Joint profiling of chromatin accessibility, DNA methylation and transcription in single cells

Stephen J. Clark[1], Ricard Argelaguet[2,3], Chantriolnt-Andreas Kapourani[4], Thomas M. Stubbs[1], Heather J. Lee[1,5,6], Felix Krueger[7], Guido Sanguinetti[4], Gavin Kelsey[1,8], John C. Marioni[2,3,5], Oliver Stegle[2,9], Wolf Reik[1,5,8,9]

1. Epigenetics Programme, Babraham Institute, Cambridge, UK
2. European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK
3. Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge
4. School of Informatics, University of Edinburgh, UK
5. Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
6. School of Biomedical Sciences and Pharmacy, The University of Newcastle, Callaghan, NSW, Australia
7. Bioinformatics Group, Babraham Institute, Cambridge, UK
8. Centre for Trophoblast Research, University of Cambridge, UK
9. Joint senior authors.

Corresponding authors: Wolf Reik (wolf.reik@babraham.ac.uk) and Oliver Stegle (oliver.stegle@ebi.ac.uk)

Parallel single-cell sequencing protocols represent powerful methods for investigating regulatory relationships, including epigenome-transcriptome interactions. Here, we report the first single-cell method for parallel chromatin accessibility, DNA methylation and transcriptome profiling. scNMT-seq (single-cell nucleosome, methylation and transcription sequencing) uses a GpC methyltransferase to label open chromatin followed by bisulfite and RNA sequencing. We validate scNMT-seq by applying it to mouse embryonic stem cells, finding links between all three molecular layers and revealing strong and widespread associations between chromatin accessibility and DNA methylation.

Understanding regulatory associations between the epigenome and the transcriptome requires simultaneous profiling of multiple molecular layers. Previously, such multi-omics analyses have been limited to bulk assays, which profile ensembles of cells. These studies have used variation in the expression of a gene across individuals[1] or between cell types[2] or conditions to assess such linkages. Alternatively, it is also possible to link chromatin state with

transcription by exploiting variability between genes within a single sample. However, insights from such an approach are limited to the discovery of genome-wide global trends[3].

With rapid advances in single-cell technologies it is increasingly possible to leverage variation between single cells in order to probe regulatory associations between molecular layers. Existing protocols allow the methylome and the transcriptome or, alternatively, the methylome and chromatin accessibility to be assayed in the same cell[4–7]. However, it is well known that DNA methylation and other epigenomic features such as chromatin accessibility do not act independently of one another. Consequently, the ability to profile, at single cell resolution, multiple epigenetic features in conjunction with gene expression is critical for obtaining a more complete understanding of how transcription, and thus cell state, is regulated[8].

To address this, we have developed a method that enables the joint analysis of the transcriptome, the methylome and chromatin accessibility. Our approach builds on previous parallel protocols such as single-cell methylation and transcriptome sequencing (scM&T-seq)[1], in which physical separation of DNA and RNA is performed first, to enable the cell's transcriptome to be profiled using a conventional Smartseq2 protocol[9]. To measure chromatin accessibility together with DNA methylation, we adapted the Nucleosome Occupancy and Methylation sequencing (NOMe-seq) method[7,10], where a methyltransferase (methylase) enzyme is used to label accessible (or nucleosome depleted) DNA prior to bisulfite sequencing (BS-seq), which distinguishes between the two chromatin states. In mammalian cells, cytosine residues in CpG dinucleotides are frequently methylated, whereas cytosines followed by either adenine, cytosine or thymine (collectively termed CpH) are methylated at a much lower rate[11]. Consequently, by using a GpC methylase enzyme (M.CviPI) to label accessible chromatin, NOMe-seq can recover endogenous CpG methylation information in parallel. NOMe-seq is particularly attractive for single-cell applications since, contrary to count-based methods such as ATAC-seq or DNase-seq, the GpC accessibility is encoded through the bisulfite conversion and hence inaccessible chromatin can be directly discriminated from missing data. Additionally, the resolution of the method is determined by the frequency of GpC sites within the genome (~1 in 16bp), rather than the size of a library fragment (>100bp) (see Fig. 1a for an illustration of the protocol).

To demonstrate the performance of scNMT-seq, we applied the method to a batch of 70 serum-grown EL16 mouse embryonic stem cells (ESCs), together with four negative (empty wells) and three scM&T-seq controls (cells processed using scM&T-seq, i.e., which did not receive M.CviPI enzyme treatment). This facilitates direct comparison with previous methods for assaying DNA methylation and transcription in the same cell[4,12].

We isolated single cells into GpC methylase reaction mixtures by FACS, before physically separating the DNA and RNA prior to bisulfite and RNA sequencing library preparation[1]. See Supplementary Table 1 for sequencing summary statistics. Alignment of the BS-seq data and other bioinformatics processing can be carried out using established pipelines, with the addition of a filter to discard G-C-G positions, for which it is intrinsically not possible to distinguish endogenous methylation from *in vitro* methylated bases (21% genome-wide). Similarly, we remove C-C-G positions to mitigate possible off-target effects of the enzyme[10] (27% genome-wide). In total, 58 out of 70 cells processed using scNMT-seq passed quality control for both bisulfite and RNA-seq.

