Stem Cell Research (2012) 8, 154–164

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/scr

Evaluating the genomic and sequence integrity of human ES cell lines; comparison to normal genomes

Walter D. Funk [a,*], Ivan Labat [b], Janani Sampathkumar [a], Pierre-Antoine Gourraud [c], Jorge R. Oksenberg [c], Elen Rosler [a], Daniel Steiger [a], Nadia Sheibani [a], Stacy Caillier [c], Birgit Stache-Crain [b], Julie A. Johnson [d], Lorraine Meisner [d], Markus D. Lacher [a], Karen B. Chapman [a], Myung Jin Park [e], Kyoung-Jin Shin [e], Rade Drmanac [f], Michael D. West [a]

[a] *BioTime, Inc., Alameda, CA, USA*
[b] *The Ernst Gallo Clinic and Research Center, Emeryville, CA, USA*
[c] *Dept. of Neurosciences, UCSF, San Francisco, CA, USA*
[d] *Cell Line Genetics, Inc., Madison, WI, USA*
[e] *Dept. of Forensic Medicine, Yonsei University, Seoul, South Korea*
[f] *Complete Genomics, Inc., Mountain View, CA, USA*

Received 20 June 2011; received in revised form 29 September 2011; accepted 1 October 2011
Available online 8 October 2011

Abstract Copy number variation (CNV) is a common chromosomal alteration that can occur during *in vitro* cultivation of human cells and can be accompanied by the accumulation of mutations in coding region sequences. We describe here a systematic application of current molecular technologies to provide a detailed understanding of genomic and sequence profiles of human embryonic stem cell (hESC) lines that were derived under GMP-compliant conditions. We first examined the overall chromosomal integrity using cytogenetic techniques to determine chromosome count, and to detect the presence of cytogenetically aberrant cells in the culture (mosaicism). Assays of copy number variation, using both microarray and sequence-based analyses, provide a detailed view genomic variation in these lines and shows that in early passage cultures of these lines, the size range and distribution of CNVs are entirely consistent with those seen in the genomes of normal individuals. Similarly, genome sequencing shows variation within these lines that is completely within the range seen in normal genomes. Important gene classes, such as tumor suppressors and genetic disease genes, do not display overtly disruptive mutations that could affect the overall safety of cell-based therapeutics. Complete sequence also allows the analysis of important transplantation antigens, such as ABO and HLA types. The combined application of cytogenetic and molecular technologies provides a detailed understanding of genomic and sequence profiles of GMP produced ES lines for potential use as therapeutic agents.

* Corresponding author.
E-mail address: wfunk@biotimemail.com (W.D. Funk).

1873-5061/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.scr.2011.10.001

Introduction

The development of cell-based therapeutics for clinical applications will continue to require rigorous quality control of the chromosomal and genetic integrity. The process of *in vitro* propagation of human cells can often lead to a disruption of DNA integrity whether analyzed at the level of whole genomes, individual chromosomes or at base pair resolution. Trisomies 12 and 17 has been reported in long-term cultures of hESCs (Draper et al., 2004), while more recently, higher resolution methods for identifying copy number variation (CNV), such as comparative genomic hybridization (CGH) and array-based profiling, have reported a much broader spectrum of chromosomal amplifications and deletions that vary substantially with both cell strain and propagation history (Laurent et al., 2011; Hussein et al., 2011; Lefort et al., 2008; Spits et al., 2008). Whole genome resequencing is now being applied to the assessment of cultured cells and has shown that induced pluripotent stem (iPS) cell methods can lead to the accumulation of point mutations in the resulting clonal population (Gore et al., 2011). Similarly, iPS cells can display substantial differences in their epigenetic profiles compared to embryonic stem cells (Lister et al., 2011).

While the current methodologies for genome and sequence analysis provide unprecedented resolution for identifying potential variance in cell cultures, the impact of any specific variance may not be easily predicted. Aneuploidy is generally considered to be an unambiguous quality control issue for human cell-based therapies given that chromosomal aberrations can result in increased risk of oncogenic transformation (McClendon et al., 2008). Similarly, rare and common DNA variations that modulate disease risk should be assessed, particularly for genes whose functional disruption can lead to transformation. Of less obvious impact on the safety and performance of cellular therapeutics are smaller scale changes in gene copy number, coding sequence variance, or disease-associated alleles. Sequence variance in the genome occurs predominantly at the level of CNV or single nucleotide variance (SNV) and for cells intended for human therapies, normal genomes can provide a useful comparator.

Here we described a comprehensive analysis of chromosomal and gene variance in 5 human ES cell lines that were derived using GMP compliant methods (Crook et al., 2007). We first analyzed the overall karyotype using G-banding to ensure overall chromosomal integrity and then used FISH probes to detect the presence of emerging clones with altered chromosomal copy number. Short tandem repeat (STR) fingerprints were derived for each line and can be used to track long-term cultures and derivatives while telomere length analysis confirmed the embryonic potential of these lines for long term propagation. Our review of copy number variation applies two orthogonal methods and then compares the resulting variants to those seen in normal genomes. Similarly, we compare sequence variance directly to normal genomes to achieve a balanced review of non-conservative variance seen in specific gene classes, such as tumor suppressors, and in disease-associated alleles, such as *APOE*. By each of these measures, early passage cultures of these lines display variance that is entirely consistent with that seen in normal individuals. Finally, we review specific transplantation-related antigens that can now be assessed by sequence-based methods.

Results

hES cell culture

The ESI lines described here were derived with GMP-compliant derivation procedures and were initially expanded on human fibroblast feeder cultures (Crook et al., 2007). We routinely culture hESC lines in ES culture medium (mTeSR1; StemCell Technologies); an assessment of hESC-related cell surface markers showed high level expression in ESI lines cultured on Matrigel (Table S1).

Analysis strategy

Our assessment of hESC cultures applied both cell-based and DNA-based analysis to the preparations. Genomic DNA samples are collected for multiple applications, including STR fingerprinting, telomere length analysis, CNV assessment and complete genome sequencing (Fig. 1). Cell cultures are used for cytogenetic assessment. G-band karyotyping provides an overall review of chromosome count and integrity, and can be particularly useful in defining several chromosomal aberrations, including balanced translocations and inversions, that can be difficult to detect by SNP-based techniques. All 5 ESI lines tested showed a normal human karyotype, with 4 female (XX) and 1 male lineages (Figs. 2A, B). To better assess culture mosaicism, we applied FISH probes to interphase nuclei that detect the short arm of chromosome 12 (ETV-6) and the centromere of chromosome 17, two of the most widely reported aneuploidies seen in cultured human cells (Meisner and Johnson, 2008). Although four of five ES cultures had a normal diploid signal pattern for both probes, culture ESI 051 demonstrated three 12p13 (ETV6) signals in 4% OF 200 interphase nuclei, consistent with an emerging clone with trisomy 12 (Figs. 2A, C). We did not observe copy number variation at the ETV-6 locus for ESI051, using either SNP-based arrays or sequencing-derived methods, as would be expected given the low frequency of these cells in the population. Homogenous assays, such as those applied to DNA prepared from cell culture, have a limited ability to detect low levels of aberrant cells within a culture compared to FISH.

