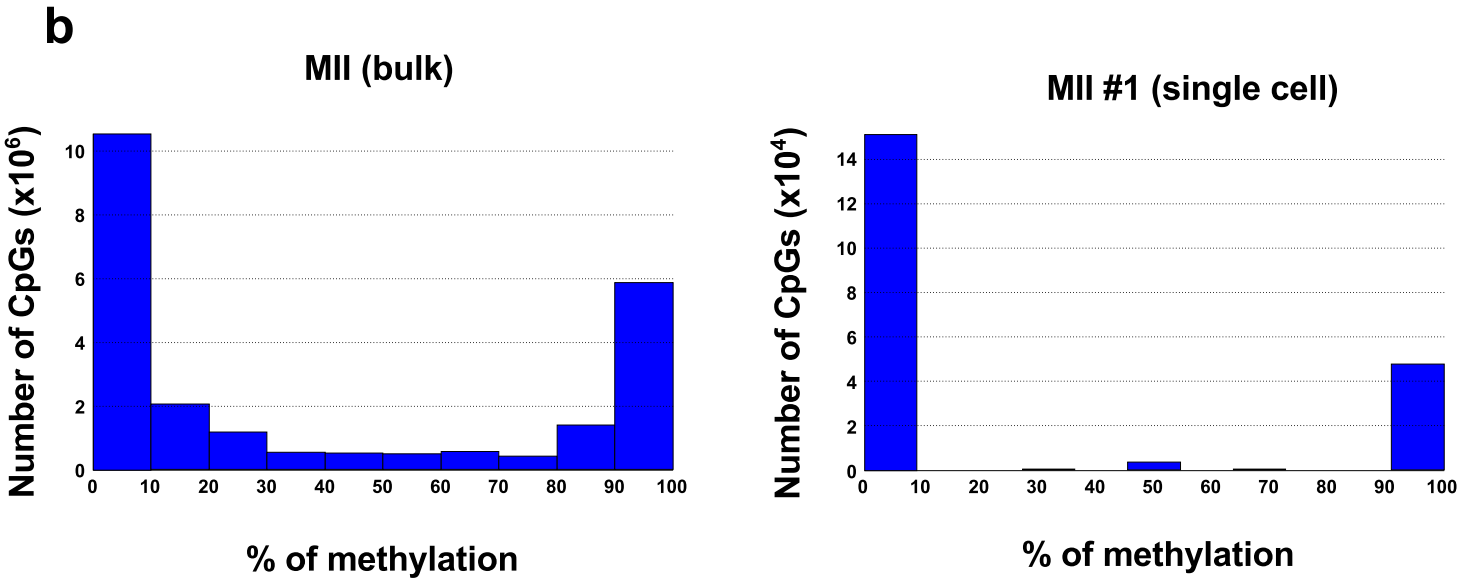


Supplementary Figure 2

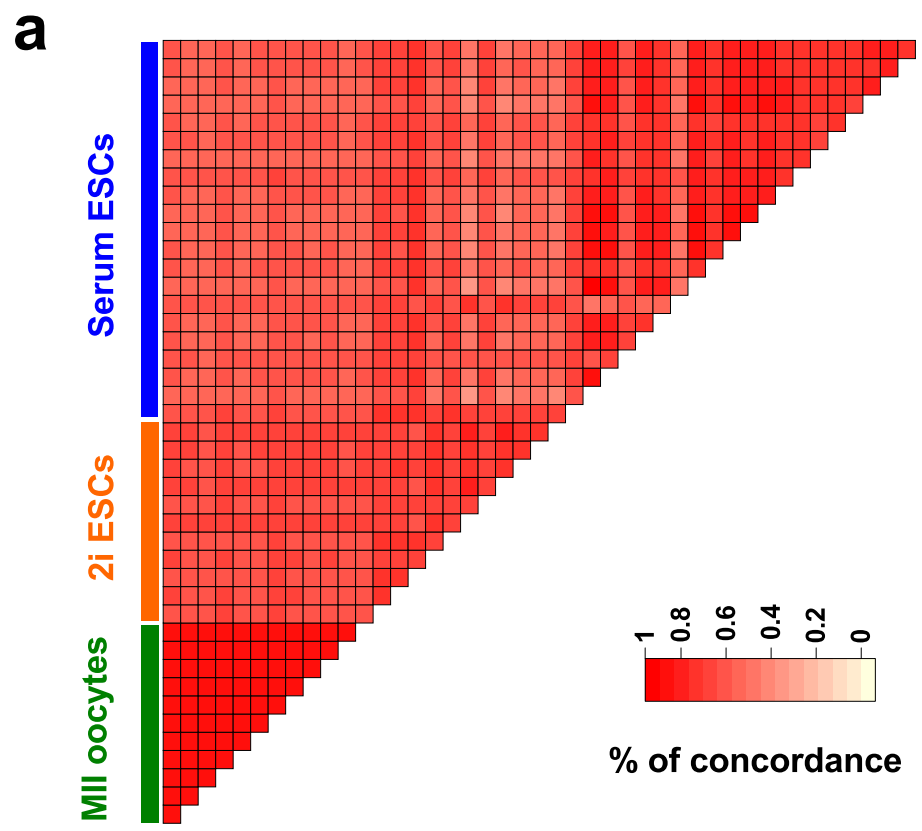


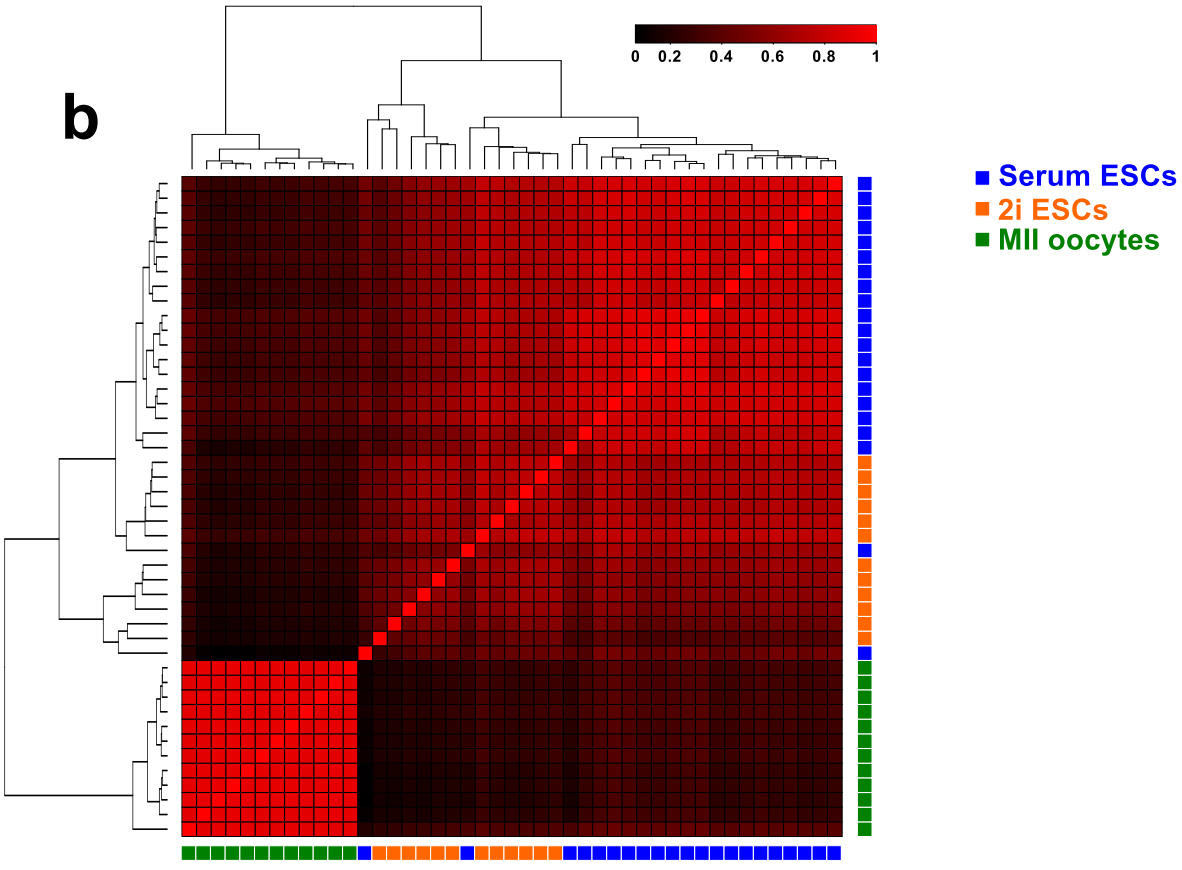
Supplementary Figure 3



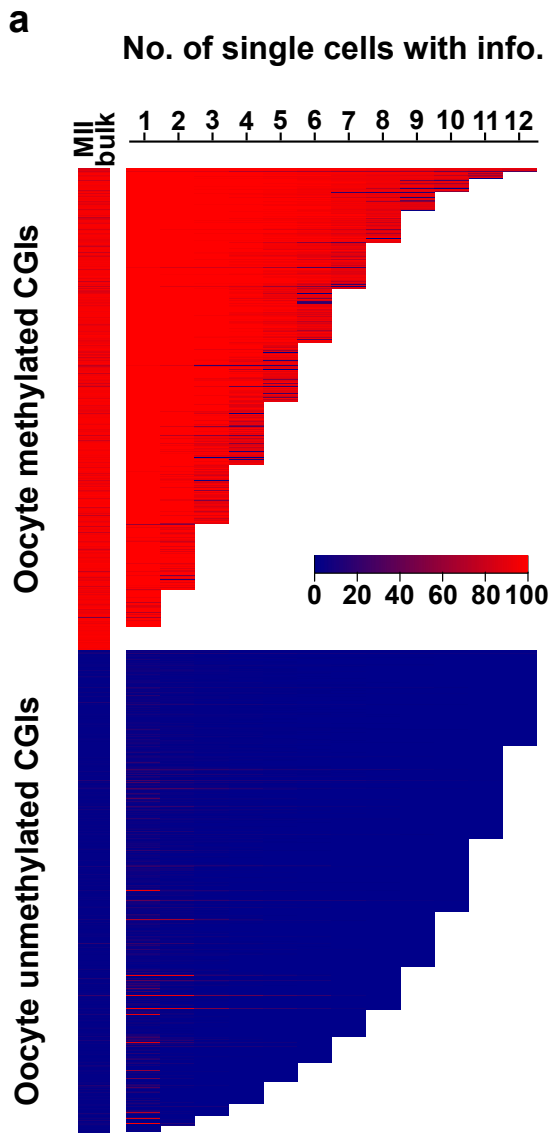


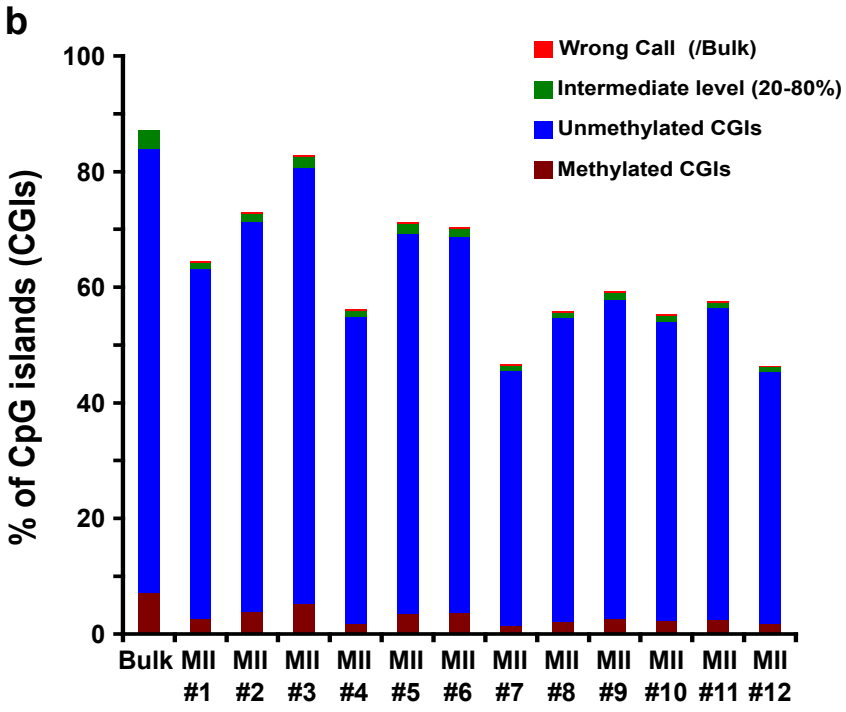
Supplementary Figure 4

