STEM CELLS®

EMBRYONIC STEM CELLS

Copy Number Variant Analysis of Human Embryonic Stem Cells

HAO WU,[a,b] KEVIN J. KIM,[a,b] KSHAMA MEHTA,[c] SALVATORE PAXIA,[d] ANDREW SUNDSTROM,[d] THOMAS ANANTHARAMAN,[d] ALI I. KURAISHY,[e] TRI DOAN,[c] JAYATI GHOSH,[c] APRIL D. PYLE,[f,g,h,i] AMANDER CLARK,[g,h,i,j] WILLIAM LOWRY,[g,h,i,j] GUOPING FAN,[g,h,i,k] TIM BAXTER,[c] BUD MISHRA,[d] YI SUN,[a,b,g,h,i] MICHAEL A. TEITELL[e,g,h,i]

Departments of [a]Psychiatry and Biobehavioral Sciences, [b]Molecular and Medical Pharmacology, [e]Pathology and Laboratory Medicine, [f]Molecular Immunology and Medical Genetics, [j]Molecular, Cell and Developmental Biology, and [k]Human Genetics, David Geffen School of Medicine, [g]Molecular Biology Institute, [h]Jonsson Comprehensive Cancer Center, and [i]Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, California, USA; [c]Agilent Laboratories, Santa Clara, California, USA; [d]New York University/Courant Bioinformatics Group, Courant Institute of Mathematical Sciences, New York University, New York, New York, USA

Key Words. Embryonic stem cells • Oligonucleotide array sequence analysis • Genome stability • Multipoint statistics • Algorithmic biology

ABSTRACT

Differences between individual DNA sequences provide the basis for human genetic variability. Forms of genetic variation include single-nucleotide polymorphisms, insertions/duplications, deletions, and inversions/translocations. The genome of human embryonic stem cells (hESCs) has been characterized mainly by karyotyping and comparative genomic hybridization (CGH), techniques whose relatively low resolution at 2–10 megabases (Mb) cannot accurately determine most copy number variability, which is estimated to involve 10%–20% of the genome. In this brief technical study, we examined HSF1 and HSF6 hESCs using array-comparative genomic hybridization (aCGH) to determine copy number variants (CNVs) as a higher-resolution method for characterizing hESCs. Our approach used five samples for each hESC line and showed four consistent CNVs for HSF1 and five consistent CNVs for HSF6. These consistent CNVs included amplifications and deletions that ranged in size from 20 kilobases to 1.48 megabases, involved seven different chromosomes, were both shared and unique between hESCs, and were maintained during neuronal stem/progenitor cell differentiation or drug selection. Thirty HSF1 and 40 HSF6 less consistently scored but still highly significant candidate CNVs were also identified. Overall, aCGH provides a promising approach for uniquely identifying hESCs and their derivatives and highlights a potential genomic source for distinct differentiation and functional potentials that lower-resolution karyotype and CGH techniques could miss. STEM CELLS *2008;26:1484–1489*

Disclosure of potential conflicts of interest is found at the end of this article.

INTRODUCTION

Human embryonic stem cells (hESCs) that have been cultured for extended periods may retain a diploid karyotype [1–3] or may demonstrate chromosomal instability, often marked by translocations and aneuploidy. This chromosome-scale instability probably reflects a selective advantage during growth over time in suboptimal conditions [1, 4, 5]. Current approaches used to evaluate hESC genome integrity include mainly G-banding metaphase karyotyping and metaphase-based comparative genomic hybridization (CGH). These techniques deliver a sensitivity estimated at 5–10 megabases (Mb) for karyotype down to 2–3 Mb for CGH [6, 7]. A recent study also evaluated single-nucleotide polymorphisms (SNPs) for early- and late-passage hESCs and determined that an SNP fingerprint could uniquely identify samples [8]. However, copy number variants (CNVs), representing amplified or deleted regions ranging in size from 1 kilobase (kb) up to 1.0 Mb or more [9, 10], have been recently recognized as a major source of human genome variability that potentially exceeds SNP differences between individuals by more than threefold and is below the detection threshold of most current techniques [11–14]. Using array-comparative genomic hybridization (aCGH), a well-established microarray-based nucleic acid hybridization method with kb-scale resolution [15–18], studies of somatic cells from phenotypically normal individuals show, on average, ~11 CNV differences between unrelated genomes [12, 14]. CNV differences among individuals may at least partially explain human uniqueness, whereas CNV similarities among individuals may indicate subpopulation relatedness. This is because particular CNVs have been shown to affect gene expression, influence phenotypic variation and adaptation by disrupting genes and altering dosage compensation, cause a variety of diseases, and confer risk to complex traits, such as HIV-1 susceptibility and glomerulonephritis [19–29]. Positive environmental selection has been postulated to impact CNV genes, such as amplification of the salivary amylase gene from populations with high-starch diets

