Published in final edited form as:

Nature. 2011 March 3; 471(7336): 63–67. doi:10.1038/nature09805.

Somatic coding mutations in human induced pluripotent stem cells

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

[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, CA, 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, CA, 92093, USA

[3]Division of Pediatric Hematology/Oncology, Children's Hospital Boston and Dana Farber Cancer Institute, Boston, MA 02115, USA

[4]Department of Anatomy, University of Wisconsin-Madison, Madison, Wisconsin, USA

[5]The Howard Hughes Medical Institute, Harvard Stem Cell Institute, Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA

[6]Department of Genetics, Harvard Medical School, Boston, MA 02135, USA

[7]The J. Craig Venter Institute, Maryland, MD, USA

[8]Immune Disease Institute, Children's Hospital Boston, Boston, MA, USA

[9]Center of Regenerative Medicine, Barcelona, Spain

[10]Salk Institute for Biological Studies, La Jolla, CA 92037, USA

Abstract

Correspondence and requests for materials should be addressed to L.S.B.G. (lgoldstein@ucsd.edu) or K.Z. (kzhang@bioeng.ucsd.edu) .

*Equal contributions.

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

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

The authors declare no competing financial interests.

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, while the rest were newly occurring 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 safety before clinical use.

Introduction

hiPS cells have the potential to revolutionize personalized medicine by allowing immunocompatible stem cell therapies to be developed1,2. However, questions remain about hiPS safety. For clinical use, hiPS lines must be reprogrammed from cultured adult cells, and could carry a mutational load due to normal *in vivo* somatic mutation. Furthermore, many hiPS reprogramming methods utilize oncogenes that may increase the mutation rate. Additionally, some hiPS lines have been observed to contain large-scale genomic rearrangements and abnormal karyotypes after reprogramming3. Recent studies also revealed that tumor suppressor genes, including those involved in DNA damage response, have an inhibitory effect on nuclear reprogramming4-9. 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 twenty-two hiPS lines and the nine matched fibroblast lines from which they originated (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 mRNA delivery into fibroblasts). All hiPS lines were extensively characterized for pluripotency and had normal karyotypes prior to DNA extraction (Supplementary Methods). Protein coding regions in the genome were captured and sequenced from the genomic DNA of hiPS lines and their matched progenitor fibroblast lines using either padlock probes10,11 or in-solution DNA or RNA baits12,13. 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 line compared with their corresponding matched progenitor fibroblast genome. A total of 124 mutations were validated with capillary sequencing (Figure 1, Table 2, Supplementary Figure S1), which revealed that each mutation was fixed in heterozygous condition in the hiPS lines. No small insertions/deletions were detected. For three hiPS lines (CV-hiPS-B, CV-hiPS-F, PGP1-

iPS), the donor's complete genome sequence obtained from whole blood is publicly available14,15; 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 variants16 were captured at high depth and quality, the predicted load is approximately 6 coding mutations per hiPS genome (see Table 1 for details). The majority of mutations were missense (83/124), nonsense (5/124), or splice variants (4/124). Fifty-three missense mutations were predicted to alter protein function17 (Supplementary Table S1). Fifty mutated genes were previously found to be mutated in some cancers18,19. For example, *ATM* is a well-characterized tumor suppressor gene found mutated in one hiPS line, while *NTRK1* and *NTRK3* (tyrosine kinase receptors) can cause cancers when mutated20 and contained damaging mutations in three hiPS lines (CV-hiPS-F, iPS29e, FiPS4F-shpRB4.5) reprogrammed in three labs from different donors. Two *NEK* kinase genes, a family related to cell division, were mutated in two independent hiPS lines. In addition to cancer-related genes, fourteen of the twenty-two lines contain mutations in genes with known roles in human Mendelian disorders21. Three pairs of hiPS lines (iPS17a and iPS17b, dH1F-iPS8 and dH1F-iPS9, CF-RiPS1.4 and CF-RiPS 1.9) shared three, two, and one mutation respectively; these most likely arose in shared common progenitor cells prior to reprogramming. However, most hiPS lines derived from the same fibroblast line did not share common mutations (Table 2 and Supplementary Table S1).

