DNA Methylation Dynamics in Human Induced Pluripotent Stem Cells over Time

Koichiro Nishino, Masashi Toyoda, Mayu Yamazaki-Inoue, Yoshihiro Fukawatase, Emi Chikazawa, Hironari Sakaguchi, Hidenori Akutsu, Akihiro Umezawa*

Department of Reproductive Biology, National Institute for Child Health and Development, Tokyo, Japan

Abstract

Epigenetic reprogramming is a critical event in the generation of induced pluripotent stem cells (iPSCs). Here, we determined the DNA methylation profiles of 22 human iPSC lines derived from five different cell types (human endometrium, placental artery endothelium, amnion, fetal lung fibroblast, and menstrual blood cell) and five human embryonic stem cell (ESC) lines, and we followed the aberrant methylation sites in iPSCs for up to 42 weeks. The iPSCs exhibited distinct epigenetic differences from ESCs, which were caused by aberrant methylation at early passages. Multiple appearances and then disappearances of random aberrant methylation were detected throughout iPSC reprogramming. Continuous passaging of the iPSCs diminished the differences between iPSCs and ESCs, implying that iPSCs lose the characteristics inherited from the parent cells and adapt to very closely resemble ESCs over time. Human iPSCs were gradually reprogrammed through the "convergence" of aberrant hyper-methylation events that continuously appeared in a de novo manner. This iPS reprogramming consisted of stochastic de novo methylation and selection/fixation of methylation in an environment suitable for ESCs. Taken together, random methylation and convergence are driving forces for long-term reprogramming of iPSCs to ESCs.

Citation: Nishino K, Toyoda M, Yamazaki-Inoue M, Fukawatase Y, Chikazawa E, et al. (2011) DNA Methylation Dynamics in Human Induced Pluripotent Stem Cells over Time. PLoS Genet 7(5): e1002085. doi:10.1371/journal.pgen.1002085

Editor: John M. Greally, Albert Einstein College of Medicine, United States of America

Received December 2, 2010; **Accepted** April 1, 2011; **Published** May 26, 2011

Funding: This research was supported by grants from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan; by Ministry of Health, Labour, and Welfare Sciences (MHLW) research grants; by a Research Grant on Health Science focusing on Drug Innovation from the Japan Health Science Foundation; by the program for the promotion of Fundamental Studies in Health Science of the Pharmaceuticals and Medical Devices Agency; by a Grant for Child Health and Development from the MHLW; by the Intramural Research Grant (22-5) for Neurological and Psychiatric Disorders of NCNP; by the Research Grant (22-2-4) for cardiovascular disease of NCVC given to AU; by a grant from New Energy and Industrial Technology Development Organization (NEDO) in Japan given to HA; and by Grant-in-Aid for Young Scientist(B)(WAKATE- B 21790372) given to KN. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: umezawa@1985.jukuin.keio.ac.jp

Introduction

DNA methylation is an important epigenetic modification and is a key component in normal differentiation, development and disease [1–3]. Expression of tissue-specific genes, such as *Oct-4* [4], *Nanog* [5], *Sry* (sex determining region on Y chromosome) [6] and *MyoD* [7], are induced by spatio-temporal demethylation during development. DNA methylation therefore specifically varies depending on tissue types and cell linage [2], indicating that information regarding cell type-specific DNA methylation profiles can enable the identification and validation of cell types. Transformation of iPSCs from somatic cells requires a process of epigenetic reprogramming promoted by transient ectopic expression of defined transcription factors expressed in ESCs [8–11]. Human iPSCs are considered to be powerful resources in regenerative medicine because of their potential of pluripotency and avoidance of rejection of their derivatives by the immune system, and for ethical issues as well [12]. Although iPSCs show pluripotency, they have different propensities for differentiation in mouse models [13]. Human iPSCs also exhibit donor cell-specific gene expression [14,15]. Moreover, iPSCs possess inherited DNA methylation states as epigenetic memories from parent cells [15–17], suggesting that these memories influence different propensities of the iPSCs. On the other hand, continuous passaging of mouse iPSCs reduces differences from each other in gene expression profiles [15]. Epigenome-wide analysis started to be used in this field [18,19], and differentially methylated regions have been identified among human iPSCs, their parent cells and ESCs [17,20]. Aberrant epigenetic reprogramming has recently been reported in human iPSCs [21,22]. However, these analyses were limited to the use of a small number of cells as a source for generation of iPS cells. Moreover, human iPSCs have only been analyzed at a single point of passage. Therefore, it has not been clarified whether human iPSCs generated from various types of cells are dissimilar from each other at different points during passage; how continuous passaging of human iPSCs influences the differences between iPSCs and ESCs; and how aberrant methylation in human iPSCs during passaging. To address these issues, we compared the epigenetic and transcriptional states of human iPSCs derived from five cell types of different origins during passage, and found random aberrant hyper-methylation at different points of adaptation into ESCs.

Results

Establishment of human iPSCs

Human iPSCs derived from fetal lung fibroblasts (MRC5), amnion (AM), endometrium (UtE), placental artery endothelium (PAE) and menstrual blood cells (Edom) were independently established in our laboratory by retroviral infection of 4 genes

Author Summary

iPSCs change to resemble ESCs via two phases: the transgene-dependent phase, in which the transcription factors act to transform somatic cells into pluripotent stem cells, and the transgene-independent phase, in which the transcription factors are silenced. In this study, we established human iPSCs derived from 5 different cell types by retroviral infection of the Yamanaka 4 factors, and we identified 8 novel epigenetic markers (*SALL4*, *EPHA1*, *PTPN6*, *RAB25*, *GBP4*, *LYST*, *SP100*, and *UBE1L*) by comprehensive DNA methylation analysis. The aberrant hyper-methylation in iPSCs occurred stochastically throughout the genome and decreased during the long-term iPSC reprogramming, suggesting that the aberrant stochastic hyper-methylation and their convergence are a direct cause of the transgene-independent phase of iPS reprogramming. These results favor the stochastic model of the Yamanaka model rather than the elite model. In addition, the stem cell–specific methylation states and the epigenetic difference between iPSCs and ESCs are useful indices for evaluating human iPSCs in therapeutic applications.

(*OCT-3/4*, *SOX2*, *c-MYC*, and *KLF4*) (Figure 1A, 1B and Table S1). These cells clearly showed human ES-like characters in terms of morphology; cell-surface antigens; gene expression of stem cell markers; teratoma formation in which these cells differentiated to various tissues including neural tissues (ectoderm), cartilage (mesoderm), and epithelial tissues (endoderm); growth (more than 20 passages); and DNA methylation patterns at *OCT-3/4* and *NANOG* promoter regions (Figures S1, S2, S3). Short tandem repeat (STR) analysis showed clonality between the respective iPSC lines and their parent cells (Table S2). Silencing of transgenes and normal karyotypes of iPSCs were also confirmed (Figure S4 and Table S3).

Analysis of DNA methylation profiles

To investigate the dynamics of DNA methylation in pluripotent stem cells, we examined 5 ESC lines (HUESCs) [23,24], 22 iPSC lines, their parent cells and 201B7, using Illumina's Infinium HumanMethylation27 BeadChip. In total, 24,273 CpG sites in 13,728 genes were analyzed, along with 33 human cell lines (Table S1). The iPSC line "201B7" was generated from human skin fibroblasts [8]. Quantitative scores of DNA methylation levels were obtained as β-values determined from the Illumina analysis, ranging from "0", for completely unmethylated, to "1", for completely methylated. We also performed genome-wide gene expression analysis using the Agilent Whole Human Genome Microarray chips. As assessed by unsupervised hierarchical clustering analysis and scatter plot of DNA methylation and gene expression data, human iPSCs could be clearly discriminated from their parent cells and were similar to ESCs (Figure 1C and Figure

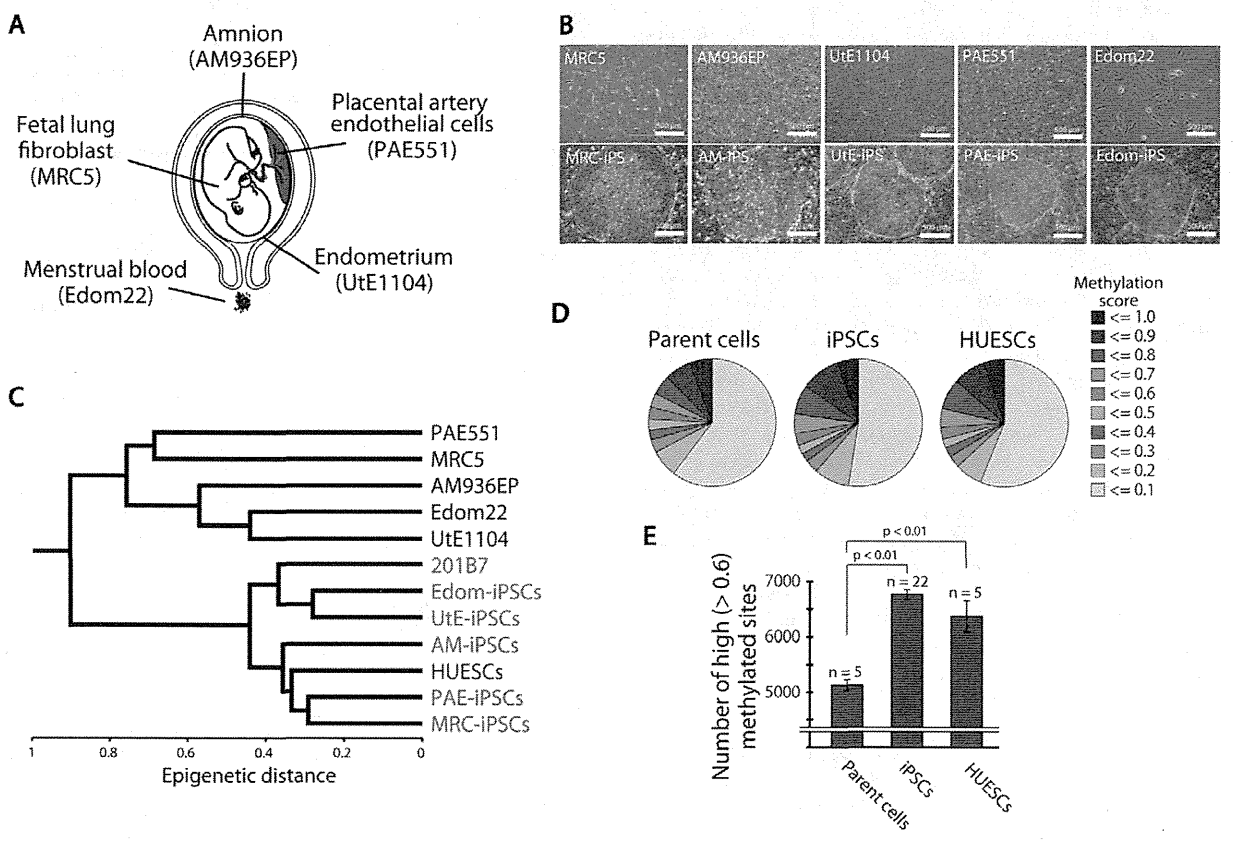


Figure 1. Pluripotent stem cells are significantly more hyper-methylated than their parent cells. (*A*) The human cell origins used for generation of iPSCs. (*B*) Morphology of the parent cells (upper panels) and iPSCs (lower panels). (*C*) Unsupervised hierarchical clustering analysis based on DNA methylation. (*D*) Distribution of 24,273 CpG sites with their methylation scores in the parent cells, iPSCs and ESCs. (*E*) The average number of high (>0.6) methylated CpG sites. The iPSCs have more highly methylated sites than the parent cells.
doi:10.1371/journal.pgen.1002085.g001

S5). The distribution of DNA methylation levels shows that the degree of global methylation in pluripotent stem cells was higher compared to the parent cells (Figure 1D, 1E), suggesting that a global gain of DNA methylation occurs during reprogramming.

Identification of stem cell-specific differentially methylated regions (DMRs)

For further analysis, we defined DMR as representing a CpG site whose score differed 0.3 points or more from the β-value between the two groups. By comparison among ESCs (average from 5 lines), iPSCs (average from 22 lines), and parent cells (average from 5 lines), about 90% of the CpG sites (17,572 sites) examined did not show differential methylation among ESCs, iPSCs and parent cells (Figure 2A), suggesting that only a small number of the CpG sites is affected during reprogramming. The number of the CpG sites has been reported to be larger by genome-wide analysis [21].

We then identified 220 sites that are pluripotent stem cell-specific DMRs (Figure 2A). The 174 sites (79.5%) of the stem cell-specific DMRs had significantly higher methylation levels in iPSCs/ESCs when compared to the parent cells (Figure 2B). Approximately 80% of the DMRs between the iPSCs and their parent cells changed to a "hyper-methylated" state from a "hypo-methylated" state in iPSCs. In contrast, 45 sites of the stem cell-specific DMRs are hypo-methylated in iPSCs/ESCs, compared with the parent cells. Gene ontology analysis indicates that the hypo-methylated stem cell-specific DMRs especially included genes related to mRNA transcription regulation (Figure 2B). Interestingly, the majority of the hypo-methylated stem cell-specific DMRs were located on CpG islands, whereas the majority of the hyper-methylated stem cell-specific DMRs were located on non-CpG islands (Figure 2C). No iPS-specific DMRs were detected. We extracted 3,123 sites that are differentially methylated in one or more parent-specific iPSCs, compared to their parent cells, because DMRs are dependent on parent cell types (Figure S6). These DMRs are here designated as stem cell-required DMRs. Distribution analysis of the stem cell-required DMRs revealed a dispersed pattern rather than specific localization on the genome (Figure S7A).

From the combined gene expression and DNA methylation data, we chose 27 genes in the stem cell-specific DMRs showing more than a 5-fold change in expression of human iPSCs/ESCs, as compared with those in the parent cells (Table S4). Nine genes with hypo-methylated stem cell-specific DMRs were found in the group "genes significantly expressed in iPSCs/ESCs," and 17 genes with hypo-methylated stem cell-specific DMRs belonged to the category "low expression or silenced in iPSCs/ESCs". In addition, the methylation state and gene expression in *EPHA1*, *PTPN6*, *RAB25*, *SALL4*, *GBP3*, *LYST*, *SP100* and *UBE1L* were confirmed by quantitative combined bisulfite restriction analysis (COBRA) [25] (Figure 2D), RT-PCR (Figure 2E) and bisulfite sequencing (Figure 2F).

We also extracted genes with stem cell-required DMRs exhibiting high expression or suppression in human iPSCs/ESCs (Tables S5, S6). Interestingly, gene ontology analysis of the genes with stem cell-required DMRs showed that genes in the transcription factor category were detected only in the hypo-methylated stem cell-required DMRs (Table S7). The top 20 transcription factor genes with hypo-methylated stem cell-required DMRs exhibiting high expression in human iPSCs are summarized in Table 1 and include *OCT-4/3* (also known as *POU5F1*), *SALL4*, *SOX8*, *ZIC5*, and *FOXD1*.