Correspondence: Michael A. Teitell, M.D., Ph.D., Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, 675 Charles Young Drive South, 4-762 MRL, Los Angeles, California 90095-1732, USA. Telephone: 310-206-6754; Fax: 310-267-0382; e-mail: mteitell@ucla.edu Received November 26, 2007; accepted for publication March 18, 2008; first published online in STEM CELLS *EXPRESS* March 27, 2008. 1066-5099/2008/$30.00/0 doi: 10.1634/stemcells.2007-0993

Table 1. HSF1 CNV intervals identified by both BuddhaCGH Pipeline and CGHAnalytics 3.4 platforms

Amp/Del	Chr	Locus	Interval start	Interval end	Size	Genes	CNV locus	Variation ID
Amp	20	q11.21	29.309999	30.780001	1.48 Mb	*DEFB115–119*, *DEFB121*, *DEFB123*, *DEFB124*, *REM1*, *HM13*, *ID1*, *COX4I2*, *BCL2L1*, *TPX2*, *MYLK2*, *FKHL18*, *DUSP15*, *TTLL9*, *PDRG1*, *XKR7*, *C20orf160*, *HCK*, *TM9SF4*, *TSPYL3*, *PLAGL2*, *POFUT1*, *KIF3B*, *ASXL1*, *C20orf112*, *COMMD7*	New	New
Del	6	p21.32	32.59	32.630001	40 kb	*HLA-DRB5*	1673	New
Del	7	q34	141.210007	141.259995	50 kb	*MGAM*	2092	10211
Del	22	q11.23	22.67	22.710001	40 kb	*GSTT1*	4208	2032

Abbreviations: Amp, amplified; Chr, chromosome; CNV, copy number variant; Del, deleted; ID, identification; kb, kilobases; Mb, megabases.

[30]. These observations on the extent and influence of CNVs in the genomes of somatic cells leave open questions regarding the extent to which CNVs account for hESC genomic variability and whether individual hESCs have characteristic gains or losses of genetic material that may influence their replication/proliferation, differentiation, or functional potentials. It is also unclear whether genomic regions below karyotype or metaphase CGH detection limits are stable over culture time or with differing culture conditions, such as positive drug selection. Therefore, in this brief technical assessment, we used a high-density aCGH platform to determine the extent to which two hESC lines with distinct neuronal differentiation potentials, HSF1 and HSF6, display CNVs.

Materials and Methods

Cell Culture

NIH-registered hESCs HSF1 (46XY; UC-0001) and HSF6 (46XX; UC-0006) were cultured on irradiated CF1 mouse embryonic fibroblasts (MEFs) in Dulbecco's modified Eagle's medium-high glucose supplemented with 20% knockout serum replacer (Invitrogen, Carlsbad, CA, http://www.invitrogen.com), basic fibroblast growth factor (4–10 ng/ml), 1 mM glutamine, 1% nonessential amino acids (with or without 1% penicillin/streptomycin), 0.1 mM 2-mercaptoethanol with daily medium changes. hESCs were passaged every 4–6 days by incubation with 1 mg/ml collagenase IV ± dispase (Invitrogen) for 10–20 minutes at 37°C. Conditions for the differentiation and maintenance of neuronal stem/progenitor cells (NPCs) from HSF1 and HSF6 have been described [31]. Drug-resistant HSF1 lines were generated by culturing with 250 μg/ml G418 or 100 μg/ml hygromycin for 14 days on DR4 MEFs following transduction with neomycin- or hygromycin-resistant lentiviruses, respectively. Medium was changed every 2 days, and after 14 days, drug-selected hESCs were passaged onto CF1 MEFs.

aCGH

Genomic DNA (gDNA) was prepared and quantified from HSF1 and HSF6 cells by standard techniques. Sample (0.5–1.0 μg) and sex-mismatched normal reference gDNA (Promega, Madison, WI, http://www.promega.com) was digested with AluI and RsaI and labeled with either Cy3- or Cy5-dUTP (Agilent Genomic Labeling Kit Plus; Agilent Technologies, Santa Clara, CA, http://www.agilent.com). Labeled gDNA products were purified using Microcon YM-30 filtration devices (Millipore, Bedford, MA, http://www.millipore.com) and volume-adjusted, and DNA yield and level of dye incorporation were measured using an ND-1000 spectrophotometer (NanoDrop, Rockland, DE, http://www.nanodrop.com). Specific Cy5- and Cy3-labeled DNA sample pairs were combined and mixed with human Cot-1 DNA (Invitrogen), 10 × blocking agent, and 2 × hybridization buffer (Agilent Technologies). Samples were heated at 95°C for 3 minutes, incubated for 30 minutes at 37°C, and then hybridized to human genome microarrays using Agilent SureHyb chambers. aCGH was performed on high-density oligonucleotide microarrays containing >236,000 (244K) coding and noncoding 60-mer oligonucleotide sequences (G4411B; Agilent Technologies). This 244K microarray was generated from human genome assembly build 35 (March 2004; UCSC hg17) and has a 6.4-kb average probe spatial resolution that is biased for known genes, microRNAs, gene promoters, intergenic regions, and telomeres. The 244K microarray is biased against regions of the genome that contain segmental duplications and repetitive elements, which is where CNVs are often located, and has 63,900 of 236,000 array probes overlapping with known CNVs, for approximately 27% known CNV coverage. The hybridization chambers were placed in a 65°C rotisserie oven and rotated at 20 rpm for 40 hours, followed by washing according to the procedures described in Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis protocol version 4. Microarray slides were scanned immediately using an Agilent microarray scanner. Data were extracted using Feature Extraction software, version 9.5.1.1 (Agilent Technologies, Santa Clara, CA, http://www.agilent.com).

