Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity

Christof Angermueller[1,7], Stephen J Clark[2,7], Heather J Lee[2,3,7], Iain C Macaulay[3,7], Mabel J Teng[3], Tim Xiaoming Hu[1,3,4], Felix Krueger[5], Sébastien A Smallwood[2], Chris P Ponting[3,4], Thierry Voet[3,6], Gavin Kelsey[2], Oliver Stegle[1] & Wolf Reik[2,3]

We report scM&T-seq, a method for parallel single-cell genome-wide methylome and transcriptome sequencing that allows for the discovery of associations between transcriptional and epigenetic variation. Profiling of 61 mouse embryonic stem cells confirmed known links between DNA methylation and transcription. Notably, the method revealed previously unrecognized associations between heterogeneously methylated distal regulatory elements and transcription of key pluripotency genes.

Multiparameter sequencing-based analysis of single cells is a powerful tool for dissecting relationships among epigenetic, genomic and transcriptional heterogeneity[1]. Recent advances have enabled single-cell genome-wide or reduced-representation bisulfite sequencing (scBS-seq or scRRBS[2–4]), making it possible to explore the intercellular heterogeneity of DNA methylation[5,6]. We and others have recently described methods for parallel genome and transcriptome sequencing in single cells[7,8]. Our method G&T-seq (genome and transcriptome sequencing) involves physical separation of RNA and DNA, which allows for bisulfite conversion of DNA without affecting the transcriptome. Here we applied scBS-seq to genomic DNA purified according to the G&T-seq protocol to generate methylomes and transcriptomes from the same single cells (**Fig. 1a** and **Supplementary Fig. 1**). Parallel profiling using scM&T-seq will enable detailed study of the complex relationship between DNA methylation and transcription in heterogeneous cell populations[9,10] and may be used to provide multidimensional information in clinical contexts where material is severely limited (for example, *in vitro* fertilization).

To demonstrate the potential of the method, we applied scM&T-seq to mouse embryonic stem cells (ESCs). In the presence of serum, these cells constitute a metastable population with stochastic switching between transcriptional states[11,12]. This transcriptional heterogeneity has been linked to the differentiation potential of ESCs, with NANOG^{lo} cells having an increased propensity to differentiate[13] and elevated expression of differentiation markers compared with NANOG^{hi} cells[12,14,15]. Experiments in sorted populations of cells have also linked transcriptional and epigenetic heterogeneity by demonstrating differences in DNA methylation between transcriptional states, such as gains in DNA methylation in NANOG^{lo} and REX1^{lo} (REX1 is also known as ZFP42) cells compared with, respectively, NANOG^{hi} and REX1^{hi} cells[11,16]. The development of single-cell techniques has allowed the transcriptional heterogeneity of ESCs to be studied in unprecedented detail, revealing a complex population structure and multiple sources of variation[17,18]. Using scBS-seq, we have also demonstrated DNA-methylation heterogeneity in ESCs at the single-cell level[3]. To further investigate the link between epigenetic and transcriptional heterogeneity in ESCs, we performed scM&T-seq on 76 individual serum ESCs and 16 ESCs grown in '2i' media, which induces genome-wide DNA hypomethylation[16].

We obtained an average of 2.7 million scRNA-seq reads per cell, and we excluded cells with fewer than 2 million mapped reads (**Supplementary Table 1**). We have previously shown that the scRNA-seq data generated by the G&T-seq method is of similar quality to that generated using the scRNA-seq protocol (Smart-seq2) alone[7]. In ESCs that met scRNA-seq quality-control criteria, we detected transcripts from between 4,000 and 8,000 genes exceeding one transcript per million, consistent with previous measurements made using the method (additional scRNA-seq quality metrics are shown in **Supplementary Fig. 2**).

To assess the quality of the scBS-seq data, we compared the resulting single-cell methylomes with published data from 20 serum and 12 2i ESCs for which stand-alone scBS-seq was performed[3]. Sequencing of the scBS-seq libraries was performed at relatively low depth (an average of 11.1 million reads), with an average of 3.15 million genomic reads mapped per cell (**Supplementary Table 1**). We excluded cells with a mapping efficiency of <7% or a bisulfite-conversion efficiency of <95% (as estimated by non-CpG methylation). Cells passing these quality-control steps had a mean mapping efficiency of 15.6% (compared to a mean of 17.2% for single ESCs with stand-alone scBS-seq[3]; **Supplementary Table 1** and **Supplementary Fig. 3**). The low mappability was not due to foreign DNA, as negative controls showed less than 2% alignment, but it can be explained by high primer contamination (**Supplementary Fig. 3**).