Aberrant and inherited methylation in iPSCs

Few changes in DNA methylation were detected between iPS and ES cells and these were not consistent among the different iPS lines (Figure 2A, Figures S6, S7). In further analyses, we compared the DNA methylation states of each iPSC line or each parent cell line with that of ESCs (averaged value) (Figure 3A). For the whole genome, the number of DMRs between ESCs and iPSCs (ES-iPS-DMRs) varied in the 22 iPSC lines (Figure 3B). A comprehensive analysis of methylation in ESCs and iPSCs identified 1,459 ES-iPS-DMRs covering 1,260 genes that were differentially methylated in one or more iPSC lines. ES-iPS-DMRs are composed of aberrant (iPS-specific) methylation sites, in comparison with ESCs and inherited methylation sites from the parent cells. The number of inherited sites as well as aberrant sites varied among iPSCs. Analysis of the ES-iPS-DMRs on each chromosome showed a characteristic distribution of the ES-iPS-DMRs on the X chromosome in XX-iPSCs (Figure 3B and Figure S8). Female XX-iPSCs demonstrate a tendency to carry a large number of ES-iPS-DMRs on the X chromosome, but male XY-iPSCs had few ES-iPS-DMRs on the X chromosome (Figure 3B, lower panel). While no ES-iPS-DMRs overlapped for all the iPSCs (Figure 2A), 20 ES-iPS-DMRs overlapped in more than 15 out of 22 lines (Figure 3C, inset). These 20 ES-iPS-DMRs include the genes for *MPG* (N-methylpurine-DNA glycosylase isoform b), *FZD10* (frizzled 10), *IREX2* (iroquois homeobox protein 2) and *ZNF248* (zinc finger protein 248), which are highly associated with aberrant methylation during reprogramming. Distribution analysis of the ES-iPS-DMRs across the genome did not show any specific localization (Figure S9). We further compared overlapping ES-iPS-DMRs in reference to a genome-wide methylation analysis [21], and found that 72 gene promoters overlapped between our data and that of Lister et al..

More than 70% of the ES-iPS-DMRs were hyper-methylated in each iPSC (Figure 3D), indicating that the iPSC genome is more methylated than the ESC genome. In addition, the majority of the ES-iPS-DMRs were located on CpG islands (Figure 3E), suggesting that aberrant methylation is biased towards CpG islands.

Effect of long-term culture on DNA methylation status in iPSCs

We investigated the effect of continuous passaging on the DNA methylation profile of human iPSCs. To address the effect, we subjected 7 iPSC lines to additional rounds of passaging under identical culture conditions, and obtained genomic DNA and RNA at passage 4 (P4) to P40 for DNA methylation and gene expression. The number of the ES-iPS-DMRs ranged from 80 in MRC-iPS-25 to 286 in UtE-iPS-11 at early passage (P10 to P20), whereas the number of the ES-iPS-DMRs dramatically decreased in all lines at late passage (P30 to P40) (Figure 4A, upper-left panel). The number of inherited and aberrant sites decreased to 30 and 70, respectively, at P30 to P40 (Figure 4A, upper-center and right panels). These decreases in the numbers of ES-iPS-DMRs indicate that iPSCs have become closer to ESCs in their DNA methylation profiles. In particular, XX-iPSC lines (AM-iPS-8, UtE-iPS-4 and -11, and Edom-iPS-2) showed decreases in the number of ES-iPS-DMRs with passaging. The XY-iPSC lines, such as MRC-iPS-25 and PAE-iPS-1, had only a small number of ES-iPS-DMRs. The number of ES-iPS-DMRs continued to decrease to approximately 100 ES-iPS-DMRs containing 30 inherited sites. Intriguingly, few ES-iPS-DMRs on the X chromosome were detected in XY-iPSCs throughout the passaging. In contrast, the number of ES-iPS-DMRs in XX-iPSCs ranged from 10 to 70 at the early passage (P4 to P20), and decreased to zero after P30 (Figure 4A, lower panels). We also

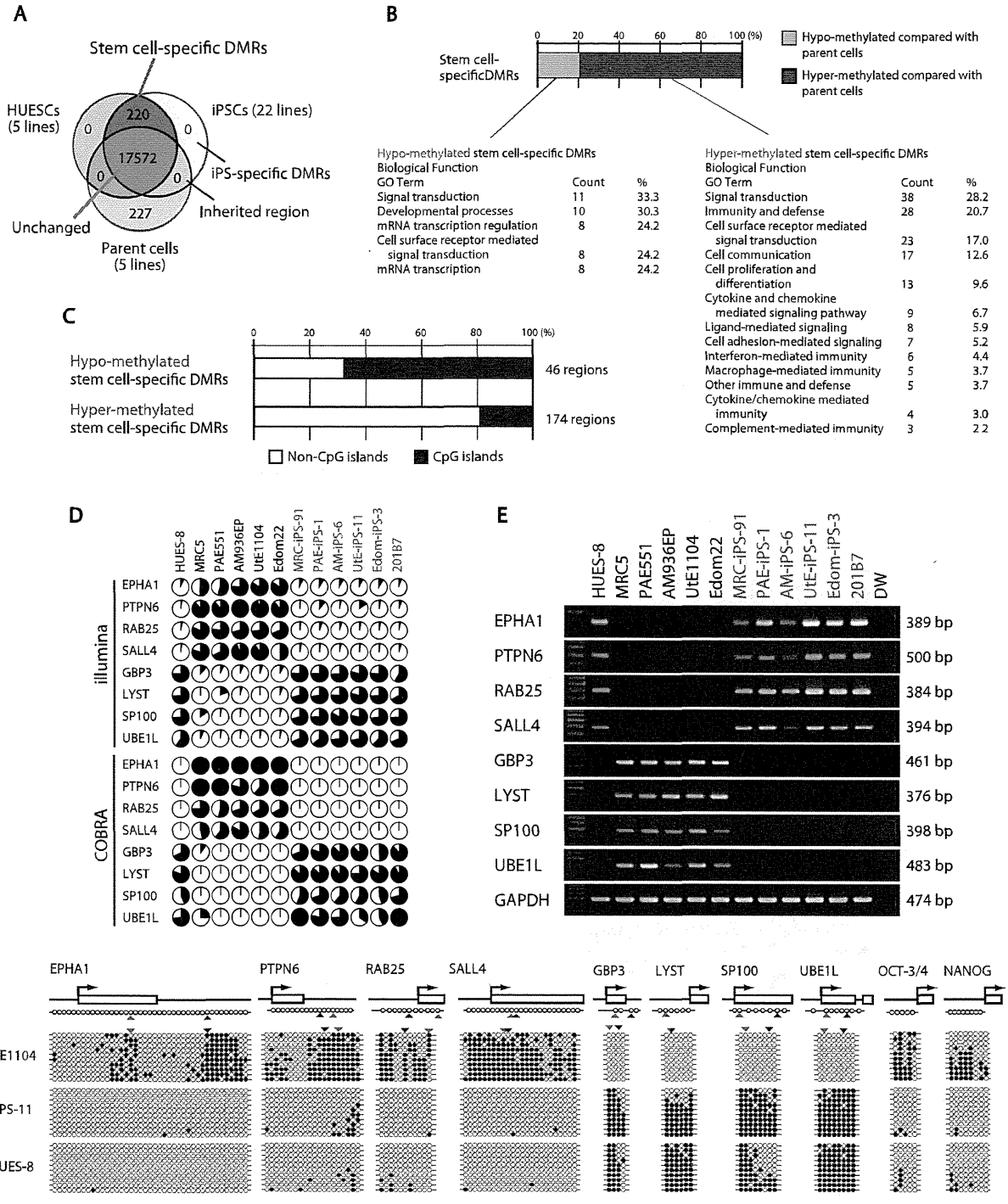


Figure 2. Defining stem cell-specific DMRs as novel epigenetic iPS markers. (*A*) Venn-like diagram showing overlapping CpG sites among ESCs, iPSCs and their parent cells. The 220 overlapping sites are stem cell-specific differentially methylated regions (DMRs). Notably, neither overlapping iPSCs-specific DMRs nor inherited regions in iPSCs from the parent cells were observed. (*B*) Proportion of the hyper- and hypo-methylated stem cell-specific DMRs and GO analysis. Approximately 80% of the regions were hyper-methylated in iPSCs, compared with that of the parent cells. (*C*) Proportion of the regions associated with CpG islands and non-CpG islands in the hypo-methylated stem cell-specific DMRs. The hypo-methylated regions were biased to CpG islands, whereas the hyper-methylated regions were biased to non-CpG islands. (*D*) DNA methylation levels in the 8 representative genes determined by Illumina Infinium HumanMethylation27 assay and Bio-COBRA. These 8 genes were defined as SS-DMRs with significant changes of expression and were described in Table S6. The relative amount of methylated and unmethylated DNA ratio is indicated as the black and white area, respectively, in the pie chart. (*E*) Expression of the 8 genes. Expression of the 8 genes had an inverse correlation with DNA methylation level. (*F*) Bisulfite sequencing analysis of the 8 genes in endometrial cells (UtE1104), UtE-iPS-11 and HUES-8 cells. (Top)

Schematic diagram of the genes. Arrows, open boxes and open circles represent transcription start site, first exon and position of CpG sites, respectively. (Bottom) Open and closed circles indicate unmethylated and methylated sites, respectively. Red and blue arrowheads represent the position of CpG sites in Infinium assay and COBRA assay, respectively.
doi:10.1371/journal.pgen.1002085.g002

investigated the effect of continuous passaging on the DNA methylation profile of the parent cells (UtE1104 and Edom22) (Figure 4B). The number of the DMRs between ESCs and parent cells (ES-parent-DMRs) increased with passaging. In addition, we also confirmed that the transgenes were silenced at each passage (Figure 4C and Figure S4), indicating that the decreasing number of the ES-iPS-DMRs in iPSCs occurred in the transgene-independent phase.

Comparative analysis of ES-iPS-DMRs dynamics

We then compared each ES-iPS-DMRs with passaging. The UtE-iPS-11 had 286 ES-iPS-DMRs at P13, 194 sites at P18, 110 sites at P31, and 55 sites at P39. The ES-iPS-DMRs detected at P13 decreased with passaging (blue bars in upper-left panel in Figure 5A). Interestingly, 66 de novo ES-iPS-DMRs appeared at P18, while at P13 these sites showed no differences between UtE-iPS-11 and ESCs (orange bars in upper-left panel in Figure 5A). These 66 ES-iPS-DMRs also decreased with passaging (P31 and P39). The 29 additional ES-iPS-DMRs at P31 also appeared and decreased with passaging (P39) (green bars in upper-left panel in Figure 5A) and 16 ES-iPS-DMRs at P39 (red bar in upper-left panel in Figure 5A) appeared. Rapid appearance and gradual disappearance of ES-iPS-DMRs was a recurring theme, but the number of newly-appearing ES-iPS-DMRs decreased with passaging (Figure 5A, upper-left panel). The same change in ES-iPS-DMRs occurred on the X chromosome, but the number of the ES-iPS-DMRs approached zero at early passages (Figure 5A, upper-center panel). Intriguingly, this change also occurred at inherited sites, which was contrary to our expectations. The inherited sites also repeatedly appeared and disappeared, and the number of newly-appearing inherited sites decreased with passaging (Figure 5A, upper-right panel). The term "inherited" is here used to mean the same methylation state found in iPSCs and their parent cells, but the "inherited" regions behaved like "aberrant" regions that had multiple appearances and disappearances. These multiple appearances/disappearances of ES-iPS-DMRs were observed in all iPSC lines regardless of parental cell type. The ES-parent-DMRs were also analyzed. The *de novo* ES-parent-DMRs appeared as well as the ES-iPS-DMRs, but did not decrease with passaging (Figure 5B).

Most ES-iPS-DMRs were hyper-methylated in iPSCs

ES-iPS-DMRs can be categorized into two groups: a, hyper-methylated and b, hypo-methylated sites in iPSCs, as compared

Table 1. List of the top 20 out of 82 transcription factor genes with hypo-methylated stem cell-required DMRs exhibiting "high" expression in human iPS cells.

		DNA methylation		
TargetID	Gene name	HUESCs	iPSCs	Expression level
cg13083810	POU5F1, POU domain; class 5; transcription factor 1 isoform 1	0.584	0.549	55543.9
cg06303238	SALL4, sal-like 4	0.032	0.026	29766.2
cg16990174	RYBP, RING1 and YY1 binding protein	0.076	0.119	10274.1
cg03589001	MORF4L1, MORF-related gene 15 isoform 2	0.176	0.173	7015.7
cg02204046	MYCN, v-myc myelocytomatosis viral related oncogene; neuroblastoma derived	0.022	0.027	5826.8
cg10705800	CITED4, Cbp/p300-interacting transactivator; with Glu/Asp-rich carboxy-terminal domain; 4	0.438	0.445	5342.2
cg21696393	SOX8, SRY (sex determining region Y)-box 8	0.074	0.061	1976.7
cg23131007	TCF12, transcription factor 12 isoform b	0.138	0.155	1930.7
cg18808261	SATB1, special AT-rich sequence binding protein 1	0.194	0.242	1634.4
cg15607672	OTX2, orthodenticle 2 isoform a	0.046	0.054	1227.5
cg05345286	MDFI, MyoD family inhibitor	0.023	0.040	1035.9
cg20909686	OVOL1, OVO-like 1 binding protein	0.215	0.204	991.0
cg26209676	ZNF581, zinc finger protein 581	0.113	0.196	916.1
cg05522383	PITX2, paired-like homeodomain transcription factor 2 isoform b	0.024	0.030	544.8
cg17675150	ZNF532, zinc finger protein 532	0.069	0.107	525.3
cg01510051	ZNF542, zinc finger protein 542	0.585	0.555	443.9
cg06154570	HEYL, hairy/enhancer-of-split related with YRPW motif-like	0.134	0.152	440.3
cg12556134	TGIF2, TGFB-induced factor 2	0.075	0.072	405.4
cg03663715	FOXD1, forkhead box D1	0.030	0.042	349.1
cg09721427	HHEX, hematopoietically expressed homeobox	0.077	0.101	206.9

"Expression level" is an average of raw data values in iPSCs from Gene Chip data.
doi:10.1371/journal.pgen.1002085.t001

