Dynamic Changes in the Copy Number of Pluripotency and Cell Proliferation Genes in Human ESCs and iPSCs during Reprogramming and Time in Culture

Louise C. Laurent,[1,3,4,*] Igor Ulitsky,[6,7] Ileana Slavin,[3,4] Ha Tran,[3,4] Andrew Schork,[2] Robert Morey,[1,3,4] Candace Lynch,[3,4] Julie V. Harness,[8] Sunray Lee,[9] Maria J. Barrero,[10,11] Sherman Ku,[5] Marina Martynova,[12] Ruslan Semechkin,[12] Vasiliy Galat,[13,14] Joel Gottesfeld,[5] Juan Carlos Izpisua Belmonte,[10,11] Chuck Murry,[15] Hans S. Keirstead,[8] Hyun-Sook Park,[9] Uli Schmidt,[16] Andrew L. Laslett,[17,18,19] Franz-Josef Muller,[3,4] Caroline M. Nievergelt,[2] Ron Shamir,[7] and Jeanne F. Loring[3,4]

[1]Department of Reproductive Medicine
[2]Department of Psychiatry
University of California, San Diego, La Jolla, CA 92093, USA
[3]Department of Chemical Physiology
[4]Center for Regenerative Medicine
[5]Department of Molecular Biology
The Scripps Research Institute, La Jolla, CA 92037, USA
[6]The Whitehead Institute, Cambridge, MA 02142, USA
[7]Department of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
[8]Department of Anatomy and Neurobiology, Sue and Bill Gross Stem Cell Center, University of California, Irvine, Irvine, CA 92697, USA
[9]Modern Cell &Tissue Technologies (MCTT) Inc., Seoul 139-240, South Korea
[10]The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[11]Centro de Medicina Regenerativa de Barcelona, Barcelona E-08003, Spain
[12]International Stem Cell Corporation, Oceanside, CA 92056, USA
[13]Department of Pathology, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA
[14]iPS and Human Stem Cell Core Facility, Northwestern University Children's Memorial Research Center, Chicago, IL 60614, USA
[15]Department of Pathology, University of Washington, Seattle, WA 98195, USA
[16]Stem Cell Laboratory, Sydney IVF, Sydney, New South Wales 2000, Australia
[17]Commonwealth Scientific and Industrial Research Organisation (CSIRO), Clayton, Victoria 3168, Australia
[18]Australian Stem Cell Centre, Clayton, Victoria 3168, Australia
[19]Department of Anatomy and Developmental Biology, Monash University, Clayton, Victoria 3168, Australia
*Correspondence: llaurent@ucsd.edu
DOI 10.1016/j.stem.2010.12.003

SUMMARY

Genomic stability is critical for the clinical use of human embryonic and induced pluripotent stem cells. We performed high-resolution SNP (single-nucleotide polymorphism) analysis on 186 pluripotent and 119 nonpluripotent samples. We report a higher frequency of subchromosomal copy number variations in pluripotent samples compared to nonpluripotent samples, with variations enriched in specific genomic regions. The distribution of these variations differed between hESCs and hiPSCs, characterized by large numbers of duplications found in a few hESC samples and moderate numbers of deletions distributed across many hiPSC samples. For hiPSCs, the reprogramming process was associated with deletions of tumor-suppressor genes, whereas time in culture was associated with duplications of oncogenic genes. We also observed duplications that arose during a differentiation protocol. Our results illustrate the dynamic nature of genomic abnormalities in pluripotent stem cells
and the need for frequent genomic monitoring to assure phenotypic stability and clinical safety.

INTRODUCTION

The tremendous self-renewal and differentiation capabilities of human pluripotent stem cells (hPSCs) make them potential sources of differentiated cells for cell therapy. Cell therapies are subject to rigorous safety trials, and high priority is placed on demonstrating that the cells are nontumorigenic (Fox, 2008). Because genetic aberrations have been strongly associated with cancers, it is important that preparations destined for clinical use are free from cancer-associated genomic alterations. Human embryonic stem cell (hESC) lines have been shown to become aneuploid in culture (Baker et al., 2007; Draper et al., 2004; Imreh et al., 2006; Maitra et al., 2005; Mitalipova et al., 2005), and the most frequent changes, trisomies of chromosomes 12 and 17, are also characteristic of malignant germ cell tumors (Atkin and Baker, 1982; Rodriguez et al., 1993; Skotheim et al., 2002). Aneuploidies can be detected by karyotyping, but less easily detectable subchromosomal genetic changes may also have adverse effects. Small abnormalities have been detected in hESCs by using comparative genomic hybridization

(CGH) and single-nucleotide polymorphism (SNP) genotyping (Lefort et al., 2008; Närvä et al., 2010; Spits et al., 2008). These studies lacked sufficient resolution and power to identify cell type-associated duplications and deletions. A recent study has reported the use of gene expression data to detect genomic aberrations in a large number of hESCs and hiPSCs (Mayshar et al., 2010). However, the methods used could reliably detect only relatively large (≥10 megabase) aberrations, and the lack of nonpluripotent samples for comparison precluded the authors from determining which regions of genomic aberration were specific to pluripotent stem cells.

In this study, we performed high-resolution SNP genotyping on a large number of hESC lines, induced human pluripotent stem cell lines (hiPSCs), somatic stem cells, primary cells, and tissues. We found that hESC lines had a higher frequency of genomic aberrations compared to the other cell types. Furthermore, we identified regions in the genome that had a greater tendency to be aberrant in the hESCs when compared to the other cell types examined. Recurrent regions of duplication were seen on chromosome 12, encompassing the pluripotency-associated transcription factor *NANOG* and a nearby *NANOG* pseudogene, and on chromosome 20, upstream of the DNA methyltransferase *DNMT3B*. Although the frequency of genomic aberrations seen in the hiPSC lines was similar to those of cultured somatic cells and tissues, we observed one of the recurrent areas of duplication characteristic of hESCs in one of the hiPSC lines.

Furthermore, comparison of 12 hiPSC lines generated from the same primary fibroblast cell line identified genomic aberrations that were present in the hiPSC lines and absent from the original fibroblast line. Analysis of early- and late-passage samples from these hiPSC lines allowed us to distinguish between events that arose during the process of reprogramming and those that accumulated during long-term passage. In general, deletions tended to occur with reprogramming and involve tumor-suppressor genes, whereas duplications accumulated with passaging and tended to encompass tumor-promoting genes. These results suggest that human pluripotent stem cell populations are prone to genomic aberrations that could compromise their stability and utility for clinical applications and that reprogramming and expansion in culture may lead to selection for particular genomic changes.

RESULTS

High-resolution SNP genotyping (1,140,419 SNPs) was performed on 324 samples, including 69 hESC lines (130 samples), 37 hiPSC lines (56 samples), 11 somatic stem cell lines (11 samples), 41 primary cell lines (41 samples), and 20 tissue types (67 samples), as well as samples of differentiated hESC lines and mixtures of known ratios of a sample with a known duplication with a sample without that duplication (Table S1 available online). Copy number variants for all samples were identified in parallel with two algorithms, CNVPartition (Illumina, Inc., Table S2A) and Nexus (Biodiscovery, Inc., Table S2B), both of which have been demonstrated to be appropriate for copy number variation (CNV) identification from SNP Genotyping data from Illumina microarrays (Kresse et al., 2010). The concordance between these two algorithms was high (76.08%

for deletions, 98.60% for loss of heterozygosity (LOH), and 93.04% for duplications on the base-pair level) (Table S2C). A subset of the CNV calls for both algorithms were validated via qPCR. For the CNVPartition calls, 82% of 3-copy gains and 43% of 1-copy losses were confirmed. For Nexus, 68% of allelic imbalance, 71% of copy number gain, 47% of copy number loss, and 100% of loss of heterozygosity calls were confirmed (Table S3, note that the allelic imbalance calls were judged to be correct if the qRT-PCR result indicated either a significant gain or a significant loss). Given the higher accuracy of the duplication calls in CNVPartition, and the ambiguity of the allelic imbalance calls in Nexus, CNVPartition was subsequently used as the primary algorithm. CNV calls that overlapped with common CNVs observed in a reference set of 450 HapMap samples (Conrad et al., 2010) were identified and removed from subsequent analyses.

Figure 1 shows a map of the areas of CNV identified in all the samples. Based on validation of the CNV calls by qRT-PCR, which indicated that duplication calls were markedly more accurate than deletion calls, we focused on duplications and large deletions. We inspected the B-allele frequency (BAF) and log R ratio (LRR) plots in order to combine adjacent areas of CNV where appropriate; it is well appreciated that CNV calling algorithms frequently break up large CNV events into multiple calls. For example, the SIVF021 line was shown to have a complete trisomy of chromosome 21 both by prenatal genetic screening (PGS) of the embryo and karyotyping of the hESC line, but CNVPartition and Nexus both call multiple noncontiguous regions of CNV for this sample on chromosome 21 (Table S2). A list of the regions mapped in Figure 1 is given in Table S4.