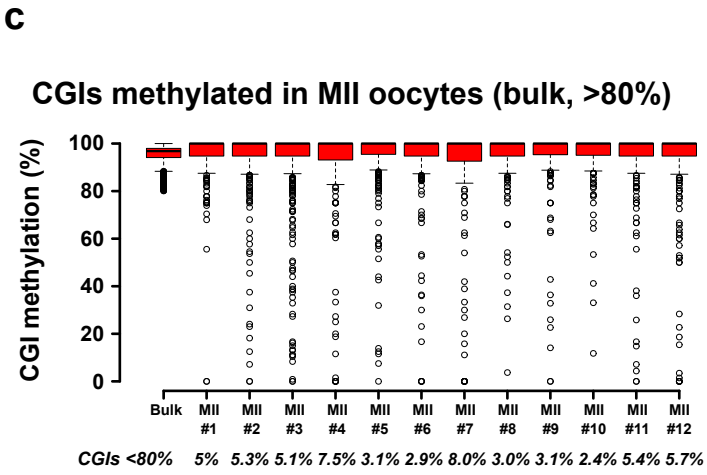


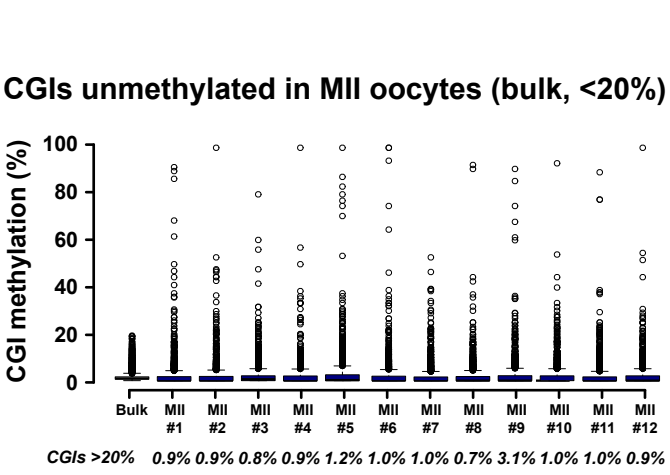


Supplementary Figure 5



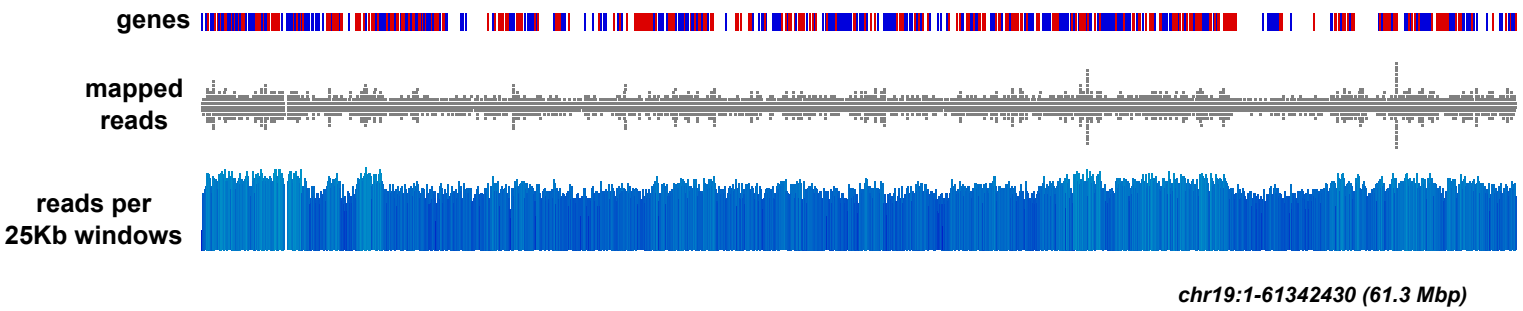




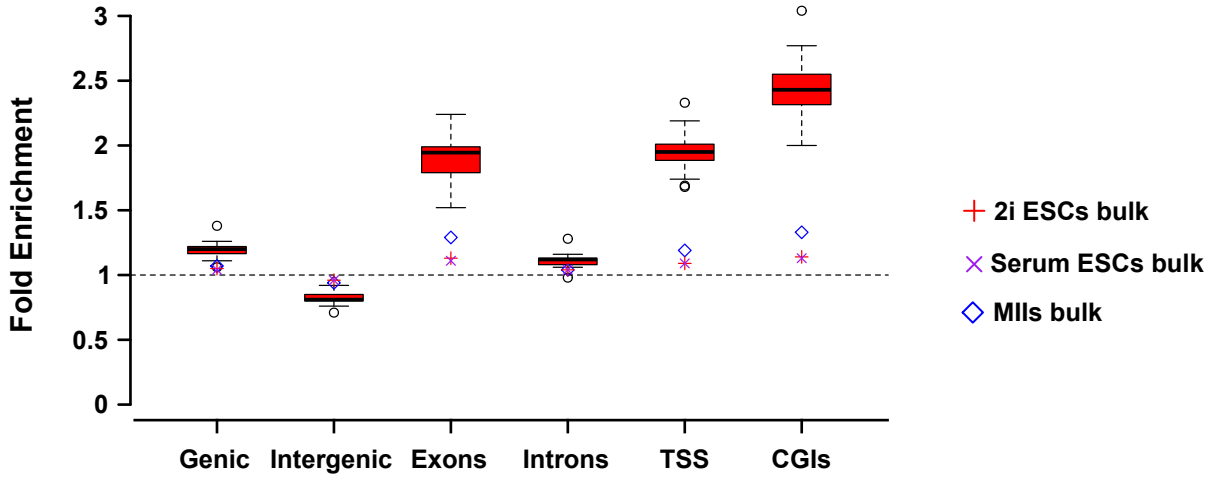


Supplementary Figure 6

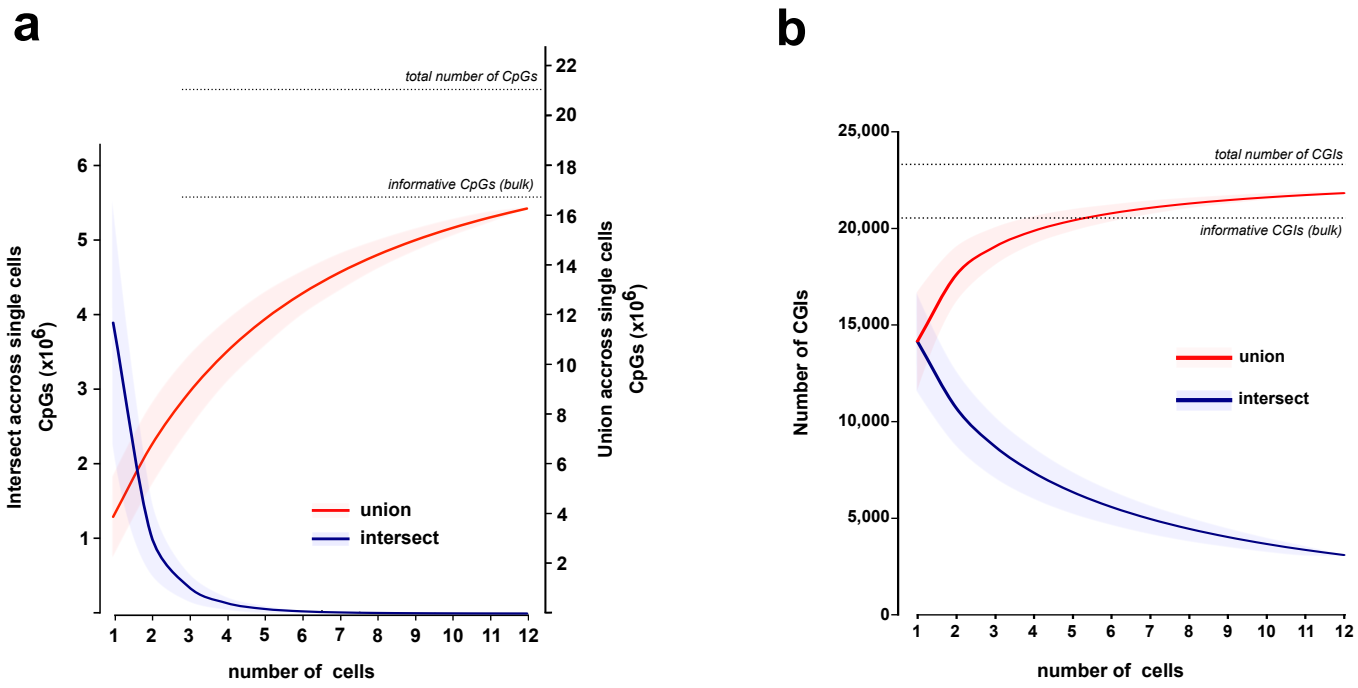
a



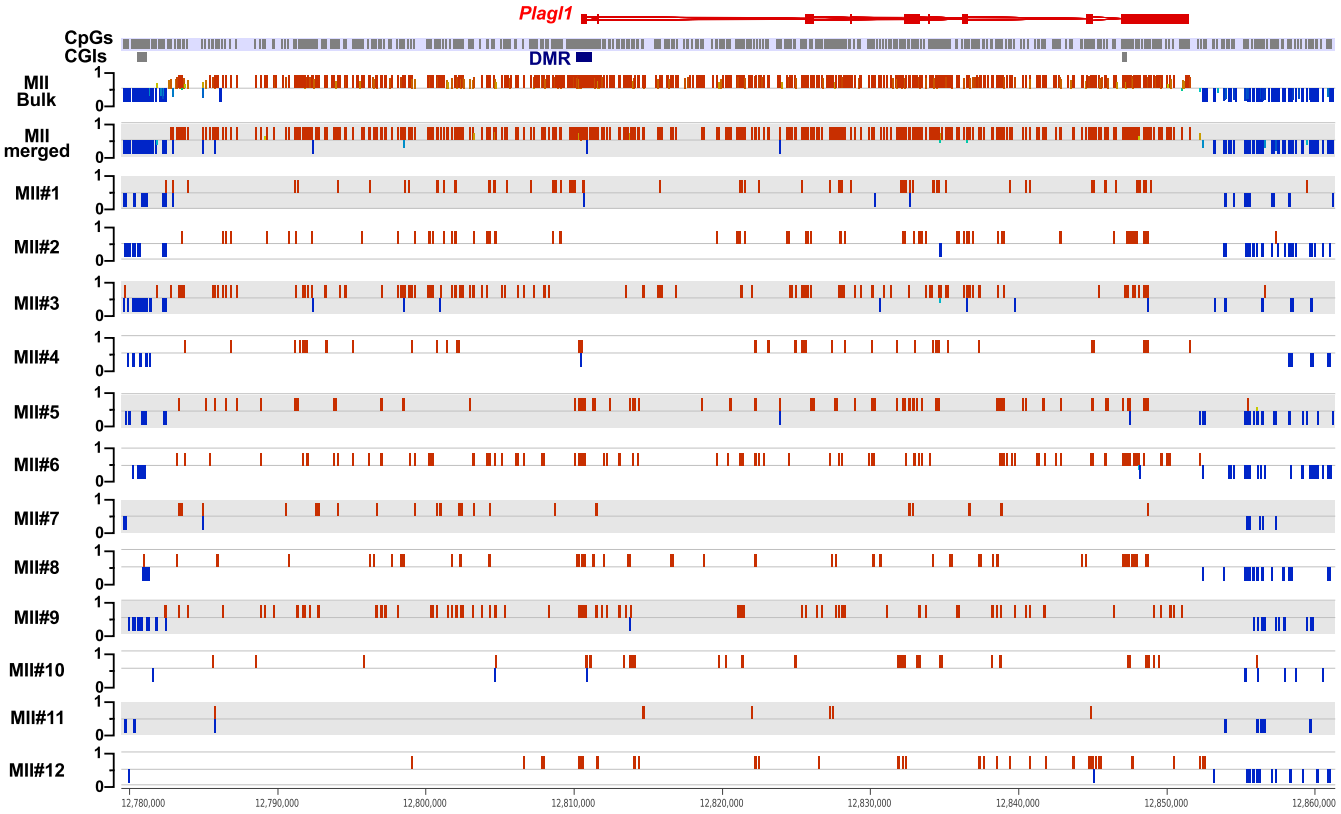
b