These data raise the possibility that a significant number of mutations are occurring 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 *in vitro* fibroblast cell culture. The skin biopsies were collected from donors at ages varying from newborn to 82 years old; biopsy heterogeneity therefore does not appear to play a primary role, as the mutational load is not correlated ($R^2 = 0.046$) with donor age (Supplementary Figure S2). We attempted to grow clonal fibroblasts in order to obtain a control for single-cell mutational load, but a direct assessment was not possible due to technical difficulties in mimicking the exact culture conditions (Supplementary Methods). Assuming the skin biopsy is mutation-free, we can 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 mutational load is high compared to normal culture mutational load. We define the term "reprogramming-associated mutations" to describe mutations observed after reprogramming. Reprogramming-associated mutations could be pre-existing at low frequencies in the fibroblast population, occurring during the reprogramming process, or occurring after reprogramming. All reprogramming-associated mutations have become fixed in the hiPS line population.

Reprogramming-Associated Mutations arise through Multiple Mechanisms

To test whether some observed mutations were present in the starting fibroblasts at low frequency prior to reprogramming, we developed a new digital quantification assay (DigiQ) to quantify the frequencies of 32 mutations in six fibroblast lines using ultra-deep sequencing (Supplementary Figure S3-4). 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 Analyzer 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 quality filtering and careful selection of controls22. For each fibroblast line, we included the mutation-carrying hiPS DNA as the positive control and another "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 a range of 0.3-1000 in ten thousand while 15 mutations were not detectable (Supplementary Table S2-3). In each fibroblast line with more than one detectable rare mutation, the frequency of each mutation was very similar, which suggests that a small sub-population of each fibroblast line appeared to contain all pre-existing hiPS mutations, while the rest of the cells lacked any of them.

We extended this analysis by asking whether all of the hiPS 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). The sequencing quality was sufficiently high at 7 of the 15 sites such that rare mutations at frequencies of 0.6-5 in 100,000 should be detectable with our assay (Supplementary Table S3). Since 30,000-100,000 fibroblast cells were used in the reprogramming experiments, we can rule out the presence of two mutated genes (*NTRK3* and *PLOR1C*) in even one cell of the starting fibroblast population, while five others were present in no more than 1-2 cells.

As another test of the hypothesis that all of the mutations pre-existed in fibroblasts prior to reprogramming, we examined the exomes of two hiPS lines derived from a fibroblast line dH1cf16, which was itself clonally derived from the dH1F fibroblast line and passaged the minimum amount to generate enough cells for reprogramming. The two hiPS lines derived from the non-clonal dH1F fibroblast line contained 8 and 3 new mutations not found in the fibroblasts respectively; we observed a very similar independent mutational load in the clonal lines (6 new mutations in the hiPS line dH1cf16-iPS1 and 2 new mutations in the hiPS line dH1cf16-iPS4). Together, these experiments establish that while some of the reprogramming-associated mutations were likely to pre-exist in the starting fibroblast cultures, the others occurred during reprogramming and subsequent culture. Specific distributions tend to vary across hiPS lines (Supplementary Table S3).

Mutations occurring during reprogramming could be due in part to a significantly elevated mutation rate during reprogramming. It is also possible that selection could play 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*, *RB1*, and other tumor suppressor genes, which are known to inhibit reprogramming and are required for normal DNA damage responses. *SV40* Large-T antigen, which inactivates tumor suppressor and DNA damage response genes (including *p53* and *p105*/*RB1*)23, was expressed during reprogramming of three analyzed hiPS lines (DF6-9-9, DF19-11, and iPS4.7).24. Another hiPS line (FiPS4F-shpRB4.5) was generated while directly knocking down *RB1* (Supplementary Figure S5). However, the observed mutational load was very similar in

these lines compared to the others, indicating that reprogramming-associated mutations cannot be explained by an elevated mutation rate caused by *p53* or *RB1* repression.

We also probed if additional mutations could become fixed during extended passaging by extending our analysis of one hiPS line. While most of our hiPS lines were sequenced at fairly low passage number (less than 20), to directly measure the effect of post-reprogramming culture we also sequenced one hiPS line (FiPS4F2) at two passages (p9 and p40). 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 is operating during hiPS generation, we performed an enrichment analysis to determine if reprogramming-associated mutated genes were more likely to be observed in cancer cells than random somatic mutation. We used the COSMIC database as a source of genes commonly mutated in cancer. We discovered that the reprogramming-associated mutated genes were significantly enriched for genes found mutated in cancer ($p=0.0019$, Supplementary Materials), which implies some mutations were selected during reprogramming.