Large Regions of CNV in hESCs and hiPSCs
Several hESC samples showed duplications of large regions: the BG01 and BG01V samples both showed trisomy 12 and trisomy 17, but only the BG01 sample contained trisomy 3 and a deletion of the long arm of chromosome 7. The MIZ13 sample also contained trisomy 3. SIVF048 had a duplication of chromosome 5, and the WA07P34MNP sample had a deletion of the same chromosome (of note, this sample was from a directed differentiation experiment from hESC to motor neuron progenitor). The FES29 sample had a duplication of the short arm, and a deletion of the long arm, of chromosome 7. Large duplications of chromosomes 12, 17, and 20 were observed in multiple samples. A large region of 2-copy LOH on chromosome 22 was identified for the HFIB2IPS5 sample. In addition, large regions of 2-copy LOH were identified on the X chromosome in several samples. Because these samples were male, these calls corresponded to duplications on the X chromosome; duplications of the entire chromosome were identified for the BG01 hESC and the TH1.60OCT4SOX2 hiPSC samples, and a large duplication of the q-arm of the chromosome was found in the BG01V sample. The aneuploidies in SIVF003 (chr16), SIVF011 (chr5), and SIVF021 (chr21) were known prior to derivation from PGS. Aneuploidies and large duplications of chromosomes 1, 12, 17, and X have been previously reported to be common in hESCs (Baker et al., 2007; Draper et al., 2004; Imreh et al., 2006; Mitalipova et al., 2005).

In a recent publication (Närvä et al., 2010), complex mosaic aneuploidy was described in one of the lines we genotyped,
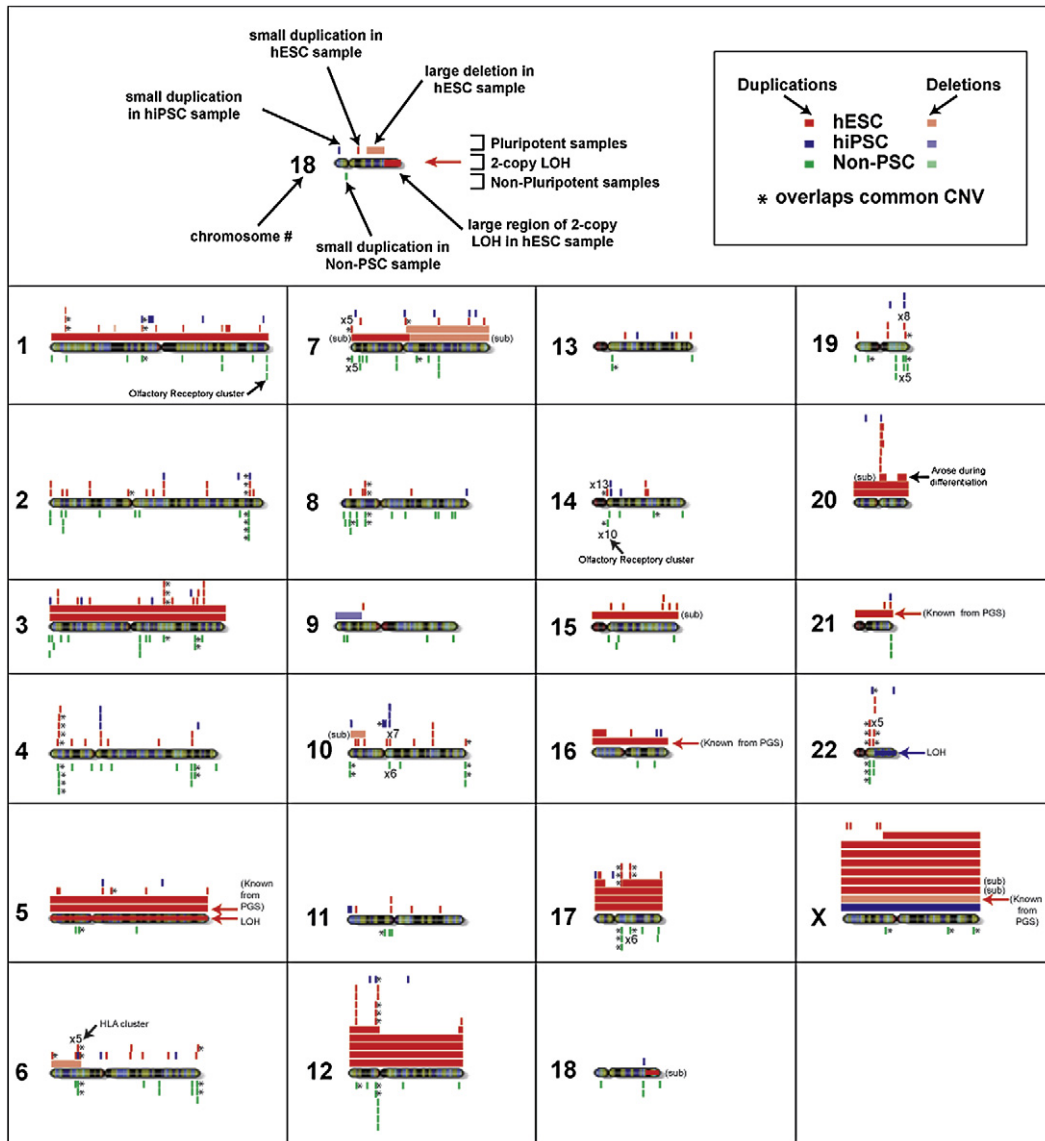
Figure 1. Duplications and Large Deletions Identified by CNVPartition Mapped onto the Genome, for All Samples

The number and extent of regions of CNV regions are shown. Duplicated regions (3 or 4 copies) are shown in the dark bars, deleted regions (0 or 1 copy) are shown in the light bars, and copy-neutral LOH regions are placed on the ideograms of the chromosomes. Where five or more samples of the same cell type have aberrations at the same region, the number of samples affected is indicated (e.g., ×5, ×10). Regions for hESC samples are shown in red, regions for hiPSC samples are shown in blue, and regions for non-PSC samples are shown in green. Some aneuploidies had been identified prior to hESC derivation and are indicated as "known from PGS." Regions where the CNV is present in only a subpopulation of the cells in a sample are denoted "(sub)." The three regions of duplication on chromosome 20 that arose in a subpopulation of the cells during differentiation of the WA07P96CMD7 sample are indicated. CNVs that overlap with the common CNVs observed in 450 HapMap samples (Conrad et al., 2010) are indicated by an asterisk. See also Figure S1 and Tables S1–S4.

FES61. In our analysis, the B-allele frequency pattern from the SNP genotyping data indicated that this line contained genetic material from three male individuals (Figure S1), which makes the data from this line uninterpretable for CNV analysis. We therefore excluded this line from further analysis.

Recurrent Regions of CNV in hESCs and hiPSCs

In addition to these large duplications and deletions, we observed multiple smaller regions of CNV, including both deletions and duplications, which we examined to identify regions of recurrent CNV in the human pluripotent stem cell samples. As noted above, the validation rate for small duplications was significantly higher than for small deletions, and therefore we focused on duplications for our analyses. We ensured that the recurrent regions identified were associated with the pluripotent state rather than with high-frequency CNVs found in the human population by comparing the CNVs found in the hPSC samples with those found in the non-PSC samples, as well as a data set identifying common CNVs via 450 HapMap samples (Figure 1; Table S2; Conrad et al., 2010).

In order to identify regions of recurrent duplication, we identified regions that were duplicated in multiple samples. Analyzing all samples, and with Fisher's exact test with a p value cutoff of 0.05, yielded 152 regions where the duplications were distributed at a statistically significantly different rate between pluripotent and nonpluripotent samples (Table S5). We then filtered for regions where the fraction of pluripotent samples was >90%, which yielded 18 regions. The two duplicated segments that fit these criteria were located on chromosome 12 and chromosome 20 and are highlighted in Figure 2. The chromosome 12 region was duplicated in 9 out of 69 hESC lines, with the smallest common duplicated region encompassing *NANOGP1* and *SLC2A3* (Figure 2A). *NANOG* itself is upstream of *NANOGP1* and was duplicated in five lines. The chromosome 20 region was identified in 7 out of 69 hESC lines and 1 out of 37 hiPSC lines. In our manual curation of the data, we identified duplications of this region in two additional samples that CNVPartition failed to detect. For one (WA07P96CMD7), the population was mosaic and for the other (BG01P67), CNVPartition called duplications of regions flanking the recurrently duplication region but missed the region itself. Six of the duplications we mapped included the *DNMT3B* gene itself (Figure 2B). In two recent publications, recurrent duplications were described in the 20q11.21 region of chromosome 20 in hESCs; these reports indicated that several hESC lines had duplications in a region near the pluripotency-associated gene *DNMT3B*, which codes for a de novo DNA methyltransferase (Lefort et al., 2008; Spits et al., 2008). Mutations in this region of chromosome 20 have been noted in a number of cancers, suggesting that genetic elements in this region may be associated with hyperproliferation (Guan et al., 1996; Hurst et al., 2004; Koynova et al., 2007; Midorikawa et al., 2006; Scotto et al., 2008; Tanner et al., 1996; Tonon et al., 2005). We also found that 5 out of 69 hESC lines and 1 out of 37 hiPSC lines had duplications in this region.