Because of the decreased sequencing depth, methylome coverage in scM&T-seq libraries was lower than that in scBS-seq libraries. However, genome-wide CpG coverage at matched sequencing depth

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK. [2]Epigenetics Programme, Babraham Institute, Cambridge, UK. [3]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [4]Medical Research Council Functional Genomics Unit, University of Oxford, Oxford, UK. [5]Bioinformatics Group, Babraham Institute, Cambridge, UK. [6]Department of Human Genetics, Katholieke Universiteit Leuven, Leuven, Belgium. [7]These authors contributed equally to this work. Correspondence should be addressed to T.V. (thierry.voet@med.kuleuven.be), G.K. (gavin.kelsey@babraham.ac.uk), O.S. (oliver.stegle@ebi.ac.uk) or W.R. (wolf.reik@babraham.ac.uk).

RECEIVED 29 OCTOBER 2015; ACCEPTED 9 DECEMBER 2015; PUBLISHED ONLINE 11 JANUARY 2016; DOI:10.1038/NMETH.3728

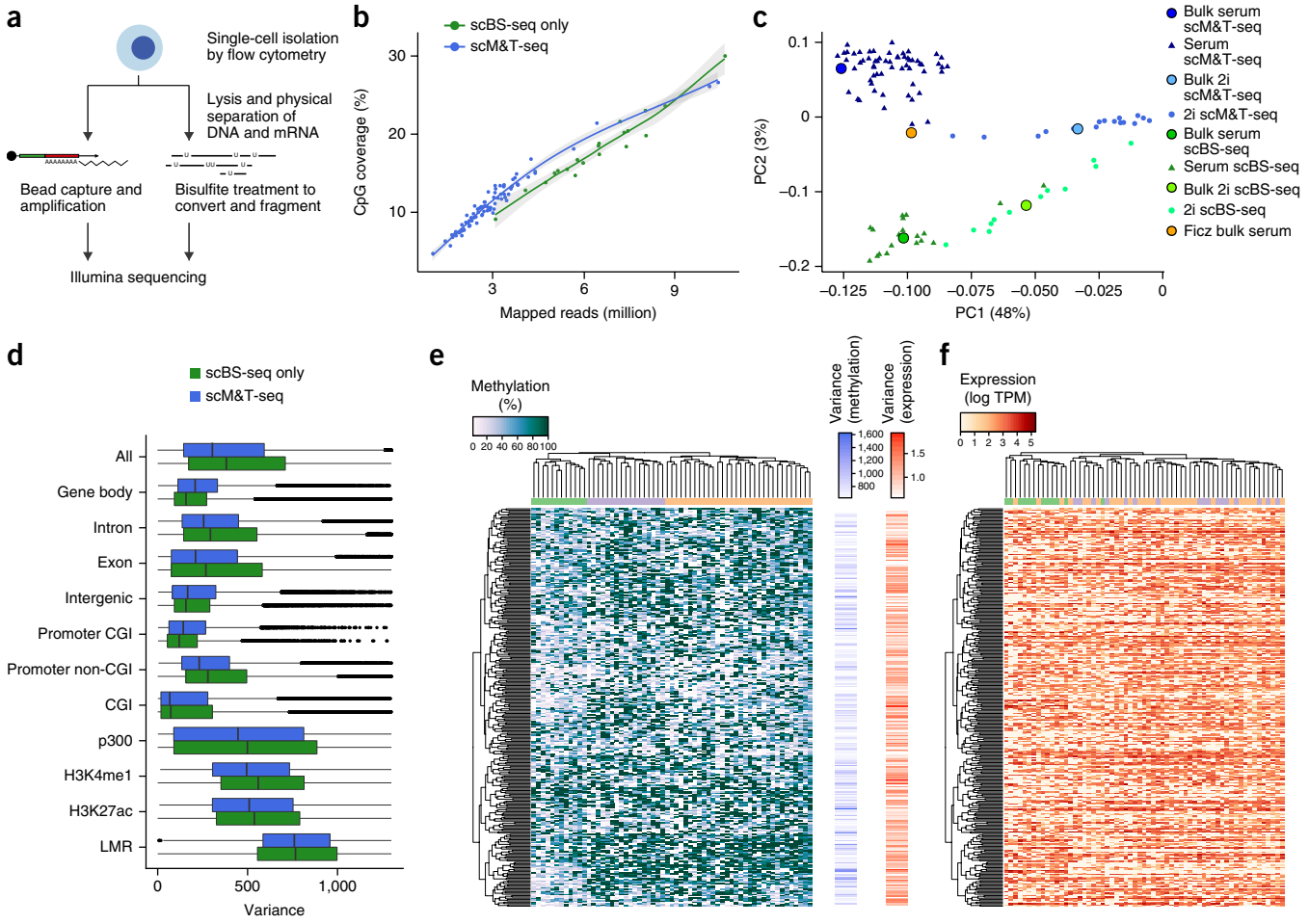

Figure 1 | Quality control and global methylation and transcriptome patterns identified in serum ESCs profiled using scM&T-seq. (**a**) Overview of the scM&T-seq protocol. (**b**) CpG coverage of single cells as a function of the number of mapped sequencing reads. Colored dots correspond to individual data points, and gray shaded areas denote the 95% confidence intervals of the locally fitted trend curves. (**c**) Joint principal-component analysis of the methylomes (gene-body methylation) of 61 serum ESCs and 16 2i ESCs obtained using scM&T-seq, as well as 20 serum ESCs and 12 2i ESCs sequenced using stand-alone scBS-seq[3]. Large outlined circles correspond to synthetic bulk data sets from the indicated cells. For comparison, we also included a bulk serum ESC DNA-methylation data set[16]. Cell type explained a substantially greater proportion of variance (PC1, 48%) than protocol did (PC2, 3%). (**d**) Comparison of epigenetic heterogeneity in different genomic contexts, considering 61 serum ESCs obtained using scM&T-seq and 20 serum ESCs sequenced using stand-alone scBS-seq[3]. (**e**,**f**) Clustering analysis of transcriptome and methylation data from 61 serum ESCs, considering gene-body methylation (**e**) and gene expression (**f**) for the 300 most heterogeneous genes (on the basis of gene-body methylation). The gene order was taken from individual clustering analysis on the basis of gene-body methylation, whereas cells were clustered separately using either DNA methylation or expression data and are color-coded by methylation cluster. The bar plots in the center show the heterogeneity in DNA methylation (left) and gene expression (right).