STR fingerprinting has become the international standard procedure for tracking cellular identity (Schweppe et al., 2008) and is particularly useful in manufacturing, where molecular authentication of derivative cell lines is required. Each of the ESI lines provides a unique identifier and as expected, H9 was returned as an identity hit when the STR-based fingerprint was queried against the NIH Embryonic Stem Cell Registry (http://grants.nih.gov/stem_cells/registry/current.htm) (Fig. 2A).

We routinely assess ES and primary cell cultures for telomere length since this parameter directly correlates with replicative lifespan in cultured somatic cells (Allsopp et al., 1992). As expected, all hESC lines tested express active telomerase (data not shown) and have mean telomere lengths in the 8–20 kbp range, typical of other hESC lines and sperm (Figs. 2A, D).

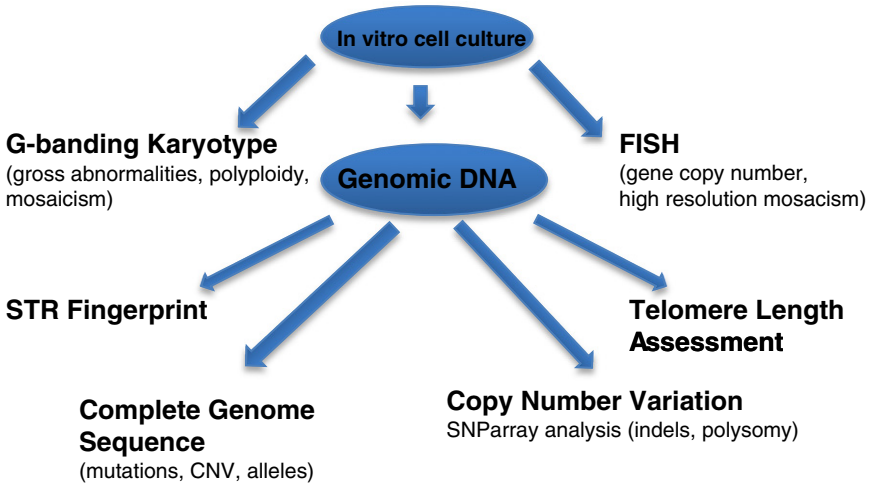

Figure 1 Overview of ES cell analysis strategy. ES lines ESI017, 035, 049, 051 and 053 were expanded using feeder-free culture from early passage frozen stocks. Cultures were evaluated for G-banded karyotype, STR fingerprinting and FISH-based assessment of chromosomes 12 and 17 (Methods). High molecular weight DNA was harvested for telomere length analysis, copy number variation determination and complete genome sequencing.

Copy number Variation in ESI cell lines

We evaluated CNVs in the ESI lines using two methods: SNP-based arrays (HumanCytoSNP-12, and GenomeStudio Genotyping Module, Illumina) and whole genome sequence-based methods (Complete Genomics, Inc.).

SNP-called CNV

The average spacing of this array (300K SNP) is approximately 10 kbp and as such, can be used to identify variants that span 50 kbp or larger. Call rates for all samples averaged >99% (data not shown). We reviewed CNVs that had high

A

Cell Line	Passage	Karyotype	STR fingerprint*	G-band (abnormal/normal)	p12, 17cen FISH (abnormal/ normal)	Mean TRF (kbp.)
ESI017	28	46, XX	Unique	0/20	0/200	13.7
ESI035	22	46, XX	Unique	0/20	0/200	14.9
ESI049	24	46, XY	Unique	0/20	0/225	14.0
ESI051	23	46, XX	Unique	0/20	9/225 p12 trisomy	9.8
ESI053	20	46, XX	Unique	0/20	0/200	14.5
H9	32	46, XX	Match-H9	0/20	0/300	15.0

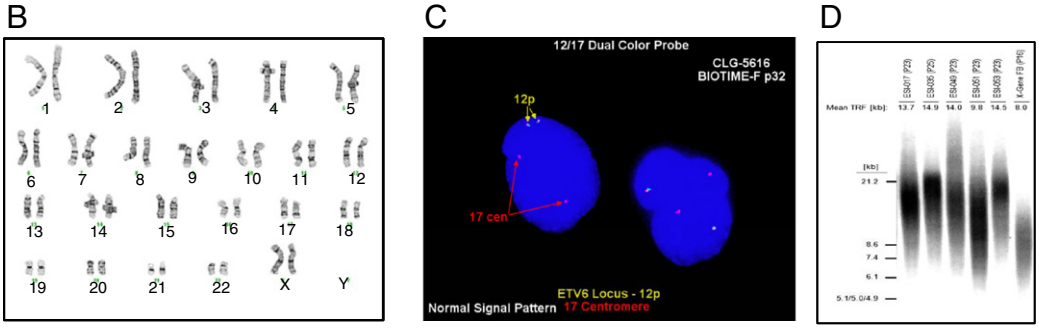

Figure 2 Karyotype, FISH, STR and TRF analyses. A: Summary of results for 5 ESI lines and H9. STR fingerprint results were compared NIH Embryonic Stem Cell Registry (http://grants.nih.gov/stem_cells/registry/current.htm) and the ATCC Cell Line Repository (www.ATCC.org). H9 scored the appropriate identity hit registered with ATCC. B: Apparently normal G-banded karyotype for cell line ESI017. C: Representative diploid interphase nuclei from cell line ESI017 using 12p13 (ETV6) and 17 centromere FISH probes. D: Southern blot analysis of TRF length. ESI lines show telomere lengths appropriate for embryonic lines, while normal dermal fibroblasts (XGene) show significantly shortened TRFs.