The occurrence of duplications near (but not including) the pluripotency-associated genes *NANOG* and *DNMT3B* suggests that the duplication of other genes in these regions are being selected for in the cultures, or that an upstream control element for these genes may be present in the duplicated regions. In several cases, the duplication event was observed in only one of multiple samples from the same cell line collected at different times. In some instances, a more "severe" aberration was present in an earlier passage sample from the same lab (see SIVF019P53 and SIVF019P67 in Figure 2B), again reinforcing the need for detailed records regarding the passage history of cultures.

Comparison of CNVs in hESCs, hiPSCs, and Non-PSCs

For comparisons of the relative number and length of CNVs among hESCs, hiPSC, and non-PSCs, we decided to eliminate possible bias resulting from having multiple samples of some of the cell lines. For such cell lines, we included the one sample that had the largest number of total CNVs in our analysis. In addition, we removed hESC lines where preimplantation genetic diagnosis on the embryo had demonstrated that there was an aneuploidy.

Although there was considerable variation in the number of regions of CNV among the samples, overall the average numbers of regions of duplication and deletion were significantly higher

in the hiPSCs compared to the non-PSCs (Figure 3). The distribution of genomic aberrations across the hiPSC samples was rather even. In contrast, the distribution among hESC samples was highly skewed, so that although the average number of regions of duplication was not significantly higher in the hESCs than in the non-PSCs, it was clear that a subset of hESC samples contained a very large number of duplications (Figure 3).

Not including calls on the X and Y chromosomes (the CNV algorithms call a 1-copy deletion of the X for male samples and a 0-copy deletion of the Y chromosome for female samples), detected aberrations ranged in size from 0.7 to 1,791 kb (0-copy deletion), from 0.6 to 12,875 kb (1-copy deletion), and from 0.9 to 6,896 kb (3-copy duplication) (Figures S4A–S4E). The average length of 3-copy duplications was higher in hESCs and hiPSCs than in non-PSCs (Wilcoxon rank sum test p values = 1.42×10^{-15} and 5.32×10^{-5}, respectively), suggesting that either the incidence of large aberrations is higher in hPSC cultures, there is positive selection for cells with large aberrations in hPSC cultures, or there is negative selection against such cells in non-PSC cultures.
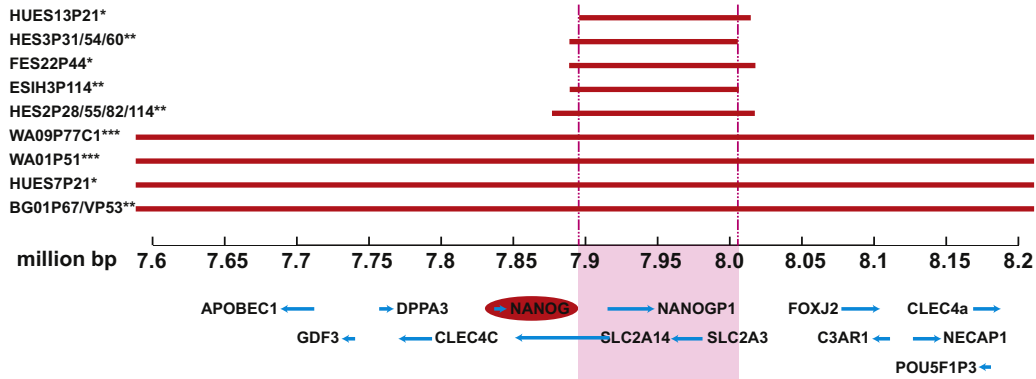
Correlation between CNVs and Data Quality or Culture Parameters

There was no correlation between the number of CNVs detected in the samples and passage number, the quality of the SNP genotyping data as measured by GenomeStudio genotyping call rate, or the Nexus quality score (Figures S4F–S4H). We did not observe a correlation between passage number or passage method and the number of aberrations, even for samples collected from the same cell line (Figures S4I–S4K). There were several very early passage hESC and hiPSC samples with large numbers of genomic aberrations, and the only noted association between passage number and the number of aberrations was in hiPSC lines that were meticulously cultured in a manner that ensured a linear path from samples collected serially during passage. In routine practice, the culture of any given line is highly branched, and investigators frequently do not know the true relationship among the various cryopreserved stocks, frozen nucleic acid samples, and live cultures for any given line. Our observations indicate that it is critical not only to record the passage number, but also the "pedigree," of each culture, in order to be able to know with certainty whether a previous assessment of the genomic stability of a line has any bearing on a current culture of that line. It is important to note that these findings do not exclude the possibility of an effect of culture conditions on genomic stability, but indicate that experiments to assess such an effect must be carefully designed and implemented.
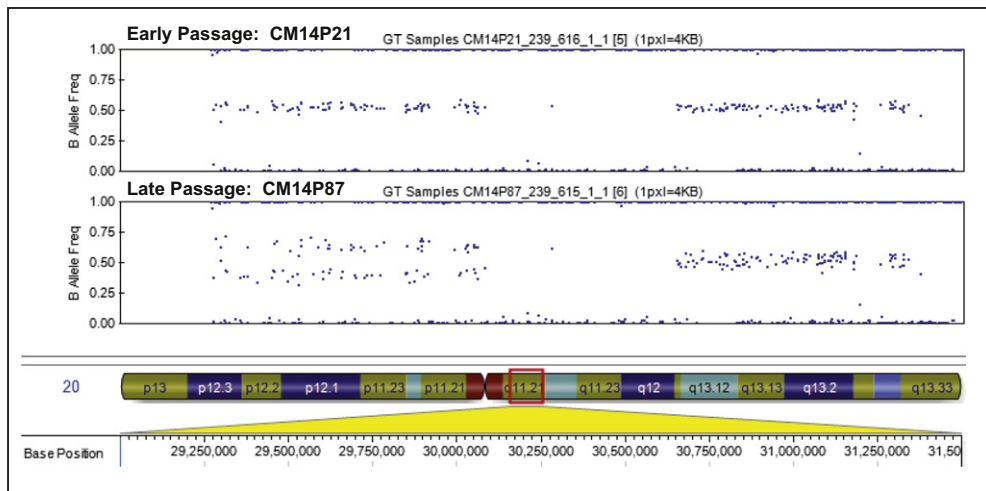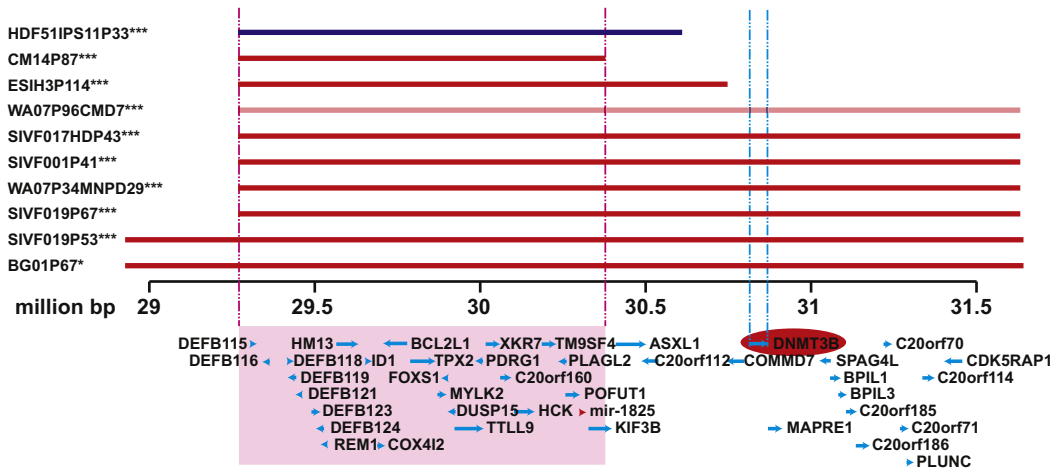
Duplications of Pseudogenes of Pluripotency-Associated Genes

