ARTICLE

doi:10.1038/nature09805

Somatic coding mutations in human induced pluripotent stem cells

Athurva Gore[1]*, Zhe Li[1]*, Ho-Lim Fung[1], Jessica E. Young[2], Suneet Agarwal[3], Jessica Antosiewicz-Bourget[4], Isabel Canto[2], Alessandra Giorgetti[5], Mason A. Israel[2], Evangelos Kiskinis[6], Je-Hyuk Lee[7], Yuin-Han Loh[3], Philip D. Manos[3], Nuria Montserrat[5], Athanasia D. Panopoulos[8], Sergio Ruiz[8], Melissa L. Wilbert[2], Junying Yu[4], Ewen F. Kirkness[9], Juan Carlos Izpisua Belmonte[5,8], Derrick J. Rossi[10], James A. Thomson[4], Kevin Eggan[6], George Q. Daley[3], Lawrence S. B. Goldstein[2] & Kun Zhang[1]

Defined transcription factors can induce epigenetic reprogramming of adult mammalian cells into induced pluripotent stem cells. Although DNA factors are integrated during some reprogramming methods, it is unknown whether the genome remains unchanged at the single nucleotide level. Here we show that 22 human induced pluripotent stem (hiPS) cell lines reprogrammed using five different methods each contained an average of five protein-coding point mutations in the regions sampled (an estimated six protein-coding point mutations per exome). The majority of these mutations were non-synonymous, nonsense or splice variants, and were enriched in genes mutated or having causative effects in cancers. At least half of these reprogramming-associated mutations pre-existed in fibroblast progenitors at low frequencies, whereas the rest occurred during or after reprogramming. Thus, hiPS cells acquire genetic modifications in addition to epigenetic modifications. Extensive genetic screening should become a standard procedure to ensure hiPS cell safety before clinical use.

Human induced pluripotent stem cells have the potential to revolutionize personalized medicine by allowing immunocompatible stem cell therapies to be developed[1,2]. However, questions remain about hiPS cell safety. For clinical use, hiPS cell lines must be reprogrammed from cultured adult cells, and could carry a mutational load due to normal *in vivo* somatic mutation. Furthermore, many hiPS cell reprogramming methods use oncogenes that may increase the mutation rate. Additionally, some hiPS cell lines have been observed to contain large-scale genomic rearrangements and abnormal karyotypes after reprogramming[3]. Recent studies also revealed that tumour suppressor genes, including those involved in DNA damage response, have an inhibitory effect on nuclear reprogramming[4–9]. These findings suggest that the process of reprogramming could lead to an elevated mutational load in hiPS cells.

To probe this issue, we sequenced the majority of the protein-coding exons (exomes) of 22 hiPS cell lines and the nine matched fibroblast lines from which they came (Table 1). These lines were reprogrammed in seven laboratories using three integrating methods (four-factor retroviral, four-factor lentiviral and three-factor retroviral) and two non-integrating methods (episomal vector and messenger RNA delivery into fibroblasts). All hiPS cell lines were extensively characterized for pluripotency and had normal karyotypes before DNA extraction (Supplementary Methods). Protein-coding regions in the genome were captured and sequenced from the genomic DNA of hiPS cell lines and their matched progenitor fibroblast lines using either padlock probes[10,11] or in-solution DNA or RNA baits[12,13]. We searched for single base changes, small insertions/deletions and alternative splicing variants, and identified 12,000–18,000 known and novel variants for each cell line that had sufficient coverage and consensus quality (Table 1).

hiPS cell lines contain a high level of mutational load

We identified sites that showed the gain of a new allele in each hiPS cell line relative to their corresponding matched progenitor fibroblast genome. A total of 124 mutations were validated with capillary sequencing (Fig. 1, Table 2 and Supplementary Fig. 1), which revealed that each mutation was fixed in heterozygous condition in the hiPS cell lines. No small insertions/deletions were detected. For three hiPS cell lines (CV-hiPS-B, CV-hiPS-F and PGP1-iPS), the donor's complete genome sequence obtained from whole blood is publicly available[14,15]; we used this information to further confirm that all 27 mutations in these lines were bona fide somatic mutations. Because 84% of the expected exomic variants[16] were captured at high depth and quality, the predicted load is approximately six coding mutations per hiPS cell genome (see Table 1 for details). The majority of mutations were mis-sense (83 of 124), nonsense (5 of 124) or splice variants (4 of 124). Fifty-three mis-sense mutations were predicted to alter protein function[17] (Supplementary Table 1). Fifty mutated genes were previously found to be mutated in some cancers[18,19]. For example, *ATM* is a well-characterized tumour suppressor gene found mutated in one hiPS cell line, and *NTRK1* and *NTRK3* (tyrosine kinase receptors) can cause cancers when mutated[20] and contained damaging mutations in three hiPS cell lines (CV-hiPS-F, iPS29e and FiPS4F-shpRB4.5) that were reprogrammed in three labs and came from different donors. Two kinase genes from the *NEK* family, which is related to cell division, were mutated in two independent hiPS cell lines. In addition to cancer-related genes, 14 of the 22 lines contained mutations in genes with known roles in human Mendelian disorders[21]. Three pairs of hiPS cell lines (iPS17a and iPS17b, dH1F-iPS8 and dH1F-iPS9, and CF-RiPS1.4 and CF-RiPS1.9) shared three, two and one mutation, respectively;

[1]Department of Bioengineering, Institute for Genomic Medicine and Institute of Engineering in Medicine, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. [2]Department of Cellular and Molecular Medicine and Howard Hughes Medical Institute, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. [3]Division of Pediatric Hematology/Oncology, Children's Hospital Boston and Dana Farber Cancer Institute, Boston, Massachusetts 02115, USA. [4]Department of Anatomy, University of Wisconsin-Madison, Madison, Wisconsin 53705, USA. [5]Center of Regenerative Medicine, 08003 Barcelona, Spain. [6]Howard Hughes Medical Institute, Harvard Stem Cell Institute, Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [7]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02135, USA. [8]Salk Institute for Biological Studies, La Jolla, California 92037, USA. [9]The J. Craig Venter Institute, Rockville, Maryland 20850, USA. [10]Immune Disease Institute, Children's Hospital Boston, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

Table 1 | Sequencing statistics for mutation discovery

Cell line	Exome capture method	Quality-filtered sequence (bp)	No. of high-quality coding variants	dbSNP percentage	Shared high-quality coding region (bp)	No. of coding mutations observed/projected
CV-hiPS-F	Padlock + SeqCap EZ	9,928,014,640	15,595	98%	16,374,878	14/15
CV-hiPS-B	SeqCap EZ	7,977,894,480	14,876	98%	21,891,518	10/12
CV fibroblast	Padlock + SeqCap EZ	7,586,731,600	15,442	98%	—	—
DF-6-9-9	Padlock + SeqCap EZ*	9,289,593,520	14,366	95%	17,806,151	6/7
DF-19-11	SeqCap EZ	3,212,662,880	13,792	95%	21,342,017	7/9
iPS4.7	SeqCap EZ	3,132,462,400	14,154	95%	21,729,562	4/5
Foreskin fibroblast	Padlock + SeqCap EZ*	8,430,654,720	14,819	95%	—	—
PGP1-iPS	SeqCap EZ	4,599,556,400	14,105	95%	19,681,915	3/4
PGP1 fibroblast	SureSelect	3,504,437,120	14,781	95%	—	—
dH1F-iPS8	SeqCap EZ	3,950,994,160	13,552	96%	16,874,057	8/10
dH1F-iPS9	SeqCap EZ	3,945,196,800	14,191	95%	21,536,158	3/4
dH1F fibroblast	SeqCap EZ	3,373,535,920	13,838	95%	—	—
iPS11a	SureSelect	1,836,303,440	13,845	95%	18,557,098	4/5
iPS11b	SureSelect	3,378,603,200	15,152	95%	17,206,934	7/8
Hib11 fibroblast	SureSelect	5,660,864,960	13,579	95%	—	—
iPS17a	SureSelect	4,805,756,800	15,039	95%	17,888,773	4/5
iPS17b	SureSelect	7,129,037,520	15,400	95%	19,902,076	5/6
Hib17 fibroblast	SureSelect	3,962,506,880	13,365	96%	—	—
iPS29A	SureSelect	4,112,237,360	13,464	94%	17,328,182	2/3
iPS29e	SureSelect	1,669,916,080	13,800	94%	18,985,791	7/9
Hib29 fibroblast	SureSelect	4,388,388,320	14,445	95%	—	—
dH1cF16-iPS1	SeqCap EZ	4,321,661,440	15,061	95%	19,601,528	2/2
dH1cF16-iPS4	SeqCap EZ	4,668,085,920	14,958	95%	23,956,732	6/7
dH1cF16 fibroblast	SeqCap EZ	4,178,664,160	14,879	95%	—	—
CF-RiPS1.4	SeqCap EZ	4,733,743,840	11,344	96%	21,272,233	2/3
CF-RiPS1.9	SeqCap EZ	3,143,591,760	13,674	95%	21,165,013	5/6
CF fibroblast	SeqCap EZ	3,204,874,880	11,855	96%	—	—
FiPS3F1	SeqCap EZ	3,397,397,360	13,333	94%	20,723,620	4/5
FiPS4F7	SeqCap EZ	3,346,801,280	14,584	94%	21,608,258	2/3
HFFXF fibroblast	SeqCap EZ	3,331,494,880	13,040	94%	—	—
FiPS4F2p9	SeqCap EZ	4,725,258,400	18,033	92%	25,188,054	7/7
FiPS4F2p40	SeqCap EZ	4,848,006,000	18,376	92%	25,411,595	11/11
FiPS4F-shpRB4.5	SeqCap EZ	4,911,008,400	19,491	92%	25,240,944	8/8
IMR90 fibroblast	SeqCap EZ	5,019,916,240	18,220	92%	—	—