First, we considered the RNA-seq component, which is directly comparable to scM&T-seq transcriptome data. On average, we detected 7,700 genes per cell (CPM ≥1), which is comparable with data from the same cell type profiled using scM&T-seq[1]. We used PCA and hierarchical clustering to jointly analyse cells across protocols and studies (using data from Angermueller *et al.* 2016[4]), and found that scM&T-seq and scNMT-seq samples prepared in parallel cluster together. This indicates that the enzyme treatment does not adversely affect the transcriptome (Supplementary Fig. 1). Larger differences were observed when comparing across studies, most likely reflecting differences in the cell lines used (male E14 versus female EL16[13], Supplementary Fig. 1).

The need to filter out C-C-G and G-C-G positions from the methylation data reduces the number of genome-wide cytosines that can be assayed from 22 million to 11 million. However, despite this filter, a large proportion of the loci in genomic regions with important regulatory roles, such as promoters and enhancers, can be profiled using scNMT-seq (Fig. 1b). Consistent with this theoretical expectation, we observed high empirical coverage: 51% of promoters and 78% of gene bodies are captured by at least 5 cytosines (Fig. 1c, Supplemental Fig. 2a). We also compared the methylation coverage to data from our previous publication[4], again finding small differences relative to conventional BS-seq, albeit these differences became more pronounced when down-sampling the total sequence coverage (e.g. the reduction in gene body coverage increased from 5% to 16% when sampling 1/10th of the reads; Supplemental Fig 2b). Due to the higher frequency of GpC compared to CpG dinucleotides in the mouse genome, the coverage of GpC accessibility was larger than that observed for endogenous CpG methylation (Fig. 1b, c and Supplementary Fig. 2a). We found, on average, that 91% of gene bodies and 79% of promoters per cell were assessable, which is the highest coverage achieved by any single-cell accessibility protocol to date (9.4% using scATAC-seq[14], and with scDNase-seq, ~50% of genes >1 RPKM, >80% of genes >3 RPKM[15]). Analogous to the analysis of the RNA-seq data, we compared the CpG methylation profiles obtained from scNMT-seq to single-cell libraries profiled using scM&T-seq[4], scBS-seq[12] and

bulk BS-seq[16], finding that cells did not cluster by protocol or by study, with most variation being attributable to difference in cell type (Supplementary Fig. 3).

To validate the accuracy of the GpC accessibility measurements, we generated a synthetic bulk dataset by merging GpC methylation data from all cells, and compared this with published bulk DNase-seq data from the same cell type[7]. Globally, we observed high consistency between datasets (Pearson R = 0.75, weighted by coverage in our merged dataset, Supplemental Fig 4). The most notable difference was that the scNMT-seq data showed oscillating profiles, with peaks spaced ~180 to ~200bp apart, consistent with the positions of nucleosomes (Fig. 1d) and similar to profiles obtained with bulk-cell NOMe-seq[2].

Next, we examined GpC methylation levels at known regulatory regions in single-cells. Across the genome, GpC accessibility was ~30%, with low cell-cell variability. However, we found a large increase in GpC accessibility at known DNase hypersensitivity sites (DHS, ~60% GpC methylated, Supplemental Fig. 5), as well as transcriptional start sites (~60% GpC methylated, Fig. 1e). We observed similar patterns for protein- and transcription factor binding sites (from p300, CTCF, Nanog and Oct4 ChIP-seq data), which were accessible at the centre of the peaks. Cells processed using the scM&T-seq control were universally low in GpC methylation (~2%) with no enrichment at regulatory regions, indicating that our accessibility data are not affected by endogenous GpC methylation (Supplementary Fig. 6). To illustrate the high-resolution GpC accessibility measurements obtained by our method, we profiled the pattern and density of nucleosomes around transcription start sites finding characteristic nucleosome depleted regions at transcription start sites and variation between cells in the position of nucleosomes (see Supplementary Fig. 7 for example plots).

To assess how differences in gene expression are associated with methylation and GpC accessibility, we stratified loci based on the expression level of the nearest gene using the RNA-seq profiles from the corresponding cells. We found that highly expressed genes were associated with the greatest GpC accessibility at promoters and at nearby regulatory sites, whereas the GpC accessibility of lowly-expressed genes was reduced (Fig. 1e; Supplementary Fig 8).

Taken together, these results demonstrate that our method is able to robustly profile gene expression, DNA methylation and GpC accessibility within the same single cell.

Having established the efficacy of our method, we next explored its potential for identifying loci with coordinated epigenetic and transcriptional heterogeneity. Globally, we observed a clear relationship between average CpG methylation rate and the GpC accessibility across cells, where methylated loci were associated with decreased accessibility (Fig. 2a). When assessing the heterogeneity of CpG methylation in different genomic contexts, enhancers

were most variable (particularly primed enhancers – H3K4me1 marked but lacking H3K27ac), followed by non-CGI promoters and inactive promoters (Supplemental Fig. 9), which is in agreement with previous data[4,12]. In contrast, heterogeneity in GpC accessibility was largest at known binding sites of transcription factors (Oct4 and Nanog) and regions of active chromatin (p300 binding sites and DNase-hypersensitive sites), indicating cell to cell differences in the accessibility of the DNA to important regulatory factors (Fig. 2a and supplemental Fig. 9).

We next jointly considered the GpC accessibility and CpG methylation data to test for correlated changes between the two layers. Significant associations were observed across all genomic contexts, with up to 98 loci showing significant patterns (FDR < 10%; Fig. 2b; Supplementary Fig. 10a and 11). The majority of significant correlations were negative, reflecting the known relationship between these two layers[17]. The largest number of individual associations was observed in intronic regions (N=98), followed by Super Enhancer regions (N=51, Fig. 2b.).

