Cell Stem Cell
Resource

Low Incidence of DNA Sequence Variation in Human Induced Pluripotent Stem Cells Generated by Nonintegrating Plasmid Expression

Linzhao Cheng,[1,*] Nancy F. Hansen,[2] Ling Zhao,[3] Yutao Du,[6] Chunlin Zou,[1] Frank X. Donovan,[4] Bin-Kuan Chou,[1] Guangyu Zhou,[6] Shijie Li,[6] Sarah N. Dowey,[1] Zhaohui Ye,[1] NISC Comparative Sequencing Program,[5] Settara C. Chandrasekharappa,[4] Huanming Yang,[6] James C. Mullikin,[2,5] and P. Paul Liu[3,*]
[1]Stem Cell Program in Institute for Cell Engineering and Division of Hematology, Johns Hopkins University, Baltimore, MD 21205, USA
[2]Comparative Genomics Unit
[3]Oncogenesis and Development Section
[4]Genomics Core
[5]NIH Intramural Sequencing Center (NISC)
NHGRI, NIH, Bethesda, MD 20892, USA
[6]Beijing Genomics Institute (BGI) in Shenzhen, Shenzhen 518000, China
*Correspondence: lcheng@welch.jhu.edu (L.C.), pliu@mail.nih.gov (P.P.L.)
DOI 10.1016/j.stem.2012.01.005

SUMMARY

The utility of induced pluripotent stem cells (iPSCs) as models to study diseases and as sources for cell therapy depends on the integrity of their genomes. Despite recent publications of DNA sequence variations in the iPSCs, the true scope of such changes for the entire genome is not clear. Here we report the whole-genome sequencing of three human iPSC lines derived from two cell types of an adult donor by episomal vectors. The vector sequence was undetectable in the deeply sequenced iPSC lines. We identified 1,058–1,808 heterozygous single-nucleotide variants (SNVs), but no copy-number variants, in each iPSC line. Six to twelve of these SNVs were within coding regions in each iPSC line, but ~50% of them are synonymous changes and the remaining are not selectively enriched for known genes associated with cancers. Our data thus suggest that episome-mediated reprogramming is not inherently mutagenic during integration-free iPSC induction.

INTRODUCTION

The iPSC technology holds great promise for human stem cell biology and regenerative medicine, but the reprogramming processes and the resulting iPSCs remain incompletely characterized. Specifically, it is not clear how many changes occur at the DNA level during reprogramming. With recently available technologies such as single-nucleotide polymorphism (SNP) array and exome sequencing, several recent studies reported first glimpses of genetic abnormalities in human iPSCs derived from fibroblasts (Gore et al., 2011; Hussein et al., 2011; Laurent et al., 2011; Martins-Taylor et al., 2011; Mayshar et al., 2010). Nucleotide substitutions, copy-number variation (CNV) changes, and other chromosomal aberrations, which are either pre-existing or generated during reprogramming, may be selected for iPSC induction and/or expansion (Pera, 2011). In addition, the origin of starting cell types may influence the quality of derived iPSCs (Kim et al., 2010; Polo et al., 2010). A whole-genome sequencing (WGS) analysis is necessary to assess potential alterations in the entire nuclear and mitochondrial genomes, because recent WGS studies confirmed the notion that DNA mutational rates in somatic cells are lower in exons than in untranscribed regions probably because of transcription-coupled DNA repair (Lee et al., 2010; Pleasance et al., 2010). In this study, we conducted WGS analysis to determine the DNA sequences of three human iPSC lines.

RESULTS

Generation of Multiple iPSC Lines from Different Tissues of the Same Adult Donor

The iPSC lines were derived with plasmid vectors based on the EBNA1/OriP episomal replicon, from two cell types of a healthy donor (a 31-year-old anonymous male of mixed ethnic backgrounds). The relationship of the four iPSC lines and their somatic cell ancestors are shown in Figure 1. Bone marrow $CD34^+$ cells were used to generate the BC1 iPSC line after 4 day culture with a single episomal vector (pEB-C5) expressing five reprogramming genes (Chou et al., 2011). The second iPSC line BCT1 was derived from the same cultured $CD34^+$ cells by using an additional episomal vector expressing SV40 large T antigen (SV40-LT) gene together with pEB-C5. The $CD34^-$ marrow mononuclear cells were used to establish adherent marrow stromal cells (MSCs) (Cheng et al., 2003). The established MSCs after primary and first passage (p1) in culture (15 days total) were then used to generate two iPSC lines, E1 and E2 (see Experimental Procedures). By using standard methods for characterization of expanded iPSC lines such as BC1 (Chou et al., 2011), we confirmed normal pluripotency and karyotyping (by G-banding) of iPSC lines E1 (Figure S1 available online), E2 (data not shown), and BCT1 (Figure S3). For the

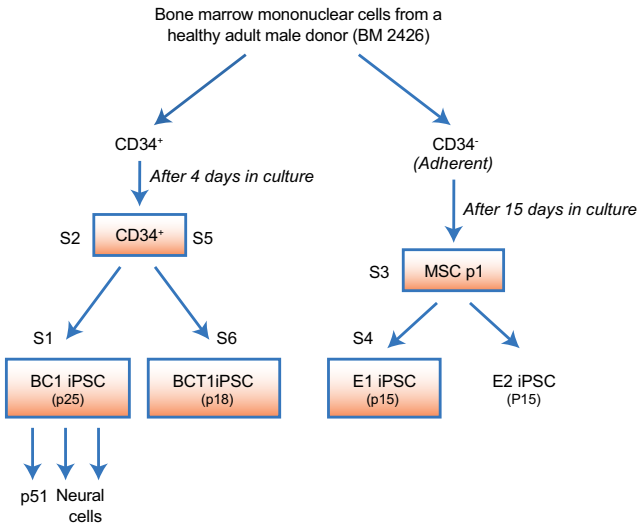

Figure 1. Relationship of iPSC Lines and Their Parental Somatic Cells Used in This Study