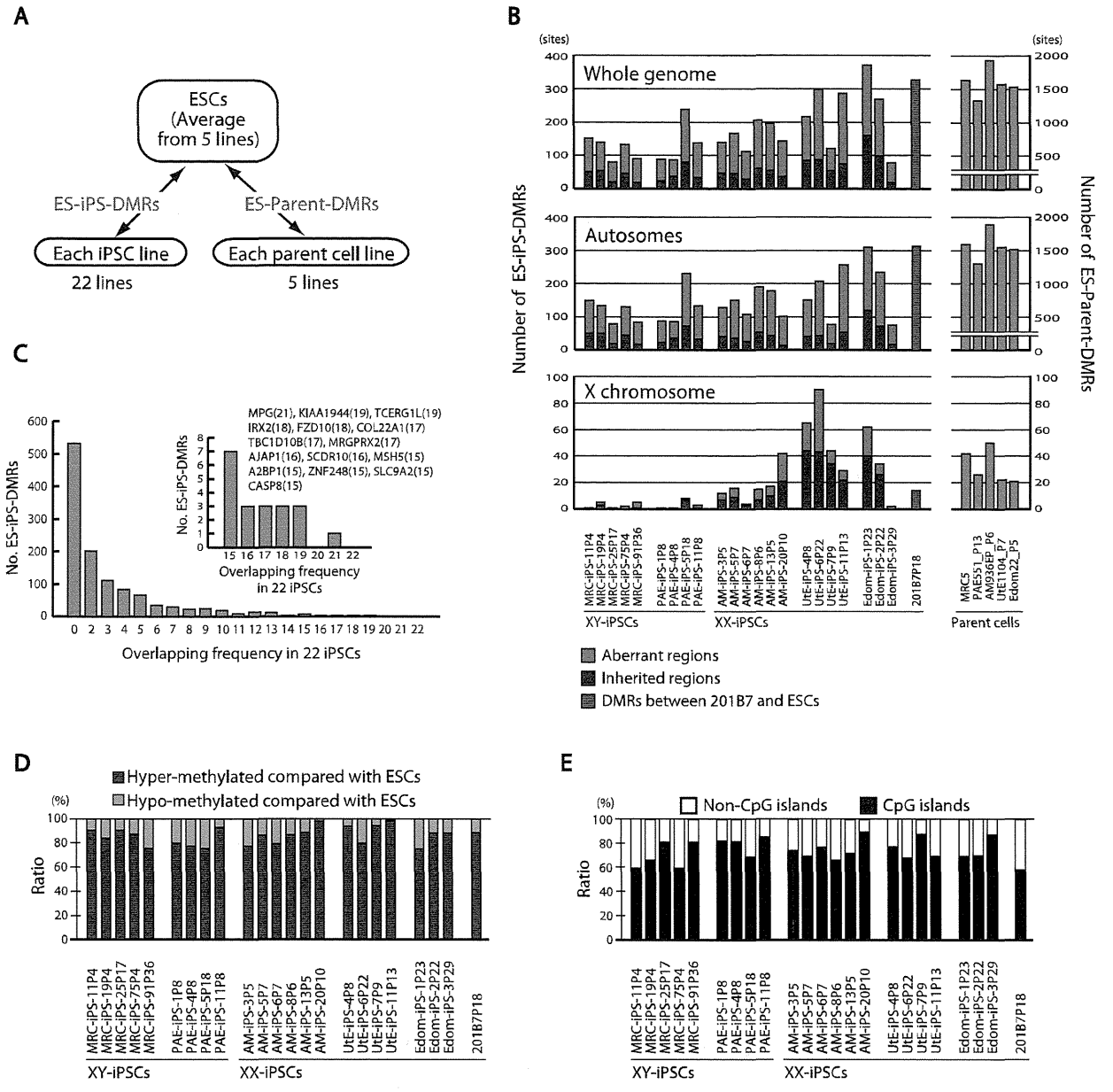
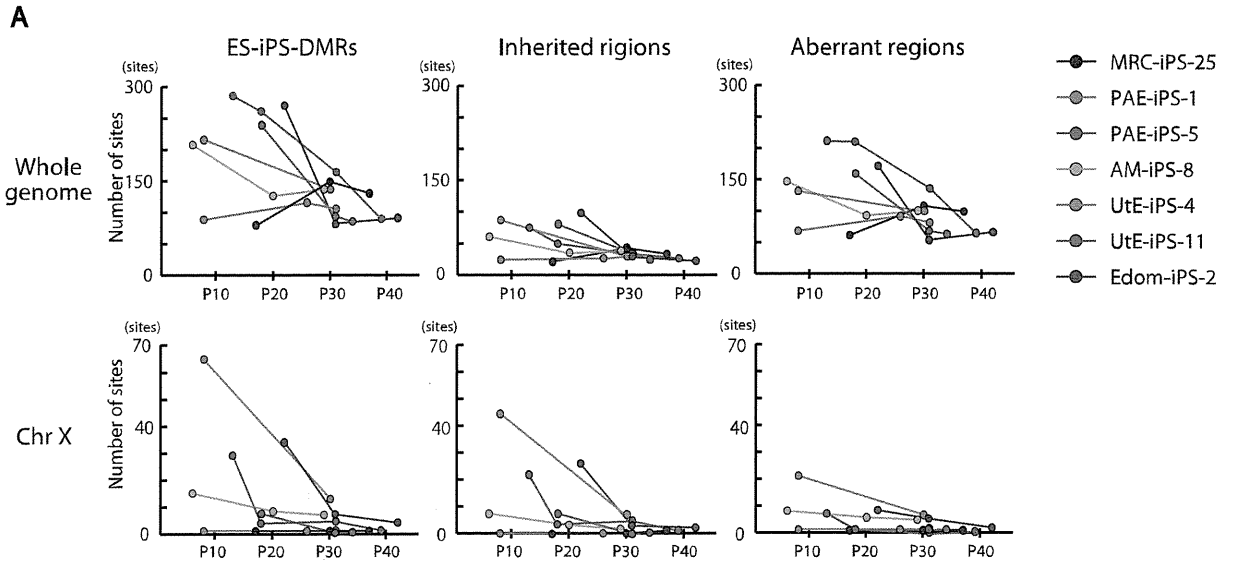
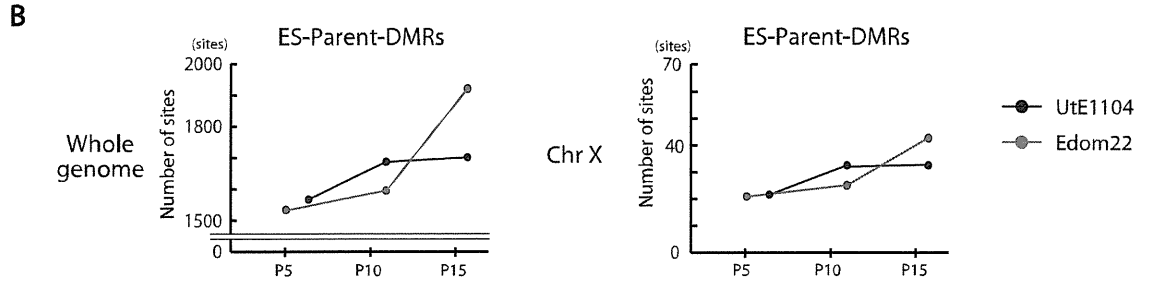


Figure 3. Aberrant methylation in human iPSCs. (*A*) Comparison of DNA methylation states of each iPSC line or each parent cell line with that of ESCs. The DMRs between ESCs and iPSCs are designated as ES-iPS-DMRs, and the DMRs between ESCs and parent cells are designated as ES-parent-DMRs. (*B*) The number of ES-iPS-DMRs and ES-parent-DMRs on whole genome (top), autosomes (middle) and X chromosome (bottom). Ratios of number of inherited regions in iPSCs from parent cells (blue) and aberrant regions in iPSCs that differ from ESCs and parent cells (red) in the ES-iPS-DMRs are shown in bars. Female iPSCs were demonstrated to carry high number of EiP-DMRs on X chromosome. (*C*) Number of overlapped ES-iPS-DMRs frequency in iPSCs. No overlapping ES-iPS-DMRs in all 22 iPSC lines. (Inlet) A small number of overlapping ES-iPS-DMRs of the frequency from 15 to 22. Overlapping frequency of each gene is indicated in parentheses. (*D*) Proportion of the hyper- and hypo-methylated ES-iPS-DMRs. More than 75% of the ES-iPS-DMRs were hyper-methylated in iPSCs. (*E*) Proportion of the ES-iPS-DMRs associated with CpG islands and non-CpG islands in ach iPSC line. ES-iPS-DMRs were biased to CpG islands.
doi:10.1371/journal.pgen.1002085.g003

with ESCs. ES-iPS-DMRs that disappeared at the last passage (P39) (blue bars in Figure 5) in both UtE-iPS-11 and Edom-iPS-2 were extracted, and each methylation score of the extracted ES-iPS-DMRs is shown (Figure 6, upper and middle panels). To compare methylation scores, a "difference value" was estimated by subtracting the scores of ESCs from those of each cell (Figure 6, lower panels). Positive and negative difference values indicate that these sites are hyper- and hypo-methylated, respectively, when compared with ESCs. Difference values of the ES-iPS-DMRs showing aberrant methylation states in iPSCs at the early passage approached zero with passaging. It should be noted that the almost all difference values became largely positive in iPSCs at


A
ES-iPS-DMRs
Inherited rigions
Aberrant regions
Whole genome
Chr X
Number of sites
(sites)
300
150
0
70
40
P10
P20
P30
P40
MRC-iPS-25
PAE-iPS-1
PAE-iPS-5
AM-iPS-8
UtE-iPS-4
UtE-iPS-11
Edom-iPS-2


B
ES-Parent-DMRs
ES-Parent-DMRs
Whole genome
Chr X
Number of sites
(sites)
2000
1800
1500
0
70
40
P5
P10
P15
UtE1104
Edom22

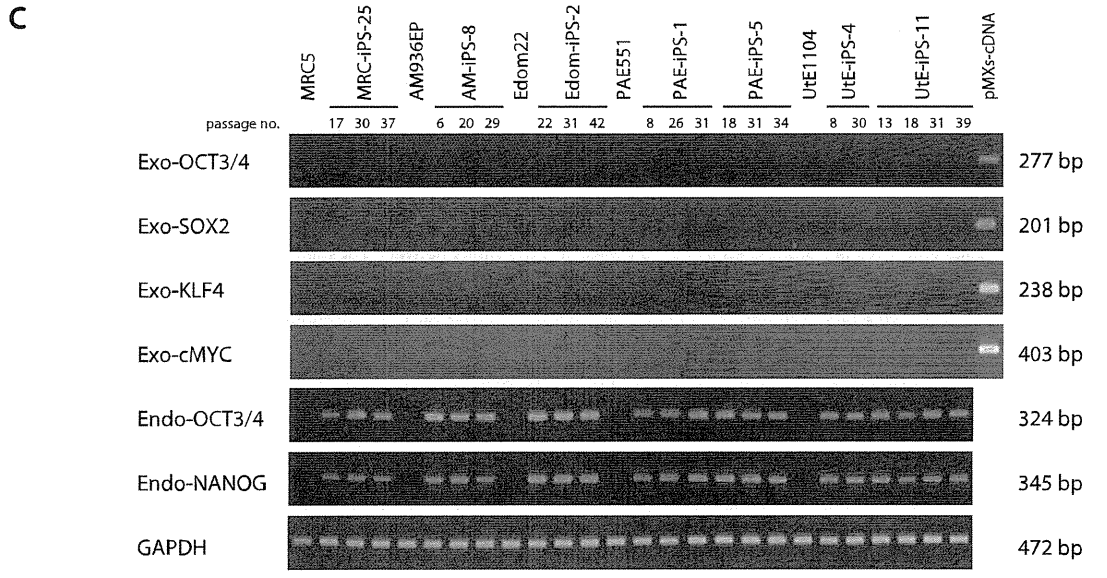

C
MRC5
MRC-iPS-25
AM936EP
AM-iPS-8
Edom22
Edom-iPS-2
PAE551
PAE-iPS-1
PAE-iPS-5
UtE1104
UtE-iPS-4
UtE-iPS-11
pMXs-cDNA
passage no.
17 30 37
6 20 29
22 31 42
8 26 31
18 31 34
8 30
13 18 31 39
Exo-OCT3/4
277 bp
Exo-SOX2
201 bp
Exo-KLF4
238 bp
Exo-cMYC
403 bp
Endo-OCT3/4
324 bp
Endo-NANOG
345 bp
GAPDH
472 bp

Figure 4. Effect of long-term cultivation on ES-iPS-DMRs. (*A*) Decrease in the number of the ES-iPS-DMRs with continuous passaging. Upper panels show change of the number of the ES-iPS-DMRs (left), the inherited regions (middle) and aberrant regions (right) on whole genome. Lower panels show change in the number of the ES-iPS-DMRs (left), inherited regions (middle) and aberrant regions (right) on X chromosome. The number of the ES-iPS-DMRs in XX-iPSCs approached zero with continuous passaging on X chromosome. In contrast, XY-iPSCs had few ES-iPS-DMRs on X chromosome throughout the passages. (*B*) The number of the ES-parent-DMRs with continuous passaging. (*C*) No expression of the transgenes in iPSCs at each passage was detected by RT-PCR.
doi:10.1371/journal.pgen.1002085.g004

early passage (P13 or P22), even though they were negative in the parent cells, and then approached zero upon further passaging. This transiently-induced hyper-methylation was observed at each passage in all iPSC lines examined. The observed transient hypermethylation patterns during iPS reprogramming did not correspond to methyflated CpGs in the parental cells. However, this observation does not rule out that transient aberrant methylation could also be observed in some cases on sites that were methylated in the parental cells.

Discussion

Identification of novel epigenetic iPS markers

OCT-4/3 and *NANOG* have been used as epigenetic markers for iPSCs [8–10,26,27]. We previously showed candidate epigenetic markers by analyzing 6 iPS lines [17]. Here we identified 8 novel epigenetic markers more closely by defining 9 genes with the hypo-methylated stem cell-specific DMRs and significantly higher expression, and 17 genes with the hyper-methylated stem cell-specific DMRs and significantly lower expression in iPSCs/ESCs from 22 iPS lines. DNA methylation and expression of these genes, especially the 8 genes, *SALL4*, *EPHA1*, *PTPN6*, *RAB25*, *GBP4*, *LYST*, *SP100* and *UBE1L*, can now be used as epigenetic markers for pluripotent stem cells. Among these 8 genes, *SALL4* has been used as an expression marker, and is revealed for the first time as an epigenetic marker. These epigenetic changes during reprogramming can be detected by 3 different methods (Illumina assay, COBRA and bisulfite sequencing), and is evident, i.e. CpG sites are methylated or unmethylated in an all-or-none fashion. The identification of these novel epigenetic markers can be another tool for the validation of pluripotent stem cells that are iPSCs and ESCs.

The hypo-methylated stem cell-required DMRs may have an important role for reprogramming as do the stem cell-specific DMRs, because reprogramming is dependent on the type of parent cells. In fact, genes associated with the hypo-methylated stem cell-required DMRs include a large number of transcription factors that are involved in pluripotency. Establishment of the stem cell-required DMRs database in iPSCs derived from different types of parent cells can help to generate human iPSCs in a fast and easy manner. Hypo-methylated stem cell-specific regions have been reported to be abundant in CpG islands [28–30]. In this study, the hypo-methylated stem cell-specific DMRs were significantly biased towards CpG islands, whereas the hyper-methylated stem cell-specific DMRs were biased to non-CpG islands, suggesting that genes with CpG islands have a propensity to be demethylated during reprogramming towards pluripotent stem cells. The higher number of the hyper-methylated stem cell-specific DMRs in iPSCs indicates that the Yamanaka factors activate only limited numbers of stem cell-specific/associated genes through demethylation of the specific DMRs shown in this study on the genome in parallel with methylating most genes associated with tissue-specific function during reprogramming.

