Cell Stem Cell
Short Article

Identification and Classification of Chromosomal Aberrations in Human Induced Pluripotent Stem Cells

Yoav Mayshar,[1,7] Uri Ben-David,[1,7] Neta Lavon,[1] Juan-Carlos Biancotti,[2] Benjamin Yakir,[3] Amander T. Clark,[4,5] Kathrin Plath,[5,6] William E. Lowry,[4,5] and Nissim Benvenisty[1,*]
[1]Department of Genetics, Silberman Institute of Life Sciences, The Hebrew University, Jerusalem 91904, Israel
[2]Regenerative Medicine Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
[3]Department of Statistics, The Hebrew University, Jerusalem 91905, Israel
[4]Department of Molecular Cell and Developmental Biology, University of California, Los Angeles, CA 90095, USA
[5]Broad Stem Cell Center, University of California, Los Angeles, CA 90095, USA
[6]Department of Biological Chemistry, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA
[7]These authors contributed equally to this work
*Correspondence: nissimb@cc.huji.ac.il
DOI 10.1016/j.stem.2010.07.017

SUMMARY

Because of their somatic cell origin, human induced pluripotent stem cells (HiPSCs) are assumed to carry a normal diploid genome, and adaptive chromosomal aberrations have not been fully evaluated. Here, we analyzed the chromosomal integrity of 66 HiPSC and 38 human embryonic stem cell (HESC) samples from 18 different studies by global gene expression meta-analysis. We report identification of a substantial number of cell lines carrying full and partial chromosomal aberrations, half of which were validated at the DNA level. Several aberrations resulted from culture adaptation, and others are suspected to originate from the parent somatic cell. Our classification revealed a third type of aneuploidy already evident in early passage HiPSCs, suggesting considerable selective pressure during the reprogramming process. The analysis indicated high incidence of chromosome 12 duplications, resulting in significant enrichment for cell cycle-related genes. Such aneuploidy may limit the differentiation capacity and increase the tumorigenicity of HiPSCs.

INTRODUCTION

Numerous studies have thus far demonstrated that induction of a small number of genes is sufficient to convert normal human somatic cells into pluripotent stem cells (HiPSCs), which very closely resemble human embryonic stem cells (HESCs) (Lowry et al., 2008; Park et al., 2008; Takahashi et al., 2007; Yu et al., 2007). This similarity has been demonstrated by gene expression profiling and epigenetic signatures as well as by differentiation potential. The expected impact of cellular reprogramming on the future of medicine cannot be overestimated given that it promises to provide an unlimited source of patient specific cells for replacement therapy and for the study of genetic diseases. However, the benefits of HiPSCs could be jeopardized by safety concerns such as their tumorigenicity. Although much effort has been recently made to reduce potential hazardous effects of the reprogramming vectors, there is a need for a thorough analysis of the reprogrammed cells' propensity for chromosomal aberrations. Thus far, no major aneuploidy has yet been demonstrated to occur in HiPSCs. In comparison, although HESCs were initially considered to have a stable normal genome copy number, it is now widely accepted that during long-term culture of the cells they stochastically acquire chromosomal aberrations that may confer a growth advantage so that the aneuploid cells quickly take over the population (Baker et al., 2007). Such progressive adaptive copy number aberrations are most commonly identified in chromosomes 12, 17 and X (Baker et al., 2007; Draper et al., 2004). So far, studies describing HiPSCs characterized them at relatively early passage numbers and mainly reported normal karyotypes (Lowry et al., 2008; Park et al., 2008; Takahashi et al., 2007; Yu et al., 2007).

The genetic integrity of ESCs and iPSCs was previously analyzed mostly by karyotype, but also at high resolution with comparative genomic hybridization (CGH) or single-nucleotide polymorphism (SNP) arrays. Such analyses directly measure either DNA content or chromosome morphology. We now suggest that the chromosomal integrity of HESCs and HiPSCs can be examined by transcriptional profiling, to identify genomic regions containing large clusters of genes with significantly higher or lower levels of gene expression. This is possible by comparison of each individual gene expression profile to a reference baseline made up of a very large number of highly similar cell lines. In recent years the correlation between copy number and levels of gene expression has been extensively recognized, primarily in cancers, but also in natural human and rodent populations (Guryev et al., 2008; Henrichsen et al., 2009; Hughes et al., 2000; Phillips et al., 2001; Pollack et al., 2002; Schoch et al., 2005; Stranger et al., 2007; Tsafrir et al., 2006). It has been further suggested that regions with biased gene expression are correlated with chromosomal abnormalities (Crawley and Furge, 2002; Furge et al., 2005; Hertzberg et al., 2007; Lilljebjörn et al., 2007; Masayesva et al., 2004; Pollack et al., 2002; Tsafrir et al., 2006). The ability to detect genomic aberrations with data from gene expression arrays enables a comprehensive study of genomic aberrations in multiple lines from various

laboratories. Such analysis cannot be currently performed by direct DNA analysis, given that much of the biological material is not available. Here, we applied two complementary tests to identify regions of presumed aneuploidy in pluripotent stem cells on the basis of their global gene expression profile.

RESULTS

Data Collection and Analysis

Gene expression data consisted of 104 unique gene expression profiles from 18 studies; 38 samples of 17 unique HESC lines and 66 samples (clones and subclones) of 46 unique HiPSC lines originating from 17 independent somatic cell lines (See Tables S1 and S2 available online). Analysis was performed on expressed genes with known chromosomal location only. Data were further filtered to retain only a single instance of each autosomal gene, resulting in a total of 12,054 data points. For each sample, the expression value of each gene was divided by the median of the same gene across the entire dataset, in order to obtain a comparative value. These values were then used in two statistical tests in order to assess aneuploidy. In the first test, overexpressed genes were determined for each sample (>1.5 fold, relative to the median) and then subjected to location enrichment analysis, using two gene expression analysis software: Expander (Sharan et al., 2003) and EASE (Hosack et al., 2003). In the second test, we applied the processed expression data to comparative genomic hybridization (CGH) analysis software, CGH-Explorer (Lingjaerde et al., 2005). Using the program's piecewise constant fit (PCF) algorithm, we were able to analyze gene expression regional bias in all our samples by using a constant set of parameters. The results of the PCF can often be clearly seen in a moving average plot of the gene expression profile. It is important to note that although both methods rely on the same expression data, they are complementary. The first method counts genes that are clearly overexpressed and tests for chromosomal enrichment, whereas the second method analyzes the spatial expression pattern of all expressed genes.

Identification and Validation of Aneuploidy in HESCs with Expression Data

To explore the potential of this system to identify aneuploidy, we first analyzed two HESC lines predetermined by preimplantation genetic screening (PGS) to harbor a single trisomic chromosome each, either 17 or 21. These trisomies were verified by high-density SNP-array copy number analysis (Figure 1A). Gene expression profiling was performed for these cell lines in parallel with the genomic analysis. The trisomic chromosomes (17 and 21) were found to have highly significant over-representation of overexpressed genes (Bonferroni corrected p values = 1×10^{-7} by Expander and 2.1×10^{-9} by EASE, for chromosome 17; 4.9×10^{-4} by Expander and 1.1×10^{-10} by EASE, for chromosome 21), (Figure 1B and Figure S1A). These results were verified by the PCF algorithm and are also clearly visible with a moving average plot relative to similarly derived normal diploid cells (Figure 1C).