A

Cell Line	Passage	Chr	Start	End	Mb	CNV Value	CNV Conf	Genes	DGV Overlap?	Tumor Suppressor/Oncogene List?	Cartilage Disease List?
ESI 017	19	X	6537147	8125052	1.59	3	272	HDHD1A...VCX2	Y	No	No
ESI 017	19	X	1.23E+08	1.23E+08	0.32	3	328	STAG2, THOC2, XIAP	Y	No	No
ESI 017	24	X	6582336	8132359	1.55	3	339	HDHD1A... VCX2	Y	No	No
ESI 017	24	X	1.23E+08	1.23E+08	0.34	3	311	STAG2, THOC2, XIAP	Y	No	No
ESI 035	25	11	50283886	51167497	0.88	1	255	LOC646813	Y	No	No
ESI 035	25	16	2826398	3203938	0.38	3	293	CCDC64B... ZSCAN10	Partial	No	No
ESI 035	25 - Tech Dup	11	48288315	51249292	2.96	1	400	FOLH1... OR4C45	Y	No	No
ESI 035	25 - Tech Dup	16	2826398	3253961	0.43	3	401	CCDC64B... ZSCAN10	Partial	No	No
ESI 051	19	11	48343394	51398585	3.06	1	379	FOLH1...OR4C46	Y	No	No
ESI 051	19	19	23623651	23634403	0.01	0	112	ZNF675	N	No	No
ESI 051	23	8	1.38E+08	1.38E+08	0.17	0	143	NONE	Y	No	No
ESI 051	23	11	48330522	51398585	3.07	1	825	FOLH1... OR4C46	Y	No	No
ESI 053	20	8	1.38E+08	1.38E+08	0.12	1	159	NONE	Y	No	No
ESI 053	20 -Tech Dup	8	1.38E+08	1.38E+08	0.16	0	189	NONE	Y	No	No
ESI 049	26	none						NONE	N/A	No	No
ESI 049	30	none						NONE	N/A	No	No
ESI 049	36	17	0	18753870	18.8	LOH	906	CSF2RA....PPEF1	N/A	Yes	Yes
H9	33	7	62127394	62531632	0.40	3	447	PRKAR1B	N	No	No
H9	33	14	63569297	63799797	0.23	3	182	SYNE2	N	No	No
H9	49	7	62127394	62531632	0.40	3	481	PRKAR1B	N	No	No
H9	49	14	63475449	63799797	0.32	3	173	SYNE2	N	No	No
H9	49	17	0	6882109	6.88	3	279	CSF2RA....VCX3A	N/A	Yes	Yes
H1	34	7	61078136	62112668	1.03	1	152	None	N	No	No
H1	34	10	46392146	47230072	0.84	3	123	STYT15....ANXA8L2	Y	No	No
H1	34	11	73140406	73397622	0.26	3	177	RAB6A...PAAF1	N	No	No

B

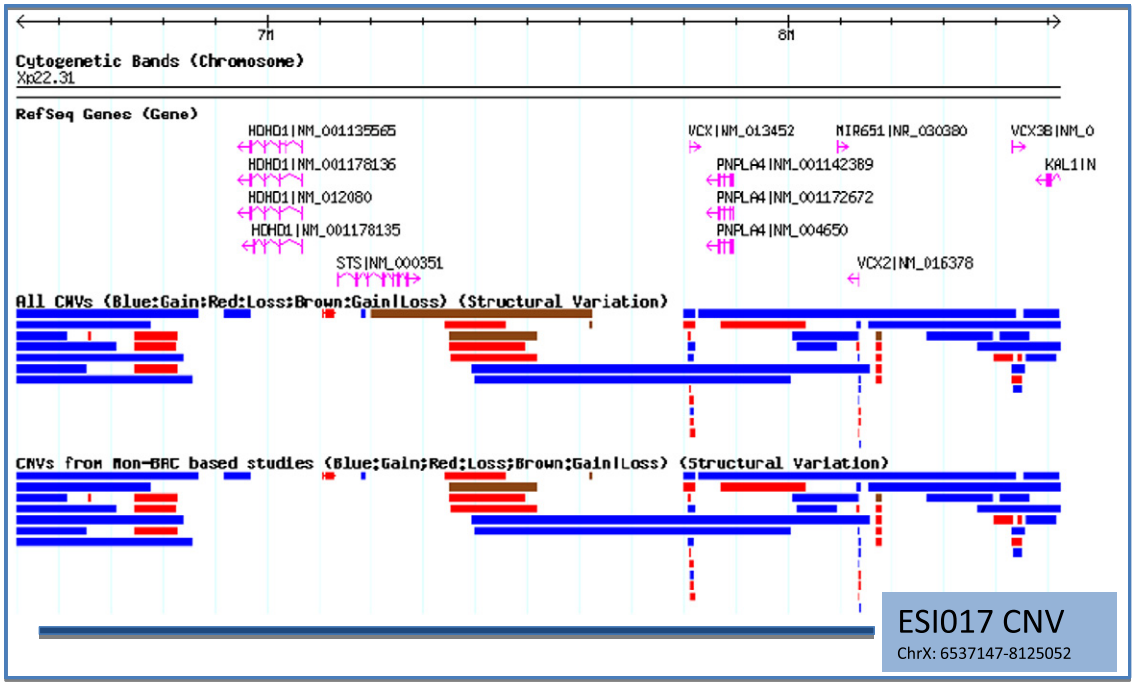

Figure 3 Copy number variance (CNV) from SNP arrays. A: Summary of SNP-based analysis of copy number variance. CytoSNP12 arrays (Illumina) were used to assess high confidence CNVs in ESI lines. Some lines were assessed at several different passages. Genomic intervals are indicated along with affected genes or gene intervals. Overlap of these intervals with those reported for normal human genomes in the Database of Genomic Variants (DGV) (Zhang et al., 2006) is indicated. Overlap with gene sets for tumor suppressors, oncogenes, or OMIM cartilage disease genes (Fig. S2) are indicated. B: Representation of Database of Genomic Variants (DGV) coverage for a chromosome X CNV identified in ESI017. Cytogenetic position, overlapping genes (pink) and CNVs seen in normal human genomes are indicated for this interval (blue: gain; red: loss; brown: gain and loss).Copy number variance (CNV) from SNP arrays. A: Summary of SNP-based analysis of copy number variance. CytoSNP12 arrays (Illumina) were used to assess high confidence CNVs in ESI lines. Some lines were assessed at several different passages. Genomic intervals are indicated along with affected genes or gene intervals. Overlap of these intervals with those reported for normal human genomes in the Database of Genomic Variants (DGV) (Zhang et al., 2006) is indicated. Overlap with gene sets for tumor suppressors, oncogenes, or OMIM cartilage disease genes (Fig. S2) are indicated. B: Representation of Database of Genomic Variants (DGV) coverage for a chromosome X CNV identified in ESI017. Cytogenetic position, overlapping genes (pink) and CNVs seen in normal human genomes are indicated for this interval (blue: gain; red: loss; brown: gain and loss).

confidence scores (>100) as this cut-off consistently captured the same CNVs in technical replicates (data not shown). Using this sift, the majority of CNVs identified in the ESI lines were larger than 100 kb. Early passage cultures of the ESI lines typically displayed fewer than 5 high-confidence CNVs with sizes ranging from 0.2 to 2 Mbp (Fig. 3A). These CNVs were found to be predominantly within regions previously annotated in databases of variance found in normal human genomes (Database of Genomic Variants (DGV); Zhang et al., 2006). For example, a 1.59 Mbp CNV seen in early passage cultures of ESI017 covers a region in which both an increase and decrease in copy number is reported in DGV (Fig. 3B). The total number of potentially affected genes in this analysis of the ESI lines ranged from 0 to 20 and similar CNVs and affected gene numbers were seen in early passages of the ES lines H1 and H9 (Fig. 3A).