Mononuclear cells from bone marrow (BM) of healthy adult donor (BM 2426) were separated into $CD34^+$ cells (~1%) and CD34-depleted ($CD34^-$) cells. The $CD34^+$ cells were cultured for 4 days with hematopoietic cytokines before being reprogrammed by episomal vectors (left). The BC1 iPSC line was derived by using a single plasmid pEB-C5 whereas the BCT1 iPSC line was derived by addition of a second plasmid pEB-Tg. The BM $CD34^-$ cells were used to establish cultures of marrow stromal cells (also called mesenchymal stem cells or MSCs) by first selecting adherent cells followed by selective expansion of MSCs for an additional 13 days (after primary and first passage). The harvested MSCs after first passage (p1) were used for reprogramming similarly by episomal vectors. Two independent iPSC lines, E1 and E2, were established and expanded for 15 passages (p15). Functional characterizations of E1 and BCT1 iPSC lines are shown in Figures S1 and S2, and characterization of BC1 iPSC line was published previously (Chou et al., 2011). Three expanded and characterized iPSC lines, BC1, BCT1, and E1 (boxed), were also analyzed by whole-genome sequencing at deep coverages. This was done in pair with the parental somatic cells (also boxed). Although S1/S2 samples were sequenced at BGI as one pair, S3/S4 and S5/S6 samples were sequenced at NIH as two pairs.

whole-genome DNA sequencing and SNP array analyses, expanded iPSC lines were cultured under a feeder-free condition (on Matrigel) for at least one passage to reduce mouse feeder cells before total DNA was extracted. In addition, DNA was extracted by the same method from the corresponding parental cells, i.e., the cultured $CD34^+$ cells and the p1 MSCs. DNA from these somatic samples and from iPSCs at various passages were also used for other analyses.

Deep and Pair-wise Whole-Genome Sequencing of Three iPSC Lines and Their Parental Somatic Cells

To reduce false positive errors of differences in shotgun DNA sequencing resulting from variations such as library preparations and sequencing procedures (Ajay et al., 2011; Kinde et al., 2011), we sequenced the iPSC lines (BC1, BCT1, and E1) with their parental somatic cells at the same time in a pair-wise fashion by Illumina HiSeq2000 technology. For all three pairs, >1.5 × 10^9 reads per sample were generated and analyzed. Table 1 shows a summary of sequence analyses of the three iPSC lines. We obtained alignable DNA sequences of >130 Gb for each iPSC line and their parental cells. The amount of sequence data is equivalent to 48× or higher coverage of the haploid (nuclear) genome. Sequence analysis indicated that we achieved high-quality coverage (i.e., sufficient coverage in all six samples to call genotypes with >99.8% accuracy) for >94% of the autosomal genome of each sample. This level of deep and pair-wise sequencing of six related genomes (from the same person) provides us a high level of confidence to detect true and small sequence differences.

Absence of Episomal Vector Sequence in the Genomes of the iPSC Lines

The deep sequencing of the total DNA (5 μg or from 800,000 iPSCs) provides more definitive evidence for the lack of vector DNA (either integrated or as episome) in these iPSC lines. We did not detect the pCEP4 vector backbone sequence above background in any of the three sequenced iPSC lines derived by episomal vectors. This conclusion was further supported by the PCR method that would detect 0.2 copies of vector DNA per cell (or a total of ~300 copies per the cell population tested) as shown in Figures S1 and S2 and in previous studies (Chou et al., 2011).

Single-Nucleotide Variations in the Genomes of the iPSC Lines

When compared to the human reference genome sequence (hg19), we identified approximately 4.2 million single-nucleotide variants (SNVs) in each of the iPSC lines as well as in their parental $CD34^+$ cells and MSCs (Table 1). Some of these SNVs were aligned to repetitive regions of the reference genome (in parentheses). This level of sequence variations is comparable to those of other sequenced human genomes (Bentley et al., 2008; Ding et al., 2010; Lee et al., 2010; Pleasance et al., 2010; Ajay et al., 2011). When the sequences between each iPSC line and its parental cells were directly compared, 1,058 to 1,808 likely SNVs were identified between each iPSC line and its parental somatic cells sequenced in pairs (Table 1). All SNVs found in iPSCs were heterozygous (i.e., single-allele) changes as compared to their parental somatic cells. Importantly, none of these iPSC-associated SNVs are shared among the three iPSC lines. Neither did we observe any clustering of these variations in specific chromosomal regions.