In addition to coupling between different epigenetic layers, we also considered associations between CpG methylation and GpC accessibility and gene expression levels. Because these effects were generally weaker than the relationship between accessibility and methylation, we used a data-driven approach to optimise the set of promoter proximal regions in which to test for such associations (Methods). This analysis identified -100bp to +100bp for accessibility and -1kb to +1kb for methylation as suitable parameters for such analyses (Supplementary Fig. 12). Notably, the strongest associations between accessibility and expression were observed upstream of the TSS, whereas the linkages for DNA methylation were most pronounced downstream of the TSS. We used these regions to assess linkages between DNA methylation and accessibility with gene expression. We found 4 significant associations between GpC accessibility and gene expression with a greater number of positive (3) compared to negative (1) correlations (Fig. 2c and Supplementary Fig. 13a and 14) and for CpG methylation and transcription, we found 39 significant associations with an enrichment for negative correlations (33/39), confirming the known negative relationship between DNA methylation and gene expression (Supplementary Fig. 15a and 16). See Supplementary Table 2 for a list of all significant correlations.

As an example, Fig. 2d displays the gene *Cth* and surrounding region, showing mean accessibility and methylation rates across the locus as well as a scatter plot, depicting significant associations between GpC accessibility or CpG and methylation at the promoter region and *Cth* expression. Notably, this relationship could also be observed in individual cells,

as shown in the zoom-in examples, revealing specific cells with either an accessible promoter and expressed transcripts or inaccessible and non-expressed.

We additionally analysed associations across genes within each cell (rather than across cells within each gene), which is similar to previous approaches used to investigate such linkages using a single bulk sample. This approach showed global correlations in different genomic contexts (Supplementary Fig. 10b, 13b, 15b), indicating that our method is accurately measuring each layer and recapitulates the expected bulk-cell results.

In conclusion, we describe a method for parallel single-cell DNA methylation, gene expression and high resolution chromatin accessibility measurements and report novel associations between each molecular layer with a strong enrichment for DNA methylation – chromatin accessibility correlations. This method will greatly expand our ability to investigate relationships between the epigenome and transcriptome in heterogeneous cell types and across developmental transitions.

Methods

Cell culture

Mouse embryonic stem cells were derived from a 129xCast/129 embryo previously[13] and cultured in serum media without feeders as previously[4]. Single-cells were collected by FACS, selecting for live cells and low DNA content (i.e., G0 or G1 phase cells) using ToPro-3 and Hoechst 33342 staining as previously described[4]. The cell line was subjected to routine mycoplasma testing using the MycoAlert testing kit (Lonza).

Library preparation

Cells were collected directly into 2.5µl methylase reaction mixture which was comprised of 1x M.CviPI Reaction buffer (NEB), 2U M.CviPI (NEB), 160 µM S-adenosylmethionine (NEB), 1U/µl RNAsein (Promega), 0.1% IGEPAL (Sigma) then incubated for 15 minutes at 37°C. The reaction was stopped and the RNA preserved with the addition of 5µl RLT plus (Qiagen) prior to scM&T-seq library preparation according to the published protocols for G&T-seq[19] and scBS-seq[20] but with the following modifications. Three G&T-seq washes were performed with 15µl volumes (steps 22 to 24 of the G&T-seq protocol [21]) and the reverse transcription reaction and PCR were performed using volumes of the published Smart-seq2 protocol[22] (i.e. 10 µl for reverse transcription and 25 µl for PCR).

Sequencing

20 of the BS-seq libraries, including 3 negative controls, were initially sequenced on 50bp single-end MiSeq run to assess quality. The negative controls were found to have substantially

reduced mapping efficiencies compared to the single cell samples (mean of 2.7% compared to 36.8%, see Supplementary Table 1). All single-cell BS-seq libraries were subsequently sequenced to a mean depth of 17 million paired-end reads and RNA-seq libraries were sequenced to a mean depth of 1.7 million paired-end reads. Both sets of libraries were sequenced on HiSeq 2500 instruments using v4 reagents and 125bp read length.

Data processing

Bisulfite-seq alignment

Single-cell bisulfite libraries were processed using Bismark[23] as described[20] but with the additional *--NOMe* option in the coverage2cytosine script which produces CpG report files containing only A-C-G and T-C-G positions and GpC report files containing only G-C-A, G-C-C and G-C-T positions.

RNA-seq alignment

Single-cell RNA-seq libraries were aligned using HiSat2[24] using options -O3 -m64 -msse2 -funroll-loops -g3 -DPOPCNT.

Allele-sorting

Since the cell-line used was derived from a hybrid embryo (129 x 129/cast) reads were separated by known SNPs between the two strains, using SNPsplit[25], however for the purposes of this study, genome-specific data was merged and therefore the allelic origin ignored.

Quality control

From the bisulfite-seq data, we discarded cells that had (1) less than 10% mapping efficiency (2) less than 500,000 CpG sites or 5,000,000 GpC sites covered. In total, 64 cells (88%) passed the quality control (supplemental Fig. 18). From the RNA-seq data we discarded cells that had (1) less than 300,000 reads mapped (2) more than 15% of total reads mapped to mitochondrial genes, (3) less than 2,000 genes expressed. In total, 66 cells (90%) passed the quality control (supplemental Fig. 17), 61 of which also passed BS-seq QC (84%) comprising 58 scNMT-seq cells and 3 scM&T-seq cells.