The CNVs seen in early passage cultures of the hESC lines were retained in later passage cultures but new variance arose with longer-term culture. For example, later passage ESI049 cultures (P36) showed a loss of heterozygosity (LOH) of the q arm of Chr 17, while a substantial amplification of q17 was also seen in later passage (p49) H9 cultures (Fig. 3A), alterations that have been reported in both cytogenetic and array-based assessments of long-term cultures of human cells (Laurent et al., 2011; Draper et al., 2004). We reviewed the potential overlap of all CNVs with subsets of genes, including 169 identified tumor suppressor genes and 198 oncogenes (http://www.binfo.ncku.edu.tw/TAG/GeneDoc.php) (Table S2). In these early passage cultures, we did not observe CNVs that overlapped with any of the tumor suppressor or oncogene exonic sequences (Fig. 3A). As a further demonstration of this approach, we analyzed subsets of genes that would be predicted to be impactful on the development of a tissue-specific cellular therapeutics. We extracted over 70 genes from OMIM (http://www.ncbi.nlm.nih.gov/omim/) reported to be mutated in human genetic diseases affecting cartilage development, maintenance and disease propensity (Table S2). Again, no early passage cultures had CNVs that affected the gene count for this set.

Genome sequence-called CNVs

Effectively, CNVs can be estimated by comparing read-depth calls to those expected from the reference genome (Methods). Sequence-called CNVs confirmed several of the SNP-called CNVs with the exception of those that were eliminated due to genome build formats (SNP data was Build 36, genome data was from Build 37), or from chromosomal locations not covered by SNPs in the array format (data not shown). The sequence-based CNV algorithms assess read depths on a 1 kbp scan window and thus can identify CNVs at finer resolution than the CytoSNP12 arrays used in this study. To provide a relevant metric for this analysis, we compared CNVs in the sequenced ESI cell lines with those identified in 46 genomes from normal human donors sequenced on the same platform (Complete Genomics, Table S3). The overall size distribution of CNVs was similar between ESI lines and human genome samples, with the majority of CNVs falling in the 5–20 kbp range (Fig. 4A). The chromosomal distribution of CNVs in the ESI lines appeared relatively unbiased and did not differ in size or distribution from that of the normal genome samples (Fig. 4B). Furthermore, the majority of CNVs detected by this analysis occurred in regions defined either as segmental duplication, or low complexity, such as LINE, SINE and LTR sequences and were annotated in the database of genomic variance (Table S4).

Complete genome sequence

We performed complete genome re-sequencing of the 5 ESI lines using a commercial adaptation of the combinatorial probe anchor ligation (cPAL) chemistry and DNA nanoballs; variants were called and scored using a local *de novo* assembly methodology (Complete Genomics, Inc; Drmanac et al., 2010). Re-sequencing maps raw reads to a reference genome (NCBI Build 37) and reports variance compared to the haploid genome reference. We concentrated our initial analysis on exonic and coding (cds) sequences as these are most typically disrupted in human genetic disease. Haploid read

A

CNV SizeRange (kbp)	ESI017	ESI035	ESI049	ESI051	ESI053	Normal Genome Median*
>1, <=5	27	25	30	23	23	23
>5, <=20	113	100	105	118	113	113
>20, <=100	45	41	40	48	38	43
>100, <=1000	7	8	3	6	6	5
>1000	1	0	0	0	0	0

B

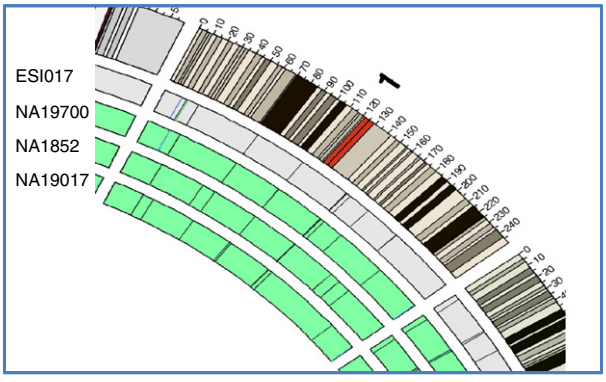

Figure 4 CNV analysis from complete genome sequencing. A: CNVs were extracted and sized from the genome sequence. Bins represent CNV interval length. Comparison to CNVs from sequence analysis of 46 normal human genomes (*Table S4) is indicated. B: Circos diagram (Krzywinski et al., 2009) representing the distribution of CNVs identifies in on Chr 1 for ESI017 and 3 normal human genomes. Bar width is proportional to CNV size interval. Outer rim shows G-banding for Chr1.

coverage and percent genome coverage were similar to those reported previously while the % of fully called sequence for exonic and cds sequences was 96–97% (Table S5). Since our sequence analysis is compared only to the reference build, we again used genome sequences from normal donors (Table S3) sequenced using the identical platform for comparison of variant profiles. To identify potentially disruptive variants in the exonic regions of the genomes, we used ANNOVAR (Wang et al., 2010) an informatics software tool, to identify and sort potential sequence variants (Fig. 5). In comparison to 46 normal human genomes analyzed using the same sequencing platform, the early passage ESI lines displayed a frequency of potentially disruptive exonic variations, including frame-shift insertion/deletions, nonsense substitutions, splice donor/acceptor changes and non-synonymous substitutions that fell within the normal genome range (Fig. 6). We then asked whether potentially disruptive variants fell within important gene classes. Similar to the results seen in the complete genome exome, non-synonymous substitutions were the predominant variants identified in the analysis of tumor suppressor genes (Table S2) (Fig. 6), however the impact of substitutions that do not directly alter coding sequence can be difficult to assess by informatics alone. Frameshift or nonsense mutations could be impactful in terms of tumor suppressor gene function and we identified a single frameshift variant affecting tumor suppressor genes in ESI017 (*EGR1*), ESI053 (*AIM2*) and ESI049 (*COL18A1*) (Fig. 6). Two of these genes (*EGR1*, *COL18A1*) have been identified as loss of copy number variants in normal genomes (DGV; data not shown). This would suggest that even complete loss of a single copy of these genes may be less likely to be considered impactful. Similarly, the distribution of variance in a large collection of defined oncogenes (Table S2) showed similar substitution frequencies in ESC and normal genomes (Fig. 6).