was consistent across protocols (**Fig. 1b**; **Supplementary Fig. 3** provides additional quality metrics, including an analysis of representation bias in different contexts), and we found that scM&T-seq covered a large proportion of sites in different genomic contexts with sufficient frequency to enable the analysis of epigenome heterogeneity across cells (**Supplementary Figs. 4** and **5**). To evaluate the potential coverage of scM&T-seq, we sequenced a randomly chosen subset of four libraries at increased depth (mean of 25.9 million raw reads), which yielded a CpG coverage in line with that of the other method (4.5 million, compared to 3.6 million mapped reads from 20.2 million raw reads for ESCs in stand-alone scBS-seq). Saturation depth was not reached in these four libraries (mean duplication rate of 25.5%), meaning that additional sequencing would yield greater coverage, as demonstrated previously[3].

As additional validation, we assessed the ability to discriminate serum and 2i ESCs using either stand-alone scBS-seq or scM&T-seq, and we found similar degrees of separation consistent with bulk data sets published previously[16] (**Fig. 1c**), with similar conclusions when using joint hierarchical clustering across all cells (**Supplementary Fig. 6**). Notably, the differences between protocols and biological batches had a substantially smaller effect (PC2, 3% variance) than cell type differences did (PC1, 48% variance), and by combining data across cells, we found that both protocols yielded genome-wide methylation profiles that accurately recapitulated bulk methylation profiles in the same cell type (**Supplementary Fig. 7**). Finally, we compared estimates of methylation heterogeneity in different genomic contexts, again finding good agreement between protocols (**Fig. 1d**). Taken together, these analyses provide confidence that the parallel scM&T-seq method yields results that are in agreement with data from stand-alone scBS-seq.