SNVs in Known Functional Elements of the Genome

To investigate functional relevance of SNVs found in iPSCs after induction and expansion, we next focused on SNVs in exons, especially those not present in dbSNP (build 132) database as reported previously. High-quality sequencing revealed six SNVs in BC1 iPSCs residing within the coding regions of six different genes (Tables 1 and 2). Three of them are nonsynonymous (Table 2). The paired sequencing data revealed six SNVs in coding regions of six different genes in BCT1 iPSCs (derived from the same $CD34^+$ cells); two of them are nonsynonymous (Tables 1 and 2). Twelve SNVs in E1 iPSCs were found in the coding regions of 12 different genes; 6 of them are nonsynonymous and one is a truncation (Tables 1 and 2). In searching for small insertions or deletions (indels) in the coding sequences that are unique to the three iPSC lines, we identified and

Table 1. Summary of Sequencing Three Pairs of iPSC Lines and Their Parental Somatic Cells

Features/iPSC Lines	BC1	BCT1	E1
Total nucleotides sequenced	172 Gb	153 Gb	217 Gb
Alignable nucleotides	165 Gb	142 Gb	134 Gb
Genome coverage (fold)	59×	51×	48×
Total SNVs (compared to the hg19 reference)	4,158,672 (2,193,600)[a]	4,207,199 (2,237,532)	4,231,439 (2,259,512)
Total SNVs (compared to parental cells)	3,234 (2,217)	4,850 (3,458)	4,470 (3,053)
High-quality, filtered SNVs	1,058 (626)	1,109 (662)	1,808 (1,029)
SNVs in intergenic regions	676 (420)	684 (438)	1,125 (678)
SNVs in conserved noncoding regions	376 (201)	420 (225)	674 (352)
SNVs in introns	367 (206)	409 (229)	664 (360)
SNVs in 5′ or 3′ UTRs	11 (4)	13 (1)	20 (3)
SNVs in coding regions	6	6	12
Synonymous (S)	3	4	5
Nonsynonymous (NS)	3	2	6
Nonsense	0	0	1
NS:S ratio	1	0.5	1.4
Coding region indels	0	0	2
SNVs in CpG islands	2	0	7
SNVs in sno/microRNA regions	0	1	0
SNV in mitochondrial genome	1 (nt 89)	0	0
Vector sequences anywhere	none	none	none

BC1 and BCT1 iPSC lines were derived from the same batch of $CD34^+$ cells cultured for 4 days. E1 iPSC line were derived from MSCs established from $CD34^-$ (adherent) cells after culture of 15 days. Parental cultured $CD34^+$ cells (twice) and MSCs were sequenced at similar depths in each pair together with an iPSC line. BC1 iPSC/$CD34^+$ pair (S1/S2) was sequenced at BGI, and the E1 iPSC/MSC (S3/S4) and BCT1 iPSC/$CD34^+$ (S6/S5) pairs were sequenced at NIH.

[a] SNVs: single-nucleotide variations. The SNV counts in parentheses are those that lie within regions annotated in the UCSC hg19 "RepeatMasker" track.

confirmed only a 5-nucleotide deletion in the *SETD8* gene in E1 iPSCs (Table 2).

In addition, there were single-nucleotide variations in the 5′ or 3′ untranslated regions (UTRs), introns, and noncoding regions deemed "conserved" by the program PhyloP when run on UCSC's 46-way multialignments (Table 1). All of these changes could potentially affect gene expression. Of the 1,058 SNVs between BC1 iPSCs and $CD34^+$ cells, only 2 lie in a CpG island near a promoter, and none lie within the sno/microRNA regions in the UCSC's sno/microRNA track. For BCT1 iPSCs, there were no SNVs in the CpG islands but one in the micro RNA mir-124-2. For other changes in the E1 iPSC line, seven of them are within CpG islands and none in the sno/microRNA regions (Table 1).

Validation of SNVs by Genomic PCR and Sanger DNA Sequencing

We set to confirm the presence of SNVs and indels located in the exons and selected ones in the introns or UTRs of the iPSC genomes by using genomic PCR (with specific primers shown in Table S2) followed by Sanger sequencing. We confirmed 48 out of total 51 SNVs tested (Table S1), for an overall confirmation rate of 94%. Two of the putative nucleotide substitutions (one in BC1 and the other in E1 iPSCs) and a single-nucleotide deletion in the *ZNF479* gene in E1 iPSCs could not be confirmed by this approach, and therefore are likely to have been false positive calls. Alternatively, the unchanged alleles may have been preferentially amplified during PCR reactions, leading to confirmation failure. Thus, the vast majority of sequence changes at nonrepetitive regions that are identified by deep WGS via the HiSeq2000 technology, along with appropriate filtering, are real.

Additionally, we analyzed discordance of two somatic cell types that were both sequenced together: $CD34^+$ cells as sample S5 and MSCs as sample S3 from the same donor (Figure 1). We found 283 possible SNVs between the two cultured somatic cell types, most of them in repetitive regions (data not shown). The subsequent PCR and Sanger sequencing failed to confirm any of the ten randomly selected putative SNVs in nonrepetitive regions; therefore, they are probably false positives. Our WGS analyses corroborate with recent findings of the noise levels even by the HiSeq2000 technology, albeit very small (Ajay et al., 2011; Kinde et al., 2011). These data also highlight the importance of validating potential SNVs revealed by WGS analysis. In addition, our data suggest that the cell cultures of $CD34^+$ cells and MSCs for 4–15 days did not introduce somatic mutations significantly.