Quality-filtered sequence represents the total amount of sequence data generated that passed the Illumina GA IIx quality filter (bp, base pair). The number of high-quality coding variants is the number of variants found with a sequencing depth of at least eight and a consensus quality score of at least 30. The dbSNP percentage represents the percentage of identified variants present in the Single Nucleotide Polymorphism Database. The shared coding region is the portion of the genome, in base pairs, that was sequenced at high depth and quality in both the iPS cell line and its progenitor fibroblast. The number of coding mutations lists both the number of identified coding mutations and a projection of the total number of identified mutations based on the fraction of Consensus Coding Sequence variants[16] (out of ~17,000 expected variants) successfully identified in both hiPS cells and fibroblasts.

* For these cell lines, mutation calling was performed individually using both padlock probe data and hybridization-capture data. Each method found five mutations, four of which were shared, leading to a total of six mutations. Padlock probe and hybridization capture have separate strengths (specificity versus unbiased coverage); it seems that these factors directly affect the ability to find separate mutations.

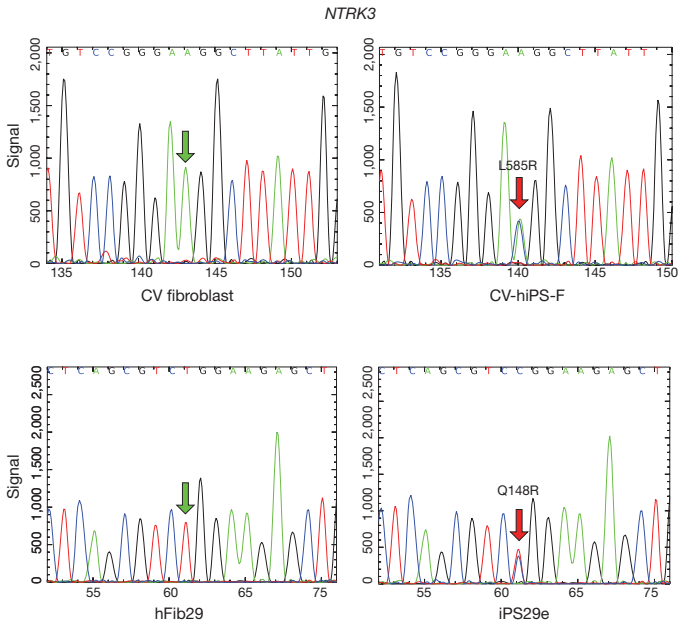

Figure 1 | hiPS cells acquired protein-coding somatic mutations. Somatic mutations in the gene *NTRK3* were found in two independent hiPS cell lines but were not present in their fibroblast progenitors. Detailed information for all mutations is in the Supplementary Information.

these most probably arose in shared common progenitor cells before reprogramming. However, most hiPS cell lines derived from the same fibroblast line did not share common mutations (Table 2 and Supplementary Table 1).

These data raise the possibility that a significant number of mutations occur during or shortly after reprogramming and then become fixed during colony picking and expansion. An alternative hypothesis is that the mutations we found are simply the result of age-accrued biopsy heterogeneity or normal somatic mutation during *in vitro* fibroblast cell culture. The skin biopsies were collected from donors of ages varying from newborn to 82 years; biopsy heterogeneity therefore does not seem to have a primary role, as the mutational load is not correlated (squared linear correlation coefficient, $R^2 = 0.046$) with donor age (Supplementary Fig. 2). We attempted to grow clonal fibroblasts to obtain a control for single-cell mutational load, but a direct assessment was not possible owing to technical difficulties in mimicking the exact culture conditions (Supplementary Methods). Assuming that the skin biopsy is mutation free, we were able to use previously published values for the typical mutation rate in culture to obtain an expectation of ten times fewer mutations per genome than we observed ($P < 1.27 \times 10^{-53}$; Supplementary Methods), indicating that hiPS cell mutational load is higher than normal-culture mutational load. We define the term 'reprogramming-associated mutations' to describe somatic mutations observed in these hiPS cell lines. Reprogramming-associated mutations could pre-exist at low frequencies in the fibroblast population, could occur during the reprogramming process or could occur after reprogramming. All reprogramming-associated mutations have become fixed in the hiPS cell population.

Table 2 | Genes found to be mutated in coding regions in hiPS cells

Cell line	Mutated genes	No. of non-silent mutations	No. detectable at low frequency in fibroblasts (present/tested)
CF-RiPS1.4	*OR52E8, TEAD4*	1	NA
CF-RiPS1.9	*OR52E8, FAM171A1, TMED9, TEAD4, RASEF*	3	NA
CV-hiPS-B	*MMP26, DYNC1H1, VMO1, DSC3, CELSR1, FLT4, UBE2CBP, ARHGEF5, IGF2BP3, DLG3*	7	7/8
CV-hiPS-F	*IQGAP3, SPEN, TNR, PBLD, OR6Q1, INTS4, GSG1, NTRK3, DNAH3, GOLGA4, FAT2, C6orf25, UBR5, SDR16C5*	12	4/7
DF19.11	*SPATA21, RGS8, LPPR4, KCNJ8, SETBP1, ZNF471, TMEM40*	5	NA
DF6-9-9	*ZZZ3, AKR1C4, NEK5, DAPL1, ITCH, PPP1R2*	5	0/5
dH1CF16-iPS1	*IRGQ, TM9SF4*	1	NA
dH1CF16-iPS4	*PKP1, MYOG, ABCA3, PTPRM, RANBP3L, CALN1*	4	NA
dH1F-iPS8	*CABC1 (ADCK3), C1orf100, OR5AN1, CACNG3, MYRIP, SLC1A3, DSP, KLRG2*	6	NA
dH1F-iPS9	*SEMA6C, MYRIP, SLC1A3*	3	NA
FiPS3F1	*SORCS3, GLRA3, CARM1, EPB41L1*	2	NA
FiPS4F7	*GDF3, ZER1*	2	NA
iPS11a	*GTF3C1, SALL1, SLC26A3, ZNF16*	3	1/1
iPS11b	*MARCKSL1, PRDM16, ATM, LRP4, TCF12, SH3PX3 (SNX33), OSBPL3*	5	0/1
iPS17a	*HK1, ANKRD12, SCN1A, IFNGR1*	4	NA
iPS17b	*HK1, CCKBR, ANKRD12, SCN1A, IFT122*	5	1/1
iPS29A	*PRICKLE1, RFX6*	2	2/2
iPS29e	*C14orf174 (SAMD15), NTRK3, VAC14, ASB3, STX7, POLR1C, LINGO2*	6	1/4
iPS4.7	*POLE, UBA2, L3MBTL2, C4orf41*	2	NA
PGP1-iPS	*C11orf67, OSBPL8, NEK11*	1	1/3
FiPS4F2	*TMEM57, RANBP6, CTSL1, SAV1, KRT25, BCL2L12, LGALS1, TTYH2*, COPA*, ARSB*, MT1B**	7	NA
FiPS4F-shpRB4.5	*NTRK1, CD1B, LRCH3, SH3TC1, GPC2, CDK5RAP2, MYH4, TRMU*	5	NA

The full details of each mutation are in Supplementary Table 1.
* Mutation was observed at passage 40 but not at passage 9. FiPS4F2 was sequenced at both passage 9 and passage 40. Seven mutations were present after reprogramming (FiPS4F2P9), and four more became fixed after extended culturing (FiPS4F2P40). All seven mutations found after reprogramming were also present after extended culturing.

Reprogramming-associated mutations arise through multiple mechanisms

To test whether some observed mutations were present in the starting fibroblasts at low frequency before reprogramming, we developed a new digital quantification assay (DigiQ) to quantify the frequencies of 32 mutations in six fibroblast lines using ultradeep sequencing (Supplementary Figs 3 and 4). We amplified each mutated region from the genomic DNA of 100,000 cells with a high-fidelity DNA polymerase and sequenced the pooled amplicons with an Illumina Genome Analyser at an average coverage of 10^6. Although the raw sequencing error is roughly 0.1–1% with the Illumina sequencing platform, detection of rare mutations at a lower frequency is possible with proper filtering and careful selection of controls[22]. For each fibroblast line, we included the mutation-carrying hiPS cell DNA as the positive control and a 'mutation-free' DNA sample as the negative control for sequencing errors (Supplementary Methods). Comparison of the allelic counts at the mutation positions between the fibroblast lines and the negative controls allowed us to distinguish rare mutations from sequencing errors and estimate the detection limit of the assay. Seventeen of the 32 mutations were found in fibroblasts in the range of 0.3–1,000 in 10,000, and 15 mutations were not detectable (Supplementary Tables 2 and 3). In each fibroblast line with more than one detectable rare mutation, the frequencies of the mutations were very similar, which suggests that a small subpopulation of each fibroblast line contains all pre-existing hiPS cell mutations and that the rest of the cells lacked any of them.