Multiple appearances/disappearances of aberrant hyper-methylation

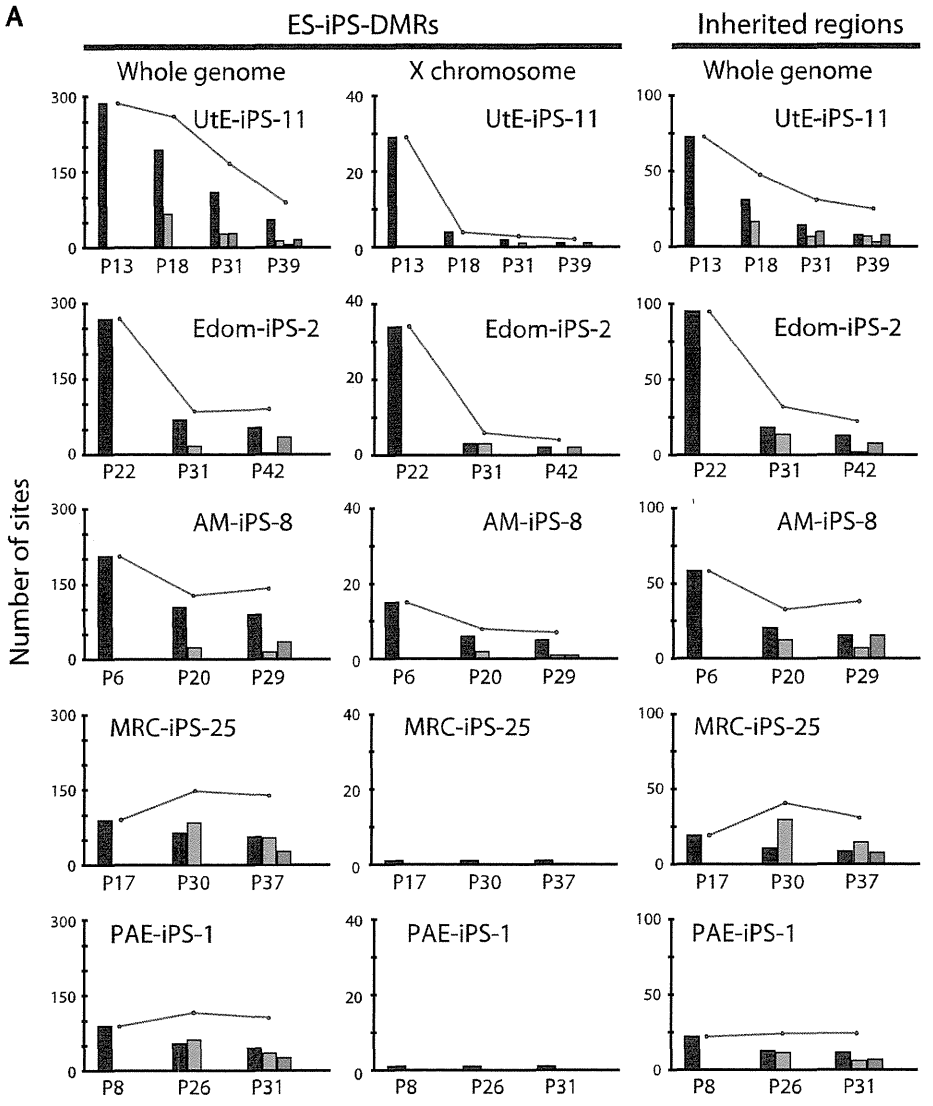
Continuous passaging of iPSCs reduces differences among clones in gene expression profiles in mouse [15] and in human [31] cells. Here we detected multiple appearances and disappearances of aberrant hyper-methylation throughout iPSC reprogramming. Furthermore, human iPSCs were gradually reprogrammed through the "convergence" of periodic aberrant hyper-methylation upon continuous passaging (Figure 7). The term "convergence" is used here to mean that amplitude of aberrant hyper-methylation (or number of ES-iPS-DMRs) decreases. The decrease of aberrant methylation suggests that iPSCs lose the characteristics inherited from the parent cells and adapt to ESCs. This aberrant and stochastic hyper-methylation and their convergence may be a direct cause of the transgene-independent phases of iPS reprogramming [15]. Aberrant hyper-methylation, for which the mechanism remains unclear, can possibly be attributed, at least in part, to up-regulation of DNMT3B, a de novo methyltransferase, at the early stages of reprogramming.

Maintenance of an epigenetic memory of their parent cells at early passage of human iPSCs (Figure 4A) is consistent with recent reports involving mouse iPSCs [15–17]. However, most inherited sites from the parent cells in iPSCs were inconsistent among iPSC clones from the same parent cells on the genome, and these sites showed periodic aberrant hyper-methylation during passaging, as well as aberrant sites. Inherited methylation is non-synchronous and stochastic, much like aberrant methylation, rather than deterministic. The inherited sites thus comprise a portion of all aberrant methylation observed in the clones.

Mouse female iPSCs as well as mouse female ESCs carry two active X chromosomes [32], but inactivation of the X chromosome in human female ESCs is variable [22,33–35]. It has been reported recently that human female iPSCs show a variable state of X-inactivation as is seen in human female ESCs [22,36]. In this study, human iPSCs exhibited a dynamic epigenetic state on the X chromosome. The ES-iPS-DMRs on the X chromosome in XY-iPSCs were rare and the average number of ES-iPS-DMRs in XY-iPSCs was significantly lower than in XX-iPSCs, suggesting that iPSCs are prone to aberrant hyper-methylation on the inactive X chromosome. A recent report showed that X inactivation in human ESCs is sensitive to the level of oxygen through culture in vitro [35]. Therefore, analysis of aberrant methylation in iPSCs that are established and cultured in low oxygen condition would be help to understand physiological relevance of X inactivation and reprogramming.

Incomplete adaptation of iPSCs to ESCs

The number of passages for "convergence" of the aberrant hyper-methylation seems be dependent on parental cell types and their sex. Disappearance of iPSCs in culture within 10 passages is occasionally observed, regardless of the cell of origin. This instability may be due to an excess of aberrant hyper-methylation at early passages in addition to the "partial reprogramming" theory [15]. The late-passage iPSCs, like the early-passage iPSCs, retained the ability to differentiate into cell types found in all three germ layers. iPSCs showed reduced aberrant methylation during adaptation to ESCs; however, iPSCs retained approximately 100 aberrant sites on autosomes, implying that iPSCs do not become identical to ESCs, although they become very close. The remaining aberrant sites were inconsistent among iPSC clones


A
ES-iPS-DMRs
Inherited regions
Whole genome
X chromosome
Whole genome
Number of sites
UtE-iPS-11
Edom-iPS-2
AM-iPS-8
MRC-iPS-25
PAE-iPS-1
P13 P18 P31 P39
P22 P31 P42
P6 P20 P29
P17 P30 P37
P8 P26 P31

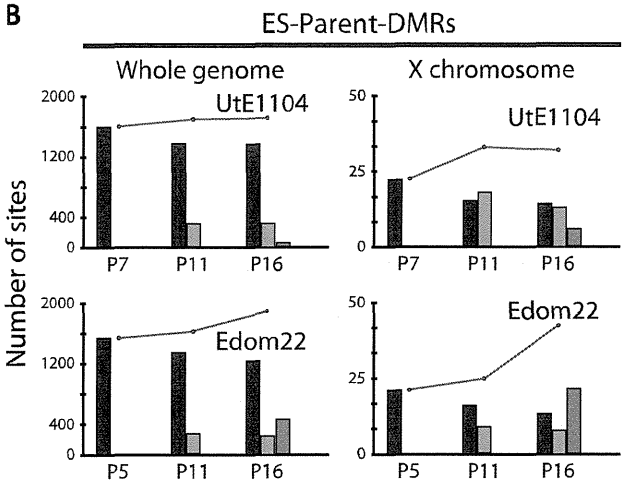

B
ES-Parent-DMRs
Whole genome
X chromosome
Number of sites
UtE1104
Edom22
P7 P11 P16
P5 P11 P16

Figure 5. Number of the ES-iPS-DMRs and ES-parent-DMRs with passaging. (*A*) Number of the ES-iPS-DMRs with passaging. Red line plots indicate total number of the ES-iPS-DMRs. Blue bars indicate the number of the ES-iPS-DMRs that appeared at the earliest passage. Orange, green and red bars indicate the number of the ES-iPS-DMRs that appeared secondarily at later passages. Appearance/disappearance of the ES-iPS-DMRs and inherited regions were repeated, but the number of newly-appeared ES-iPS-DMRs was decreased with passaging. (*B*) Number of the ES-parent-DMRs with passaging. Blue bars indicate the number of the ES-parent-DMRs at P5 (or P7). Orange and green bars indicate de novo ES-parent-DMRs at P11 and P16, respectively.
doi:10.1371/journal.pgen.1002085.g005

with different parent cell types, but the numbers were consistent among iPSC clones after a 42-week cultivation. The quantity (or number) of ES-iPS-DMRs would be another validation index for iPSC identity as well as quality analysis (or methylation ratio) of pluripotent stem cell-specific methylation.

Abnormalities of imprint genes, *MEG3* genes, and *H19* genes in human iPSCs

Genomic imprinting of *H19*, *IGF2* and *MEG3* has been reported to be unstable in human ESCs [37,38]. The *Dlk1-Dio3* genes were aberrantly silenced in most of the mouse iPSC lines. But mouse iPSCs without *MEG3* expression still have the ability to differentiate into cell type of three germ layers *in vitro* [39]. In humans, IG-DMR and MEG3-DMR are relevant to upd(14)pat-like and upd(14)mat-like phenotypes [40]. In this study, only *MEG3* and *H19*, out of 87 imprinted genes examined showed aberrant methylation in human iPSCs (Figure S10). Six out of 15 human iPSC lines were aberrantly methylated at MEG3-DMR. *MEG3* expression was silenced in those six lines regardless of their parent cell type, although all parent cells showed about 50% methylation at MEG3-DMR and expression of *MEG3* (Figure S10A, S10B). However, MEG3-negative iPS lines are almost indistinguishable from MEG3-positive iPS lines in DNA methylation and gene expression in human. Continuous passaging

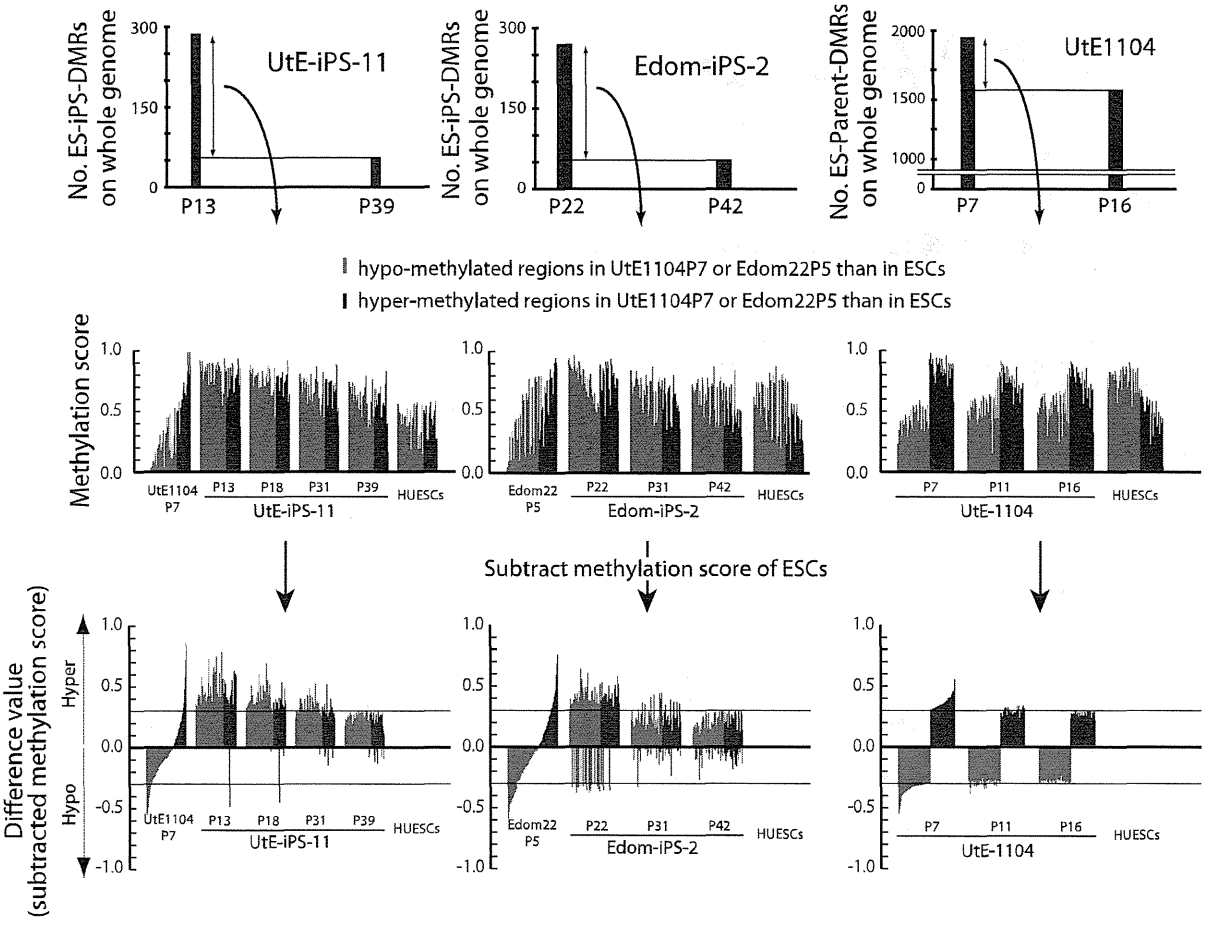


Figure 6. Hyper-methylation in the ES-iPS-DMRs and ES-parent-DMRs. ES-iPS-DMRs that disappeared in UtE-iPS-11 and Edom-iPS-2 at the latest passage (upper) were analyzed and the methylation score of each ES-iPS-DMR was plotted on bar graph (middle). To clearly compare methylation scores, difference value were estimated by subtracting the scores of ESCs from that of each sample (lower). Red and blue bars represent hypo- and hyper-methylated regions, respectively, in the parent cells, compared with ESCs. Notably, almost all the regions, even though their difference values were hypo-methylated in the parent cells, became hyper-methylated in iPSCs at the early passage, and then their methylation levels were adjusted to the level of ESCs with passaging, i.e. subtracted methylation score became close to zero. This transiently-induced hyper-methylation was not detected in parent cells.
doi:10.1371/journal.pgen.1002085.g006