Data Analysis

aCGH data sets were analyzed with two independently developed software packages, BuddhaCGH Pipeline and CGHAnalytics 3.4 (Agilent Technologies). BuddhaCGH Pipeline was developed to analyze genomic array data in a technology-agnostic manner (S. Paxia et al., manuscript in preparation) and has been used to evaluate CNVs in cancer and to identify oncogenes and tumor-suppressor genes from genomic data, either alone or in conjunction with transcriptome data [32, 33]. CGHAnalytics 3.4, a software platform developed by Agilent Technologies, provided an independent statistical approach for determining genomic amplified and deleted regions. These two statistical tools differed in subtle ways in how they normalized the data, removed noise, partitioned the data into segments of equal copy number values, and interpreted segments. Nevertheless, they were expected to concur on the statistically most significant conclusions, to thereby increase confidence in the results of the joint analyses.

In CGHAnalytics 3.4, a data normalization step enables the identification of probes that behave similarly with respect to both the test and control samples (i.e., logarithms of their intensity ratios are 0). Consequently, normalization of the log ratios also provides a basis for identifying and characterizing those probes that significantly deviate in copy number. Under the assumption that almost all pairs of consecutive probes sample regions of similar copy number, the log ratios of most pairs of consecutive probes are expected to be equal, and significant deviations from near-equality (estimated with respect to the expected probe-to-probe noise) were used to identify genomic aberrations. The deviation is measured as the SD of the derivative of the log ratio for an array and is used in estimating the statistical significance of observed aberrations. All regions of statistically significant copy number changes are identified using the aberration detection module-1 (ADM-1) algorithm [34]. Stable regions of amplification or deletion across multiple samples of each hESC type were called visually, with a penetrance of 100% reported in Tables 1 and 2.

Table 2. HSF6 CNV intervals identified by both BuddhaCGH Pipeline and CGHAnalytics 3.4 platforms

Amp/Del	Chr	Locus	Interval start	Interval end	Size	Genes	CNV locus	Variation ID
Amp	1	q21.3	149.380005	149.399994	20 kb	*LCE3C*	234	1018
Del	4	q13.2	69.270004	69.310005	40 kb		1134	1084
Del	6	p21.32	32.59	32.630001	40 kb	*HLA-DRB5*	1673	new
Del	19	p12	20.42	20.49	70 kb		3923	10546
Del	22	q11.23	22.67	22.710001	40 kb	*GSTT1*	4208	2032

Abbreviations: Amp, amplified; Chr, chromosome; CNV, copy number variant; Del, deleted; ID, identification; kb, kilobases.

BuddhaCGH uses a Bayesian maximum a posteriori (MAP) algorithm to find segments of genomic amplifications and deletions, under a set of noninformative priors regarding the distributions of segmental breakpoints and segment lengths with penalty terms determined by the intersegment *t* statistics. The optimization problem admits an efficient implementation under a dynamic programming formulation and leads to a robust estimator by accounting for outlier probe values. The data sets were normalized prior to segmentation so that different hESC data sets could be directly compared. Since in both analyses each segment was constrained to be no smaller than three contiguous probes, the algorithms can catch a segmental change at ~10–15-kb resolution.

The BuddhaCGH pipeline introduced a multilocus scoring function to identify genomic intervals that show a stable or possibly time-dependent signature, to detect the possible evolution of amplifications or deletions over culture time. For example, a score function "rewarded" an amplified interval that emerged at an early period in time but stabilized over time. The same score function also "penalized" these values when measurements or biological noise corrupted the signals. The results of this analysis for HSF1 and HSF6 amplifications and deletions were tabulated, with the entries sorted by score and filtered to include only score values above a threshold. By visually comparing the spatial distributions of the high-scoring intervals and by aiming to select no more than one or two regions per chromosomal arm, we determined that a threshold of 0.2 or above was reasonable. For a conservative interplatform comparison, we used a more stringent threshold of 0.7 to lower the number of false-positive intervals and genes.