As an alternative test of the selection hypothesis, we asked whether mutations associated with reprogramming could be functional based on the nonsynonymous:synonymous (NS:S) ratio. Traditionally, the analysis of the NS:S ratio is applied to germline mutations evolved over a long period of evolutionary time, which is thus not directly applicable to somatic mutations. However, functional mutations are known to be positively selected in cancers, allowing us to make a direct comparison to mutation characteristics found in cancer genomes. Strikingly the NS:S ratio is very similar between mutations identified in three recent cancer genome sequencing projects[25,26,27] 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 if reprogramming-associated mutations could be providing a common functional advantage using a pathway enrichment analysis through 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[25,26,27]. 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 true for reprogramming-associated mutations. Nonetheless, these analyses suggest that selection of potentially functional mutations could play a role in amplifying rare mutation-carrying cells and, when coupled with the single-cell bottleneck in hiPS colony picking, could contribute to the fixation of initially low-frequency mutations throughout the entire hiPS cell population.

Discussion

Taken together, our results clearly demonstrate that pre-existing and new mutations during and after reprogramming all contribute to the high mutational load we discovered in hiPS lines. Although we cannot completely rule out the possibility that reprogramming itself is "mutagenic", our data argue that selection during hiPS reprogramming, colony-picking, and subsequent culture 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 accurately identify and characterize reprogramming-associated mutations, 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-derived cells/tissues are used in the clinic. These two aspects are not necessarily connected. Although the functional effects of the 124 mutations remained 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 for diseases other than cancer needs to be evaluated for hiPS-based therapeutic methods. Future long-term studies must focus on functional characterization of reprogramming-associated mutations in order to further aid the creation of clinical safety standards.

Because safe hiPS cells are critical for clinical application, just as previous findings of large-scale genome rearrangements in hiPS lines led to the introduction of karyotyping as a standard post-reprogramming protocol, routine genetic screening of hiPS lines to ensure that no obviously deleterious point mutations are present must become a standard procedure. Complete exome or genome sequencing of hiPS lines might be an efficient way to screen out hiPS lines that have a high mutational load or that have mutations in genes implicated in development, disease, or tumorigenesis. Further rigorous work on mutation rates and distributions during *in vitro* culture 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 4-factor retroviral vectors. PGP1-iPS cells were reprogrammed by Cellular Dynamics using the same four factors in a lentiviral vector from PGP1F fibroblasts29. dH1F-iPS8, dH1F-iPS9, dH1cF16-iPS1, dH1cF16-iPS4, dH1cF16, and dH1F cells were obtained from previous cultures30 reprogrammed with retroviral vectors containing the same factors31. DF-6-9-9, DF-19-11, iPS4.7, and FS cells were obtained from previously existing cultures; the

reprogramming process and characterization of lines has been described previously24. iPS11a, iPS11b, iPS17a, iPS17b, iPS29A, iPS29e, Hib11, Hib17, and Hib29 cells were obtained from previous cultures reprogrammed using retroviral vectors encoding three or four factors32. FiPS3F1 and FiPS4F7 were reprogrammed from HFFxF fibroblasts using similar protocols33-35. FiPS4F2 and FiPS4F-shpRB4.5 were reprogrammed using the same 4-factor protocol from IMR90 fibroblasts. The mRNA-derived lines (CF-RiPS1.4, CF-RiPS1.9, and CF Fibroblasts) were obtained from previous cultures36. All hiPS 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. Putative mutations were rejected 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 one million independent base calls for each location. A binomial test was then used to determine if the observed minor allele frequency could be separated from error and estimate the frequency of each mutation.

Detailed methods are available in the Supplementary Materials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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 on DNA shearing, G.L. Boulting and S. Ratansirintrawoot for assistance on hiPS cell culturing. This study is supported by NIH R01 HL094963 and UCSD new faculty startup 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.G. is supported by the Focht-Powell Fellowship and a CIRM pre-doctoral Fellowship. Y.H.L. is supported by the A*Star Institute of Medical Biology and Singapore stem cell consortium. Work in the laboratory of J.C.I.B. was supported by grants from MICINN, Sanofi-Aventis and the G. Harold and Leila Y. Mathers and Cellex Foundations.

REFERENCES

1. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006; 126:663–76. [PubMed: 16904174]
2. Yu J, et al. Induced pluripotent stem cell lines derived from human somatic cells. Science. 2007; 318:1917–20. [PubMed: 18029452]
3. Mayshar Y, et al. Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. Cell Stem Cell. 2010; 7:521–31. [PubMed: 20887957]
4. Hong H, et al. Suppression of induced pluripotent stem cell generation by the p53-p21 pathway. Nature. 2009; 460:1132–5. [PubMed: 19668191]

