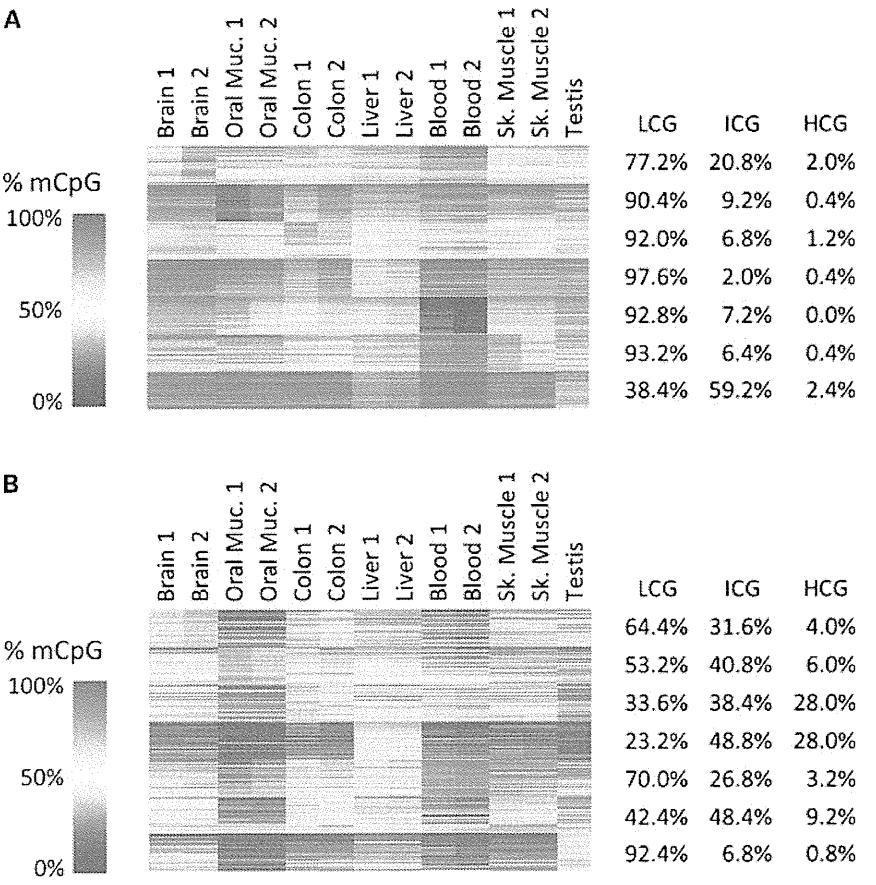

Figure 2. Methylation panel for tissue-specific differential methylation across seven human normal tissues. Left panels indicate the methylation levels of probe sets selected as tissue-specific hypomethylation (**A**) and as tissue-specific hypermethylation (**B**) among seven human tissues. Each row represents a CpG locus (250 for each tissue) and each column represents a tissue sample. The color scale bar at the left side shows the percentage of the methylation level (0–100%). The percentages of the LCG, ICG and HCG in a given probe set are represented at the right side.

specific manner. In contrast, tissue-specific hypermethylated genes are suppressed among all tissues. These results indicate that hypomethylation in the CpG-poor promoters identified here underlie tissue-specific expression in a given cell type.

Local epigenetic modification and recruitment of transcription factors are a fundamental part of the system for appropriate transcriptional regulation (21). We performed enrichment analysis of 746 recognition motifs for transcription factors to examine the relationship between cis-regulatory elements of promoters and tissue-specific hypomethylation. As shown in Figure 4, some matrices are significantly enriched (Z-score $>$ 8.0) in the hypomethylated regions. In oral-mucosa-specific hypomethylated regions, the binding motifs of p53 family genes are highly enriched. p63, the master regulator of keratinocyte differentiation, has similar DNA-binding domains to p53 and half of p63-bound regions in the squamous cell carcinoma cell line have p53 consensus motifs (22). In liver-specific hypomethylated regions, the matrices for the C4 zinc finger domain of the PPAR family (PPARA, PPARG and RXRs) and the NR2F family (HNF4A) are enriched compared with the background sequences. In the blood set, the matrices for the ETS domain of ETS factors (ETS1, ETS2, ELF2, ELK1) and the Runt domain of AML factors (RUNX1) are enriched. Similarly, MyoD-binding motifs are enriched in skeletal muscle-specific hypomethylated regions. In contrast, we could not find significant enrichment of transcription factor-binding motifs in tissue-specific hypermethylated regions (data not shown). Although the molecular mechanism of *de novo* hypermethylation and hypomethylation remains unknown, it is suggested that selective binding of transcription factors are at least significantly associated with regional hypomethylation during terminal differentiation.

Dynamic changes of CpG-poor promoter methylation during *in vitro* differentiation and cellular reprogramming

Although tissue-specific hypomethylation in CpG-poor promoters are closely related to gene specification for the tissue phenotype, when and how these variable methylation statuses are established remain unknown. To elucidate the methylation changes during cellular differentiation, we performed clustering analysis of human somatic tissue and normal cells

Table 1. GO analysis of tissue-specific hypomethylation

Tissue	GO term (biological process)	No. of genes	*P*-value
Brain	Nervous system development	28	4.22E − 05
	Multicellular organismal development	50	4.33E − 04
	Developmental process	53	4.64E − 04
Oral mucosa	Ectoderm development	20	4.89E − 14
	Epidermis development	18	2.19E − 12
	Tissue development	24	2.72E − 07
Colon	Defense response	31	3.68E − 11
	Response to stress	49	3.42E − 09
	Response to stimulus	71	7.31E − 07
Liver	Acute inflammatory response	20	9.00E − 19
	Response to wounding	34	7.22E − 16
	Response to external stimulus	43	3.72E − 15
Blood	Immune system process	55	6.86E − 24
	Immune response	42	2.53E − 19
	Defense response	33	2.62E − 13
Skeletal muscle	Muscle contraction	19	6.02E − 14
	Muscle system process	19	2.70E − 13
	Striated muscle contraction	9	2.64E − 08
Testis	Reproductive process in a multicellular organism	16	1.82E − 04
	Multicellular organism reproduction	16	1.82E − 04
	Gamete generation	14	2.40E − 04

including human ES cells, iPS cells and primary fibroblast cells using tissue-specific hypomethylation sites (Fig. 5). The heatmap shows distinct methylation patterns between the pluripotent cells and somatic tissues composed of the terminally differentiated cells. Seven human ES cell lines and two iPS cell lines show similar methylation patterns. Intriguingly, most genes representing specific hypomethylation in differentiated cells are densely methylated in both ES cells and iPS cells, raising the possibility that the default state of low CpG promoters in the embryonic stage is totally methylated and erasure of methylation may occur during terminal differentiation in a cell-type-specific manner.

We next compared the methylation status of the adult human liver and the fetal liver. Liver-specific hypomethylated genes are heavily methylated in KhES3, a human ES cell line, but are hypomethylated in the adult liver tissue (Fig. 6B and C). In the fetal liver, the methylation level of these genes shows a mild decrease in these regions. Bisulfite sequencing also revealed the partial hypomethylation of *ITIH3* and *APOA1* promoters (Supplementary Material, Fig S8A and B).

To further analyze the demethylation dynamics during hepatic differentiation, we analyzed methylation during *in vitro* differentiation toward hepatic lineages (23). On day 7, the cells began to express an endoderm marker, *SOX17* (Fig. 6A). *AFP* expression was detected on day 13 and *ALB* expression was detected on day 21. The methylation status of liver-specific hypomethylated genes showed a slight decrease during hepatic differentiation (Fig. 6B and C). Indeed, bisulfite sequencing of the *APOA1* promoter region demonstrated that CpG sites in this promoter region are fully hypermethylated in KhES3 and gradually become demethylated during *in vitro* differentiation (Supplementary Material, Fig. S8B). Demethylated regions are observed only in the vicinity of *APOA1* TSSs at day 21 of differentiation, and spread over 1 kb beyond the *APOA1* TSS in adult liver tissues. Sparse non-CpG methylation is observed in KhES3 and lost at day 21 of differentiation and also in adult liver tissues. This demethylation in non-CpG sites in KhES3 is also observed in the promoter region of *CD6* in adult blood and of *STMN4* in the adult brain (Supplementary Material, Fig. S9).

We then analyzed further the methylation status over the entire *APOA1* gene locus to determine the extent of demethylation events (Fig. 6D). Demethylation starts from the vicinity of *APOA1* TSSs at day 13 and extends to 200 bp around the TSS on day 21. Hypomethylated regions in human liver tissues spread over the *APOA1* region, from TSSs to the CpG island of the 3′ end and the further downstream region, suggesting the correlation of extensive demethylation with the stable expression of specific gene sets and cell fate determination.

Epigenetic reprogramming using defined factors enables terminally differentiated cells to gain pluripotency (24). Re-expression of pluripotency genes associated with these promoters, which are methylated in differentiated somatic cells, is important for iPS cell generation (25). The heatmap shows that the four human primary fibroblast cell lines (IMR90, MRC-9, KMS-6 and TIG-103) share specific hypomethylation. After cellular reprogramming into iPS cells, the IMR90 cells show restoration of methylation in these fibroblast-specific hypomethylated sites (Fig. 5). These results suggest that regaining promoter methylation in tissue-specific hypomethylated genes, as well as erasure of methylation in pluripotency genes, is important for this process.

DISCUSSION

In this study, we analyzed inclusive gene sets for tissue-specific hypomethylation and hypermethylation among human normal tissues. Of note, the former gene subsets are remarkably associated with cellular functions characterizing the tissue phenotypes. Although we have examined the limited sites of promoter regions, we reveal here that these hypomethylated genes display tissue-specific patterns of gene expression and specific enrichment of transcription factor recognition motifs in their promoters. This indicates the methylation changes in these regulatory regions might have functional roles in spatiotemporal transcriptional control. Furthermore, the hypomethylation panel showed an unexpected dense methylation pattern in pluripotent stem cells and regional hypomethylation in differentiated cells, suggesting this type of tDMRs might be a consequence of methylation erasure or a dilution process.