Analysis of SNVs in Different Passages of the Same iPSC Line or between Different iPSC Lines Derived from the Same Somatic Cell Population

The sequence variants were stably maintained in the BC1 iPSC line, because we detected the same variants at an earlier passage (p11) and a later passage (p51) as in the sequenced

Table 2. Sequence Variants in the Coding Regions in BC1 and E1 iPSCs

iPSC Lines	Gene Name	Chromosome	Position	Ref Allele	Variant Allele	Variant Type
BC1	CNBD1	8	88365930	G	A	nonsynonymous
	ITGA11	15	68650878	T	C	nonsynonymous
	SLIT2	4	20490512	C	T	nonsynonymous
	DHX34	19	47865827	C	T	synonymous
	DOCK1	10	128836023	C	T	synonymous
	PRPS1L1	7	18066668	A	G	synonymous
BCT1	ADAMTS17	15	100657067	C	T	nonsynonymous
	TG	8	133961089	C	T	nonsynonymous
	CDH4	20	60318830	G	A	synonymous
	EPPK1	8	144941146	G	C	synonymous
	MCM10	10	13213236	C	A	synonymous
	OR8K1	11	56113928	C	T	synonymous
E1	SDHA	5	231082	C	T	stop
	DHX33	17	5359463	C	A	nonsynonymous
	FAM3B	21	42717750	T	A	nonsynonymous
	KCNH8	3	19389396	C	A	nonsynonymous
	HRH1	3	11301698	C	G	nonsynonymous
	IRF4	6	397216	C	T	nonsynonymous
	ZNF226	19	44677036	G	C	nonsynonymous
	CYP11B2	8	143999035	C	T	synonymous
	HIST3H2A	1	228645201	A	G	synonymous
	HJURP	2	234746286	C	T	synonymous
	PRKD2	19	47197220	G	A	synonymous
	RPL13AP6	10	112696395	G	A	synonymous
	SETD8	12	123892105	CCAAA	–	del CCAAA

BC1 iPSCs (p25). They are also present in neural progenitor cells differentiated from BC1 iPSCs (Table S1). Importantly, none of the confirmed 16 variants tested were detected in the parental $CD34^+$ cells or in the sibling BCT1 iPSC line (at least at the level of <0.1%). Similar results were obtained with MSC-derived iPSCs: none of the confirmed 25 variants found in E1 iPSCs were detected in a sibling iPSC line E2 or in their common parental MSCs used for reprogramming (Table S1). In addition, the six heterozygous variants found in BCT1 iPSCs were not presented in the sibling BC1 iPSC line. Therefore, none of the SNVs found by WGS and further confirmed by Sanger sequencing are shared among the three iPSC lines.

Analysis of the Mitochondrial Genome

We also obtained high-quality and deep coverage of the mitochondrial genomes. No iPSC-specific substitutions were found in the BCT1 and E1 iPSCs. In the BC1 iPSC line, however, there is a single substitution at nt89 (T>C) in the 5′ highly variable, noncoding region of the mitochondrial genome (16.5 kb). This nucleotide substitution was present in all the sequencing reads of the BC1 iPSCs (see Discussion below).

Absence of CNV Alterations in the iPSC Lines

The WGS with deep-depth and paired-end read mapping data also provides us a new way to assess CNVs changes after reprogramming as compared to parental somatic cells. We used three prediction programs for CNV detection: RDXplorer (Yoon et al., 2009), CNVseq (Xie and Tammi, 2009), and BreakDancer (Chen et al., 2009) on the three pairs of DNA samples (iPSCs versus respective parental somatic cells) from the same person. No tenable examples of CNV differences between an iPSC line and pair-wise sequenced parental cells were found by more than one of the three methods.

To validate the absence of new CNVs in the integration-free iPSC lines derived by episomal vectors, we also used HumanOmni2.5-Quad BeadChips that measure 2.5 million SNP markers to detect CNVs and other structural changes in iPSC lines BC1 and E1, as compared to their corresponding parental somatic cells from the same person. For BC1 iPSCs, we analyzed the iPSCs at three different passages: 11, 28, and 51, together with the parental $CD34^+$ cells. For E1 iPSCs, p17 was used together with its parental MSCs. The call rates of SNP alleles for all the DNA samples were >99%, and the data met the desired quality control requirements (data not shown).

Potential CNVs were detected with cnvPartition, PennCNV (Wang et al., 2007), and Nexus and summarized in Table S3. With the 2.5M SNP array, we were able to detect CNVs in the range of few kilobases in length. The two smallest CNVs (2.02 kb and 2.75 kb) are shown in Figure S3, which are shared by all six samples of iPSCs and parental somatic cells. The overall numbers of CNVs in the iPSCs and their parental cells are similar to other human somatic cells, because the numbers of

