nature biotechnology

Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction

Dylan Mooijman[1,2,4], Siddharth S Dey[1,2,4], Jean-Charles Boisset[1,2], Nicola Crosetto[1–3] & Alexander van Oudenaarden[1,2]

The epigenetic DNA modification 5-hydroxymethylcytosine (5hmC) has crucial roles in development and gene regulation[1–7]. Quantifying the abundance of this epigenetic mark at the single-cell level could enable us to understand its roles. We present a single-cell, genome-wide and strand-specific 5hmC sequencing technology, based on 5hmC glucosylation and glucosylation-dependent digestion of DNA, that reveals pronounced cell-to-cell variability in the abundance of 5hmC on the two DNA strands of a given chromosome. We develop a mathematical model that reproduces the strand bias and use this model to make two predictions. First, the variation in strand bias should decrease when 5hmC turnover increases. Second, the strand bias of two sister cells should be strongly anti-correlated. We validate these predictions experimentally, and use our model to reconstruct lineages of two- and four-cell mouse embryos, showing that single-cell 5hmC sequencing can be used as a lineage reconstruction tool.

In the past few years single-cell sequencing technologies have been developed to enable genome-wide quantification of mRNA or genomic DNA (gDNA) molecules in thousands of individual cells[8,9]. These techniques have enabled assessment of cell-to-cell gene expression variability and unbiased identification of novel cell types on a genome-wide level[10–13]. Whereas variability in gene expression has been extensively studied[14], the upstream mechanisms regulating cell-to-cell heterogeneity have been more difficult to study and are poorly understood.

Quantifying epigenetic marks at the single-cell level could increase our understanding of gene regulation and the underlying sources of cell-to-cell variability in gene expression. Recently, there has been rapid progress in genome-wide quantification of DNA methylation (5mC) in single cells using bisulfite sequencing methods[15–17]. Furthermore, it has been shown that 5mC can be converted to 5hmC by the TET family of enzymes, which has been proposed to represent an intermediate towards an unmethylated cytosine[1–3]. To date, genome-wide distribution of 5hmC in bulk samples has been quantified using 5hmC-specific antibodies, restriction enzymes or modified bisulfite sequencing approaches[4–7]. Extending this to understanding 5hmC occupancy in single cells could provide unique insights into the dynamics of DNA methylation turnover and the extent of cellular heterogeneity in this epigenetic mark[18].

To address these questions, we developed a technique that allows genome-wide detection of 5hmC marks in single cells, using the restriction endonuclease AbaSI. First, single cells are sorted into 384-well plates and, adopting a previously published method for bulk 5hmC sequencing[7], 5hmC marks are glucosylated using T4 phage β-glucosyltransferase. Next, this glucosylated form of 5hmC is recognized by AbaSI, which generates a double-stranded break with a 2-nucleotide overhang, 11–13 bp downstream of its binding site[7,19]. Subsequently, the digested gDNA is ligated to double-stranded adapters containing a 2-nucleotide random 3′ overhang, together with a cell-specific barcode, the Illumina 5′ adaptor and a T7 promoter. Thereafter, *in vitro* transcription is used to amplify the gDNA fragments linearly in a strand-specific orientation. Finally, the amplified RNA is fragmented and subjected to directional RNA library preparation (**Fig. 1a**).

We first applied single-cell hydroxymethylation sequencing (scAba-seq) to 480 E14tg2a (E14) mouse embryonic stem (mES) cells. For cells that were sequenced successfully we detected between 20,000 and 450,000 unique 5hmC sites per cell, with a median of 44,000 unique 5hmC sites per cell (**Supplementary Table 1** and **Supplementary Fig. 1**). As previously reported, we found that the restriction enzyme AbaSI cuts 11–13 bp downstream of the cut site, with greater than 92% of the reads displaying CpG dinucleotides at the correct position[7] (**Fig. 1b**). Furthermore, we did not observe a strong preference for symmetrical cytosines around the cut site as has been reported before[7] (**Supplementary Fig. 2**). Using synthetic 5hmC molecules (for a detailed explanation see **Supplementary Note 1**), we were able to determine a detection efficiency of approximately 10% and a false-positive detection rate of about 2% (**Supplementary Fig. 3**). Similarly, the distribution of 5hmC over different genomic elements in single cells was similar to that observed in previous bulk assays[7] (**Supplementary Fig. 4**). Finally, the averaged single-cell distribution of 5hmC over all chromosomes correlated strongly with previous bulk 5hmC sequencing data performed on the same cell line (Pearson $r = 0.89$ with bulk Aba-Seq and Pearson $r = 0.88$ with bulk TAB-Seq for 10-kb bin sizes; **Supplementary Fig. 5**). Taken together, these controls indicate that 5hmC sites detected by our scAba-seq method represent

[1]Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), Utrecht, the Netherlands. [2]University Medical Center Utrecht, Cancer Genomics Netherlands, Utrecht, the Netherlands. [3]Science for Life Laboratory, Division of Translational Medicine and Chemical Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [4]These authors contributed equally to this work. Correspondence should be addressed to A.v.O. (a.vanoudenaarden@hubrecht.eu).