To date, the exploration of tDMRs was performed on the premise that stepwise addition of promoter methylation contributes to cell fate determination during early embryogenesis (26,27). It has been widely accepted that the genomic DNA of the embryo, which has pluripotency to differentiate into multiple lineages, is initially unmethylated and subsequent accumulations of hypermethylation in CpG island promoters are important for lineage restriction by reinforcing transcriptional repression of the unnecessary genes (28). Although

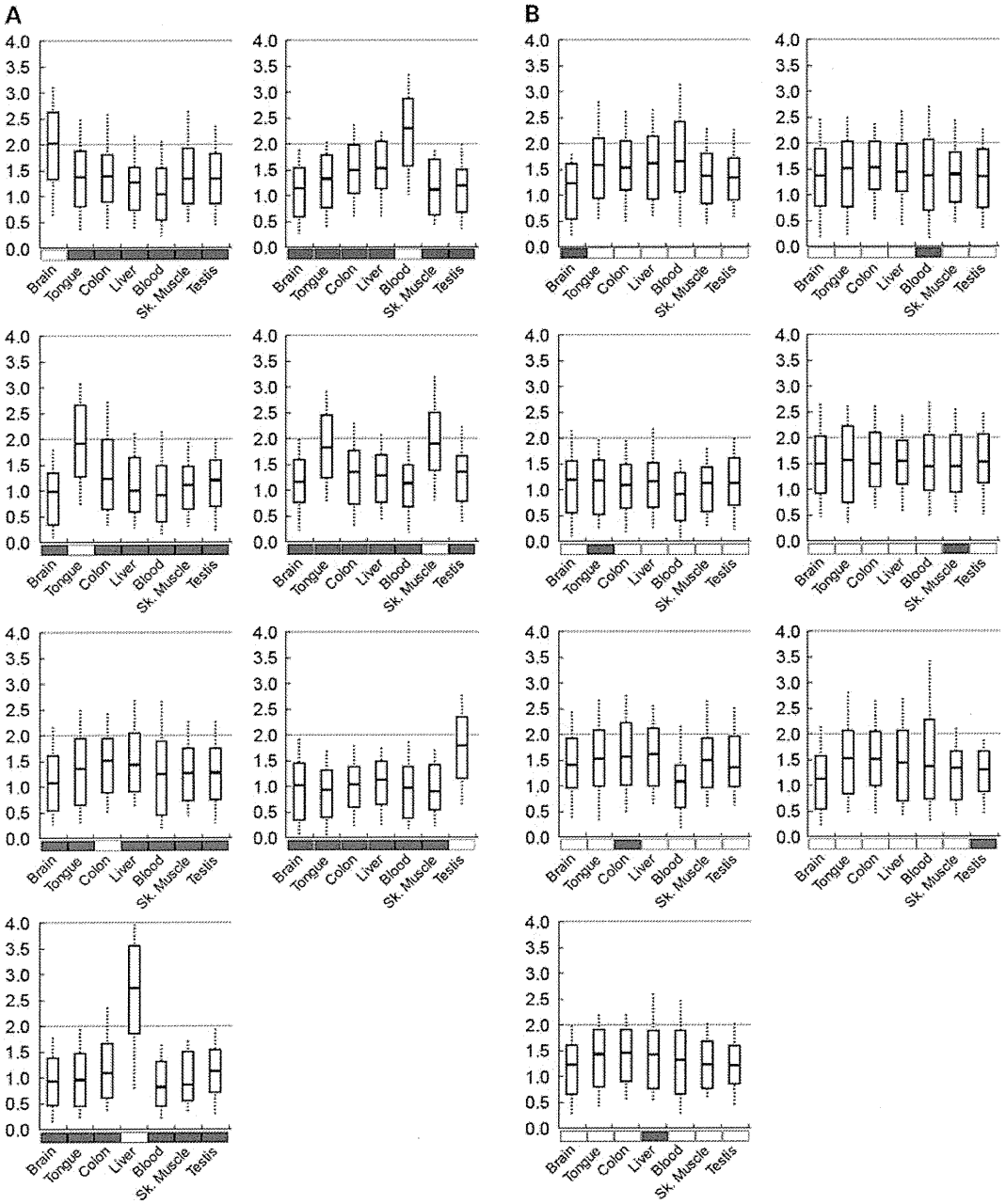

Figure 3. The gene expression level of tissue-specific differentially methylated genes. Shown box plots (from 25th percentile to the 75th percentile with heavy lines at the median) represent average gene expression levels (the log scale of the GeneChip score) of tissue-specific hypomethylated genes (**A**) and tissue-specific hypermethylated genes (**B**), for each tissue. The dotted lines extend above and below the box to show the first and ninth deciles. Black and white boxes below the bar graphs represent hypermethylation and hypomethylation of the given tissue, respectively.

this concept was true for some validated examples, it cannot adequately explain the global control of gene expression. In fact, consistent with the previous studies (6,10), we observed that most CpG island promoters are invariably unmethylated among normal tissues. In contrast with tissue-specific hypermethylation in CpG island promoters, tissue-specific hypomethylation in CpG-poor promoters has been underestimated so far and is significantly associated with the tissue phenotype.

These observations raise a new question about the molecular mechanism of tissue-specific hypomethylation established during terminal differentiation. Promoter demethylation in the differentiated cells is an old concept (29,30), but it has been forgotten while mammalian DNA demethylase was yet to be discovered. Now, two types of mechanisms for DNA demethylation, namely active demethylation and passive demethylation, are widely accepted for mammals (31,32).

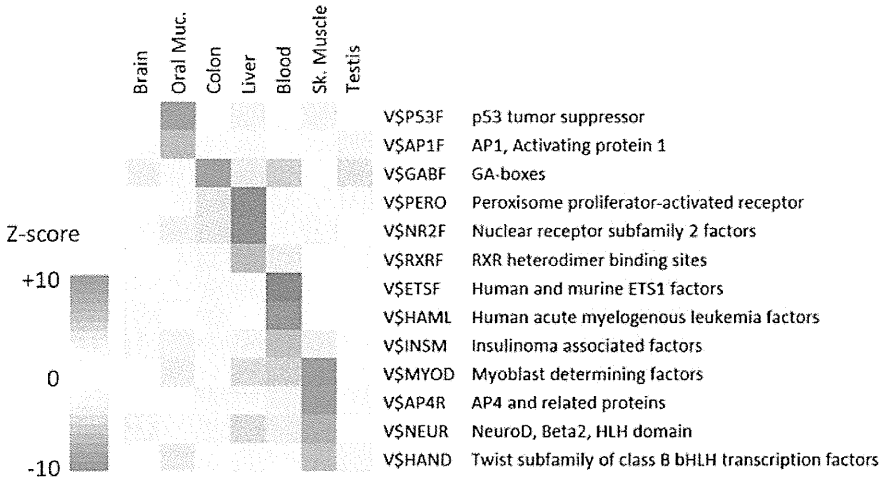

Figure 4. Enrichment of transcription factor recognition motifs in the tissue-specific hypomethylated regions. Each row represents a cis-regulatory module family with significant over-representation relative to a random set of mammalian promoters (Z-score > 8.0). Each column represents a tissue type. Four tissues (oral mucosa, liver, blood and skeletal muscle) show some specific enrichment of their master regulators binding motifs, respectively.

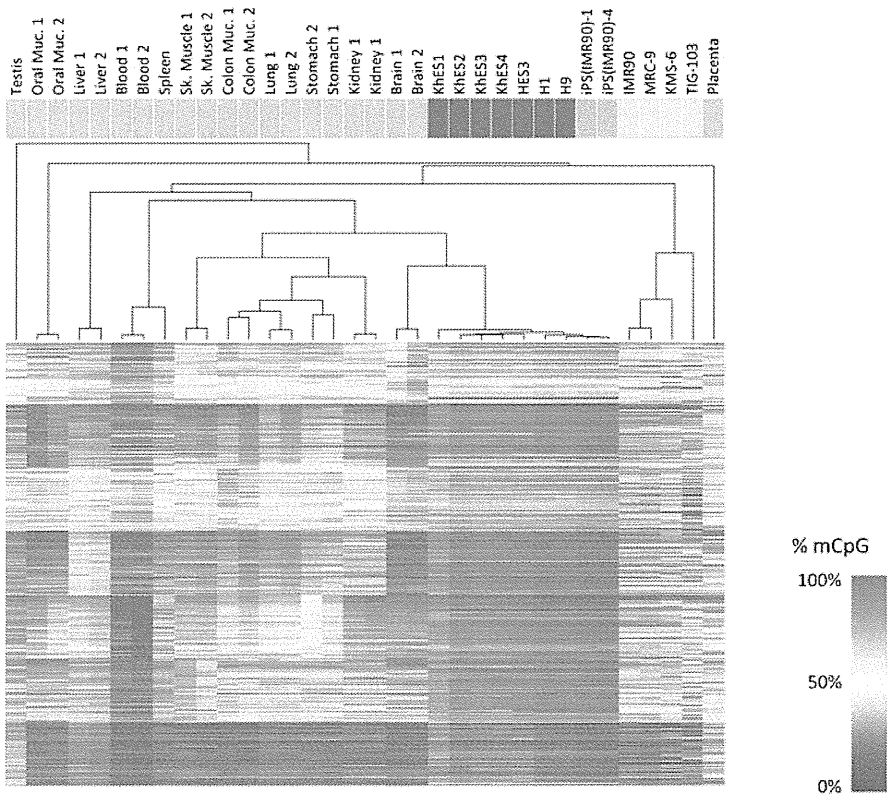

Figure 5. Hierarchical clustering analysis of human somatic tissues and normal cells. The dendrogram in the upper panel was obtained on the basis of the representative gene sets of tissue-specific hypomethylation using average linkage correlation. Each row represents a CpG locus (250 tissue-specific hypomethylation for each) and each column represents a sample. The colored boxes above the dendrogram indicate the nature of the samples; human somatic tissues (blue), human ES cells (red), human iPS cells (orange) and human primary fibroblast (green). The color scale bar at the right side shows the percentage of the methylation level (0–100%).

Active demethylation is observed in the paternal genome of an embryo during the first few days (33,34). In this process, demethylation occurs globally except for the limited foci such as imprinting control regions and centromeric and pericentromeric heterochromatin (35). Although recent reports suggested the ten-eleven translocation (TET) family proteins,

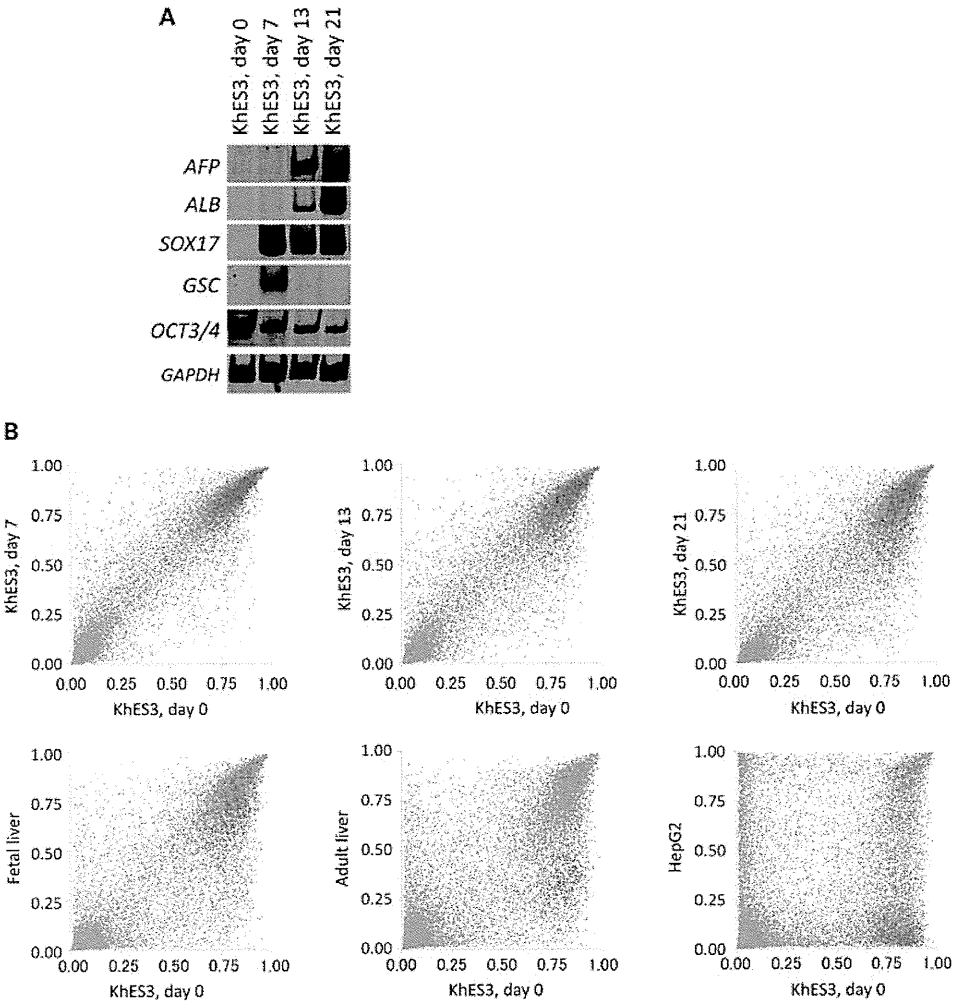

Figure 6. *In vitro* demethylation of liver-specific hypomethylated genes during hepatic differentiation (**A**) RT–PCR analysis of endodermal and hepatic differentiation markers in ES cells and differentiated cells (**B**) Global comparison among undifferentiated ES cells and differentiated cells, human fetal liver, adult liver and HepG2 cells. Liver-specific hypomethylated genes are indicated as red dots, overlapping with the others (blue). (**C**) Examples of gradually demethylated genes during *in vitro* differentiation into hepatic lineages. The bar graphs show the methylation levels of the genes that show gradual demethylation (~20% decrease) in day 21 of *in vitro* differentiation. (**D**) The liver-specific hypomethylated region around the *APOA1* gene. In the upper panel of the UCSC browser, nine black boxes indicate the position of PCR amplicons in a MassARRAY analysis. The methylation levels around the *APOA1* gene among ES cells and adult liver tissues are shown in the lower panel.