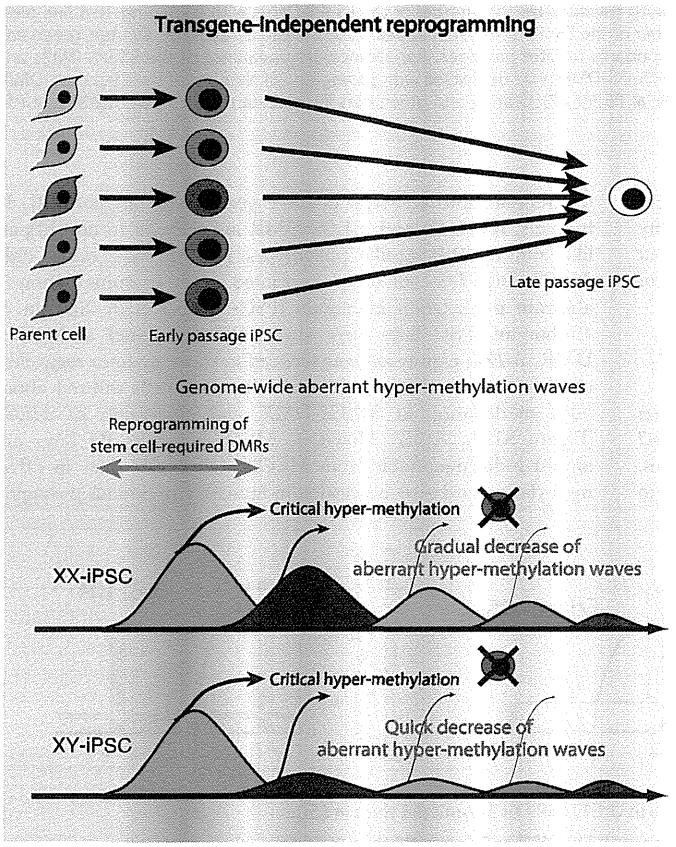


Figure 7. Model of mechanism for transgene-independent reprogramming. During reprogramming from somatic cells to iPSCs, the cells undergo dynamic change of methylation of SS-DMRs and genome. The cells with incomplete reprogramming or excessive hyper-methylation of the genome fail to maintain pluripotency at early passages. Human iPSCs are transgene-independently reprogrammed gradually through "convergence" of periodic aberrant hyper-methylation and become closer to ESCs upon continuous passaging. Due to the sensitivity to aberrant methylation on X chromosome, XY-iPSCs become close to ESCs faster than XX-iPSCs do.
doi:10.1371/journal.pgen.1002085.g007

did not resolve the aberrant hyper-methylation at MEG3-DMR, suggesting that these abnormalities occur at early passage and are fixed at later stages. In addition, aberrant hyper-methylation at *H19* in all iPSCs and ESCs was observed (Figure S10C), and *H19* was not expressed in all iPSCs and their parent cells.

We revealed that transgene-independent reprogramming is a convergence of periodic hyper-methylation. The aberrant hyper-methylation in iPSCs occurs stochastically throughout the genome. Early-stage iPSC clones with different propensities due to stochastic hyper-methylation may be used after selection of desirable phenotypes to treat a wide range of target diseases using cell-based therapy, and would thus have advantages for clinical use. In this sense, the number of ES-iPS-DMRs and methylation states of the stem cell-specific DMRs are useful epigenetic indices for evaluating human iPSCs in therapeutic applications.

Materials and Methods

Ethics statement

Human endometrium, amnion, placental artery endothelium and menstrual blood cells were collected by scraping tissues from surgical specimens, under signed informed consent, with ethical approval of the Institutional Review Board of the National Institute for Child Health and Development, Japan. Signed informed consent was obtained from donors, and the surgical specimens were irreversibly de-identified. All experiments handling human cells and tissues were performed in line with Tenets of the Declaration of Helsinki.

Human cell culture

Endometrium (UtE1104), amnion (AM936EP), placental artery endothelium (PAE551) and menstrual blood cell (Edom22) cell lines were independently established in our laboratory [41,42]. UtE1104, AM936EP, Edom22, and MRC-5 [43] cells were maintained in the POWEREDBY10 medium (MED SHIROTORI CO., Ltd, Tokyo, Japan). PAE551 cells were cultured in EGM-2MV BulletKit (Lonza, Walkersville, MD, USA) containing 5% FBS. Human iPSCs were generated in our laboratory, via procedures described by Yamanaka and colleagues [8] with slight modification [17,41,44–46]. The human cells were infected with retroviruses produced from the retroviral vector pMXs, which encodes the cDNA for human *OCT3/4*, *SOX2*, *c-MYC*, and *KLF4*. Human iPSCs were established from MRC-5, AM936EP, UtE1104, and PAE551, which were designated as MRC-iPSCs, AM-iPSCs, UtE-iPSCs and PAE-iPSCs [17,41,44–46]. Edom-iPSCs were established from Edom22 in this study. Human iPSCs were maintained on irradiated MEFs in 0222 medium (MED SHIROTORI CO., Ltd, Tokyo, Japan) supplemented with

10 ng/ml recombinant human basic fibroblast growth factor (bFGF, Wako Pure Chemical Industries, Ltd., Osaka, Japan). The 201B7 human iPSC line [8] that was generated from human skin fibroblasts by retroviral transfection with 4 transcription factors was also used. Frozen pellets of human ESCs (HUESCs) [23,24] were kindly gifted from Drs. C. Cowan and T. Tenzan (Harvard Stem Cell Institute, Harvard University, Cambridge, MA).

DNA methylation analysis

DNA methylation analysis was performed using the Illumina infinium assay with the HumanMethylation27 BeadChip (Illumina) and the BeadChip was scanned on a BeadArray Reader (Illumina), according to the manufacturer's instructions. Methylated and unmethylated signals were used to compute a β-value, which was a quantitative score of DNA methylation levels, ranging from "0", for completely unmethylated, to "1", for completely methylated. On the HumanMethylation27 BeadChip, oligonucleotides for 27,578 CpG sites covering more than 14,000 genes are mounted, mostly selected from promoter regions. CpG sites with ≥0.05 "Detection p value" (computed from the background based on negative controls) were eliminated from the data for further analysis, leaving 24,273 CpGs (13,728 genes) valid for use with the 51 samples tested. Average of methylation was calculated from HUESCs, MRC-iPSCs, AM-iPSCs, UtE-iPSCs, PAE-iPSCs and Edom-iPSCs, in which DMRs among each line in the each set were removed. Analyzed data sets (list of stem cell-specific DMRs and stem cell-required DMRs) can be obtained from http://www.nch.go.jp/reproduction/e/thdmds.html.

Gene expression analysis

Gene expression analysis was performed using the Agilent Whole Human Genome Microarray chips G4112F (Agilent, Santa Clara, CA), which contains over 41,000 probes. Raw data were normalized and analyzed by GeneSpringGX11 software (Silicon Genetics, Redwood City, CA). For RT-PCR, an aliquot of total RNA was reverse-transcribed using Random Hexamer primers. The cDNA template was amplified using specific primers for *EPHA1*, *PTPN6*, *RAB25*, *SALL4*, *GBP3*, *LYST*, *SP100*, *UBE1L*, OCT3/4 and NANOG. For detecting RNA derived from transgenes, specific primer sets, FY-11 and OCT3/4-SR, FY-11 and SOX2-SR, KLF4-SF and FY-12, cMYC-SF and FY-12, were used. Expression of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as a control. Primers used in this study are summarized in Table S8.

Quantitative combined bisulfite restriction analysis (COBRA) and bisulfite sequencing

To confirm the DNA methylation state, bisulfite PCR-mediated restriction mapping (known as the COBRA method) was performed. Sodium bisulfite treatment of genomic DNA was carried out using EZ DNA Methylation-Gold kit (Zymo Research). PCR amplification was performed using BIOTAQ HS DNA polymerase (Bioline Ltd; London, UK) with specific primers for *EPHA1*, *PTPN6*, *RAB25*, *SALL4*, *GBP3*, *LYST*, *SP100*, and *UBE1L*. Primers used in this study are summarized in Table S8. After digestion with restriction enzymes, HpyCH4IV or Taq I, quantitative-COBRA coupled with the Shimadzu MCE-202 MultiNA Microchip Electrophoresis System (Shimadzu, Japan) was carried out for quantitative DNA methylation level. To determine the methylation state of individual CpG sites, the PCR product was gel extracted and subcloned into pGEM T Easy vector (Promega, Madison, WI), and then sequenced. The promoter regions of the *OCT3/4* and *NANOG* [41,44] were also amplified and sequenced. Methylation sites were visualized and quality control was carried out by the web-based tool, "QUMA" (http://quma.cdb.riken.jp/) [47].

Web tools

The following web tools were used in this study: NIA Array [48] (http://lgsun.grc.nia.nih.gov/ANOVA/) for hierarchical clustering, DAVID Bioinformatics Resources [49] (http://david.abcc.ncifcrf.gov/home.jsp), PANTHER Classification System [50] (http://www.pantherdb.org/).

Accession numbers

NCBI GEO: HumanMethylation27 BeadChip data and gene expression microarray data have been submitted under accession number GSE 20750, GSE24676 and GSE24677.

Supporting Information

Figure S1 Immunohistochemistry of stem cell-specific surface antigens, NANOG, OCT3/4, SOX2, SSEA-4 and TRA-1-60 in AM-iPSCs, MRC-iPSCs and Edom-iPSCs, and teratoma formation of those iPSCs by subcutaneous implantation into NOD/Scid mice. The iPSCs differentiated to various tissues including ectoderm (neural tissues and retinal pigment epithelium), mesoderm (cartilage) and endoderm (gut). Immunostaining and teratoma formation were carried out as previously described [41,44].
(PDF)

Figure S2 Immunohistochemistry of stem cell-specific surface antigens, NANOG, OCT3/4, SOX2, SSEA-4 and TRA-1-60 in PAE-iPSCs and UtE-iPSCs, and teratoma formation of those iPSCs by subcutaneous implantation into NOD/Scid mice. The iPSCs differentiated to various tissues including ectoderm (neural tissues and retinal pigment epithelium), mesoderm (cartilage) and endoderm (gut). Immunostaining and teratoma formation were carried out as previously described [41,44].
(PDF)

Figure S3 Bisulfite sequencing at the OCT3/4 and NANOG promoter regions in ESCs, iPSCs and their parent cells.
(PDF)

Figure S4 Expression of the transgenes in iPSCs. (A) RT-PCR for transgenes in 22 iPSC lines. No expression of the transgenes in each iPSC lines was detected. (B) Quantitative RT-PCR for the transgenes at each passage. Relative expression of each transgene normalized to GAPDH was calculated. P0(D2), RNA from UtE1104 cells that were infected with the retroviruses and were cultured for 2 days. No expression of the transgenes at each passage was detected.
(PDF)

Figure S5 (A) Unsupervised hierarchical clustering analysis based on DNA methylation (left) and gene expression (right) in each ESC line, iPSC line and their parent cell line. (B) Unsupervised hierarchical clustering analysis based on DNA methylation (left) and gene expression (right) of average of ESCs, iPSCs and parent cells. (C) Scatter plot of DNA methylation (left) and gene expression data (right) in ESCs, iPSCs and their parent cells.
(PDF)

Figure S6 (A) Venn-like diagram showing seven categories (aa-gg) overlapped CpG sites among ESCs, iPSCs and their parent cells. (B) Number of CpG sites involved in each seven category from the five ESCs-iPSCs-the parent cell sets. "Overlapped"

indicates a number of sites that overlap in all iPSCs examined. The 220 overlapping sites in "ee" are designated as stem cell-specific differentially methylated regions (DMRs) and 3,123 total sites in "ee" are designated as stem cell-required DMRs. Notably, no overlapping sites were observed in "bb" that is a category involved in iPSCs-specific DMRs and in "ff" that is a category involved in inherited regions in iPSCs from the parent cells.
(PDF)

Figure S7 (A) Distribution of stem cell-required DMRs on each chromosome (upper) and frequency on each chromosome (bottom). (B) The number of parent cell specific DMRs (left) and the number of iPSC derived from different parent cells specific DMRs (left).
(PDF)

Figure S8 The number of DMRs between ESCs and each iPSC line (ES-iPS-DMRs) on each chromosome. ES-iPS-DMRs between 201B7 (iPSCs from Yamanaka) and ESCs are shown for comparison.
(PDF)

Figure S9 Distribution of the ES-iPS-DMRs on each chromosome. Distribution of the EiP-DMRs overlapped in less than 9 lines (light blue bars), in more than 10 and less than 14 lines (blue bars), and in more than 15 lines (red bars) among 22 lines.
(PDF)

Figure S10 DNA methylation at human *MEG3* and *H19*. (A) DNA methylation at MEG3-DMR (CG7) and expression of MEG3. (Top) Schematic diagram of the MEG3 gene. The arrow, open boxes and open circles represent transcription start site, first exon and position of CpG sites, respectively. Red and blue arrowheads represent the position of CpG sites in Infinium assay and COBRA assay, respectively. DNA methylation scores of MEG3 were determined by Illumina Infinium HumanMethylation27 assay (upper bar graph) and Bio-COBRA (lower bar graph). (Bottom) Expression of MEG3 and GAPDH was determined by RT-PCR. Information of MEG3 primers for COBRA and RT-PCR is described by Kagami et al. [40]. (B) Bisulfite sequencing analysis of MEG3-DMRs (CG7). (C) Methylation scores of H19 were determined by Illumina Infinium HumanMethylation27 assay.
(PDF)

Table S1 List of human cells analyzed for a methylation state in this study.
(PDF)

Table S2 STR analysis of iPSCs.
(PDF)

Table S3 Karyotypic analysis of iPSCs.
(PDF)

Table S4 List of genes with stem cell-specific DMRs exhibiting significant changes in expression in human iPS cells.
(PDF)

Table S5 List of the top 100 genes with hypo-methylated stem cell-required DMRs exhibiting 'high' expression in human iPS cells.
(PDF)

Table S6 List of top 100 genes with hyper-methylated stem cell-required DMRs exhibiting suppression in human iPS cells.
(PDF)

Table S7 List of top 5 categories of GO Term in "Stem cell-required DMRs".
(PDF)

Table S8 Primer list.
(PDF)

Acknowledgments

We would like to express our sincere thanks to Drs. C. Cowan and T. Tenzan for HUESC lines; to Drs. K. Hata and K. Nakabayashi for COBRA; to Dr. H. Makino for establishing the AM936EP, UtE1104, PAE551, and Edom22 cells; to Mr. M. Machida for immunohistochemical analysis; to Ms. Y. Takahashi for bioinformatics analyses; and Dr C. Ketcham for critical proofreading.

Author Contributions

Conceived and designed the experiments: KN AU. Performed the experiments: KN MT MY-I. Analyzed the data: KN. Contributed reagents/materials/analysis tools: KN MT MY-I YF EC HS HA. Wrote the paper: KN AU.

