Subtelomeric hotspots of aberrant 5-hydroxymethylcytosine-mediated epigenetic modifications during reprogramming to pluripotency

Tao Wang1,2, Hao Wu3, Yujing Li3, Keith E. Szulwach1, Li Lin1, Xuekun Li3, I-Ping Chen4, Ian S. Goldlust1, Stormy J. Chamberlain5, Ann Dodd1, He Gong1, Gene Ananiev6, Ji Woong Han7, Young-sup Yoon7, M. Katharine Rudd1, Miaoyu Yu8, Chun-Xiao Song8, Chuan He8, Qiang Chang8, Stephen T. Warren1,9,10 and Peng Jin1,10

Mammalian somatic cells can be directly reprogrammed into induced pluripotent stem cells (iPSCs) by introducing defined sets of transcription factors. Somatic cell reprogramming involves epigenomic reconfiguration, conferring iPSCs with characteristics similar to embryonic stem cells (ESCs). Human ESCs (hESCs) contain 5-hydroxymethylcytosine (5hmC), which is generated through the oxidation of 5-methylcytosine by the TET enzyme family. Here we show that 5hmC levels increase significantly during reprogramming to human iPSCs mainly owing to TET1 activation, and this hydroxymethylation change is critical for optimal epigenetic reprogramming, but does not compromise primed pluripotency. Compared with hESCs, we find that iPSCs tend to form large-scale (100 kb–1.3 Mb) aberrant reprogramming hotspots in subtelomeric regions, most of which exhibit incomplete hydroxymethylation on CG sites. Strikingly, these 5hmC aberrant hotspots largely coincide (~80%) with aberrant iPSC–ESC non-CG methylation regions. Our results suggest that TET1-mediated 5hmC modification could contribute to the epigenetic variation of iPSCs and iPSC–hESC differences.

Pluripotency is defined as a stem cell state with the potential to differentiate into any of the three germ layers. Somatic cells can be reprogrammed to a pluripotent state by defined factors such as OCT4, SOX2, KLF4, c-MYC, NANOG and LIN28 (refs 1–3). These iPSCs are extremely similar to ESCs. During the reprogramming process, the global epigenetic landscape in somatic cells has to be reset to reach a pluripotent state through DNA methylation/demethylation and chromatin remodelling processes.

Besides 5-methylcytosine (5mC), which is known to exhibit dynamic changes during early embryonic and germ cell development as well as the reprogramming process, the mammalian genome also contains 5hmC, which is generated by oxidation of 5mC by the TET family of enzymes4,5. The Tet proteins function in ESC regulation, myelopoiesis and zygote development6–10. 5hmC was found to be widespread in many tissues and cell types at different levels11,12. In particular, 5hmC is abundant in the central nervous system and ESCs. Several reports have explored the genome-wide distribution of 5hmC modification in mouse ESCs and hESCs, and suggest that it is enriched in gene bodies and enhancers13,14.

Reprogramming towards pluripotency involves a dynamic epigenetic modification process. 5hmC has been implicated in the DNA demethylation process15, pointing to a potential role for 5hmC modification during reprogramming towards pluripotency. Thus, understanding the dynamic 5hmC changes during reprogramming will provide further insight into somatic cell reprogramming mechanisms.

Multiple studies suggest that there are subtle yet substantial genetic and epigenetic differences between iPSCs and hESCs (refs 16,17). The present consensus is that iPSCs and ESCs are two overlapping classes of heterogeneous cells, with iPSCs being more variable than hESCs (ref. 18). Although iPSCs and hESCs are functionally equivalent in

1Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia 30322, USA. 2Genetics and Molecular Biology Graduate Program, Emory University, Atlanta, Georgia 30322, USA. 3Department of Biostatistics and Bioinformatics, Emory University Rollins School of Public Health, Atlanta, Georgia 30322, USA. 4Department of Oral Health and Diagnostic Sciences, School of Dental Medicine, University of Connecticut Health Center, Farmington, Connecticut 06030, USA. 5Department of Genetics and Developmental Biology, University of Connecticut Health Center, Farmington, Connecticut 06030, USA. 6Department of Genetics and Neurology, Waisman Center, University of Wisconsin-Madison, Madison, Wisconsin 53705, USA. 7Department of Medicine, Division of Cardiology, Emory University School of Medicine, Atlanta, Georgia 30322, USA. 8Department of Chemistry and Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois 60637, USA. 9Department of Biochemistry and Pediatrics, Emory University School of Medicine, Atlanta, Georgia 30322, USA. 10Correspondence should be addressed to S.T.W. or P.J. (e-mail: swarren@emory.edu or peng.jin@emory.edu)

Received 21 May 2012; accepted 5 April 2013; published online 19 May 2013; DOI: 10.1038/ncb2748
**RESULTS**

**TET1-mediated hydroxymethylation plays a critical role during reprogramming to pluripotency in human cells**

DNA methylation is a major barrier to iPSC reprogramming. Several lines of evidence suggest that 5hmC is involved in the process of DNA demethylation\(^\text{20,21}\). We found a significant increase of the 5hmC level in human iPSCs when compared with their original fibroblasts, with the amount in iPSCs being similar to that in hESCs (Fig. 1a).

TET family proteins (TET1, TET2 and TET3) could convert 5mC to 5hmC (ref. 6). We found a statistically significant increase in the level of TET1 and TET3, with a more pronounced increase of TET1, and a slight decrease of TET2 expression (Fig. 1b). RNA-seq reveals that TET1 is at a comparable level to NANOG in pluripotent cells, but the expression levels of TET2 and TET3 are significantly lower (Fig. 1c). Depletion of TET1 but not TET2 and TET3 by short interfering RNA (siRNA) could significantly decrease total 5hmC levels in human iPSCs (Fig. 1d and Supplementary Fig. S1a,b). Therefore, we conclude that TET1 is the main TET protein regulating hydroxymethylation during human iPSCs reprogramming.

As cellular reprogramming is an epigenetic-state reconfiguring process, we next examined whether TET1-mediated hydroxymethylation changes are critical in human iPSC reprogramming. Introducing TET1 short hairpin RNA (shRNA) lentivirus with Yamanaka factors demonstrated a slight decrease of 5hmC expression of sequencing data\(^\text{22}\), we found 267,664 regions in the genome showing differential 5-hydroxymethylation modification between iPSCs and fibroblasts (false discovery rate (FDR): 0.01), denoted as differential 5-hydroxymethylated regions (DhMRs). Among them, 231,866 are hyperDhMRs (5hmC level is higher in iPSCs), and 35,798 are hypoDhMRs (5hmC level is lower in iPSCs; Fig. 2b). The hyperDhMRs show a higher gain of 5hmC than the loss of 5hmC observed at hypoDhMRs (Fig. 2c). The hyperDhMRs are distributed across all autosomes, but largely missing in sex chromosomes (Fig. 2d). In particular, of the top 20,000 hyperDhMRs (ranked by adjusted \(P\) values), they have a higher probability (\(P < 0.0001\)) of being located in the telomere-proximal regions (Fig. 2e), as shown by example of chromosome 1 and chromosome X (Fig. 2f).

**5hmC is bi-directionally correlated with DNA methylation changes and associated with pluripotency-related gene networks**

The analysis described above suggests a global hydroxymethylation change during reprogramming. 5hmC has been suggested to be linked with gene expression in ESCs and neurons\(^\text{13,14,23–26}\). To assess the correlation between 5hmC modifications and gene expression changes during reprogramming, we stratified genes into 9 categories based on gene expression changes between iPSCs and fibroblasts (category 1: high expression in iPSCs, low expression in fibroblasts; category 2: medium expression in iPSCs, low expression in fibroblasts, and so on). We then quantified the amount of 5hmC around the transcription start site (TSS). As a result, those 9 categories can be clustered into 3 distinct patterns (Fig. 3a). Of note, most genes expressed during reprogramming show a bimodal distribution with a depletion of 5hmC in TSSs, whereas genes that remain silenced after reprogramming show a peak in TSSs. Among the 3 clusters, cluster 1 has the lowest 5hmC levels in TSSs; cluster 3 has the highest levels of 5hmC in TSSs, but has the lowest 5hmC levels in gene bodies (Fig. 3b).

We then examined the correlation between the absolute amount of transcripts and 5hmC enrichment. We noticed that hyperDhMRs tend to form a bimodal distribution associated with gene activity in iPSCs, with the lowest level similar to the level in fibroblasts in TSS regions (Fig. 3c and Supplementary Fig. S2). Transcription end site (TES) regions also show a bimodal distribution; the depletion is more pronounced in a narrower region centred on the TES (Supplementary Fig. S2). However, when compared with hypoDhMRs, hyperDhMRs are more enriched in TSSs, exons and TEs (Supplementary Fig. S3a). We observed a significant negative correlation between the 5hmC level of regions surrounding TSSs (±200 base pairs (bp)) and gene expression levels in iPSCs (Supplementary Fig. 3b).