TET1, TET2 and TET3, are candidate proteins responsible for the erasure process through an oxidative demethylation pathway (32,36), further investigations are needed. The unexpected dynamics of DNA methylation during cellular differentiation might give us an important clue to elucidate the mechanism of cell fate determination during embryogenesis.

An alternative explanation for the tissue-specific demethylation seen in CpG-poor promoters is passive demethylation, which is usually observed in asymmetric cell division or highly proliferating cells like cancer cells. Inhibiting maintenance of cytosine methylation of the template strand could result in dilution of methylation in differentiated daughter cells. According to this scenario, transcription factor-related inhibition of DNA methyltransferase at the timing of cell division might be necessary because the developmental hypomethylation we observed here occurs not in a genome-wide manner but in a regional manner. Indeed, the enrichment of transcription factor-binding motifs is seen at the demethylated regions in a tissue-specific manner. Recently, it was shown that mitotically retained transcription factors are associated with the asymmetric cell division in some contexts (37,38). If sustained binding of transcription factors inhibits propagation of DNA methylation into the newly synthesized strand, transcription factor-driven demethylation will be inherited in proliferating cells. In our study, we examined *in vitro* differentiation in a series of promoters and found that a wave of demethylation develops from the TSS of *APOA1* and *ITIH3* promoters. Once the binding of transcription factors at demethylated regions induces gene expression in the tissue progenitor cells, sustained induction in response to appropriate extrinsic stimuli may result in loss of propagation of DNA methylation marks in the promoter regions for

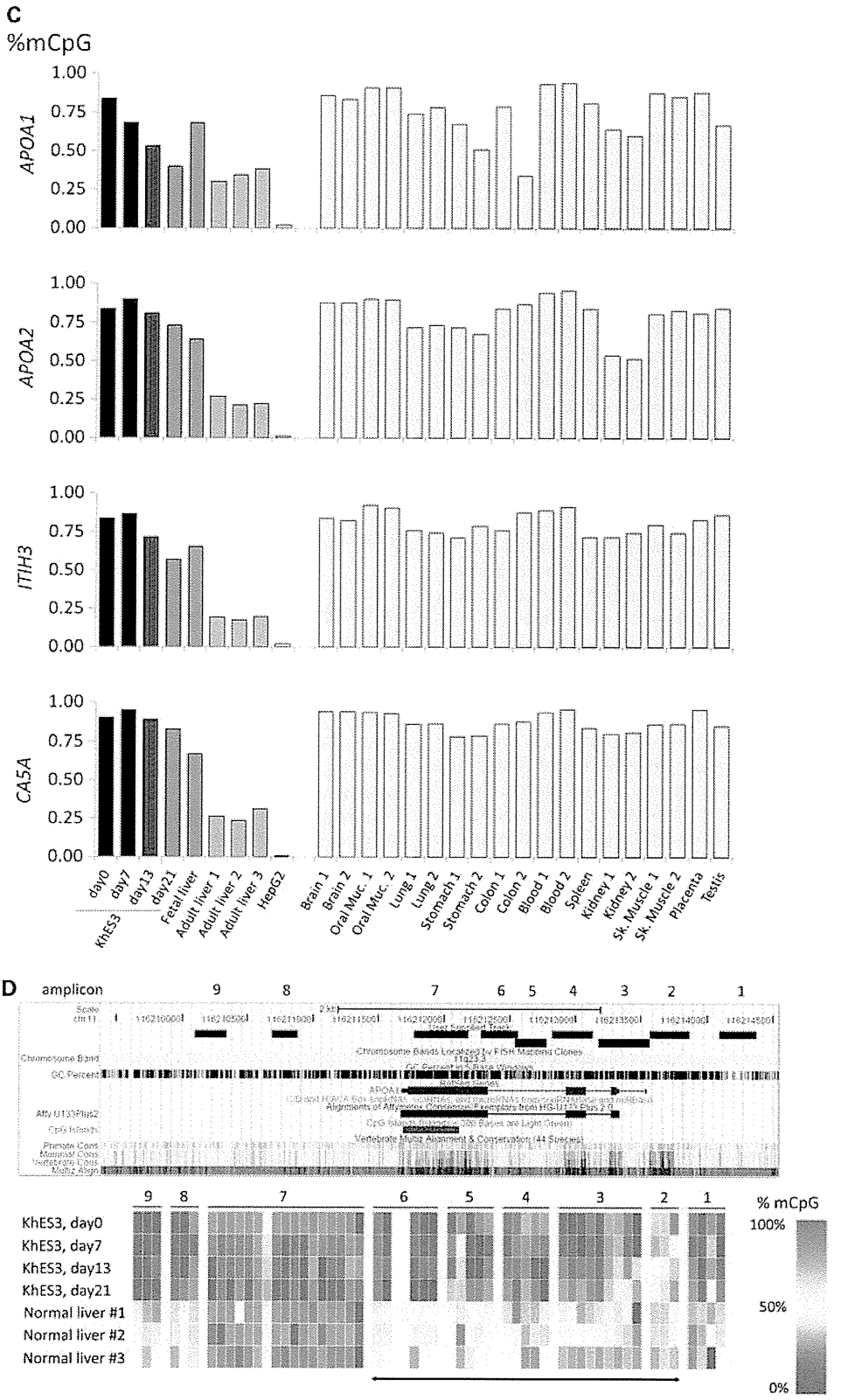

Figure 6. (Continued).

long-lasting maintenance of a transcriptionally active state. Subsequently, in this model, chromatin conformation changes in terminally differentiated cells would expand the demethylated regions and contribute to the establishment of stable and highly efficient expression of specific gene subsets.

Growing evidence suggests that forced induction of master regulator genes has the potential to change the fate of lineage-restricted cells, even in terminally differentiated cells (39–41). We identified restoration of methylation during reprogramming into iPS cells. The feasibility of cell reprogramming suggests that differentiated cells still have much more plasticity in the epigenetic status including DNA methylation than we had expected. Further analysis of methylation changes might provide novel insight into mechanisms that will generate a transcriptional repertoire for variable cell lineages and give us useful clues to control cell fate fixation, which might be applicable for regenerative medicine.

MATERIALS AND METHODS

Genomic DNA from human normal tissues

Frozen tissues of the brain, lung, liver and kidney were obtained from surgical specimens. Patients undergoing surgical resection at the Tokyo University General Hospital provided tissue after obtaining informed consent. Buccal swabs of oral mucosa, peripheral blood and placental tissue were from healthy volunteers. This study was certified by the Ethics Committee of Tokyo University. Genomic DNA from these clinical samples was extracted using the QIAamp DNA Mini Kit (QIAGEN). Genomic DNA of further individuals was purchased from BioChain (details are listed in Supplementary Material, Table S1). For the methylation-negative control, totally unmethylated genomic DNA was synthesized by a whole-genome amplification system, GenomiPhi (GE healthcare). For a positive control, fully methylated genomic DNA was generated by *Sss*.I CpG methylase (New England Biolabs) treatment of lymphocyte DNA.

Human ES cell lines

Human ES cell lines, KhES1, KhES2, KhES3, KhES4, were established and maintained as described previously (42). Human ES cell lines (H1, H9) and human iPS cell lines [iPS(IMR90)-1 and iPS(IMR90)-4] were obtained from WiCell Research Institute. HES3 cell line was obtained from ES Cell International.

Briefly, undifferentiated human ES cells were maintained on a feeder layer of MEF in DMEM/F12 (Sigma) supplemented with 20% KSR, l-Glu, NEAA and β-ME under 3% CO_2. To passage ES cells, ES cell colonies were detached from the feeder layer by treatment with 0.25% trypsin and 0.1 mg/ml of collagenase IV in PBS containing 20% KSR and 1 mm of $CaCl_2$ at 37°C for 5 min, followed by the addition of culture medium. ES cell clumps were disaggregated into smaller pieces by gentle pipetting.

An *in vitro* differentiation experiment was performed following the reported method, with some modification (43). Briefly, KhES3 cells were cultured in differentiation medium [RPMI supplemented with human recombinant activin A (100 ng/ml) and defined FBS]. FBS concentrations were 0% for the first 24 h, 0.2% for the second 48 h and 2.0% for subsequent days of differentiation. Media were replaced every 2 days with fresh differentiation medium supplemented with growth factors. ES cells were cultured in differentiation medium (DMEM supplemented with 10% KSR, Dex and HGF) for up to 30 days.

Methylation profiling

Methylation status was analyzed using HumanMethylation27 BeadChip (Illumina). Genomic DNA for methylation profiling was quantified using the Quant-iT dsDNA BR Assay Kit (Invitrogen). Five hundred nanograms of genomic DNA was bisulfite-converted using an EZ DNA Methylation Kit (Zymo Research). The converted DNA was amplified, fragmented and hybridized to a BeadChip according to the manufacturer's instructions. The raw signal intensity for both methylated (M) and unmethylated (U) DNA was measured using a BeadArray Scanner (Illumina). The methylation level of the each individual CpG is obtained using the formula (M)/(M)+(U)+100 by the GenomeStudio (Illumina).

Quantitative methylation analysis using the MassARRAY system

Bisulfite treatment of genomic DNA was performed using an EZ Methylation Kit (Zymo Research). Primer sequences are given in Supplementary Material, Table S4. This system utilizes MALDI-TOF mass spectrometry in combination with RNA base-specific cleavage (MassCLEAVE). A detectable pattern is analyzed for the methylation status. Mass spectra were acquired using a MassARRAY Compact MADLI-TOF (Sequenom) and spectra's methylation ratios were generated using Epityper software v1.0 (Sequenom).

Bisulfite sequencing

Bisulfite sequencing analysis was performed as described previously (44). Bisulfite treatment of genomic DNA was performed using an EZ Methylation Kit (Zymo Research). All primer sequences and melting temperatures for the polymerase chain reaction (PCR) are given in Supplementary Material, Table S4. PCR amplicons were subcloned into the pGEM-T vector (Promega). Clones were sequenced using PRISM3100 Sequencer (Applied Biosystems).

RNA extraction and gene expression microarray analysis

Genome-wide analysis of mRNA expression levels using U133plus2.0 human expression array® (Affymetrix) was done essentially as described previously (45). Briefly, total RNA was isolated using TRIzol reagent (Invitrogen), according to the manufacturer's instructions. One microgram of RNA was used for the generation of double-stranded cDNA with the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen) according to the manufacturer's protocol. Double-stranded cDNAs were hybridized to the microarray.