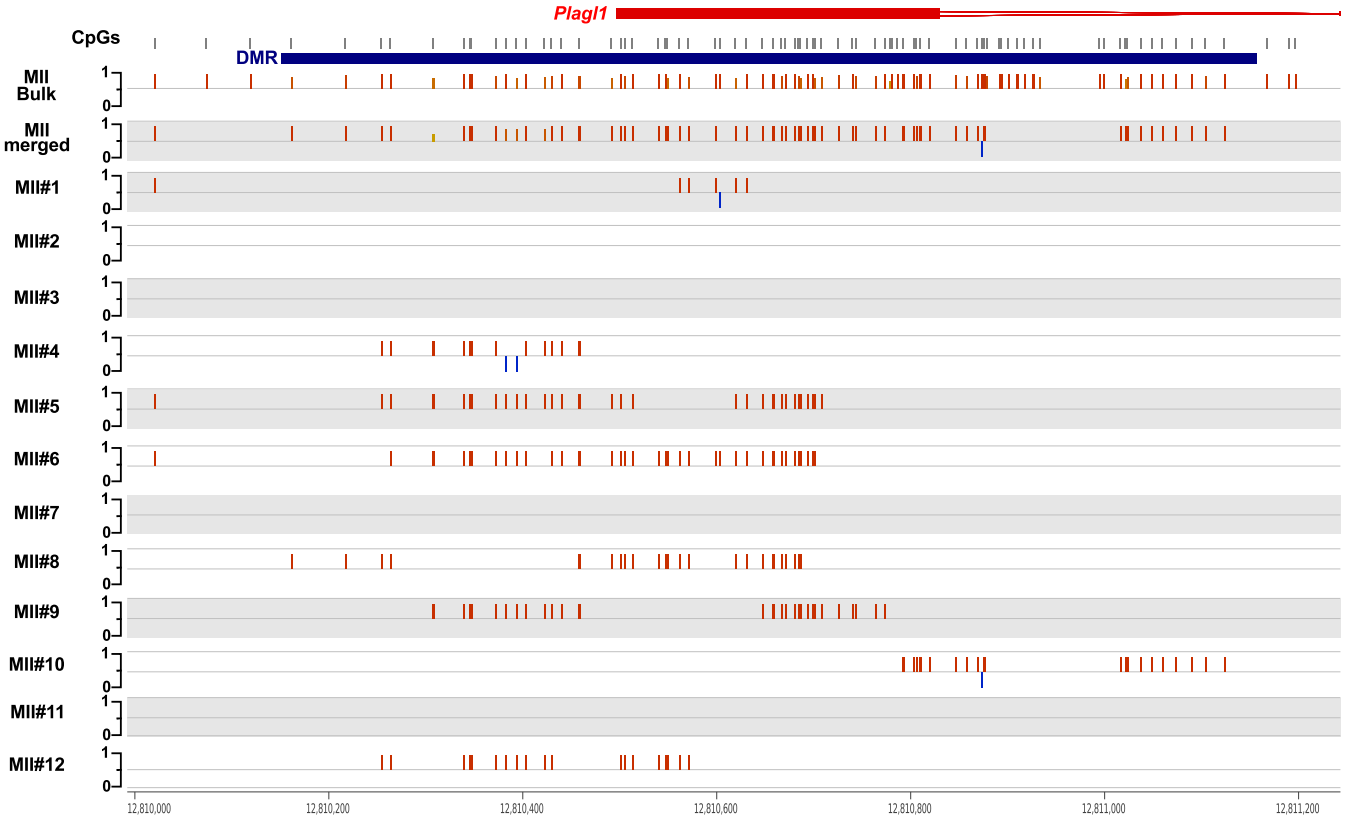


Supplementary Figure 7

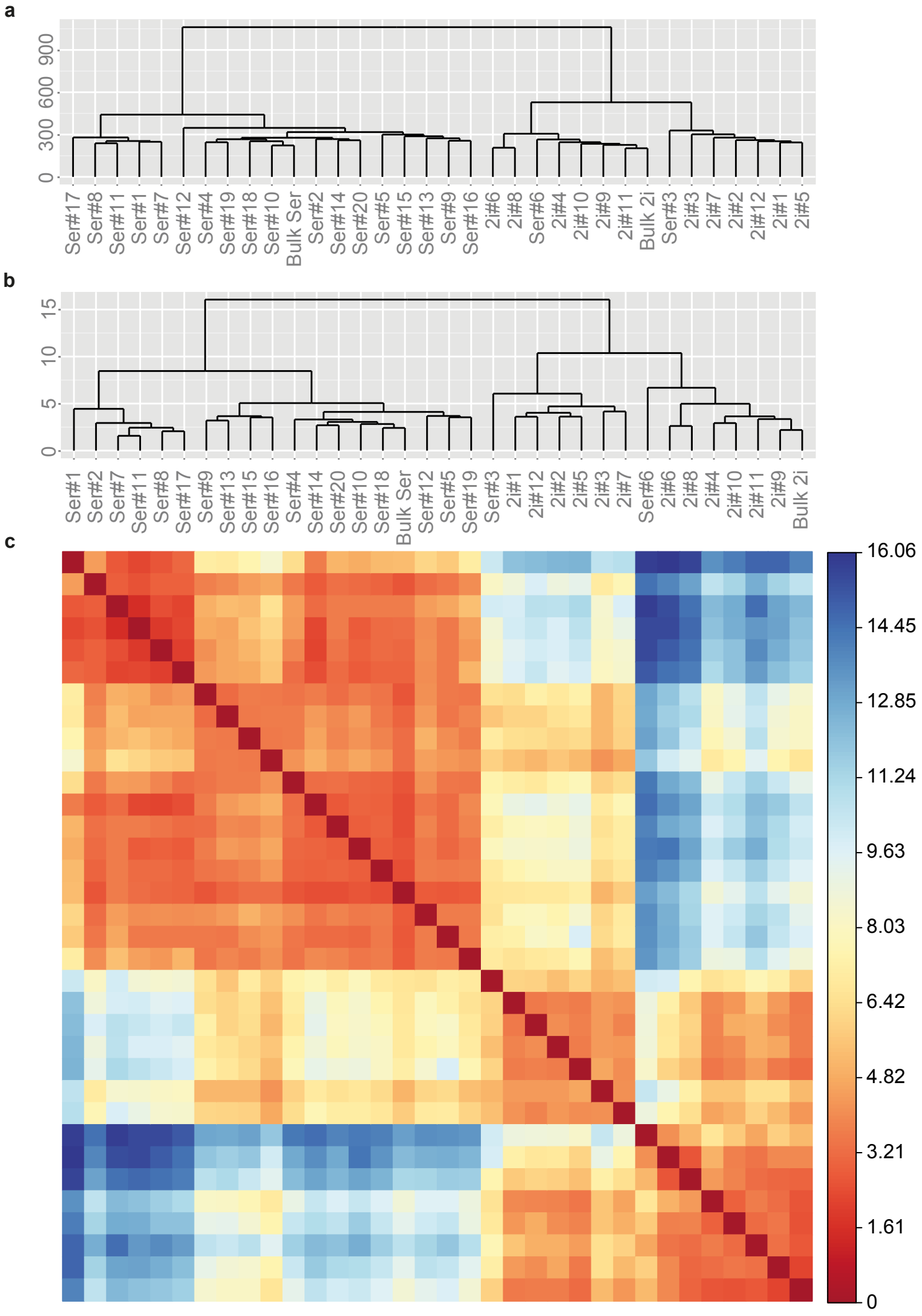


Supplementary Figure 8

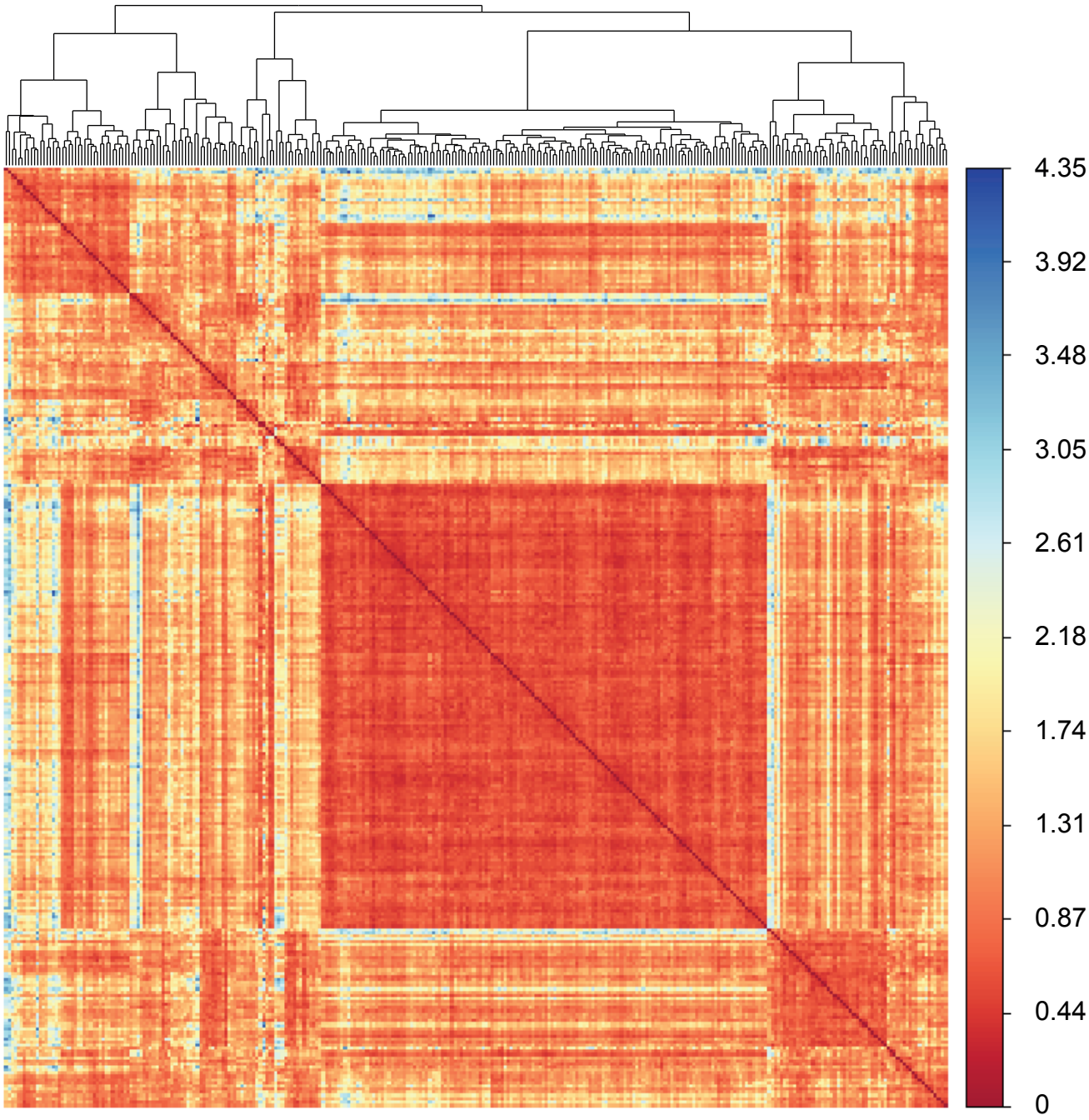




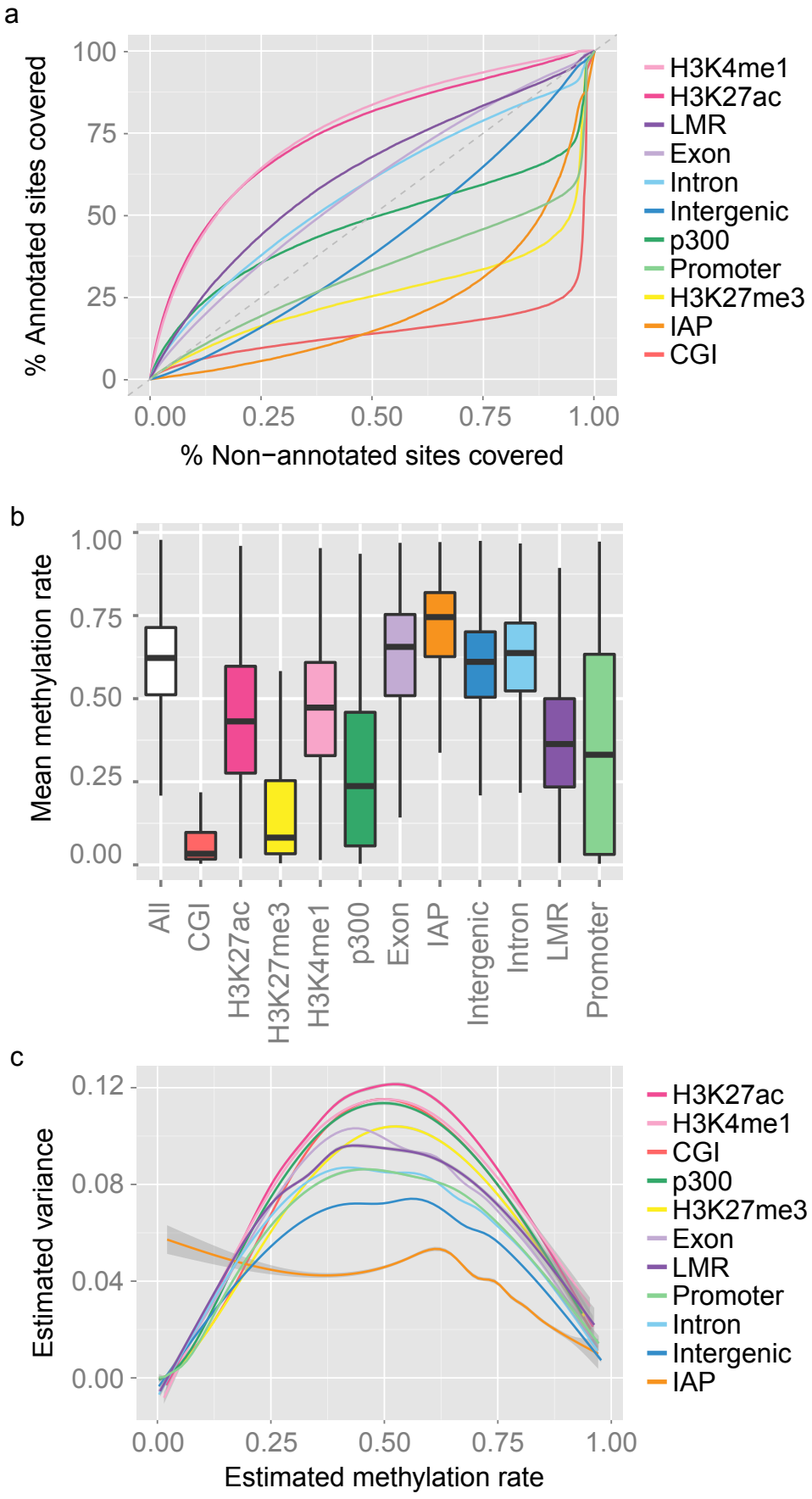
Supplementary Figure 9



Supplementary Figure 10

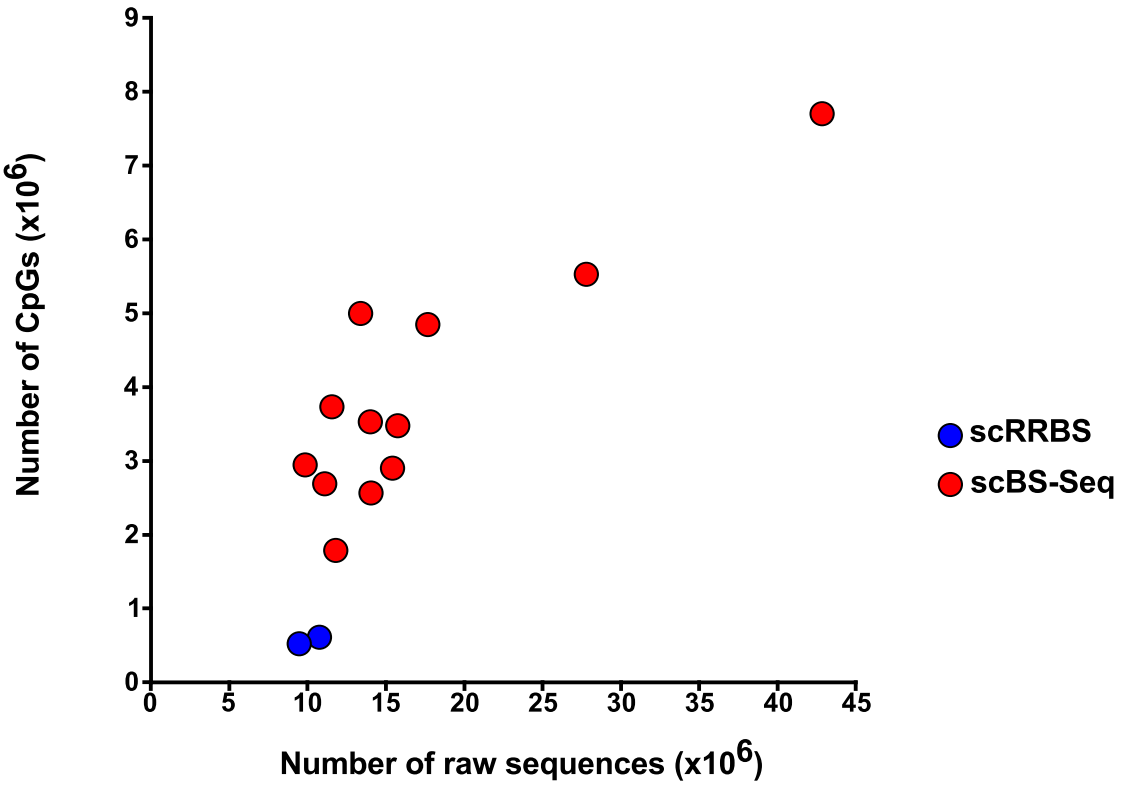