We also observe a bidirectional correlation between the 5hmC level and DNA methylation during the reprogramming process. Eighty percent of the partially methylated domains (PMDs), which exhibit lower levels of CG methylation in somatic cells than stem cells,\(^\text{27}\), have
Figure 1  TET1 is associated with increased hydroxymethylation during human iPSC reprogramming. (a) Measurement of 5hmC levels in genomic DNA from fibroblasts, hiPSCs and hESCs by dot blot using anti-5hmC antibody. Mouse cerebellum genomic DNA was used as a control. For each sample, 225, 450 and 1,000 ng of DNA was used. (b) Quantitative RT–PCR to detect messenger RNA levels of TET1, TET2, TET3 and NANOG in fibroblasts (CRL2097) and hiPSCs (iPSC-B21, iPSC-B22). Error bars represent the standard error of the mean (s.e.m.) collected from three independent experiments. Scale bar, 300 μm. (c) Box plot of transcript copy numbers of TET1, TET2, TET3 and NANOG in IMR90 (broblasts) and hESCs (iPSC-B22) by dot blot using anti-5hmC antibody. Mouse cerebellum genomic DNA was used as a control. For each sample, 225, 450 and 1,000 ng of DNA was used. (d) Knocking down TET1 by siRNA significantly decreases 5hmC levels in hiPSCs. Left panel, TET1 siRNA knockdown efficiency by quantitative RT–PCR (* t-test, P < 0.05). Right panel, the effect of total 5hmC levels 48 h post TET1 siRNA transfection. Error bars represent s.e.m. collected from three independent experiments. (e) Alkaline phosphatase (AP) staining of reprogrammed cells treated with either TET1 shRNA lentivirus or an equal titre of control shRNA lentivirus after OSKM (OCT4, SOX2, KLF4, c-MYC) retroviral transduction of 100,000 CRL2097 cells on day 20. Cells used for staining were grown in 10 cm dishes. The image on the right shows a representative AP-positive colony and TET1 transcript level in TET1-shRNA- or control-shRNA-treated cells 10 days post transduction in one representative experiment of three independent experiments. Scale bar, 300 μm. (f) Summary of quantitative analysis of AP-positive colonies in three different experiments (* t-test, P < 0.05). Controls were normalized to 100%. Error bars represent the standard deviation (s.d.). (g) Real-time PCR analysis of TET1 and the pluripotency marker NANOG. TET1-shRNA-treated reprogrammed colonies maintained normal levels of NANOG, but shows decreased TET1 expression (* t-test, P < 0.05). Colonies were picked and maintained in puromycin medium (0.5 μg ml⁻¹) on puromycin-resistant mouse embryonic fibroblasts. (h) Real-time PCR analysis of normalized gene expression levels of TET1 and selected pluripotency-related factors in stable TET1 shRNA or control shRNA iPSC-B22 cells under puromycin selection (0.5 μg ml⁻¹); *** t-test, P < 0.05. Error bars represent the s.e.m. of three independent experiments. The raw values of related statistical test in this figure are listed in Supplementary Table S1.
Figure 2 Reprogramming confers a 5hmC epigenome in a pattern with a bias towards telomere-proximal regions in autosomes. (a) Pearson correlation analysis and clustering among fibroblasts and fibroblast-derived iPSCs. The values close to 1 indicate greater similarity. (b) Summary of the numbers of 5hmC differentially modified between fibroblasts and iPSCs, indicated by hyperDhMR (iPSCs > fibroblasts) and hypoDhMR (iPSCs < fibroblasts). The regions enriched either in fibroblasts or in iPSCs were subjected to DhMR calling. 5hmC-enriched regions in 3 fibroblast lines and 5 fibroblast-derived iPSC lines were coalesced into a union window. The reads in these windows were then recounted and normalized to the total read count from the respective cell line. A total of 267,664 DhMRs were called with a FDR of 0.01 by the Bioconductor Deseq package, which uses a negative binomial model for testing differential expression of sequencing data. Among them, 231,866 are hyperDhMRs, and 35,798 are hypoDhMRs. (c) Composite 5hmC enrichment profile for fibroblasts and iPSCs in the upstream regions of DhMRs, DhMRs and downstream regions of DhMRs. The length for upstream and downstream regions of DhMRs is 5 kb. (d) Chromosome ideograms showing the genome-wide distribution of the top 20,000 Fib–iPSC DhMRs ranked by the lowest adjusted P value. Blue lines indicate location of DhMRs. (e) Observed and expected numbers of hyperDhMRs occurring at telomere-proximal regions (χ² test, P value < 0.0001). Telomere-proximal regions were defined as regions at either end of a chromosome with a length equal to 1/20 of the total length of that chromosome. The observed number occurring at telomere-proximal regions is called by overlapping with the top 20,000 hyperDhMRs. The expected number is calculated on the basis of the proportion of total telomere-proximal region length compared with the whole length of all chromosomes. The top 20,000 hyperDhMRs were based on the 5hmC profiles of 3 fibroblast lines and 5 fibroblast-derived iPSC lines. (f) The distribution of the top 20,000 Fib–iPSC hyperDhMRs in chr1 and chrX.
Figure 3 5hmC is associated with gene activity and pluripotency regulatory networks in stem cells. (a) Three distinct clusters of 5hmC density pattern at TSS regions (±3 kb) in iPSCs and fibroblasts among 9 categories. The 9 categories were classified on the basis of the gene expression changes between iPSCs and fibroblast. Category 1: high expression in iPSCs, low expression in fibroblasts; category 2: medium expression in iPSCs, low expression in fibroblasts, and so on. (b) Box plots of hydroxymethylation levels between iPSCs and fibroblasts. Category 1: high expression in iPSCs, low expression in fibroblasts; category 2: medium expression in iPSCs, low expression in fibroblasts; category 3: medium expression in iPSCs, medium expression in fibroblasts; category 4: low expression in iPSCs, high expression in fibroblasts; category 5: low expression in iPSCs, medium expression in fibroblasts; category 6: medium expression in iPSCs, high expression in fibroblasts; category 7: medium expression in iPSCs, medium expression in fibroblasts; category 8: low expression in iPSCs, low expression in fibroblasts; category 9: high expression in iPSCs, high expression in fibroblasts. (c) Box plots of hydroxymethylation levels in TSS regions and gene bodies among the three clusters. *** indicates significantly more 5hmC levels compared with all others (P < 0.001, Wilcoxon rank test). Similarly, * indicates lowest 5hmC levels, ** indicates intermediate 5hmC level. (d) 5hmC density at the NANOG locus in input, iPSC and fibroblast cell lines. The position of the loci within the chromosome and the scale are shown above the gene tracks. Black lines indicate the DhMRs. (e) The overlap between NANOG-, OCT4-, KLF4- and SOX2-binding sites in ESCs and 5hmC significant change tracks. Black lines indicate the DhMRs. (f) The overlap between DhMRs in iPSC and fibroblast cell lines. The position of the loci within the chromosome and the scale are shown above the gene tracks. Black lines indicate the DhMRs. (g) Gene ontology analysis for genes overlapped with most significant DhMRs. (h) Plot of hyperDhMR and hypoDhMR densities in the context of the percentage of C+G, CG, CH and CHG.
Aberrant 5hmC reprogramming hotspots cluster at subtelomeric regions. (a) Pearson correlation analysis and clustering among 9 iPSCs and hESCs. Values close to 1 indicate greater similarity. (b) Chromosome ideograms showing the genome-wide distribution of 113 iPSC-ESC DhMRs. Red lines indicate locations of DhMRs. (c) The number of iPSC-ESC hyperDhMRs and iPSC-ESC hypoDhMRs. The 372,423 5hmC-enriched regions either in 9 iPSC lines or 4 hESC lines were subjected to DhMR calling by the Bioconductor Deseq package. This analysis led to the identification of 113 iPSC-ESC DhMRs that were differentially hydroxymethylated in at least one iPSC or ESC line (FDR < 0.01). Of the 113 iPSC-ESC DhMRs, 105 are hypo-hydroxymethylated, with 5hmC levels similar to their respective progenitors. (d) Complete linkage hierarchical clustering of 5hmC density within the iPSC-ESC DhMRs. The raw count values are scaled by rows during clustering. (e) Hierarchical cluster analysis using the top 1,000 most variable 5hmC-enriched regions across all iPSC and hESC samples. Arrows indicate hESCs.