References

1. Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet 3: 662–673.
2. Reik W (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447: 425–432.
3. Feng S, Jacobsen SE, Reik W (2010) Epigenetic reprogramming in plant and animal development. Science 330: 622–627.
4. Hattori N, Nishino K, Ko YG, Ohgane J, Tanaka S, et al. (2004) Epigenetic control of mouse Oct-4 gene expression in embryonic stem cells and trophoblast stem cells. J Biol Chem 279: 17063–17069.
5. Hattori N, Imao Y, Nishino K, Ohgane J, Yagi S, et al. (2007) Epigenetic regulation of Nanog gene in embryonic stem and trophoblast stem cells. Genes Cells 12: 387–396.
6. Nishino K, Hattori N, Tanaka S, Shiota K (2004) DNA methylation-mediated control of Sry gene expression in mouse gonadal development. J Biol Chem 279: 22306–22313.
7. Zingg JM, Pedraza-Alva G, Jost JP (1994) MyoD1 promoter autoregulation is mediated by two proximal E-boxes. Nucleic Acids Res 22: 2234–2241.
8. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, et al. (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131: 861–872.
9. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, et al. (2007) Induced pluripotent stem cell lines derived from human somatic cells. Science 318: 1917–1920.
10. Park IH, Zhao R, West JA, Yabuuchi A, Huo H, et al. (2008) Reprogramming of human somatic cells to pluripotency with defined factors. Nature 451: 141–146.
11. Woltjen K, Michael IP, Mohseni P, Desai R, Mileikovsky M, et al. (2009) piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. Nature 458: 766–770.
12. Park IH, Arora N, Huo H, Maherali N, Ahfeldt T, et al. (2008) Disease-specific induced pluripotent stem cells. Cell 134: 877–886.
13. Miura K, Okada Y, Aoi T, Okada A, Takahashi K, et al. (2009) Variation in the safety of induced pluripotent stem cell lines. Nat Biotechnol 27: 743–745.
14. Ghosh Z, Wilson KD, Wu Y, Hu S, Quertermous T, et al. (2010) Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. PLoS ONE 5: e8975. doi:10.1371/journal.pone.0008975.
15. Polo JM, Liu S, Figueroa ME, Kulalert W, Eminli S, et al. (2010) Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. Nat Biotechnol 28: 848–855.
16. Kim K, Doi A, Wen B, Ng K, Zhao R, et al. (2010) Epigenetic memory in induced pluripotent stem cells. Nature.
17. Nishino K, Toyoda M, Yamazaki-Inoue M, Makino H, Fukawatase Y, et al. (2010) Defining Hypo-Methylated Regions of Stem Cell-Specific Promoters in Human iPS Cells Derived from Extra-Embryonic Amnions and Lung Fibroblasts. PLoS ONE 5: e13017. doi:10.1371/journal.pone.0013017.
18. Fazzari MJ, Greally JM (2004) Epigenomics: beyond CpG islands. Nat Rev Genet 5: 446–455.
19. Fazzari MJ, Greally JM (2010) Introduction to epigenomics and epigenome-wide analysis. Methods Mol Biol 620: 243–265.

20. Doi A, Park IH, Wen B, Murakami P, Aryee MJ, et al. (2009) Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat Genet 41: 1350–1353.
21. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, et al. (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471: 68–73.
22. Bock C, Kiskinis E, Verstappen G, Gu H, Boulting G, et al. (2011) Reference Maps of Human ES and iPS Cell Variation Enable High-Throughput Characterization of Pluripotent Cell Lines. Cell 144: 439–452.
23. Cowan CA, Klimanskaya I, McMahon J, Atienza J, Witmyer J, et al. (2004) Derivation of embryonic stem-cell lines from human blastocysts. N Engl J Med 350: 1353–1356.
24. Osafune K, Caron L, Borowiak M, Martinez RJ, Fitz-Gerald CS, et al. (2008) Marked differences in differentiation propensity among human embryonic stem cell lines. Nat Biotechnol 26: 313–315.
25. Brena RM, Auer H, Kornacker K, Plass C (2006) Quantification of DNA methylation in electrofluidics chips (Bio-COBRA). Nat Protoc 1: 52–58.
26. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126: 663–676.
27. Huangfu D, Osafune K, Maehr R, Guo W, Eijkelenboom A, et al. (2008) Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. Nat Biotechnol 26: 1269–1275.
28. Fouse SD, Shen Y, Pellegrini M, Cole S, Meissner A, et al. (2008) Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. Cell Stem Cell 2: 160–169.
29. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454: 766–770.
30. Sato S, Yagi S, Arai Y, Hirabayashi K, Hattori N, et al. (2010) Genome-wide DNA methylation profile of tissue-dependent and differentially methylated regions (T-DMRs) residing in mouse pluripotent stem cells. Genes Cells 15: 607–618.
31. Chin MH, Pellegrini M, Plath K, Lowry WE (2010) Molecular analyses of human induced pluripotent stem cells and embryonic stem cells. Cell Stem Cell 7: 263–269.
32. Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, et al. (2007) Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. Cell Stem Cell 1: 55–70.
33. Hall LL, Byron M, Butler J, Becker KA, Nelson A, et al. (2008) X-inactivation reveals epigenetic anomalies in most hESC but identifies sublines that initiate as expected. J Cell Physiol 216: 445–452.
34. Shen Y, Matsuno Y, Fouse SD, Rao N, Root S, et al. (2008) X-inactivation in female human embryonic stem cells is in a nonrandom pattern and prone to epigenetic alterations. Proc Natl Acad Sci U S A 105: 4709–4714.
35. Lengner CJ, Gimelbrant AA, Erwin JA, Cheng AW, Guenther MG, et al. (2010) Derivation of pre-X inactivation human embryonic stem cells under physiological oxygen concentrations. Cell 141: 872–883.
36. Tchieu J, Kuoy E, Chin MH, Trinh H, Patterson M, et al. (2010) Female human iPSCs retain an inactive X chromosome. Cell Stem Cell 7: 329–342.
37. Rugg-Gunn PJ, Ferguson-Smith AC, Pedersen RA (2005) Epigenetic status of human embryonic stem cells. Nat Genet 37: 585–587.
38. Rugg-Gunn PJ, Ferguson-Smith AC, Pedersen RA (2007) Status of genomic imprinting in human embryonic stem cells as revealed by a large cohort of independently derived and maintained lines. Hum Mol Genet 16 Spec No. 2: R243–251.
39. Stadtfeld M, Apostolou E, Akutsu H, Fukuda A, Follett P, et al. (2010) Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. Nature 465: 175–181.
40. Kagami M, Sekita Y, Nishimura G, Irie M, Kato F, et al. (2008) Deletions and epimutations affecting the human 14q32.2 imprinted region in individuals with paternal and maternal upd(14)-like phenotypes. Nat Genet 40: 237–242.
41. Nagata S, Toyoda M, Yamaguchi S, Hirano K, Makino H, et al. (2009) Efficient reprogramming of human and mouse primary extra-embryonic cells to pluripotent stem cells. Genes Cells 14: 1395–1404.
42. Cui CH, Uyama T, Miyado K, Terai M, Kyo S, et al. (2007) Menstrual blood-derived cells confer human dystrophin expression in the murine model of Duchenne muscular dystrophy via cell fusion and myogenic transdifferentiation. Mol Biol Cell 18: 1586–1594.
43. Jacobs JP, Jones CM, Baille JP (1970) Characteristics of a human diploid cell designated MRC-5. Nature 227: 168–170.
44. Makino H, Toyoda M, Matsumoto K, Saito H, Nishino K, et al. (2009) Mesenchymal to embryonic incomplete transition of human cells by chimeric OCT4/3 (POU5F1) with physiological co-activator EWS. Exp Cell Res 315: 2727–2740.
45. Saito S, Onuma Y, Ito Y, Tateno H, Toyoda M, et al. (2010) Potential linkages between the inner and outer cellular states of human induced pluripotent stem cells. BMC Bioinformatics, in press.
46. Toyoda M, Yamazaki-Inoue M, Itakura Y, Kuno A, Ogawa T, et al. (2010) Lectin microarray analysis of pluripotent and multipotent stem cells. Genes Cells, in press.
47. Kumaki Y, Oda M, Okano M (2008) QUMA: quantification tool for methylation analysis. Nucleic Acids Res 36: W170–175.
48. Sharov AA, Dudekula DB, Ko MS (2005) A web-based tool for principal component and significance analysis of microarray data. Bioinformatics 21: 2548–2549.
49. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.
50. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, et al. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33: D284–288.

RESOURCE

nature biotechnology

Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage

The International Stem Cell Initiative[1]

The International Stem Cell Initiative analyzed 125 human embryonic stem (ES) cell lines and 11 induced pluripotent stem (iPS) cell lines, from 38 laboratories worldwide, for genetic changes occurring during culture. Most lines were analyzed at an early and late passage. Single-nucleotide polymorphism (SNP) analysis revealed that they included representatives of most major ethnic groups. Most lines remained karyotypically normal, but there was a progressive tendency to acquire changes on prolonged culture, commonly affecting chromosomes 1, 12, 17 and 20. DNA methylation patterns changed haphazardly with no link to time in culture. Structural variants, determined from the SNP arrays, also appeared sporadically. No common variants related to culture were observed on chromosomes 1, 12 and 17, but a minimal amplicon in chromosome 20q11.21, including three genes expressed in human ES cells, *ID1*, *BCL2L1 and HM13,* occurred in >20% of the lines. Of these genes, *BCL2L1* is a strong candidate for driving culture adaptation of ES cells.

In human ES cell cultures, somatic mutations that generate a selective advantage, such as a greater propensity for self-renewal, can become fixed over time[1]. This selection may be the reason for the nonrandom genetic changes found in human ES cells maintained for long periods in culture. These changes, mostly detected by karyotypic analyses, commonly involve nonrandom gains of chromosomes 12, 17, 20 and X, or fragments of these chromosomes[2–12]. The embryonal carcinoma (EC) stem cells of human teratocarcinomas, the malignant counterparts of ES cells, though typically highly aneuploid, always contain amplified regions of the short arm of chromosome 12 and, commonly, gains of chromosomes 1, 17 and X[13–16]. Gain of chromosome 20q has also been noted in yolk sac carcinoma and nonseminomatous germ cell tumors, which contain EC cells[17–19]. Such observations suggest that these specific genetic changes in ES cells may be related to the nature of pluripotent stem cells themselves rather than the culture conditions. Mouse ES cells also undergo karyotypic changes upon prolonged passage[20], often with gain of mouse chromosomes 8 and 11 (ref. 21); mouse chromosome 11 is highly syntenic with human chromosome 17 (ref. 22).

Structural variants in otherwise karyotypically normal human ES cells have also been described[10,11,23,24]. These structural variants include gains on chromosome 4, 5, 15, 18 and 20 and losses on chromosome 10, although only gains on chromosome 20 were commonly observed in multiple cell lines.

Marked epigenetic changes have also been noted on prolonged passage; studies of global DNA methylation in human ES cells found considerable instability with time in culture[25,26]. Functional gain of the X chromosome, resulting from loss of X-chromosome inactivation in culture-adapted ES cells with two karyotypically normal X chromosomes has been reported[27]. On the other hand, some imprinted genes retain their monoallelic expression over long-term culture of human ES cells, although this stability is not invariant for all loci[28–31].

Because stem cells can adopt alternative fates (that is, self-renewal, differentiation or death), it might be expected that those maintained in the pluripotent state for many passages would be subject to strong selection favoring variants that enhance the probability of self-renewal[32]. Viewed in this light, the increased frequency of genetic variants in ES cell cultures over time might be considered inevitable[33]. Indeed, ES cell lines do often show progressive 'adaptation' to culture, with the result that late-passage cells may be maintained more easily, showing enhanced plating efficiencies[27]. Similarly, some mouse and human EC cell lines derived from germ cell tumors are nullipotent, as if selected for the capacity for self-renewal exclusively[34,35]. Taken together, these observations suggest that acquisition of extra copies of portions of chromosomes 12, 17, 20 and X by human ES and EC cells is driven by increased dosage of a gene or genes that favor self-renewal, independent of culture conditions.

However, there are also reports of human ES cell lines that have been maintained for many passages *in vitro* without overt karyotypic changes. It has been argued that some culture techniques, such as manual 'cutting and pasting' of ES cell colonies, favor maintenance of cells with a diploid karyotype[3,6]. As the appearance of a genetic variant in an ES cell culture must involve both mutation and selection, the low population size in cultures maintained by these methods may simply beat the mutation frequency[33]. Nevertheless, culture conditions themselves might influence the mutation rate independently of selection, and a population bottleneck, such as cloning, could allow a viable genetic variant to dominate in the absence of a selective advantage.

[1]A full list of authors and affiliations is provided at the end of the paper.

Received 6 September; accepted 26 October; published online 27 November 2011; doi:10.1038/nbt.2051

Candidate genes from the commonly amplified regions can be posited to provide the driving force for selection of variant ES cells, but direct evidence for the involvement of any specific gene is lacking. For example, *NANOG*, on human chromosome 12p, promotes the self-renewal of ES cells when overexpressed[36–38], but one of the two minimal amplicons of chromosome 12p in EC cells has been reported to exclude the *NANOG* locus[39]. It is also unclear to what extent changes affecting different loci are selected independently of one another or whether alterations at some loci act synergistically. Further, overexpression of disparate genes affecting a common pathway(s) could lead to an increased proliferative potential. Although the frequent gain of chromosomes 12, 17, 20 and X in both ES and EC cells argues for a selective advantage independent of culture conditions, changes affecting other regions might be more likely to depend upon culture conditions.

To provide better insight into the frequency and types of genetic changes affecting human ES cells on prolonged passage, the International Stem Cell Initiative (ISCI) surveyed by karyology and high-resolution SNP array 125 independent human ES cell lines, provided by 38 laboratories in 19 countries around the world, particularly to identify the common genetic changes that occur during prolonged culture (**Supplementary Table 1**). An opportunity was also taken to screen the samples against a specialized custom DNA methylation array focused on polycomb-target genes. These likely play a role in controlling ES cell differentiation and could be primary targets for the types of epigenetic change observed in cancer cells[40]. Thus, they may provide a source of selective advantage for variant stem cells. In most cases, each line was analyzed at both an early- and a late-passage level, using all three types of assay. The scale and design of this screen helped ensure that the ES cell lines sampled were representative of the world population. A group of 11 human iPS cell lines from three laboratories was also included to provide a pilot comparison of these pluripotent cells derived by reprogramming. Our results indicate that the common gains of chromosomes 12 and 17 in human ES cells are unlikely to be driven by the gain of single genes, but that the gain of chromosome 20 may be driven by the gain of a single gene, *BCL2L1*.

RESULTS

Diversity and population structure of the cell lines surveyed

To define the range of ethnicity represented by the human ES cell lines included in this study, we first analyzed the SNP calls identified in the SNP array data by referencing them to ethnically defined human genotyping data sets. Of the samples submitted for SNP analysis, three cell lines were included twice, and four pairs of ES cell lines and a set of three lines were identified as having a full sibling relationship (**Supplementary Table 1**). After accounting for these, 112 genetically unrelated ES cell lines passed SNP quality-control criteria. Subsequent analysis allowed us to determine whether specific structural variants found in particular cell lines are limited to the population from which they were derived or common to all human ES cell lines studied.