Most aneuploidies in HESCs were identified in culture-adapted cells (Baker et al., 2007; Draper et al., 2004). Such cytogenetically verified chromosomes and chromosome arms were correctly identified by enrichment analysis of the overexpressed genes (Figures 1D and 1E and Figures S1B–S1D). The entire set of identified trisomies analyzed by the expression profile of HESCs is shown in Table S1. Next, we explored the possibility of using PCF to detect subchromosomal gains or losses in the HESC data set. This method could detect both cytogenetically verified multiple copy amplification (Figure 1F) and single-copy duplication (Figure 1G). Analysis of the entire HESC data set revealed chromosomal aberrations in 12 out of the 38 cell lines. Eight of these aberrations involved entire chromosomes or large chromosomal regions that were thus apparent by both expression analyses, three of which could be confirmed by karyotype (Table S1). Not all the aberrations identified by gene expression could be analyzed cytogenetically because of lack of corresponding cells or DNA from similar passage. Notably, the most recurrent duplications were in chromosomes 12 and 17 (three cell lines each), coinciding with their previously suggested role in embryonic tumors and ESC adaptation (Baker et al., 2007; Reuter, 2005) (Table S1, Figures 1D–1F, and Figures S1B–S1D).

Aneuploidy Detection in HiPSCs with Expression Data

HiPSCs are similar to HESCs in many respects, even though some differences have been observed between the two cell types (Chin et al., 2009). Although it might be presumed that HiPSCs would acquire genetic abnormalities in a similar manner to HESCs, the unique nature of their derivation may make them different in this regard. We thus aimed to analyze the genetic instability of HiPSCs. Of the 66 HiPSC samples included in this study, two lines were previously reported to harbor two trisomic chromosomes each (chromosomes 1 and 9). Using the same methods and parameters as for HESCs, we correctly identified these aberrations (Figures 2A and 2B). Two other HiPSC studies reported subchromosomal genomic aberrations. In the study of Yu et al. (2009), 16 HiPSC clones and subclones were determined to be normal by karyotype and a single small deletion in one subclone was identified by high-density CGH arrays (Yu et al., 2009). Here, by examining the corresponding gene expression data, we identified the same small deletion in chromosome 15 as the sole aneuploidy in the entire set of samples from this study (Figure 2C). In another study (Chin et al., 2009), CGH-array analysis revealed a number of possible aberrations at varying confidence levels, the three most significant of which were correctly reproduced by our gene expression analysis (Figures S2A–S2C). However, the events that were suggested by CGH at low confidence were not detected by this analysis. This could suggest that gene expression-based analysis might not be adequate to detect changes in a minority of cells within a mixed population.

In addition to the verification of previously reported genetic abnormalities, we could identify many chromosomal aberrations that were established early on during the isolation of HiPSCs, or during their prolonged culture. As seen in Figure 2D, the HiPSC line, HiPSC1-8 (Masaki et al., 2007), contains an abnormally high frequency of overexpressed genes in chromosome 12, which is present already at passage 14, and the level of overexpressed genes is even higher at passage 31. The evidence of trisomy 12 is also very clear in the respective moving average plots, in which the aneuploidy is evident at low passage and takes over the culture at the higher passage (Figure 2E). Another

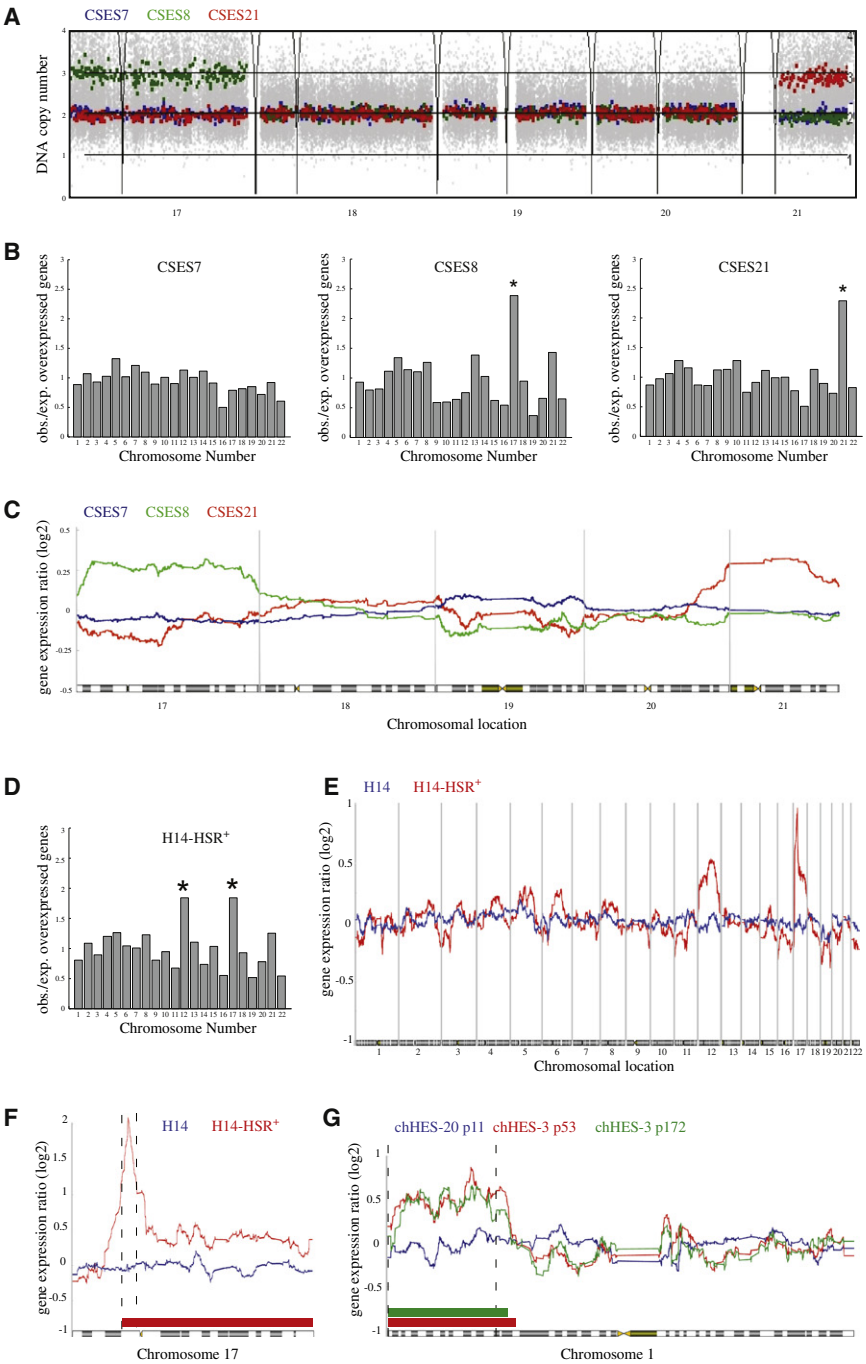

Figure 1. Identification of Chromosomal Aberrations in HESCs with Gene Expression Analysis

(A–C) Identification of aneuploidy in preimplantation genetic screening (PGS) derived HESC lines:

(A) High density SNP-array copy number analysis of: CSES7 (46XX); CSES8 (47XX+17); and CSES21 (47XY+21).

(B) Whole chromosome gain analysis of overexpressed genes. Bars represent fold enrichment of overexpressed genes in each particular chromosome relative to the expected random frequency. CSES8 shows significant enrichment of overexpressed genes in chromosome 17 (Bonferroni corrected p value = $1*10^{-7}$ by Expander and 2.1×10^{-9} by EASE), CSES21 shows significant enrichment in chromosome 21 (Bonferroni corrected p value = 4.9×10^{-4} by Expander and $1.1*10^{-10}$ by EASE).

(C) Gene expression profile moving average plot demonstrates homogenous overexpression of genes along the abnormal chromosomes.

(D–G) Identification of aneuploidy in culture adapted HESCs.

(D) Identification of trisomies 12 and 17 in the H14-HSR$^+$ cell line (Baker et al., 2007) by whole-chromosome analysis (Bonferroni corrected p value = 5.5×10^{-18} by Expander and 3.6×10^{-18} by EASE, for chromosome 12; 2.6×10^{-18} by Expander and 1.5×10^{-12} by EASE, for chromosome 17), but not in the normal parental H14 cell line (Table S2).