We extended this analysis by asking whether all of the hiPS cell mutations could have pre-existed in the fibroblast populations. For the 15 mutations not detected with the DigiQ assay, the detection limits can be estimated (Supplementary Methods). At seven of the 15 sites, the sequencing quality was high enough that rare mutations at frequencies of 0.6–5 in 100,000 should be detectable with our assay (Supplementary Table 3). Because 30,000–100,000 fibroblast cells were used in the reprogramming experiments, we can rule out the presence of two mutated genes (*NTRK3* and *POLR1C*) in more than one cell of the starting fibroblast population, and five others were present in no more than one or two cells.

As another test of the hypothesis that all of the mutations pre-existed in fibroblasts before reprogramming, we examined the exomes of two hiPS cell lines derived from fibroblast line dH1cf16, which was clonally derived from the dH1F fibroblast line and passaged the minimum amount to generate enough cells for reprogramming. The two hiPS cell lines derived from the non-clonal dH1F fibroblast line contained eight and, respectively, three new mutations not found in the fibroblasts; we observed a very similar independent mutational load in the clonal lines (six new mutations in the hiPS cell line dH1cf16-iPS1 and two new mutations in the hiPS cell line dH1cf16-iPS4). Together, these experiments establish that although some of the reprogramming-associated mutations were likely to pre-exist in the starting fibroblast cultures, the others occurred during reprogramming and subsequent culturing. Specific distributions tend to vary across hiPS cell lines (Supplementary Table 3).

Mutations that occur during reprogramming could be due in part to a significantly elevated mutation rate during reprogramming. It is also possible that selection could have an important role. We tested the possibility that an elevated mutation rate might occur because the reprogramming process might be inducing transient repression of *p53* (also known as *TP53*), *RB1* and other tumour suppressor genes, which are known to inhibit reprogramming and are required for normal DNA damage responses. Simian virus 40 large-T antigen, which inactivates tumour suppressor and DNA damage response genes[23] (including *p53* and *RB1*), was expressed during reprogramming of three analysed hiPS cell lines (DF6-9-9, DF19-11 and iPS4.7)[24]. Another hiPS cell line (FiPS4F-shpRB4.5) was generated while directly knocking down *RB1* (Supplementary Fig. 5). However, the observed mutational load was very similar in these lines in comparison with the others, indicating that reprogramming-associated mutations cannot be explained by an elevated mutation rate caused by *p53* or *RB1* repression.

We also probed whether additional mutations could become fixed during extended passaging by extending our analysis of one hiPS cell line. Although most of our hiPS cell lines were sequenced at fairly low passage number (less than 20), to measure the effect of post-reprogramming culturing directly we also sequenced one hiPS cell line (FiPS4F2) at two passages (9 and 40). We discovered that all seven mutations identified in the passage-9 line remained fixed in the passage-40 line, but that four additional mutations were found to be fixed in the passage-40 cell line.

To test the possibility that selection operates during hiPS cell generation, we performed an enrichment analysis to determine whether reprogramming-associated mutated genes were more likely to be

observed than random somatic mutation in cancer cells. We used the COSMIC database as a source of genes commonly mutated in cancer[18]. We discovered that the reprogramming-associated mutated genes were significantly enriched for genes found mutated in cancer ($P = 0.0019$; Supplementary Information), which implies that some mutations were selected during reprogramming.

As an alternative test of the selection hypothesis, we asked whether mutations associated with reprogramming could be functional, on the basis of the non-synonymous/synonymous (NS/S) ratio. Traditionally, the analysis of the NS/S ratio is applied to germline mutations that have evolved over a long period of evolutionary time, and is not directly applicable to somatic mutations. However, functional mutations are known to be positively selected in cancers, allowing us to make a direct comparison with mutation characteristics found in cancer genomes. Strikingly, the NS/S ratio is very similar between mutations identified in three recent cancer genome sequencing projects[25–27] and the reprogramming-associated mutations we found (2.4/1 and 2.6/1, respectively), indicating that a similar degree of selection pressure may be present.

We also checked whether reprogramming-associated mutations could provide a common functional advantage, through a pathway enrichment analysis using Gene Ontology terms[28]. No statistically significant similarity was identified, indicating that mutated genes have varied cellular functions. Again, identical results were found when performing the same analysis on mutations identified during the genome sequencing of melanoma, breast cancer and lung cancer samples[25–27]. This lack of enrichment in cancer genomes is generally thought to be due to the presence of many passenger mutations in cancer cells, which could also be the cause for reprogramming-associated mutations. Nonetheless, these analyses suggest that selection of potentially functional mutations could have a role in amplifying rare-mutation-carrying cells and, when coupled with the single-cell bottleneck in hiPS cell colony picking, could contribute to the fixation of initially low-frequency mutations throughout the entire hiPS cell population.

Discussion

Taken together, our results demonstrate that pre-existing and new mutations that occur during and after reprogramming all contribute to the high mutational load we discovered in hiPS cell lines. Although we cannot completely rule out the possibility that reprogramming itself is 'mutagenic', our data argue that selection during hiPS cell reprogramming, colony picking and subsequent culturing may be contributing factors. A corollary is that if reprogramming efficiency is improved to a level such that no colony picking and clonal expansion is necessary, the resulting hiPS cells could potentially be free of mutations.

Despite the power of our experimental approach to identify and characterize reprogramming-associated mutations accurately, their functional significance remains to be shown. This issue parallels a general problem facing the genomics community: high-throughput sequencing technologies have allowed data generation rates to greatly outpace functional interpretation. Additionally, when considering the biological significance of reprogramming-associated mutations, there are two separate functional aspects to consider: whether some of these mutations contributed functionally to the reprogramming of cell fate, and whether some of these mutations could increase disease risk when hiPS-cell-derived cells/tissues are used in the clinic. These two aspects are not necessarily connected. Although the functional effects of the 124 mutations remain to be characterized experimentally, it is nonetheless striking that the observed reprogramming-associated mutational load shares many similarities with that observed in cancer. Furthermore, the observation of mutated genes involved in human Mendelian disorders suggests that the risk of diseases other than cancer needs to be evaluated for hiPS-cell-based therapeutic methods. Future long-term studies must focus on functional characterization of reprogramming-associated mutations to aid further the creation of clinical safety standards.

Safe hiPS cells are critical for clinical application. Therefore, just as previous findings of large-scale genome rearrangements in hiPS cell lines led to the introduction of karyotyping as a standard post-reprogramming protocol, routine genetic screening of hiPS cell lines to ensure that no obviously deleterious point mutations are present must become a standard procedure. Complete exome or genome sequencing of hiPS cell lines might be an efficient way to screen out hiPS cell lines that have a high mutational load or have mutations in genes implicated in development, disease or tumorigenesis. Further rigorous work on mutation rates and distributions during *in vitro* culturing and reprogramming of hiPS cells, and perhaps human embryonic stem cells, will be essential to help establish clinical safety standards for genomic integrity.

METHODS SUMMARY

CV-hiPS-F and CV-hiPS-B were reprogrammed from CV fibroblasts using four-factor retroviral vectors. PGP1-iPS cells were reprogrammed by Cellular Dynamics using the same four factors in a lentiviral vector from PGP1F fibroblasts[29]. We obtained dH1F-iPS8, dH1F-iPS9, dH1cF16-iPS1, dH1cF16-iPS4, dH1cF16 and dH1F cells from previous cultures[30] reprogrammed with retroviral vectors containing the same factors[31]. We obtained DF-6-9-9, DF-19-11, iPS4.7 and FS cells from previously existing cultures; the reprogramming process and characterization of lines has been described previously[24]. We obtained iPS11a, iPS11b, iPS17a, iPS17b, iPS29A, iPS29e, Hib11, Hib17 and Hib29 cells from previous cultures reprogrammed using retroviral vectors encoding three or four factors[32]. FiPS3F1 and FiPS4F7 were reprogrammed from HFFxF fibroblasts using similar protocols[33–35]. FiPS4F2 and FiPS4F-shpRB4.5 were reprogrammed using the same four-factor protocol from IMR90 fibroblasts. We obtained the mRNA-derived lines (CF-RiPS1.4, CF-RiPS1.9 and CF fibroblasts) from previous cultures[36]. All hiPS cell lines were extensively characterized for pluripotency. Fourteen lines were tested for teratoma formation and shown to express all embryonic germ layers *in vivo*. DNA was extracted from each cell type using Qiagen's DNeasy kit.