For subsequent analyses, we focused on serum ESCs only, as transcription and DNA methylation are uncoupled in 2i ESCs[16,19]. A comparison of the principal components derived from the two data types—gene-body methylation and gene expression—showed

that the global sources of variation were partially linked (**Supplementary Figs. 8** and **9**). However, a hierarchical clustering analysis of gene-body methylation and gene expression for the 300 most variable genes (on the basis of DNA-methylation variance; **Supplementary Fig. 10** presents alternatives) showed distinct clustering of cells when either source of information was used (**Fig. 1e**,**f**). This suggests that global methylome and transcriptome profiles reveal complementary, but distinct, aspects of cell state. This is also

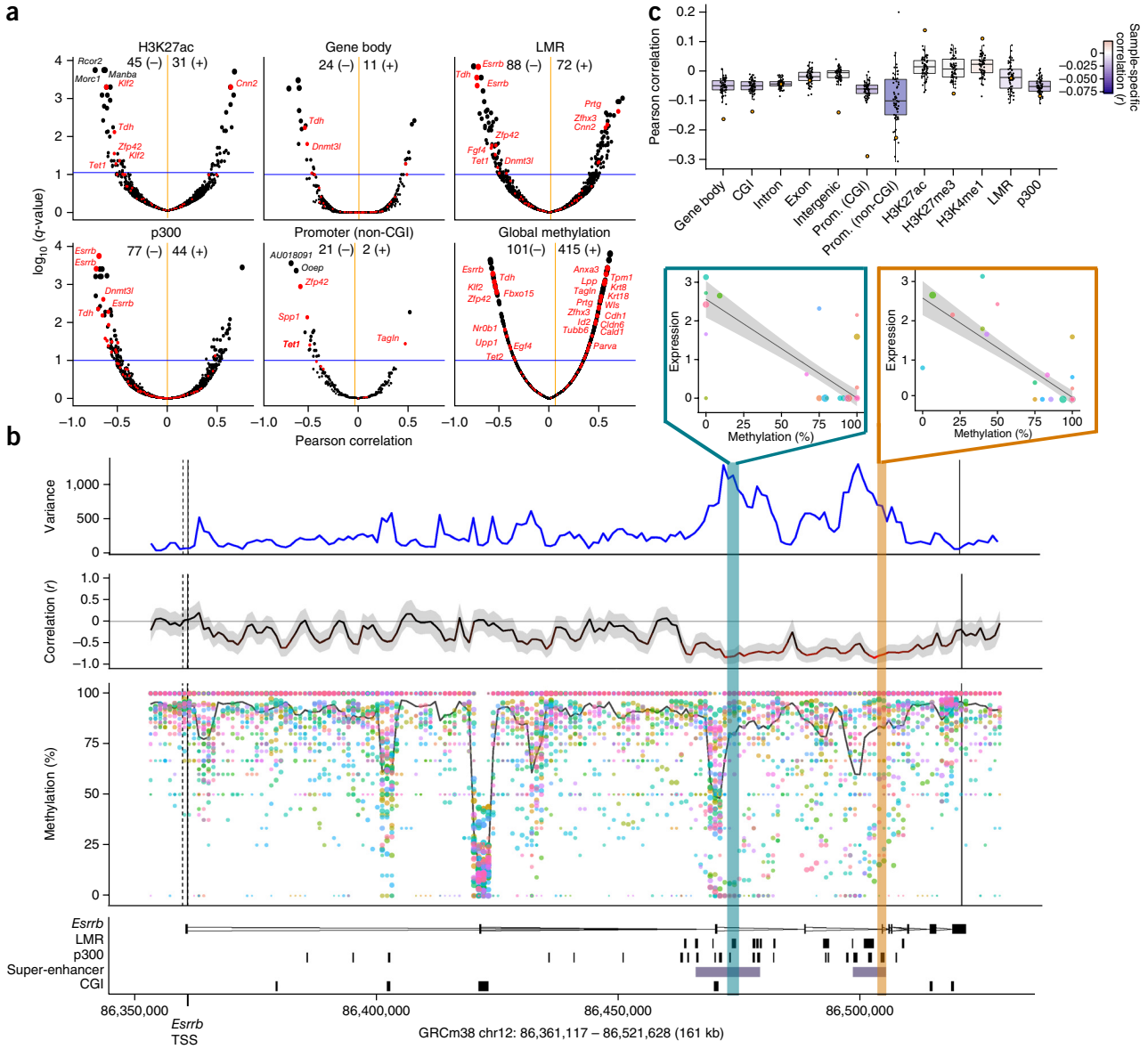

Figure 2 | Genome-wide associations between methylation and transcriptional heterogeneity in mouse ESCs. (**a**) Correlation coefficients (Pearson r^2) from association tests between gene expression heterogeneity of individual genes and DNA-methylation heterogeneity in alternative genomic contexts. Shown are the correlation coefficients for all genes versus the adjusted *P* value (obtained using Benjamini-Hochberg correction and denoted by dot size). A set of 86 known pluripotency and differentiation genes[18] are highlighted in red. The blue horizontal line in each plot corresponds to the FDR 10% significance threshold. The total number of significant positive (+) and negative (–) correlations (FDR < 10%) for each annotation is shown at the top of each plot. Orange vertical bars correspond to the average correlation coefficient across all genes for a given context. (**b**) Representative zoomed-in analysis for *Esrrb*. Shown from bottom to top are the annotation of the *Esrrb* locus with LMR, p300, super-enhancer and CGI sites indicated; the estimated methylation rate of 3-kb windows for each cell, with dot size corresponding to CpG coverage, dot colors corresponding to single cells, the solid black curve denoting the weighted mean methylation rate across all cells, and solid and dashed vertical lines delineating the position and transcription start site (TSS) of *Esrrb*, respectively; the correlation between the methylation rate and *Esrrb* expression for each region, with red shading in the curve corresponding to significant correlations (brighter red denotes higher significance) and the gray shaded area denoting the 95% confidence interval of the correlation coefficient; and the estimated weighted DNA-methylation variance between cells. The two scatter plots at the top right depict the association between DNA methylation at a p300 region (outlined in yellow) and at an LMR (outlined in blue) and *Esrrb* expression. (**c**) Gene-specific association analysis of correlations between DNA methylation in different genomic contexts and gene expression in individual cells. Shown are methylation-expression correlations for all variable genes in single cells, for each annotation, with the correlation obtained from matched RNA-seq and BS-seq of a bulk cell population superimposed[16] (orange circles). Prom., promoter. Upper and lower hinges correspond to 75th and 25th percentiles, upper and lower whiskers correspond to maximum and minimum values within the 1.5× interquartile range, and dots denote outliers.