(E) The same trisomies shown by moving average plot.

(F) Tandem multiple copy number gain in proximal 17p in the adapted cell line H14-HSR$^+$.

(G) Single copy number gain in distal 1p in the chHES-3 cell line (Yang et al., 2008). Two samples of different passages carrying the same aberration 46XX,dup(1)(p32p36) relative to a normal cell line from the same study are shown.

Asterisks indicate p value $< 1 \times 10^{-4}$, judged by Expander and EASE location analyses. All significance tests are presented after Bonferroni multiple test correction. Vertical dashed lines represent cytogenetic boundaries of chromosomal aberrations as described in the respective studies. Horizontal colored bars represent piecewise constant fit (PCF) abnormality detection as described in the methods section.

For further examples of detection of aneuploidy in HESCs, see also Figure S1. For a detailed list of HESC lines, see Table S1.

cell line, hiPSC18 (Chin et al., 2009; Lowry et al., 2008), was found normal by gene expression analysis at passage 9, whereas by passage 56 both PCF and Expander/EASE analyses predicted gains in chromosomes 3 and 12 (Figures S2D and S2E).

Because chromosomal duplications can occur in some cell lines after a relative small number of passages, as described above, we speculated it might be possible to enhance or generate them de novo by growing these cells in culture. Therefore, we decided to re-examine the hiPSC18 line (Chin et al., 2009; Lowry et al., 2008), which seemed susceptible to chromosomal aberrations. This cell line had a normal karyotype at passage 45 (20/20 metaphases, Figure 3A) and was further grown until passage 63, when it was analyzed in parallel by karyotype and gene expression. Interestingly, these cells acquired a full trisomy of chromosome 12, as demonstrated by karyotype analysis (Figure 3C). Moreover, both the Expander/EASE and the PCF analyses correctly identified the trisomy while not identifying any false duplication or deletion (Figures 3D–3F). We can thus suggest that this trisomy may confer a major growth

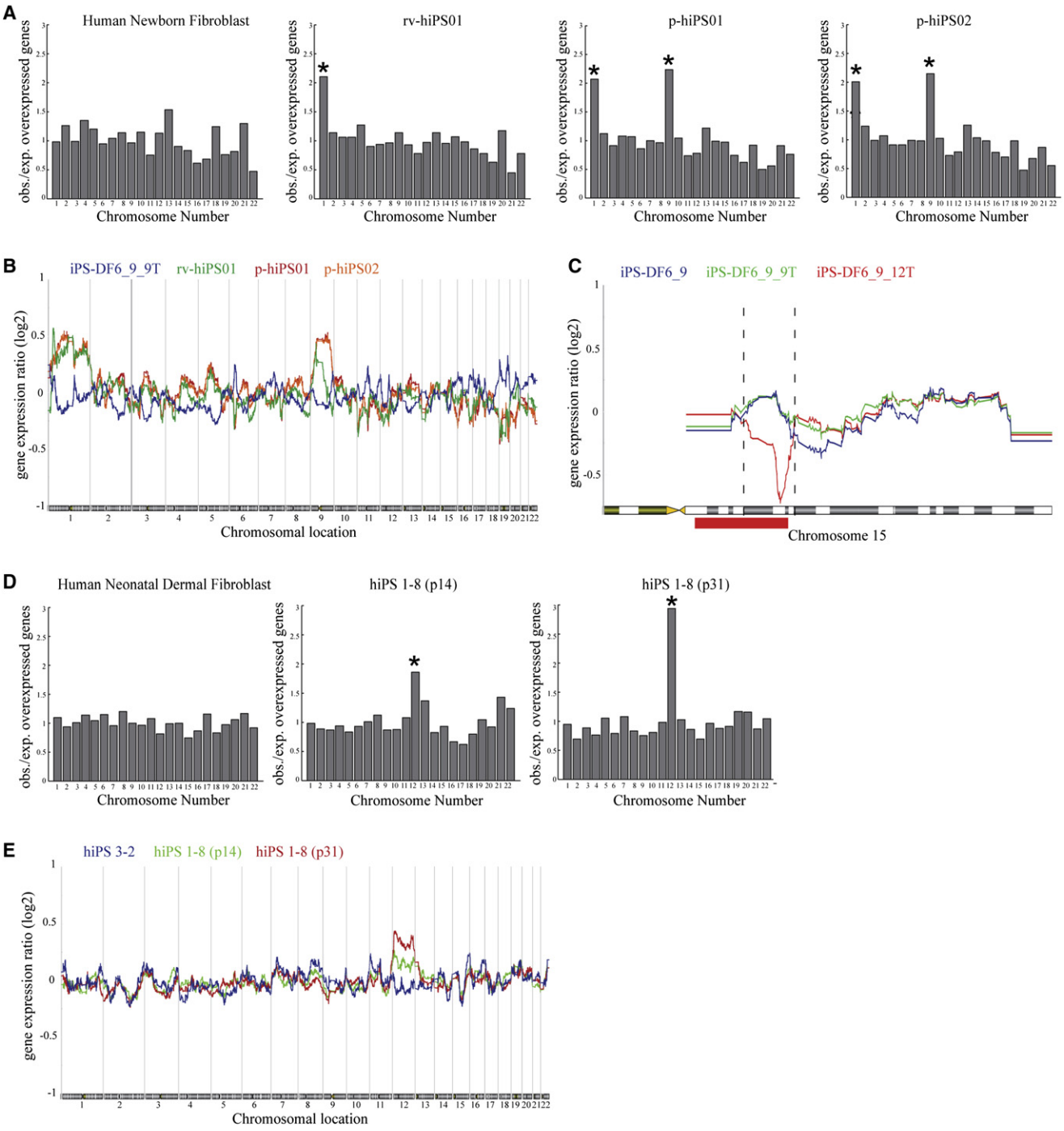

Figure 2. Identification of Chromosomal Aberrations in HiPSCs

(A) Detection of chromosome 1 and chromosome 9 trisomies in p-hiPS01 (Bonferroni corrected p values = 5.0×10^{-32} by Expander and 1.1×10^{-30} by EASE, for chromosome 1; 6.0×10^{-13} by Expander and 3.2×10^{-18} by EASE, for chromosome 9) and p-hiPS02 (Bonferroni corrected p values = 4.0×10^{-33} by Expander and 4.5×10^{-44} by EASE, for chromosome 1; 2.7×10^{-12} by Expander and 2.2×10^{-16} by EASE, for chromosome 9), as well as detection of chromosome 1 and chromosome 9p trisomies in rv-hiPS01 (Bonferroni corrected p values = 1.3×10^{-30} by Expander and 5.0×10^{-46} by EASE, for chromosome 1; 1.4×10^{-7} by Expander and 8.7×10^{-13} by EASE, for chromosomal arm 9p). The original study described trisomy in chromosomes 1 and 9 in p-hiPS01 and p-hiPS02 as well as in the parental somatic cell line (Kim et al., 2009).

(B) Moving average plot of the gene expression profile of these cell lines; iPS-DF6_9 is displayed as reference.

(C–E) Identification of aneuploidy acquired in culture.

(C) Identification of a small deletion described by Yu et al. (2009) in the subclone iPS-DF6_9_12T. All the other clones described in this study were found in our analysis to be normal, congruent with the original report. The parental HiPSC line iPSC-DF6_9 and another subclone isolated from this line (iPS-DF6_9_9T) are presented as normal reference.

advantage, as is evident from the rapid rate in which it is acquired, coinciding with the high frequency at which this particular aberration has been found in our study (Figure 4A). In order to learn more about the selection for the aneuploid cells, we performed karyotype analysis at passage 58 of the same line (five passages prior to the full trisomy detection). At this passage, the culture was found to be mosaic, comprising both normal and trisomic cells (47,XY,+12[9]/46,XY[11], Figure 3B). These results suggest a rapid selection for cells with trisomy 12, as was previously shown in HESCs for trisomy 17 (Olariu et al., 2010).