Sequence preferences of 5hmC modification during reprogramming

We compared the CG, CH (CA, CT, CC) and CHG preference of hyperDhMRs and hypoDhMRs. HyperDhMRs tend to be located at higher C- and G-enriched regions, as well as CHG- and CH-enriched regions, whereas hypoDhMRs have the same level as the genome background (Fig. 3h). Previous observations suggest that 5hmC modification is related to CpG density. We find that in iPSCs, the low-CpG-content group of CpG islands tend to have more 5hmC modifications (Supplementary Fig. S3c), which is consistent with the observation that DNA methylation occurs more frequently in CpG islands with a low CpG content. Furthermore, 5hmC modifications acquired during reprogramming tend to occur within the unique sequence in which the methylation is evolutionarily less conserved (Supplementary Fig. S3d-f).

Aberrant 5hmC reprogramming hotspots cluster in telomere-proximal regions

Reprogramming of somatic cells to a pluripotent state requires complete reversion of the somatic epigenome into the pluripotent epigenome, which is an ESC-like state. iPSCs retain some type of somatic memory from their previous identity. We further determined the genome-wide 5hmC modification differences between iPSCs and ESCs, aiming to understand whether 5hmC modifications underlie the differences between hESCs and iPSCs. To reduce the biases of tissue origins, we used 9 iPSCs derived from different origins, associated with core pluripotency regulatory networks. We found that pluripotent master regulators, such as OCT3/4 and NANOG, bear this typical modification in iPSCs but not in fibroblasts (Fig. 3e). We further investigated the relation between 5hmC and key pluripotency factor-binding sites. Our results suggest that OCT4 and KLF4 regulatory networks may require 5hmC to regulate pluripotency during reprogramming. Furthermore, gene ontology analysis shows that genes acquiring most 5hmC are involved in stem cell differentiation and the patterning process, suggesting that 5hmC levels in stem cells are highly correlated with pluripotency.
Figure 5 5hmC DhMRs largely overlap with non-CG-DMRs in a large-scale pattern. (a) 5hmC density at the iPSC–ESC DhMR SIGLEC6, SIGLEC12 locus, in fibroblast (CRL2097), blood, iPSC and ESC lines. The position of the loci within the chromosome and the scale are shown above the gene tracks. Black bars indicate DhMRs. (b) The number of 5hmC DhMRs that overlap with CG-DMRs. CG-DMRs were categorized by methylation state relative to the ESCs. (c) The number of 5hmC large-scale hypoDhMRs that overlap with non-CG-DMRs. Non-CG-DMRs were categorized by methylation state relative to the ESCs reported previously17. The overlap was called for overlapping length larger than 1 kb. First bar summarizes the overlap for large-scale hypoDhMRs with hypo-non-CG-DMRs. The second bar summarizes the overlap for hypo-non-CG-DMRs with large-scale hypoDhMRs. The blue colour represents overlap between non-CG-DMR and hypoDhMRs. The red colour represents no overlap. (d) 5hmC density at the iPSC–ESC DhMR TCERG1L locus in fibroblast (CRL2097), blood, iPSC and ESC lines. The position of the loci within the chromosome and the scale are shown above the gene tracks. The lower parts show the 5mC levels in CH reported previously17. Black colour indicates H1 stem cells; green depicts iPSCs.

6 of which are from fibroblasts as mentioned earlier, 2 are derived from peripheral blood cells, and 1 is derived from human exfoliated deciduous teeth cells (SHED).

In general, global DNA hydroxymethylation patterns are very similar between iPSCs and ESCs (Fig. 4a). A comprehensive analysis of 372,423 5hmC-enriched regions between 4 hESC and 9 iPSC lines led to the identification of 113 iPSC–ESC DhMRs that were differentially hydroxymethylated in at least one iPSC or ESC line (FDR < 0.01), as shown for the SIGLEC6 and SIGLEC12 locus in Fig. 5a. Surprisingly, these regions are not randomly located across the genome; instead, they tend to cluster at the telomere-proximal regions, in particular, at chromosomes 3, 7, 8, 12 and 20 (Fig. 4b).

In contrast to the symmetric pattern of DMRs between iPSCs and ESCs (ref. 17), 105 of the 113 iPSC–ESC DhMRs are hypo-hydroxymethylated, with 5hmC levels similar to their respective progenitor blood cells or fibroblasts (Fig. 4c,d). Of these DhMRs, the 5hmC patterns are more variable when compared with hESCs (Fig. 4d). Unsupervised hierarchical clustering using the top 1,000 most variable 5hmC-modified regions among all samples could not distinguish hESCs from hiPSCs, suggesting that the variability among iPSCs is not due to...
different levels of pluripotency, and the 5hmC deviation of iPSCs is not a key determinant to distinguish hESCs from iPSCs (Fig. 4e).

Copy-number variation (CNV) has been reported to contribute to the variations of iPSCs (refs 34,35). As DhMRs cluster at subtelomeric regions and show depletion of hydroxymethylation, we further examined whether the DhMRs were simply due to genetic variation, such as CNV, instead of real aberrant 5mC epigenetic modification. To this end, we used a high-density comparative genomic hybridization (aCGH) array to examine 3 iPSCs and 2 hESC lines. We then compared DhMRs with the DMRs identified previously by Tet-assisted-bisulphite sequencing (TAB-seq) and the chemical capture approach is well correlated both genome-wide and within the proximal regions, forming aberrant reprogramming hotspots. To better define these large-scale regions, we developed a statistical method to identify potential large-scale aberrant reprogramming hotspots. An aberrant reprogramming hotspot is defined as a genomic region satisfying the following conditions: large variability of 5hmC levels among iPSCs; the average 5hmC difference between iPSCs and ESCs is statistically significant, and longer than 100 kb. Twenty large-scale regions were identified. Among them, 19 are hypoDhMRs, all of which have the same epigenetic status as their parent cells, pointing to a somatic memory during reprogramming, and 1 is hyperDhMRs (Table 1).

We then compared DhMRs with the DMRs identified previously using whole-genome single-base bisulphite sequencing, which would not be able to distinguish 5mC from 5hmC (ref. 17). Of the total 113 DhMRs, only 5 overlap with 1,175 CG-DMRs (Fig. 5b). Surprisingly, out of the 19 hypo large-scale hotspots, 84.2% overlap with the 24 mega-scale hypo-non-CG-DMRs, whereas the expected percentage is 1.6% based on permutation (Fig. 5c). Figure 5d shows one of these regions, chr10: 132,010,001–133,270,001; 5mCH are depleted in iPSC but not hESC lines. Similarly, of the 9 total iPSCs, only iPSC-S1 and iPSC-S2 derived from blood bear similar levels of 5mC compared to hESC counterparts. Of note, the variances from iPSCs are significantly larger than ESCs (Fig. 6a and Supplementary Fig. S5a,b). None of the iPSC lines has all of the 19 hypo large-scale DhMRs restored to the same level as the 4 hESC lines (Fig. 6b). This indicates that these large-scale regions tend to form aberrant reprogramming hotspots that were resistant to reprogramming. We did not observe a statistically significant ($P = 0.54$) correlation between passage number of iPSCs and the number of aberrant hotspots (Supplementary Fig. S5c), implying that passage number may not be a key determinant of hotspot number in each iPSC line.

The aberrant 5mC reprogramming hotspots we identified may also explain the transcription level variability in iPSCs. Notably, some of the genes, such as TCERG1L and FAM19A (Table 1), have been reported to be expressed at a significantly lower level in many but not all iPSCs as compared with ESCs (refs 36,37).

### Base-resolution 5mC analyses reveal large-scale hotspots are mainly caused by aberrant CG hydroxymethylation

The observed extremely high concordance between hypo large-scale DhMRs and non-CG-DMRs is surprising, and might indicate that of the previously identified aberrant 5mCH hotspot regions, a significant portion of CH consists of 5hmC; alternatively, these regions could contain both non-CG (mC) and CG (hmC) aberrant modification. Most 5hmC in ESCs is found at CG sites$^{38}$. In addition, 5mC quantification by Tet-assisted-bisulphite sequencing (TAB-seq) and the chemical capture approach is well correlated both genome-wide and within the 20 large-scale hotspot regions (Supplementary Fig. S6a,b). Therefore, it is very likely that the aberrant 5hmC is caused by CG modification.