Exome capture was performed with either a library of padlock probes, commercial hybridization-capture DNA baits (NimbleGen SeqCap EZ) or RNA baits (Agilent SureSelect), and the resulting libraries were sequenced on an Illumina GA IIx sequencer. We rejected putative mutations if they were known polymorphisms or contained any minor allele presence in the fibroblast. All candidate mutations were confirmed using capillary Sanger sequencing.

For digital quantification, mutations were PCR-amplified and sequenced using an Illumina GA IIx. These libraries were sequenced to obtain on average 1,000,000 independent base calls for each location. A binomial test was then used to determine whether the observed minor allele frequency could be separated from error and to estimate the frequency of each mutation.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 April 2010; accepted 12 January 2011.

1. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
2. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–1920 (2007).
3. Mayshar, Y. *et al.* Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell* **7,** 521–531 (2010).
4. Hong, H. *et al.* Suppression of induced pluripotent stem cell generation by the p53–p21 pathway. *Nature* **460,** 1132–1135 (2009).
5. Li, H. *et al.* The Ink4/Arf locus is a barrier for iPS cell reprogramming. *Nature* **460,** 1136–1139 (2009).
6. Kawamura, T. *et al.* Linking the p53 tumour suppressor pathway to somatic cell reprogramming. *Nature* **460,** 1140–1144 (2009).
7. Utikal, J. *et al.* Immortalization eliminates a roadblock during cellular reprogramming into iPS cells. *Nature* **460,** 1145–1148 (2009).
8. Marión, R. M. *et al.* A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity. *Nature* **460,** 1149–1153 (2009).
9. Ruiz, S. *et al.* A high proliferation rate is required for somatic cell reprogramming and maintenance of human embryonic stem cell identity. *Curr. Biol.* **21,** 45–52 (2011).
10. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4,** 931–936 (2007).
11. Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature Biotechnol.* **27,** 353–360 (2009).
12. Bashiardes, S. *et al.* Direct genomic selection. *Nature Methods* **2,** 63–69 (2005).
13. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnol.* **27,** 182–189 (2009).

14. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
15. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2009).
16. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).
17. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* **4,** 1073–1081 (2009).
18. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protocols Hum. Genet.* **10,** 10.11 (2008).
19. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461,** 809–813 (2009).
20. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4,** 177–183 (2004).
21. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33,** D514–D517 (2005).
22. Druley, T. E. *et al.* Quantification of rare allelic variants from pooled genomic DNA. *Nature Methods* **6,** 263–265 (2009).
23. Ahuja, D., Saenz-Robles, M. T. & Pipas, J. M. SV40 large T antigen targets multiple cellular pathways to elicit cellular transformation. *Oncogene* **24,** 7729–7745 (2005).
24. Yu, J. *et al.* Human induced pluripotent stem cells free of vector and transgene sequences. *Science* **324,** 797–801 (2009).
25. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
26. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465,** 473–477 (2010).
27. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
28. Dennis, G. Jr *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4,** P3 (2003).
29. Lee, J. H. *et al.* A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet.* **5,** e1000718 (2009).
30. Park, I. H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451,** 141–146 (2008).
31. Chan, E. M. *et al.* Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. *Nature Biotechnol.* **27,** 1033–1037 (2009).
32. Dimos, J. T. *et al.* Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. *Science* **321,** 1218–1221 (2008).
33. Rodriguez-Piza, I. *et al.* Reprogramming of human fibroblasts to induced pluripotent stem cells under xeno-free conditions. *Stem Cells* **28,** 36–44 (2010).
34. Aasen, T. *et al.* Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nature Biotechnol.* **26,** 1276–1284 (2008).
35. Stewart, S. A. *et al.* Lentivirus-delivered stable gene silencing by RNAi in primary cells. *RNA* **9,** 493–501 (2003).
36. Warren, L. *et al.* Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* **7,** 618–630 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. M. Akey, G. M. Church, S. Ding, J. B. Li and J. Shendure for discussions and suggestions, S. Vassallo for assistance with DNA shearing, and G. L. Boulting and S. Ratansirintrawoot for assistance with hiPS cell culture. This study is supported by NIH R01 HL094963 and a UCSD new faculty start-up fund (to K.Z.), a training grant from the California Institute for Regenerative Medicine (TG2-01154) and a CIRM grant (RC1-00116) (to L.S.B.G.). L.S.B.G. is an Investigator of the Howard Hughes Medical Institute. A. Gore is supported by the Focht-Powell Fellowship and a CIRM predoctoral fellowship. M.L.W. is supported by an institutional training grant from the National Institute of General Medical Sciences (T32 GM008666). Y.-H.L. is supported by the A*Star Institute of Medical Biology and the Singapore Stem Cell Consortium. Work in the laboratory of J.C.I.B. was supported by grants from MICINN, Sanofi-Aventis, the G. Harold and Leila Y. Mathers Foundation and the Cellex Foundation. G.Q.D. is an investigator of the Howard Hughes Medical Institute and supported by grants from the NIH.

Author Contributions L.S.B.G. and K.Z. co-directed the study. A. Gore, Z.L., L.S.B.G. and K.Z. designed the experiments. J.E.Y., S.A., J.A.-B., I.C., A. Giorgetti, M.A.I., E.K., J.-H.L., Y.-H.L., P.D.M., N.M., A.D.P., S.R., M.L.W., J. Yu, J.C.I.B., D.J.R., J.A.T., K.E., G.Q.D. and L.S.B.G. biopsied, cultured and derived hiPS cell lines. Z.L. performed DNA extraction. A. Gore, Z.L. and K.Z. performed exome library construction, DigiQ library construction and validation Sanger sequencing. H.-L.F. performed Illumina sequencing. A. Gore and K.Z. performed bioinformatic and statistical analysis with contributions from E.F.K. A. Gore, Z.L., L.S.B.G., G.Q.D. and K.Z. wrote the manuscript with contributions from all other authors.

Author Information Sequencing results for the mutations reported here are included in Supplementary Figure 1. Raw Illumina sequencing reads are available from the NCBI Short Read Archive, accession SRP005709, except for lines derived from Hib11, Hib17, Hib29, CF, HFFxF, dH1F fibroblasts as the original donors were not consulted about public release of their genome data. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to L.S.B.G. (lgoldstein@ucsd.edu) or K.Z. (kzhang@bioeng.ucsd.edu).

METHODS

CV fibroblast derivation. Primary fibroblasts were established from a 4-mm dermal punch biopsy of a 63-year-old male using a protocol based on Takashima's method[37]. The biopsy and subsequent reprogramming protocols and the informed-consent documents were reviewed and approved by the UCSD institutional ESCRO and IRB. Briefly, collagenase type 1A (Sigma) was used to dissociate the biopsy and cells were cultured in fibroblast media (DMEM containing 15% FBS, penicillin/streptomycin, sodium pyruvate, non-essential amino acids and L-glutamine). Fibroblasts were reprogrammed at passage 5. DNA was isolated for sequencing from 3,000,000 fibroblasts at passage 9.

CV-hiPS-B and CV-hiPS-F derivation. For reprogramming, ~100,000 fibroblasts per well were transduced with pCX4 retroviral vectors encoding *OCT4* (*POU5F1*), *SOX2*, *KLF4*, *c-MYC* (*MYC*) and ±*EGFP*. CV-hiPS-B and CV-hiPS-F were derived from the +*EGFP* and −*EGFP* transductions, respectively. Transduced fibroblasts were trypsinized and seeded onto irradiated mouse embryonic fibroblasts (MEFs) and cultured in HUES media[38]. Cultures were treated with 2 mM valproic acid for the first seven days post-transduction and 10 nM Y-27632 for the first three weeks (both from EMD Chemicals). After about three weeks post-transduction, individual colonies that morphologically resembled hES were isolated and expanded. Established hiPS cell lines were maintained in HUES media and dissociated cultures for subculturing using 0.05% trypsin/EDTA. DNA for sequencing was isolated from CV-hiPS-B and CV-hiPS-F at passages 13 and 9, respectively.

CV-hiPS characterization. For PCR analysis with reverse transcription, hiPS cells were purified away from MEFs by passage onto Matrigel. Cells were collected and total RNA was isolated with the Ambion PaRIS kit following manufacturer's protocols. First-strand complementary DNA was generated with Superscript II (Invitrogen) following manufacturer's protocols. cDNA was amplified with primers specific for endogenous *SOX2*, *NANOG* and *OCT4* for 30 cycles. For immunofluorescence experiments, cells were passaged onto Matrigel-coated coverslips and samples were processed using standard methods. Antibodies were used at the following dilutions: *NANOG* (Santa-Cruz Biotechnology, 1:200), Tra-1-81 (BD Biosciences, 1:500), *SOX2* (Chemicon, 1:2,000). Cell Line Genetics performed karyotype analysis of CV hiPS cell lines. For embryoid body generation, hiPS cells were passaged with dispase and plated in suspension culture in embryoid body media (DMEM, 20% FBS, L-glutamine and NEAA) for eight days. On day eight, embryoid bodies were plated onto either Matrigel- or polyornithine/laminin-coated coverslips and cultured in either embryoid body media (for endoderm/mesoderm) or neural differentiation media (DMEM-F12, glutamax, N2 and B27) supplemented with dbcAMP, *BDNF* and *GDNF* (for neuroectoderm) for eight days. On day nine, cells were fixed and processed for immunofluorescence as described above. Cell Line Genetics performed karyotype analysis of CV-hiPS cell lines.