Supplementary Figure 11



Supplementary Figure 12

Technique	Sample	Raw sequences	Number of CpGs	Number of informative CGIs (5CpGs) [1]	Number of informative CGIs (10CpGs) [2]
scRRBS	MII_oocyte1	9,572,299	540,260	9,860	7,563
scRRBS	MII_oocyte2	10,722,272	623,080	9,737	7,215
scBS-Seq	MII #1	11,526,952	3,750,723	17,490	14,832
scBS-Seq	MII #2	27,712,173	5,567,568	19,055	16,781
scBS-Seq	MII #3	42,830,232	7,726,619	20,938	19,056
scBS-Seq	MII #4	11,085,152	2,737,377	15,780	12,896
scBS-Seq	MII #5	13,370,171	4,998,613	18,888	16,370
scBS-Seq	MII #6	17,640,105	4,852,784	18,616	16,178
scBS-Seq	MII #7	11,828,558	1,808,915	13,528	10,704
scBS-Seq	MII #8	9,797,740	2,938,955	15,570	12,821
scBS-Seq	MII #9	13,921,551	3,496,702	16,232	13,617
scBS-Seq	MII #10	15,447,641	2,899,725	15,489	12,705
scBS-Seq	MII #11	14,038,511	2,578,951	13,387	13,242
scBS-Seq	MII #12	15,544,734	3,482,271	15,917	10,654