ABO, ApoE and HLA variants

Complete genome sequence can also serve to provide initial information on disease propensity alleles and transplantation-related markers of cellular therapeutics. The ABO blood type is defined by the 3 alleles of the ABO gene (Fig. 7A). A review of these in the sequence of the ESI lines showed unambiguous allele calls reflecting OO, AO and AA blood types. These were confirmed using a multiplex PCR assay (Lee et al., 2011)

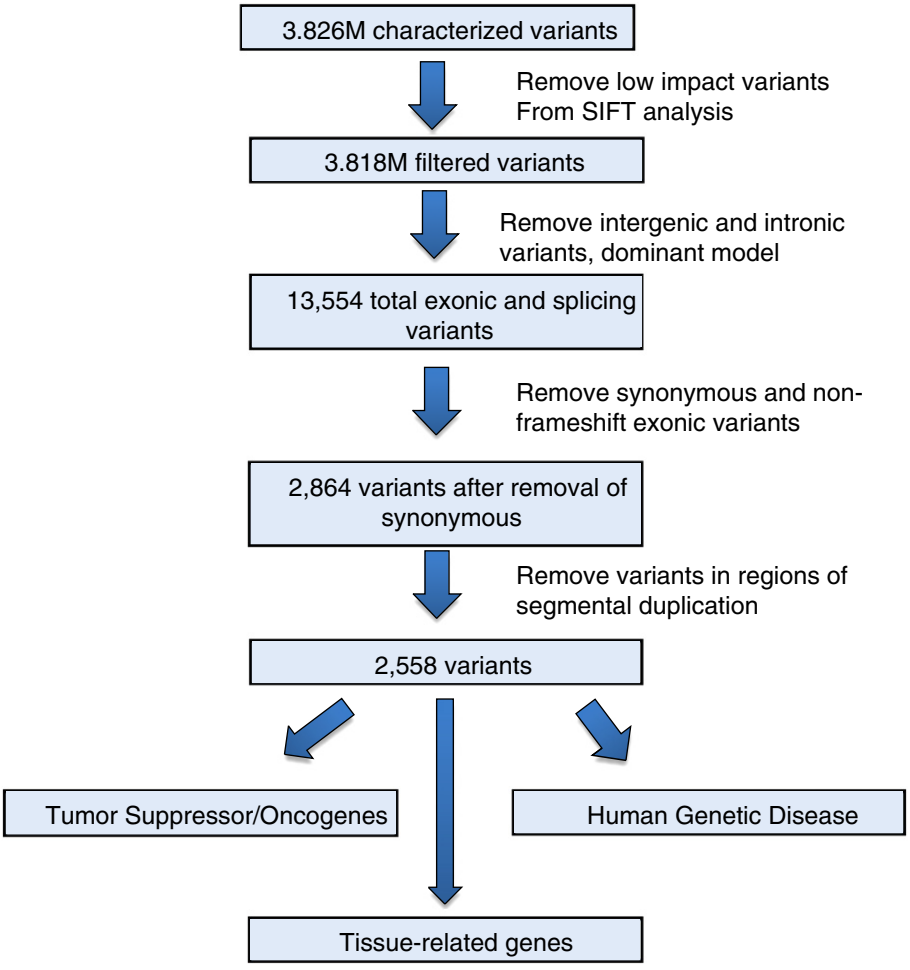

Figure 5 Sequence variance analysis strategy. Sequence variants that differ from the reference build (NCBI Bld37) were assessed using a modified version of ANNOVAR (Wang et al., 2010). SIFT (Ng and Henikoff, 2001) identifies and removes low impact variants. Exonic variants are then contrasted against specific gene sets.

Exome Variance	Normal Genome* Range(freq)	ESI017	ESI035	ESI049	ESI051	ESI053
frameshift deletion	63-102	69	86	74	76	78
frameshift insertion	97-134	115	122	111	119	105
frameshift substitution	9-25	16	18	12	16	16
nonsynonymous SNV	1916-2672	2011	2018	2071	2020	1973
splicing	141-232	145	162	166	170	157
stopgain SNV (nonsense)	28-57	48	51	41	54	45
Total genes affected (ave)	2110	1911	1922	1971	1933	1886
Genes with 1 mutation(ave)	1697	1560	1537	1611	1564	1533
Genes with >1 mutation (ave)	413	351	385	360	369	353
Tumor Suppressor Variance						
frameshift deletion	0-1 (9/46)	0	0	1 (COL18A1)	0	1 (AIM2)
frameshift insertion	0-1 (9/46)	1 (EGR1)	0	0	0	0
frameshiftsubstitution	0-1 (2/46)	0	0	0	0	0
nonsynonymous SNV	9-26 (46/46)	17	11	20	14	15
splicing	0	0	0	0	0	0
stopgain SNV (nonsense)	0-2 (2/46)	0	0	0	0	0
Total genes w variant	8-24	17	11	17	14	16
Genes with 1 variant	7-22	16	11	14	14	16
Genes with >1 variant	0-4	1	0	3	0	0
Oncogene Variance						
frameshift deletion	0-1 (3/46)	0	0	0	0	0
frameshift insertion	0-1 (13/46)	0	0	0	0	0
frameshiftsubstitution	0-1 (3/46)	0	0	0	0	0
nonsynonymous SNV	4-21 (46/46)	9	3	13	5	11
splicing	0	0	0	0	0	0
stopgain SNV (nonsense)	0-1 (2/46)	0	1	0	1	0
Total genes affected	8-22 (46/46)	11	8	11	11	7
Genes with 1 mutation	7-17 (46/46)	10	8	10	11	6
Genes with >1 mutation	0-4 (39/46)	1	0	1	0	1

Figure 6 Summary of exonic sequence variants from analysis strategy (Fig. 5). Sequence variants of the indicated type were scored for each ESI line in comparison to a range (or average) of the same variant types in 46 normal human genomes (*Table S3), including total exonic variants, exonic variants found in tumor suppressor genes (Fig. S2), and exonic variants found in protooncogenes (Fig. S2). For frameshift and stopgain (nonsense) variants in the ESI lines, the affected genes are indicated.