CV-hiPS-B was purified away from MEFs by culturing on Matrigel (BD Biosciences) for two passages. CV-hiPS-F was purified by dissociation with Accutase (Innovative Cell Technologies), staining with TRA-1-81 antibody (BD Biosciences) and purifying 5,000,000 TRA-1-81^{+} cells using a BD Biosciences FACSAria II flow cytometer.

dH1F-iPS8 and dH1F-iPS9 derivation. The dH1F fibroblast line was derived from the H1-OGN line previously[30]. dH1F-iPS8 and dH1F-iPS9 were reprogrammed[31] with human *OCT4*, *SOX2*, *KLF4* and *c-MYC* retroviral vectors from dH1F at passage 5. Briefly, 293T cells in 15-cm plates were transfected with 6.25 μg of retroviral vector, 0.75 μg of VSVG vector and 5.625 μg of Gag-Pol vector using FUGENE 6 reagents. Three days after transfection, supernatants were filtered through a 0.45-μm cellulose acetate filter, concentrated by centrifugation at 23,000 r.p.m. for 90 min and stored at −80 °C until use. Transductions were carried out on dH1F fibroblast cells in six-well plates (100,000 cells per well). Viruses were added at a multiplicity of infection of five. Three days after infection, cells were split into plates pre-seeded with MEFs. The medium was changed to human ES culture medium five days after infection. hiPS cell clones stared to emerge about two to three weeks later and were picked and expanded in standard human ES cell culture medium (DMEM/F12 containing 20% KOSR, 10 ng ml^{-1} human recombinant basic fibroblast growth factor, ×1 NEAA, 5.5 mM 2-ME, 50 units ml^{-1} penicillin and 50 μg ml^{-1} streptomycin). During cell collection, MEFs were removed by suction pump and collagenase (Gibco) was used to lift the cells. For dH1F, cells were cultured in 10% FBS DMEM. Trypsin-EDTA was used to lift the cells from the plate for collection. DNA was extracted using a Qiagen DNeasy kit at the following passage numbers: 12 (dH1F), 19 (dH1F-iPS8), 17 (dH1F-iPS9).

hiPS 11a, 11b, 17a, 17b, 29A and 29e derivation. Human fibroblasts were generated from 3-mm forearm dermal biopsies following informed consent under an IRB approved by Harvard University. The murine leukaemia retroviral vector pMXs containing the human cDNAs for *KLF4*, *SOX2* and *OCT4*[32] were modified to produce higher-titer virus by including the woodchuck post-transcriptional responsive element of FUGW (Addgene plasmid 14883) downstream of the cDNA. VSV-g pseudotyped viruses were packaged and concentrated by the Harvard Gene Therapy Initiative at Harvard Medical School. To produce hiPS cells, 30,000 human fibroblasts were transduced at a multiplicity of infection of 10–15 with viruses containing all three genes in hES medium with 8 μg ml^{-1} polyprene. Cells were incubated with virus for 24 h before medium was changed to standard fibroblast medium for 48 h. Cells were subsequently cultured in standard hES medium and hiPS cell colonies were manually picked on the basis of morphology within 2–4 weeks. Derived hiPS cell lines (11a, 11b, 17a, 17b and 29e) have been extensively characterized by standard assays including staining for markers of pluripotency by immunocytochemistry, cell cycle analysis, three-germ-layer differentiation potential *in vitro* and *in vivo*, and karyotype analysis[39]. All cell cultures were maintained at 37 °C in 5% CO_2. Human fibroblasts were cultured in KO-DMEM (Invitrogen), supplemented with 20% Earl's salts 199 (Gibco) and 10% hyclone (Gibco), ×1 GlutaMax, penicillin/streptomycin (Invitrogen) and 100 μM 2-mercaptoethanol. hiPS cells were maintained on gelatinized tissue culture plastic on a monolayer of irradiated CF-1 MEFs (GlobalStem), in hES media[38], supplemented with 20 ng ml^{-1} of bFGF. The medium was changed every 24 h and lines were passaged by trypsinization (0.5% trypsin EDTA, Invitrogen) or dispase (Gibco, 1mg ml^{-1} in hES media for 30 min at 37 °C). hiPS cell lines 11a, 11b, 17a, 17b, 29A and 29e were purified from MEFs by using dispase, which selectively detaches stem cells, and then were washed twice to ensure removal of any contaminating MEFs. Genomic DNA was extracted with a Qiagen DNeasy kit at the following passages: 7 (hFib17), 20 (iPS17A), 23 (iPS17B), 7 (hFib11), 24 (hFib11a), 20, (hFib11b), 8 (hFib29), 21 (hFib29e), 36 (hFib29A).

HFFXF fibroblast derivation. Primary fibroblasts were established from a foreskin biopsy of a three-year-old individual as detailed in ref. 33. Briefly, a skin sample was placed in sterile saline solution, divided into small pieces and allowed to be attached to cell culture dishes before the addition of xeno-free human foreskin fibroblast growth medium. Fibroblasts generated under xeno-free conditions (HFFxF) were reprogrammed at passage 3. DNA was isolated for sequencing from 4,000,000 HFFxF fibroblasts at passage 4 with a Qiagen DNeasy kit.

FiPS3F1 and FiPS4F7 generation. For reprogramming, about 100,000 fibroblasts per six-well plate were transduced with 1 ml of retroviral supernatants encoding FLAG-tagged *OCT4*, *SOX2*, *KLF4*, and *c-MYC*(T58A) as described in ref. 34. High-titer VSV-G-pseudotyped retroviruses expressing a polycistronic vector encoding for *OCT4*, *SOX2*, *KLF4* and *GFP* (pMXs OSKG) and containing 5 mg ml^{-1} polybrene were produced as described in ref. 35. Infection was performed as indicated previously[33]. Colonies were picked on the basis of morphology 25–35 days after the initial infection and plated onto fresh irradiated XF HFF (iXF HFF) cells. Xeno-free iPS cell lines FiPS3F1 and FiPS4F7 were maintained by mechanical dissociation in XF-hESm, which is composed of KO-DMEM (Dulbecco's modified Eagle's medium; Invitrogen) supplemented with 15% xeno-free KO-SR (Invitrogen), xeno-free KO-SR growth factor cocktail (×1), 2 mM glutamax, 50 mM 2-mercaptoethanol, penicillin/streptomycin (×0.5, all from Invitrogen), non-essential amino acids (Cambrex) and 20 ng ml^{-1} bFGF (Peprotech).

FiPS3F1 and FiPS4F7 characterization. Derived hiPS cell lines FiPS3F1 and FiPS4F7 have been extensively characterized by staining for markers of pluripotency by immunofluorescence analyses. The following antibodies were used: MAB4360 for Tra-1-60 (1:200), MAB4381 for Tra-1-81 (1:200) and AB5603 for *SOX2* (1:500, all from Chemicon); MC-813-70 for SSEA-4 (1:2) and MC-631 for SSEA-3 (1:2, both from the Developmental Studies Hybridoma Bank at the University of Iowa); C-10 for *OCT4* (1:100, Santa Cruz); EB06860 for *NANOG* (1:100, Everest Biotechnology); and Anti-FLAG (Sigma M2). Three-germ-layer differentiation potential *in vitro* was conducted by means of embryoid body formation, which was induced from colony fragments mechanically collected. For endoderm, embryoid bodies were cultured in KO-DMEM medium supplemented with 10% FBS, 2 mM L-glutamine, 0.1 mM 2-β-mercaptoethanol, non-essential amino acids and penicillin/streptomycin. For mesoderm differentiation, the same medium described above in the presence of ascorbic acid (0.5 mM) was used. For ectoderm induction, embryoid bodies were cultured in N2/B27 medium with the stromal cell line PA6 for two weeks. The medium for each condition was changed every other day. On day 15, cells were fixed and processed for immunofluorescence for the following antibodies: Tuj1 (1:500, Covance), α-fetoprotein (1:400), α-actinin (1:100, Sigma). Teratoma formation assay was performed by injecting about 0.5 × 106 XF-iPS cells into the testes of severe combined immunodeficient beige mice (Charles River Laboratories). Mice were euthanized eight weeks after cell injection, and tumours were processed and analysed following conventional immunohistochemistry protocols (Masson's trichromic stain) and immunofluorescence staining for Tuj1 (1:500, Covance), α-fetoprotein (1:400) and α-actinin (1:100, Sigma). Expression of retroviral transgenes and endogenous pluripotency-associated factors by quantitative PCR with reverse transcription

were conducted as described previously[33]. hiPS cell lines FiPS3F1 and FiPS4F7 were purified from iXF HFF by mechanical dissociation and further culturing on Matrigel (BD Biosciences) for two more passages. DNA for sequencing was isolated from passage 9 for both FiPS3F1 and FiPS4F7 with a Qiagen DNeasy kit.