Reverse transcription–polymerase chain reaction analysis

RNA extraction and reverse transcription–polymerase chain reaction (RT–PCR) were done as described (46). Total RNA was extracted using TRI Reagent (Sigma-Aldrich) or the RNeasy micro-kit (Qiagen) and then treated with DNase (Sigma-Aldrich). Three micrograms of RNA was reverse-transcribed using Moloney Murine Leukemia Virus reverse transcriptase (Toyobo, Japan) and oligo(dT) primers (Toyobo). The primer sequences are shown in Supplementary Material, Table S4. The PCR conditions for each cycle were as follows: denaturation at 96°C for 30 s, annealing at 60°C for 2 s and extension at 72°C for 45 s. RT–PCR products were separated by 5% non-denaturing polyacrylamide gel electrophoresis, stained with SYBR Green I (Molecular Probes), and visualized using a Gel Logic 200 Imaging System (Kodak).

Definition of probe classes and promoter classes

We classified 27 578 probes into three categories: HCG, ICG and LCG. Each probe position was defined with respect to the position of a given CpG site. We determined the GC content and the ratio of observed versus expected CpG dinucleotides in a surrounding 500 bp window. The CpG ratio was calculated using the following formula: (number of CpGs × number of bp) (number of Cs × number of Gs). Three categories of probes were determined as follows: (i) HCGs (8098 probes) covering a 500 bp area with a CpG ratio above 0.75 and GC content above 55%; (ii) LCGs (8374 probes) excluded from a 500 bp area with a CpG ratio above 0.48; and (iii) ICGs (11 106 probes) that could not be categorized as either HCGs or LCGs.

Clustering analysis

To analyze the similarity of the methylation levels among human somatic tissues, ES cells and iPS cells, we used the data set of tissue-specific hypomethylation selected in Figure 2A for the cluster analysis. We applied a hierarchical clustering algorithm using the uncentered correlation coefficient as the measure of similarity and average linkage clustering (47) and visualized the dendrogram and the heatmap using TreeView (48).

GO functional annotation analysis

GO functional annotations for differentially hypomethylated and hypermethylated gene sets were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatic Resources v6.7 (http://niaid.abcc.ncifcrf.gov/ home.jsp). The lists of 250 gene symbols that show specific hypermethylation or hypermethylation for each tissue were submitted and DAVID default population background (*Homo sapiens*) was chosen to detect significantly over-represented GO biological processes (GOTERM BP-FAT). *P*-values were calculated by a modified Fisher's exact test and adjusted for multiple hypotheses testing using Bonferroni correction. The three GO terms with the most significant *P*-value and the number of genes involved in the term were listed for each tissue.

Enrichment analysis of transcription factor-binding motifs

To determine over-represented transcription factor-binding sites in tissue-specific hypomethylated and hypermethylated regions, sequences around the probe within a 500 bp window were screened for the presence of binding sites using Genomatix RegionMiner (http://www.genomatix.de, matrix library version 7.1). The number of binding site motifs was determined and over-representation over the background of random mammalian promoter sequences was calculated as the *Z*-score. Transcription factor families with a Z-score greater than 8.0 were considered highly significant. The *Z*-scores of these representative TF modules are visualized in the heatmap using TreeView (48).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We are grateful to Hiroko Meguro for microarray experiment, Kaoru Nakano for MassARRAY analysis, Elodie Lebredonchel for bisulfite sequencing experiment and Michael Jones for critical reading of the manuscript.

Conflict of Interest statement. None declared.

FUNDING

This work was mainly supported by a Grant-in-Aid for Scientific Research (S) 20221009 (H.A.) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, and the Program of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (NIBIO), Japan.

REFERENCES

1. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
2. Bernstein, B.E., Meissner, A. and Lander, E.S. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.
3. Li, E., Bestor, T.H. and Jaenisch, R. (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.
4. Okano, M., Bell, D.W., Haber, D.A. and Li, E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.
5. Jackson-Grusby, L., Beard, C., Possemato, R., Tudor, M., Fambrough, D., Csankovszki, G., Dausman, J., Lee, P., Wilson, C., Lander, E. *et al.* (2001) Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nat. Genet.*, **27**, 31–39.
6. Rakyan, V.K., Down, T.A., Thorne, N.P., Flicek, P., Kulesha, E., Graf, S., Tomazou, E.M., Backdahl, L., Johnson, N., Herberth, M. *et al.* (2008) An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.*, **18**, 1518–1529.
7. Khulan, B., Thompson, R.F., Ye, K., Fazzari, M.J., Suzuki, M., Stasiek, E., Figueroa, M.E., Glass, J.L., Chen, Q., Montagna, C. *et al.* (2006)

Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res.*, **16**, 1046–1055.
8. Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R.A. and Issa, J.P. (2007) Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet.*, **3**, 2023–2036.
9. Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z. and Cedar, H. (2009) Developmental programming of CpG island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.*, **16**, 564–571.
10. Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378–1385.
11. Illingworth, R., Kerr, A., DeSousa, D., Jorgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J. *et al.* (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.*, **6**, e22.
12. Laird, P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
13. Waterland, R.A., Kellermayer, R., Rached, M.T., Tatevian, N., Gomes, M.V., Zhang, J., Zhang, L., Chakravarty, A., Zhu, W., Laritsky, E. *et al.* (2009) Epigenomic profiling indicates a role for DNA methylation in early postnatal liver development. *Hum. Mol. Genet.*, **18**, 3026–3038.
14. Saxonov, S., Berg, P. and Brutlag, D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA*, **103**, 1412–1417.
15. Irizarry, R.A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S.A., Jeddeloh, J.A., Wen, B. and Feinberg, A.P. (2008) Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.*, **18**, 780–790.
16. Barrera, L.O., Li, Z., Smith, A.D., Arden, K.C., Cavenee, W.K., Zhang, M.Q., Green, R.D. and Ren, B. (2008) Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. *Genome Res.*, **18**, 46–59.
17. Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R. and Gunderson, K.L. (2009) Genome-wide DNA methylation profiling using infinium assay. *Epigenomics*, **1**, 177–200.
18. Jones, P.A. and Takai, D. (2001) The role of DNA methylation in mammalian epigenetics. *Science*, **293**, 1068–1070.
19. Walsh, C.P. and Bestor, T.H. (1999) Cytosine methylation and mammalian development. *Genes Dev.*, **13**, 26–34.
20. Baek, D., Davis, C., Ewing, B., Gordon, D. and Green, P. (2007) Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res.*, **17**, 145–155.
21. Kadonaga, J.T. (1998) Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell*, **92**, 307–313.
22. Yang, A., Zhu, Z., Kapranov, P., McKeon, F., Church, G.M., Gingeras, T.R. and Struhl, K. (2006) Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Mol. Cell*, **24**, 593–602.
23. Shiraki, N., Umeda, K., Sakashita, N., Takeya, M., Kume, K. and Kume, S. (2008) Differentiation of mouse and human embryonic stem cells into hepatic lineages. *Genes Cells*, **13**, 731–746.
24. Takahashi, K. and Yamanaka, S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.
25. Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E. and Jaenisch, R. (2007) *In vitro* reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature*, **448**, 318–324.
26. Cedar, H. and Bergman, Y. (2009) Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.*, **10**, 295–304.
27. Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M. and Schubeler, D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, **39**, 457–466.
28. Reik, W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425–432.
29. Bergman, Y. and Mostoslavsky, R. (1998) DNA demethylation: Turning genes on. *Biol. Chem.*, **379**, 401–407.
30. Eden, S. and Cedar, H. (1994) Role of DNA methylation in the regulation of transcription. *Curr. Opin. Genet. Dev.*, **4**, 255–259.
31. Ooi, S.K.T. and Bestor, T.H. (2008) The colorful history of active DNA demethylation. *Cell*, **133**, 1145–1148.
32. Wu, S.C. and Zhang, Y. (2010) Active DNA demethylation: many roads lead to Rome. *Nat. Rev. Mol. Cell Biol.*, **11**, 607–620.
33. Mayer, W., Niveleau, A., Walter, J., Fundele, R. and Haaf, T. (2000) Embryogenesis: demethylation of the zygotic paternal genome. *Nature*, **403**, 501–502.
34. Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W. and Walter, J. (2000) Active demethylation of the paternal genome in the mouse zygote. *Curr. Biol.*, **10**, 475–478.
35. Reik, W., Dean, W. and Walter, J. (2001) Epigenetic reprogramming in mammalian development. *Science*, **293**, 1089–1093.
36. Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C. and Zhang, Y. (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, **466**, 1129–1133.
37. Young, D.W., Hassan, M.Q., Yang, X.-Q., Galindo, M., Javed, A., Zaidi, S.K., Furcinitti, P., Lapointe, D., Montecino, M., Lian, J.B. *et al.* (2007) Mitotic retention of gene expression patterns by the cell fate-determining transcription factor Runx2. *Proc. Natl Acad. Sci. USA*, **104**, 3189–3194.
38. Zaidi, S.K., Young, D.W., Montecino, M.A., Lian, J.B., van Wijnen, A.J., Stein, J.L. and Stein, G.S. (2010) Mitotic bookmarking of genes: a novel dimension to epigenetic control. *Nat. Rev. Genet.*, **11**, 583–589.
39. Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K. and Yamanaka, S. (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, **131**, 861–872.
40. Vierbuchen, T., Ostermeier, A., Pang, Z.P., Kokubu, Y., Sudhof, T.C. and Wernig, M. (2010) Direct conversion of fibroblasts to functional neurons by defined factors. *Nature*, **463**, 1035–1041.
41. Ieda, M., Fu, J.-D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B.G. and Srivastava, D. (2010) Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, **142**, 375–386.
42. Suemori, H., Yasuchika, K., Hasegawa, K., Fujioka, T., Tsuneyoshi, N. and Nakatsuji, N. (2006) Efficient establishment of human embryonic stem cell lines and long-term maintenance with stable karyotype by enzymatic bulk passage. *Biochem. Biophys. Res. Commun.*, **345**, 926–932.
43. D'Amour, K.A., Agulnick, A.D., Eliazer, S., Kelly, O.G., Kroon, E. and Baetge, E.E. (2005) Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nat. Biotech.*, **23**, 1534–1541.
44. Hayashi, H., Nagae, G., Tsutsumi, S., Kaneshiro, K., Kozaki, T., Kaneda, A., Sugisaki, H. and Aburatani, H. (2007) High-resolution mapping of DNA methylation in human genome using oligonucleotide tiling array. *Hum. Genet.*, **120**, 701–711.
45. Hippo, Y., Watanabe, K., Watanabe, A., Midorikawa, Y., Yamamoto, S., Ihara, S., Tokita, S., Iwanari, H., Ito, Y., Nakano, K. *et al.* (2004) Identification of soluble NH2-terminal fragment of glypican-3 as a serological marker for early-stage hepatocellular carcinoma. *Cancer Res.*, **64**, 2418–2423.
46. Shiraki, N., Yoshida, T., Araki, K., Umezawa, A., Higuchi, Y., Goto, H., Kume, K. and Kume, S. (2008) Guided differentiation of embryonic stem cells into Pdx1-expressing regional-specific definitive endoderm. *Stem Cells*, **26**, 874–885.
47. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
48. Saldanha, A.J. (2004) Java Treeview–extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.

RESOURCE

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.