(Figs. 7C, S1). Given the significant impact of ApoE genotypes on cardiovascular and Alzheimer's disease propensities, we reviewed the ESI line genomes for the 3 major ApoE alleles (E2, E3, E4). The processed sequence suggested calls for cell lines ESI017 and 051 (Fig. 7A), while half calls and no calls prevented the determination of the genotypes at the high confidence threshold for the remaining lines. A review of the evidence files for these genomes revealed a very low read depth across these positions (4–6 reads) (data not shown). We performed multi-plex PCR-based analysis of *APOE* alleles and confirmed the determination calls for the two lines and assigned ApoE genotypes for the remaining lines (Figs. 7B, S2). None of the ESI lines, nor H1 and H9, carry the potentially deleterious ApoE4/E4 genotype.

Similarly, sequence calls across the classical HLA loci were insufficient for unambiguous typing, as both the complex variance at these loci and issues of phase determination represent significant technical hurdles for this application. Amplification and re-sequencing of the genomic loci for the *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1* and *HLA-DQB1* genes provided allele calls for each cell line (Fig. 7D). Three of the most frequently occurring HLA haplotypes in American populations (Maiers et al., 2007) are compatible with the observed phenotypes of ESI017 (A*01:01–B1*08:01C*07:02 DRB1*03:01), ESI035 (A*03:01–B*35:01C*04:01 DRB1*01:01), and ESI051, 053 (A*03:01–B*07:02C*07:02 DRB1*15:01).

Discussion

We report here a set of methodologies intended to provide a complete and thorough review of the genome integrity of cellular therapeutics. Our approach includes a review of the chromosomal integrity and for the first time, a careful review of the genome sequence integrity of cell lines derived using GMP-compliant processes and documentation. Importantly, we have compared the occurrence of variance in terms of type, size and distribution and conclude that at early passage, hESCs are entirely consistent with variance seen in normal human genomes. Given the importance of cellular integrity for cell therapeutic applications, reviewing the genome and sequence integrity for these same products will likely become a standard requirement for non-autologous therapeutics.

Extended *in vitro* expansion of hESC and other cultured human cells most often results in large scale chromosomal abnormalities, as has been documented using cytological methods (Draper et al., 2004) and more recently, SNP and

A

Chromosome	Offset Based	Genotype Strand	Alleles	Annotation	Reference	ESI017	ESI035	ESI049	ESI051	ESI053
chr9	136131314	-	C/G	268 G-Gly, C-Ala (rev)	C	CC	CC	CC	CC	CC
chr9	136131591	+	C/G	176 C-Arg, G-Gly (rev)	G	GG	GG	GG	GG	GG
chr9	136132908	-	-/G	87 frameshift (rev)	-	--	CC	CC	C-	C-
					ABO type	OO	AA	AA	AO	AO
chr19	45411940	+	C/T	112 (130) T-Cys, C-Arg (fwd)	T	TT	NN	NN	CT	NT
chr19	45412078	+	C/T	158 (176) C-Arg, T-Cys (fwd)	C	CC	CC	CC	CC	CC
					APOE	e3e3	e3 or e4	e3 or e4	e3e4	e3e3 or e3e4

B

Cell Line	rs429358 (aa112) T/C	rs7412 (aa158) C/T	Genotype
ESI-017	TT	CC	E3/E3
ESI-035	TC	CC	E3/E4
ESI-049	TT	CC	E3/E3
ESI-051	TC	CC	E3/E4
ESI-053	TC	CC	E3/E4
H1	TT	CC	E3/E3
H9	TC	CC	E3/E4

C

Cell Line	261	771	796	802	803	ABO type
ESI-017	del	C/T	C	G	G	O01O02
ESI-035	G	C	C	G	G	AA
ESI-049	G	C	C	G	G	AA
ESI-051	G/del	C/T	C	G	G	AO02
ESI-053	G/del	C/T	C	G	G	AO02
H1	del	C/T	C	G	G	O01O02
H9	G/del	C	C	G	G	AO01

D

Cell line	A1	A2	B1	B2	C1	C2	DRB11	DRB12	DQB11	DQB12
ESI017	**0101g**	*3201*	**0801g**	*1501g*	*0303g*	**0701g**	**030101**	*040101*	**0201g**	*030201*
ESI035	0301g	1101g	3501g	4402g	0401g	0704g	010101	130201	050101	0604g
ESI049	0201g	2402g	4101	5101g	0701g	140201	030101	1401g	0201g	050301
ESI051	**0101g**	0301g	0702g	**380101**	0702g	**1203**	150101	1401g	050301	060201
ESI053	**0101g**	3201	1401	**380101**	0802	**1203**	150101	1401g	050301	060201

Figure 7 Evaluation of ABO, ApoE and HLA allelic variants. A: Genome sequence was evaluated directly for ABO and ApoE allele status. ABO alleles could be called unambiguously from the sequence reads. ApoE alleles for ESI 017 and 051 were called, alleles for ESI035, 049 and 053 were not called due to low read coverage. B: TaqMan assay results for ApoE alleles (E2, E3, E4) (Fig. S2). C: Multiplex PCR assay results for ABO blood type alleles (Fig. S1). D: HLA typing of 10 major HLA alleles. Genomic DNA was amplified across 10 major HLA alleles and re-sequenced.

FISH-based methodologies (Laurent et al., 2011; Caisander et al., 2006). We confirm here the propensity of later stage cultures to show examples of classical karyotypic abnormalities, such as chromosome 12 and 17 aneuploidy. Importantly, cytological methodologies can detect these variances even when present in a small minority of cells in a culture and provide early evidence of cultures that begun to drift (Meisner and Johnson, 2008). The application of multiple FISH probes should allow an even sharper detection of the emergence of common chromosomal variants in cultured populations.

Array and complete genomic hybridization-based methods are now being used to detect deletions and amplifications that are considerably shorter than the resolution of cytogenetic techniques. Our data provides additional evidence of accrual of CNVs with extended culture, potentially as a result of replicative stress and culture adaptation (Narva et al., 2010; Baker et al., 2007). We applied two orthogonal methods (SNP array and sequence-based detection) to identify these variances. Replicate analyses using array-based methods showed good reproducibility for high confidence CNVs, while sequence-based methods confirmed some, but not all CNVs detected by arrays, suggesting that alternate methods can provide complementary information. Importantly, the vast majority of CNVs identified in these early passage cultures have been previously described in databases that curate CNVs in normal human genomes (Iafrate et al., 2004) and these sequence intervals typically show low sequence complexity, as demonstrated by the preponderance of LINE elements and regions of segmental duplication in these intervals. Sequence-based detection of CNVs allowed for the direct comparison of hESC lines to normal human genomes and by all measure of size, distribution and affected genes, CNVs found in these early passage ES cultures were entirely within the range of normal genome variation.