CF-Fib, CF-RiPS1.4 and CF-RiPS1.9 derivation. CF fibroblasts (CF-Fib) were previously obtained from a skin biopsy taken from an adult with cystic fibrosis, with proper informed consent[36]. CF-induced pluripotent stem cell lines were derived using modified mRNAs coding reprogramming factors *OCT4*, *SOX2*, *KLF4*, *c-MYC* and *LIN28* (OSKML) with molar concentrations in the ratio 3:1:1:1:1, in an atmosphere with 5% oxygen, as previously described[36]. Briefly, 50,000 fibroblasts were plated onto γ-irradiated human neonatal fibroblast feeders (GlobalStem) seeded at 33,00 cells cm^{-2}. For CF-RiPS derivations, the cationic lipid delivery system RNAiMAX was used. First, pooled RNA from the five factors OSKML (100 ng ml^{-1}) was diluted ×5 and the reagent (5 μl of RNAiMAX per microgram of RNA) was diluted ×10 in Opti-MEM basal media (Invitrogen). These components were pooled and incubated for 15 min at room temperature before being dispensed to culture media. Nutristem medium was replaced daily 4 h after transfection, and supplemented with 100 ng ml^{-1} bFGF and 200 ng ml^{-1} B18R (eBioscience). CF-RiPS derivation was performed in low oxygen (5%) in a NAPCO 8000 WJ incubator (Thermo Scientific). Medium was equilibrated in 5% oxygen for approximately 4 h before use and cultures were passaged with TrypLE Select recombinant protease (Invitrogen) on days five and six. The daily RNA dose applied in the RiPSC derivations was 1,200 ng per well (six-well plate format). On day 21, RiPS colonies were mechanically picked and transferred to MEF-coated 24-well plates with standard hESC medium (DMEM/F12 containing 20% KOSR (Invitrogen), 10 ng ml^{-1} bFGF (Gembio), ×1 NEAA (Invitrogen), 0.1 mM b-ME (Sigma), 1 mM L-glutamine (Invitrogen), 50 units ml^{-1} penicillin and 50 μg ml^{-1} streptomycin) containing 5 mM Y27632 (BioMol). Clones were mechanically passaged once more to MEF-coated six-well plates, and then expanded via enzymatic passaging with collagenase IV (Invitrogen). Genomic DNA was extracted with a Qiagen DNeasy kit at the following passages: 9 (CF-Fib), 5 (CF-RiPS1.4), 5 (CF-RiPS1.9).

FiPS4F2 and FiPS4F-shpRb4.5 plasmid construction. pMX-Oct4, pMX-SOX2, pMX-KLF4, pMX-cMyc and pLVTHM were obtained from Addgene (plasmids 17217, 17218, 17219, 17220 and 12247, respectively). For the generation of the mammalian lentiviral plasmid encoding small hairpin RNAs against pRb-specific oligonulceotides (forwards, 5′-CGCGTGTTTCCTCTTCCAAAGTAATTCAAGAGATTACTTTGGAAGAGGAAACTTTTTTGGAAAT-3′; reverse, 5′-CGATTTCCAAAAAAGTTTCCTCTTCCAAAGTAATCTCTTGAATTACTTTGGAAGAGGAAACA-3′), were annealed, phosphorylated with T4 kinase and ligated into MluI/ClaI-linearized pLTVHM plasmid. The design of the small hairpin RNA was carried out using the SFOLD software (http://sfold.wadsworth.org/). All constructs generated were subjected to direct sequencing to rule out the presence of mutations.

FiPS4F2 and FiPS4F-shpRb4.5 retroviral and lentiviral production. Moloney-based retroviral vectors (pMX-) were co-transfected with packaging plasmids (pCMV-gag-pol-PA and pCMV-VSVg) in 293T cells using Lipofectamine (Invitrogen). Retroviral supernatants were collected 24 h after transfection, and passed through a 0.45 mM filter. Second-generation lentiviral vectors (pLVTHM-) were co-transfected with packaging plasmids (psPAX2 and pMD2.G, obtained from Addgene, 12260 and 12259, respectively) in 293T cells using Lipofectamine (Invitrogen). Lentiviral supernatants were collected 36 h after transfection.

FiPS4F2P9, FiPS4F2P40 and FiPS4F-shpRb4.5 derivation. Briefly, for the formation of hiPS cells IMR90 fibroblasts were infected with equal proportions of retroviruses encoding for *OCT4*, *SOX2*, *KLF4* and *c-MYC* plus empty lentiviruses (used to generate the FiPS4F2 line) or lentiviruses encoding small hairpin RNA against pRb (used to generate the line FiPS4F-shpRb4.5) by spinfection of the cells at 1,850 r.p.m. for 1 h at room temperature in the presence of polybrene (4 μg ml^{-1}). After two serial infections, cells were passaged onto fresh MEFs and switched to hES cell medium (DMEM/F12 (Invitrogen) supplemented with 20% Knockout serum replacement (Invitrogen), 1 mM L-glutamine, 0.1 mM non-essential amino acids, 55 mM β-mercaptoethanol and 10 ng ml^{-1} bFGF (Joint Protein Central)) four days after the first infection. For the derivation of hiPS cell lines, colonies were manually picked and maintained on fresh MEF feeder layers for five passages before the growth in Matrigel/mTesR1 (Stem Cell Technologies) conditions. DNA was extracted after nine passages for FiPS4F2P9 and FiPS4F-shpRB4.5 and 40 passages for FiPS4F2P40.

FiPS4F2 and FiPS4F-shpRb4.5 characterization. Cell pellets were lysed in 10 mM Tris-HCl (pH 8), 150 mM NaCl, 1% Triton X100, 1 mM Na_3VO_4, 1 mM PMSF and the Complete protease inhibitor mixture (Roche). Total protein extracts (25 μg) were used for SDS–PAGE, transferred to nitrocellulose membranes (Amersham Biosciences) and analysed using primary antibodies against *OCT4* (sc-5279, Santa Cruz), *SOX2* (AB5603, Chemicom), *RB1* (554136, Pharmingen) and Tubulin (T5168, Sigma). Horseradish-peroxidase-conjugated secondary anti-mouse or rabbit were purchased from Cell Signaling and used at 1:5,000 dilution. Tubulin was used as a loading control. Immunoblots were visualized using SuperSignal solutions following the manufacturer's instructions (Thermo Scientific). Total RNA was isolated using TRIzol Reagent (Invitrogen), and cDNA was synthesized using the SuperScript II Reverse Transcriptase kit for RT–PCR (Invitrogen). Real-time PCR was performed using the SYBR-Green PCR Master mix (Applied Biosystems). Values of gene expression were normalized using GAPDH expression and are shown as fold change relative to the value of the sample control. All the samples were done in triplicate. Primer sequences are available upon request. The hiPS cell lines were cultured in the presence of 20 ng ml^{-1} colcemid for 45 min. The cells were trypsinized, washed with PBS and resuspended in a hypotonic solution by drop-wise addition while vortexing at low speed. After 10 min of incubation at 37 °C, cells were fixed by drop-wise addition of 1 ml of cold Carnoy's fixative. Stained metaphases were analysed with CYTOVISION software (Applied Imaging). Teratoma analyses were performed as described in ref. 34.

Preparation of padlock probes. The design and preparation of padlock probes was based on published methods[10,11,40]. Libraries of long oligonucleotides (140 nucleotides) that cover different exonic regions were synthesized from programmable microarrays (Agilent Technologies). The libraries were amplified by performing 48–96 PCR reactions (100 μl each) with 0.02 nM template oligonucleotides, 200 nM Ap1V4IU primer (G*T*AGACTGGAAGAGCACTGTU), 200 nM Ap2V4 primer (/5Phos/TAGCCTCATGCGTATCCGAT), ×0.2 SybrGreen I and 50 μl Econo Taq PLUS master mix (Lucigen), at 94 °C for 2 min, and then 17 cycles at 94 °C for 30 s, 58 °C for 30 s, 72 °C for 30 s and 72 °C for 3 min. The amplicons were then purified by ethanol precipitation. Libraries were then digested with 40 units of Lambda Exonuclease (5 U μl^{-1}, NEB) in ×1 Lambda Exonuclease buffer (NEB) at 37 °C for 2 h, followed by purification with four Qiagen Qiaquick PCR purification columns for every 48 wells of PCR products. Approximately 8 μg of the purified PCR amplicons were digested with ten units of DpnII (50 U μl^{-1}) and ×1 DpnII buffer at 37 °C for 2 h, followed by the addition of four units of USER enzyme (1 U μl^{-1}, NEB) at 37 °C for 4 h. The DNA was digested with 6% PAGE and purified into single-stranded, 102-nucleotide probes.

Multiplex capture of exonic regions. Padlock probes (600 nM total concentration), 250 ng of genomic DNA, 1 nM suppressor oligonucleotides and ×1 Ampligase buffer (Epicentre) were mixed in a 15-μl reaction and denatured at 95 °C for 10 min, then gradually cooled at the rate of 0.1 °C s^{-1} to 60 °C. The mixture was hybridized at 60 °C for 24 h. To circularize the captured targets, the reactions were then incubated at 60 °C for another 24 h after adding 2 μl of gap-filling mix (two units of AmpliTaq Stoffel (Life Technology), four units of Ampligase (Epicentre), and 500 pmol total dNTP). After circularization, 2 μl of exonuclease mix containing 10 U μl^{-1} exonuclease I (USB) and 100 U μl^{-1} exonuclease III (USB) was added to digest the linear DNA, and the reactions were incubated at 37 °C for 2 h and then inactivated at 94 °C for 5 min.