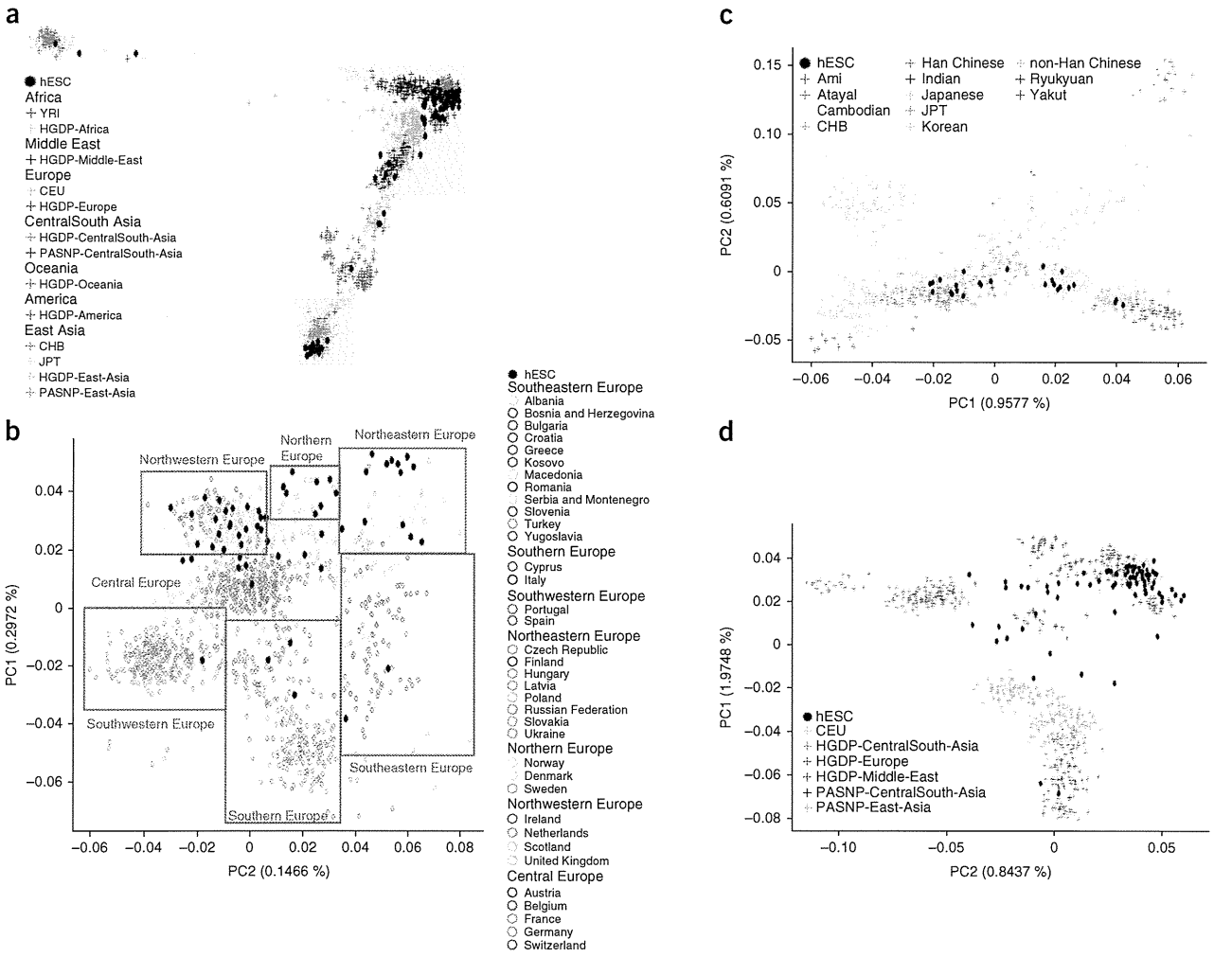

Figure 1 Population structure of the human ES cell lines analyzed. Principal component (PC) analyses were conducted on the entire final merged data set. PC1 and PC2 are plotted on the *y* and *x* axes, respectively. (**a**) The overall distribution of the human ES cell lines studied compared to the major 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

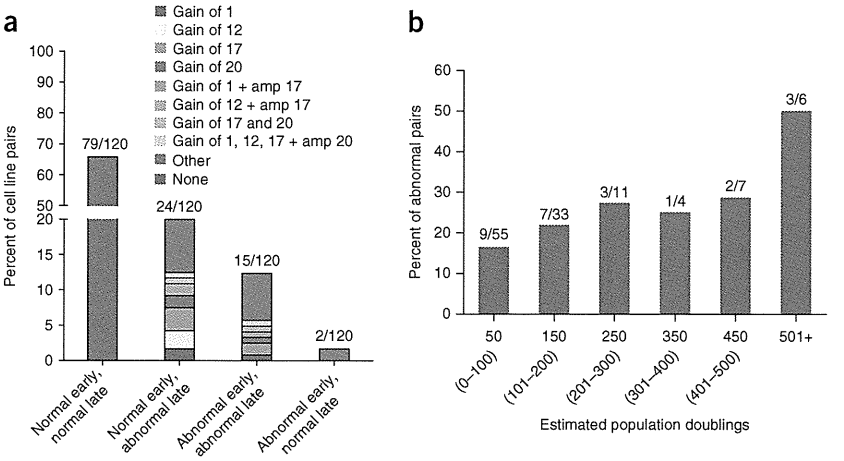

Figure 2 Cytogenetic changes occurring during prolonged passage of human ES cells. (**a**) Percentage of human ES cell line pairs that exhibited a karyotypic abnormality in either early or late passages, or both. Cell lines were excluded if they were known to be derived from karyotypically abnormal embryos. 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

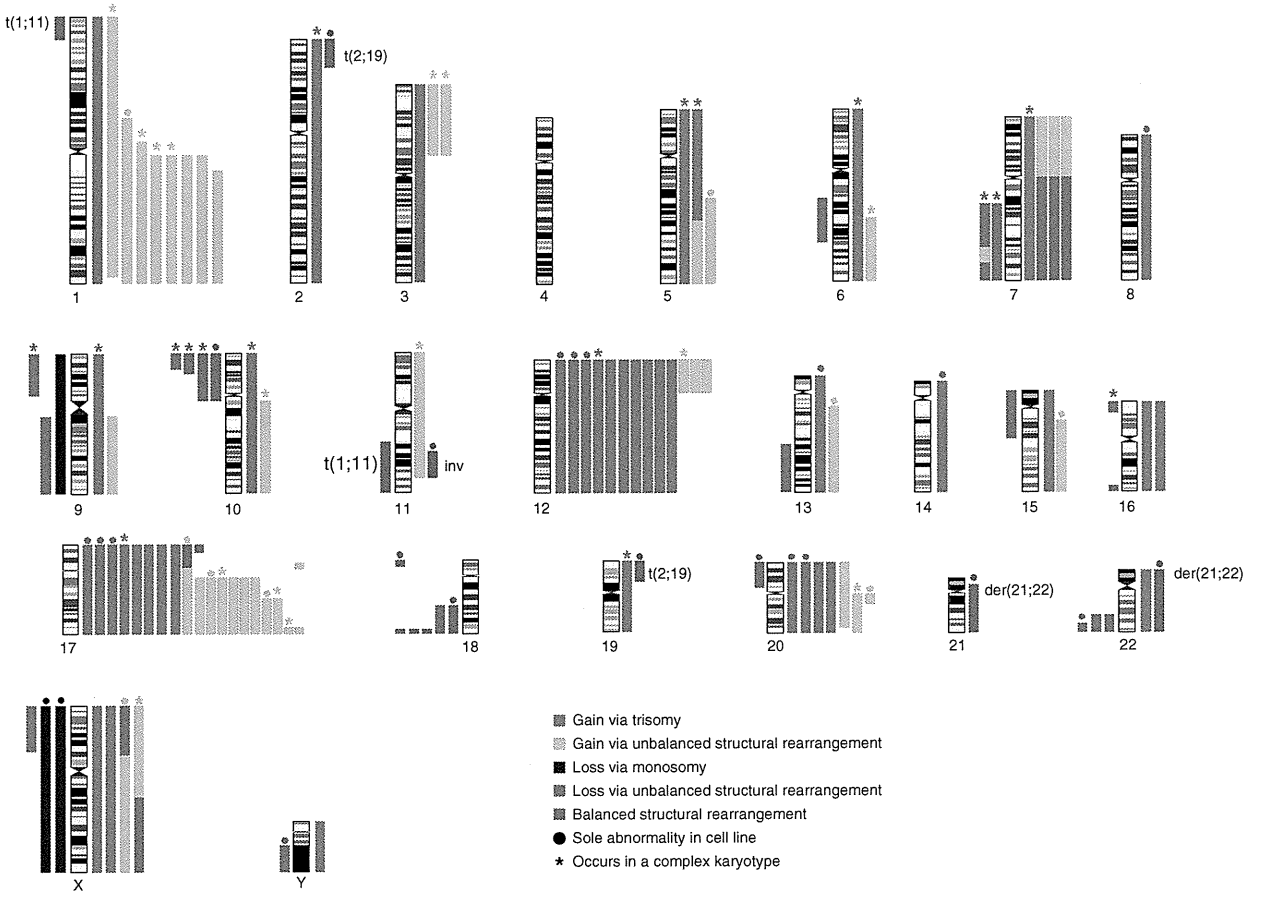

Figure 3 Ideogram demonstrating the chromosome changes found in this study. Each colored bar represents one chromosome change occurrence in one cell line. Chromosome losses and gains are shown to the left and right of the ideogram, respectively, except that those instances where a single chromosome rearrangement 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 (CO2 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