We compared the results from BuddhaCGH and CGHAnalytics 3.4 to identify regions of concordance. Supplemental online Table 1 includes all of the data from the two different software systems, including regions where only one or the other software platform identified an amplified or deleted region above the cutoff threshold.

Quantitative Real-Time Polymerase Chain Reaction

Quantitative real-time polymerase chain reaction (qPCR) was performed with an iCycler using iQ SYBR Green Supermix (Bio-Rad, Hercules, CA, http://www.bio-rad.com) and male and female reference, HSF1, and HSF6 gDNA samples. Specificity for each primer pair was examined by melting curve functionality and agarose gel electrophoresis. To calculate gene copy number differences between reference normal gDNA (Promega) and HSF1 or HSF6 gDNA, the threshold cycle of each sample was normalized to a negative control gene, *SRC*, which did not show copy number differences between reference and HSF1 or HSF6 gDNAs in 10 independent aCGH experiments. Fold change between reference gDNA and HSF1 or HSF6 gDNA was calculated based on the $2^{-\Delta\Delta C_t}$ method. Primer sequences used for qPCR are listed in supplemental online Table 2. Error bars represent the SEM.

RESULTS

HSF1 and HSF6 hESCs were recently shown to have normal diploid karyotypes at passage 25 [35], with HSF1 further characterized as 46XY to at least passage 29 (University of California, San Francisco, NIH Registry Stem Cells, http://www.escells.ucsf.edu) and HSF6 further determined as 46XX to passage 120 [31], indicating chromosome level genomic stability. To determine subkaryotypic genome heterogeneity, aCGH using >236,000 feature, 60-mer oligonucleotide probe microarrays was used. The choice of this platform versus an SNP array was not straightforward, as each approach offers advantages and disadvantages. For example, SNP arrays can provide both copy number estimations and genotype information in cases of copy-neutral aberrations, such as uniparental disomy [36, 37]. In a recent comparison, Agilent 60-mer oligonucleotide microarrays provided the highest sensitivity and specificity of CNV detection, whereas the newer SNP arrays (e.g., Affymetrix SNP 6.0 [Affymetrix, Santa Clara, CA, http://www.affymetrix.com] and Illumina Linkage IV [Illumina Inc., San Diego, http://www.illumina.com]) require polymerase chain reaction-probe complexity reduction, and their performance has yet to be determined, precipitating our platform choice [38, 39].

HSF1 gDNA was assessed from two pluripotent controls grown in different laboratories with no in vitro manipulations, from stably selected G418 or hygromycin-resistant subclones following retroviral integration of resistance genes, and from NPCs differentiated from HSF1 hESCs at passage 2, providing five independent assessments under a variety of growth conditions. HSF6 cells were differentiated to NPCs as recently described [31], and gDNA was evaluated at passages 1, 2, 3, 4, and 9, also providing five independent assessments over a 2-month period of continuous culturing after hESCs were converted/differentiated into NPCs. An important aspect of this study was the distinct culturing conditions and passage numbers used for aCGH analysis, because a main study goal was to determine CNV region stability in a variety of culturing contexts, as occurs in laboratories worldwide.

Resulting aCGH data were analyzed with two independently developed software packages. CGHAnalytics 3.4 (http://www.chem.agilent.com/Scripts/PDS.asp) uses the ADM-1 algorithm [34], with a stringent scoring threshold of 8. BuddhaCGH Pipeline is a custom data analysis package that includes a substrate-specific data normalization module, a MAP segmenter module [32] to partition the normalized copy-number data, and a multipoint statistical analysis module [33] that scores genomic intervals for how well they support a particular biological hypothesis (supplemental online text). With each analysis package, a CNV interval required a minimum of three consecutive concordant probes to score positive for amplification or deletion. Variant segment chromosome mapping and identification of genes contained within variant segments used human genome assembly build 35 (hg17). CNV tags were obtained from the most recent (October 24, 2007) The Centre for Applied Genomics (TCAG) Database of Genomic Variants using build 35, which currently contains 11,784 CNVs (http://projects.tcag.ca/variation/).

HSF1 CNVs

Table 1 shows consistently amplified and deleted regions in HSF1 over five biological and technical replicates, performed once for each growth condition, as determined by both CGHAnalytics 3.4 and BuddhaCGH Pipeline. Four candidate genomic intervals, one amplification region and three deletion