To investigate this possibility experimentally, we applied TAB-seq, which can detect hydroxymethylation status at base resolution, to 2 hESC and 4 iPSC lines. We performed base-resolution analysis

| Chr        | Range (bp) | Length (bp) | Non-CG-DMR | No. of aberrant lines | Somatic memory | Genes                  |
|------------|------------|-------------|------------|-----------------------|----------------|------------------------|
| Chr1       | 4,533,001–5,059,001 | 526,001      | Y          | 5                     | Y              | AJAP1                  |
| Chr3       | 474,001–592,001      | 118,001      | N          | 9                     | Y              | Intergenic             |
| Chr3       | 2,515,001–2,907,001  | 392,001      | N          | 7                     | Y              | CNTN4                  |
| Chr7       | 152,805,001–153,016,001 | 211,001    | Y          | 8                     | Y              | Intergenic             |
| Chr7       | 153,184,001–153,312,001 | 128,001    | Y          | 8                     | Y              | DPP6                   |
| Chr7       | 153,461,001–153,856,001 | 395,001    | Y          | 6                     | Y              | DPP6                   |
| Chr7       | 154,010,001–154,317,001 | 307,001    | Y          | 6                     | Y              | DPP6                   |
| Chr8       | 2,681,001–3,289,001  | 608,001      | Y          | 7                     | Y              | CSMD1                  |
| Chr8       | 138,881,001–139,209,001 | 328,001    | Y          | 7                     | Y              | CSMD1                  |
| Chr8       | 139,536,001–139,818,001 | 282,001    | Y          | 5                     | Y              | FAM135B, COL22A1       |
| Chr10      | 132,010,001–133,270,001 | 1,260,001  | Y          | 7                     | Y              | TCERG1L, MIR378c       |
| Chr12      | 125,969,001–126,071,001 | 301,001    | Y          | 7                     | Y              | Intergenic             |
| Chr12      | 127,355,001–127,814,001 | 459,001    | Y          | 5                     | Y              | Intergenic             |
| Chr16      | 6,803,001–7,330,001  | 527,001      | Y          | 5                     | Y              | RBXO1                  |
| Chr17      | 73,780,001–74,420,001 | 640,001     | N          | 4                     | Y              | Intergenic             |
| Chr20      | 40,395,001–40,593,001  | 198,001     | Y          | 7                     | Y              | PTRPT                  |
| Chr20      | 41,004,001–41,305,001  | 301,001     | Y          | 7                     | Y              | PTRPT                  |
| Chr20      | 53,591,001–53,742,001  | 151,001     | Y          | 7                     | Y              | Intergenic             |
| Chr22      | 46,433,001–46,536,001  | 103,001     | Y          | 4                     | Y              | Intergenic             |

| Chr        | Range (bp) | Length (bp) | Non-CG-DMR | No. of aberrant lines | Somatic memory | Genes                  |
|------------|------------|-------------|------------|-----------------------|----------------|------------------------|
| Chr22      | 46,005,001–46,204,000 | 199,000     | N          | 6                     | Y              | LOC339685              |
of 5hmC in 3 randomly chosen large-scale regions, chr10, chr18, chr22, and amplified 5hmC-enriched regions by PCR (Fig. 7a and Supplementary Tables S6 and S7). We then subjected them to deep sequencing. Deep sequencing of PCR amplicons after traditional bisulfite conversion confirmed that there is epigenetic variation in non-CG sites but not CG sites (Fig. 7b,d). Consistent with the results obtained by the capture method, we observed similar 5hmC variations in iPSCs (Fig. 7c and Supplementary Fig. S6c,d). Importantly, this incomplete hydroxymethylation is caused by CG modification, but not CH modification (Fig. 7c and Supplementary Fig. S6c,d). For example, in the chr10 hotspot, iPSC-B22 and -B23 show incomplete 5hmC in CG dinucleotides, but not in CH dinucleotides (Fig. 7e). Therefore, our results suggest the coexistence of aberrant non-CG methylation and CG aberrant hydroxymethylation in subtelomeric hotspots (Fig. 7f). The concordance of aberrant CG hydroxymethylation with those aberrant CH large-scale regions suggests that there might be crosstalk between the epigenetic pathway that regulates hydroxymethylation and the pathway that regulates CH methylation; this crosstalk may behave more stochastically in those subtelomeric regions.

DISCUSSION

Our study suggests that the significant increase of 5hmC during reprogramming is mainly due to the activation of the TET1 protein in human iPSCs, which is in contrast to the previous observations that both Tet1 and Tet2 are upregulated in mouse iPSCs. Mouse ESCs are different from hESCs in many aspects, such as X-chromosome inactivation status in female lines. From a cell-signalling perspective, human pluripotency (primed pluripotency) depends mainly on FGF and activin–nodal signalling pathways, whereas mouse pluripotency (naive/ground-state pluripotency) is maintained by LIF–STAT pathways. The difference between human and mouse TET family proteins involved in reprogramming may be caused by FGF signalling selection of a subpopulation of hiPSCs. Several studies of generating naive human iPSCs under LIF signalling have been reported. Thus, it is possible that TET1 and TET2 have distinct roles in regulating pluripotency, with TET2 being involved in naive pluripotency and TET1 functioning in primed pluripotency. On the other hand, it is possible that TET1-mediated 5hmC modification is unique in humans regardless of different pluripotent stages. As TET1/2 is dispensable for maintaining the pluripotency of stem cells, and their loss is compatible with embryonic and postnatal development, it is likely that TET2 expression is not under positive selection for stem cell functions during evolution, and is thus eventually silenced in human pluripotent stages.

Reprogramming induces a remarkable epigenomic reconfiguration throughout the somatic cell genome. Recently, it was shown that TET1 and TET2, in synergy with NANOG, enhance the efficiency of mouse iPSC reprogramming. Here we show that TET1-mediated hydroxymethylation change is critical for optimal human iPSC reprogramming. We further show that TET1-mediated 5hmC modification affects
Figure 7 Large-scale hotspots are caused predominantly by aberrant CpG hydroxymethylation. (a) Summary of PCR-based TAB-seq. (b) 5hmC + 5mC single-base density in one of the amplicons by traditional bisulphite sequencing in 2 hESC and 2 iPSC lines. Bisulphite sequencing shows the CH methylation (or methylation plus hydroxymethylation) variation in iPSCs. The position of the loci within the chromosome and the scale are shown above the gene tracks. (c) 5hmC single-base density on CG sites in 15 amplicons by TAB-seq in 2 hESC and 4 iPSC lines. iPSC-B22 and -B23 show incomplete CG hydroxymethylation. Green indicates iPSCs bearing the same hydroxymethylation detected by 5hmC Capture-seq. Blue indicates iPSCs bearing incomplete hydroxymethylation detected by 5hmC Capture-seq in this region. (d) 5hmC + 5mC single-base density in 15 amplicons by traditional bisulphite sequencing in 2 hESC and 2 iPSC lines. (e) 5hmC single-base density on CG dinucleotides and CH dinucleotides in one of the amplicons (marked by a black dot in c) by TAB-seq in 2 hESC and 4 iPSC lines. Green indicates iPSCs bearing the same hydroxymethylation detected by 5hmC Capture-seq. Blue indicates iPSCs bearing incomplete hydroxymethylation detected by 5hmC Capture-seq in this region. (f) Schematic summary of large-scale incomplete hydroxymethylation on CG dinucleotides in iPSCs.

only the reprogramming efficiency, but does not alter the essential pluripotency in human stem cells. The pathways involving TET1 regulation largely remain unknown. It would be interesting to determine whether the known epigenetic factors such as DOT1L, Kdm2b and so on\(^{44,45}\) that are negative and positive modulators for reprogramming are linked to TET1-regulated hydroxymethylation modification.

Human iPSCs hold great promise for regenerative medicine and for establishing models of specific diseases. iPSCs and ESCs are known...
to share key features of pluripotency, including the expression of pluripotency markers, teratoma formation, cell morphology, the ability to differentiate into germ layers, and tetraploid complementation. Two models depict the equivalence, or lack thereof, between iPSCs and ESCs. One model posits that there may be small but consistent differences between ESCs and iPSCs, as suggested previously; the other model states that iPSCs and ESCs should be treated as two partially overlapping groups that share unique features. In this second model, single iPSC lines cannot be distinguished from ESC lines, although iPSCs show more epigenetic variance. Mounting evidence supports the latter model. Therefore, each iPSC may represent a unique epigenetic status with variable differentiation potential. The cause and degree of variation remain to be determined. Our study integrates the 5hmC epigenetic mark into the investigation of ESC-iPSC equivalence. We find that 5hmC occurs extensively in iPSCs at levels similar to ESCs, and there are no consistent 5hmC markers that can distinguish iPSCs from hESCs; however, we identified 20 regions in iPSCs that tend to form large-scale (100 kb–1 Mb) aberrant reprogramming hotspots, suggesting the present consensus that iPSCs are more epigenetically variable than ESCs. Remarkably, these regions with 5hmC variations tend to cluster in telomere-proximal regions. The close proximity of the hotspots to telomeres indicates there may be a distinct cellular process that could impede the reprogramming process.