Interestingly, we found a high frequency of duplications in pseudogenes of the pluripotency-associated *NANOG* and *OCT4/POU5F1* genes, including *NANOGP1* (Figure 2A). It has been noted that genes active in early embryogenesis, such as *OCT4/POU5F1*, *NANOG*, *GDF3*, and *STELLA*, tend to have many pseudogenes (Booth and Holland, 2004; Elliman et al., 2006; Liedtke et al., 2007; Pain et al., 2005). *NANOG* has an unusually large number of pseudogenes (11) of which *NANOGP1* is the only unprocessed pseudogene, retaining the

A Chromosome 12

B Chromosome 20

Early Passage: CM14P21 GT Samples CM14P21_239_616_1_1 [5] (1pxl=4KB)

Late Passage: CM14P87 GT Samples CM14P87_239_615_1_1 [6] (1pxl=4KB)

*one culture available for analysis
**multiple cultures available for analysis, all contain duplication
***multiple cultures available for analysis, only one culture contains duplication

Duplication in hESC line
Duplication hiPSC line
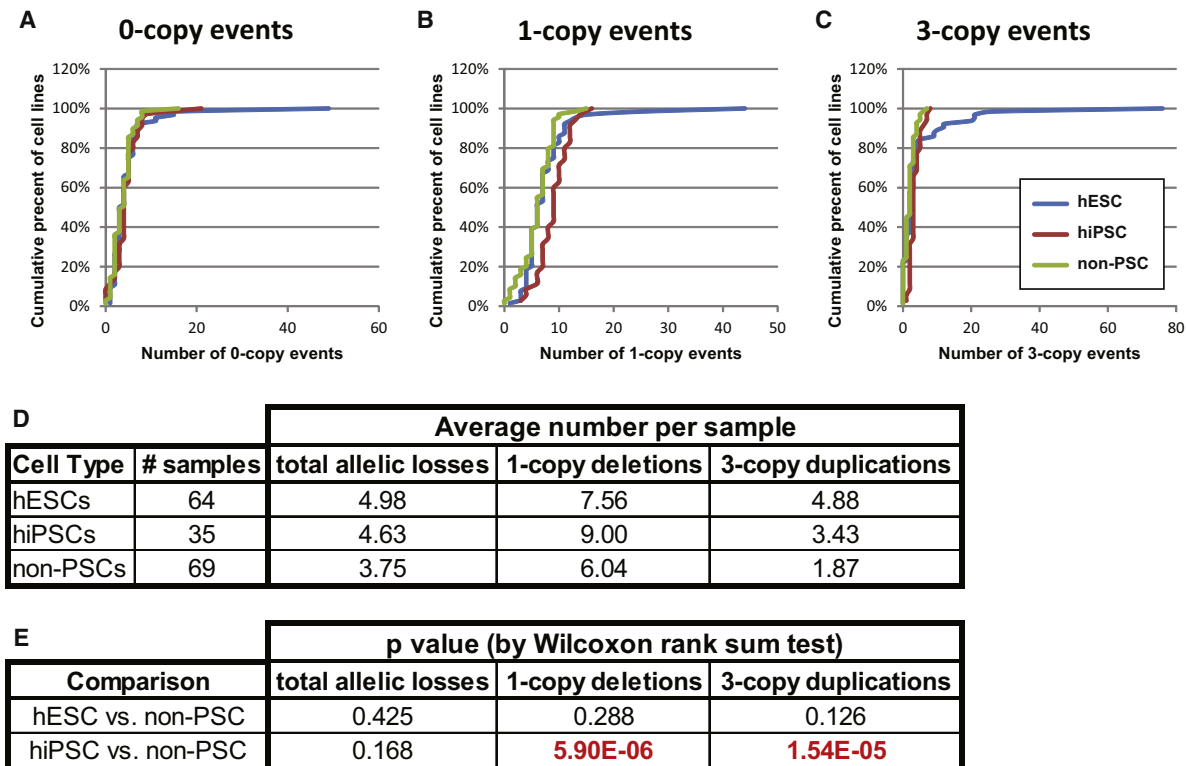
Figure 3. Number of Regions of Duplication and Deletion, as Identified by CNVPartition

(A–C) Cumulative distribution function plots of the numbers of 0-copy (total allelic loss), 1-copy, and 3-copy, and total CNVs for each sample type (hESCs, hiPSCs, and non-PSCs).

(D) Average number per sample of each type of CNV for the hESC, hiPSC, and non-PSC samples.

(E) Wilcoxon rank sum p values for each type of CNV, comparing hESC versus nonpluripotent and hiPSC versus nonpluripotent. Significant p values (<0.05) are highlighted in red.

See also Figure S4.

exon-intron structure of the coding gene. Of the other *NANOG* pseudogenes, *NANOGP4* is in the region of chromosome 7 duplicated in the FES29P39 sample, and *NANOGP8* is in the region of chromosome 15 that was duplicated in a subpopulation of the late-passage MIZ4P88 line (Figure 4A). *NANOGP9* and *NANOGP10* are on the X chromosome and were duplicated in a subpopulation of the late-passage UC06P112 sample (Figure 4B). In terms of *OCT4/POU5F1* pseudogenes, *POU5F1P4* is located on chromosome 1, which was trisomic in the WA07P95 sample; *POU5F1P6* is located in a region of chromosome 3 that is duplicated in the SIVF002P17 and the MEL2P13 samples; and *POU5F1P3* is located on chromosome 12, which was trisomic in samples from five hESC lines (Figure 3). The ESI051P37 sample is interesting, in that it possessed a large deletion that encompasses the *OCT4/POU5F1* and *NANOGP3* genes. There is little known about the role that transcribed

pseudogenes may play in cellular function. In one report (Hirotsune et al., 2003), a pseudogene was shown to stabilize the transcript of its protein-coding homolog, although its mechanism of action was unclear. It is intriguing to speculate that the pseudogenes of the pluripotency-associated genes may exert positive or negative regulatory influence over these genes.

Dynamic Changes in Genomic Structure in hPSC Populations

We observed cases where duplications appeared and took over hESC cultures. In the MIZ4 line, there was evidence that a trisomy of chromosome 15 had arisen in a subpopulation of cells between passage 33 and passage 88 (Figure 4A). In the UC06 line, the subpopulation of cells that had a trisomy of the X chromosome at passage 59 had taken over a larger proportion of the population by passage 112 (Figure 4B). These instances

Figure 2. Details of Regions of CNV on Chromosome 12 and Chromosome 20

Chromosome 12 shown in (A) and chromosome 20 shown in (B). Areas of duplication are shown in red bars for hESC samples and blue for hiPSC samples. Areas of overlap between the hPSC samples are highlighted in pink. The pluripotency-associated genes *NANOG* and *DNMT3B* are highlighted by red ovals. The dashed vertical blue lines in (B) indicate the boundaries of the *DNMT3B* gene. The lower panel of (B) shows the BAF plots demonstrating that a duplication on chromosome 20 arose during long-term passage of the hESC line CM14. See also Figures S2 and S3 and Table S5.