Amplification of capture circles. The 15-μl circularization products were placed in 100-μl PCR reactions with 200 nM of each primer (NH2-CAGATGTTATCGAGGTCCGAC, NH2-GGAACGATGAGCCTCCAAC, ×0.2 SybrGreen I and ×1 Phusion High-Fidelity PCR Master Mix (NEB) at 98 °C for 1 min, and then 16 cycles at 98 °C for 10 s, 58 °C for 20 s, 72 °C for 20 s and 72 °C for 3 min. The amplicons of the expected size range (200 bp) were purified using Qiagen Qiaquick columns.

Shotgun sequencing library construction. Purified PCR products with the four probe sets on the same template DNA were pooled in equal molar ratio. The PCR products were transferred into Covaris microTubes with snap caps for Covaris AFA shearing using a 10% duty cycle, an intensity setting of 5 and 200 cycles per burst. The sheared DNA was concentrated to 85 μl using a vacufuge, and was then prepared for sequencing library construction using NEBNext DNA Sample Prep Master Mix Set 1 (NEB). The fragmented DNA was end-repaired at room temperature for 30 min in 100-μl reaction consisting of ×1 NEBNext End Repair Reaction Buffer and 5 μl of NEBNext End Repair Enzyme Mix. The DNA was then purified with Qiagen Qiaquick columns. Approximately 500 ng to 1 μg of the end-repaired blunt DNA was incubated in a thermal cycler for 30 min at 37 °C along with ×1 NEBNext dA-Tailing Reaction Buffer and 3 μl of Klenow fragment. The DNA was again purified using Qiagen Qiaquick columns. The purified DNA was size-selected (125–150 nucleotides) using E-Gel SizeSelect 2% (Invitrogen) and concentrated to 36 μl using a vacufuge (Eppendorf). The dA-tailed DNA was then ligated at room temperature for 15 min with ×1 Quick Ligation Reaction Buffer, 1.6 nM Illumina ligation adaptors and 2 μl of Quick T4 DNA ligase. Ligation products were purified using Qiagen Qiaquick columns and amplified by PCR in 100-μl reactions with a 15-μl template, 200 nM Illumina PCR primers, ×0.2 SybrGreen I and ×1 Phusion High-Fidelity PCR Master Mix

(NEB) at 98 °C for 1 min, and then eight cycles at 98 °C for 10 s, 65 °C for 20 s, 72 °C for 15 s and 72 °C for 3 min. The PCR amplicons were purified with Qiaquick PCR purification columns, size-selected (200–275 nucleotides) using 6% PAGE and sequenced on an Illumina Genome Analyser IIx.

Hybridization capture with DNA or RNA baits. Liquid exome capture was performed using the commercial Roche NimbleGen SeqCap EZ Exome kit or the commercial Agilent SureSelect kit (Table 1). Experiments were performed following the manufacturers' protocols. Briefly, genomic DNA was sheared and ligated to Solexa sequencing adaptors. DNA was then hybridized with the SeqCap EZ Exome library or SureSelect RNA baits to capture exomic regions. Exome regions were captured with streptavidin beads and then PCR-amplified with Illumina sequencing adaptors. The resulting libraries were sequenced on an Illumina Genome Analyser IIx.

Consensus sequence generation and variant calling. Reads obtained from the Illumina Genome Analyser were post-processed and quality filtered using GERALD. The end of each read was then mapped to the padlock-probe capturing arm sequences using Bowtie; any reads that successfully mapped were discarded to prevent bias from capturing arms. Reads were then mapped to the whole genome using Bowtie or BWA. Any read that could not be mapped uniquely was discarded to reduce false positives due to sequence homology. The 5′ and 3′ ends of reads were then trimmed to reduce the effect of sequencing errors, which tend to occur near the beginnings and ends of reads on the Illumina platform. (No trimming was performed when GATK was used for variant calling.) To reduce errors introduced by pre-sequencing amplification, mapped reads that started and ended at identical locations were then removed using SAMtools or Picard to account for these clonal reads. SAMtools or GATK was then used to generate a consensus sequence for each sample by combining the results of each read that mapped to each exomic location. A minimum read depth of eight and consensus quality of 30 was required at every examined location. The consensus sequences were then compared to look for candidate novel mutations in hiPS cells. Variants that occurred at locations present in the dbSNP database (version 130) were removed from consideration to reduce the false-positive rate, as a novel mutation in the hiPS cell line is very unlikely to have been previously characterized in other cell lines and was most probably just not observed in the fibroblast line owing to stochastic sequencing bias. Because sequencing depth was relatively low in a small fraction of exomic regions, allelic imbalance can also lead to false positives, as sites in the fibroblast genome could, for example, be heterozygous but be sequenced as seven copies of the major allele and one copy of the minor allele and called as homozygous. To prevent these false positives, sites in which the fibroblast genome showed even a very small presence of minor allele were removed from consideration as candidate sites for novel mutations (as these sites are most probably truly heterozygous in both lines). Several locations were identified in which the hiPS cell sample consensus sequence showed a heterozygous call but the fibroblast sample consensus sequence showed a homozygous call; these were identified as candidate mutations, as it is expected that during mutational processes, the hiPS cell sample would most probably gain an additional allele. These candidate mutations were then validated by capillary sequencing as below.

Sanger validation of candidate mutations. Genomic DNA (6 ng) was amplified in a 50-μl PCR reaction with 100 nM specifically designed primers near the mutation site and 25 μl Taq ×2 master mix (NEB) at 94 °C for 2 min, followed by 35 cycles at 94 °C for 30 s, 57 °C for 30 s and 72 °C for 30 s, and final extension at 72 °C for 3 min. The PCR products were then purified with Qiagen Qiaquick columns, and 10 ng of purified DNA was pre-mixed with 8 pmol of the sequencing primer for capillary Sanger sequencing by Genewiz.

Clonal fibroblast experiments. In an attempt to determine the mutational load present in single fibroblasts, we performed a reprogramming-like clonal colony purification strategy on fibroblasts. CV fibroblasts were thawed at passage 14 and cultured in fibroblast media (DMEM containing 15% FBS, penicillin/streptomycin, sodium pyruvate, non-essential amino acids and L-glutamine). A confluent 6-cm plate was trypsinized and cells were plated in three 96-well dishes, in the presence (two plates) or absence (one plate) of MEF feeder cells, at limiting dilutions. Another 96-well plate was plated as a reference plate. Using Poisson calculations, cells were diluted and plated such that it was extremely unlikely (<1%) for one well to contain more than one cell (leading to an expectation of eight wells per plate with one cell). These wells were cultured and progressively passaged from the 96-well dish to a 6-cm plate (96-well, 48-well, 24-well, 12-well, 6-well, 6-cm). For cells growing on MEFs, all passages from a 12-well dish to a 6-cm dish were done without MEFs to minimize contamination with mouse cells in the sequencing analysis. Only three MEF-free wells and nine MEF-containing wells successfully grew; using Poisson calculations, 24 wells should have successfully grown.

All fibroblasts grown from single cells showed heavy signs of stress. Cells grew very slowly (with passaging needed approximately every one to two weeks). MEF-free cells had a flattened morphology, whereas MEF-plated cells maintained a normal, spindle-shaped morphology. Cells tended to senesce very soon after plating; only a few cells grew successfully. Seven clonal lines were sequenced (three grown without MEFs and four grown with MEFs). Six of the lines contained a very high number of putative mutation candidates (~100), and no mutations were found in one line grown on MEFs. We randomly selected 21 of the 600 mutation candidates for Sanger validation, and found that approximately 50% were true positives. This leads to a projection of ~50 protein-coding mutations in six clonal fibroblast lines, which is tenfold more than what was observed in hiPS cells and not consistent with the observations on the other clonal fibroblast line, which was completely mutation free. We proposed that the mutations in the six clonal fibroblast lines were due to the stress associated with expanding single fibroblast cells. Because fibroblast growing conditions are very different from those found in reprogramming, we cannot estimate the background somatic mutation rate in such an experiment. We therefore instead used published estimates of fibroblast mutation rate to estimate clonal fibroblast mutational load (see below).