Supplementary Figure 13

Supplementary Table 1

ID	Sample	Raw Seq.	Seq. For Mapping[1]	Seq. Mapped	% Mapped	% Duplication[2]	Nb. CpGs	% of total CpGs	% mCpG	% mCpG chr.MT	% mCHH/mCHG[2]	Nb. $CH^H/_G$
MII#1	MII oocyte	11,526,952	10,888,102	4,018,891	34.9	31.8	3,750,723	17.6	32.4	-	3.6	76,790,741
MII#2	MII oocyte	27,712,173	25,911,001	9,740,275	35.1	47.2	5,567,568	26.1	32.9	-	3.6	121,863,785
MII#3	MII oocyte	42,830,232	38,070,664	14,311,873	33.4	45.0	7,726,619	36.3	34.9	-	3.8	179,003,920
MII#4	MII oocyte	11,085,152	9,765,795	2,543,732	22.9	27.4	2,737,377	12.9	32.2	-	3.8	54,085,393
MII#5	MII oocyte	13,370,171	12,641,680	5,365,931	40.1	26.4	4,998,613	23.5	33.4	-	3.8	107,137,487
MII#6	MII oocyte	17,640,105	16,739,216	6,358,367	36.0	37.5	4,852,784	22.8	33.8	-	3.8	103,505,143
MII#7	MII oocyte	11,828,558	9,021,093	1,474,690	12.5	25.4	1,808,915	8.5	32.7	-	4.2	33,863,808
MII#8	MII oocyte	9,797,740	8,817,125	2,786,462	28.4	27.1	2,938,955	13.8	32.4	-	3.8	58,138,180
MII#9	MII oocyte	13,921,551	11,206,678	3,940,050	28.3	31.9	3,496,702	16.4	33.0	-	4.2	71,394,196
MII#10	MII oocyte	15,447,641	12,838,519	2,825,097	18.3	28.8	2,899,725	13.6	32.9	-	4.2	57,533,300
MII#11	MII oocyte	14,038,511	12,025,091	2,637,416	18.8	28.1	2,578,951	12.1	33.6	-	3.7	53,180,806
MII#12	MII oocyte	15,544,734	12,162,655	3,980,617	25.6	32.6	3,482,271	16.3	32.9	-	3.5	71,679,934
Bulk MII*	MII oocyte	874,735,536	874,735,536[a]	451,714,706	51.6	38.0	17,302,720	81.2	39.0	6.9	3.9	820,485,200
MII#2 deep 100bp	MII oocyte	54,185,479	49,481,208	19,158,847	35.4	62.6	6,837,514	32.6	33.6	-	3.5	156,542,649
MII#2 deep 150bp	MII oocyte	54,185,479	49,464,979	17,415,934	32.1	60.8	7,527,693 (8,361,588) #	35.8 (39.8) #	34.2	-	3.9	178,001,406 (202,092,274)
MII#5 deep 100bp	MII oocyte	49,015,151	45,213,029	19,574,720	39.9	52.1	8,609,618	41.0	35.4	-	3.9	207,292,025
MII#5 deep 150bp	MII oocyte	49,015,151	45,185,861	17,743,442	36.2	50.5	9,461,486 (10,155,982) #	45.1 (48.4) #	36.0	-	4.0	235,657,882 (257,986,613)
2i#1	2i ESC	20,121,114	15,289,712	2,944,748	14.6	13.0	3,385,387	15.9	41.4	1.9	1.4	67,926,519
2i#2	2i ESC	11,230,949	9,146,963	1,956,177	17.4	10.8	2,543,593	11.9	39.3	1.3	1.4	49,414,224
2i#3	2i ESC	19,654,356	15,285,350	2,782,814	14.2	16.9	3,052,094	14.3	51.0	2.7	1.7	61,138,407
2i#4	2i ESC	22,995,697	17,787,837	3,605,918	15.7	16.7	3,794,177	17.8	29.4	1.2	1.5	77,565,269
2i#5	2i ESC	23,458,154	18,974,001	3,441,439	14.7	17.6	3,515,813	16.5	39.6	1.0	1.3	71,579,160
2i#6	2i ESC	25,434,842	19,302,532	3,066,821	12.1	20.4	2,998,026	14.1	10.1	1.5	2.1	60,462,444
2i#7	2i ESC	19,978,754	16,169,870	3,269,980	16.4	12.7	3,766,721	17.7	39.4	1.7	1.3	76,255,902
2i#8	2i ESC	15,966,034	12,678,501	3,026,791	19.0	13.4	3,510,739	16.5	15.5	2.3	1.3	69,570,951
2i#9	2i ESC	18,065,844	14,605,593	3,387,244	18.7	13.2	3,925,357	18.4	25.2	1.3	1.5	79,508,007
2i#10	2i ESC	21,749,327	16,332,840	3,147,704	14.5	16.3	3,343,369	15.7	25.1	2.1	1.7	66,980,939
2i#11	2i ESC	19,066,379	15,720,901	2,786,491	14.6	14.1	3,185,725	15.0	18.2	1.4	1.2	55,621,101
2i#12	2i ESC	17,740,157	17,094,458	2,618,865	14.8	18.4	2,796,397	13.1	41.6	1.1	1.1	64,491,228
Bulk 2i	50ng 2i ESC DNA	106210222	102,876,577[a]	42,900,227[b]	40.4	5.0	17,981,120	84.2	29.52	2.1	0.9	629,221,268
Ser#1	Serum ESC	19,430,498	16,639,883	4,678,381	24.1	23.4	3,914,706	18.4	79.0	1.3	1.8	84,385,397
Ser#2	Serum ESC	15,905,437	12,516,204	2,772,528	17.4	17.5	2,939,191	13.8	69.7	1.1	1.9	59,493,066
Ser#3	Serum ESC	22,116,506	19,392,071	5,005,230	22.6	21.9	4,203,584	19.7	43.6	1.1	1.5	88,717,640
Ser#4	Serum ESC	17,572,229	14,942,208	4,264,558	24.3	15.4	4,363,212	20.5	65.9	1.2	1.4	88,736,245
Ser#5	Serum ESC	20,725,654	15,693,723	3,530,118	17.0	18.1	3,459,164	16.2	62.2	1.5	1.8	71,445,210
Ser#6	Serum ESC	21,358,923	18,034,542	5,049,863	23.6	16.5	5,034,972	23.6	25.2	1.8	1.4	102,504,193
Ser#7	Serum ESC	22,524,548	18,906,367	3,547,019	15.7	22.4	3,121,433	14.7	77.6	1.3	2.5	66,218,273
Ser#8	Serum ESC	19,473,176	14,648,071	2,779,012	14.3	12.9	3,186,898	15.0	74.0	1.7	2.4	64,276,721
Ser#9	Serum ESC	19,387,056	15,320,411	3,884,644	20.0	10.4	4,612,155	21.7	59.4	1.8	1.5	93,448,346
Ser#10	Serum ESC	22,371,353	17,674,299	4,106,887	18.4	13.9	4,309,970	20.2	70.3	1.8	1.9	89,604,311
Ser#11	Serum ESC	21,700,471	16,533,242	3,081,947	14.2	14.9	3,298,523	15.5	75.0	1.7	2.3	68,271,355
Ser#12	Serum ESC	21,329,965	14,384,685	1,869,023	8.8	21.0	1,955,445	9.2	66.9	2.1	3.4	38,456,484
Ser#13	Serum ESC	18,620,980	13,624,798	2,267,317	12.2	13.9	2,733,166	12.8	57.4	1.5	2.4	53,435,567
Ser#14	Serum ESC	16,040,768	13,835,704	5,627,139	35.1	8.5	6,400,100	30.0	66.5	1.2	1.7	137,535,583
Ser#15	Serum ESC	16,089,156	12,546,745	2,993,816	18.6	17.8	3,263,367	15.3	59.4	1.1	1.9	65,515,607
Ser#16	Serum ESC	21,825,607	15,969,399	4,663,460	21.4	17.4	4,909,235	23.0	55.8	1.1	1.8	100,964,537
Ser#17	Serum ESC	22,770,588	16,039,208	3,908,172	17.2	20.1	3,718,731	17.5	73.2	1.1	2.2	76,787,874
Ser#18	Serum ESC	21,326,695	15,404,043	3,466,764	16.3	20.1	3,558,205	16.7	69.0	1.0	2.3	73,406,852
Ser#19	Serum ESC	27,893,745	19,648,668	4,078,175	14.6	24.0	3,931,433	18.5	63.8	1.3	2.1	79,338,599
Ser#20	Serum ESC	23,375,132	16,644,765	3,909,666	16.7	20.5	4,007,045	18.8	64.4	1.6	2.1	81,881,127
Bulk Serum	50ng Serum ESC DNA	86,457,361	82,429,400 [a]	54,414,142[b]	62.9	6.6	18,574,322	87.0	69.87	2.1	1.2	685,417,520
CT#1	FACS Beads	2,507,335	2,008,981	40,017	1.6	34.6	10,977	-	73.8	-	43.0	-
CT#2	FACS Beads	3,190,123	2,127,796	103,094	3.2	52.4	31,856	-	69.5	-	30.9	-
CT#4	Empty ESCs	2,424,673	1,896,897	30,356	1.3	33.3	10,438	-	70.7	-	36.0	-
CT#5	Empty ESCs	2,086,192	1,448,973	45,238	2.2	29.3	25,270	-	65.1	-	24.1	-
CT#6	Empty ESCs	4,425,725	2,736,905	156,468	3.5	61.5	27,903	-	74.4	-	45.3	-
CT#7	Empty MIIs	2,818,868	2,351,449	28,441	1.0	13.1	24,715	-	30.0	-	20.0	-
CT#8	Empty MIIs	2,231,558	1,409,497	46,355	2.1	37.7	13,654	-	69.5	-	44.7	-

1. Sequences left after trimming and mapping against human reference genome (GRCh37)

2. Values obtained from the Bismark reports

** Our data combined with Shirane et al. (2013)*

 # the number in brackets correspond to MII#2 / MII#5 and MII#2 deep / MII#5 deep merged datasets

[a] *Bulk datasets were not aligned to the human genome.*

[b] *Bulk datasets were first mapped in paired-end mode, before unaligned reads were mapped in single-end mode. These values are the sum of all reads mapped.*

Supplementary Table 2: Representation of regulatory regions in ESC scBS-Seq datasets.

Informative Regions / features are defined as covered by at least 5CpGs.