Complete genome sequencing methodologies and annotation provide sequence resolution that allows for the initial detection and cataloging of variance, although the accuracy or

completeness of calling a diploid genome is limited by the requirement for relatively deep sequence coverage. Clearly, orthogonal methods that confirm or complete sequence-based variance determinations remain a necessity. We confined our initial analyses to variance that could reliably be predicted to affect the expression of gene products from protein-encoding genes.

As opposed to iPS cell studies, where *de novo* mutations that have been cataloged between progenitor and iPS derivatives (Gore et al., 2011), hESC lines have not yet been directly compared to the genomes of the donor embryos (nor can they be compared to the parental IVF donors). However we show that by all measures of nucleotide-level variance, these early passage ES lines show variance that is entirely within the range seen in normal human genomes. hESC lines have been shown to have effective means to ensure genome integrity, such as enhanced apoptosis of mutated cells and enhanced mitotic recombination (Filion et al., 2009; Hong et al., 2007), however maintenance of this integrity is challenged continually in long term *in vitro* cultures. The analysis of genome and sequence integrity can be applied to select an appropriate progenitor line, and should then be repeated at regular intervals during the expansion and derivation of cell populations.

Further focusing these analyses on gene sets that are of specific concern for developing therapeutic products (tumor suppressor/oncogenes), we see no appreciable difference in the variance found in hESC *vs*. normal human genomes. This same type of analysis can be applied to inherited disease genes (OMIM), disease propensity alleles (GWAS (Hindorff et al., 2009)), or mutation databases (HGMD) (Stenson et al., 2009) and would then allow for the selection of cell lines that lack deleterious variants or mutations. The evaluation of *APOE* alleles is a prime example of this approach. Although the direct cellular mechanisms by which the E4/E4 allelic combination contributes to late onset Alzheimer's disease are likely complex, clearly avoiding this combination when choosing a cell line for developing neuronal applications would seem prudent.

We demonstrate here that the evaluation of transplantation-related alleles, such as ABO and HLA can also be evaluated directly from sequence data and this can allow for the selection of specific cell lines for specific therapeutic applications (*e.g.* Blood type OO for the development of erythrocyte progenitors). In spite of high reported accuracy of the cPAL sequencing methods (Drmanac et al., 2010), orthogonal methods that confirm or complete sequence-based variance remain a necessity until an expected clinical genome quality is achieved in this rapidly developing field.

Experimental procedures

Cell culture

The ESC lines ESI017, 035, 049, 051 and 053 (Crook et al., 2007) and H1, H9 were cultured from research-grade cell banks on matrigel-coated culture plastic (Ludwig et al., 2006; Rosler et al., 2004). Briefly, hESC lines were cultured on tissue culture plastic coated with Growth Factor Reduced Matrigel (Becton-Dickenson, Inc) in mTeSR1™ medium (Stem Cell Technologies, Inc.) hESC were passaged weekly (1:3 or 1:4 split ratio) using 5 mM EDTA (Invitrogen, Inc.) or Accutase (Innovative Cell Technologies, Inc.), and supplemented with 10 μM of Rock Inhibitor, Y-27632 (Tocris Bioscience) at plating. FACS-based assessment of markers was performed using an Accuri C6 Flow Cytometer with CSampler and CFlow Sampler Software (Accuri) and antibodies listed in Table S1.

Cytogenetic analyses

Cultures of ESI lines were evaluated at Cell Line Genetics, Inc. (Madison, WI) for G-banded karyotype analysis and FISH assessment of interphase nuclei using probes for ETV6 (12p) and the chromosome 17 centromere as described (Meisner and Johnson, 2008).

DNA preparation

Total genomic DNA was prepared from by using the QIAGEN Blood & Cell Culture DNA Kit (Qiagen).

STR fingerprinting

Analysis of genomic DNA for 15 STR loci plus Amelogelin was performed at Cell Line Genetics, Inc. (Madison WI).

TRF analysis

Mean TRF lengths were determined by non-radioactive Southern blotting using the *Telo**TAGGG* Telomere Length Assay kit (Roche).

SNP array CNV analysis

Genomic DNA samples were analyzed at the Biomedical Genomics Center usingCytoSNP12 BeadChip kits (Illumina) using the BeadStudio2009.2 software (Illumina). Call rates were greater than 99.1%. Normalized intensity ratios (logR), B allele frequencies (BAF) and CNVs were extracted using the GenomeStudio Genotyping Module (Illumina). CNVs with confidence scores greater than or equal to 100 were included in subsequent analyses.

Genome sequencing

Genomic DNA samples were analyzed at Complete Genomics, Inc. (Mountain View, CA). Paired end library preparation and sequence-by-ligation methods were applied as described (Drmanac et al., 2010). Reads were mapped to the NCBI reference genome (Build 37) and data for each genome sample were provided as listing of sequence variants relative to the reference genome.

Genome sequence CNV analysis

The "cnvSegmentsBeta files" from the Complete Genomics data files were first filtered to remove potential CNVs that were annotated as having "invariant counts": (non diploid copy number that is invariable in normal genomes), or "hypervariable counts" (highly variable in normal human genomes) or were from mitochondrial sequence. The remaining calls indicating an increase or decrease from the reference copy number were tabulated.

ABO genotyping

Genomic DNA samples were analyzed using multiplex allele-specific primer sets for the ABO gene as described (Lee et al., 2011).

ApoE genotyping

Genomic DNA samples were analyzed using TaqMan primers (Roche) for alleles rs7412 and rs429358 according as described (Julian et al., 2009) using an ABI7500 BioAnalyzer.

HLA typing

Genomic DNA samples were analyzed by high resolution re-sequencing of the indicated loci (Histogenetics, Inc.) as described (Noreen et al., 2001).

Supplementary materials related to this article can be found online at doi:10.1016/j.scr.2011.10.001.

Conflict of interest

W.D.F., J.S, E.R., D.S., N.S., M.D.L., K.B.C., and M.D.W. are employees of BioTime Inc.

J.J. and L.M. are employees and shareholder of Cell Line Genetics, Inc.

R.D. is Chief Scientific Officer and Founder of Complete Genomics, Inc., and owns a substantial number of its shares and stock options.

Acknowledgments

We acknowledge and thank Archana Deshpande (University of Minnesota) for CytoSNP12 array processing.