All together, of the 66 HiPSC samples analyzed, we identified thirteen (~20%) abnormal lines, of which six (~9%) carried at least one full trisomy (Table S2). Cytogenetic information from similar passage was available for nine of the aberrations and all agreed with these findings (Table S2). Thus, we found many more aberrations than were reported in the original studies. This may be due to the higher resolution of the present analysis, but in some cases can also be attributed to the fact that in many cases gene expression and cytogenetic analyses presented by the original studies were conducted at different passages. The complete list of genomic aberrations identified in HiPSCs is presented as Table S2. A schematic representation of the PCF results is presented in Figure 4A, and gene level PCF calls are presented in Tables S3A and S3B.

Adaptation of HiPSCs Is Associated with Elevated Expression of Pluripotent and Cell Cycle-Related Genes

Investigation of the frequency of copy number gains in HiPSCs demonstrates that chromosome 12 (and specifically 12p) is the most recurrent (Figure 4A). This abnormality has been described as a hallmark of testicular germ cell tumors (Reuter, 2005). The most recurrent autosomal gains in abnormal culture adapted HESCs are of chromosomes 12 and 17 (Baker et al., 2007). Interestingly, here we did not find a single instance of gain in chromosome 17 in HiPSCs, whereas three were found in the smaller HESC data set (Figure 4A). In all five HiPSC lines that acquired a gain in the 12p region as a result of prolonged time in culture, the hallmark pluripotency genes *NANOG* and *GDF3* are included in the duplication (Figure 4B). Moreover, *NANOG* and *GDF3* are significantly overexpressed in these cell lines relative to all other HiPSC lines (p values = 4.3×10^{-5} and 2.9×10^{-10}, respectively; average fold change ×1.6 and ×4.4, respectively) (Figures 4C–4E). The higher levels of these genes seem to be directly correlated with copy number gain because other pluripotency genes, such as *OCT4* (p value = 0.38) and *SOX2* (p value = 0.95), are not differentially expressed. The increase in expression of *NANOG* and *GDF3* upon the selection for cells with trisomy 12 was confirmed in the hiPSC18 line by quantitative RT-PCR. In this line, a significant gradual increase in the expression of these genes was measured between the normal culture, the mosaic culture, and the fully trisomic culture (Figure 4D).

Functional analysis of the expressed genes in recurring aberrant chromosomal regions in HiPSCs was performed with the DAVID Functional Annotation tool (Dennis et al., 2003; Huang et al., 2009). Expressed genes in chromosomal regions gained in at least two independent HiPSC lines were significantly enriched for cell cycle annotations (p value = 2.8×10^{-5}) (Table S3C). Such enrichment was not found in randomly chosen chromosomal regions of the same size.

In order to further investigate the functional implications of the recurring aberrant chromosomal regions, we examined the effect of gains in chromosome 12 on the gene expression profile of the entire genome. A total of 135 genes residing outside chromosome 12 were found to be significantly differentially expressed in the aberrant HiPSC lines that carry gains in chromosome 12, relative to all the other HiPSC lines (p value < 0.01). However, these genes were not significantly enriched for any functional annotation. In order to determine whether the trisomic lines more closely resemble one another in terms of overall gene expression, we performed unsupervised hierarchical clustering of the entire data set of HESC and HiPSC lines (HG-U133plus2 platform). Although it is evident that the aberrant lines show consistently high gene expression levels on the aberrant chromosomes, they do not cluster together. The main contribution to the way the samples cluster together seems to be the laboratory and study of origin (Figure S3).

Sensitivity and Specificity of Statistical Tests

In order to determine the success rate of our analysis, we analyzed the gene expression profiles of control and trisomic HESC lines isolated from aneuploid PGS embryos. These lines were cytogenetically analyzed by high-density SNP arrays and karyotype at the same passage of RNA extraction. In the three lines shown in Figure 1, we did not find a single false positive event in which a diploid chromosome was identified as trisomic. Because we do not expect trisomies that are established within the cell culture to disappear, all known trisomies from the study (both in HESCs and in HiPSCs) were used for assessing the false negative rate. All such duplication events (of whole chromosomes or chromosome arms) were identified by both Expander/EASE and PCF analyses of the expression profile (11/11). Because of the tendency of cultured cells to acquire de novo genetic alterations, for verification purposes it is imperative to analyze gene expression data alongside cytogenetic data obtained from very similar passages. Thus, in order to further validate our methods, we used the lines from Yu et al. (2009). In this study, 16 low passage HiPSC clones and subclones were subjected to both array CGH and gene expression analyses at very close passages. Out of these 16 HiPSC samples, our analysis detected only one small deletion in a single

(D and E) Acquired chromosome 12 trisomy in two separately grown samples of HiPSC 1-8 from Masaki et al. (2007).
(D) The parental somatic cell and HiPSC 1-8 from different passages (14 and 31), showing trisomy in chromosome 12 (Bonferroni corrected p-values = 6.0×10^{-10} by Expander and $1.0*10^{-18}$ by EASE, for passage 14; $5.0*10^{-36}$ and $1.3*10^{-26}$, for passage 31).
(E) Moving average plot of the gene expression profile of clone 1-8 at passages 14 and 31, clone 3-2 is presented as normal reference. These results suggest a trisomy of chromosome 12, which was acquired very early in culture and gradually took over.
Asterisks indicate p value < $1*10^{-4}$, judged by Expander and EASE location analyses after Bonferroni multiple test correction.
For further examples of detection of aneuploidy in HiPSCs, see also Figure S2. For a detailed list of HiPSC lines, see Table S2.