Received 23 July 2015; accepted 11 May 2016; published online 27 June 2016; doi:10.1038/nbt.3598

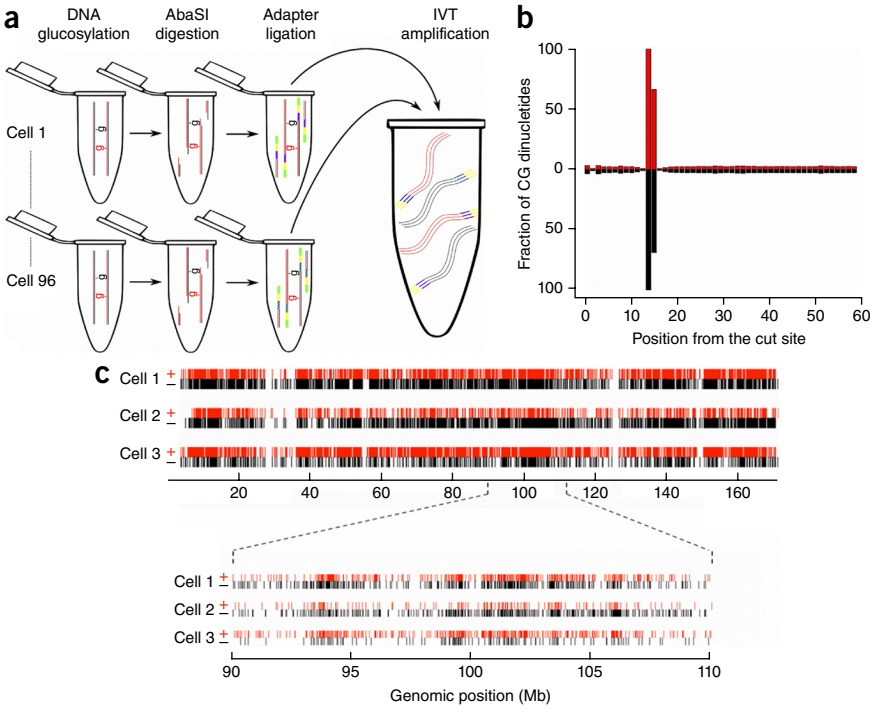

Figure 1 Schematic of the scAba-seq method. (**a**) Glucosylated DNA (plus strand in red, minus strand in black) of individual cells was digested with AbaSI. Digested DNA was ligated to an adaptor (blue) containing a cell-specific barcode (colored stripes), Illumina 5′ adaptor (yellow) and a T7 promoter (green). Ligated DNA from different cells were pooled and amplified using *in vitro* transcription (IVT) mediated by T7 polymerase. Amplified RNA containing cell-specific barcodes is used for generation of directional RNA sequencing libraries. (**b**) Histogram shows the relative amount of CG dinucleotides for each position along a 60-bp DNA stretch starting from the AbaSI cut site. The plus strand is indicated in red and the minus strand in black. (**c**) Density of 5hmC sites along the X chromosome in three individual mES cells, with plus strand 5hmC sites in red and minus strand 5hmC sites in black.

a faithful sampling of 5hmC distributions obtained from bulk experiments.

Upon closer inspection of the distribution of 5hmC sites in single cells, we observed substantial variability between the number of 5hmC sites on the plus and minus strands of the same chromosome (**Fig. 1c**). For example, data from chromosome X of three single cells show that although cell 1 had similar numbers of 5hmC sites on the plus and minus strands, the distribution of 5hmC along the entire chromosome was highly skewed toward the minus strand in cell 2 and the plus strand in cell 3 (**Fig. 1c**). Bulk sequencing approaches to quantify 5hmC have previously reported conflicting results on the existence of 5hmC strand bias in mES cells[6,7]. Because 5hmC strand bias can only be indirectly inferred from bulk sequencing approaches, we used single-cell 5hmC sequencing to investigate the strand bias more systematically.

To quantify this bias (denoted by f), we calculated the ratio of 5hmC present on the plus strand divided by the total number of 5hmC sites on both strands of each chromosome (**Fig. 2a**, top panel). For all chromosomes we observed symmetric distributions centered around $f = 0.5$. However, the width of these distributions was considerably greater than would be expected by random sampling of the reads from the two strands (**Fig. 2a**, middle panel). To exclude technical artifacts, we downsampled the number of 5hmC sites detected per chromosome from a bulk experiment (10,000 E14 mES cells) to the number of 5hmC sites observed in single cells, and confirmed that the experimentally observed strand bias distributions in single cells were considerably wider compared to the downsampled bulk distributions (**Fig. 2a**, bottom panel). When we downsampled the number of 5hmC sites per chromosome in single cells, we observed that above approximately 50 detected 5hmC sites/chromosome, the variance of the strand bias distributions was independent of the number of 5hmC sites (**Fig. 2b**). As we typically detect thousands of 5hmC sites per chromosome, the broad f distributions cannot be explained by a sampling artifact. Notably, we observed that whereas the autosomes displayed a wide unimodal distribution of strand bias, chromosome X in this male cell line showed a bimodal distribution (**Fig. 2a**,**c**).