Table 3. Genuine CNVs Detected in the iPSCs and Parental Somatic Cells

Chr	Copy Number	Start	End	Length	Method of Detection
1	0	8105122	8114476	9354	PennCNV, CNVPartition, Nexus
1	1	190093507	190134936	41429	PennCNV, CNVPartition, Nexus
1	1	195089923	195146893	56970	PennCNV, CNVPartition, Nexus
2	3	95884787	95890722	5935	CNVPartition
3	3	77913197	77920280	7083	PennCNV, CNVPartition
3	1	113584277	113599333	15056	PennCNV, CNVPartition
4	1	11983684	11993214	9530	PennCNV, CNVPartition
5	0	33143879	33149247	5368	CNVPartition
6	1	79020533	79092458	71925	CNVPartition
6	1	93631870	93635080	3210	CNVPartition
6	1	113012462	113023906	11444	CNVPartition
7	1	49144752	49150490	5738	CNVPartition
7	1	139841470	139847922	6452	CNVPartition
7	1	145424894	145427315	2421	CNVPartition
7	1 (deletion)[a]	66762793	66803212	40419	CNVPartition, Nexus
7	3	151512420	151745955	233536	Nexus
7	0	20717299	20720053	2754	PennCNV, CNVPartition, Nexus
8	1	63378480	63385988	7508	CNVPartition
8	1	39351738	39499553	147815	PennCNV, CNVPartition
9	1	580454	598622	18168	PennCNV, CNVPartition
9	1	4491352	4494129	2777	PennCNV, CNVPartition
10	1	91988554	91991990	3436	CNVPartition
10	1	26725084	26730905	5821	PennCNV, CNVPartition
10	3	46561933	47173875	611942	PennCNV, CNVPartition, Nexus
11	0	102254658	102261874	7216	CNVPartition
12	1	12424251	12433136	8885	CNVPartition
12	1	121202830	121215756	12926	CNVPartition
12	0	128624494	128628185	3691	CNVPartition
12	1	103185002	103193085	8084	Nexus
12	3	34211024	34744278	533254	PennCNV, CNVPartition, Nexus
13	3	62488915	62512304	23389	CNVPartition, Nexus
13	0	31430761	31435768	5007	PennCNV, CNVPartition, Nexus
13	1	36968549	37018275	49726	PennCNV, CNVPartition, Nexus
15	1	22615864	22635687	19823	CNVPartition
15	1	79809091	79958158	149067	PennCNV, CNVPartition, Nexus
15	1	22010825	22104326	93,502	PennCNV, Nexus
16	0	163625	165653	2028	CNVPartition, Nexus
16	1	20421133	20440897	19764	PennCNV, CNVPartition
16	1	33849764	33929519	79755	PennCNV, CNVPartition
16	3	44973739	45446855	473116	PennCNV, CNVPartition, Nexus
17	1	76309874	76312824	2950	CNVPartition
17	1	46346285	46353510	7225	PennCNV, CNVPartition
19	1	23974942	23977215	2273	CNVPartition
19	3	21623408	21639452	16,045	PennCNV
20	0	55019459	55021582	2124	Nexus

[a] Asterisk denotes a deletion found only in BC1 p51 (also see Figure S4). All others were shared in all six samples.

CNVs predicted in these cells are comparable with the median value for a larger cohort of 84 unrelated human samples analyzed by PennCNV with the same parameters (Table S3).

The potential CNVs were further evaluated by visual examination with both GenomeStudio and Nexus, and only 45 of them were deemed to be real (Table 3). Importantly, all of these 45 CNVs are shared among all 6 tested samples (both iPSCs and parental cells), except for one that is unique to BC1p51 iPSCs. After detailed analyses, this putative CNV was found to be an incomplete deletion in a region (40.42 kb) containing many segmental duplications within chromosome 7q11.21. Visual demonstrations of the incomplete deletion are shown in Figure S4. The putative CNV or alteration was found only in the 51st passage of the BC1 iPSCs (not sequenced) but not in other five samples including BC1 iPSCs of earlier passages (p11 and p28; p25 was sequenced). The exact nature of the incomplete deletion in this region remains to be determined. It is also well known that extensively cultured human ESCs and iPSCs often contain CNV alterations, although not at this locus (Laurent et al., 2011; Martins-Taylor et al., 2011). Therefore, our SNP analysis effectively validated the WGS data: as compared to parental cells, essentially no new CNVs were introduced to the episome-mediated, integration-free iPSC lines.

DISCUSSION

We report here results of WGS analyses of three iPSC lines and their parental somatic cells that were reprogrammed by episomal vectors. We confirmed that the episomal vector DNA did not integrate in the genome or persist in the three characterized iPSC lines, nor did it alter the structures of the iPSC genomes at detectable levels. The three fully sequenced iPSC lines derived from two different cell types of the same person would provide valuable references or standards with DNA sequence information, for future studies of genomic integrity and epigenomics (such as DNA methylation) of iPSCs before and after differentiation.

The present study corroborates with a recent report that ~6 SNVs or small sequence changes were found in exonic regions of examined iPSC lines derived from fibroblasts by various reprogramming methods (Gore et al., 2011). However, our deep DNA sequencing of the whole genome also allowed us to detect SNVs and other sequence changes in nonexonic regions (>98%) of the nuclear genome and in the mitochondrial genome. We found 1,058–1,808 sequence changes, mostly SNVs, per genome in the three iPSCs after induction and expansion. Currently it is unclear exactly where and when these iPSC-associated SNVs arose. At least two possibilities can be envisioned, which are not mutually exclusive. First, these SNVs are simply normal mitotic mutations during iPSC induction (presumably from a single cell) and/or subsequent expansion before WGS. Second, a founder somatic cell from adult tissues that has been reprogrammed to a clonal iPSC line may already contain most if not all of these SNVs. Both are related to the fact that somatic DNA mutations occur along mitotic cell divisions both in vivo and in vitro.