References

Allsopp, R.C., Vaziri, H., Patterson, C., Goldstein, S., Younglai, E.V., Futcher, A.B., Greider, C.W., Harley, C.B., 1992. Telomere length predicts replicative capacity of human fibroblasts. Proc. Natl. Acad. Sci. U. S. A. 89, 10114–10118.

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

Caisander, G., Park, H., Frej, K., Lindqvist, J., Bergh, C., Lundin, K., Hanson, C., 2006. Chromosomal integrity maintained in five human embryonic stem cell lines after prolonged *in vitro* culture. Chromosome Res 14, 131–137.

Crook, J.M., Peura, T.T., Kravets, L., Bosman, A.G., Buzzard, J.J., Horne, R., Hentze, H., Dunn, N.R., Zweigerdt, R., Chua, F., Upshall, A., Colman, A., 2007. The generation of six clinical-grade human embryonic stem cell lines. Cell Stem Cell 1, 490–494.

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

Drmanac, R., et al., 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327, 78–81.

Filion, T.M., Qiao, M., Ghule, P.N., Mandeville, M., van Wijnen, A.J., Stein, J.L., Lian, J.B., Altieri, D.C., Stein, G.S., 2009. Survival responses of human embryonic stem cells to DNA damage. J. Cell. Physiol. 220, 586–592.

Gore, A., Li, Z., Fung, H.L., Young, J.E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M.A., Kiskinis, E., Lee, J.H., Loh, Y.H., Manos, P.D., Montserrat, N., Panopoulos, A.D., Ruiz, S., Wilbert, M.L., Yu, J., Kirkness, E.F., Izpisua Belmonte, J.C., Rossi, D.J., Thomson, J.A., Eggan, K., Daley, G.Q., Goldstein, L.S., Zhang, K., 2011. Somatic coding mutations in human induced pluripotent stem cells. Nature 471, 63–67.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., Manolio, T.A., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. 106, 9362–9367.

Hong, Y., Cervantes, R.B., Tichy, E., Tischfield, J.A., Stambrook, P.J., 2007. Protecting genomic integrity in somatic cells and embryonic stem cells. Mutat. Res. 614, 48–55.

Hussein, S.M., Batada, N.N., Vuoristo, S., Ching, R.W., Autio, R., Närvä, E., Ng, S., Sourour, M., Hämäläinen, R., Olsson, C., Lundin, K., Mikkola, M., Trokovic, R., Peitz, M., Brüstle, O., Bazett-Jones, D.P., Alitalo, K., Lahesmaa, R., Nagy, A., Otonkoski, T., 2011. Copy number variation and selection during reprogramming to pluripotency. Nature 471, 58–62.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., Lee, C., 2004. Detection of large-scale variation in the human genome. Nat. Genet. 36, 949–951.

Julian, L.J., Vella, L., Frankel, D., Minden, S.L., Oksenberg, J.R., Mohr, D.C., 2009. ApoE alleles, depression and positive affect in multiple sclerosis. Mult. Scler. 15, 311–315.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.

Laurent, L.C., Ulitsky, I., Slavin, I., Tran, H., Schork, A., Morey, R., Lynch, C., Harness, J.V., Lee, S., Barrero, M.J., Ku, S., Martynova, M., Semechkin, R., Galat, V., Gottesfeld, J., Izpisua Belmonte, J.C., Murry, C., Keirstead, H.S., Park, H.S., Schmidt, U., Laslett, A.L., Muller, F.J., Nievergelt, C.M., Shamir, R., Loring, J.F., 2011. Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. Cell Stem Cell 8, 106–118.

Lee, H.Y., Park, M.J., Kim, N.Y., Yang, W.I., Shin, K.J., 2011. Rapid direct PCR for ABO blood typing. J. Forensic Sci. 56 (Suppl 1), S179–S182.

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

Lister, R., Pelizzola, M., Kida, Y.S., Hawkins, R.D., Nery, J.R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., Downes, M., Yu, R., Stewart, R., Ren, B., Thomson, J.A., Evans, R.M., Ecker, J.R., 2011. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471, 68–73.

Ludwig, T.E., Bergendahl, V., Levenstein, M.E., Yu, J., Probasco, M.D., Thomson, J.A., 2006. Feeder-independent culture of human embryonic stem cells. Nat. Methods 3, 637–646.

Maiers, M., Gragert, L., Klitz, W., 2007. High-resolution HLA alleles and haplotypes in the United States population. Hum. Immunol. 68, 779–788.

McClendon, R., et al., 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061–1068.

Meisner, L.F., Johnson, J.A., 2008. Protocols for cytogenetic studies of human embryonic stem cells. Methods 45, 133–141.

Närvä, E., Autio, R., Rahkonen, N., Kong, L., Harrison, N., Kitsberg, D., Borghese, L., Itskovitz-Eldor, J., Rasool, O., Dvorak, P., Hovatta, O., Otonkoski, T., Tuuri, T., Cui, W., Brüstle, O., Baker, D., Maltby, E., Moore, H.D., Benvenisty, N., Andrews, P.W., Yli-Harja, O.,

Lahesmaa, R., 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.

Ng, P.C., Henikoff, S., 2001. Predicting deleterious amino acid substitutions. Genome Res. 11, 863–874.

Noreen, H.J., Yu, N., Setterholm, M., Ohashi, M., Baisch, J., Endres, R., Fernandez-Vina, M., Heine, U., Hsu, S., Kamoun, M., Mitsuishi, Y., Monos, D., Perlee, L., Rodriguez-Marino, S., Smith, A., Yang, S.Y., Shipp, K., Hegland, J., Hurley, C.K., 2001. Validation of DNA-based HLA-A and HLA-B testing of volunteers for a bone marrow registry through parallel testing with serology. Tissue Antigens 57, 221–229.

Rosler, E.S., Fisk, G.J., Ares, X., Irving, J., Miura, T., Rao, M.S., Carpenter, M.K., 2004. Long-term culture of human embryonic stem cells in feeder-free conditions. Dev. Dyn. 229, 259–274.

Schweppe, R.E., Klopper, J.P., Korch, C., Pugazhenthi, U., Benezra, M., Knauf, J.A., Fagin, J.A., Marlow, L.A., Copland, J.A., Smallridge, R.C., Haugen, B.R., 2008. Deoxyribonucleic acid profiling analysis of 40 human thyroid cancer cell lines reveals cross-contamination resulting in cell line redundancy and misidentification. J. Clin. Endocrinol. Metab. 93, 4331–4341.

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

Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S., Cooper, D.N., 2009. The Human Gene Mutation Database: 2008 update. Genome Med. 1, 13.

Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164.

Zhang, J., Feuk, L., Duggan, G.E., Khaja, R., Scherer, S.W., 2006. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res. 115, 205–214.