Previous work has suggested that 5hmC marks are not maintained, in contrast to 5mC marks, which are copied to the newly replicated strand by the DNA maintenance methyltransferase DNMT1 (refs. 20,21). This prompted us to hypothesize that differences in strand age between the plus and minus strands of a chromosome could be a potential mechanism for generating the observed 5hmC strand bias. To gain quantitative understanding of how the dynamics of DNA hydroxymethylation and strand age affect strand bias, we constructed

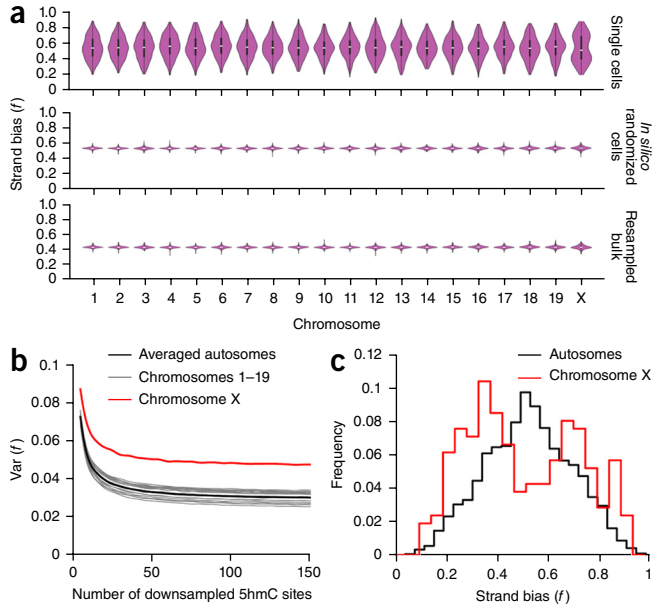

Figure 2 The levels of 5hmC reflect strong bias between the two strands of DNA of a chromosome. (**a**) Distribution of strand bias shown as violin plots for each chromosome from single cells (top panel), from *in silico* randomized single cells (middle panel) and from bulk 5hmC sequencing downsampled to reflect levels of 5hmC detected in single cells (bottom panel). (**b**) Variance of the 5hmC strand bias (f) distribution as a function of the number of downsampled 5hmC sites in single cells for the autosome average (black), individual autosomes (gray) and chromosome X (red). (**c**) Strand-bias (f) distribution for autosomes and chromosome X.

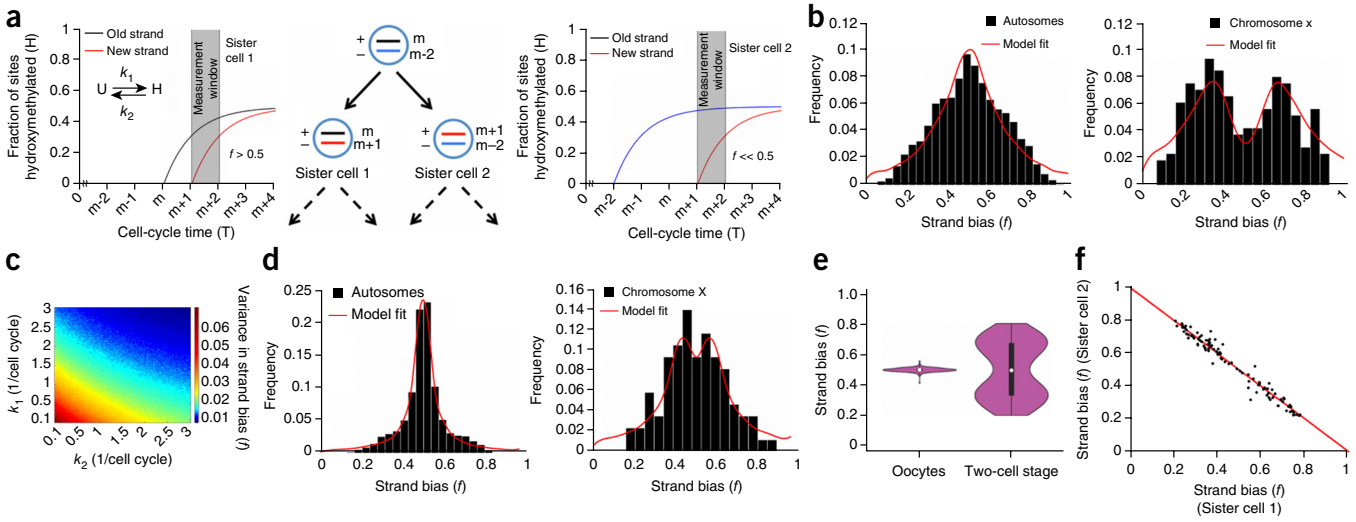

Figure 3 Stochastic model explains chromosome-wide strand bias. (**a**) Schematic showing a cell dividing to produce two sister cells, each of which inherits one old strand (black or blue) from the mother cell and a new strand (red) that is synthetized in the most recent S phase of the cell cycle (middle panel). + and – signs indicate the plus and minus strands, respectively. The age of each strand is indicated to the right of each cell (middle panel). The age of the strand generated in the mother cell is *m*. Gain and loss of 5hmC is modeled as first-order reactions with rate constants k_1 and k_2, respectively. The panels on the left and right indicate the fraction of sites that are hydroxymethylated for hypothetical values of k_1 and k_2 on the two strands of sister cell 1 and 2, respectively. The gray region indicates a time window of one cell cycle during which strand bias is measured. The model shows that $f > 0.5$ for sister cell 1 and $f < 0.5$ for sister cell 2, thereby capturing the bimodal distributions observed for chromosomes present in one copy. (**b**) The stochastic model (red) captures the wide experimental (black) strand bias distribution of autosomes and the bimodal distribution of chromosome X in E14 cells. (**c**) Theoretical simulations showing the variance of the strand bias distribution in autosomes for different mean values of k_1 and k_2. The variance of the strand bias distribution is reduced with increasing values of k_1 or k_2. For two-dimensional visualization of simulation results, the coefficient of variation of the distribution for k_1 and k_2 was fixed at 0.25. (**d**) As predicted by the stochastic model, strand bias in E14 cells is reduced upon treatment with vitamin C with the rates of 5hmC turnover, $k_1 + k_2$, increasing approximately twofold. (**e**) Distribution of strand bias shown as violin plots for oocytes ($n = 9$) and cells from haploid two-cell-stage embryos ($n = 11$). (**f**) Strand bias for each chromosome of sister cell 1 on the *y*-axis and sister cell 2 on the *x*-axis for each haploid two-cell embryo.