For population structure analysis, the international breadth of this study required the use of a diverse set of reference samples to compare to these 112 genetically unrelated cell lines. The reference samples were pooled from the HapMap[41], the human genome diversity panel (HGDP)[42] and the Pan-Asian SNP Initiative[43] to generate an ethnically diverse set of 1,868 reference samples. We performed cluster analysis[44] of the human ES cell samples against these reference samples, using the CEU (European), Chinese, Japanese and African HapMap populations as references, to arrive at the population structure of the human ES cell lines analyzed (**Fig. 1a**).

Of the 112 genetically unrelated ES cell lines, 61 (54%) were of European ancestry (excluding Middle East–East European and Central-South Asia–South European), 31 (28%) of Asian ancestry, 3 (3%) of African ancestry, 12 (10%) of Middle East and East European ancestry, and 4 (4%) of Central-South Asian and South European ancestry (**Table 1**). The European ES cell lines were further stratified using a recently described comprehensive European reference set[45] and were found to match subpopulations from many different regions of Europe (**Fig. 1b**). The cell lines of Asian descent were stratified into those of East Asian origin, including those of Han Chinese, Korean, Japanese and Indian origin, and those of Central or Central-South Asian origin (**Fig. 1c,d**). Five of the cell lines classified as Middle East and East European clustered with one another but not particularly close to any of the reference samples used in this study, namely clusters belonging to HGDP-Central/South-Asia, HGDP-Middle East and the HGDP-European samples (**Fig. 1d**). Four of these five lines were derived in Iran, and are most likely of Persian ancestry, a population not represented in the reference samples. It is notable that the nine ES cell lines most commonly cited in the scientific literature are representative of the genetic backgrounds of populations from northern, northwestern and central European, Han Chinese, Indian and Middle Eastern populations (**Table 1**).

Karyotype analysis

Stability of the cell lines. Analyses were carried out on all 120 human ES cell lines (including duplicate and sibling cell lines) provided for karyotyping at both early- and late-passage levels ('paired' lines), as well as on five additional lines that were provided only in early passage (**Supplementary Table 1**). Among this total of 125 lines, 42 (34%) had abnormal karyotypes (defined as at least two metaphases with identical, abnormal karyotypes of at least 30 metaphases screened) in at least one passage level. The data from this study confirm that human ES cells are commonly diploid soon after derivation, and that many do retain a normal karyotype after many passages (**Fig. 2a**).

Late-passage cultures of the paired lines were approximately twice as likely to have a chromosome abnormality (39/120, 33%) as those from the early-passage cultures (17/120, 14%). Among the five lines submitted only at an early-passage level, one (20%) had an abnormal karyotype with an extra copy of chromosome 17q. Of the 39 paired lines with abnormal karyotypes at late passage, 24 were normal at the early passage, whereas the remaining 15 also had abnormalities at both passage levels. In this case, the abnormalities seen at the late passage were mostly similar to those seen at the early passage. About half of all the abnormalities involved combinations of chromosomes 1, 12, 17 and 20 (**Fig. 2a**).

A number of cultures were mosaic with, mostly, two populations of cells, one with a normal karyotype and one with a particular abnormal karyotype; 10 of 24 with abnormalities only at late passage, and 8 out of 15 with abnormalities at both passage levels were mosaic (**Supplementary Table 1**). Five lines that were mosaic at early passage showed an increase in the abnormal cell population at late passage. In all of these cases, the abnormality involved extra copies of chromosomes 1, 12, 17, 20 or X. One pair showed additional chromosome changes in the late passage and one pair had unrelated abnormal karyotypes at each passage level. Two lines were scored as abnormal in early passage but normal at late passage. However, both were mosaic, with 3/30 metaphases in one case with a translocated chromosome t(2:19), and 5/30 metaphases in the other with a duplication on chromosome 13. Both chromosomal rearrangements were unique to these lines and most likely represent random changes that were outcompeted by the normal cells over time.

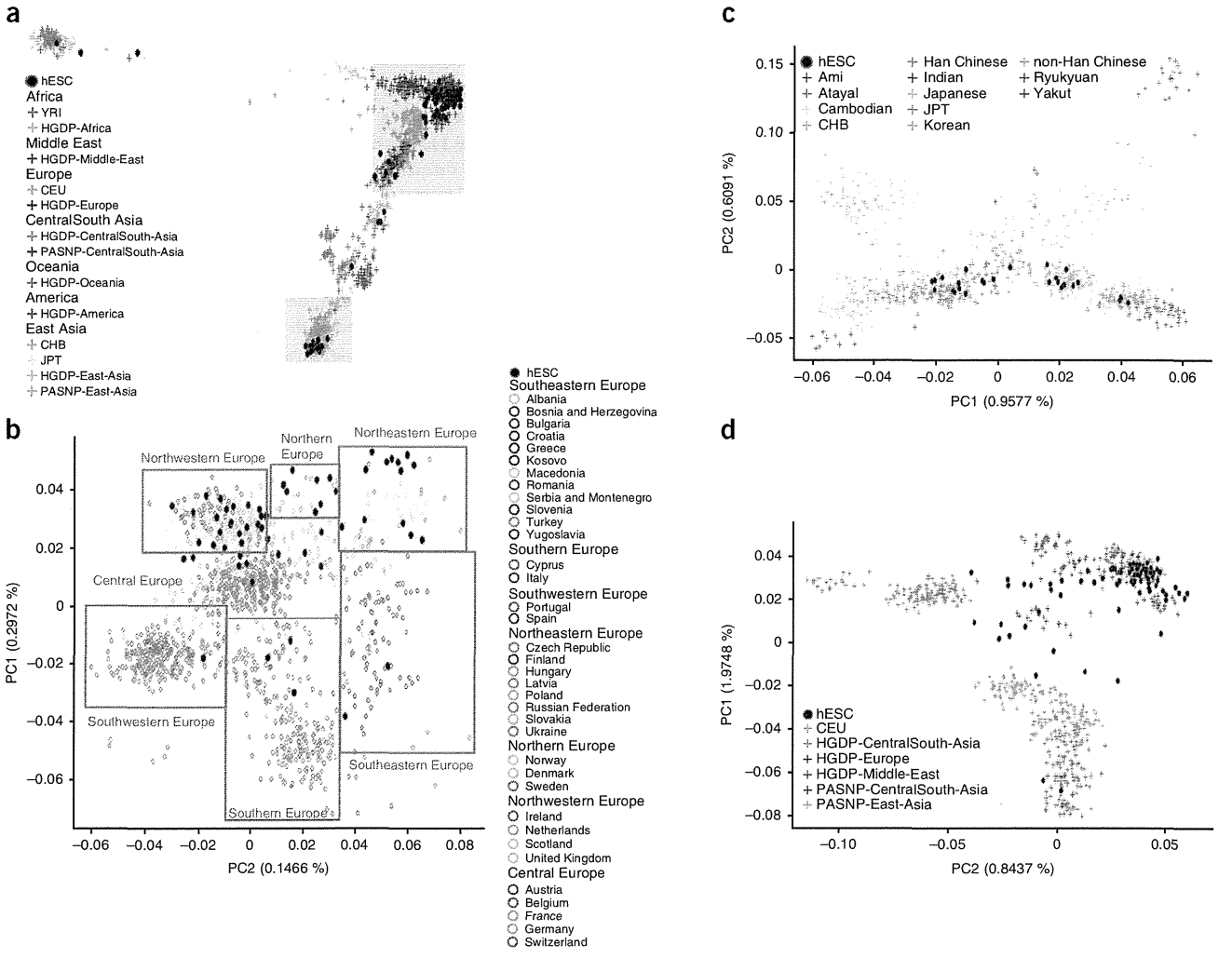


Figure 1 Population structure of the human ES cell lines analyzed. Principal component (PC) analyses were conducted on the entire final merged data set. PC1 and PC2 are plotted on the *y* and *x* axes, respectively. (**a**) The overall distribution of the human ES cell lines studied compared to the major ethnic groups identified in the HapMap study[41], the human genome diversity panel (HGDP)[42] and the Pan-Asian SNP Initiative[43]. (**b**–**d**) The cell lines were further subdivided to show their relationships to European (**b**), East Asian and Indian (**c**) and Middle East-European–Central South Asian populations (**d**).

Among the 11 iPS cell lines examined, three exhibited chromosome abnormalities, a frequency (27%) comparable to that found in ES cell lines. Of these, one line (RR01) exhibited trisomy 12 at both early and late passage. The other two lines were provided only at one passage level; one had a trisomy 12 (RR05) and one an inversion on chromosome 5 (RR03). None of these abnormalities were present in the somatic cells from which they were derived. These results are consistent with recent analysis of human iPS cell chromosomal instability both in the general frequency of aberrations and over-representation of chromosome 12 alterations[12,46].

A common source of cells with abnormal karyotypes. The proportion of cell lines with abnormal karyotypes did increase with delta, the difference in estimated number of population doublings ($P = 0.048$) (**Fig. 2b**). There was also a marked variation in the proportion of abnormal ES cell lines submitted by the different participating laboratories. The 42 abnormal lines were among cell lines submitted by 21 laboratories, whereas no abnormal lines were found among the other 38 lines submitted from the remaining 11 laboratories. This was not directly linked to the delta of the submitted lines and might simply reflect the stochasticity of mutation, or could imply a laboratory effect. The cell lines in each category were from diverse ethnic origins, and were cultured under very similar conditions, although a role for subtle variations in culture technique cannot be excluded. Nevertheless, consistent with suggestions that enzymatic mass-passaging techniques may favor the generation of abnormalities, a twofold higher proportion of the paired lines that had an initially normal karyotype but became abnormal at late passage were passaged by enzymatic methods (18/58, 31%), relative to those passaged by the manual cut-and-paste technique (6/43, 14%) (χ^2, $P = 0.009$). This effect is significant even after adjusting for delta ($P = 0.017$).

Candidate regions/genes. Aberrations of all chromosomes with the exception of chromosome 4 were observed (**Fig. 3**). However, most chromosomes were affected in very few instances, and four cell lines with particularly abnormal karyotypes accounted for many of these sporadic changes (**Supplementary Table 1**). In addition, there were three instances of balanced rearrangements seen as sole aberrations,

Table 1 Ethnic origin of human ES cell lines analyzed indicating ancestry of those most often cited

Ancestry	Number of cell lines[a]	Most commonly used cell lines	No. citations (2008 to 2009)[b]
European	**63 (61[c])**		
Italian	4		
Southwestern European	2		
Southeastern European	2		
Northeastern European	14[d]		
Northern European	8	BG01	13
Northwestern European	24[d]	HUES7	18
Central European	11	H1	95
Asian	**33 (32[c])**		
Central Asian	3		
Central-South Asian	1		
Han Chinese	14	HES2	16
		HES3	14
Japanese	3		
Korean	9		
Indian	3[d]	HES-1	6
African	**4 (3[c])**		
East African	1		
West African	3[d]		
Middle East–East European	**14[e] (12[c])**		
		H9	122
		H7	25
		HSF-6	12
Central-South Asia South European	**4**		
Total cell lines	**118 (112[c])**		

[a]The numbers of cell lines shown includes only those that passed quality control for SNP analysis. [b]UMass Stem Cell Registry (http://www.umassmed.edu/iscr/hESCusage.aspx). [c]Total number of genetically unrelated cell lines. [d]Includes two cell lines from siblings. [e]Includes three cell lines from siblings.

a translocation between 2 and 19 in an early-passage human ES cell culture, an inversion of 11 in a late-passage culture, for which the early passage was normal, and a Robertsonian translocation between chromosome 21 and 22 in both passages of one line. There were also abnormalities affecting chromosome 7 in seven ES cells, but five came from one laboratory, suggesting an unknown cause particularly associated with that group, perhaps related to their derivation of ES cells from prenatal genetic screening material. By contrast, in most abnormal lines (25/42), the changes involved one or more of chromosomes 1, 12, 17 and 20. Of the 17 lines that were abnormal in early passage, eight had abnormalities involving these chromosomes, whereas, of the 24 lines that acquired abnormalities between early and late passage, 16 lines had changes involving acquisition of one or more of these chromosomes (**Fig. 2a**). Among the gains, there were minimal amplicons affecting the telomeric region of chromosome 17 (17q25) in two lines, and another affecting 20q11.2 was apparent in another line (**Fig. 3**). Gains of only the short arm of chromosome 12 were found in three cell lines.

The large differential in frequency between gain and loss of chromosomes is remarkable. In contrast to the 39 ES cell lines that showed gains of chromosomal material in late passage, 20 lines showed losses of chromosomal material. However, only two lines exhibited chromosomal deletions that were not caused by unbalanced translocations (one, UU03, had two unrelated deletions of chromosomes 6 and 18), although even in these there were also unrelated chromosome gains. Excepting the deletions on chromosome 7, which only occurred in the lines from one laboratory, three regions showed recurrent loss, 10p13-pter (five cases), 18q21-qter (five cases) and 22q13-qter (three cases); in several cases these were the sole changes (**Fig. 3**).

Structural changes determined by molecular karyotyping

Identification of ES cell–associated structural variants. As genomic structural changes do occur below the ~5 MB detectable limit of karyotyping, we used SNP data to identify structural variants and detect structural changes down to a minimum of 1 kb in length. We identified structural variants for all samples that passed quality control, but restricted our detailed analyses to those cells judged to have a normal karyotype, because of the difficulty of ascribing functional significance to a small structural genomic change in a background of a much larger karyotypic abnormality. Nevertheless, we did examine the breakpoints in six cases of balanced rearrangements (PP-107, NN-12, J-02, CC-05, AA-03, RR-03) but found no evidence of structural variants associated with these (**Supplementary Table 1**). In addition, although loss of heterozygosity can be detected with the SNP platform, we focused our attention primarily on structural variant analysis as this is the more likely structural change to lead to a selective advantage. Nonetheless, we provide a spreadsheet of overlapping loss of heterozygosity across the 225 human ES cell samples and an associated .bed file with all loss-of-heterozygosity calls (**Supplementary Data Sets 1** and **2**). Structural variants were identified in the 200 DNA samples from karyotypically normal ES cells that passed quality control by comparison with the reference genome (hg18). Further quality controls removed one sample due to an extremely high number of structural variants called and two more for extremely high total length of structural variants (**Supplementary Fig. 1**). A total of 27,409 deletions with an average size of 40.2 kb, and 7,413 duplications with an average size of 95.4 kb, were detected. The sizes of these structural variants and the total number of differences between deletions and duplications are consistent with previous structural variant studies of human populations[47]. As structural variants are a common feature of variation between individuals, the majority of structural variants detected in the human ES cells most likely represent the condition of the genomes of the respective embryos from which they were derived, and are unrelated to human ES cell culture.