Digital quantification of mutations. Thirty-two pairs of DigiQ-PCR primers were designed such that the forward or reverse primers are roughly 25 base pairs away from the 5′ end of each mutation site. This ensured that the mutations of interest were sequenced in the part of the read length that had the highest accuracy. Primers also contained an annealing region for Illumina Solexa sequencing primers at the 5′ ends. Each primer corresponding to a different mutation was amplified with a high-fidelity polymerase in three samples: the mutated hiPS cell line, the progenitor fibroblast line and a clean control. To sample DNA from 100,000 cells, 600 ng of DNA was used for each mutated hiPS cell line and fibroblast line. In cases where a separate clonal hiPS cell line not containing the mutation in question was available, this line was used as a clean control, as the chance of this line acquiring the same mutation during clonal expansion is extremely low ($\sim 10^{-9}$ for one mutation). In other cases, a 'low-input' sample using 300 pg of DNA (~50 cells) was used, as rare mutations are unlikely to be present in such a small quantity of DNA. If any mutated DNA was sampled, it would be immediately obvious in the sequencing results and the experiment could be repeated. First-round PCR amplification was performed with 600 ng (~100,000 cells) of DNA, 500 nM of each DigiQ-PCR primer and ×1 iProof High-Fidelity Master Mix (Bio-Rad) at 98 °C for 30 s, followed by ten cycles at 98 °C for 10 s, 59 °C for 20 s and 72 °C for 15 s, 18 cycles at 98 °C for 10 s and 72 °C for 20 s, and final extension at 72 °C for 3 min. The PCR amplicons were purified using Qiaquick columns (Qiagen). Roughly 100 ng of the first-round PCR product was used as a template for second-round PCR amplification, together with ×1 Phusion High-Fidelity PCR Master Mix (NEB) and 200 nM of each Illumina PCR primer, at 98 °C for 30 s, followed by ten cycles at 98 °C for 10 s and 64 °C for 30 s, and final extension at 72 °C for 30 s. The amplicons were purified again with Qiaquick columns (Qiagen) and size-selected (roughly 150–200 nucleotides) using an E-Gel SizeSelect 2% system (Invitrogen). PCR reactions were performed with the iProof High-Fidelity Master Mix (Bio-Rad) and Phusion High-Fidelity PCR Master Mix (NEB) to minimize amplification errors. All size-selected products were pooled together at equal ratio; these libraries were then mixed with the Illumina PhiX control library in a roughly equal ratio to balance the fluorescent signals at all four bases and improve the base-calling accuracy, and sequenced using an Illumina GA IIx. Each pair of libraries from the fibroblasts and negative controls was sequenced in two non-adjacent lanes of a same flow cell. Extreme care was taken in sample handling to ensure no cross-contamination from the positive control libraries to the other libraries. Alleles identified at each mutation position by the sequencer were counted and evaluated. The specific sample choices for each mutation (and raw allele counts) are listed in Supplementary Table 2 (for details, see Supplementary Fig. 3 and Supplementary Table 3). To verify the robustness of the DigiQ assay, the assay was repeated on CV fibroblasts. The obtained read proportions were extremely similar (Supplementary Fig. 4).

Statistical analysis—probability of mutations occurring naturally. We evaluated the likelihood that the mutations found were generated during fibroblast culturing and reprogramming (assuming a clean starting population of fibroblasts) at the normal estimated somatic mutation rate of between 10^{-6} and 10^{-7} non-synonymous coding mutations per gene per cell division, which corresponds to a rate of 6.7×10^{-10} (using the average human coding-region size of 1,500 base pairs per gene[41]). Assuming that mutations are independent events that occur uniformly across the genome, the number of expected mutations during fibroblast culturing and reprogramming can be estimated using a Poisson distribution with expected value $\lambda = 6.7 \times 10^{-10} ns$, where n is the number of cell divisions and s is the observed sequence. Although accurate records of the number of cell divisions experienced by each line during expansion and reprogramming are not available, we estimated that 30–35 doublings had occurred for six lines with well-documented culture histories. In these lines, a total of 206,227,380 base pairs were pairwise-sequenced (at a depth of at least eight and quality of at least 30). This

leads to a Poisson distribution with $\lambda = 4.13–4.81$ for the expected number of mutations. In this case, we observed 54 coding mutations, leading to a P value of 1.29×10^{-40}–2.72×10^{-37}. If this calculation is extrapolated to all 22 lines, we expect $\lambda = 8.7–10.1$ coding mutations; we observed 91, leading to a P value of 4.29×10^{-59}–1.27×10^{-53}. We can therefore say that these mutations did not occur by chance with more than 99% confidence for all 22 lines.

Statistical analysis—digital quantification. To quantify the frequency of each mutation in the fibroblast samples, a one-tailed binomial distribution test was used. Reads were quality-filtered; only base calls with a Phred-like quality score of 30 or greater were considered. We denote by p the probability of obtaining a sequencing read containing the minor allele. The fibroblast sample was compared with either the clean low-input sample or a clean clonal hiPS cell line. Because the two hiPS cell lines are clonally independent, they will not share any mutations. Therefore, for example, FS-low can be used as a negative control for FS and CV-hiPS-B can be used as a negative control for CV-hiPS-F. Any minor allele obtained from the clonal hiPS cell or low-input fibroblast sample will be purely due to sequencing error. We denote by H0 the event that the minor allele frequency in the fibroblast sample was less than or equal to the minor allele frequency in the other clonal/low-input sample, and denote by H1 the event that the minor allele frequency in the fibroblast sample was greater. If H0 is found to be true, the mutation cannot be detected in the fibroblast, as any presence of the minor allele cannot be distinguished from sequencing error. If H1 is found to be true, the presence of the minor allele is detectable and can be quantified. We denote by n the total number of reads that called the mutated position. A critical value of $a = 0.01$ was chosen (99% confidence). Because the number of reads for each sample was very high, both np and $n(1-p)$ were greater than five, meaning that the minor allele presence could be approximated with a normal distribution. We can therefore set a criterion for rejection of the null hypothesis of $Z = (x-\mu)/s > 2.33$, where x is the minor allele count, μ is the mean of the minor allele counts of the fibroblast and low-input/clonal samples, and s is the standard deviation of the minor allele counts of the fibroblast and low-input/clonal samples. For a binomial-distribution approximation, n is the number of reads in the fibroblast sample, p is the minor allele frequency if the fibroblast and low-input/clonal data are merged, $\mu = np$, and $s = np(1-p)$. If the value of Z is greater than 2.33, we are capable of distinguishing the observed fraction of minor alleles in the fibroblast sample from that observed in the clonal/low-input sample. These results are presented in Supplementary Table 3.

We can also construct a 99% confidence interval using the normal approximation for the binomial distribution. Although we observed a value for the minor allele in each fibroblast sample, due to sequencing error, this value may overestimate or underestimate the true minor allele frequency. We can counteract this error using a normal distribution. The confidence-interval values are derived from the normal probability density function and represent the boundaries that we are 99% sure the true minor allele frequency lies within: lower bound, $\min((-2.57s + x)/n, 0)$; upper bound, $\min((2.57s + x)/n, 0)$. These estimates for the minor allele fraction in fibroblasts are shown in Supplementary Table 3. An example of calculation is shown in Supplementary Note.

Statistical analysis—NS/S mutation ratio. To determine whether selection pressure could have a role in reprogramming-associated mutations, we compared the mutational load associated with reprogramming with that associated with tumorigenesis. The NS/S ratio found in several previously conducted pairwise cancer sequencing analyses[25–27] was found to be 2.4:1. The load found here out of 124 identified mutations is 2.6:1, meaning that hiPS cell lines carry a very similar mutational pattern to cancer lines.

Statistical analysis—pathway and COSMIC gene enrichment. To check for enrichment of reprogramming-associated mutated genes in cancer-related genes, the fraction of genes mutated in hiPS cells found mutated in the COSMIC[18] database was identified as 50/124. As 4,471 of the 16,017 genes well targeted by our exome sequencing pipeline are considered to be commonly mutated in cancer, a χ^2 test with one degree of freedom can be used to test for equivalency of distribution. The obtained χ^2 value is 9.67, indicating that the fraction of mutated hiPS cell genes in the COSMIC set is statistically significantly greater than the normally obtained number with a P value of 0.001873. This indicates that hiPS cell mutations are enriched in COSMIC set genes at approximately 1.5-fold the normal level, of 28%, with >99% confidence. To check for commonly mutated pathways, reprogramming-associated mutated genes and mutated genes identified in three cancer sequencing papers[25,26,27] were analysed using DAVID[28]. No statistically significant pathway Gene Ontology terms were identified; the lowest Benjamini P value found was 0.6, which is well above the cut-off value, of 0.01, required for 99% confidence. Therefore, no common pathways seem to be mutated in hiPS cells.

37. Akagi, T., Sasai, K. & Hanafusa, H. Refractory nature of normal human diploid fibroblasts with respect to oncogene-mediated transformation. *Proc. Natl Acad. Sci. USA* **100,** 13567–13572 (2003).
38. Cowan, C. A. *et al.* Derivation of embryonic stem-cell lines from human blastocysts. *N. Engl. J. Med.* **350,** 1353–1356 (2004).
39. Boulting, G. L, *et al.* A functionally characterized test set of human induced pluripotent stem cells. *Nature* advance online publication, doi:10.1038/nbt.1783 (3 February 2011).
40. Zhang, K. *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nature Methods* **6,** 613–618 (2009).
41. Meena Kishore, S., Vincent, T. K. C. & Pandjassarame, K. Distributions of exons and introns in the human genome. *In Silico Biol.* **4,** 387–393 (2004).