Estimation of mutational rates in somatic cells varies significantly by previous methods, which typically depend on the expression of a functional gene (Araten et al., 2005; Lynch, 2010). Recent WGS analysis provides a more definitive and selection-independent method to measure mutational frequencies in both exonic and nonexonic regions in normal primary cells as well as in cultured human cell lines (Kinde et al., 2011). Based on these studies, we used the estimation of 3 to 30 mutations per haploid genome per mitotic division for somatic cells in the following discussion. Formation of a sizable iPSC colony (~1,000 cells), presumably from a single somatic cell during our episome-mediated reprogramming, takes ~2 weeks and requires ~10 cell divisions. Subsequently we expanded iPSCs for 15–25 passages, with an estimated >3 cell divisions per passage, before sequencing. Therefore, it is conceivable that some of the SNVs we observed in iPSCs may have arisen during iPSC induction and subsequent expansion (the first possibility). However, the SNVs found in iPSCs could have also been inherited from a given somatic cell that was reprogrammed successfully (the second possibility). Considering that a typical human somatic cell is derived from a fertilized egg after 46–47 cell divisions during embryonic, fetal, and postnatal development, each somatic cell is expected to harbor 138–1,410 spontaneous mutations that differ from those in another cell in the population. Because we set the algorithms to consider a sequence change to be real if it is present in at least 10% of the reads (both nuclear and mitochondrial DNA sequences), the detected SNVs should have been either pre-existing in a founder somatic cell that was reprogrammed or acquired within ~3–4 cell divisions after iPSC induction (when a single colony was formed and picked). Otherwise, a later acquired mutation cannot be present in $\geq$10% of the derived iPSCs unless it offers a growth advantage in subsequent expansions and is preferentially retained. In our present study, we found 1,058–1,808 sequence changes, mostly SNVs, per genome in the 3 iPSC lines, which are within the expected ranges for a normal human somatic cell in adults. The present study therefore corroborates with a recent report that at least 50% of the SNVs or small sequence changes found by exome sequencing of the examined iPSC lines can be found in parental somatic cell populations (Gore et al., 2011). The somatic origin of SNVs was also supported by our data of the mitochondrial DNA sequencing. In the BC1 iPSC line where a single SNV was found in the noncoding region, the variant sequence was found in all the reads although a cell contains hundreds or thousands of mitochondrial DNA genomes. Overall, our WGS data indicate that reprogramming of iPSCs by episomal vectors is not inherently mutagenic. We predict that the level of SNVs found in the iPSC lines is probably of the same magnitude as those found in other adult somatic (stem) cells after extended proliferation.

It is possible that somatic cells harboring sequence variations that favor iPSC induction and expansion could have been selected for iPSC reprogramming. Those iPSCs with additional sequence variants generated during early passages that favor iPSC growth may be enriched during clonal expansion. Although we cannot rule out these possibilities completely, they do not seem to be likely for the three iPSC lines studied here, based on the following two reasons. First, bioinformatic analysis of the genes harboring nonsynonymous, premature termination and deletion variants did not reveal a recurrent pattern of mutating a single common gene or signal pathway, similar to the recent exome sequencing study (Gore et al., 2011). None of the SNVs we found (in exons and other regions) are shared among the three iPSC lines sequenced here or with those found in the previous exome sequencing study. Second, the ratios of nonsynonymous:synonymous substitutions (NS:S), which was traditionally applied to germline mutations that have evolved over a long period of evolutionary time, are 1, 0.5, and 1.4 for each iPSC line, respectively (Table 2), or 1.1 (13:12) as a group. These ratios are lower than 2.6 as reported recently for fibroblast-derived iPSC lines by exome sequencing (Gore et al., 2011), whereas the three cited cancer WGS studies reported NS:S ratios of 0.97, 1.78, and 2.64, respectively (Lee et al., 2010; Pleasance et al., 2010; Ding et al., 2010). Although the significance or relevance of using NS:S ratios in analyzing somatic mutations in cancers or iPSCs remains to be determined, the observed NS:S ratio (1:1) of SNVs in our integration-free three iPSC lines derived by episomal vectors did not suggest that they bear such a characteristic if it is proven to be a cancer cell signature. Notably, the SNVs found in all the iPSCs sequenced by us and Gore et al. (2011) did not cluster in few genes or a group of genes encoding proteins related to a common functional pathway, such as recurrent mutations in *TP53*, *RAS*, *RAF*, *PTEN*, and *PIK3CA* genes found in multiple cancer samples sequenced (Lee et al., 2010; Pleasance et al., 2010; Ding et al., 2010). The analysis of SNV patterns such as

NS:S ratios in known functional regions further supports the notion that iPSC derivation by an improved method is not inherently tumorigenic or mutagenic. Future experimental studies, however, are still needed to determine the functional importance of the observed DNA sequence variations in iPSCs, for their growth and differentiation to various cell lineages, and whether these variations have any adverse consequences.

The robust, unbiased, and increasingly affordable WGS technology for analyzing genome integrity of iPSCs and other cell types is not without limitations. Because read lengths and library fragment lengths span only a few hundred bases and read depths are randomly distributed, accurately detecting CNVs and other structural rearrangements (such as inversions, translocations, transposon insertions, and others associated with repetitive sequences) by the current WGS technology remains difficult. Because WGS analysis alone cannot fully rule out the existence of CNV changes between iPSCs and their parental somatic cells, we also detected CNVs with the 2.5M SNP array. The data from both SNP array and WGS technologies were analyzed with multiple algorithms. We did not observe CNV changes (up to 51 passages) in these three iPSC lines derived by episomal vectors, although others found CNV alterations in a fraction of iPSCs derived from fibroblasts by integrating viral vectors (Hussein et al., 2011; Laurent et al., 2011; Martins-Taylor et al., 2011). It is unclear what factors contribute to the discrepancy. The absence of detectable CNVs in our iPSC lines could be due to the fact that different types of parental somatic cells and/or reprogramming vectors were used. We used human $CD34^+$ cells and MSCs after short cultures and episomal vectors for generating integration-free iPSCs, whereas the previously reported iPSCs were derived from extensively cultured fibroblasts with integrating viral vectors.