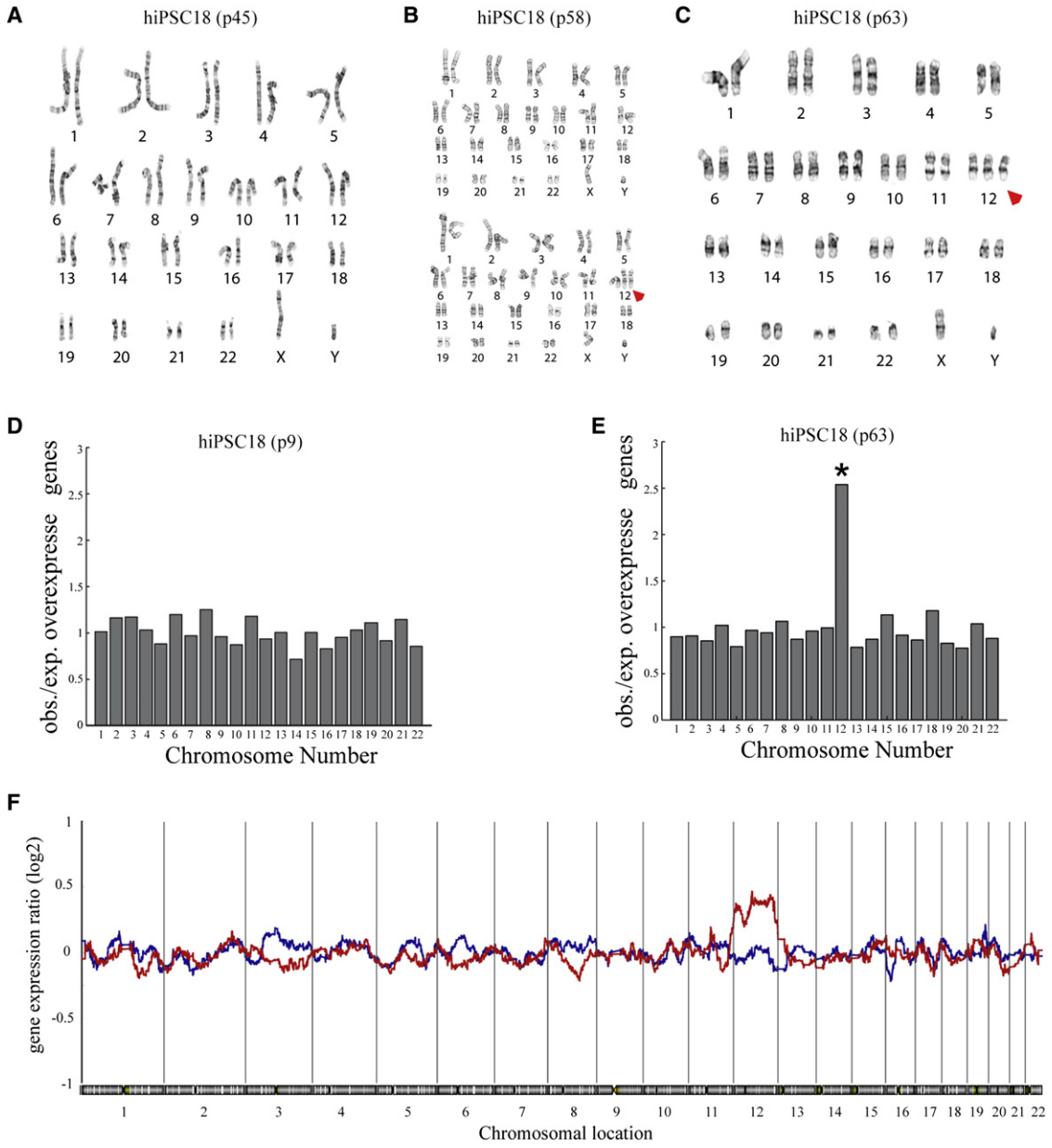

Figure 3. Generation of Trisomy 12 in HiPSCs upon their Growth in Culture

(A) Karyotype analysis of hiPSC18 at passage 45. This cell line was also found normal by Array CGH-analysis at passage 48 (Chin et al., 2009).

(B) hiPSC18 at passage 58 acquired a full trisomy of chromosome 12 in approximately half the population (9/20 metaphases). Normal and trisomic karyotypes from the same analysis are presented.

(C) By passage 63 of hiPSC18 trisomy 12 cells had taken over the culture.

(D) Whole chromosome gain analysis of hiPSC18 showing this cell line was normal at passage 9.

(E) Full trisomy of chromosome 12 was detected at passage 63 (Bonferroni corrected p value = 4.3×10^{-47} by Expander and 1.6×10^{-38} by EASE).

(F) Moving average plot of the gene expression profile of hiPSC18 at passages 9 and 63.

subclone, whereas all the other lines were found to be completely normal (Figure 2C). These results are identical to the reported results from the CGH arrays, thus supporting both the sensitivity and the specificity of our methods. Finally, we performed parallel gene expression and SNP-array analyses of four HiPSC lines generated by us. All four lines were determined to be normal diploid lines in both the SNP-array and the Expander/EASE and PCF tests, further confirming the specificity and low false positive rate of these tests.

In total, the combination of Expander/EASE and PCF analyses identified 19 aberrations in the test data. Nine of these aberrations were also confirmed by direct chromosomal or DNA-based analysis, whereas none was refuted by such analyses. In order to assess the false positive rate of the Expander/EASE analysis

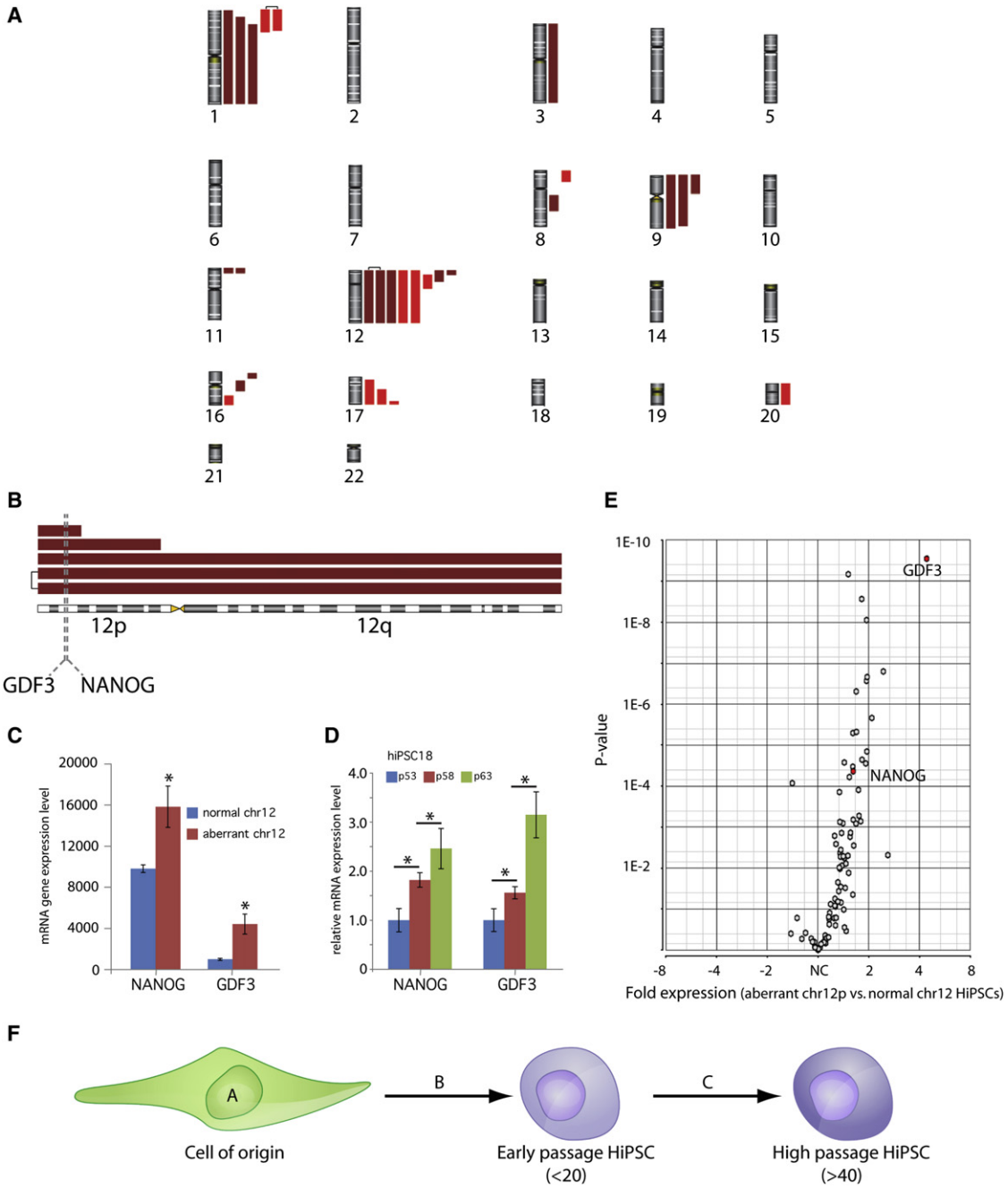

Figure 4. Functional Analysis of Recurring Chromosomal Aberrations in HiPSCs

(A) Ideogram representing gained chromosomal regions identified by PCF analysis. Red bars represent a gain of the respective chromosomal region in one line. Dark and light bars represent gains in HiPSCs and HESCs, respectively. Similar chromosomal aberrations in different passages of the same cell line are interconnected by a line.

(B) Representation of chromosome 12 gains identified by PCF analysis. Each bar represents a gain of the respective chromosomal region.