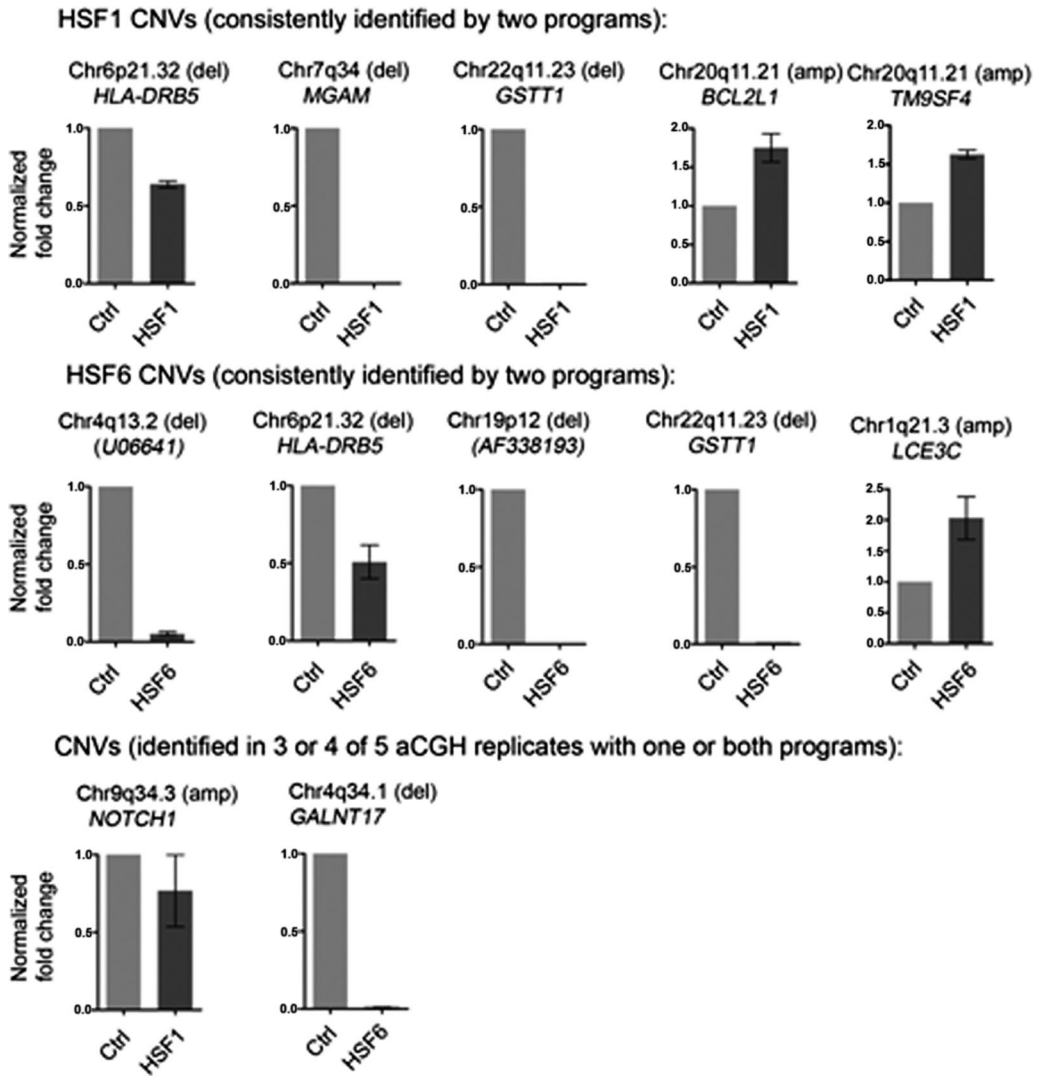

Figure 1. Quantitative real-time polymerase chain reaction validation of consistent and variably detected CNV region genes. Abbreviations: aCGH, array-comparative genomic hybridization; amp, amplified; Chr, chromosome; CNV, copy number variant; Ctrl, control; del, deleted.

regions, provide stable CNVs that could represent the genetic makeup of the inner cell mass from which this line was derived or that may be part of an early adaptation to derivation and subsequent culturing in conditions that did not select against these CNVs. This number of consistent CNVs is a conservative underestimate of the likely number of CNVs because both programs had to positively score at least a three-probe interval, and each program was set with high stringencies for detection. Thirty additional candidate HSF1 CNV intervals, at slightly lower (and therefore detected in three or four aCGH runs instead of all five) but still highly significant threshold scores, are shown in supplemental online Table 1.

A 1.48-Mb stable amplified region involving 30 genes was detected at 20q11.21 (supplemental online Figs. 1, 2), whereas three stable deleted regions averaging 40–50-kb each and containing at least part of a single coding gene were detected at 6p21.32, 7q34, and 22q11.23. Among the affected genes are copy number variants in the *BCL2L1* gene, which regulates cell survival and death [40]; *ID1*, which encodes a protein that binds helix-loop-helix transcription factors to inhibit lineage commitment and affect cell growth, senescence, and differentiation; and *HLA-DRB5*, which is part of the locus that encodes histocompatibility antigens (Table 1). The 1.48-Mb amplification at 20q11.21 is a new CNV locus providing a novel CNV variation ID not previously reported in the TCAG Database of Genomic Variants, whereas the 40-kb deletion at 6p21.32 provides a novel CNV variation ID within the reported 1,673 CNV locus.

qPCR was used to evaluate five genes within the four consistently identified HSF1 CNV regions, including two genes from amplified regions and three genes from deleted regions (Fig. 1). In each case, the aCGH result was confirmed. Interestingly, CNV-associated genes at 7q34 and 22q11.23 showed homozygous deletions, whereas *HLA-DRB5* at 6p21.32 was haploinsufficient and two genes at 20q11.21 showed low-level copy number amplification.