a stochastic model. Gain and loss of hydroxymethylation on each strand of a chromosome was modeled as a simple reversible reaction with k_1 and k_2 as the first-order rate constants for hydroxymethylation gain- and replication-independent loss, respectively (**Fig. 3a**). When a cell divides, its two daughter cells each inherit one strand of the original mother chromosome (black and blue strands, **Fig. 3a**). The other strands in the daughter cells (red strands, **Fig. 3a**) are newly replicated during the last S phase of the cell cycle. We assume that at this point the new strand does not have any 5hmC marks because of the absence of 5hmC maintenance[20,21]. Over time this new strand accumulates 5hmC marks and exponentially approaches a steady state where 5hmC gain and loss are balanced. We assume similar dynamics for the old strand with the important difference that the old strand is created *n* generations before the new strand. Assuming an exponentially dividing culture of mES cells, *n* is a random variable sampled from the distribution $p(n) = 2^{-n}$. Next, the bias *f* is calculated assuming that cells are sampled uniformly from a time window of one cell cycle after the birth of the new strand (**Fig. 3a**). We found that models with deterministic rate constants of k_1 and k_2 or those where the reversible reaction was modeled stochastically failed to explain the experimental data (**Supplementary Note 2** and **Supplementary Figs. 6–10**). Therefore, we constructed a stochastic model where k_1 and k_2 are treated as normally distributed random variables, which are independently sampled for the two strands. Variability in these rate constants could, for example, be caused by differences in the epigenetic state between the two strands of a chromosome[22].

We first applied this model to the diploid autosomes (**Fig. 3b**, left panel). Assuming random chromosome segregation, the model nicely fit the experimental data (red line, **Fig. 3b**) with best fits obtained for $(k_1 + k_2) = (0.85 \pm 0.14)$/division (**Supplementary Fig. 11**). When the model was fitted to the haploid X chromosome, we obtained a very similar value for this fit parameter: $(k_1 + k_2) = (1.02 \pm 0.18)$/division.

Further exploration of the model suggested that increasing the rate constants for hydroxymethylation gain and loss would result in a reduction of the variance of the strand bias distribution (**Fig. 3c**). To validate this model prediction, we treated E14 cells with vitamin C for 18 h before cell sorting, a factor previously shown to increase Tet activity and cause a rapid increase of 5hmC[23]. scAba-seq on 192 vitamin C-treated E14 cells showed that strand bias was reduced considerably for both the autosomes and chromosome X, as predicted by the model (**Fig. 3d**). Fitting the vitamin C-treated strand bias distributions to the model showed that the fit parameter $(k_1 + k_2)$ increased approximately twofold compared to untreated cells (**Fig. 3d**). Similar to the results of the untreated cells, fitting the autosomes and chromosome X to the model led to comparable hydroxymethylation turnover rates between the chromosomes $(k_1 + k_2) = (1.83 \pm 0.12)$/division and $(k_1 + k_2) = (2.29 \pm 0.61)$/division, respectively. When individual chromosomes were fit to the model, we obtained similar rates of hydroxymethylation (**Supplementary Fig. 12**).

To further confirm the effect of strand age on 5hmC strand bias, we induced parthenogenesis of mouse oocytes using $SrCl_2$ and subsequently analyzed uninduced mouse oocytes, and haploid two-cell-stage embryos isolated after cytokinesis. The haploid embryos allowed us to assess the bias of individual chromosomes without the presence of the confounding homologous pair. Comparing the distribution of the strand bias between oocytes and single cells from haploid two-cell embryos revealed that whereas oocytes showed no significant strand