A hESC line MIZ4, chromosome 15
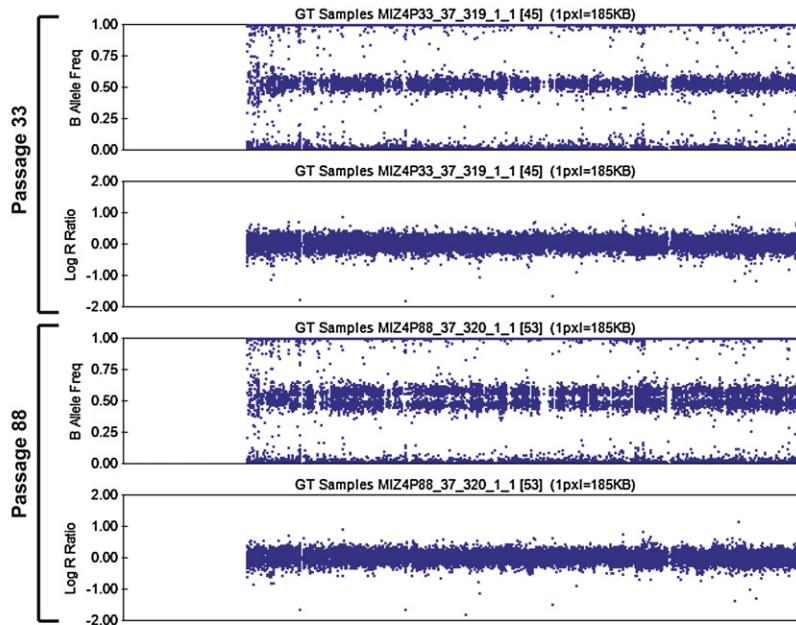
B hESC line UC06, X chromosome
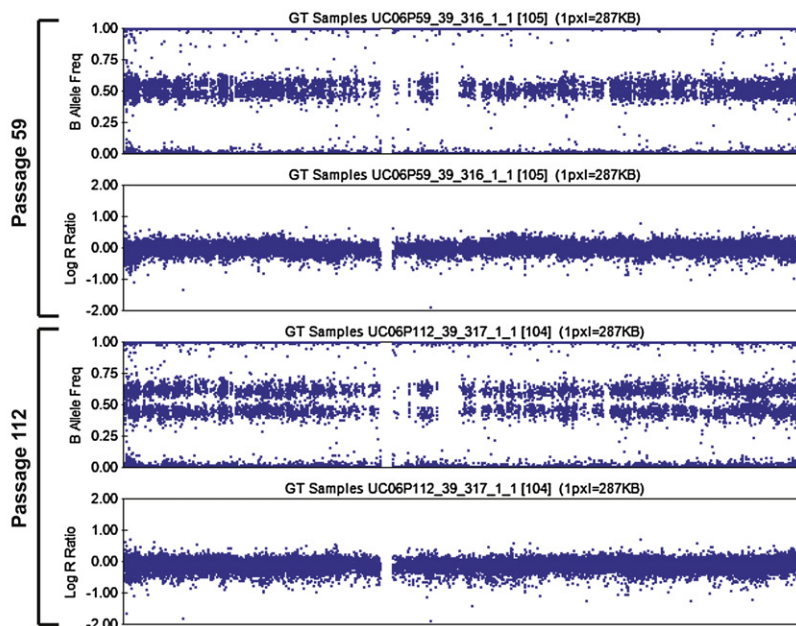
Figure 4. Dynamic Copy Number Changes over Long-Term Passage

(A) BAF and LRR plots of chromosome 15 for early- and late-passage samples of the MIZ4 hESC line. The early-passage plots show normal autosomal BAF and LRR distributions, whereas the late-passage plots indicate that a subpopulation of the cells have a duplication of chromosome 15.

(B) BAF and LRR plots of the X chromosome for early- and late-passage samples of the UC06 hESC line. There is a subtle widening of the band of heterozygous SNPs in the BAF plot for the early-passage sample, which has separated into two distinct bands in the BAF plot for the late-passage sample, indicating that the small subpopulation of cells carrying a duplication of the X chromosome in the early-passage population has outcompeted the cells without the duplication over long-term passage.

the duplication when the percentage of HDF51IPS11P33 cells was ≥70% of the cells. However, calculating BAF distance can be used to detect the presence of the duplication when ≥20% of the population is affected (Figure 5B; Figure S5A), indicating that improvements in CNV calling algorithms may be possible and would be very useful.

Genomic Aberrations during Reprogramming and Passage of hiPSCs

hiPSCs present an ideal system for distinguishing between the effects of reprogramming and passage on genomic stability. They also confer the ability to determine with certainty whether a given alteration is new, because the parental differentiated cells can also be analyzed. Accordingly, we analyzed 3 samples from a primary human fetal fibroblast line, HDF51, and 12 independent hiPSC clones generated from that line. For the hiPSC clones, we collected samples at early (passage 5–8), mid (passage 12–15), and late (passage 25–34) passage and analyzed at least the early- and late-passage samples. In addition to identifying duplications via CNVPartition, we identified deletions by using a combination of CNVPartition and replicate error

highlighted the need for improved methods for detecting CNVs in mosaic populations of cells. We analyzed mixtures of cells, where we varied the proportion of HDF51IPS11P33 cells, which contain a duplication in chromosome 20, and the parental HDF51 fibroblast line, which is genomically normal in this region. By using CNVPartition, we were able to detect the presence of

detection, which identifies the discrepancies between SNP calls from two samples (Table S6). Because all of the samples originated from the same individual, the replicate error detection represented a way of improving our confidence in our deletion calls. Inspecting the duplication and deletion calls for the HDF51 and HDF51IPS samples (Figure 6), we noticed that all 11 deletions

A

Day 2 Differentiation
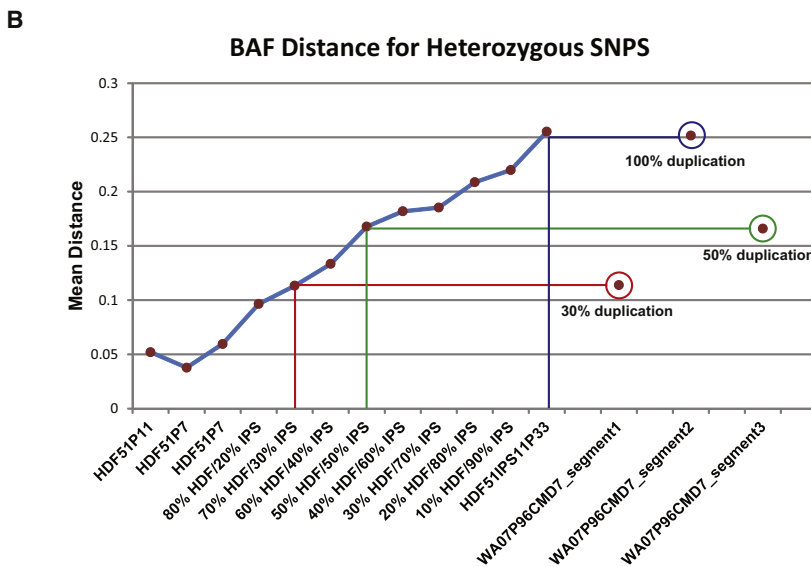
Day 7 Differentiation
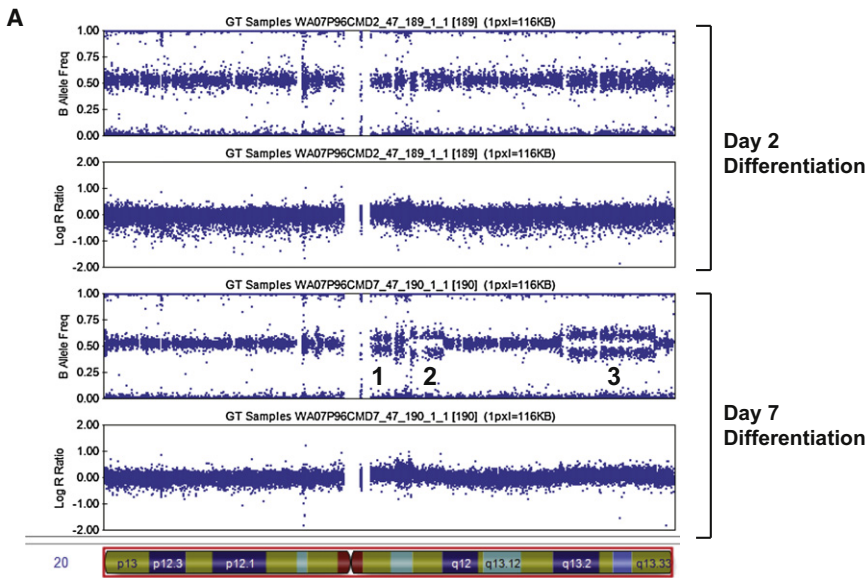
Figure 5. Duplications on Chromosome 20 Arising over a 5 Day Period during Directed Differentiation of hESCs to Cardiomyocytes
(A) The top two panels show the BAF and LRR plots at day 2 of the differentiation protocol; the bottom two panels show the plots at day 7. Three segments showing different degrees of separation of the "cloud" of BAF values for heterozygous SNPs are labeled 1, 2, and 3.
(B) The BAF distance for heterozygous SNPs are shown for regions duplicated on chromosome 20. The BAF distance for mixtures of known ratios of HDF51 cells (which have no duplication on chromosome 20) and HDF51IPS11P33 cells (which have a duplication of the proximal portion of the q-arm) are shown on the left (BAF and LRR plots are shown in Figure S4A). The BAF distances for the three partially duplicated segments (corresponding to the segments labeled 1 (red), 2 (blue), and 3 (green) in [A]) are shown on the right and have been used to estimate the percent of the population carrying the duplication.
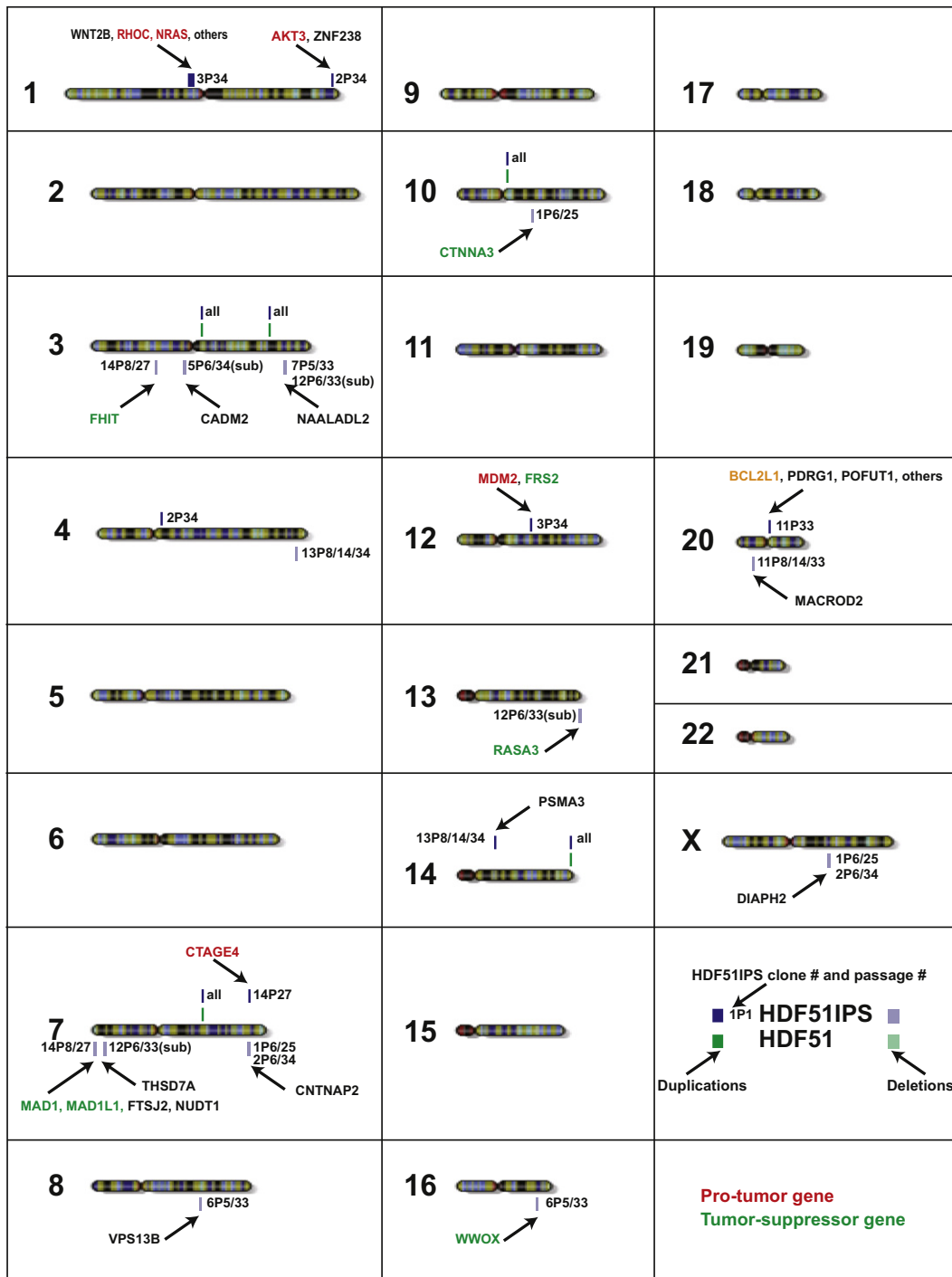See also Figure S5.