CpG Methylation and GpC accessibility quantification

Following the approach of Smallwood et al[7] individual CpG or GpC sites in each cell were modelled using a binomial model where the number of successes is the number of reads that support methylation and the number of trials is the total number of reads. A CpG methylation or GpC accessibility rate for each site and cell was calculated by maximum *a posteriori* assuming a beta prior distribution. Subsequently, CpG methylation and GpC accessibility rates were computed for each genomic feature assuming a Normal distribution across cells and

accounting for differences in the standard errors of the single site estimates. The coverage (number of observed CpG or GpC sites) was recorded and used as weight in subsequent analysis. See Supplementary Table 3 for details of genomic contexts used in this study.

RNA quantification

Gene expression counts were quantified from the mapped reads using featureCounts[26]. Gene annotations were obtained from Ensembl version 87[26]. Only protein-coding genes matching canonical chromosomes were considered. Following[27] the count data was log-transformed and size-factor adjusted based on a deconvolution approach that accounts for variation in cell size[28].

Statistical analysis

CpG Methylation and GpC accessibility profiles

CpG methylation and GpC accessibility profiles were visualised by taking predefined windows around the genomic context of interest. For each cell and feature, CpG methylation and GpC accessibility values were averaged using running windows of 50 bp. The information from multiple cells was combined by calculating the mean and the standard deviation for each running window. Profiles were calculated using a subset of 20 cells with similar mean methylation rate values. Genes were split into three classes according to a histogram of the log2 normalised expression counts (x): Low (x<2), Medium (2<x<6) and High (x>6). For genomic features that are not directly linked to genes (i.e. enhancers or transcription factor binding sites), all possible relationships between genes and features within 5kb of the gene (upstream and downstream of gene start and stop) were considered.

GpC accessibility profiles around the TSS in a single cell (as displayed in Supplementary Fig. 9a and Fig. 2e) were generated using a generalised linear model (GLM) of basis function regression coupled with a Bernoulli likelihood using BPRMeth[29].

Correlation analysis

For the correlation analysis across cells, genes with low expression levels and low variability were discarded, according to the rationale of independent filtering[30]. Genomic features observed in less than 50% of the cells and with a coverage of less than 3 sites were discarded. Furthermore, only the top 50% of the most variable loci were considered for analysis and a minimum number of 20 cells was required to compute a correlation. Only genomic contexts with more than 100 features that passed the filtering criteria were considered for the analysis. A minimum coverage of 3 sites was required per feature. For association tests, all possible relationships between genes and genomic features within 8kb of the gene (upstream and downstream) were considered. Following our previous approach[4], we tested for linear associations by computing a weighted Pearson correlation coefficient, thereby accounting for

differences in the coverage between cells. When assessing correlations between GpC accessibility with CpG methylation, we used the average CpG methylation coverage as a weight.

Two-tailed Student's t-tests were performed to test for nonzero correlation, and P-values were adjusted for multiple testing for each context using the Benjamini-Hochberg procedure.

To improve the correlations of promoter methylation or accessibility with expression, we optimized the genomic window used to define the CpG methylation or GpC accessibility rate as follows. First, we selected 20 random cells and we extracted +/-4kb regions around the transcription start site of all genes and we divided them into overlapping 200bp windows with a stride of 50bp (Supplementary figure 12). Then, for each cell and window, we performed a correlation across all genes between the CpG methylation or GpC accessibility rates and the corresponding gene expression. Finally, we selected the regions for which the correlation is maximized, in the case of accessibility being +/-100bp and in the case of methylation +/- 1kb.

References

1. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (2012).
2. Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23,** 555–567 (2013).
3. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* **13,** R53 (2012).
4. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13,** 229–232 (2016).
5. Hou, Y. *et al.* Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26,** 304–319 (2016).
6. Hu, Y. *et al.* Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* **17,** 88 (2016).
7. Pott, S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *bioRxiv* (2016). doi:10.1101/061739

8. Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G. & Reik, W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol.* **17,** 72 (2016).
9. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1098 (2013).
10. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22,** 2497–2506 (2012).
11. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).
12. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11,** 817–820 (2014).
13. Lee, J., Davidow, L. S. & Warshawsky, D. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat. Genet.* **21,** 400–404 (1999).
14. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).
15. Jin, W. *et al.* Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528,** 142–146 (2015).
16. Ficz, G. *et al.* FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency. *Cell Stem Cell* **13,** 351–359 (2013).
17. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
18. Iurlaro, M., von Meyenn, F. & Reik, W. DNA methylation homeostasis in human and mouse development. *Curr. Opin. Genet. Dev.* **43,** 101–109 (2017).
19. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12,** 519–522 (2015).
20. Clark, S. J. *et al.* Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat. Protoc.* **12,** 534–547 (2017).

21. Macaulay, I. C. *et al.* Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nat Protoc.* **11,** 2081–2103 (2016).

22. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,** 171–181 (2014).

23. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinforma. Oxf. Engl.* **27,** 1571–1572 (2011).

24. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–360 (2015).

25. Krueger, F. & Andrews, S. R. SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Research* **5,** (2016).

26. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma. Oxf. Engl.* **30,** 923–930 (2014).

27. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5,** 2122 (2016).

28. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17,** 75 (2016).