Almost none of the DhMRs overlap with CG-DMRs, suggesting that CG-DMRs identified previously are primarily caused by DNA methylation. DNA methylation in non-CG contexts is abundant in pluripotent stem cells (mCHG and mCHH, where H = A, C or T), comprising almost 25% of all cytosines at which DNA methylation is identified. Strikingly, ~80% of large-scale iPSC–ESC DHMR regions coincide with previously reported non-CG DNA methylation aberrant hotspots. Reciprocally, ~50% of non-CG DMRs overlaps with our identified DhMRs. It was reported that non-CG DMRs also occur in the peri-centromeric zones. Notably, these peri-centromeric regions contain a low level of 5hmC (stem cells have similar levels of 5hmC as fibroblasts), suggesting that cells do not need to establish 5hmC in these regions during reprogramming (Supplementary Fig. S7). Thus, the concordance occurs mainly at telomere-proximal regions. By applying TAB-seq, we show that incomplete hydroxymethylation occurs predominantly at CG sites, but not CH sites, suggesting the coexistence of aberrant non-CG methylation and aberrant CG hydroxymethylation in these regions. During reprogramming, both CH methylation and hydroxymethylation need to be established from the somatic epigenome. It is known that non-CG cytosine methylation is exclusively catalysed by Dnmt3a and Dnmt3b (ref. 48). The concordance suggests that there might be crosstalk between epigenetics pathways that regulate the activities of Tet and DNMT3, which may behave more stochastically in those subtelomeric regions.

Our results indicate that TET1-mediated 5hmC modification contributes to both the human iPSC reprogramming process and differences between iPSCs and hESCs. In particular, we identified 20 large-scale aberrant hotspots, suggesting that iPSCs are more epigenetically variable than ESCs in terms of 5hmC modification. Our data suggest that, when studying aberrant epigenetic reprogramming events, as well as their functional consequences, at the DNA level, 5hmC modification merits particular consideration, in addition to 5mC.

**METHODS**

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

**Note:** Supplementary Information is available in the online version of the paper.

**ACKNOWLEDGEMENTS**

We thank J. Suhl, M. Santoro, S. Bray and C. Strauss for critical reading of the manuscript. We thank X. Huang from the Viral Vector Core of the Emory Neuroscience NINDS Core Facilities for preparing the retrovirus/ lentivirus used in this study. We are grateful to J. Mowrey, V. Patel, C. Street and S. Namburi for support. Illumina HiSeq2000/MiSeq sequencing. This study was supported in part by the National Institutes of Health (NS079625 and HD073162 to P.J.; MH089606 and HD24064 to S.T.W.), the Emory Genetics Discovery Fund, and the Autism Speaks grant (#7660 to X.L.).

**AUTHOR CONTRIBUTIONS**

T.W., S.T.W. and P.J. designed the study and interpreted the results. T.W. and H.W. analysed the data. T.W. performed most experiments; Y.L., L.L., X.L. performed 5mC capture and parts of library preparation. M.Y. C-X.S., H.G. and C.H. assisted with the TAB-seq experiment and 5hmC capture experiment. A.D. and K.E.S. contributed to the Illumina sequencing; I.S.G. and M.K.R. contributed array CGH experiments. I.P.C., S.I.C., I.W.H., G.A., Y.-x.Y. and Q.C. provided some of the hESC and hiPSC lines. T.W., S.T.W. and P.J. wrote the paper with assistance from H.W.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

Published online at www.nature.com/doifinder/10.1038/ncb2748
Reprints and permissions information is available online at www.nature.com/reprints

1. Takahashi, K. et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131, 861–872 (2007).
2. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663–676 (2006).
3. Yu, J. et al. Induced pluripotent stem cell lines derived from human somatic cells. Science 318, 1917–1920 (2007).
4. Tahiliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324, 930–935 (2009).
5. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science 324, 929–930 (2009).
6. Ito, S. et al. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature 466, 1129–1133 (2010).
7. Ko, M. et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature 468, 839–843 (2010).
8. Koh, K. P. et al. Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell 8, 200–213 (2011).
9. Wossidlo, M. et al. 5-hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nature Commun. 2, 241 (2011).
10. Dawlaty, M. M. et al. Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. Cell Stem Cell 9, 166–175 (2011).
11. Song, C. X. et al. Selective chemical labelling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat. Biotechnol. 29, 68–72 (2011).
12. Globisch, D. et al. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. PloS One 5, e15367 (2010).
13. Pastor, W. A. et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature 473, 394–397 (2011).
14. Szulwach, K. E. et al. Integrating 5-hydroxymethylcytosine into the epigenetic landscape of human embryonic stem cells. PLoS Genet. 7, e1002154 (2011).
15. Wu, H. & Zhang, Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. Genes Dev. 25, 2436–2452 (2011).
16. Bock, C. et al. Reference maps of human ES and iPSC cell line variation enable high-throughput characterization of pluripotent cell lines. Cell 144, 439–452 (2011).
17. Lister, R. et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471, 68–73 (2011).
18. Robijnfrute, D. A. & Daley, G. Q. The promise of induced pluripotent stem cells in research and therapy. Nature 481, 295–305 (2012).
19. Huang, Y. et al. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PLoS One 5, e8888 (2010).
20. Guo, J. U., Su, Y., Zhong, C., Ming, G. L. & Song, H. Hydroxylation of 5-methylcytosine by Tet1 promotes active DNA demethylation in the adult brain. Cell 145, 423–434 (2011).
21. Cortellino, S. et al. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146**, 67–79 (2011).
22. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
23. Szulwach, K. E. et al. 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and ageing. *Nat. Neurosci.* **14**, 1607–1616 (2011).
24. Williams, K. et al. TET1 and hydroxymethylcytidosine in transcription and DNA methylation fidelity. *Nature* **473**, 343–348 (2011).
25. Ficz, G. et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398–402 (2011).
26. Wu, H. et al. Genome-wide analysis of 5-hydroxymethylcytidosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev.* **25**, 679–684 (2011).
27. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
28. Xu, Y. et al. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Mol. Cell* **42**, 451–464 (2011).
29. Meissner, A. et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
30. Cohen, N. M., Kenigsberg, E. & Tanay, A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145**, 773–786 (2011).
31. Kim, K. et al. Epigenetic memory in induced pluripotent stem cells. *Nature* **467**, 285–290 (2010).
32. Kim, K. et al. Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. *Nat. Biotechnol.* **29**, 1117–1119 (2011).
33. Polo, J. M. et al. Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat. Biotechnol.* **28**, 848–855 (2010).
34. Hussein, S. M. et al. Copy number variation and selection during reprogramming to pluripotency. *Nature* **471**, 58–62 (2011).
35. Laurent, L. C. et al. Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell* **8**, 106–118 (2011).
36. Chin, M. H. et al. Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* **5**, 111–123 (2009).
37. Guenther, M. G. et al. Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell* **7**, 249–257 (2010).
38. Yu, M. et al. Base-resolution analysis of 5-hydroxymethylcytidosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
39. Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell Stem Cell* **4**, 487–492 (2009).
40. Wang, W. et al. Rapid and efficient reprogramming of somatic cells to induced pluripotent stem cells by retinoic acid receptor gamma and liver receptor homolog 1. *Proc. Natl Acad. Sci. USA* **108**, 18283–18288 (2011).
41. Hanna, J. H., Saha, K. & Jaenisch, R. Pluripotency and cellular reprogramming: Facts, hypotheses, unresolved issues. *Cell* **143**, 508–525 (2010).
42. Dawlaty, M. M. et al. Combined deficiency of tet1 and tet2 causes epigenetic abnormalities but is compatible with postnatal development. *Dev. Cell* **24**, 310–323 (2013).
43. Costa, Y. et al. NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. *Nature* **495**, 370–374 (2013).
44. Oder, T. T. et al. Chromatin-modifying enzymes as modulators of reprogramming. *Nature* **483**, 598–602 (2012).
45. Liang, G., He, J. & Zhang, Y. Kdm2b promotes induced pluripotent stem cell generation by facilitating gene activation early in reprogramming. *Nat. Cell Biol.* **14**, 457–466 (2012).
46. Daley, G. Q. et al. Broad implications of defining standards for the pluripotency of iPSCs. *Cell Stem Cell* **4**, 200–201 (2009).
47. Stadtfeld, M. et al. Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* **465**, 175–181 (2010).
48. Dyachenko, O. V., Schevchuk, T. V., Kretzner, L., Buryanov, Y. I. & Smith, S. S. Human non-CG methylation: are human stem cells plant-like? *Epigenetics* **5**, 569–572 (2010).
METHODS

iPSC reprogramming and cell culturing. Human fibroblasts IMR90 and CRL2097 were obtained from ATCC, and GM0011 was obtained from Coriell Cell Repositories. The fibroblasts were cultured in DMEM medium containing 10% FBS, 1x glutamine, 1x penicillin/streptomycin. The H1 hESC and iPSC IMR90 were obtained from WiCell. HUES48, HUES49 and HUES53 were obtained from the Human Embryonic Stem Cell Collection at Harvard University. The cells were maintained in hESC/hiPSC standard medium (DMEM/F12, 20% KnockOut Serum Replacement, 1x non-essential amino acids, 1x glutamine, 0.11 mM 2-mercaptoethanol, 10 ng/ml β-EGF) on irradiated mouse embryonic fibroblast (MEF) feeders. The medium was changed every 48 h.