To aid in distinguishing culture-associated structural variants, we compared the human ES cell structural variants to those identified using the same platform to analyze a set of 267 HapMap samples (raw data directly supplied by Illumina). Though relatively restricted in population diversity compared to our human ES cell data set, the HapMap samples provide a set of common reference structural variants. Our subsequent analyses focused only on variant regions enriched in human ES cell lines over the HapMap samples. We identified 504 regions of gain and 860 regions of deletion in the karyotypically normal ES cell lines as 'ES cell associated' (**Supplementary Data Set 3** and **Supplementary Table 2**).

Genome-of-origin variants. The apparent ES cell–associated structural variants most likely include some rare and/or localized variants absent in the HapMap set, yet unrelated to human ES cell culture selection. There are a number of examples in which a particular variant occurs in a single line in both the early and late passage. Although we cannot exclude that such variants arose in culture before the early-passage samples being obtained, it is more likely they represent rare and/or localized variants present in the genomes of the donated embryos. We did see such a case among the iPS cell lines for which we have SNP data from the parental somatic cell line. For instance, in three iPS cell lines derived from the same parental fibroblast, the same rare gain (chr12:106,928,902-107,008,902) was detected in both the early and late passages and the parental line (**Supplementary Data Set 3**). Also, among the sibling human ES cells lines, we found recurring rare variants specific to each family. For instance, a gain at chr3:45,220,749-45,263,539 was found in the early and late passages of human ES cell lines G02 and G05, although this allele was absent in G04, the third of these sibling lines. At another

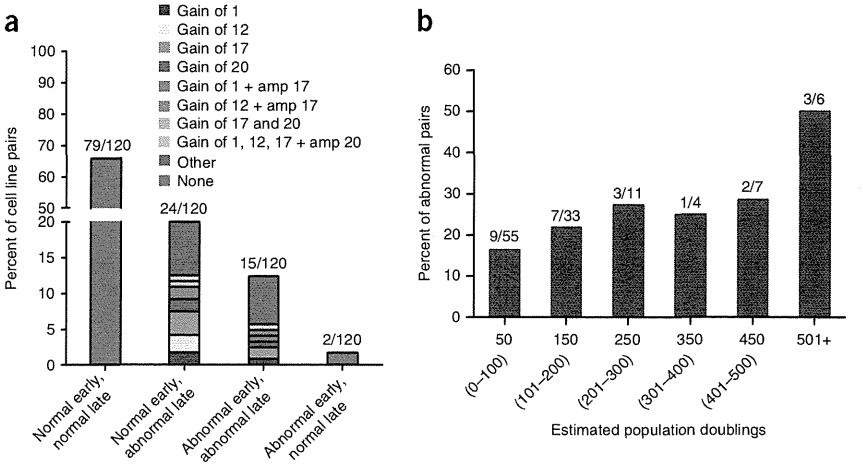


Figure 2 Cytogenetic changes occurring during prolonged passage of human ES cells. (**a**) Percentage of human ES cell line pairs that exhibited a karyotypic abnormality in either early or late passages, or both. Cell lines were excluded if they were known to be derived from karyotypically abnormal embryos. The ES cell pairs are grouped according to whether the chromosome change was observed at late passage only (normal early, abnormal late), both at early and late passages (abnormal early, abnormal late) or early passage only (abnormal early, normal late) and no chromosomal change (normal early, normal late). The percentage of cell lines that have individual gains of chromosomes 1, 12, 17 and 20, gain of chromosomes 1 and 17, or gain of chromosomes 1, 12, 17 and 20 are highlighted. Chromosome changes not involving 1, 12, 17 and 20 are indicated as 'Other'. The numbers above each bar indicate the total number of lines that fall into the four categories out of the total number of pairs of lines analyzed. Two cell lines (CO2 and CC05) in the 'abnormal early, abnormal late' category were known to be derived from karyotypically abnormal embryos (a trisomy 13 and ring chromosome 18). One abnormal cell line (AA06) has been excluded from this figure as only one passage was available for analysis. (**b**) Proportion of pairs of lines that acquired karyotypic abnormalities over different periods in culture. The pairs of lines are grouped according to 'Delta', the difference in estimated population doublings between the early and late passages. Only those lines that had a normal karyotype at the early-passage level were included in the analysis, and of those only 115 pairs could reliably be assigned an estimated population doubling time estimate.

location, chr3:167,536,633-167,837,107, a gain occurs in the early and late passage of all three of these sibling lines. For the purposes of this study, we have assumed that none of these rare variants arose during ES cell culture, and we define them as 'genome-of-origin' variants (**Supplementary Table 2**).

Dynamically changing variants. Some structural variants that were detected are represented in the HapMap population and change dynamically in ES cell culture, suggesting the labile nature of at least some genomic elements. For example, 18 human ES cell lines had a gain at chr17:75,289,455-75,296,305 (**Supplementary Table 2**, labile structural variant), but this was also present in four HapMap samples. Among the human ES cell samples, this structural variant was present in the late but not early passage of four lines, but in five other definitive cases it was present in the early but not late passage. Thus, this represents a dynamically changing variant with no evidence for positive selection in human ES cell culture but provides an example of the labile nature of the human genome.

Structural variants enriched in late-passage cultures. In the subset of structural variants enriched in the ES cells, there was no overall trend toward a gain of total structural variant numbers between early-passage and late-passage samples: that is, there was an increase in the total number in the late passage of some lines, but a decrease in others (**Supplementary Table 2**). Among the particular structural variants that did show increases in several lines in a late passage, a number encompassed regions known to encode genes that may be relevant to human ES cell behavior, but they were isolated instances. For example, a deletion variant spanning the *SOX21* locus, a gene encoding a transcription factor associated with differentiation of human ES cells, was found in one line (UU03-E), and a minimal deleted region at chr4:983425-997875, which spans the promoter and first exon of *FGFRL1*, was present in the late but not early passage of two lines (L03-l, TT20-l). FGFRL1 is expressed in human ES cells and may act as an inhibitory sink for FGF2, which is important for human ES cell maintenance[48]. Late-passage samples of both the MM01 and TT20 lines share a minimal overlapping deletion variant of ~540 bp at chr3:196,472,618-196,473,157. This spans a highly conserved open reading frame (C3orf21) that is expressed in human ES cells but has no known function[48].

Structural variants in karyotypically normal ES cells

We next analyzed structural variants in regions subject to common karyotypic abnormalities. In one region of chromosome 1q, two cell lines (V09 and FF01) in late, but not early, passage, have an overlapping 3.1 MB gain (chr1:199,985,282-203,092,388), which spans *JARID1B*, a polycomb-related gene encoding a histone H3 lysine-4-demethylase[49,50]. On chromosome 12, two cell lines (B02 and F04) have an overlapping gain of 1.1 MB in chr12:5,592,150-6,749,326 in the late-passage samples. This structural variant is within a minimal amplicon identified by karyology (12p13.31) and includes *NANOG*, *CD9* and *GAPDH*, all of which are expressed in human ES cells. There was little evidence for repeated occurrence of gains below the resolution of standard banding techniques in regions of chromosome 17 (**Supplementary Fig. 2**).

In contrast, there was a striking occurrence of a structural variant gain within chromosome 20 in 22 karyotypically normal cell lines. Notably, these gains, many validated by quantitative PCR (**Supplementary Fig. 3**), are within the minimal amplicon 20q11.2 found by karyology (**Fig. 4**). Among these 22 cell lines, there were five instances where the gain was detected in both early and late passage but 17 instances where it was detected only in the late passage. There were no instances of this gain in early passage but absence in late passage of the same cell line. This gain was also present in an ES cell line (J01) that had an abnormal karyotype at late passage and in an iPS cell line (RR01) that contained an extra copy of chromosome 12 (**Supplementary Table 1**). This strongly suggests that once genomic rearrangements occur in this region, they provide a positive selective advantage during subsequent culture. The least difference in passage number between the early and late passage from the same cell line, which had the gain in the late passage alone, was 22. The apparent strong positive selection for gain of this region suggests that a gene providing a cell-autonomous functional advantage under normal human ES cell culture conditions is encoded within the DNA of the

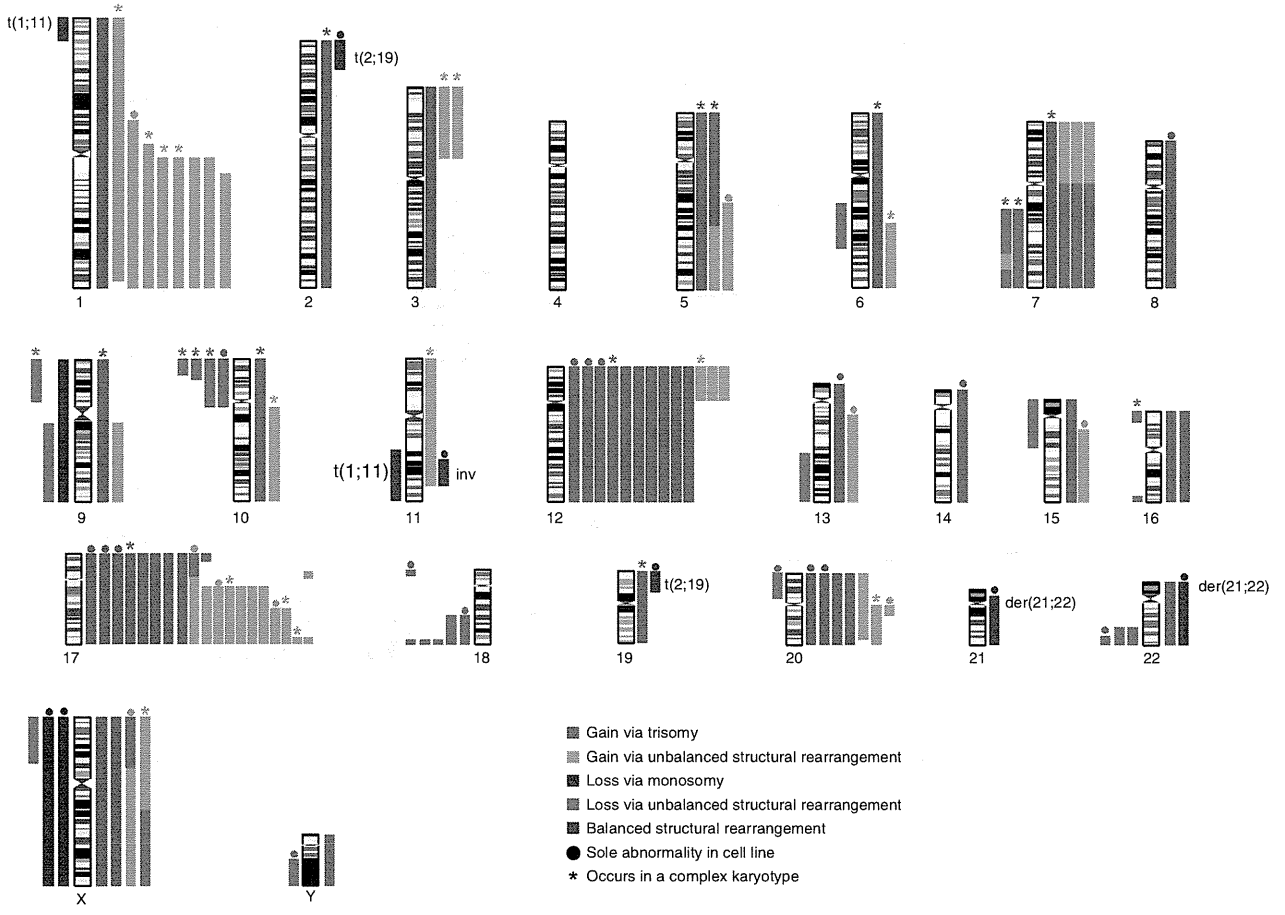


Figure 3 Ideogram demonstrating the chromosome changes found in this study. Each colored bar represents one chromosome change occurrence in one cell line. Chromosome losses and gains are shown to the left and right of the ideogram, respectively, except that those instances where a single chromosome rearrangement results in a gain and a loss the colored bars are shown together for clarity. The cytogenetic changes are color coded: Maroon, loss of a whole chromosome (monopsony); red, loss via a structural chromosome rearrangement (unbalanced translocation or interstitial deletion); dark green, gain of a whole chromosome (trisomy); light green is gain via a structural chromosome rearrangement (unbalanced translocation or interstitial duplication); blue represents the occurrence of an apparently balanced rearrangement the nature of which is labeled. Instances in which a change affected only a single chromosome are denoted by ●, whereas changes associated with complex karyotypes (>5 unrelated chromosome aberrations) are denoted by ★. Two cell lines (C02 and CC05) were known to be derived from karyotypically abnormal embryos and contain a trisomy 13 and ring chromosome 18 respectively. iPS cell lines are excluded from this figure. Based upon these studies the minimal critical chromosomal regions subject to gain in culture adapted human ES cell lines were 1q21-qter, 12p11-pter, 17q21.3-qter and 20q11.2. The minimal regions subject to loss were 10p13-pter, 18q21-qter and 22q13-qter.

shared overlapping region. Moreover, three cell lines (F-01, Q-02 and K-05) that had a normal karyotype and a 20q11.21 structural variant gain in early passage acquired an abnormal banded karyotype in samples from later passage. The late-passage abnormal karyotypes of F-01, Q-02 and K-05 were 46,XX,der(15)t(15;17)(p11;q21); (47,XX,+der(1)(t(1;1)(p?21.2;q11); and 47,XX,t(1;11)(p?36;q13),trp(17)(p11.2),+20, respectively. This preliminary evidence suggests that early gains in 20q11.21 might promote further subsequent genetic change.

The duplicated regions of chromosome 20 in the various cell lines and the minimal amplicon are diagrammed in **Figure 4b**. The proximal ends of each of the structural variant gains within chromosome 20 are presumed to lie in a nonbridged sequencing gap sized at 1 MB near the centromeric region of the long arm. The most proximal SNP identified in all these gains is the first occurring after this gap, at position 29,267,954. The distal end of the gain varies across the lines. The longest gain extends to 31,793,485 with a measured length of 2.5 MB, similar to the shortest karyotypically identified gain in this region, dup(20)q11.21 in cell line UU01 (**Fig. 3**). The shortest gain is 0.55 MB extending to 29,821,940 and contains 13 genes (**Fig. 4c**). Three of these genes, *ID1*, *BCL2L1* and *HM13*, are known to be expressed in human ES cells based on mRNA-Seq data (**Fig. 4c**) and published microarray data[27]. Although a single RefSeq-annotated microRNA lies in this region, there is no evidence for its expression in human ES cells[51]. Further, combined with the mRNA-Seq data, ChIP-Seq data from H1 human ES cells of histone modifications, considered universal predictors of enhancer and promoter activity (H3K4me3, H3K27ac), do not suggest additional functional regions other than those associated with the three RefSeq genes identified as expressed in human ES cells (**Fig. 4c**).

When four pairs of cell lines with and without the chromosome 20 gain were analyzed, there was no clear correlation between increased expression and the presence of the 20q11.21 gain for these three expressed genes (**Fig. 4d**). Nevertheless, preliminary results indicated a strong selective advantage in culture for cells with the gain