In summary, we have conducted a deep whole-genome sequencing of three independently derived and functionally characterized iPSC lines from a single donor to comprehensively document the number and type of sequence variations in these iPSC lines. Unbiased whole-genome DNA sequencing is a definitive way to identify vector DNA integration and mutations in iPSC lines. The data presented in this report suggest that the genome of an iPSC line derived by episomal vectors could be largely intact: less than one SNV per megabase of DNA when compared to the parental somatic cells, despite clonal selection and many cycles of mitotic cell divisions during iPSC reprogramming and subsequent expansion. The existence of these DNA sequence variations in thousands as compared to the starting somatic cell population, however, suggests that each established iPSC line needs to be characterized thoroughly at the genomic DNA level before it is used for comprehensive functional studies and clinical applications.

EXPERIMENTAL PROCEDURES

Anonymous Adult Human Somatic Cells Used for Reprogramming

Human primary mononuclear cells (MNCs) obtained from bone marrow and blood of anonymous donors were collected and processed at AllCells, LLC (Emeryville, CA). The consent form signed by the healthy adult male donor coded as BM2426 is available upon request. The practice of obtaining bone marrow aspirates and blood from adult donors was approved by Institutional Review Board (IRB) at AllCells. Use of anonymous human samples for laboratory research including iPSC derivation was approved by IRB and the Institutional Stem Cell Research Oversight (ISCRO) committee at Johns Hopkins University.

Human MNCs from the marrow donor BM2426 were isolated with a standard gradient protocol by Ficoll-Paque Plus ($p = 1.077$). The human MNCs expressing a high level of the CD34 surface marker ($CD34^+$) were purified with the MACS magnet system and CD34 isolation beads (Miltenyi, Auburn, CA). The CD34-depleted ($CD34^-$) MNCs were used to establish marrow stromal cells (also called mesenchymal stem cells or MSCs) by a standard protocol (Cheng et al., 2003; Mali et al., 2010). In brief, total unfractionated $CD34^-$ MNCs were first cultured for 2 days in DMEM (low glucose) plus 10% FBS in standard adherent tissue culture flasks. After discarding hematopoietic cells that remained in suspension at day 2, MSCs as adherent cells were then selectively expanded until subconfluence before harvest by trypsin digestion. The resulting adherent cells (called passage zero or p0) were replated under the same condition, expanded, and harvested at subconfluence as p1 (15 days in culture). The cultured p1 MSCs were used for cell reprogramming as well as DNA analysis as described below.

Human iPSC Lines Derived from Adult Marrow $CD34^+$ Cells and MSCs from BM2426

The BC1 iPSC line was derived from the BM2426 $CD34^+$ cells (after in a hematopoietic culture for 4 days) by a single episomal vector pEB-C5 as previously described (Chou et al., 2011). The BCT1 iPSC line was derived from the same cultured marrow $CD34^+$ cells as BC1, except that the second episomal vector pEB-Tg expressing SV40-LT transiently was also used in addition to pEB-C5. For reprogramming MSCs, 0.5×10^6 cells were nucleofected by up to 5 μg DNA plasmid with Lonza/Amaxa's recommended MSC solution and electroporation parameter as we previously used in the DNA transposon vector study (Mali et al., 2010). In this study, we used episomal vectors (such as combination #6) for reprogramming MSCs as described previously (Chou et al., 2011; Yu et al., 2009). Four days after transfection by two or three plasmids, human embryonic stem cell medium was added in the presence of sodium butyrate as we previously described (Mali et al., 2010; Chou et al., 2011). The efficiency of iPSC derivation from adult MSCs was ~1 per 10^6 transfected MSCs even when three episomal vectors were used. Among various clones we picked and expanded, two (E1 and E2) were fully characterized by the functional assays such as pluripotency and karyotyping. G-banded karyotyping was conducted by a certified cytogeneticist (Cheng et al., 2003; Mali et al., 2010). In brief, at least 20 metaphases for each sample were counted and partially analyzed. At least 5 of the 20 spreads were fully analyzed in detail. Resolution of 300–450 bands was obtained. This analysis rules out mosaicism of greater than 14% with 95% confidence.

Whole-Genome Sequencing and Analysis

Whole-genome DNA libraries suitable for sequencing on Illumina's sequencing platform were generated from 5 μg of genomic DNA with the TruSeq Sample Prep Kit from Illumina. The DNA was sheared to approximately 450 bp with a Covaris E210. Size selection was achieved on a Pippin Prep with 1.5% agarose cassettes (Sage Science). The libraries were sequenced on HiSeq2000 DNA Sequencers (Illumina). Although samples of S1 (BC1 iPSC) and S2 ($CD34^+$ cells) were sequenced at BGI with 90 bp paired-end reads, samples S3 (MSC), S4 (MSC-derived E1 iPSC), S5 ($CD34^+$ cells again), and S6 (BCT1 iPSC) were sequenced at NISC with 101 bp paired-end reads (Figure 1). Details of WGS analyses are provided in Supplemental Information.

Genomic DNA PCR and Sanger Sequencing

Sequence variants found by whole-genome sequencing were confirmed by PCR-Sanger sequencing with an ABI 3100 Genetic Analyzer (Applied Biosystems). In brief, 20 ng genomic DNA samples from iPSC lines, their parental cells and an unrelated control blood sample were used for PCR reactions with primers located about 100–200 bp at either side of the selected sequence variants (Table S2). PCR reactions were cleaned with Exo-SAP-IT (Affymetrix) and followed by sequencing reactions with BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). The sequence data were analyzed with Sequencher 4.10.1.