We focus on efficient reprogramming methods mainly by retrovirus, lentivirus and Sendai virus, all known to have distinct behaviors in establishing iPSCs (refs 49, 70). As the stoichiometry of reprogramming factors can influence the epigenetic status of iPSCs (refs 47, 50), we included the iPSCs reprogrammed by Yamanaka factors and Thomson factors either in polycistronic vectors or separate vectors.

For human iPSC-A2, -B2 and -B23 reprogramming, 2 x 10^5 fibroblasts were seeded in a well of a 6-well plate on day 1. On day 2, 10 μl of concentrated pMXs-hoCT4, hSoX2, hε-MYC and hKLF4 retrovirus were added to the cells in the presence of 6 μg/ml Polybrene. A second round of transduction was repeated on day 3. On day 7, the cells were reseded in 10 cm dishes with irradiated MEF feeders. The potential hiPSC colonies were picked between days 18 and 25. The established iPSC cell lines were subsequently confirmed with alkaline phosphatase staining, immunofluorescence staining of pluripotent markers and the ability to differentiate into 3 germ layers. iPSC-RX35i and iPSC-RX35i2 were derived from fibroblasts by STEMCCA lentivirus. hPSC51 and hiPSC52 lines were generated from peripheral blood mononuclear cells (PBMCs) of 2 healthy volunteers using Sendai virus (CytoTune-iPS Kit; kindly provided by and property of DNAVEC), which are presumably free of transgene integration. To transduce cells, 4 separate sets of Sendai viruses containing hOct3/4, hSoX2, hNANOG and hLIN28 lentiviruses, and iPSC-RX35i were delivered from fibroblasts by STEMCCA lentivirus. hPSC51 and hiPSC52 were established from fibroblasts by STEMCCA lentivirus. hPSC51 and hiPSC52 were subsequently confirmed with alkaline phosphatase staining, immunofluorescence staining of pluripotent markers and the ability to differentiate into 3 germ layers. IPSC-AG2.3 and iPSC-RX35i were reprogrammed in a similar protocol. The final concentration for siRNA is 50 nM or 100 nM. At 48 h post transfection, cells were denatured with NaOH according to the Illumina protocol (final template concentration is 1 nM). Denatured libraries were diluted to a final concentration was further confirmed with real-time PCR by using the DNA 2.1 PCR purification kit (Qiagen). Click chemistry was performed with the addition of 150 μM disulpahide-biotin linker, and the mixture was incubated for 2 h at 37 °C. The biotin-labelled DNA samples were then captured by using Dynabeads MyOne Streptavidin C1 (Invitrogen) following the manufacturer’s recommendations except eluted in dithiothreitol. Subsequently, the 5hmC-enriched DNA was purified by the MinElute PCR Purification Kit (Qiagen) and eluted with 12 μl elution buffer.

Quantitative RT–PCR. Total RNA was extracted using the RNeasy Kit (Qiagen). Total RNA (2 μg) was converted to complementary DNA by using the Script cDNA Synthesis Kit (Bio-Rad). The cDNA was then diluted to 1:200 and 8 μl of each diluted template was subjected to PCR amplification in a 20 μl volume mixed with Power SYBR Green Master Mix (Applied Biosystems). The PCR conditions were an initial 5 °C denaturation for 10 min followed by amplification cycles consisting of 95 °C for 10 s, 60 °C for 10 s, and 72 °C for 10 s for 40 cycles. For data analysis, the results were normalized with GAPDH signal. For iPSC colony analyses, cells were lysed and subjected to reverse transcription by using a Cells-to-Ct Kit (Life Technologies) according to the manufacturer's protocol. Primer sequences are listed in Supplementary Table S5.

Library preparation and Illumina sequencing for 5hmC captured DNA. Approximately 50 ng of each 5hmC-enriched DNA was used for Illumina SR library preparation by NEBNext ChIP-seq Sample Prep Reagent Set (NEB) according to the manufacturer’s standard protocol. The sequences of the adapters used for ligation are: 5'-P-GATCGGAAGAGTCGCTTAGTCTGGCTTCTGTTG-3' and 5'-ACATCTTCTTCTCCTAAGGCGCTCCGGGATCTC-3'. The PCR primers used for the amplification step are: PCR1.1, 5'-AATGATACGGGACACAGAGATCT-ACACCTTCCTCTACAGGCGCTCCGGGATCTC3'-3'; PCR2.1, 5'-CAACAGAAGAGAGCGCTCAGGCGCTCCGGGATCTC3'-3'. The concentration of each DNA library was determined in DNA 1000 chip by Agilent 2100 Bioanalyzer. The library concentration was further confirmed by real-time PCR using the DNA Library Quantification Kit (KapaBiosystems). The libraries were then diluted to 10 pm for sequencing on Illumina HiSeq/SiHiScanQ systems. Briefly, DNA libraries were denatured with NaOH according to the Illumina protocol (final template concentration is 1 nM). Denatured libraries were diluted to a final concentration of 6 pm. Each sample was then spiked with 1% PhiX control. The libraries were clustered to a single read flow cell according to the Illumina cBot Cluster Generation System procedures. On completion of cluster generation the flow cell was run for 50 cycles (36 cycles for H1).

Sequence alignment and 5hmC peak calling. All FASTQ sequence files were aligned to the human reference genome (hg18) using Bowtie 0.12.7 (ref 52) with the same criteria: genomic matches and no more than 3 mismatches. The aligned tags were further processed to filter the duplicate reads. 5hmC peaks were

5hmC dot blot. DNA was spotted on an Amersham Hybond-N+ membrane (GE Healthcare) and then fixed to the membrane by drying at 80 °C for 30 min. The membrane was then blocked with 3% BSA and incubated with a polyclonal antibody against 5hmC (1:5,000 dilution, Active Motif) as the primary antibody overnight at 4 °C. A horseradish peroxidase-conjugated secondary antibody against rabbit (1:5,000 dilutions, Sigma) was used to incubate the membrane for 1 h at room temperature.

Immunofluorescence staining. Human iPSCs treated with either TET1 shRNA or control shRNA were plated onto coverslips that were pre-coated with Matrigel in mTeSR1 medium under puromycin selection (0.5 μg/ml). Cells were fixed in 4% paraformaldehyde for 10 min. Then cells were permeabilized for 10 min with 0.25% Triton X-100. Cells were then incubated for 1 h with 4% donkey serum blocking buffer. Next, cells were incubated with primary antibody overnight at 4 °C. After washing with PBS 3 times for 5 min, secondary antibodies conjugated to Alexa Fluor 555 (Invitrogen) were used. The primary antibodies used were anti-NANOG (Cell Signaling, 5508s, 1:100), SOX2 (Santa Cruz, sc-20088, 1:100) and TRA-1-81 (Millipore; 90233, 1:100).

Genomic DNA preparation and 5hmC capture. Before isolation of genomic DNA, hiPSCs/hESCs were treated with collagenase to detach from feeder cells, and transferred to Matrigel-coated culture plates in mTeSR1 medium (Stemcell Technologies) for at least 3 passages to eliminate the contamination of feeder cells. Genomic DNA was extracted and purified with the DNeasy Kit (Qiagen). Genomic DNA was transduced with pLKO.1 Quick Titer Lentivirus Quantitation Kit (Cell Biolabs). The titre of concentrated virus was then determined by the Phosphatase (Millipore) around 20 days after initial infection with Yamanaka factors. The labelled DNA was purified by the QiAgquick PCR purification kit (Qiagen). Click chemistry was performed with the addition of 150 μM disulpahide-biotin linker, and the mixture was incubated for 2 h at 37 °C. The bioin labelled DNA samples were then captured by using Dynabeads MyOne Streptavidin C1 (Invitrogen) following the manufacturer’s recommendations except eluted in dithiothreitol. Subsequently, the 5hmC-enriched DNA was purified by the MinElute PCR Purification Kit (Qiagen) and eluted with 12 μl elution buffer.