29. Kapourani, C.-A. & Sanguinetti, G. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics* **32,** i405–i412 (2016).

30. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 9546–9551 (2010).

31. Kolodziejczyk, A *et al.* Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell stem cell.* **17(4)**, 471-485 (2015).

32. Zvetkova, I *et al*. Global hypomethylation of the genome in XX embryonic stem cells. *Nat.Gen.* **37**, 1274-1275 (2005).

Author contributions

S.J.C and W.R. conceived the project. S.J.C, T.M.S and H.J.L performed experiments. R.A., S.J.C and C-A.K performed statistical analysis. F.K. processed and managed sequencing data. S.J.C, R.A, J.C.M, O.S, W.R interpreted results and drafted the manuscript. G.S., G.D.K, J.C.M, O.S. and W.R supervised the project.

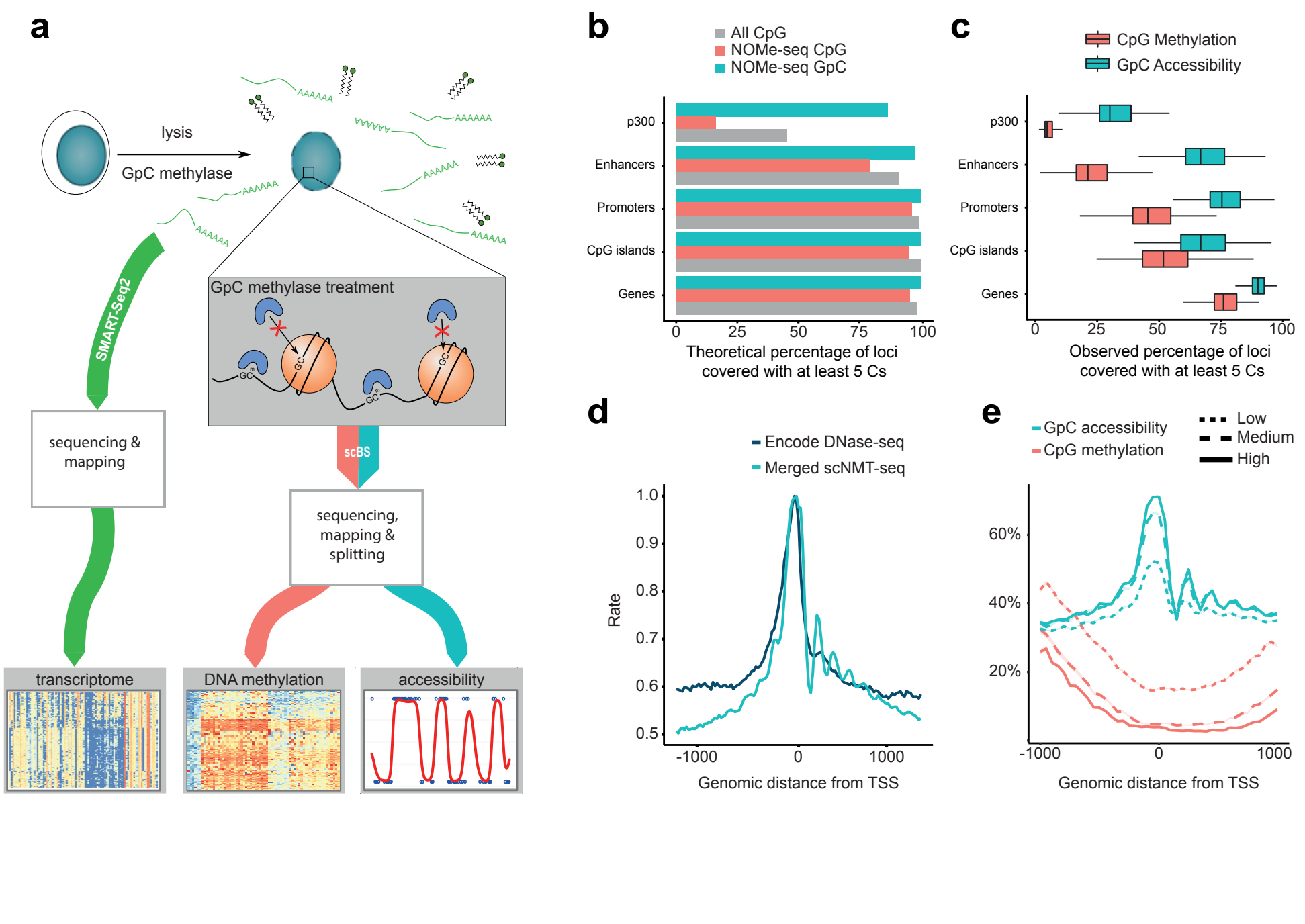

a
lysis
GpC methylase
GpC methylase treatment
SMART-Seq2
scBS
sequencing & mapping
sequencing, mapping & splitting
transcriptome
DNA methylation
accessibility
b
All CpG
NOMe-seq CpG
NOMe-seq GpC
p300
Enhancers
Promoters
CpG islands
Genes
0
25
50
75
100
Theoretical percentage of loci covered with at least 5 Cs
c
CpG Methylation
GpC Accessibility
Observed percentage of loci covered with at least 5 Cs
d
Encode DNase-seq
Merged scNMT-seq
Rate
1.0
0.9
0.8
0.7
0.6
0.5
-1000
0
1000
Genomic distance from TSS
e
GpC accessibility
CpG methylation
Low
Medium
High
60%
40%
20%
Genomic distance from TSS

Fig. 1. Coverage and accuracy of scNMT-seq.