	CGIs		Promoters		LMRs		H3K4me1		p300		H3K27ac		H3K27me3		IAPs	
	Number	% total	Number	% total	Number	% total	Number	% total	Number	% total	Number	% total	Number	% total	Number	% total
Total	*23,020*		*32,071*		*26,335*		*25,029*		*30,236*		*14,574*		*7,953*		*21,824*	
2i#1	**16,319**	**70.9%**	**12,863**	**40.1%**	**3,210**	**12.2%**	**7,205**	**28.8%**	**5,315**	**17.6%**	**5,717**	**39.2%**	**4,859**	**61.1%**	**1,226**	**5.6%**
2i#2	**14,808**	**64.3%**	**10,786**	**33.6%**	**2,334**	**8.9%**	**5,793**	**23.1%**	**4,150**	**13.7%**	**4,724**	**32.4%**	**4,516**	**56.8%**	**871**	**4.0%**
2i#3	**15,695**	**68.2%**	**12,095**	**37.7%**	**2,871**	**10.9%**	**6,533**	**26.1%**	**4,886**	**16.2%**	**5,259**	**36.1%**	**4,704**	**59.1%**	**1,100**	**5.0%**
2i#4	**16,839**	**73.1%**	**13,496**	**42.1%**	**3,637**	**13.8%**	**8,046**	**32.1%**	**5,487**	**18.1%**	**6,156**	**42.2%**	**4,950**	**62.2%**	**1,506**	**6.9%**
2i#5	**16,004**	**69.5%**	**12,558**	**39.2%**	**3,290**	**12.5%**	**7,619**	**30.4%**	**5,288**	**17.5%**	**5,883**	**40.4%**	**4,781**	**60.1%**	**1,284**	**5.9%**
2i#6	**14,961**	**65.0%**	**11,271**	**35.1%**	**2,622**	**10.0%**	**6,297**	**25.2%**	**4,385**	**14.5%**	**4,962**	**34.0%**	**4,525**	**56.9%**	**1,296**	**5.9%**
2i#7	**17,203**	**74.7%**	**13,790**	**43.0%**	**3,512**	**13.3%**	**7,874**	**31.5%**	**5,664**	**18.7%**	**6,135**	**42.1%**	**5,044**	**63.4%**	**1,575**	**7.2%**
2i#8	**16,658**	**72.4%**	**12,998**	**40.5%**	**3,312**	**12.6%**	**7,402**	**29.6%**	**5,309**	**17.6%**	**5,813**	**39.9%**	**4,856**	**61.1%**	**1,390**	**6.4%**
2i#9	**17,547**	**76.2%**	**14,149**	**44.1%**	**3,602**	**13.7%**	**8,197**	**32.8%**	**5,954**	**19.7%**	**6,316**	**43.3%**	**5,141**	**64.6%**	**1,625**	**7.4%**
2i#10	**16,508**	**71.7%**	**12,430**	**38.8%**	**3,025**	**11.5%**	**7,203**	**28.8%**	**4,968**	**16.4%**	**5,655**	**38.8%**	**4,761**	**59.9%**	**1,337**	**6.1%**
2i#11	**13,774**	**59.8%**	**10,519**	**32.8%**	**2,499**	**9.5%**	**5,866**	**23.4%**	**4,063**	**13.4%**	**4,544**	**31.2%**	**4,172**	**52.5%**	**1,169**	**5.4%**
2i#12	**15,446**	**67.1%**	**12,102**	**37.7%**	**2,907**	**11.0%**	**6,944**	**27.7%**	**4,813**	**15.9%**	**5,442**	**37.3%**	**4,682**	**58.9%**	**1,183**	**5.4%**
Ser#1	**15,641**	**67.9%**	**13,122**	**40.9%**	**3,516**	**13.4%**	**8,028**	**32.1%**	**5,456**	**18.0%**	**6,136**	**42.1%**	**4,743**	**59.6%**	**1,613**	**7.4%**
Ser#2	**14,855**	**64.5%**	**11,438**	**35.7%**	**2,761**	**10.5%**	**6,329**	**25.3%**	**4,607**	**15.2%**	**5,033**	**34.5%**	**4,514**	**56.8%**	**1,097**	**5.0%**
Ser#3	**15,900**	**69.1%**	**13,520**	**42.2%**	**3,986**	**15.1%**	**8,214**	**32.8%**	**5,749**	**19.0%**	**6,204**	**42.6%**	**4,756**	**59.8%**	**1,992**	**9.1%**
Ser#4	**17,770**	**77.2%**	**15,091**	**47.1%**	**4,810**	**18.3%**	**9,552**	**38.2%**	**6,940**	**23.0%**	**7,361**	**50.5%**	**5,169**	**65.0%**	**1,566**	**7.2%**
Ser#5	**15,497**	**67.3%**	**12,386**	**38.6%**	**3,337**	**12.7%**	**7,246**	**29.0%**	**5,063**	**16.7%**	**5,573**	**38.2%**	**4,691**	**59.0%**	**1,375**	**6.3%**
Ser#6	**18,601**	**80.8%**	**16,468**	**51.3%**	**5,563**	**21.1%**	**10,536**	**42.1%**	**7,844**	**25.9%**	**8,061**	**55.3%**	**5,388**	**67.7%**	**1,970**	**9.0%**
Ser#7	**13,991**	**60.8%**	**11,059**	**34.5%**	**2,726**	**10.4%**	**6,283**	**25.1%**	**4,568**	**15.1%**	**4,819**	**33.1%**	**4,218**	**53.0%**	**1,361**	**6.2%**
Ser#9	**18,890**	**82.1%**	**16,192**	**50.5%**	**6,948**	**26.4%**	**9,993**	**39.9%**	**7,430**	**24.6%**	**7,618**	**52.3%**	**5,456**	**68.6%**	**1,465**	**6.7%**
Ser#8	**16,429**	**71.4%**	**12,930**	**40.3%**	**3,083**	**11.7%**	**6,864**	**27.4%**	**5,280**	**17.5%**	**5,549**	**38.1%**	**4,856**	**61.1%**	**1,041**	**4.8%**
Ser#10	**17,931**	**77.9%**	**15,062**	**47.0%**	**5,326**	**20.2%**	**9,225**	**36.9%**	**7,009**	**23.2%**	**7,115**	**48.8%**	**5,215**	**65.6%**	**1,558**	**7.1%**
Ser#11	**16,018**	**69.6%**	**12,720**	**39.7%**	**3,837**	**14.6%**	**7,055**	**28.2%**	**5,238**	**17.3%**	**5,540**	**38.0%**	**4,752**	**59.8%**	**1,204**	**5.5%**
Ser#12	**12,534**	**54.4%**	**8,964**	**28.0%**	**3,529**	**13.4%**	**5,134**	**20.5%**	**3,512**	**11.6%**	**4,282**	**29.4%**	**3,827**	**48.1%**	**512**	**2.3%**
Ser#13	**15,613**	**67.8%**	**11,695**	**36.5%**	**2,726**	**10.4%**	**6,033**	**24.1%**	**4,679**	**15.5%**	**4,988**	**34.2%**	**4,685**	**58.9%**	**925**	**4.2%**
Ser#14	**20,556**	**89.3%**	**18,960**	**59.1%**	**4,890**	**18.6%**	**12,274**	**49.0%**	**9,522**	**31.5%**	**9,102**	**62.5%**	**5,887**	**74.0%**	**2,362**	**10.8%**
Ser#15	**16,338**	**71.0%**	**12,804**	**39.9%**	**3,008**	**11.4%**	**7,239**	**28.9%**	**5,270**	**17.4%**	**5,708**	**39.2%**	**4,874**	**61.3%**	**1,249**	**5.7%**
Ser#16	**19,147**	**83.2%**	**16,621**	**51.8%**	**5,326**	**20.2%**	**10,538**	**42.1%**	**7,613**	**25.2%**	**7,945**	**54.5%**	**5,527**	**69.5%**	**1,710**	**7.8%**
Ser#17	**16,294**	**70.8%**	**13,549**	**42.2%**	**3,837**	**14.6%**	**8,336**	**33.3%**	**5,987**	**19.8%**	**6,627**	**45.5%**	**4,843**	**60.9%**	**1,255**	**5.8%**
Ser#18	**16,645**	**72.3%**	**13,365**	**41.7%**	**3,529**	**13.4%**	**7,799**	**31.2%**	**5,737**	**19.0%**	**6,151**	**42.2%**	**4,918**	**61.8%**	**1,284**	**5.9%**
Ser#19	**17,861**	**77.6%**	**14,612**	**45.6%**	**4,263**	**16.2%**	**8,743**	**34.9%**	**6,691**	**22.1%**	**6,963**	**47.8%**	**5,193**	**65.3%**	**1,330**	**6.1%**
Ser#20	**17,809**	**77.4%**	**14,679**	**45.8%**	**3,967**	**15.1%**	**8,421**	**33.6%**	**6,279**	**20.8%**	**6,565**	**45.0%**	**5,202**	**65.4%**	**1,492**	**6.8%**
Bulk_2i	**22,908**	**99.5%**	**28,220**	**88.0%**	**21,933**	**83.3%**	**19,705**	**78.7%**	**17,198**	**56.9%**	**12,457**	**85.5%**	**6,789**	**85.4%**	**15,552**	**71.3%**
Bulk_Serum	**22,937**	**99.6%**	**28,565**	**89.1%**	**23,545**	**89.4%**	**19,972**	**79.8%**	**17,947**	**59.4%**	**12,579**	**86.3%**	**6,830**	**85.9%**	**15,918**	**72.9%**
Average single cells	***16,440***	***71.4%***	***13,259***	***41.3%***	***3,681***	***14.0%***	***7,776***	***31.1%***	***5,649***	***18.7%***	***6,061***	***41.6%***	***4,866***	***61.2%***	***1,373***	***6.3%***