SNP Array Analysis

Genotyping was performed with the HumanOmni2.5_Quad v1.0 DNA analysis BeadChip kit (Illumina, Inc.) and 300 ng of genomic DNA per the Illumina "infinium assay" protocol (Gunderson et al., 2005). CNVs were detected with the Illumina GenomeStudio "in-house" algorithm, cnvPartition v3.1.6, PennCNV, and Nexus 5.1. Details of CNV analyses are provided as Supplemental Information.

ACCESSION NUMBERS

The raw SNP data from the HumanOmni2.5_Quad v1.0 array are deposited at http://research.nhgri.nih.gov/. The three lists (A, B, and C) of all CNV calls we analyzed by each of the three methods (PennCNV, cnvPartition, and Nexus) are also available at the website.

The raw sequencing data reported in this study have been submitted to the Sequence Read Archive (SRA) at NCBI (accession number SRA048525), which can be accessed at http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and three tables and can be found with this article online at doi:10.1016/j.stem.2012.01.005.

ACKNOWLEDGMENTS

We thank Raman Sood and Blake Carrington for PCR and Sanger sequencing for confirming mutations and Xianmin Zeng in The Buck Institute in California for providing DNA from neural progenitor cells derived from BC1 iPSCs. This study was supported in part by Johns Hopkins University and NIH grants (RC2 HL101582, U01 HL099775, and R01 HL073781), an award from the NIH Center for Regenerative Medicine (NCRM), and The Intramural Research Program of The National Human Genome Research Institute at NIH.

Received: June 6, 2011
Revised: December 5, 2011
Accepted: January 10, 2012
Published: March 1, 2012

REFERENCES

Ajay, S.S., Parker, S.C., Abaan, H.O., Fajardo, K.V., and Margulies, E.H. (2011). Accurate and comprehensive sequencing of personal genomes. Genome Res. *21*, 1498–1505.

Araten, D.J., Golde, D.W., Zhang, R.H., Thaler, H.T., Gargiulo, L., Notaro, R., and Luzzatto, L. (2005). A quantitative measurement of the human somatic mutation rate. Cancer Res. *65*, 8111–8117.

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

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677–681.

Cheng, L., Hammond, H., Ye, Z., Zhan, X., and Dravid, G. (2003). Human adult marrow cells support prolonged expansion of human embryonic stem cells in culture. Stem Cells *21*, 131–142.

Chou, B.K., Mali, P., Huang, X., Ye, Z., Dowey, S.N., Resar, L.M.S., Zou, C., Zhang, Y.A., Tong, J., and Cheng, L. (2011). Efficient human iPS cell derivation by a non-integrating plasmid from blood cells with unique epigenetic and gene expression signatures. Cell Res. *21*, 518–529.

Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., McLellan, M.D., Fulton, R.S., Fulton, L.L., et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature *464*, 999–1005.

Gore, A., Li, Z., Fung, H.L., Young, J.E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M.A., Kiskinis, E., et al. (2011). Somatic coding mutations in human induced pluripotent stem cells. Nature *471*, 63–67.

Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G., and Chee, M.S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. Nat. Genet. *37*, 549–554.

Hussein, S.M., Batada, N.N., Vuoristo, S., Ching, R.W., Autio, R., Närvä, E., Ng, S., Sourour, M., Hämäläinen, R., Olsson, C., et al. (2011). Copy number variation and selection during reprogramming to pluripotency. Nature *471*, 58–62.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I., et al. (2010). Epigenetic memory in induced pluripotent stem cells. Nature *467*, 285–290.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W., and Vogelstein, B. (2011). Detection and quantification of rare mutations with massively parallel sequencing. Proc. Natl. Acad. Sci. USA *108*, 9530–9535.

Laurent, L.C., Ulitsky, I., Slavin, I., Tran, H., Schork, A., Morey, R., Lynch, C., Harness, J.V., Lee, S., Barrero, M.J., et al. (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, W., Jiang, Z., Liu, J., Haverty, P.M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K.P., Bhatt, D., et al. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature *465*, 473–477.

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. USA *107*, 961–968.

Mali, P., Chou, B.K., Yen, J., Ye, Z., Zou, J., Dowey, S., Brodsky, R.A., Ohm, J.E., Yu, W., Baylin, S.B., et al. (2010). Butyrate greatly enhances derivation of human induced pluripotent stem cells by promoting epigenetic remodeling and the expression of pluripotency-associated genes. Stem Cells *28*, 713–720.

Martins-Taylor, K., Nisler, B.S., Taapken, S.M., Compton, T., Crandall, L., Montgomery, K.D., Lalande, M., and Xu, R.H. (2011). Recurrent copy number variations in human induced pluripotent stem cells. Nat. Biotechnol. *29*, 488–491.

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

Pera, M.F. (2011). Stem cells: The dark side of induced pluripotency. Nature *471*, 46–47.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordóñez, G.R., Bignell, G.R., et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191–196.

Polo, J.M., Liu, S., Figueroa, M.E., Kulalert, W., Eminli, S., Tan, K.Y., Apostolou, E., Stadtfeld, M., Li, Y., Shioda, T., et al. (2010). Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. Nat. Biotechnol. *28*, 848–855.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.

Xie, C., and Tammi, M.T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics *10*, 80.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. *19*, 1586–1592.

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.