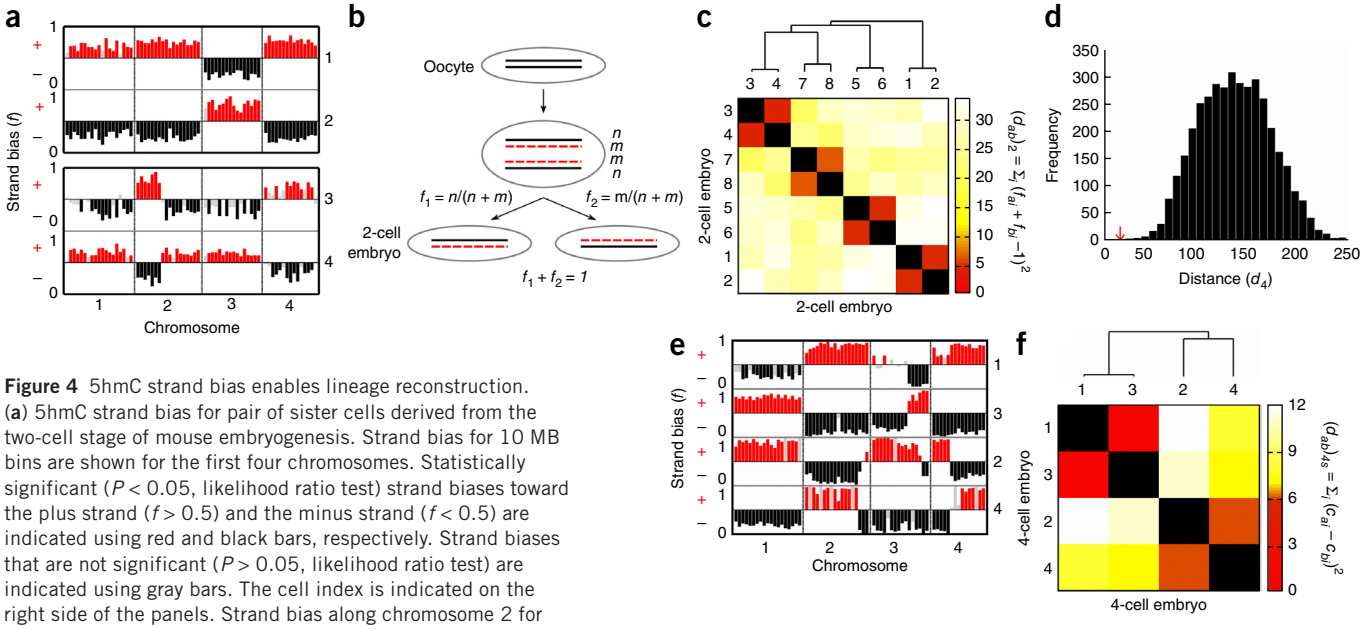

Figure 4 5hmC strand bias enables lineage reconstruction. (**a**) 5hmC strand bias for pair of sister cells derived from the two-cell stage of mouse embryogenesis. Strand bias for 10 MB bins are shown for the first four chromosomes. Statistically significant ($P < 0.05$, likelihood ratio test) strand biases toward the plus strand ($f > 0.5$) and the minus strand ($f < 0.5$) are indicated using red and black bars, respectively. Strand biases that are not significant ($P > 0.05$, likelihood ratio test) are indicated using gray bars. The cell index is indicated on the right side of the panels. Strand bias along chromosome 2 for sister cells 3 and 4 shows that 5hmC strand bias can be used to identify sister chromatid exchange events. The panels show that strand bias between sister cells is anti-correlated over all chromosomes. (**b**) Schematic demonstrating the theoretical basis of anti-correlation in 5hmC strand bias between two sister cells. If the old strand (black) and new strands (dashed red) contain n and m 5hmC sites per unit length of the genome, it can be shown that the strand bias between the two sister cells at each location in the genome sums to 1. (**c**) Heat map shows that the distance metric, $(d_{ab})_2$ can be used to identify pairs of sister cells in two-cell embryos. The distance between cells a and b derived from two-cell embryos, $(d_{ab})_2$ is estimated as the sum over all bins of the deviation of the sum of strand bias between cells a (f_{ai}) and b (f_{bi}) from 1. The lineage tree is constructed using the nearest neighbor joining algorithm. (**d**) Histogram of the distance (d_4) for groups of 4 cells chosen from the 19 cells sequenced from the four-cell stage of mouse embryogenesis. The distance is computed as $d_4 = \Sigma_i(f_{ai} + f_{bi} + f_{ci} + f_{di} - 2)^2$, where a, b, c and d represent the indices of the 4 cells chosen to compute the distance d_4. The red arrow indicates the distance between four cells $a = 1$, $b = 2$, $c = 3$ and $d = 4$ that derive from the same oocyte. (**e**) Strand bias for the first four chromosomes of the four cells that derive from the same oocyte. Strand biases are computed over 10 MB bins with statistically significant biases ($P < 0.05$, likelihood ratio test) on the plus ($f > 0.5$) and minus ($f < 0.5$) strands indicated using red and black bars, respectively. The cell index is indicated on the right side. The panels show that sister cells at the four-cell stage of mouse embryogenesis can be identified visually using sister chromatid exchange events where 5hmC strand biases transition abruptly along the length of a chromosome with the corresponding sister cell showing a mirrored pattern of strand bias about $f = 0.5$. For example, such a pattern of 5hmC strand bias can be observed by comparing chromosome 4 between cells 2 and 4. (**f**) The genomic location of sister chromatid exchange events can be used to identify sister cells within four-cell embryos. The heatmap shows that the distance metric, $(d_{ab})_{4s}$ can be used to identify cells 1-3 and 2-4 as sister cells. $c_{xi} = 1$, if 5hmC shows a transition in strand bias from one strand to another or 0 otherwise, where x indicates the cell index. The occurrence of transitions in 5hmC strand bias is estimated using the circular binary segmentation (CBS) algorithm. The lineage tree is constructed using the nearest neighbor joining algorithm.

bias, single cells from two-cell embryos displayed strong strand bias, resulting in a bimodal distribution (**Fig. 3e**). This resembled the f distribution of chromosome X that is present in one copy in the E14 mES cells (**Fig. 2c**).