consistent with previous observations that the transcriptome and methylome are partially uncoupled in serum ESCs[16].

Next we tested for associations between the expression of individual genes and DNA-methylation variation in several genomic contexts (Online Methods and **Supplementary Table 2**), and we identified a total of 1,493 associations (false discovery rate (FDR) < 10%; **Fig. 2a** and **Supplementary Tables 3** and **4**), which were robust when we used a bootstrapping approach to subsample the set of cells (**Supplementary Fig. 11**). We found both positive and negative associations, highlighting the complexity of interactions between the methylome and the transcriptome[9,10]. Although methylation of non–CpG island (CGI) promoters is known to be associated with transcriptional repression, the role of enhancer methylation is less clear. Accordingly, negative correlations between DNA methylation and gene expression were predominant for non-CGI promoters, whereas distal regulatory elements including low-methylation regions[20] (LMRs) had a more even balance of positive and negative associations (**Fig. 2a** and **Supplementary Figs. 12** and **13**). Associated genes were enriched for known pluripotency and differentiation genes[18] (FDR < 1%, Fisher's exact test; **Supplementary Table 5**). To our knowledge, our results provide the first evidence that heterogeneous methylation of distal regulatory elements (for example, LMRs) accompanies heterogeneous expression of key pluripotency factors in stem cell populations[6,21]. As an example, the expression of *Esrrb*, a known hub gene in pluripotency networks[22], negatively correlates with the methylation of several LMR and p300 sites overlapping 'super-enhancers' in the genomic neighborhood[23] (**Fig. 2b**). We also found 516 genes whose expression correlated with the overall methylation level (FDR < 10%), indicating substantial links between transcriptional heterogeneity and global methylation levels (**Fig. 2a**).

In addition to its utility in between-cell analyses, scM&T-seq can be used to correlate the methylome and transcriptome between genes in individual cells (**Fig. 2c** and **Supplementary Table 6**). We found that correlation between methylation and gene expression varied substantially between cells but was consistent in direction with matched RNA-seq and BS-seq data from a population of cells[16]. Again, this attests to scM&T-seq being sufficiently accurate to reliably study epigenome-transcriptome linkages. Our results also point to the possibility of heterogeneity between cells in the degree of coupling between the methylome and the transcriptome. Although we ruled out obvious confounding factors such as average methylation rate and sequence coverage (**Supplementary Figs. 14** and **15**), more data will be required for an understanding of the possible technical components in these linkages.