(C) Comparison of *NANOG* and *GDF3* expression between the five aberrant HiPSCs and HiPSCs lines with normal chromosome 12 copy number, showing that both genes are significantly overexpressed in the lines with gains in chromosome 12 (p values = 4.3×10^{-5} and 2.9×10^{-10}; average fold changes ×1.6 and ×4.4, respectively; error bars represent SEM).

(D) qRT-PCR expression of *NANOG* and *GDF3* at three different passages of hiPSC18, demonstrating their expression increase upon selection for trisomy 12 in the culture (normalized to β-Actin). Asterisks indicate p value < 0.05.

(E) Volcano plot showing overexpression and underexpression of genes residing in the minimal gained region common to all five aberrant HiPSCs, in these aberrant lines relative to lines with normal chromosome 12 copy number.

(F) Schematic model of the different types of chromosomal aberrations found in HiPSCs: aberrations with somatic cell origin ("A"); aberrations present in early passage but without apparent somatic cell origin ("B"); and aberrations acquired during prolonged culture ("C").

(i.e., the false positive rate of full trisomies based on gene expression profiles), we examined all lines that underwent direct chromosomal or DNA-analysis and gene expression profiling at similar passages. None of the 613 normal chromosomes was falsely detected as trisomic (false positive rate between 0 and 0.006, with 95% confidence). In order to assess the false positive rate of the PCF method for the detection of subchromosomal aberrations, we generated randomized data for each cell line using its own gene expression data. This was performed five times and a total of only two false positive aberrations were found (at an average size of ~90 probes). In the analysis of HESC and HiPSC lines, we identified a total of 45 discreet aberrations. Thus, the estimated false positive rate is 0.0038 (between 0.0005 and 0.0138, with 95% confidence). The smallest reported aberrations in the published data we examined were gain of 11.7 Mb and losses of 8.8 Mb and 11.2 Mb (corresponding to 86, 44, and 85 expressed genes, respectively). These were all correctly identified by our analysis, indicating that the current validated resolution of the analysis is ~10 Mb.

Because regional epigenetic modifications and coregulation might lead to misinterpretation of the gene expression data, we set out to further validate that our findings indeed represent chromosomal aberrations and not epigenetic effects. We conducted functional GO analysis on the regions found by PCF to be aberrant in one or more HiPSC lines and could not detect enrichment for any functional annotation that might suggest coregulation. Moreover, we could not detect aberration calls in regions of genes that are known to be clustered together and coregulated, such as the imprinting locus in chromosome 15.

However, epigenetic effects are very significant in the case of the X chromosome, precluding its analysis by this methodology. Upon analysis of chromosome X, some of the female lines were found to have the same level of gene expression as the male lines, whereas other female lines showed regional overexpression in chromosome X (Figure S2F). These results may suggest that X chromosome inactivation is variable in the female cell lines.

DISCUSSION

In this study, we examined the chromosomal stability of multiple HiPSC lines through the analysis of gene expression data. Previous studies comparing gene expression with copy number alterations indicated a strong correlation between copy number and gene expression. Here, we identified aneuploidy through the detection of regions of biased gene expression in multiple HESC and HiPSC lines.

The validity of our methodology was established in a prospective study in which SNP and gene expression analyses of HESCs and HiPSCs were conducted simultaneously. The methodology was further confirmed in a retrospective analysis of chromosomal aberrations which were previously reported in pluripotent stem cells. Thus, we identified 19 aberrations by both PCF and Expander/Ease analyses. In all cases in which direct DNA content analysis was available from similar passage as the gene expression, they were in agreement (nine aberrations). The high confidence of our results can be attributed to a large degree to the quality of the data set, which consisted of a large set of lines that demonstrate homogenous gene expression profiles.

Expression-based analysis has several important advantages: First, the functionality of the genomic abnormalities thus identified becomes immediately apparent with the identification of the genes whose expression is actually aberrant; second, the same biological material is used both for gene profiling and for assessment of the chromosomal integrity; third, it allows conducting retrospective examinations of genetic integrity; and lastly, the method allows detection of chromosomal duplications and deletions at higher resolution than standard karyotype analysis.

However, this methodology also has several limitations: First, only cell lines whose gene expression profile has been analyzed with the same platform can be compared to each other; second, because of the higher noise of gene expression data, this method is not as sensitive as CGH arrays, SNP arrays, or karyotype analysis in identifying abnormalities that exist only in a subpopulation in the culture. Furthermore, resolution is limited by the number of expressed genes in the sample. The unbalanced distribution of genes along the genome also dictates that euchromatic regions with higher gene abundance will be detected at higher resolution than heterochromatic regions; lastly, epigenetic regional modifications may also affect the interpretation of the data (Stransky et al., 2006).

Despite these limitations, we could successfully use gene expression profiling to detect chromosomal aberrations in pluripotent cells, by combining two statistical methods. These were consistently verified by direct DNA content analysis, on both HESCs and HiPSC lines, where such material was available.

Prior to this study, no comprehensive evaluation of the genomic integrity of HiPSCs has been reported. Most of the karyotypes of HiPSCs were conducted and reported during their initial characterization as pluripotent cells. Here, we identified thirteen (~20%) abnormal lines, of which six (~9%) carried at least one full trisomy (Figure S2). Because of the nature of gene expression profiling, abnormalities can only be identified if present in the majority of the cell population. Thus, any abnormality found should either have been present in the parental somatic cell line or to have been selected for. We could therefore divide the genomic aberrations identified in HiPSCs by this method to three general categories: A, aberrations with somatic cell origin; B, aberrations present in early passage but without apparent somatic cell origin; and C, aberrations acquired during passaging of the cells (Figure 4F).

In the first group are the HiPSC lines from the study by Kim et al. (2009). In their study, they report the derivation of three HiPSC lines by either direct reprogramming using proteins or by retroviral induction. The karyotypes for both the direct reprogrammed lines and the somatic cells (human newborn fibroblasts) were shown to be abnormal with trisomies in both chromosomes 1 and 9 (Kim et al., 2009). Here, we identify both trisomic chromosomes in the direct delivery lines as well as in the retroviral induced line, for which karyotype analysis was not provided in the original study (Figures 2A and 2B). Surprisingly, we could not detect corresponding aneuploidy in the somatic cell. This would indicate possible mosaicism in the parental somatic cells and a selective advantage conferred to the fibroblasts that carried these trisomies. Interestingly, chromosome 9 is known

to be a hallmark of germ cell tumors (Looijenga et al., 2006; McIntyre et al., 2007; Reuter, 2005).

The second group of aberrations consists of a number of early passage HiPSCs (passages 5 to 16) harboring sub-chromosomal duplications or deletions (Table S2). In contrast to the previous group, these aberrations seem to originate from events of genetic instability during the isolation and culture of the HiPSC lines, given that additional clones from the same source do not share the same aberrations. Moreover, no corresponding aberrations were found in the parent somatic cells. Thus, it seems that a rapid selection for certain aneuploidies occurs during the reprogramming process and the establishment of HiPSCs. This category of genomic instability, seen in HiPSCs but not in HESCs, adds to the accumulating data regarding the differences between these types of pluripotent cells.

The third group contains HiPSC lines that acquired chromosomal abnormalities upon prolonged time in culture (Figures 2D and 2E and Figures S2D and S2E). Of the four independent lines for which there is data from multiple passages, three acquired major aberrations. Notably, all these lines acquired copy number gains of either entire chromosome 12 or part of the short arm 12p.

The aberrations identified are non-randomly distributed, with the highest incidence occurring in chromosome 12, which has been previously shown to be involved in HESC adaptation. Importantly, gains in the 12p region were found in all five adapted high passage HiPSC lines, and full trisomy of chromosome 12 could be produced de novo upon culturing. Interestingly, in this study we did not detect aberrations of chromosome 17 in any of the HiPSC lines, whereas three such trisomies were detected in the smaller data set of HESC lines.