(a) Protocol overview. Single-cells are lysed and accessible DNA is labelled using GpC methylase. RNA is then separated and sequenced using Smart-seq2, whilst DNA undergoes scBS-seq library preparation and sequencing. Methylation and chromatin accessibility data are separated bioinformatically (splitting).

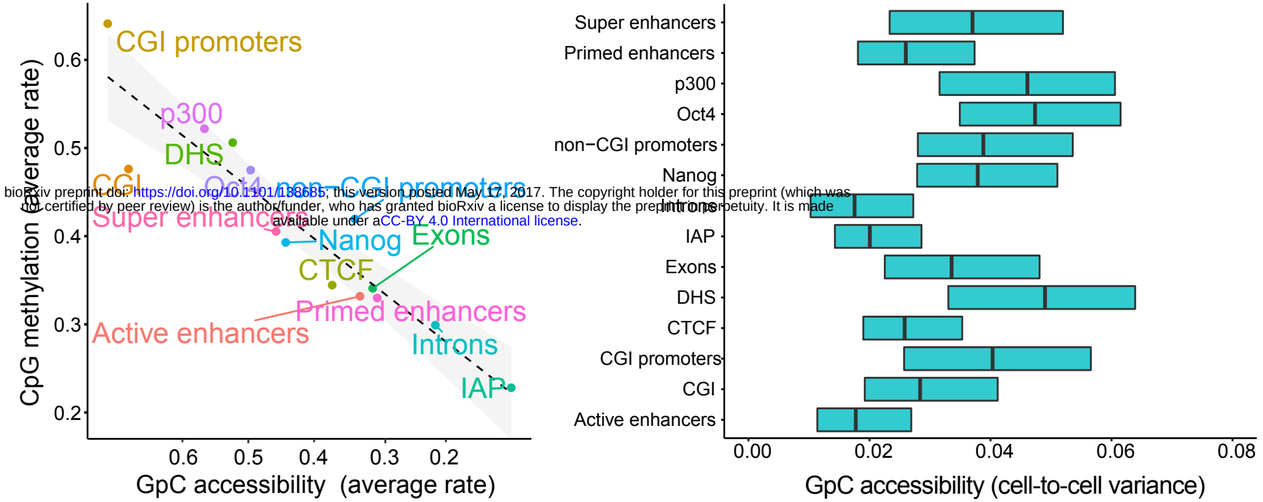
(b) Theoretical maximum coverage of representative genome contexts used in this study. Shown is the proportion of loci in different contexts that can be covered by at least 5 cytosines. All CpG considers any C-G dinucleotides (e.g. as in scBS-seq), NOMe-seq CpG considers A-C-G and T-C-G trinucleotides and NOMe-seq GpC considers G-C-A, G-C-C and G-C-T trinucleotides.

(c) Empirical coverage of individual loci, considering the same contexts as in b. Shown is the coverage across each of 58 single-cells (after QC); box plots show median coverage and the first and third quartile, whiskers show 1.5 x the interquartile range above and below the box.

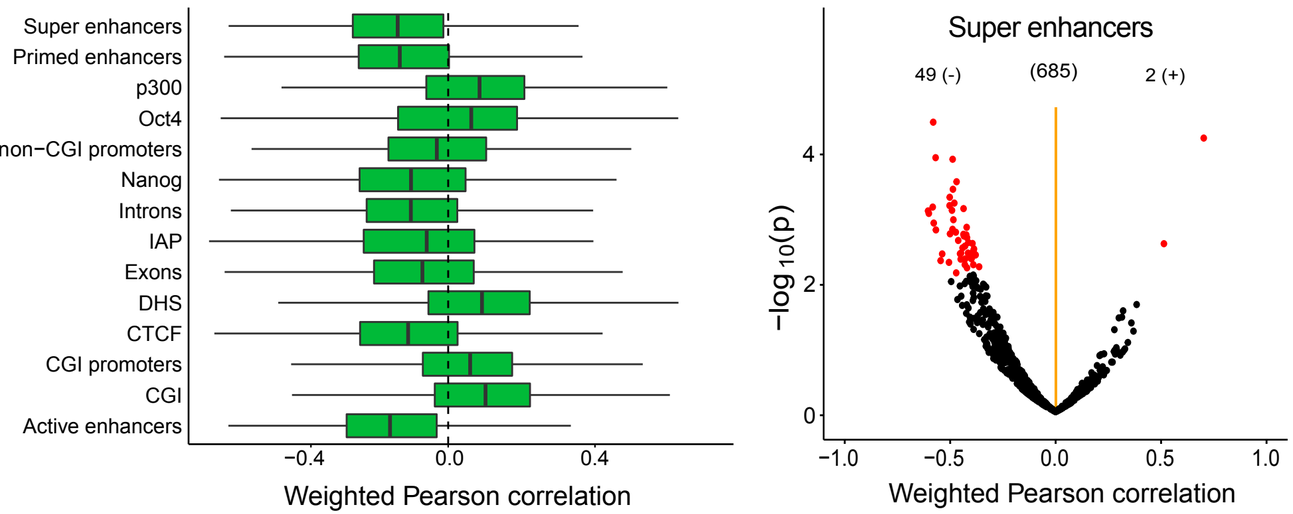
(d) GpC accessibility profiles at gene promoters compared to published DNase-seq data. GpC accessibility is the mean rate of all cells in 25bp windows, DNase-seq is the number of reads within the same 25bp windows. Both were scaled to the fraction of the maximum in all windows to enable a comparison of the two different data types.