Analysis of individual chromosomes in the sister cells of the two-cell embryos showed a strong anti-correlation between the bias of the chromosomes between sister cell 1 and sister cell 2 (**Fig. 3f**). This strongly suggests that each sister cell indeed receives an old strand containing 5hmC marks and a new strand containing fewer 5hmC marks. Further, the f distribution for these haploid chromosomes is bimodal because one sister cell receives the old plus strand whereas the other sister cell receives the old minus strand. Taken together, these results suggest that differences in strand age arising from replication and subsequent cell division are likely to cause variability in 5hmC strand bias between chromosomes in single cells. Previous work by Huh *et al.*[24] using immunofluorescence with anti-5hmC antibodies has shown that in asymmetrically dividing cells, newly synthesized DNA strands have lower amounts of 5hmC. Our observations suggest that this is a more generally occurring phenomenon, not exclusive to asymmetric cell divisions.

Based on the low rates of hydroxymethylation inferred from the model, we postulated that information about previous cell divisions would be retained, allowing us to infer sister-cell relationships and potentially reconstruct cellular lineages. Using two-cell embryos, we analyzed 5hmC strand bias within 10-Mb bins for each individual chromosome and found that each chromosome showed a mirrored pattern of strand bias about $f = 0.5$ between sister cells (**Fig. 4a**). Notably, we also observed that 5hmC strand bias can flip about $f = 0.5$ within a chromosome (**Fig. 4a**, bottom panel chromosome 2). This sharp transition in 5hmC strand bias is consistent with a putative sister chromatid exchange (SCE) event that occurs during the G2 phase of the cell cycle in the mother cell (**Supplementary Fig. 13**).

These observations of anti-correlation in 5hmC strand bias allowed us to develop a theoretical basis for identifying sister cells (**Fig. 4b** and **Supplementary Fig. 14**), where the sum of the strand bias between two sister cells at any location in the genome should equal 1. In general, as chromosomes exhibited similar strand biases along its entire length, such a strategy would enable accurate identification of sister cells from a large population of cells. Assuming random chromosome segregation, the probability of two random non-sister cells having mirrored strand bias pattern is 2^{-N}, where N is the number of chromosomes. For haploid murine cells ($N = 19$ autosomes) this probability is approximately 2×10^{-6}. The presence of SCE events results in an even greater discriminative power. We tested this strategy by analyzing eight cells obtained from four

two-cell mouse embryos. For all possible cell pairs (*a*,*b*) we calculated the following distance metric: $(d_{ab})_2 = \Sigma_{bins}(f_a + f_b - 1)^2$. Thus, if cell *a* and *b* are a sister pair $(d_{ab})_2 \approx 0$, whereas $(d_{ab})_2 >> 0$ for non-sister pairs. When hierarchical clustering is performed using this distance metric, all four sister pairs are correctly identified (**Fig. 4c**).

Generalizing this further, we next attempted to infer the lineage of four-cell embryos using a similar strategy, where the sum of 5hmC strand bias, at any location in the genome, for all four cells that derive from the same oocyte should equal 2 (**Supplementary Fig. 15**). We sequenced 19 single cells from the four-cell stage of mouse embryogenesis, where cells with labels 1 through 4 derived from the same oocyte. For all possible cell quadruplets (*a*, *b*, *c*, *d*) we calculated the distance metric: $d_4 = \Sigma_{bins}(f_a + f_b + f_c + f_d - 2)^2$. Indeed, we found that the combination of cells labeled 1, 2, 3 and 4 showed the smallest value for d_4, allowing us to accurately predict the four cells that were derived from the same oocyte out of a total of 3,876 different combinations of cell quadruplets (**Fig. 4d**). Next, to identify sister relationships between cells 1, 2, 3 and 4, we took advantage of the existence of SCE events that occur at the same location in the genome between sister, but not cousin, cells. Such a strategy is necessary to identify sister cells within four-cell embryos as 5hmC strand bias between sister cells at this stage does not sum to 1 (**Supplementary Fig. 15**). For the representative four-cell embryo (consisting of cells 1, 2, 3 and 4), we observed visually that the genomic location of SCE events are shared between cell pairs 1-3 and 2-4, respectively (**Fig. 4e**), indicating that these pairs of cells experienced an SCE event in their mother cell and therefore share a common history. By automating the identification of these putative SCE events over the entire genome and using a distance metric that utilizes the location of these SCE events to identify sister cells, we were able to assign cells 1-3 and 2-4 as sister cells within the four-cell embryo (**Fig. 4f**).

In summary, we developed a new single-cell sequencing technique to profile 5hmC on a genome-wide scale that led to the observation that in mouse embryonic stem cells and early mouse embryos, the two opposite strands of a chromosome can display dramatically different levels of 5hmC, with differences up to tenfold. Although proteins have been identified that specifically bind to 5hmC[25,26], it is currently not known if these proteins can detect 5hmC sites strand-specifically. Thus, it is possible that 5hmC strand bias, and therefore strand age, may serve as a source of chromosome-wide epigenetic memory to determine downstream protein activity and instruct biological processes such as chromosome segregation[24] or DNA repair[27]. Thus, scAba-seq could be a complementary method to Strand-seq[28] for analyzing strand age and SCE events in cell types containing 5hmC. Additionally, our observation that information about previous cell divisions is retained in 5hmC profiles allows for endogenous lineage reconstruction, bypassing the need for invasive cell labeling requirements. This could facilitate studies to assess sister-cell relationships in an unbiased manner in tissues not amendable to cell labeling, such as human primary material. Finally, our method of 5hmC detection is highly specific and because amplification is facilitated using *in vitro* transcription, combinations with other *in vitro* transcription-based single-cell technologies would allow for integration of multiple measurements from the same cell.