B

the other is antiapoptotic (Boise et al., 1993). All 12 deletion regions overlapped at least one gene, and 5 of them contained genes that have evidence of tumor-suppressor activity.

It is notable that the presence of the transduced copies of the reprogramming factors did not confound our analysis by appearing as duplications in the reprogramming genes. This is due to the fact that the transduced genes included only the coding sequences (which have few SNPs), and that to identify a CNV region the CNV-calling algorithms require longer stretches of consecutive SNPs to be affected.

Genomic Aberrations Arising during Differentiation
The most rapidly arising genomic aberrations in our data set were identified in samples from a directed differentiation experiment. Parallel differentiations were performed with WA07 cells at passage 95 and 96, with samples collected from the undifferentiated cells (WA07P95), on differentiation day 2 (WA07P95CMD2 and WA07P96CMD2), and differentiation day 7 (WA07P95CMD7 and WA07P96CMD7). Partial duplications of three segments of chromosome 20 were found in the WA07P96CMD7 sample only (Figure 5A; Figure S5B), indicating that they arose between day 2 and day 7 of differentiation. Comparing the BAF plots for WA07P96CMD7 to those from mixtures of known ratios of cells with and without a duplication of a smaller region of chromosome 20 (Figure 5B; Figure S5A), we estimated the percent of cells in the population carrying duplications of the three segments to be 30%, 100%, and 50%. This finding points out that differentiation can be

appeared by the earliest passage time point, whereas 5 out of 6 duplications arose during the course of long-term passage. In fact, some of the deletions receded from the population over long-term passage, suggesting that they were positively selected during reprogramming and negatively selected during passage (Figure S6).

Of the seven duplicated regions that were present in an HDF51IPS line, but absent from the parental HDF51s, six contained the coding region and/or promoter region of at least one gene. The overexpression of five of these genes (in red in Figure 6) were positively associated with tumorigenicity or cell proliferation, whereas for one (*FRS2*, in green in Figure 6), low expression has been associated with poor prognosis in non-small cell lung cancer (Iejima et al., 2010). *BCL2L1* (in orange in Figure 6) has two isoforms, one of which is proapoptotic and

Figure 6. Regions of Duplication and Deletion in the HDF51 Fibroblast Line and the HDF51IPS Lines

Duplicated regions (3 or 4 copies) (identified by CNVPartition) are shown in the dark bars above the ideogram of the chromosomes, and deleted regions (0 or 1 copy) (identified by both CNVPartition and replicate error analysis) are shown in the light bars below the ideograms of the chromosomes. The line number and passage number of the HDF51IPS line (blue) are shown adjacent to each bar, with regions where the CNV is present in only a subpopulation of the cells in a sample denoted "(sub)." HDF51 fibroblast line shown in green. The names of genes overlapping the regions of CNV are indicated.
See also Figure S6 and Table S6.

a highly selective process and that genomically aberrant cells can rapidly take over a population undergoing differentiation. We suggest that it is important to assess the genomic normality of cells frequently, not only in the pluripotent state but also at the endpoint of differentiation experiments or other treatments.

Correlations between Genomic Aberrations and Gene Expression

To determine whether the regions of frequent duplication in hESCs might have common features linked to the pluripotent phenotype, we used our large-scale mRNA expression database, which contains gene expression levels for a large number of pluripotent and nonpluripotent cell lines. We found that many of the genes in the recurrently duplicated region on chromosome 12 were more highly expressed in human pluripotent cells compared to multiple nonpluripotent cell types (Figures S2 and S3A). There was not a statistically significant difference in the expression of these genes between the hPSC samples that contained duplications and those that did not. However, this result could have been confounded by the differences in genetic background and culture conditions among the lines.

We therefore examined the expression of genes found within areas of duplication in samples in which we had genetically matched controls (Figure S3). There was higher expression of many genes on chromosome 20 in the WA07P96CMD7 samples, which had partial duplications of large stretches of this chromosome (shown in the BAF plot on the lower panel of Figure S3A), compared to the WA07P95CMD7 samples, which were euploid for this chromosome. One of the genes that was most highly affected was *DNMT3B*, as seen on the panel on the right. We noted that the higher expression was not restricted to the areas involved in the duplications, indicating potential long-range effects of chromosomal aberrations on gene expression. These effects appeared to be weaker, but still present, on other chromosomes (see chromosome 12 panel in Figure S6A). We ensured that this effect was not simply due to variations in gene expression between biological replicates by examining the corresponding data for the samples collected at day 0 and day 2 of the same experiment (upper two panels of Figure S3A). We also had matched controls for the HDF51IPS lines, and we did see correlation between gene expression and presence of duplications for these samples as well (Figure S3B). These findings suggest that duplications do result in increases in gene expression, both at the site of duplication as well as at distant sites, which can be detected when a genetically matched sample is used for comparison. Even though these gene expression changes are not apparent when comparing samples from unrelated cell lines, this is unlikely to be relevant, because a cell containing a genomic aberration is competing in culture with a population of otherwise genetically matched cells.

DISCUSSION

This study is the most comprehensive and highest-resolution study of the genomic stability of hPSCs to date and includes samples from a large number of both hESCs and hiPSCs, as well as somatic stem cells, primary cell lines, and tissues for comparison. In addition, we analyzed a primary HFF line and 12 hiPSC clones generated from it, collected at early and late passage, which allowed us to distinguish between genomic aberrations that arose during derivation versus long-term culture of hiPSCs.