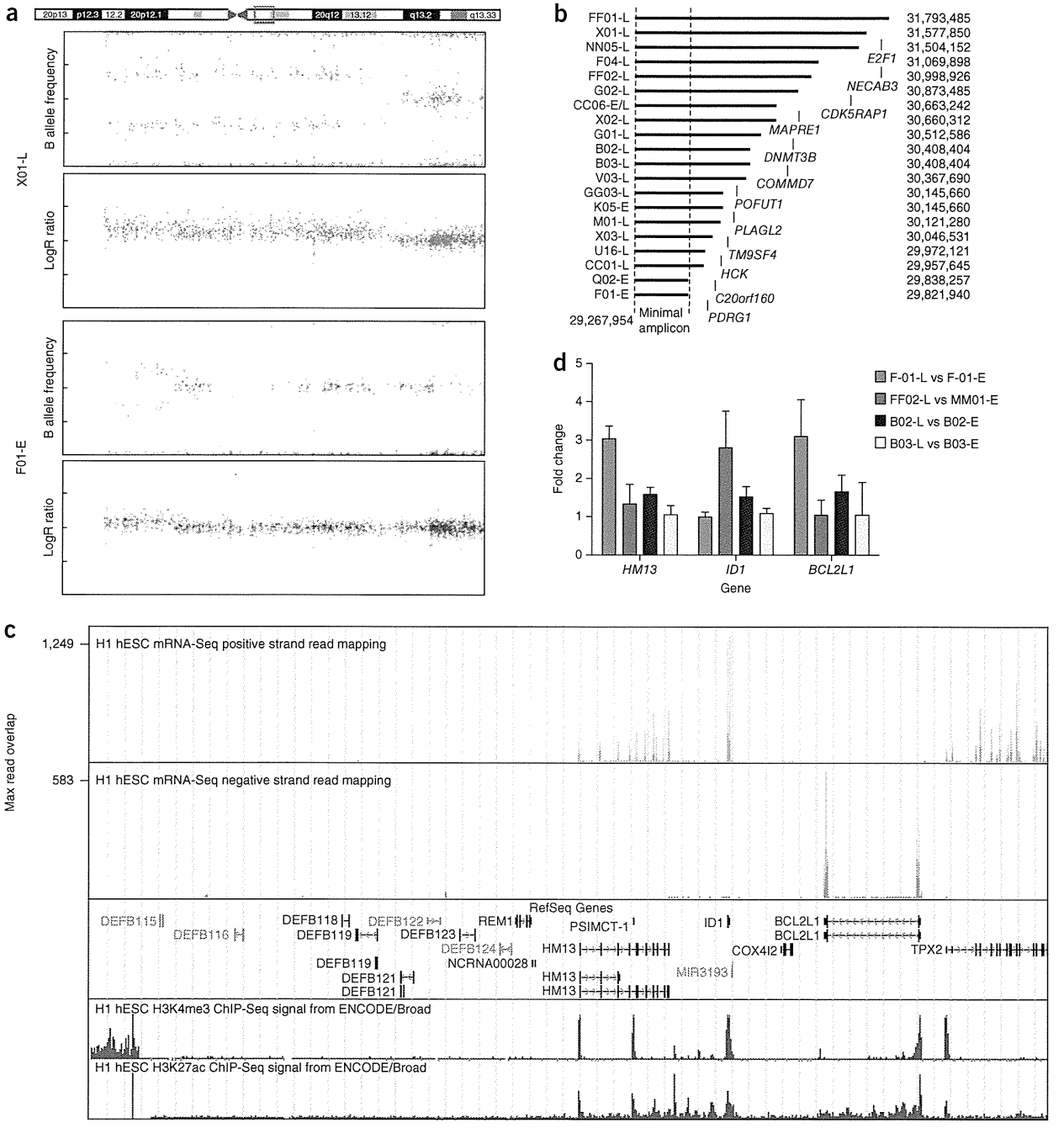

Figure 4 Copy number variation occurrence in human ES cell lines during prolonged passage. (**a**) 20q11.21 gain. The region on chromosome 20 frequently found to experience gain over extended human ES cell culture is indicated by the red boxed region in the chromosome ideogram. Also shown are the B allele frequency and logR ratio plots representing instances of one of the longest and one of the shortest 20q11.21 structural variants. (**b**) Length representation of all individual occurrences of gains in the 20q11.21 region. Samples from which the structural variant was derived are indicated on the left-hand column. The invariant 5′ region and the variable 3′ positions are indicated. Position of genes outside of the minimal amplicon that show greater than 20 RPKM level of expression in human ES cells are shown (RPKM = number of reads that map per kilobase of exon model per million mapped reads for each gene). (**c**) Expression, RefSeq gene, and regulation tracks in the minimal amplicon. Positive and negative strand mRNA-Seq data from H1 human ES cells indicating polyA RNA transcripts expressed within the minimal amplicon region (chr20:29,267,954-29,853,264) are shown together with H1 human ES cell ChIP-Seq data of histone modifications considered universal predictors of enhancer and promoter activity. (**d**) Comparison of expression levels of three genes (*HM13*, *ID1*, *BCL2L1*) contained within the identified minimal 20q11.2 amplicon between early- (normal) and late-passage (20q11.2 CNV carrying) samples. MM01 and FF02 are genetically identical sub-lines from two separate laboratories, MM01 has no amplification at 20q11.2, whereas FF02 possesses a copy number change at 20q11.2 that includes the identified minimal amplicon (**b**).

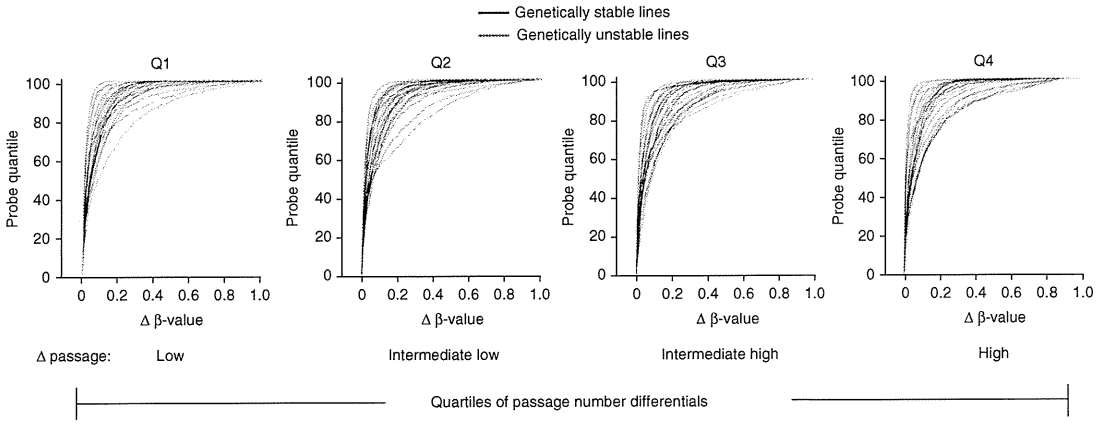

Figure 5 Cumulative distribution function of methylation changes in human ES cells in this study. The change in DNA methylation is represented by empirical CDF curves of the absolute difference in DNA methylation between early- and late-passage cell-line pairs for all 1,536 analyzed probes. The black curves denote genetically stable lines; the red curves denote genetically unstable lines. All analyzed lines were divided into quartiles based on the passage-number difference between the early and late member of each pair. The first quartile contains the lines with the lowest difference in passage number between the early and late sample (range 4 to 47), whereas the fourth quartile contains the lines with the highest difference in estimated population doublings (range 210 to 1,482).

over those without (**Supplementary Fig. 4**). It has also been recently reported that Bcl-X_L, the long, anti-apoptotic isoform encoded by the *BCL2L1* locus, can suppress apoptosis in human ES cells and increase their cloning efficiency[52]. Further, when we transfected MM01 ES cells with a constitutive vector encoding Bcl-X_L, the predominant isoform expressed in human ES cells, these cells showed a distinct growth advantage with respect to the parental cells (**Supplementary Fig. 4**).

DNA methylation analysis

To examine whether cell lines that are genetically unstable at the karyotype level tend to show higher levels of epigenetic instability, we analyzed DNA methylation patterns, focusing on developmentally relevant genes known to be targets of abnormal promoter DNA methylation in cancer[40], and thus most likely to be subjected to selection for altered expression during culture adaptation. For this we used a custom GoldenGate DNA methylation array developed to interrogate DNA methylation changes in known polycomb group protein (PcG) targets in human ES cells[53]. In general, the DNA methylation patterns of the human ES cells tended to be unstable, with both increases or decreases depending upon the locus (**Fig. 5** and **Supplementary Data Set 4**). **Table 2** summarizes those genes that were most frequently subject to gain or loss of methylation during passage, or that showed the least change. Overall, we did not observe any hot spots for DNA methylation at the ~1,500 loci interrogated in the array used in this study, and chromosomes 12, 17 and 20 were not any more methylated, on average, than the rest of the genome.

Table 2 The top 20 genes that were most frequently gained, lost or showed no change in DNA methylation levels in the 120 ES cell lines analyzed at early and late passage

Gained DNA methylation	Lost DNA methylation	No change in DNA methylation
GPC3	*CBLN4*	*NR4A3*
RAB9B	*HIST1H3C*	*EPHA4*
TCEAL4	*LY6H*	*COL12A1*
IL1RAPL2	*HIST1H4L*	*TIGD3*
ESX1	*ANKRD20B*	*SNX7*
TCEAL3	*HIST1H4F*	*PIP5K1B*
AMMECR1	*DMRT2*	*KCNJ2*
MGC39900	*TTLL7*	*T*
LRCH2	*FOXD4L1*	*ZBTB7A*
ZCCHC12	*FOXD4L2*	*IL20RA*
REPS2	*ONECUT1*	*GNAO1*
SOX3	*MAL*	*EPB41L4A*
RP13-360B22.2	*SYT6*	*VDR*
TSC22D3	*BHLHB4*	*HS6ST3*
NHS	*HIST1H3I*	*VGLL2*
TCEAL7	*XTP7*	*SIX1*
MGC4825	*NEUROG1*	*SFT2D2*
GPR50	*TFAP2D*	*BCAN*
BCL2L10	*DRD5*	*ELMOD1*
CDX4	*ASCL2*	*PTGER4*

GPC3 gained more than 5% DNA methylation (range: 98–5%) in over 70% of the samples analyzed, whereas *CBLN4* lost more than 5% DNA methylation (range: 70–5%) in over 60% of them. The genes listed in the "No change" column showed fluctuations in DNA methylation <1% in all samples profiled.

As shown by cumulative distribution function (CDF) curves, most cell lines underwent extensive DNA methylation changes during their time in culture (Online Methods). However, there was a marked difference between the cell lines. For example, in some cell lines there were few changes observed even if there was a large difference in passage level between the early- and late-passage samples (**Fig. 5** Q4 and **Supplementary Table 3**), whereas with other pairs there were large differences observed even when the passage-level difference between the samples was small (**Fig. 5** Q1 and **Supplementary Table 3**). However, the causes of the variation in methylation stability between the lines were not evident. There was no obvious laboratory effect, and the karyotypically abnormal cell lines were not any more unstable than their karyotypically normal counterparts. This suggests that genetic instability played little to no role in the epigenetic instability of the cell lines analyzed. In addition, the DNA methylation patterns of the sibling ES cell lines were as different between themselves as they were between unrelated lines (**Supplementary Data Set 4**), suggesting that the genetic background of human ES cells plays a minor role in the degree of their epigenetic instability.

DISCUSSION

The occurrence of genetic and epigenetic change in human ES cells on prolonged passage is clearly important with respect to their use in regenerative medicine. Understanding the key genes involved and the mechanisms that drive change is important, not only for minimizing the impact of such variants in applications of ES and iPS cells, but also

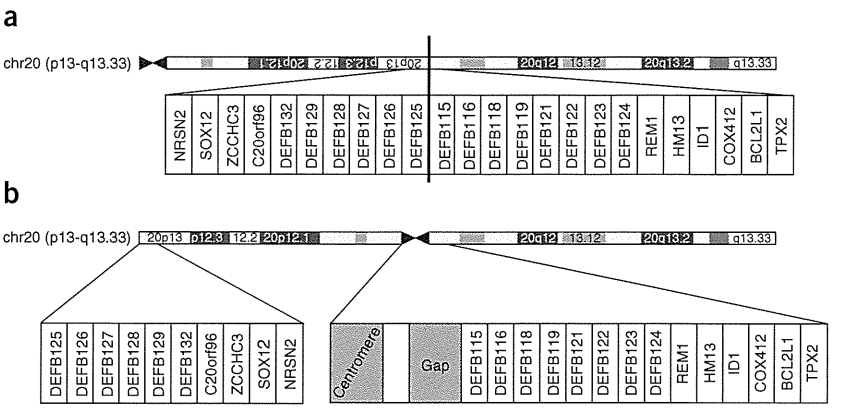

Figure 6 Recent pericentric inversion associated with 20q11.21 susceptibility to gain. (**a**) The ancestral condition of chromosome 20 before a pericentric inversion in the last common ancestor of the gorilla, chimp and human. (**b**) Structure of human chromosome 20 with the location of the gap indicated in which the proximal end of all 20q11.21 amplicons lie.

for exploring the mechanisms that control the fate decisions of pluripotent cells between self-renewal, death and differentiation. Nevertheless, given the scale of the present study, it is striking that most of the ES cell lines studied (79/120 pairs, 66%) remained karyotypically normal, even after many passages, whereas it was only with respect to chromosome 20 that evidence for structural variants in a specific region offering a strong selective advantage could be deduced. Among the small number of iPS cell samples studied, 3 out of 11 had abnormal karyotypes, with 1 of the 3 having the 20q11 gain in the late-passage sample.