Using the gene expression data, we further demonstrated that these nonrandom aberrations indeed have functional implications. The hallmark pluripotency genes *NANOG* and *GDF3* were significantly overexpressed in the lines that possessed a gain in 12p. Moreover, functional annotation analysis revealed a considerable enrichment for cell cycle genes among the aberrant HiPSC lines. This is congruent with recent findings that cell division rate is a crucial parameter in the success and efficiency of the reprogramming process (Hanna et al., 2009). Interestingly, the same upregulation of *NANOG*, *GDF3*, and cell cycle genes were observed in the de novo generated trisomy 12 HiPSC line. The rapid rate at which this trisomy appeared demonstrates the functional significance of these gene expression abnormalities in conferring selective growth advantage to the cells that carry them, consistent with previous findings from chromosomal aberrations in HESCs (Olariu et al., 2010).

The relatively high incidence of aneuploidy in HiPSCs could be a side effect of the reprogramming process, in which integrating viral vectors are often used. Viral integrations have been previously shown to cause chromosomal aberrations in a proximity to the site of integration (Kadaja et al., 2009). In this study we did not observe higher incidence of aneuploidy in viral-derived HiPSC lines relative to lines derived without viral integration. Out of 14 confirmed aberrations in HiPSCs, five were evident in lines that were derived without viruses. Concurrently, of the 48 lines of HiPSCs derived by viral integrations, most were found to be normal. Moreover, many of the aneuploidies identified were of whole chromosomes or of very large chromosomal regions that are unlikely to have arisen because of viral integration-mediated process.

In conclusion, evaluation of the extent and nature of HiPSC genomic instability is important for both basic research and future clinical use. As was previously suggested and demonstrated regarding HESCs, such chromosomal aberrations might affect the differentiation capacity of the cells and increase their tumorigenicity (Blum and Benvenisty, 2009; Draper et al., 2004; Enver et al., 2005). More immediate implications apply for basic research, given that these chromosomal aberrations are likely to influence the interpretation of biological studies of HiPSCs. Thus, careful analysis of the genetic integrity of HiPSCs during culture is required, even at early passages.

EXPERIMENTAL PROCEDURES

Cell Culture

HESCs and HiPSCs were cultured on mitomycin C-treated mouse embryonic fibroblast (MEF) feeder layer and passaged as detailed in Supplemental Experimental Procedures.

Copy Number Variation Analysis

DNA was isolated from the HESC and HiPSC lines and analyzed by Genechip Human Mapping 250K Sty Array or by Human Mapping 500K Array in accordance with the manufacturer's protocol (Affymetrix, CA). Copy-number variation results were analyzed using the Partek Genomics Suite version 6.3 (Partek, MO; http://www.partek.com) with the 270 samples of the Human HapMap Project used as baseline. Data from the 250K arrays were analyzed with the Hidden Markov Model default settings in the copy number workflow in Partek Genomics Suite. The working definition of trisomy is a chromosome with multiple contiguous loci in which the relative copy number is three instead of two.

Gene Expression Profiles Database

Gene expression profiles from 18 studies that involved HESCs and/or HiPSCs and that conducted DNA microarray analysis with HG-U133plus2 or HG-ST1.0 microarray platforms (Affymetrix), were obtained from the GEO (Gene Expression Omnibus; http://www.ncbi.nlm.nih.gov/geo) and EMBL-EBI (http://www.ebi.ac.uk) databases. Raw .CEL files for all samples were analyzed with MAS5 or RMA probe set condensation algorithm with Expression Console (Affymetrix). Arrays were analyzed for quality control and outliers were removed. Further outliers were removed following hierarchical clustering analysis. Probes were filtered to retain a single instance for each gene expressed by >80% of the samples (see Supplemental Experimental Procedures), resulting in 12,054 data points. Values under 50 (for HG-U133plus2) or 6.0 (for HG-ST1.0) were collectively raised to this level. For reduction of possible bias from any given experiment, groups of similar samples with highly similar gene expression profiles (as judged by hierarchical clustering) were averaged for the sake of calculating the grand population median. This median then served as the baseline for examining expression bias.

Location Enrichment Analysis

For each chromosome for each cell line, the percentage of genes overexpressed >1.5-fold relative to the median expression of that gene was calculated. Enrichment of whole chromosomes or chromosomal arms was determined by subjecting the list of overexpressed genes of each line to location enrichment analysis with the software Expander (http://acgt.cs.tau.ac.il/expander) and EASE (http://david.abcc.ncifcrf.gov/ease/ease1.htm). Significance was defined as Bonferroni corrected p values $< 1.0 \times 10^{-4}$, the default value of the Expander program.

False positive and negative rates were calculated using exact binomial confidence intervals (http://statpages.org/confint.html).

CGH-PCF Overexpression Analysis

For each sample, the expression value of each gene was divided by the median of the same gene across the entire data set, in order to obtain

a comparative value. Data were then processed with a freely available CGH analysis program, CGH-Explorer (http://www.ifi.uio.no/forskning/grupper/bioinf/Papers/CGH). Gene expression regional bias was detected with the program's piecewise constant fit (PCF) algorithm, with a constant set of parameters (detailed in Supplemental Experimental Procedures).

Gene Ontology Analysis

For detection of significantly over-represented GO biological processes, the DAVID functional annotation clustering tool (http://david.abcc.ncifcrf.gov) was used (GO_TERM_BP_5). Expressed genes in chromosomal regions with copy number gain according to PCF analysis in at least two independent HiPSC lines were analyzed. General annotation categories with over 2000 members were excluded from the analysis. Enrichment was determined at DAVID calculated Benjamini value <0.05. Significance of overexpression of individual genes was determined with a standard Student's t test.

Comparison of Gene Expression between Normal and Aberrant Lines

Differentially expressed genes between the HiPSC lines with aberrations in chromosome 12 and the rest of the lines was determined with the t test probe filter function of Expander, with a 0.01 p value threshold and FDR multiple tests correction. *NANOG* and *GDF3* expression values were compared using Student's t test. Hierarchical clustering was performed with Partek Genomics Suite version 6.3 (Partek, MO; http://www.partek.com).

ACCESSION NUMBERS

Original microarray data are accessible at the GEO database under accession numbers GSE21243 and GSE21244. References to all microarray data used in this study are found in Tables S1 and S2.

SUPPLEMENTAL INFORMATION

Supplemental Information includes three tables and three figures and can be found with this article online at doi:10.1016/j.stem.2010.07.017.

ACKNOWLEDGMENTS

We would like to thank Tamar Golan-Lev for excellent technical assistance, Dr. Danny Kitzberg for critically reading this manuscript and Dr. Sagiv Shifman for stimulating discussions. N.B. is the Herbert Cohn Chair in Cancer Research. This research was partially supported by funds from the European Community (ESTOOLS, Grant number 018739) and by funds from the Morasha-ISF (Grant number 943/09). We gratefully acknowledge support for this project provided by a grant from the Legacy Heritage Fund of New York.

Received: December 7, 2009
Revised: May 27, 2010
Accepted: July 9, 2010
Published: October 7, 2010

REFERENCES

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.

Blum, B., and Benvenisty, N. (2009). The tumorigenicity of diploid and aneuploid human pluripotent stem cells. Cell Cycle *8*, 3822–3830.

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.

Crawley, J.J., and Furge, K.A. (2002). Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. Genome Biol. *3*, H0075.

Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, 3.

Draper, J.S., Moore, H.D., Ruban, L.N., Gokhale, P.J., and Andrews, P.W. (2004). Culture and characterization of human embryonic stem cells. Stem Cells Dev. *13*, 325–336.