This study is unique in combining a sufficient numbers of both pluripotent and nonpluripotent samples to detect cell-type-specific recurrent genomic aberrations in a statistically significant manner and a high-resolution analysis platform that enables the detection of kilobase-length aberrations. A recently published study using gene expression data to detect genomic aberrations did not include nonpluripotent samples for comparison and was limited to detection of duplications at least 10 megabases in length (Mayshar et al., 2010). In our results, >90% of duplications in hPSCs and 100% of duplications in non-hPSCs were <10 megabases (Figure S2, Table S2), indicating that gene expression-based methods are unable to detect small genomic aberrations. Moreover, the genomic locations assigned via gene expression data correspond to the location of the coding sequences of the perturbed genes, rather than the actual genomic coordinates of the genomic aberrations.

The results presented here indicate that hESC lines contain numerous genomic aberrations, most of which would not be detected by karyotyping or other microscopy-based methods. Some regions of CNV occurred multiple times in unrelated hESC and hiPSC lines, suggesting that certain changes may be characteristic of self-renewing pluripotent cells. It should be noted that it was not possible to establish with certainty the stage at which the genomic changes occurred in the hESC samples for which there was not an earlier passage sample demonstrating genomic normality; some of the abnormalities may have been present in the embryos from which the cells were derived. The analysis of hiPSCs does not suffer from this shortcoming, provided that the parental cells collected prior to reprogramming are analyzed. It is also important to consider other differences between hPSCs and cultured somatic cells. In general, because they do not undergo senescence, the hPSC lines had been in continuous culture longer than the primary cell lines, so some of the genetic changes seen may be a function of the selection pressures of cell culture in general, rather than specific to pluripotent stem cell culture.

The relatively high frequency of duplications in hPSCs raises the concern that these genetic aberrations may increase the risk of oncogenesis. The recurrent regions of copy number variation on chromosomes 12 and 20, which lie in close proximity to known pluripotency genes, are particularly worrisome, because a major issue in cell therapy is the elimination of pluripotent precursors in populations destined for transplantation. Three out of the 10 duplications on chromosome 12, and 9 out of 10 duplications on chromosome 20, developed over the course of long-term culture of hPSCs, raising the concern that expansion of pluripotent cells may inevitably lead to increased genetic abnormality. However, the *NANOG* and *NANOGP1* duplications were seen in cell lines as early as passage 21 (HUES7), 21 (HUES13), and 28 (HES2), which suggests that low passage number does not in itself ensure genetic integrity.

Our data indicate that the pattern of genomic aberrations in hiPSCs and hESCs may differ slightly, but that both cell types are prone to developing such changes, and that one of the two most significant recurrent duplications seen in hESCs, on chromosome 20, was also found in one of the hiPSC lines. The other region of recurrent duplication, encompassing the *NANOG/NANOGP1* region of chromosome 12, was detected in a late-passage hiPSC line by means of array CGH by Chin et al. (2009).

Our results and those of others (Lefort et al., 2008; Maitra et al., 2005; Mayshar et al., 2010; Spits et al., 2008; Wu et al., 2008) highlight the need for optimization of derivation and culture

conditions that promote genetic stability of pluripotent stem cells. These results also underscore the need to perform further studies that include larger numbers of pluripotent cell lines and careful phenotypic assessments in order to distinguish genetic variations that are harmless from those that pose clinical risks. The evidence for accumulation of genetic aberrations in culture of existing hPSC lines makes it clear that new hPSC lines need to be generated now and on a continuing basis, and emphasizes the necessity of frequent assessments of genomic stability in hPSC lines, both in the pluripotent state and when the cells are subjected to other potentially selective conditions, such as differentiation procedures.

EXPERIMENTAL PROCEDURES

Cell Culture
All cell types were derived and propagated as described in the references listed in Table S1. This work was approved by the Embryonic Stem Cell Research Oversight Committee at the University of California, San Diego, which oversees pluripotent stem cell research at both UCSD and TSRI.

DNA Purification
Genomic DNA was purified with the DNeasy Blood & Tissue Kit (QIAGEN).

SNP Genotyping
SNP genotyping was performed on the Illumina OmniQuad version 1, which interrogates 1,140,419 SNPs across the human genome. 1 μg input genomic DNA (the yield from approximately 200,000 cells) was amplified and labeled according to the manufacturer's instructions. The DNA was quantified with the PicoGreen reagent (Invitrogen, Inc.). The labeled product was then hybridized to the array and scanned on a BeadArray Reader (Illumina, Inc.). Genotyping calls were made with BeadStudio (Illumina, Inc.), via the standard cluster files provided by the manufacturer. The GenCall threshold was set to 0.15, and the call rates were between 0.979 and 0.999.

Copy Number Variation Assessment
For the SNP Genotyping data, data preprocessing was performed in BeadScan (Illumina, Inc.). Data cleaning, SNP calling, and replicate error identification was performed in GenomeStudio (Illumina, Inc.). CNVPartition v2.4.4 (Illumina, 2008) was used as the primary CNV-calling algorithm for the results presented in this paper. CNV regions were also identified with the SNPRank Segmentation aligorithm in Nexus (Biodiscovery, Inc.) to assess concordance between the two methods. The CNVPartition CNV score threshold was set at 50, with a minimum number of SNPs per CNV region of 10. The Nexus parameters included a significance threshold of 1×10^{-8} and a minimum number of probes per segment of 10.

We chose to remove data from probes on the array that were designated as "CNV" probes prior to using CNVPartition and Nexus. We did this for two reasons: first, the CNV probes were designed as monoallelic probes, and hence provide no B-allele-frequency information, potentially reducing their accuracy in calling duplications and deletions; second, we were interested in detecting genomic aberrations that occurred with derivation and passage of cell lines (and potentially with tissue-specific differentiation), rather than CNVs that vary among individuals and are carried in the germline, which are the ones targeted by the CNV probes.

Because the average spacing of SNPs on the SNP genotyping microarrays used was 3 kb, the shortest detectable CNV regions were expected to be approximately 30 kb. These two algorithms generally identified similar regions of duplication (97% agreement on the individual SNP level for 3-copy duplications and 99% for 1-copy deletions) (Tables S1 and S2).

Overlap between CNV Calls and Common CNVs
An overlap was identified between the CNVs called by CNVPartition in our data set and the common CNVs observed in 450 reference HapMap samples (Conrad et al., 2010) when the common region of the CNVs exceeded both 20% of the CNV identified in our samples and 20% of the common CNV in

the reference set. The CNVs in our data set that overlap with common CNVs are indicated by an asterisk in Figure 1 and were also removed from subsequent analyses.

Validation of CNV Calls
CNV calls for CNVPartition and Nexus were validated by performing qRT-PCR for a subset of the CNV calls. TaqMan CNV assays (Life Technologies, Inc.) were performed according to the manufacturer's instructions. Assays were performed in triplicate, with the HDF51IPS1P25 sample used as the reference. The predicted copy number was calculated with the equation

$$CN = 2 * \left(2^{\char`\^}(-(\text{DeltaDeltaCt})) \right).$$

Validation of SNP Calls
Because replicate errors could be identified only where samples were derived from the same original cell line, replicate error calling was performed only for the HDF51-derived lines. For these samples, SNP calls were validated by performing qPCR for a subset of the loci where replicate errors were called. TaqMan SNP assays (Life Technologies, Inc.) were performed according to the manufacturer's instructions. The HDF51P11 sample was used as the reference. There were 8 homozygous-to-homozygous replicate errors identified (0/4 tested were confirmed), 313 homozygous-to-heterozygous replicate errors (0/14 were confirmed), and 310 heterozygous-to-homozygous replicate errors (11/11 were confirmed) (Table S3). These results indicate that the large majority of apparent SNP mutations identified by replicate error analysis are in fact due to SNP genotyping error; this result is not unexpected based on reports that the discrepancy between SNP calls by sequencing and microarray-based SNP genotyping is ~0.1%–0.05% (Bentley et al., 2008). Based on the average number of heterozygous and homozygous SNPs in the SNP genotyping data (~20% heterozygous and 80% homozygous), we would have expected an excess of homozygous-to-heterozygous replicate error calls. The reason for the larger than expected number of heterozygous-to-homozygous calls was due to the fact that deletions and some duplications (when the cluster separation is poor) appear to result in replicate error calls; this is also the reason that heterozygous-to-homozygous replicate error calls are also expected to be better validated.

Calculation of BAF Distance
For intervals of interest, homozygous SNPs were removed by eliminating SNPs with BAF values >0.8 or <0.2. The heterozygous SNPs were separated into two clusters, with the median BAF value of the heterozygous SNPs as a cutoff. The "AAB" cluster had BAF values < median BAF, and the "ABB" cluster had BAF values > median BAF. The difference between the mean BAF for the AAB cluster and the mean BAF for the ABB cluster was the BAF distance.