Our work demonstrates that parallel profiling of the methylome and transcriptome from the same single cell is feasible and can yield data similar in quality to those obtained with methods profiling either feature in isolation. Use of scM&T-seq allows the relationship between DNA methylation and expression to be studied at specific genes in single cells. We have confirmed a negative association between non–CGI promoter methylation and transcription in single cells and identified both positive and negative associations at distal regulatory regions. The expression levels of many pluripotency factors, such as *Esrrb*, were found to be negatively associated with DNA methylation, suggesting that an important mechanistic component of fluctuating pluripotency in serum ESCs is epigenetic heterogeneity. Finally, we have demonstrated that the strength of the connection between the methylome and the transcriptome can vary from cell to cell. scM&T-seq is a powerful approach for investigating the poorly understood connectivity between transcriptional and DNA-methylation heterogeneity in single cells and provides the potential to identify factors that regulate this relationship.

METHODS

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

Accession codes. scRNA-seq and scBS-seq data from all 92 ESC libraries and four negative controls are available in the Gene Expression Omnibus under accession GSE74535.

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

ACKNOWLEDGMENTS

We thank A. Kolodziejczyk and S.A. Teichmann for providing a list of 86 ESC pluripotency and differentiation genes[18]. We thank W. Haerty for his supervision and valuable advice to T.X.H. We thank the Wellcome Trust Sanger Institute sequencing pipeline team for assistance with Illumina sequencing. We thank the members of the Sanger–European Bioinformatics Institute (EBI) Single-Cell Genomics Centre for general advice. W.R. is supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC), the Wellcome Trust and the EU. G.K. is supported by the BBSRC, the UK Medical Research Council (MRC) and the EU. C.P.P. is supported by the Wellcome Trust and the MRC. T.V. is supported by the Wellcome Trust and KU Leuven (SymBioSys, PFV/10/016). H.J.L. is supported by EU Network of Excellence EpiGeneSys. O.S. is supported by the European Molecular Biology Laboratory (EMBL), the Wellcome Trust and the EU. This study reused some data from Smallwood *et al.*[3] available in the Gene Expression Omnibus under accession GSE56879.

AUTHOR CONTRIBUTIONS

C.A. performed all statistical analyses of the data. H.J.L., I.C.M., S.J.C. and S.A.S. developed the protocol and performed experiments. H.J.L., I.C.M., C.A., S.J.C., O.S., W.R. and C.P.P. interpreted the results. M.J.T. contributed to method development. T.X.H. processed RNA-seq data. F.K. processed BS-seq data. W.R., G.K., I.C.M. and T.V. contributed protocols and reagents. H.J.L., I.C.M., W.R. and T.V. conceived the project. W.R., O.S., T.V. and G.K. jointly supervised the project. O.S., H.J.L., S.J.C., W.R. and I.C.M. wrote the paper with input from all other authors. Names of authors who contributed equally to this work are ordered alphabetically on the first page.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Shapiro, E., Biezuner, T. & Linnarsson, S. *Nat. Rev. Genet.* **14**, 618–630 (2013).
2. Guo, H. *et al. Genome Res.* **23**, 2126–2135 (2013).
3. Smallwood, S.A. *et al. Nat. Methods* **11**, 817–820 (2014).
4. Farlik, M. *et al. Cell Rep.* **10**, 1386–1397 (2015).
5. Levsky, J.M., Shenoy, S.M., Pezo, R.C. & Singer, R.H. *Science* **297**, 836–840 (2002).
6. Yan, L. *et al. Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
7. Macaulay, I.C. *et al. Nat. Methods* **12**, 519–522 (2015).
8. Dey, S.S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. *Nat. Biotechnol.* **33**, 285–289 (2015).
9. Schübeler, D. *Nature* **517**, 321–326 (2015).
10. Jones, P.A. *Nat. Rev. Genet.* **13**, 484–492 (2012).
11. Singer, Z.S. *et al. Mol. Cell* **55**, 319–331 (2014).
12. Kalmar, T. *et al. PLoS Biol.* **7**, e1000149 (2009).
13. Chambers, I. *et al. Nature* **450**, 1230–1234 (2007).
14. Singh, A.M., Hamazaki, T., Hankowski, K.E. & Terada, N. *Stem Cells* **25**, 2534–2542 (2007).
15. Torres-Padilla, M.E. & Chambers, I. *Development* **141**, 2173–2181 (2014).
16. Ficz, G. *et al. Cell Stem Cell* **13**, 351–359 (2013).
17. Klein, A.M. *et al. Cell* **161**, 1187–1201 (2015).
18. Kolodziejczyk, A.A. *et al. Cell Stem Cell* **17**, 471–485 (2015).
19. Habibi, E. *et al. Cell Stem Cell* **13**, 360–369 (2013).
20. Stadler, M.B. *et al. Nature* **480**, 490–495 (2011).
21. Lee, H.J., Hore, T.A. & Reik, W. *Cell Stem Cell* **14**, 710–719 (2014).
22. Papp, B. & Plath, K. *EMBO J.* **31**, 4255–4257 (2012).
23. Whyte, W.A. *et al. Cell* **153**, 307–319 (2013).