© 2013 Macmillan Publishers Limited. All rights reserved.
identified using MACS (ref. 53) with the following parameters: effective genome size = 2.7 × 10^6; Tag size = 38 or 50; bandwidth = 250; P-value cutoff = 1.00 × 10^-5 with H1 genomic DNA input as a control.

**DHMR identification.** To compare peaks between samples, the 5hmC-enriched regions in each sample were coalesced into a union window. We recounted the total aligned reads for each window, and then further normalized with each aligned total count. To call differential 5hmC-enrichment regions, the Bioconductor Deseq package was used for analysis, and a FDR of 0.01 was used for positive calling. When using iPSCs compared with original fibroblasts, we found a significant number of peaks. When using fibroblasts compared with a repeat experiment of fibroblasts yielded only background level of peaks, suggesting that we captured bona fide 5hmC modification in fibroblasts (Supplementary Table S4).

The large-scale aberrant reprogramming hotspots are defined as genomic regions satisfying the following conditions: the 5hmC levels are highly variable among iPSCs but relatively consistent among ESCs; the average difference of 5hmC levels between iPSCs and ESCs is large, and longer than 100 kb. To assess the variability of 5hmC levels, the whole genome was binned into 1 kb windows and the read counts within each window were obtained. The biological variation in each window was then calculated using a method of moment estimator (detailed in estimating 5hmC variations). We then smoothed the estimated variance by moving window average with 100 kb. The smoothed variances from iPSCs are significantly larger than from ESCs (Supplementary Fig. S5a). We then pooled the smoothed variances of iPSCs and ESCs, and used the 99th quantile as the threshold to detect variable 5hmC regions (DHMR) in iPSCs and ESCs. Thirty-three VMRs were detected from iPSCs, and one VMR was detected from ESCs. We then assessed the average 5hmC levels in these VMRs. First the counts are normalized by total reads and average 5hmC levels are computed for iPSC and ESC. The average 5hmC levels are greater in ESCs for most VMRs (Supplementary Fig. S5b). The large-scale aberrant reprogramming hotspots are identified as genomic regions satisfying the following criteria: smoothed variances of iPSCs greater than the 99th quantile of the pooled variances; smoothed variances of ESCs smaller than the 50th quantile of the pooled variances; differences of smoothed averages between iPSCs and ESCs greater than the 95th quantile of all absolute differences; and minimum length greater than 100 kb. Detected regions closer than 50 kb are merged into one.

**Estimating 5hmC variations.** We first obtained read counts in non-overlapping 1 kb windows. We denote the count for window i and sample j by X_{ij}. X_{ij} is assumed to follow a negative binomial distribution: X_{ij} ∼ NB(θ_{ij}, φ_{ij}). Here, θ_{ij} is the true 5hmC level, s_{ij} is the library size and φ_{ij} is the dispersion parameter. The negative binomial is a γ-Poisson compound distribution. It assumes that the true 5hmC level, θ_{ij}, follows a gamma distribution, and conditional on θ_{ij} the observed counts follow a Poisson distribution. A negative binomial distribution accounts for over-distributions (sample variance greater than sample mean) so it is often used for modelling sequencing data from biological replicates. The dispersion parameter φ_{ij} is the squared coefficient of variation of the true 5hmC level θ_{ij}, and represents the variability among biological replicates. It can be shown that directly using the sample variances of normalized reads to estimate φ_{ij} will lead to erroneous results. Samples with larger library sizes will have smaller variance estimation. We designed the following moment estimator. First, define a new variable Y_{ij} = (X_{ij} - X_{ij})/s_{ij}. We have E(Y_{ij}) = μ_{ij}/φ_{ij} - 1; here μ_{ij} is the expected value of θ_{ij}. We first estimate μ_{ij} as μ_{ij} = X_{ij}/s_{ij}, and then use the method of moment to obtain the estimates for φ_{ij} as V(μ_{ij})/μ_{ij} - 1. Detailed proofs and derivations for the estimators are presented in a statistical paper.41

**Tet-assisted bisulphite-based PCR amplicon sequencing.** To investigate 5hmC distribution at a single-base resolution, HEK293 iPSC genomic DNA was subjected to glycosylation and cyclisation by Tet as described previously42 and the processed DNA was eluted in a ~50 μl (500 ng) volume. The treated gDNA was bisulphite converted and eluted in 30 μl H2O. One microlitre of converted DNA was PCR amplified by using Phusion Turbo CX Hotstart DNA polymerase under the following conditions: 2.5 U polymerase, 5 μl 10× Phusion Turbo CX reaction buffer 4 μl 2.5 mM dNTPs and 1 μl primers. The PCR cycling conditions were: 95 °C 2 min, 40 cycles of 95°C 30 s, 55°C 30 s, 72°C 1 min, followed by 72°C 5 min. Primers used for amplifying bisulphite converted genomic DNA were designed by Methyl Primer Express Software v1.0 (Invitrogen) targeting chr10, chr18 and chr22 large-scale hotspots, and confirmed by specific bands in agarose gel electrophoresis. The average amplicon size is around 200 bp. The primer sequence and amplified region coordinates are listed in Supplementary Tables S6 and S7.

The PCR amplicon was further purified by AMPure XP bead, and eluted in 50 μl H2O. The concentration was quantified with a Qubit High Sensitivity kit and then pooled together in an equal molar concentration for each sample. Then the mixed amplicons were subjected to library preparation and MiSeq deep sequencing. Briefly, samples were treated by end repair, A-tailing and the ligation of TruSeq adaptors containing compatible indexes by using NEBNext library preparation for the Illumina kit. Libraries were then filtered using KAPA SYBR FAST qPCR kits and pooled together with an equal molar concentration. MiSeq sequencing was performed as standard procedures recommended by Illumina: the concentration of pooled library used was 8 pM, and the run was initiated for 2 × 150 bases of SBS sequencing. Image analysis and base calling were performed with the standard Illumina pipeline. To call m5Cm5C status, the Bismark application was used.

**Array CGH.** Two microarrays of HUES48, HUES49, hiPSC-B22, hiPSC-B23 and hiPSC-RX35i DNA was co-hybridized with 2 μg H1 hESC reference DNA to 1 × 1 M Agilent SurePrint G3 Human Catalog oligonucleotide arrays (Agilent Technologies). The arrays span the entire genome with an oligonucleotide backbone spaced, on average, every 3 kb the unique identifier (AMADID) for the design is 021529. Arrays were hybridized according to the manufacturer’s instructions and scanned using the Agilent high-resolution C scanner (Agilent Technologies). Signal intensities were evaluated using Feature Extraction Version 9.5.1.1 software (Agilent Technologies) and analysed with Genomic Workbench 5.0 software (Agilent Technologies). To detect the maximum number of CNVs, we used a minimum absolute log ratio of 0.25 on at least 4 aberrant probes. To generate Supplementary Fig. S4, the stringency was raised to 20 aberrant probes.

**Genomic analysis.** Microarray data on fibroblasts and iPSCs were obtained from a previous study37. The microarray data were normalized and analysed using Bioconductor’s oligo26 and siggens packages within R (http://www.r-project.org/). The differentially expressed genes were called by SAM (significance analysis of microarrays) with corrected P-value <0.01.

ReSeq genes and CpG islands were defined on the basis of NCBI build 36/hg18 coordinates downloaded from the UCSC Genome Browser website. Core promoters were arbitrarily defined as 200 bp upstream and downstream of the TSS of ReSeq genes. Gene bodies are defined as the transcribed regions, from the start to the end of transcription sites for each ReSeq gene.

**Association of DHMR with genomic features was performed by overlapping defined sets of DHMRs with known genomic features obtained from UCSC Tables for NCBI36/hg18: ReSeq Genes, 5’ UTR, Exon, Intron, 3’ UTR, ±300 bp of TSS, ReSeqIntergenic, ±500 bp of TES, CGIs, Islands, BGC CGIs, Islands, hypodeaminated CpG islands were defined from a previous study37.** Transmission factor binding sites for KLF4, NAT, OCT4 and SOX2 and RNA-seq RPMK values in H1 hESCs were described previously39. Data analysis was performed by R (http://www.r-project.org/) scripts.