Since the first reports of nonrandom chromosomal gain in human ES cells, many studies by standard karyology and by various molecular techniques, including CGH and SNP arrays, have found that, indeed, certain regions of the genome of both ES and, more recently, iPS cells are particularly subject to such genetic change upon prolonged passage in culture. Recently, it was also shown that iPS cells acquire mutations during their derivation, although many such mutations are lost on subsequent passaging[54]. It is commonly assumed that those genetic changes that repeatedly appear in pluripotent stem cells provide variant cells with a growth advantage, but the nature of the selective advantage is unclear. At the molecular karyotype level, it is difficult to disentangle changes that simply reflect variants existing in the human population from those acquired during culture. To address this, we explicitly sought to compare the genomes of a large set of human ES cell lines at two different passage levels and from as diverse a set as possible of the principal laboratories isolating these cells around the world. Although the number of human ES cell lines that have been derived worldwide is uncertain, the 125 ES cell lines analyzed in this study represent a substantial proportion of those commonly available. Notably, our data show that these lines include representatives of most major ethnic groups, reflecting far greater ethnic diversity than previously reported[55,56].

One feature of the human genome emphasized by the current study is that some regions are especially dynamic, particularly but not exclusively those including repetitive elements. In the current panel of ES cells, many regions showed gains or losses between the passage levels, but with no consistency, suggesting that there is no common selection pressure driving the copy number changes. That such dynamically variable regions were readily detected suggests that human ES cell cultures may go through population size restrictions more often than appreciated. Indeed, the cell cycle time of human ES cells is about 18–20 h, but common culture practice involves splitting cultures at low split ratios every 4–5 d or longer. This implies a very large proportion of undifferentiated cells, maybe as many as 90%, are lost between passages of stock cultures[33].

Likewise, the DNA methylation status of the ES cell lines also appeared to change dynamically. Although there was a marked increase in differential DNA methylation with time, indicated by the greater number of DNA methylation changes in the cell lines with the highest differences in passage number, there was also a substantial variation between lines that had undergone similar differences in passage numbers. Thus, human ES cells change not only genetically, but also epigenetically in culture. This conclusion is consistent with several other smaller scale studies that have interrogated human ES cells with respect to either general DNA methylation[25], or imprinting[29,31]. These studies all found DNA methylation and imprinting changes that appeared to be variable between lines and were locus dependent. However, we could not identify specific recurring regions subject to methylation in the genome and there was no observed correlation between DNA methylation changes and chromosomal abnormalities. This suggests that, in general, changes in DNA methylation may be a dynamic process and not necessarily associated with adaptation as such. This point is reinforced by the observation that DNA methylation is markedly different between sibling lines.

In addition to these apparently stochastic and dynamic changes in the genome and epigenome, we did detect marked nonrandom changes in certain parts of the genome. The karyotypic changes seen in the current study match well with other published reports (**Supplementary Fig. 5**)[1]. Gains of chromosomes 1, 12, 17 and 20 and losses of chromosomes 10p and 18q are common in both data sets, and it is only gains of chromosomes 12, 17 and 20 that are often seen as a sole karyotypic change. However, recurrent deletion of chromosome 22q is a novel finding. On the other hand, the gain of chromosome X is a relatively common finding in published studies, whereas only two instances of gain and three instances of loss were observed in the present study. In the light of their relatively frequent occurrence, the minimal amplicons 1q21-qter, 12p11-pter, 17q25-qter and 20q11.2, and perhaps minimal deletions 10p13-pter, 18q21-qter and 22q13-qter deserve special attention as being likely to harbor genes of particular importance for the culture adaptation of human ES cells.

The frequent nonrandom gain of chromosomes 1, 12, 17 and 20 suggests that these chromosomes include a gene(s) that, when overexpressed, confers a growth advantage. Yet, it is striking that in our current extensive study, as in previous studies, structural variant analysis did not point to any frequent repetitive minimal amplicon occurring on chromosomes 1, 12 and 17. Obvious candidate genes are located on these chromosomes—for example, *NANOG* on chromosome 12—but none seems to be more subject to structural variants than other genes on these chromosomes in the absence of karyotypic change. We did see gains spanning the neighboring *SLC2A3/NANOGP1* region described in a recent study[46] but this is just as prevalent, if not more so, within our reference samples and spread across most major ethnic groups, suggesting it is a common structural variant in the human population rather than specific to human ES cells. Together, these observations suggest that the selective advantage attributable to the

gain of chromosomes 1, 12 and 17 may depend upon overexpression of genes or genetic elements at multiple, spatially separated loci, or upon the combination of a structural gene with a long range *cis*-acting regulatory element such that both units must be amplified together to yield an increased function. Alternatively, the appearance of gains within smaller regions may be restricted by chromosomal structure less susceptible to this form of mutation.

By contrast, and in agreement with other studies[5,10,11,23,46,57], our karyotypic and structural variant data point to a region (20q11.21) that, when amplified, apparently drives selection. In this study, because of the much larger number of cell lines and our ability to compare early and late passage, we were able to map the gain to a specific region. Other studies have also reported that gains in this region are associated with enhanced growth characteristics[23], and at least some of the lines in the present study were reported by their contributors to have increased population growth rates (data not shown). The frequency of this gain (25% of the karyotypically normal cell lines), combined with the enrichment in late-passage samples, clearly indicates its selective advantage in human ES cell culture. The mechanism for the selective advantage presumably lies in the minimal region shared by all 22 affected lines, a region containing 13 genes, only three of which are known to be expressed in human ES cells: *HM13*, *ID1* and *BCL2L1*.

A recent genome-wide RNA interference (RNAi) screen highlights the functional importance of *BCL2L1*, an anti-apoptotic factor, in human ES cell biology[58]. This RNAi screen ranked *BCL2L1* twenty-second of 21,121 genes in reducing proliferation after knockdown, whereas *HM13* and *ID1* were ranked 6,679th and 4,224th, respectively[58]. Additionally, a recent structural variant screen of >3,000 specimens from two dozen cancer cell types similarly identified a reoccurring gain on 20q11.21 in which *BCL2L1* was also contained within the minimal amplicon, and knockdown experiments indicated a role for *BCL2L1* in cancer cell proliferation[59]. Recently, it has also been reported that overexpression of the related anti-apoptotic gene, *BCL2*, enhances the survival of human ES cells[60], although *BCL2* is encoded with the region of chromosome 18 subject to recurrent loss in the current data set. Taken together, these observations suggest that similar mutations shared between ES and cancer cells lead to a selective advantage during clonal evolution. The temporal component of our study, where we see[17] instances of early/normal to late/mutated transitions, provides additional support for the notion that the 20q11.21 mutation is the driver mutation in the clonal evolution of these adapted stem cells. Although a role for *ID1* (ref. 61) and *HM13* cannot be excluded, enhanced cell survival due to elevated expression levels of *BCL2L1* offers the most likely mechanism.

The repeated appearance of a structural variant across multiple lines requires both a selective advantage for the variant (e.g., increased expression of *BCL2L1*), and a predisposition for the respective mutation to occur. It is noteworthy that the proximal end of all human ES cell 20q11.21 gains lies within a gap region of the current human assembly[62]. The presumption is that the highly repetitive sequence within this gap predisposes the region to structural rearrangement. With the link between genome rearrangements, primate evolution and disease association[63], it is notable that this gap coincides with a recent chromosomal rearrangement, a pericentric inversion[64], occurring in the last common ancestor of gorilla, chimp and human (**Fig. 6**). The gap region, possibly a centromeric remnant of a tandem duplication[62], introduces the repetitive sequence creating 20q11.21 rearrangement (or amplification) susceptibility. The frequency of appearance that is created by this combination of mutability and the decreased apoptosis warrants routine surveillance similar to that now done in karyotypic analysis.

The identification of genes that drive both cancer progression of EC cells in germ cell tumors and the progressive culture adaptation of ES cells has been a goal since the first clear recognition that gain of sections of the short arm of chromosome 12 is an invariant feature of EC cells[14]. The commonality of the changes in the tumors and in the ES cell in culture suggests common underlying mechanisms. However, the identification of a specific driver gene on chromosomes 1, 12 and 17 has been elusive, suggesting that more than one gene may be involved in the growth advantage of the aneuploid cells. Our present results now point to a specific gene subject to gain, most likely the anti-apoptotic gene, *BCL2L1*, on chromosome 20, that may promote the survival of ES cells *in vitro* and EC cells *in vivo*, thereby providing a strong growth advantage, whether in cancers or *in vitro*.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The International Stem Cell Initiative is funded by The International Stem Cell Forum. The authors would like to acknowledge the following: Medical Research Council, UK (P.W.A., H.M.); Mohammad Pakzad & Adeleh Taei, Royan Institute (H.B., G.H.S.); California Institute for Regenerative Medicine (CIRM) (E.C., P.W.L.); Institute of Medical Biology, A*STAR, Singapore (J.M.C.); Ministry of Education, Youth and Sports of the Czech Republic (P.D., A.H.); Stem Cell Research Center of the 21st Century Frontier Research Program, Ministry of Education, Science & Technology, Republic of Korea (SC-1140) (D.R.L., S.K.O.); Ministry of Science and Technology of China (863 program 2006AA02A102) (L.G.); Swedish Research Council, Cellartis (O.H.); Department of Biotechnology, Government of India, UK-India Education and Research Initiative and the Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India (M.I.); Program for Promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation, Leading Project of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program) of the Japan Society for the Promotion of Science (JSPS), Grants in-Aid for Scientific Research of JSPS and MEXT (T.I., S.Y., K.T.); Swiss National Science Foundation (grant no. 4046-114410) (M.J.); Shanghai Science and Technology Developmental Foundation (06DJ14001), Chinese Ministry of Science and Technology (2007CB948004) (Y.J.); funding from the North West Science Fund, UK (S.K.); One North East Regional Developmental Agency, Medical Research Council, UK, Newcastle University (M.L.); research funding from the Australian Stem Cell Centre (A.L.L.); The Netherlands Proteomics Consortium grant T4-3 (C.M.); Stem Cell Network, Canada (A.N.); National BioResource Project, MEXT, Japan (N.N.); Singapore Stem Cell Consortium (SSCC) & the Agency for Science Technology and Research (A*STAR) (S.K.W.O., P.R.) and the Genome Institute of Singapore Core Genotyping Lab (P.R.); Academy of Finland, Sigrid Juselius Foundation (T.O.); Conselho Nacional de Desenvolvimento Científico e Tecnológico/Departamento de Ciência e Tecnologia do Ministério da Saúde (CNPq/MS/DECIT), and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (L.V.P.); supported by the kind donation of Judy and Sidney Swartz (B.R.); financial support from the Faculty of Medicine, University of New South Wales (UNSW) and the National Health and Medical Research Council (NHMRC) Program Grant no. 568969 (Perminder Sachdev), South Eastern Sydney and Illawarra Area Health Service (SEIAHS) for making hES cell line Endeavour-2 available for this study, and H. Chung and J. Kim for their help in preparing the samples (K.S.); Academy of Finland (grant 218050), the Competitive Research Funding of the Tampere University Hospital (grant 9F217) (H. Skottman).