ONLINE METHODS

Sample collection and single-cell sequencing. E14 mouse ESCs (the E14 cell line was a generous gift from A. Smith) were cultured in serum and leukemia inhibitory factor or in 2i media as described previously[16] and subject to routine mycoplasma testing using the MycoAlert testing kit (Lonza). Single cells were collected by flow cytometry after ToPro-3 and Hoechst 33342 staining to select for live cells with low DNA content (i.e., G_0 or G_1 phase cells). Cells were collected in RLT Plus lysis buffer (Qiagen) containing 1 U/µl SUPERase-In (Ambion) and processed using the G&T-seq protocol[7], except that after physical separation of mRNA and genomic DNA from single cells, the DNA was eluted into 10 µl of H_2O.

Single-cell bisulfite libraries were then prepared as previously described[3], with the following modifications. Conversion was carried out using EZ Methylation Direct bisulfite reagent (Zymo) on purified DNA in the presence of AMPure XP beads (Beckman Coulter) after G&T-seq. Purification and desulfonation of converted DNA were performed with magnetic beads (Zymo) on a Bravo workstation (Agilent), eluting into the master mix for the first-strand synthesis. Primers for first- and second-strand synthesis contained a 3′-random hexamer, and biotin capture of first-strand products was omitted, but an extra 0.8× AMPure XP purification was performed between second-strand synthesis and PCR. Each pre-PCR AMPure XP purification was carried out using a Bravo workstation. To avoid batch effects, we prepared all libraries in parallel in a 96-well plate. Purified scBS-seq libraries were sequenced in pools of 16–20 per lane of an Illumina HiSeq2000 using 125-bp paired-end reads.

RNA-seq libraries were prepared from the single-cell cDNA libraries using the Nextera XT kit (Illumina) as per the manufacturer's instructions, but using one-fifth volumes. Multiplexed library pools were sequenced on one lane of an Illumina HiSeq2000 generating 125-bp paired-end reads.

Sequence data processing and raw data analysis. *BS-seq read alignment.* Sequencing data were processed as previously described[3], with minor modifications. Briefly, we trimmed raw sequence reads to remove the first 6 bp (the 6N random priming portion of the reads), adaptor contamination and poor-quality base calls using Trim Galore (v0.3.8; parameters: --clip_r1 6 (or 9) --clip_r2 6 (or 9)). We aligned trimmed reads in single-end mode to the GRCm38 mouse genome assembly using Bismark[24] (v0.13.1; parameters: --bowtie2 --non-directional). Methylation calls were extracted after duplicate alignments had been removed. (Note: because of the multiple rounds of random priming with oligo 1, the single-cell bisulfite libraries were nondirectional.)

RNA-seq read-alignment and gene expression quantification. GSNAP[25] (version 2014.02.28) was used to align all RNA-seq libraries onto mouse genome assembly GRCm38 (with the --use-splicing option). For computation of the table of transcriptome raw read counts, an aligned read was counted toward a gene if it overlapped with any exonic region of that gene. To normalize transcriptome counts for library size, we used library size estimates obtained from DESeq2 (ref. 26). For computation of the transcriptome TPM (transcripts per million mapped reads) table, the output from cufflinks[27] (with the --frag-bias-correct --compatible-hits-norm --multi-read-correct option) was normalized to TPM values. Ensembl annotation (version 75) was used whenever gene annotations were required.

BS-seq and RNA-seq quality assessment. We included four negative controls (empty wells) in the library-preparation procedure to exclude the possibility of DNA or RNA contamination. Single-cell BS-seq libraries from negative controls had <2% mapping efficiency (the percentage of raw sequencing reads aligned), and scRNA-seq libraries from these samples had an alignment rate of <1%.