Enver, T., Soneji, S., Joshi, C., Brown, J., Iborra, F., Orntoft, T., Thykjaer, T., Maltby, E., Smith, K., Abu Dawud, R., et al. (2005). Cellular differentiation hierarchies in normal and culture-adapted human embryonic stem cells. Hum. Mol. Genet. *14*, 3129–3140.

Furge, K.A., Dykema, K.J., Ho, C., and Chen, X. (2005). Comparison of array-based comparative genomic hybridization with gene expression-based regional expression biases to identify genetic abnormalities in hepatocellular carcinoma. BMC Genomics *6*, 67.

Guryev, V., Saar, K., Adamovic, T., Verheul, M., van Heesch, S.A., Cook, S., Pravenec, M., Aitman, T., Jacob, H., Shull, J.D., et al. (2008). Distribution and functional impact of DNA copy number variation in the rat. Nat. Genet. *40*, 538–545.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature *462*, 595–601.

Henrichsen, C.N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F., Ruedi, M., Kaessmann, H., and Reymond, A. (2009). Segmental copy number variation shapes tissue transcriptomes. Nat. Genet. *41*, 424–429.

Hertzberg, L., Betts, D.R., Raimondi, S.C., Schäfer, B.W., Notterman, D.A., Domany, E., and Izraeli, S. (2007). Prediction of chromosomal aneuploidy from gene expression data. Genes Chromosomes Cancer *46*, 75–86.

Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol. *4*, R70.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Hughes, T.R., Roberts, C.J., Dai, H., Jones, A.R., Meyer, M.R., Slade, D., Burchard, J., Dow, S., Ward, T.R., Kidd, M.J., et al. (2000). Widespread aneuploidy revealed by DNA microarray expression profiling. Nat. Genet. *25*, 333–337.

Kadaja, M., Isok-Paas, H., Laos, T., Ustav, E., and Ustav, M. (2009). Mechanism of genomic instability in cells infected with the high-risk human papillomaviruses. PLoS Pathog. *5*, e1000397.

Kim, D., Kim, C.H., Moon, J.I., Chung, Y.G., Chang, M.Y., Han, B.S., Ko, S., Yang, E., Cha, K.Y., Lanza, R., and Kim, K.S. (2009). Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins. Cell Stem Cell *4*, 472–476.

Lilljebjörn, H., Heidenblad, M., Nilsson, B., Lassen, C., Horvat, A., Heldrup, J., Behrendtz, M., Johansson, B., Andersson, A., and Fioretos, T. (2007). Combined high-resolution array-based comparative genomic hybridization and expression profiling of ETV6/RUNX1-positive acute lymphoblastic leukemias reveal a high incidence of cryptic Xq duplications and identify several putative target genes within the commonly gained region. Leukemia *21*, 2137–2144.

Lingjaerde, O.C., Baumbusch, L.O., Liestøl, K., Glad, I.K., and Børresen-Dale, A.L. (2005). CGH-Explorer: A program for analysis of array-CGH data. Bioinformatics *21*, 821–822.

Looijenga, L.H., Hersmus, R., Gillis, A.J., Pfundt, R., Stoop, H.J., van Gurp, R.J., Veltman, J., Beverloo, H.B., van Drunen, E., van Kessel, A.G., et al. (2006). Genomic and expression profiling of human spermatocytic seminomas: Primary spermatocyte as tumorigenic precursor and DMRT1 as candidate chromosome 9 gene. Cancer Res. *66*, 290–302.

Lowry, W.E., Richter, L., Yachechko, R., Pyle, A.D., Tchieu, J., Sridharan, R., Clark, A.T., and Plath, K. (2008). Generation of human induced pluripotent stem cells from dermal fibroblasts. Proc. Natl. Acad. Sci. USA *105*, 2883–2888.

Masaki, H., Ishikawa, T., Takahashi, S., Okumura, M., Sakai, N., Haga, M., Kominami, K., Migita, H., McDonald, F., Shimada, F., et al. (2007).

Heterogeneity of pluripotent marker gene expression in colonies generated in human iPS cell induction culture. Stem Cell Res. *1*, 105–115.

Masayesva, B.G., Ha, P., Garrett-Mayer, E., Pilkington, T., Mao, R., Pevsner, J., Speed, T., Benoit, N., Moon, C.S., Sidransky, D., et al. (2004). Gene expression alterations over large chromosomal regions in cancers include multiple genes unrelated to malignant progression. Proc. Natl. Acad. Sci. USA *101*, 8715–8720.

McIntyre, A., Summersgill, B., Lu, Y.J., Missiaglia, E., Kitazawa, S., Oosterhuis, J.W., Looijenga, L.H., and Shipley, J. (2007). Genomic copy number and expression patterns in testicular germ cell tumours. Br. J. Cancer *97*, 1707–1712.

Olariu, V., Harrison, N.J., Coca, D., Gokhale, P.J., Baker, D., Billings, S., Kadirkamanathan, V., and Andrews, P.W. (2010). Modeling the evolution of culture-adapted human embryonic stem cells. Stem Cell Res. *4*, 50–56.

Park, I.H., Zhao, R., West, J.A., Yabuuchi, A., Huo, H., Ince, T.A., Lerou, P.H., Lensch, M.W., and Daley, G.Q. (2008). Reprogramming of human somatic cells to pluripotency with defined factors. Nature *451*, 141–146.

Phillips, J.L., Hayward, S.W., Wang, Y., Vasselli, J., Pavlovich, C., Padilla-Nash, H., Pezullo, J.R., Ghadimi, B.M., Grossfeld, G.D., Rivera, A., et al. (2001). The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis. Cancer Res. *61*, 8143–8149.

Pollack, J.R., Sørlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Børresen-Dale, A.L., and Brown, P.O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc. Natl. Acad. Sci. USA *99*, 12963–12968.

Reuter, V.E. (2005). Origins and molecular biology of testicular germ cell tumors. Mod. Pathol. *18* (*Suppl 2*), S51–S60.

Schoch, C., Kohlmann, A., Dugas, M., Kern, W., Hiddemann, W., Schnittger, S., and Haferlach, T. (2005). Genomic gains and losses influence expression levels of genes located within the affected regions: A study on acute myeloid leukemias with trisomy 8, 11, or 13, monosomy 7, or deletion 5q. Leukemia *19*, 1224–1228.

Sharan, R., Maron-Katz, A., and Shamir, R. (2003). CLICK and EXPANDER: A system for clustering and visualizing gene expression data. Bioinformatics *19*, 1787–1799.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science *315*, 848–853.

Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., de Medina, S.G., Segraves, R., de Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., et al. (2006). Regional copy number-independent deregulation of transcription in cancer. Nat. Genet. *38*, 1386–1396.

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.

Tsafrir, D., Bacolod, M., Selvanayagam, Z., Tsafrir, I., Shia, J., Zeng, Z., Liu, H., Krier, C., Stengel, R.F., Barany, F., et al. (2006). Relationship of gene expression and chromosomal abnormalities in colorectal cancer. Cancer Res. *66*, 2129–2137.

Yang, S., Lin, G., Tan, Y.Q., Zhou, D., Deng, L.Y., Cheng, D.H., Luo, S.W., Liu, T.C., Zhou, X.Y., Sun, Z., Xiang, Y., Chen, T.J., Wen, J.F., and Lu, G.X. (2008). Tumor progression of culture-adapted human embryonic stem cells during long-term culture. Genes Chromosomes Cancer *47*, 665–679.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., et al. (2007). Induced pluripotent stem cell lines derived from human somatic cells. Science *318*, 1917–1920.

Yu, J., Hu, K., Smuga-Otto, K., Tian, S., Stewart, R., Slukvin, I.I., and Thomson, J.A. (2009). Human induced pluripotent stem cells free of vector and transgene sequences. Science *324*, 797–801.