(e) CpG methylation and GpC accessibility profiles at gene promoters. Promoters are stratified by expression of the corresponding gene (low = logCPM < 2, medium = 2 < logCPM < 6 , high = logCPM > 6) within the same cell. The profile is generated by computing a running mean and standard deviation in 50bp windows across 20 random cells.

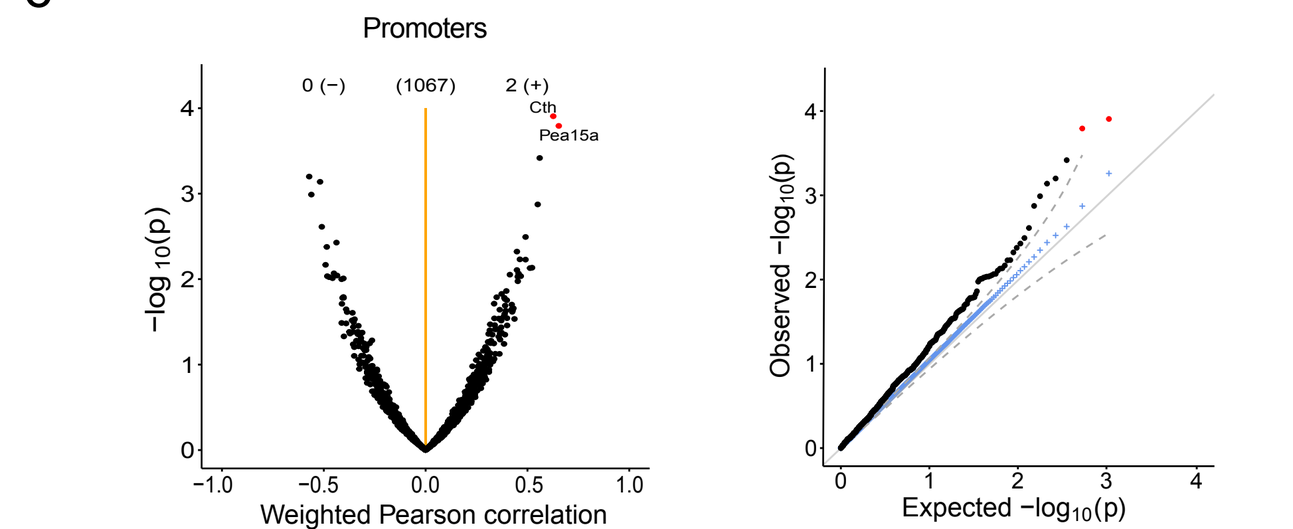
a



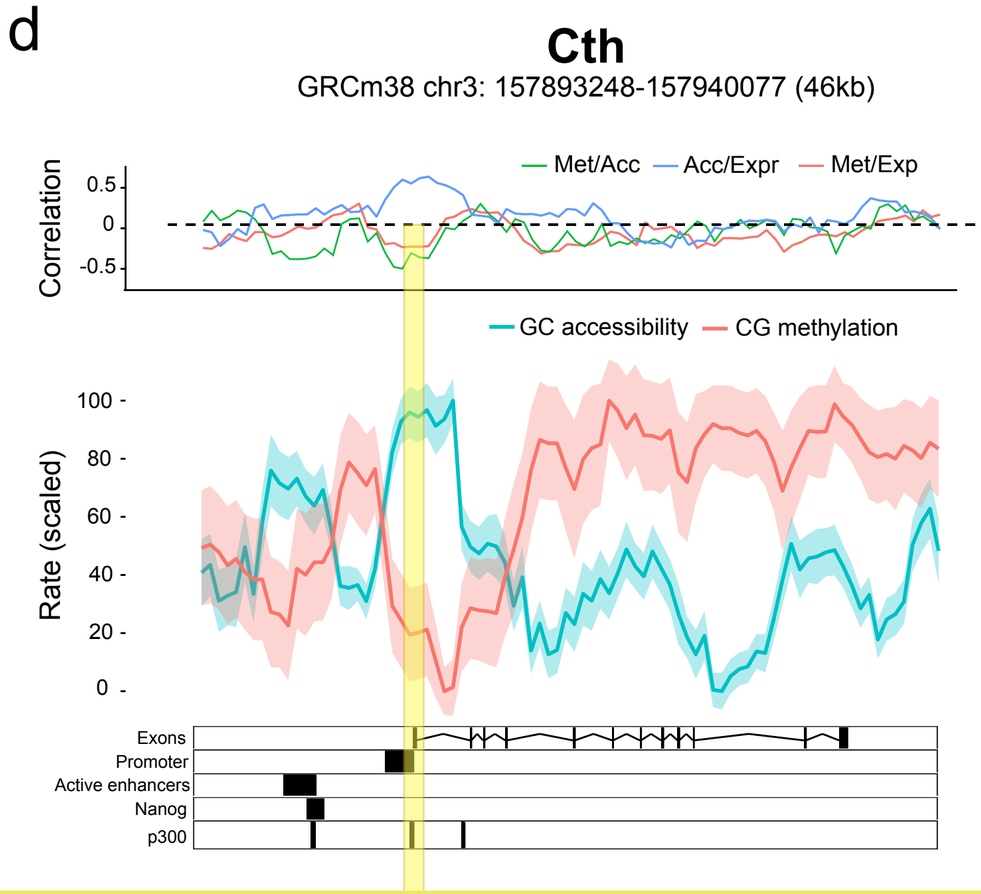
b



c



d



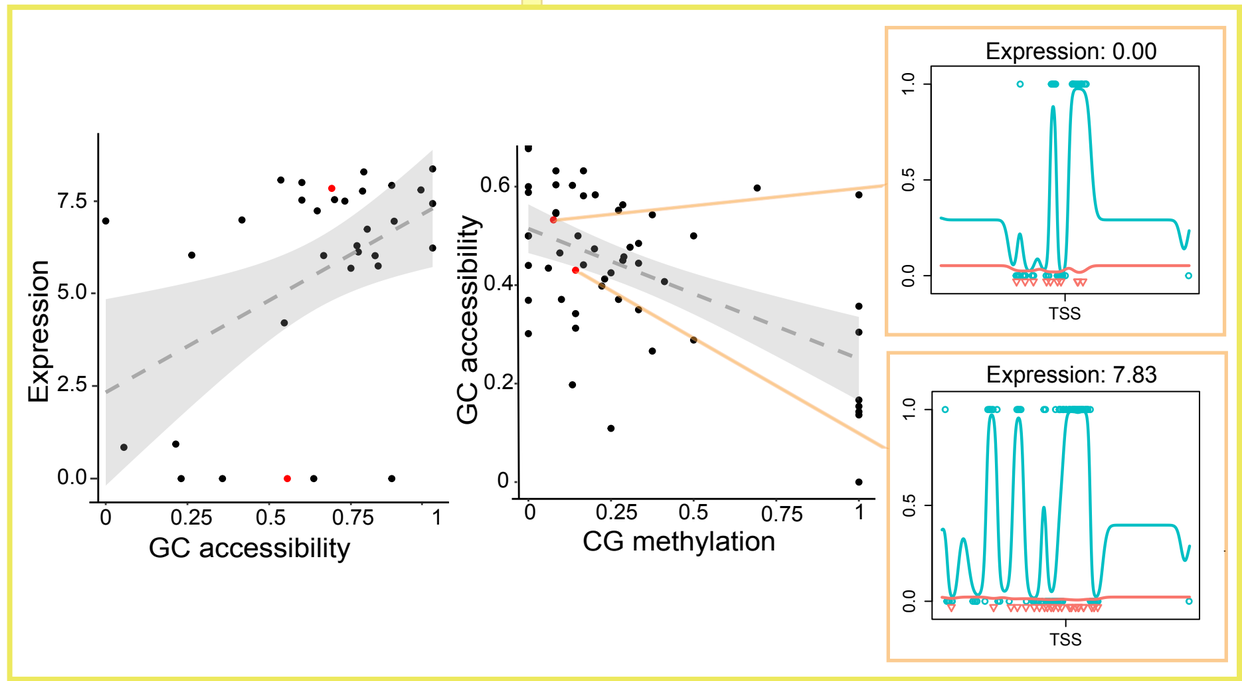


Fig. 2. Linking heterogeneity in GpC accessibility, DNA methylation and gene expression using scNMT-seq.

(a) Left: Scatter plots between average CpG methylation and GpC accessibility for different genomic contexts. Shown are, for each context, averages across cells and loci, revealing a negative association between both layers. Right: Box plots of the cell-to-cell variance of GC chromatin accessibility for the corresponding contexts. Boxplots highlight the median accessibility across loci and the lower and upper hinges correspond to the first and third quartiles, respectively.

(b) Associations between GpC accessibility and CpG methylation across cells, for different genomic contexts. Left: Boxplots of Pearson r-values for individual loci and for different contexts (shown are medians and first and third quartiles). Right: Pearson correlation coefficients versus P-values for loci in the super enhancer context (N=685). Significant associations (FDR<0.1, Benjamini-Hochberg correction), are highlighted in red. The top panel shows the number of significant positive (+) and negative (−) correlations (FDR < 0.1).

(c) Associations between GpC accessibility gene expression across cells, for gene promoters. Left: Pearson correlation coefficients versus P-values for individual loci (N=1,067). Right: Q-Q plot, showing observed p-values versus the random expectation. Solid points correspond to actual P-values from 1,067 promoter associations, with significant association (FDR<0.1, Benjamini Hochberg adjusted) highlighted in red. Blue '+' symbols show analogous results from permuted data, revealing no association.

(d) Zoom-in view for the gene *Cth*. Shown from top to bottom are: Pairwise Pearson correlation coefficients between each of the three layers. CpG methylation (red) and GpC accessibility (blue) profiles; mean rates (solid line) and standard deviation (shade) were calculated using a running window of 4kb with a step size of 500bp; to show relative instead of absolute changes and to bring the two layers into the same scale, CpG methylation and GpC accessibility rates were separately scaled to 0 and 100. Track with genomic annotations, highlighting the position of several regulatory elements: promoters, enhancers, Nanog binding sites and p300 binding sites. The scatter plots show the correlation between accessibility and gene expression as well as methylation and accessibility around the transcription start site. Finally, two cells were selected to display the actual methylation (red) and accessibility (blue) profile around the transcription start site, dots display empirical data points, lines represent fitted profiles (see online methods).