Single-cell BS-seq libraries with low alignment rates (<7% raw sequencing reads aligned) or poor bisulfite conversion (<95% on the basis of Bismark CHH and CHG methylation estimates) were excluded. Out of a total of 92 single-cell libraries, 81 passed this quality filter.

To identify low-quality scRNA-seq libraries, we required a minimum of 2 million mapped reads. Four serum and two 2i ESCs were excluded on the basis of this criterion (**Supplementary Table 1**).

Of the 92 single-cell samples, 75 (61 serum ESCs and 14 2i ESCs; 81.5%) passed quality assessment for both methylome and transcriptome sequencing. Complete quality-control data for both scRNA-seq and scBS-seq are provided as **Supplementary Table 1**.

Statistical analyses. *Clustering analyses.* The PCA analysis in **Figure 1c** was performed jointly on gene-body methylation of 12 2i and 20 serum cells profiled by stand-alone scBS-seq[3], 61 serum and 16 2i cells profiled by scM&T-seq, and a bulk BS-seq sample[16] and single-cell bulk methylation rates corresponding to genome-wide averages.

DNA methylation–gene expression association analysis. For association analyses, gene expression levels were considered on a logarithmic scale, using $\log_{10}$ normalized TPM counts (described above). Binary single-base-pair CpG methylation states were estimated from the ratio of methylated read counts to total read counts. The methylation rate in different genomic contexts, such as gene-body, promoter and enhancer annotations, was estimated as the mean CpG methylation rate in the region defined by the context (**Supplementary Table 2**). Following the approach of Smallwood *et al.*[3], we obtained weighted arithmetic mean and variance estimates for each context and cell, thereby accounting for differences in CpG coverage between cells.

For correlation analysis, genes with low expression levels or low expression and methylation variability between cells were discarded, according to the rationale of independent filtering[28]. First, a minimum expression level (at least 10 TPM counts) in at least 10% of all cells was required. From these, the 7,500 most variable genes were considered for analysis. Second, methylated regions were required to be covered by at least one read in at least 50% of all cells. For association tests, all possible relationships between genes and methylated regions within 10 kb of the gene (upstream and downstream of gene start or stop) were considered. Association tests were based on the weighted Pearson correlation coefficient, thereby accounting for differences in CpG coverage between cells. Precisely, let e be a vector with expression rates of cells for a particular gene, m be the methylation rate of the associated region, and w be the weight corresponding to the number of covered CpGs within the region. Then the weighted Pearson correlation $\mathrm{cor}(e,m;w)$ between gene expression e and methylation m is

$$\mathrm{cor}(e,m;w) = \frac{\mathrm{cov}(e,m;w)}{\sqrt{\mathrm{cov}(e,e;w)\,\mathrm{cov}(m,m;w)}}$$

Here cov(x,y;w) is the weighted covariance,

$$\mathrm{cov}(x, y; w) = \frac{\sum_i w_i (x_i - m(x; w))(y_i - m(y; w))}{\sum_i w_i},$$

and m(x;w) is the weighted arithmetic mean,

$$\mathrm{m}(x; w) = \frac{\sum_i x_i w_i}{\sum_i w_i}$$

Two-sided 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. For the zoom-in plot in **Figure 2b**, we considered a sliding-window approach (3-kb windows with a step size of 1 kb) to estimate the methylation rate in consecutive regions. Each region was tested for association with gene expression, again using weighted correlation coefficients as defined above.

To correlate the methylation and expression of a single cell across genes (**Fig. 2c**), we filtered genes in the same way as described above and again used the weighted Pearson correlation to test for associations.

R version 3.2.2 was used for all the analyses. The corresponding source code is available on GitHub (https://github.com/PMBio/scMT-seq.git) and as **Supplementary Software**. SeqMonk version 0.30 was used to compute methylation rates and CpG coverage for different regions (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/). Ensembl annotation (version 75) was used whenever gene annotations were required.

24. Krueger, F. & Andrews, S.R. *Bioinformatics* **27**, 1571–1572 (2011).
25. Wu, T.D. & Nacu, S. *Bioinformatics* **26**, 873–881 (2010).
26. Love, M.I., Huber, W. & Anders, S. *Genome Biol.* **15**, 550 (2014).
27. Trapnell, C. *et al. Nat. Biotechnol.* **28**, 511–515 (2010).
28. Bourgon, R., Gentleman, R. & Huber, W. *Proc. Natl. Acad. Sci. USA* **107**, 9546–9551 (2010).

doi:10.1038/nmeth.3728