Supporting a 1.48-Mb amplification at 20q11.21, 11 genes within this CNV region, including a cluster of 6 linearly arranged genes (*HM13*, *ID1*, *COX412*, *BCL2L1*, *TPX2*, and *MYLK2*) and a smaller cluster of 3 linearly arranged genes (*KIF3B*, *ASXL1*, and *C20orf112*), are all overexpressed by ~1.5–2.5-fold in HSF1 NPCs versus HSF6 NPCs [31], which lacks this unique CNV amplification (Table 2). Tissue typing also shows that HLA-DRB5 is not an expressed allele of the HLA locus in HSF1 hESCs, providing an example of a biallelic gene that is deleted in one allele (Fig. 1) without a known functional consequence (E.F. Reed and G. Fan, data not shown). HSF1 hESCs grown under selection or

induced to NPCs at passage 2 demonstrated the same four CNVs, indicating stability for these previously altered genomic regions under a variety of culturing conditions.

HSF6 CNVs

Table 2 shows the stable CNV regions for HSF6 hESCs detected by both analysis programs over five biological and technical replicate samples and includes one amplified region and four deleted regions. A 20-kb amplified region was detected at 1q21.3, whereas a 70-kb deleted region was seen at 19p12, and 40-kb deleted regions were identified at 4q13.2, 6p21.32, and 22q11.23. Deletions at 6p21.32 and 22q11.23 were shared with HSF1 hESCs and include a copy number change in the biallelically expressed *HLA-DRB5* histocompatibility antigen locus, which is also not expressed in HSF6 hESCs (Reed and Fan, data not shown). Two CNV regions, 4q13.2 and 19p12, lack any known coding genes, whereas CNV regions at 1q21.3 and 22q11.23 affect single genes whose expression is unaltered in HSF1 NPCs versus HSF6 NPCs with gene expression microarray analysis [31]. Forty additional candidate HSF6 CNV intervals, at slightly lower (and therefore detected in three or four instead of all five aCGH runs) but still highly significant threshold scores are shown in supplemental online Table 1. Because of the serial-passage component of HSF6 hESC analysis, 18 regions that were altered in only one or two early passages but not later passages were considered technical or scoring artifacts. One small amplified region on 11p15.5 appeared during passages 4 and 9 (supplemental online Table 1), suggesting a potential subkaryotpic region of instability with extended culture time.

qPCR was used to evaluate five consistently identified HSF6 CNV regions (Table 2), including one amplified CNV gene and four loci with two identified genes from four CNV deleted regions (Fig. 1). In each case, the aCGH result was again confirmed. Interestingly, homozygous deletions were identified at 4q13.2, 19p12, and 22q11.23 (*glutathione S-transferase 1* [*GSTT1*]), whereas *HLA-DRB5* at 6p21.32 was haploinsufficient and *LCE3C* at 1q21.3 showed low-level copy number amplification. Shared CNVs containing *HLA-DRB5* and *GSTT1* showed identical copy number aberrations between HSF1 and HSF6 cells for unknown reasons. Two additional genes from lower threshold scoring CNV regions (supplemental online Table 1) were also evaluated by qPCR, with *GALNT17* from HSF6 showing a homozygous deletion and *NOTCH1* from HSF1 not confirming the lower scoring threshold aCGH analysis. These results reveal the detection robustness associated with stringent scoring from two independent data analysis packages and the likelihood that multiple lower threshold scoring candidates also identify CNV regions.

DISCUSSION

The International Stem Cell Initiative has undertaken the initial characterization of 75 hESC lines, which includes DNA fingerprinting of 10 short tandem repeat loci for each line to establish its unique identity [35]. Thus far, the genomes of most hESCs have been characterized by relatively low-resolution, 2–10-Mb genome-wide techniques. Here, we evaluated the utility of identifying CNVs in hESCs using kb-resolution aCGH. Although SNP analysis provides nucleotide resolution, the sequence variation in absolute nucleotide numbers between two independent human genomes, including by extension two different hESC genomes, is estimated to be 5–10-fold higher because of amplified or deleted CNV regions compared to SNP variations alone [41]. Therefore, CNV differences could contribute as much or more genetic variability than SNP differences to controlling hESC differentiation and function.