METHODS

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

Accession codes. GEO: GSE80973.

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

ACKNOWLEDGMENTS

We thank L. Kaaij, M. Bienko, M. Staps, M. Welling, W. Reik and members of the van Oudenaarden laboratory for constructive feedback. We would also like to thank S. van der Elst and R. van der Linden for their assistance with flow sorting and M. Verheul and E. de Bruijn at the Utrecht DNA Sequencing Facility for assistance with Illumina sequencing. This work was supported by NWO (VICI award) and ERC (ERC-AdG 294325-GeneNoiseControl) grants.

AUTHOR CONTRIBUTIONS

D.M., S.S.D. and A.v.O. designed the study. D.M., S.S.D. and J.-C.B. performed experiments. D.M., S.S.D. and A.v.O. performed analysis. S.S.D. and A.v.O. performed modeling. N.C. provided key inputs for initial technology development. D.M., S.S.D. and A.v.O. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
2. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
3. Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692–706 (2013).
4. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398–402 (2011).
5. Booth, M.J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
6. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
7. Sun, Z. *et al.* High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. *Cell Rep.* **3**, 567–576 (2013).
8. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
9. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. & Teichmann, S.A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
10. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
11. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
12. Jaitin, D.A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
13. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
14. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
15. Guo, H. *et al.* Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126–2135 (2013).
16. Smallwood, S.A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
17. Farlik, M. *et al.* Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.* **10**, 1386–1397 (2015).
18. Lee, H.J., Hore, T.A. & Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14**, 710–719 (2014).
19. Horton, J.R. *et al.* Structure of 5-hydroxymethylcytosine-specific restriction enzyme, AbaSI, in complex with DNA. *Nucleic Acids Res.* **42**, 7947–7959 (2014).
20. Valinluck, V. & Sowers, L.C. Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1. *Cancer Res.* **67**, 946–950 (2007).
21. Inoue, A. & Zhang, Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334**, 194 (2011).
22. Zhao, L. *et al.* The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome Res.* **24**, 1296–1307 (2014).
23. Blaschke, K. *et al.* Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature* **500**, 222–226 (2013).
24. Huh, Y.H., Cohen, J. & Sherley, J.L. Higher 5-hydroxymethylcytosine identifies immortal DNA strand chromosomes in asymmetrically self-renewing distributed stem cells. *Proc. Natl. Acad. Sci. USA* **110**, 16862–16867 (2013).
25. Spruijt, C.G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146–1159 (2013).
26. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14**, R119 (2013).
27. Shukla, A., Sehgal, M. & Singh, T.R. Hydroxymethylation and its potential implication in DNA repair system: A review and future perspectives. *Gene* **564**, 109–118 (2015).
28. Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).

ONLINE METHODS

Cell culture. E14tg2a mouse embryonic stem cells were obtained from American Type Culture Collection (ATCC CRL-182) and recently tested for mycoplasma contamination. Cells were grown on 0.1% gelatin in ES cell culture media; DMEM (1×) high glucose + glutamax (Gibco), supplemented with 10% FCS (Greiner) 100 µM β-mercaptoethanol (Sigma), 100 µM Non-essential amino acids (Gibco), 50 µg/mL Pen/Strep (Gibco) and 1000 U/mL ESGRO mLIF (Millipore). Cells were harvested before sorting by washing 3 times with 1× PBS with calcium and magnesium and incubated with 0.05% trypsin (Life Technologies). Cell were resuspended in ES culture media and cell clumps were removed by passing the cells through a BD Falcon 5mL polystyrene tube with a filter top. Cells were split every 2 days and media changed every day.

Vitamin C treatment. Vitamin C (L-ascorbic acid 2-phosphate, Sigma, A8960) was added to E14tg2a cells in a final concentration of 1 µg/mL 24 h after splitting the cells. After 18 h, the cells were washed three times with 1× PBS with calcium and magnesium and harvested using 0.05% trypsin (Life Technologies). Cell were resuspended in ES culture media and cell clumps were removed by passing the cells through a BD Falcon 5 mL polystyrene tube with a filter top.

Cell sorting. One cell per well was sorted into 4titude Framestar 384-well plates containing 4 µL VaporLock (Qiagen) and 0.2 µL lysis buffer. Doublets were removed by selecting for forward and side-scatter properties. Plates were centrifuged at 1,000 r.p.m. for 1 min directly after sorting to ensure that cells reach the aqueous phase.

Embryo isolation. B6/CBA mouse oocytes were obtained from four 3-month-old superovulated mothers (injected with pregnant mare serum gonadotropin (PMSG) and human chorionic gonadotropin (HCG) 22 h later) and incubated in 10 mM $SrCl_2$ (Sigma) for 2 h in M16 medium to induce parthenogenic activation. Developing parthenogenic embryos were monitored hourly and sister cells were isolated using hyaluronic acid (Sigma) and trypsin (Life Technologies) after cytokinesis.