ACCESSION NUMBERS

The microarray data are available in the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/gds) under the accession number GSE25925.

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and six tables and can be found with this article online at doi:10.1016/j.stem.2010.12.003.

REFERENCES

Atkin, N.B., and Baker, M.C. (1982). Specific chromosome change, i(12p), in testicular tumours? Lancet 2, 1349.

Baker, D.E., Harrison, N.J., Maltby, E., Smith, K., Moore, H.D., Shaw, P.J., Heath, P.R., Holden, H., and Andrews, P.W. (2007). Adaptation to culture of human embryonic stem cells and oncogenesis in vivo. Nat. Biotechnol. 25, 207–215.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59.

Boise, L.H., González-García, M., Postema, C.E., Ding, L., Lindsten, T., Turka, L.A., Mao, X., Nuñez, G., and Thompson, C.B. (1993). bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell 74, 597–608.

Booth, H.A., and Holland, P.W. (2004). Eleven daughters of NANOG. Genomics 84, 229–238.

Chin, M.H., Mason, M.J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G., Aimiuwu, O., Richter, L., Zhang, J., et al. (2009). Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell 5, 111–123.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al; Wellcome Trust Case Control Consortium. (2010). Origins and functional impact of copy number variation in the human genome. Nature 464, 704–712.

Draper, J.S., Smith, K., Gokhale, P., Moore, H.D., Maltby, E., Johnson, J., Meisner, L., Zwaka, T.P., Thomson, J.A., and Andrews, P.W. (2004). Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. Nat. Biotechnol. 22, 53–54.

Elliman, S.J., Wu, I., and Kemp, D.M. (2006). Adult tissue-specific expression of a Dppa3-derived retrogene represents a postnatal transcript of pluripotent cell origin. J. Biol. Chem. 281, 16–19.

Fox, J.L. (2008). FDA scrutinizes human stem cell therapies. Nat. Biotechnol. 26, 598–599.

Guan, X.Y., Horsman, D., Zhang, H.E., Parsa, N.Z., Meltzer, P.S., and Trent, J.M. (1996). Localization by chromosome microdissection of a recurrent breakpoint region on chromosome 6 in human B-cell lymphoma. Blood 88, 1418–1422.

Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A., and Yoshiki, A. (2003). An expressed pseudo-gene regulates the messenger-RNA stability of its homologous coding gene. Nature 423, 91–96.

Hurst, C.D., Fiegler, H., Carr, P., Williams, S., Carter, N.P., and Knowles, M.A. (2004). High-resolution analysis of genomic copy number alterations in bladder cancer by microarray-based comparative genomic hybridization. Oncogene 23, 2250–2263.

Iejima, D., Minegishi, Y., Takenaka, K., Siswanto, A., Watanabe, M., Huang, L., Watanabe, T., Tanaka, F., Kuroda, M., and Gotoh, N. (2010). FRS2beta, a potential prognostic gene for non-small cell lung cancer, encodes a feedback inhibitor of EGF receptor family members by ERK binding. Oncogene 29, 3087–3099.

Illumina (2008). Technical Note: DNA Copy Number Analysis Algorithms. http://wwwilluminacom/downloads/CNValgorithms_TechNotepdf.

Imreh, M.P., Gertow, K., Cedervall, J., Unger, C., Holmberg, K., Szöke, K., Csöregh, L., Fried, G., Dilber, S., Blennow, E., and Ahrlund-Richter, L. (2006). In vitro culture conditions favoring selection of chromosomal abnormalities in human ES cells. J. Cell. Biochem. 99, 508–516.

Koynova, D.K., Jordanova, E.S., Milev, A.D., Dijkman, R., Kirov, K.S., Toncheva, D.I., and Gruis, N.A. (2007). Gene-specific fluorescence in-situ hybridization analysis on tissue microarray to refine the region of chromosome 20q amplification in melanoma. Melanoma Res. 17, 37–41.

Kresse, S.H., Szuhai, K., Barragan-Polania, A.H., Rydbeck, H., Cleton-Jansen, A.M., Myklebost, O., and Meza-Zepeda, L.A. (2010). Evaluation of high-resolution microarray platforms for genomic profiling of bone tumours. BMC Res Notes 3, 223.

Lefort, N., Feyeux, M., Bas, C., Féraud, O., Bennaceur-Griscelli, A., Tachdjian, G., Peschanski, M., and Perrier, A.L. (2008). Human embryonic stem cells reveal recurrent genomic instability at 20q11.21. Nat. Biotechnol. 26, 1364–1366.

Liedtke, S., Enczmann, J., Waclawczyk, S., Wernet, P., and Kögler, G. (2007). Oct4 and its pseudogenes confuse stem cell research. Cell Stem Cell 1, 364–366.

Maitra, A., Arking, D.E., Shivapurkar, N., Ikeda, M., Stastny, V., Kassauei, K., Sui, G., Cutler, D.J., Liu, Y., Brimble, S.N., et al. (2005). Genomic alterations in cultured human embryonic stem cells. Nat. Genet. 37, 1099–1103.

Mayshar, Y., Ben-David, U., Lavon, N., Biancotti, J.-C., Yakir, B., Clark, A.T., Plath, K., Lowry, W.E., and Benvenisty, N. (2010). Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. Cell Stem Cell 7, 521–531.

Midorikawa, Y., Yamamoto, S., Ishikawa, S., Kamimura, N., Igarashi, H., Sugimura, H., Makuuchi, M., and Aburatani, H. (2006). Molecular karyotyping of human hepatocellular carcinoma using single-nucleotide polymorphism arrays. Oncogene 25, 5581–5590.

Mitalipova, M.M., Rao, R.R., Hoyer, D.M., Johnson, J.A., Meisner, L.F., Jones, K.L., Dalton, S., and Stice, S.L. (2005). Preserving the genetic integrity of human embryonic stem cells. Nat. Biotechnol. 23, 19–20.

Närvä, E., Autio, R., Rahkonen, N., Kong, L., Harrison, N., Kitsberg, D., Borghese, L., Itskovitz-Eldor, J., Rasool, O., Dvorak, P., et al. (2010). 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.

Pain, D., Chirn, G.W., Strassel, C., and Kemp, D.M. (2005). Multiple retropseudogenes from pluripotent cell-specific gene expression indicates a potential signature for novel gene identification. J. Biol. Chem. 280, 6265–6268.

Rodriguez, E., Houldsworth, J., Reuter, V.E., Meltzer, P., Zhang, J., Trent, J.M., Bosl, G.J., and Chaganti, R.S. (1993). Molecular cytogenetic analysis of i(12p)-negative human male germ cell tumors. Genes Chromosomes Cancer 8, 230–236.

Scotto, L., Narayan, G., Nandula, S.V., Arias-Pulido, H., Subramaniyam, S., Schneider, A., Kaufmann, A.M., Wright, J.D., Pothuri, B., Mansukhani, M., and Murty, V.V. (2008). Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in

cervical cancer: Potential role in progression. Genes Chromosomes Cancer *47*, 755–765.

Skotheim, R.I., Monni, O., Mousses, S., Fosså, S.D., Kallioniemi, O.P., Lothe, R.A., and Kallioniemi, A. (2002). New insights into testicular germ cell tumorigenesis from gene expression profiling. Cancer Res. *62*, 2359–2364.

Spits, C., Mateizel, I., Geens, M., Mertzanidou, A., Staessen, C., Vandeskelde, Y., Van der Elst, J., Liebaers, I., and Sermon, K. (2008). Recurrent chromosomal abnormalities in human embryonic stem cells. Nat. Biotechnol. *26*, 1361–1363.

Tanner, M.M., Tirkkonen, M., Kallioniemi, A., Isola, J., Kuukasjärvi, T., Collins, C., Kowbel, D., Guan, X.Y., Trent, J., Gray, J.W., et al. (1996). Independent amplification and frequent co-amplification of three nonsyntenic regions on the long arm of chromosome 20 in human breast cancer. Cancer Res. *56*, 3441–3445.

Tonon, G., Wong, K.K., Maulik, G., Brennan, C., Feng, B., Zhang, Y., Khatry, D.B., Protopopov, A., You, M.J., Aguirre, A.J., et al. (2005). High-resolution genomic profiles of human lung cancer. Proc. Natl. Acad. Sci. USA *102*, 9625–9630.

Wu, H., Kim, K.J., Mehta, K., Paxia, S., Sundstrom, A., Anantharaman, T., Kuraishy, A.I., Doan, T., Ghosh, J., Pyle, A.D., et al. (2008). Copy number variant analysis of human embryonic stem cells. Stem Cells *26*, 1484–1489.