Genomic views of 5hmC relative enrichment intensity were generated using IGV 2.10 and igtools (Integrated Genomics Viewer tools and browser, http://www. broadinstitute.org/igv/).46

**Accession number.** Sequencing data have been deposited to GEO with the accession number GSE37050.

**Methods**

49. Ohta, K. & Yamanaka, S. Induced pluripotent stem cells: opportunities and challenges. Phil. Trans. R. Soc. Lond. Ser. B 366, 2198–2207 (2011).
50. Carey, B. W. et al. Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. Cell Stem Cell 9, 588–598 (2011).

51. Somers, A. et al. Generation of transgene-free lung disease-specific human induced pluripotent stem cells using a single excisable lentiviral stem cell cassette. Stem Cells 28, 1728–1740 (2010).
52. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Gen. Bio. 10, 25 (2009).
53. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). Gen. Bio. 9, R137 (2008).
54. Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics 14, 232–243 (2013).
55. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. Bioinformatics 26, 2363–2367 (2010).
56. Robinson, J. T. et al. Integrative genomics viewer. Nat. Bio. 29, 24–26 (2011).
Figure S1  TET1 is associated with increased hydroxymethylation during human iPSC reprogramming, but reduction of TET1 does not compromise the pluripotency of human iPSCs. (a) Knocking down TET1 by siRNA significantly decreases 5hmC levels in iPS-A2. The left panel represents siTET1 knock down efficiency by quantitative RT-PCR (*t-test, p<0.05). Error bars represent S.E.M. collected from three independent experiments. Right panel depicts the effect of total 5hmC levels 48 hours post siTET1 transfection. (b) siTET1 only, but not TET2 or TET3 could affect 5hmC levels in iPSCs. (c) shTET1 lentivirus (two shTET1 vectors, 75024 and 75026) could efficiently knock down TET1 (*t-test, p<0.05). Error bars represent S.E.M. collected from three independent experiments. (d) Quantitative analysis of AP-positive colonies in three different experiments: left, day 20; middle, day 25; right, day 20. shGFP lentivirus was used as control for shTET1 lentiviral transduction. (e) Dot blot analysis of 5hmC levels in stable shTET1 or shControl iPS-B22 cells. (f) Wells stained for alkaline phosphatase for the shControl and shTET1 cells. Cells used for staining were grown in 6-well plate. (g) Immunostaining for pluripotency markers NANOG, SOX2 and TRA1-81 in both cell groups (Scale bars: 120 μm). The raw values of related statistical test in this figure are listed in supplementary table S1.
Figure S2: Bimodal distribution of 5hmC around TSS and TES. Normalized 5hmC and input read densities among TSS- and TES- surrounding regions. Reads were summed in 50-bp windows 3 kb upstream and downstream of TSS and TES. (a) All genes with expression level changes between iPSCs and fibroblasts interrogated by Affymetrix hg133plus2 array. (b) Genes with no expression changes. (c) Genes with expression increased in iPSCs. (d) Genes with expression decreased in iPSCs. (e) Genes stratified by 5 groups according to H1 RPKM values.
Figure S3 Sequence preference of hydroxymethylation modification during reprogramming. (a) Functional annotation of hyperDhMRs and hypoDhMRs between iPSCs and fibroblasts. TSS represents ±1kb of transcription start site; TES represents ±1kb of transcription end site. (b) 5hmC density around core TSS is negatively correlated with gene activity. Genes were divided into four groups (Highest, High, Low, Lowest) according to the RPKM of H1 hESC data. The core TSS region is defined as ±200 bp of transcription start site. (c) 5hmC density is negatively correlated with CG dinucleotide percentage in UCSC CpG islands. CpG islands are artificially divided into three groups (High, Intermediate, and Low) according to their CG dinucleotide percentages, and then plotted with the 5hmC-normalized counts within these regions. Box shows the center quartile, with the outliers suppressed. (d) Chromosomal distribution. A different classification of CpG islands was reported based on the conservation of CpGs across species during evolution. One classification is hypodeaminated CpG islands, which are CpG-rich regions characterized by evolutionarily slow rates of CpG loss, and represent genomic regions with low levels of methylation. Shown is the chromosomal layout of CpG-rich loci that were classified as hypodeaminated islands, having low levels of DNA methylation and low deamination rates. (e) Chromosomal distribution of Biased Gene Conversion (BGC) islands. Shown is the chromosomal layout of CpG-rich loci that were classified as methylated and hyperdeaminated islands. They exhibit more rapid rates of CpG loss evolutionarily as well as higher methylation levels. The distribution is shown to be highly non-uniform, with hotspots on most subtelomeric regions. (f) Fib-ES-DhMR favours evolutionarily less conserved BGC CpG island groups. The relative Fib-ES-DhMR overlapping regions with BGC and hypodeaminated CpG islands; expected number is also plotted. Of the top 20,000 (lowest adjusted P-values) hyperDhMRs, more regions than expected overlaps with 26,058 BGC islands; in contrast, less regions than expected are located in hypodeaminated CpGs. This distribution suggests a significant (p-value < 2.2e-16) bias for hydroxymethylation to occur at BGC islands. Therefore, 5hmC modifications acquired during reprogramming tend to occur within the unique sequence context of BGC islands, in which the methylation is evolutionarily less conserved.
**Figure S4** The 5hmC aberrant reprogramming hotspots are not due to genomic instability. (a) Number of CNVs called by four consecutive probes with the average of each probe spanning 3 kb across the genome. (b) Density plot of log ratios of signal of two identified CNV regions in iPS-B23 and iPS-RX35i, both of which show a deletion of sequence compared with H1. (c) Chromosome ideograms show no overlap of CNVs with 5hmC aberrant reprogramming hotspots. CNV regions shown are indicated by figure annotated bottom right; the hotspot regions are labeled with each individual black line on chromosomes.
Figure S5 Large-scale DhMRs in iPS cells are more variable than in hES cells. (a) Density plot of smooth variance of identified VhMR in ESCs and iPSCs. (b) Difference of smoothed 5hmCs. (c) Correlation between the number of large-scale DhMRs and passage number in 9 iPSC lines. Table summarizes the linear regression on the parameter passage number, producing coefficient of passage number with a p-value 0.544, which is not significant.
Figure S6  Correlation and confirmation analyses between TAB-Seq and 5hmC capture approach. Correlation analysis between TAB-Seq and 5hmC capture approach in H1 ES cells (a, b). (a) Genome-wide Pearson correlation. Correlation coefficient: 0.65. (b) Pearson correlation within 20 large-scale regions. Correlation coefficient: 0.6. The window size used for analysis is 3000 bp. Chr18 and 22 large-scale hotspots validated at single base resolution by TAB-Seq (c, d). (c) Summary of PCR-based TAB-Seq in Chr18 large-scale hotspot. The first three tracks are 5hmC intensity determined by capture-Seq, showing iPS-B22 bearing incomplete hydroxymethylation. Below is 5hmC intensity either in CG or CH format determined by PCR based TAB-Seq. Black circles indicate the PCR amplicon mapped loci. (d) Summary of PCR based TAB-Seq in Chr22 large-scale hotspot. The first three tracks are 5hmC Capture-Seq results, in some of the regions; both iPS-B22 and iPS-S2 show incomplete hydroxymethylation. In amplified regions by TAB-Seq, both iPS-B22 and iPS-S2 became completely hydroxymethylated. This is also confirmed by TAB-PCR-Seq. Black circles indicate the PCR amplicon mapped loci.
Figure S7 Low level of 5hmC at peri-centromeric non-CG DMRs. 5hmC density in fibroblasts, iPS, and ES cell lines at two of the non-CG large-scale DMRs. The position of the loci within the chromosome and the scale are shown above the gene tracks.
Supplementary Table Legends

**Table S1** Statistical test raw data.
**Table S2** Summary of iPSC lines used in this study. O,S,K,M indicates reprogrammed by OCT4, KLF4, SOX2 and c-MYC; OSNL indicates reprogrammed by OCT4, SOX2, NANOG and LIN28, [OKSM] means reprogrammed by a polycistronic formatted in the order of OCT4, KLF4, SOX2 and c-MYC separated by T(F)2As.
**Table S3** Summary of 5hmC sequencing statistics.
**Table S4** DhMRs pairwise comparison between fibroblast biological replicates, and between iPSCs and original fibroblasts.
**Table S5** Summary of quantitative RT-PCR primers used in this study.
**Table S6** Primers used for PCR-based TAB-Seq targeting large-scale hotspot in chromosome 10 and corresponding amplicon coordinates. The primers were designed to amplify the bisulfite converted forward strand of genomic DNA.
**Table S7** Primers used for PCR-based TAB-Seq targeting large-scale hotspot in chromosome 18 and 22, and corresponding amplicon coordinates.