Two independent analytic programs using five replicate samples with stringent scoring criteria showed that HSF1 and HSF6 have both shared and unique amplified and deleted genomic regions. Two of the four stringently identified CNVs in HSF1 and five stringently identified CNVs in HSF6 were shared between these hESCs. Shared homozygous loss of *GSTT1* (Fig. 1), which catalyzes the conjugation of reduced glutathione to various electrophilic and hydrophobic compounds and has been implicated in carcinogenesis, could influence the selection of specific hESCs for study or therapy. Additional CNV differences between these hESCs are likely to emerge from the 30 candidate HSF1 CNVs and 40 candidate HSF6 CNVs that scored positively in three or four of the five replicate samples for each hESC (supplemental online Table 1; Fig. 1), such as a confirmed homozygous deletion in the *GALNT17* gene in HSF6, thereby increasing genome uniqueness for these lines. In fact, the number of CNVs that ultimately distinguish HSF1 and HSF6 hESCs, and by extension that are likely to distinguish all hESCs from one another, are likely to be similar to the estimated average of 11 CNV differences between individuals reported for somatic cells [12, 14].

HSF1 and HSF6 hESCs were chosen for this study because they were derived under similar protocols from a single institution and yet they exhibit unique neuronal differentiation potentials under identical culturing conditions [31]. Although the HSF6 line tends to produce midbrain GABAergic (PAX2- and GAD67-positive) neurons and other subtypes of neurons with posterior regional identities (e.g., dopaminergic, serotonergic, and cholinergic neurons from midbrain, hindbrain, and spinal cord), the HSF1 line tends to generate forebrain-like glutamatergic and GABAergic neurons. It is not yet known what genetic or epigenetic factors influence these distinct neuronal fate decisions. However, CNV differences could participate in these differential outcomes. Although it is tempting to consider a direct copy number-gene expression relationship for controlling distinct differentiation fates, and approximately twofold increased expression of *BCL2L1* is retained in 20q11.21 amplified HSF1 hESCs and NPCs compared with HSF6 hESCs and NPCs (data not shown), this is almost certainly an oversimplification. For example, there are numerous reports of wide variation in gene and protein expression associated with genes within CNV regions. Because gene expression depends on lineage, state of differentiation within a lineage, and a variety of genetic (copy number, point mutation, and rearrangement) and epigenetic (DNA methylation, histone modification, and post-transcriptional stability) factors, a strictly linear correlation between gene expression level and CNV status is unlikely. This is also true for cells in two different differentiation states, such as hESCs and NPCs. If such a linear relationship did occur, it would tend to negate the impact for all of these affiliated factors beyond simple increased or decreased gene copy number. Clearly, additional studies are required to determine whether the amplified or deleted genes present in distinct CNV regions between HSF1 and HSF6 hESCs control differentiation potential, such as the impact of specific gain-of-function and loss-of-function alterations followed by NPC differentiation.

CONCLUSION

CNVs detected from somatic cells and tissues, like SNPs detected from similar sources, could represent variations of normality, or so-called "benign CNVs, " as opposed to "pathogenic CNVs" [42]. A benign CNV or SNP is the likely interpretation when a genomic imbalance is detected in an individual and that person's healthy parent(s). However, for most if not all hESCs, tests on the parent(s) of origin are not available, and the CNV in question may or may not affect not only the differentiation potential and function of a

particular hESC but also any derivative therapeutic cells or tissues. The observation that 16 of 34 (47%) of the genes listed in the consistent CNV intervals of Tables 1 and 2 are represented in the Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/omim) database, without even considering any noncoding RNAs within these CNV intervals, suggests that many copy number variant regions within HSF1 or HSF6 could have significant disease relevance as well.

ACKNOWLEDGMENTS

This study was supported by California Institute for Regenerative Medicine (CIRM) Predoctoral Training Grant T1-00005 (H.W.) and CIRM Seed Grant RS1-00313 (to M.A.T.). Work from the New York University/Courant Bioinformatics group was supported by two National Science Foundation (NSF) ITR grants and an NSF-EMT grant (to B.M.). M.A.T. is a Scholar of the Leukemia and Lymphoma Society (White Plains, NY). H.W. and K.J.K. contributed equally to this work.

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

K.M., T.D., J.G., and T.B. are employees of Agilent Technologies Inc.