AUTHOR CONTRIBUTIONS

Project coordination: P.W.A. **Cytogenetic analyses**: D.B., A.D., E.M., K.D.M. and T.G.-L. **Molecular karyotyping by SNP BeadChip**: P.R. **DNA methylation arrays**: R.M.B. and P.W.L. **Administration and data curation**: A. Ford and P.J.G. **Data analysis and manuscript drafting**: P.W.A., S.A., D.B., N.B., R.M.B., P.J.G., K.H., L.H., B.B.K., Y. Mayshar, S.K.W.O., M.F.P. and P.R. **The scientific management of the ISCI project was provided by a steering committee comprising:** P.W.A., N.B., B.B.K., S.K.W.O., M.F.P., J.R. and G.N.S. **Sample contribution**: A. Colman, A. Robins, A. Hampl, A. Bosman, A.M. Fraga, A. Nagy, A.B.H. Choo, A.L. Laslett, A. Feki, A. Kuliev, A. Kresentia Irwanto, B. Reubinoff, B. Sun, C. Denning,

C. Mummery, C. Li, C. Olson, C. Spits, D. Ben-Yosef, D. Collins, D.J. Weisenberger, D. Ryul Lee, D. Ward-van Oostwaard, E. Chiao, E. Sherrer, Fei Pan, F. Holm, G. Anyfantis, G.Q. Daley, G.H. Salekdeh, G. Selva Raj, G. Caisander, H. Gourabi, H. Moore, H. Skottman, H. Suemori, H. Baharvand, H. Shen, I. Mateizel, In-Hyun Park, J. Sheik Mohamed, J. Downie, J. Eun Lee, J.M. Crook, J. Chen, J. Hyllner, J.-C. Biancotti, J. Baker, K. Sermon, K. Amps, K. Narwani, K. Takahashi, K. Sidhu, L. Ge, L.S. Lim, L. Young , Q. Zhou, L. Guangxiu, L.V. Pereira, L. Armstrong, M. Lako, M.S. Inamdar, M.A. Lagarkova, M.B. Munoz, M. Mileikovsky, M.V. Camarasa, M. Jaconi, M. Gropp, N. Lavon, N. Strelchenko, N. Nakatsuji, O. Kopper, O. Hovatta, O. Qi, P. Venu, P.A. De Sousa, P. Dvorak, R. Strehl, R. Suuronen, S. Kiselev, S. Yong Moon, S. Yamanaka, S. Sivarajah, S. Beil, S.L. Minger, S.K.W. Oh, S. Pells, S. Kyung Oh, S. Kimber, T. Miyazaki, T.E. Ludwig, T. Ishii, T.C. Schulz, T. Otonkoski, T. Tuuri, T. Frumkin, V. Kukharenko, V. Fox, W. Herath, Y. Jin, Y. Min Choi, Y. Ma, Y. Wu and Y. Verlinsky.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/nbt/index.html.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Baker, D.E. *et al.* Adaptation to culture of human embryonic stem cells and oncogenesis *in vivo*. *Nat. Biotechnol.* **25**, 207–215 (2007).
2. Draper, J.S. *et al.* Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. *Nat. Biotechnol.* **22**, 53–54 (2004).
3. Mitalipova, M.M. *et al.* Preserving the genetic integrity of human embryonic stem cells. *Nat. Biotechnol.* **23**, 19–20 (2005).
4. Hoffman, L.M. & Carpenter, M.K. Characterization and culture of human embryonic stem cells. *Nat. Biotechnol.* **23**, 699–708 (2005).
5. Maitra, A. *et al.* Genomic alterations in cultured human embryonic stem cells. *Nat. Genet.* **37**, 1099–1103 (2005).
6. Buzzard, J.J., Gough, N.M., Crook, J.M. & Colman, A. Karyotype of human ES cells during extended culture. *Nat. Biotechnol.* **22**, 381–382, author reply 382 (2004).
7. Caisander, G. *et al.* Chromosomal integrity maintained in five human embryonic stem cell lines after prolonged in vitro culture. *Chromosome Res.* **14**, 131–137 (2006).
8. Inzunza, J. *et al.* Comparative genomic hybridization and karyotyping of human embryonic stem cells reveals the occurrence of an isodicentric X chromosome after long-term cultivation. *Mol. Hum. Reprod.* **10**, 461–466 (2004).
9. Rosler, E.S. *et al.* Long-term culture of human embryonic stem cells in feeder-free conditions. *Dev. Dyn.* **229**, 259–274 (2004).
10. Lefort, N. *et al.* Human embryonic stem cells reveal recurrent genomic instability at 20q11.21. *Nat. Biotechnol.* **26**, 1364–1366 (2008).
11. Spits, C. *et al.* Recurrent chromosomal abnormalities in human embryonic stem cells. *Nat. Biotechnol.* **26**, 1361–1363 (2008).
12. Mayshar, Y. *et al.* Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell* **7**, 521–531 (2010).
13. Wang, N., Trend, B., Bronson, D.L. & Fraley, E.E. Nonrandom abnormalities in chromosome 1 in human testicular cancers. *Cancer Res.* **40**, 796–802 (1980).
14. Atkin, N.B. & Baker, M.C. Specific chromosome change, i(12p), in testicular tumours? *Lancet* **320**, 1349 (1982).
15. Rodriguez, E. *et al.* Molecular cytogenetic analysis of i(12p)-negative human male germ cell tumors. *Genes Chromosom. Cancer* **8**, 230–236 (1993).
16. Skotheim, R.I. *et al.* New insights into testicular germ cell tumorigenesis from gene expression profiling. *Cancer Res.* **62**, 2359–2364 (2002).
17. Mostert, M. *et al.* Comparative genomic and in situ hybridization of germ cell tumors of the infantile testis. *Lab. Invest.* **80**, 1055–1064 (2000).
18. Schneider, D.T. *et al.* Genetic analysis of childhood germ cell tumors with comparative genomic hybridization. *Klin. Padiatr.* **213**, 204–211 (2001).
19. Looijenga, L.H. *et al.* Comparative genomic hybridization of microdissected samples from different stages in the development of a seminoma and a non-seminoma. *J. Pathol.* **191**, 187–192 (2000).
20. Longo, L., Bygrave, A., Grosveld, F.G. & Pandolfi, P.P. The chromosome make-up of mouse embryonic stem cells is predictive of somatic and germ cell chimaerism. *Transgenic Res.* **6**, 321–328 (1997).
21. Liu, X. *et al.* Trisomy eight in ES cells is a common potential problem in gene targeting and interferes with germ line transmission. *Dev. Dyn.* **209**, 85–91 (1997).
22. Zody, M.C. *et al.* DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. *Nature* **440**, 1045–1049 (2006).
23. Werbowetski-Ogilvie, T.E. *et al.* Characterization of human embryonic stem cells with features of neoplastic progression. *Nat. Biotechnol.* **27**, 91–97 (2009).
24. Narva, E. *et al.* High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. *Nat. Biotechnol.* **28**, 371–377 (2010).
25. Allegrucci, C. *et al.* Restriction landmark genome scanning identifies culture-induced DNA methylation instability in the human embryonic stem cell epigenome. *Hum. Mol. Genet.* **16**, 1253–1268 (2007).
26. Calvanese, V. *et al.* Cancer genes hypermethylated in human embryonic stem cells. *PLoS ONE* **3**, e3294 (2008).
27. Enver, T. *et al.* Cellular differentiation hierarchies in normal and culture-adapted human embryonic stem cells. *Hum. Mol. Genet.* **14**, 3129–3140 (2005).
28. Rugg-Gunn, P.J., Ferguson-Smith, A.C. & Pedersen, R.A. Epigenetic status of human embryonic stem cells. *Nat. Genet.* **37**, 585–587 (2005).
29. Adewumi, O. *et al.* Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. *Nat. Biotechnol.* **25**, 803–816 (2007).
30. Rugg-Gunn, P.J., Ferguson-Smith, A.C. & Pedersen, R.A. 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–R251 (2007).
31. Kim, K.P. *et al.* Gene-specific vulnerability to imprinting variability in human embryonic stem cell lines. *Genome Res.* **17**, 1731–1742 (2007).
32. Andrews, P.W. *et al.* The International Stem Cell Initiative: toward benchmarks for human embryonic stem cell research. *Nat. Biotechnol.* **23**, 795–797 (2005).
33. Olariu, V. *et al.* Modeling the evolution of culture-adapted human embryonic stem cells. *Stem Cell Res.* **4**, 50–56 (2010).
34. Martin, G.R. & Evans, M.J. The morphology and growth of a pluripotent teratocarcinoma cell line and its derivatives in tissue culture. *Cell* **2**, 163–172 (1974).
35. Andrews, P.W., Bronson, D.L., Benham, F., Strickland, S. & Knowles, B.B. A comparative study of eight cell lines derived from human testicular teratocarcinoma. *Int. J. Cancer* **26**, 269–280 (1980).
36. Chambers, I. *et al.* Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113**, 643–655 (2003).
37. Mitsui, K. *et al.* The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**, 631–642 (2003).
38. Darr, H., Mayshar, Y. & Benvenisty, N. Overexpression of NANOG in human ES cells enables feeder-free growth while inducing primitive ectoderm features. *Development* **133**, 1193–1201 (2006).
39. Korkola, J.E. *et al.* Down-regulation of stem cell genes, including those in a 200-kb gene cluster at 12p13.31, is associated with in vivo differentiation of human male germ cell tumors. *Cancer Res.* **66**, 820–827 (2006).
40. Widschwendter, M. *et al.* Epigenetic stem cell signature in cancer. *Nat. Genet.* **39**, 157–158 (2007).
41. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
42. Li, J.Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
43. Abdulla, M.A. *et al.* Mapping human genetic diversity in Asia. *Science* **326**, 1541–1545 (2009).
44. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
45. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
46. Laurent, L.C. *et al.* Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell* **8**, 106–118 (2011).
47. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
48. Assou, S. *et al.* A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. *Stem Cells* **25**, 961–973 (2007).
49. Peng, J.C. *et al.* Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* **139**, 1290–1302 (2009).
50. Nottke, A., Colaiacovo, M.P. & Shi, Y. Developmental roles of the histone lysine demethylases. *Development* **136**, 879–889 (2009).
51. Morin, R.D. *et al.* Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.* **18**, 610–621 (2008).
52. Bai, H. *et al.* Bcl-xL enhances single-cell survival and expansion of human embryonic stem cells without affecting self-renewal. *Stem Cell Res. (Amst.)* (in the press).
53. Lee, T.I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301–313 (2006).
54. Hussein, S.M. *et al.* Copy number variation and selection during reprogramming to pluripotency. *Nature* **471**, 58–62 (2011).
55. Mosher, J.T. *et al.* Lack of population diversity in commonly used human embryonic stem-cell lines. *N. Engl. J. Med.* **362**, 183–185 (2010).
56. Laurent, L.C. *et al.* Restricted ethnic diversity in human embryonic stem cell lines. *Nat. Methods* **7**, 6–7 (2010).
57. Wu, H. *et al.* Copy number variant analysis of human embryonic stem cells. *Stem Cells* **26**, 1484–1489 (2008).
58. Chia, N.Y. *et al.* A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature* **468**, 316–320 (2010).
59. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
60. Ardehali, R. *et al.* Overexpression of BCL2 enhances survival of human embryonic stem cells during stress and obviates the requirement for serum factors. *Proc. Natl. Acad. Sci. USA* **108**, 3282–3287 (2011).
61. Martins-Taylor, K. *et al.* Recurrent copy number variations in human induced pluripotent stem cells. *Nat. Biotechnol.* **29**, 488–491 (2011).
62. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).
63. Shaw, C.J. & Lupski, J.R. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum. Mol. Genet.* **13** Spec No 1, R57–R64 (2004).
64. Misceo, D. *et al.* Evolutionary history of chromosome 20. *Mol. Biol. Evol.* **22**, 360–366 (2005).