scABa-seq. Robotic preparation: 4 µL Vapor-Lock (Qiagen) was dispensed manually into each well of a 384-well plate using a multichannel pipet. 0.2 µL of lysis buffer (1× New England BioLabs buffer 4, 0.1 µg Qiagen protease) was dispensed at 12 p.s.i. pressure into each well using a Innovadyne NanoDrop II robot. Single cells were sorted into each well, and the plate was incubated at 50 °C for 3 h, 75 °C for 20 min and 80 °C for 5 min. 0.2 µL of glucosylation mix (1× NEB buffer 4, 2 U NEB T4-BGT, 1x NEB UDP-Glucose) was added, and the plate was incubated at 37 °C for 18 h. Next, 0.2 µL of lysis buffer was added, and the plate was incubated at 50 °C for 3 h, 75 °C for 20 min and 80 °C for 5 min. 0.2 µL of AbaSI (1× NEB buffer 4, 2 U NEB AbaSI) was added, and the plate was incubated at 25 °C for 1 h and 65 °C for 20 min. 0.2 µL of 0.01 µM adaptor was added, followed by 0.2 µL of ligation mix (80 U T4-Ligase, 2X T4-ligase buffer, 0.6 mM ATP), and the plate was incubated for 16 h at 16 °C. Contents of all the wells with different adapters were pooled and incubated with 0.8× Agencourt Ampure XP beads for 30 min, washed twice with 80% ethanol and resuspended in 6.4 µL nuclease-free water. 9.6 µL of *in vitro* transcription mix was added (1.6 µL of each ribonucleotide, 1.6 µL T7 buffer, 1.6 µL T7 enzyme mix) and incubated at 37 °C for 14 h. Thereafter, library preparation was performed as described in the CEL-Seq protocol with minor adjustments[29]. After *in vitro* transcription, the aRNA (amplified RNA) was size-selected by incubating with 0.8× Agencourt RNAClean beads for 30 min, washed twice with 70% ethanol and resuspended in 16 µL nuclease-free water. Next, the aRNA is fragmented using fragmentation buffer (200 mM Tris-acetate, pH 8.1, 500 mM KOAc, 150 mM MgOAc) at 94 °C for 1.5 min and quenched by putting the reaction on ice and adding 2 µL 0.5 M EDTA. Fragmented aRNA is then purified using 1× Agencourt RNAClean beads and eluted in 16 µL nuclease-free water. After these steps, library preparation was done as described in the CEL-Seq protocol[29].

Manual preparation: cells were transferred with a mouth pipette into 20 µL Vapor-Lock in the cap of a 0.5 mL tube. Thereafter, all steps were performed using 1 µL volumes (instead of 0.2 µL volumes used in the robotic preparation) with a P2 Gilson pipette. Samples are pooled, purified using 0.8× Agencourt Ampure XP beads, and the library was prepared as described in the CEL-Seq protocol[29].

scAba-seq adapters. Adapters contain a T7 promoter, Illumina 5′ adaptor, cell-specific barcode and a random 2-nucleotide 3′ overhang. Top and bottom oligos were ordered separately and resuspended to 100 µM. The general sequence of the oligos used for the mES cell experiments are:

Top oligo:

5′-CGATTGAGGCCGGTAATACGACTCACTATAGGGGTTCAGAGTTCTACAGTCCGACGATCCA[6bp barcode]NN - 3′

Bottom oligo:

5′-[6bp barcode]GCGTGATGGATCGTCGGACTGTAGAACTCTGAACCCCTATAGTGAGTCGTATTACCGGCCTCAATCG - 3′

The 96 top and bottom oligo sequences can be found in **Supplementary Table 2**. Bottom oligos were phosphorylated by incubation in 10 µL kinase mix (1× NEB Ligase buffer, 20 units T4-PNK, 2 µL 10 mM ATP) at 37 °C for 1 h.

The general sequence of the oligos used for the mouse oocyte and haploid embryo experiments are:

Top oligo:

5′-CGATTGAGGCCGGTAATACGACTCACTATAGGGGTTCAGAGTTCTACAGTCCGACGATCNNN[8bp barcode]NN - 3′

Bottom oligo:

5′-5Phos[8bp barcode]NNNGATCGTCGGACTGTAGAACTCTGAACCCCTATAGTGAGTCGTATTACCGGCCTCAATCG - 3′

The 96 top and bottom oligo sequences can be found in **Supplementary Table 2**. These adapters contain an additional 3-nucleotide unique molecular identifier (UMI) between the cell barcode and 5′ Illumina adaptor and are synthesized with 5′ phosphorylation on the bottom oligo. All barcodes were designed with a minimal hamming distance of 2 to prevent a sequencing error from misidentifying a cell barcode. Finally, top and bottom oligos are annealed in a thermocycler starting at 98 °C, ramping down to 4 °C at 1 °C per min. Adapters are subsequently diluted to 0.01 µM in nuclease free water.

5hmC analysis pipeline. Libraries were sequenced on the Illumina Nextseq 500, and the fastq files were parsed for library barcodes. Read 1 was mapped to the mm10 build using the Burrows-Wheeler Aligner (BWA), filtered for unique mapping full-length reads and demultiplexed. 5hmCG positions were obtained from the sequencing data using custom scripts in R and Perl identifying CG dinucleotides in the mm10 genome at the expected distance from the mapped read position. All PCR duplicates were removed, and finally cells with fewer than 20,000 unique 5hmC sites were removed. Computer codes will be made available upon request.

Synthetic 5hmC molecules. Synthetic 5hmC molecules were generated via PCR amplification, using a hydroxymethylated forward primer amplifying a random 20-bp region on a plasmid library, generating a 514-bp molecule containing a single 5hmC site corresponding to the forward primer location. A detailed description can be found in the **Supplementary Note 1**.

The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

29. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).