REFERENCES

1 Amit M, Carpenter MK, Inokuma MS et al. Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. Dev Biol 2000;227:271–278.
2 Reubinoff BE, Pera MF, Fong CY et al. Embryonic stem cell lines from human blastocysts: Somatic differentiation in vitro. Nat Biotechnol 2000;18:399–404.
3 Thomson JA, Itskovitz-Eldor J, Shapiro SS et al. Embryonic stem cell lines derived from human blastocysts. Science 1998;282:1145–1147.
4 Baker DE, Harrison NJ, Maltby E et al. Adaptation to culture of human embryonic stem cells and oncogenesis in vivo. Nat Biotechnol 2007;25:207–215.
5 Draper JS, Smith K, Gokhale P et al. Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. Nat Biotechnol 2004;22:53–54.
6 Kallioniemi A, Kallioniemi OP, Sudar D et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992;258:818–821.
7 Uhrig S, Schuffenhauer S, Fauth C et al. Multiplex-FISH for pre- and postnatal diagnostic applications. Am J Hum Genet 1999;65:448–462.
8 Maitra A, Arking DE, Shivapurkar N et al. Genomic alterations in cultured human embryonic stem cells. Nat Genet 2005;37:1099–1103.
9 Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet 2006;7:85–97.
10 Freeman JL, Perry GH, Feuk L et al. Copy number variation: New insights in genome diversity. Genome Res 2006;16:949–961.
11 Hinds DA, Stuve LL, Nilsen GB et al. Whole-genome patterns of common DNA variation in three human populations. Science 2005;307:1072–1079.
12 Iafrate AJ, Feuk L, Rivera MN et al. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949–951.
13 Redon R, Ishikawa S, Fitch KR et al. Global variation in copy number in the human genome. Nature 2006;444:444–454.
14 Sebat J, Lakshmi B, Troge J et al. Large-scale copy number polymorphism in the human genome. Science 2004;305:525–528.
15 Pinkel D, Segraves R, Sudar D et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 1998;20:207–211.
16 Pollack JR, Perou CM, Alizadeh AA et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 1999;23:41–46.
17 Snijders AM, Nowak N, Segraves R et al. Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet 2001;29:263–264.
18 Solinas-Toldo S, Lampel S, Stilgenbauer S et al. Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. Genes Chromosomes Cancer 1997;20:399–407.
19 Aitman TJ, Dong R, Vyse TJ et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 2006;439:851–855.
20 Buckland PR. Polymorphically duplicated genes: Their relevance to phenotypic variation in humans. Ann Med 2003;35:308–315.
21 Gonzalez E, Kulkarni H, Bolivar H et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 2005;307:1434–1440.
22 Jongmans MC, Admiraal RJ, van der Donk KP et al. CHARGE syndrome: The phenotypic spectrum of mutations in the CHD7 gene. J Med Genet 2006;43:306–314.
23 Lupski JR, Stankiewicz P. Genomic disorders: Molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet 2005;1:e49.
24 McCarroll SA, Hadnott TN, Perry GH et al. Common deletion polymorphisms in the human genome. Nat Genet 2006;38:86–92.
25 Nguyen DQ, Webber C, Ponting CP. Bias of selection on human copy-number variants. PLoS Genet 2006;2:e20.
26 Repping S, van Daalen SK, Brown LG et al. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat Genet 2006;38:463–467.
27 Rovelet-Lecrux A, Hannequin D, Raux G et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat Genet 2006;38:24–26.
28 Shaw-Smith C, Redon R, Rickman L et al. Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. J Med Genet 2004;41:241–248.
29 Singleton AB, Farrer M, Johnson J et al. alpha-Synuclein locus triplication causes Parkinson's disease. Science 2003;302:841.
30 Perry GH, Dominy NJ, Claw KG et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet 2007;39:1256–1260.
31 Wu H, Xu J, Pang ZP et al. Integrative genomic and functional analyses reveal neuronal subtype differentiation bias in human embryonic stem cell lines. Proc Natl Acad Sci U S A 2007;104:13821–13826.
32 Daruwala RS, Rudra A, Ostrer H et al. A versatile statistical analysis algorithm to detect genome copy number variation. Proc Natl Acad Sci U S A 2004;101:16292–16297.
33 Ionita I, Daruwala RS, Mishra B. Mapping tumor-suppressor genes with multipoint statistics from copy-number-variation data. Am J Hum Genet 2006;79:13–22.
34 Lipson D, Aumann Y, Ben-Dor A et al. Efficient calculation of interval scores for DNA copy number data analysis. J Comput Biol 2006;13:215–228.
35 Adewumi O, Aflatoonian B, Ahrlund-Richter L et al. Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. Nat Biotechnol 2007;25:803–816.
36 Bignell GR, Huang J, Greshock J et al. High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res 2004;14:287–295.
37 Gunderson KL, Steemers FJ, Lee G et al. A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 2005;37:549–554.
38 Coe BP, Ylstra B, Carvalho B et al. Resolving the resolution of array CGH. Genomics 2007;89:647–653.
39 Greshock J, Feng B, Nogueira C et al. A comparison of DNA copy number profiling platforms. Cancer Res 2007;67:10173–10180.
40 Cory S, Adams JM. The Bcl2 family: Regulators of the cellular life-or-death switch. Nat Rev Cancer 2002;2:647–656.
41 Shianna KV, Willard HF. Human genomics: In search of normality. Nature 2006;444:428–429.
42 Lee C, Iafrate AJ, Brothman AR. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. Nat Genet 2007;39:S48–S54.


See www.StemCells.com for supplemental material available online.