5. Li H, et al. The Ink4/Arf locus is a barrier for iPS cell reprogramming. Nature. 2009; 460:1136–9. [PubMed: 19668188]
6. Kawamura T, et al. Linking the p53 tumour suppressor pathway to somatic cell reprogramming. Nature. 2009; 460:1140–4. [PubMed: 19668186]
7. Utikal J, et al. Immortalization eliminates a roadblock during cellular reprogramming into iPS cells. Nature. 2009; 460:1145–8. [PubMed: 19668190]
8. Marion RM, et al. A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity. Nature. 2009; 460:1149–53. [PubMed: 19668189]
9. Ruiz S, et al. A high proliferation rate is required for somatic cell reprogramming and maintenance of human embryonic stem cell identity. Current Biology. (In Press).
10. Porreca GJ, et al. Multiplex amplification of large sets of human exons. Nat Methods. 2007; 4:931–6. [PubMed: 17934468]
11. Deng J, et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol. 2009; 27:353–60. [PubMed: 19330000]
12. Bashiardes S, et al. Direct genomic selection. Nat Methods. 2005; 2:7.
13. Gnirke A, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009; 27:182–9. [PubMed: 19182786]
14. Levy S, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007; 5:e254. [PubMed: 17803354]
15. Drmanac R, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2009; 327:78–81. [PubMed: 19892942]
16. Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–276. [PubMed: 19684571]
17. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4:1073–81. [PubMed: 19561590]
18. Forbes SA, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008 Chapter 10, Unit 10 11.
19. Shah SP, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature. 2009; 461:809–13. [PubMed: 19812674]
20. Futreal PA, et al. A census of human cancer genes. Nat Rev Cancer. 2004; 4:177–83. [PubMed: 14993899]
21. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research. 2005; 33:D514–D517. [PubMed: 15608251]
22. Druley TE, et al. Quantification of rare allelic variants from pooled genomic DNA. Nat Methods. 2009; 6:263–5. [PubMed: 19252504]
23. Ahuja D, Saenz-Robles MT, Pipas JM. SV40 large T antigen targets multiple cellular pathways to elicit cellular transformation. Oncogene. 2005; 24:7729–7745. [PubMed: 16299533]
24. Yu J, et al. Human induced pluripotent stem cells free of vector and transgene sequences. Science. 2009; 324:797–801. [PubMed: 19325077]
25. Pleasance ED, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010; 463:191–6. [PubMed: 20016485]
26. Lee W, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature. 2010; 465:473–7. [PubMed: 20505728]
27. Ding L, et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature. 2010; 464:999–1005. [PubMed: 20393555]
28. Dennis G Jr. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003; 4:P3. [PubMed: 12734009]
29. Lee JH, et al. A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. PLoS Genet. 2009; 5:e1000718. [PubMed: 19911041]
30. Park IH, et al. Reprogramming of human somatic cells to pluripotency with defined factors. Nature. 2008; 451:141–6. [PubMed: 18157115]

31. Chan EM, et al. Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. Nat Biotechnol. 2009; 27:1033–7. [PubMed: 19826408]
32. Dimos JT, et al. Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. Science. 2008; 321:1218–21. [PubMed: 18669821]
33. Rodriguez-Piza I, et al. Reprogramming of human fibroblasts to induced pluripotent stem cells under xeno-free conditions. Stem Cells. 2010; 28:36–44. [PubMed: 19890879]
34. Aasen T, et al. Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. Nat Biotechnol. 2008; 26:1276–84. [PubMed: 18931654]
35. Stewart SA, et al. Lentivirus-delivered stable gene silencing by RNAi in primary cells. RNA. 2003; 9:493–501. [PubMed: 12649500]
36. Warren L, et al. Highly Efficient Reprogramming to Pluripotency and Directed Differentiation of Human Cells with Synthetic Modified mRNA. Cell stem cell. 2010
37. Akagi T, Sasai K, Hanafusa H. Refractory nature of normal human diploid fibroblasts with respect to oncogene-mediated transformation. Proc Natl Acad Sci U S A. 2003; 100:13567–72. [PubMed: 14597713]
38. Cowan CA, et al. Derivation of embryonic stem-cell lines from human blastocysts. N Engl J Med. 2004; 350:1353–6. [PubMed: 14999088]
39. Rodríguez-Pizà I, et al. Reprogramming of Human Fibroblasts to Induced Pluripotent Stem Cells under Xeno-free Conditions. STEM CELLS. 2010; 28:36–44. [PubMed: 19890879]
40. Aasen T, et al. Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. Nat Biotech. 2008; 26:1276–1284.
41. Zhang K, et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. Nat Methods. 2009; 6:613–8. [PubMed: 19620972]
42. Kishore, S. Meena; Vincent, TKC.; Pandjassarame, K. Distributions of Exons and Introns in the Human Genome. In Silico Biology. 2004; 4:387–393. [PubMed: 15217358]

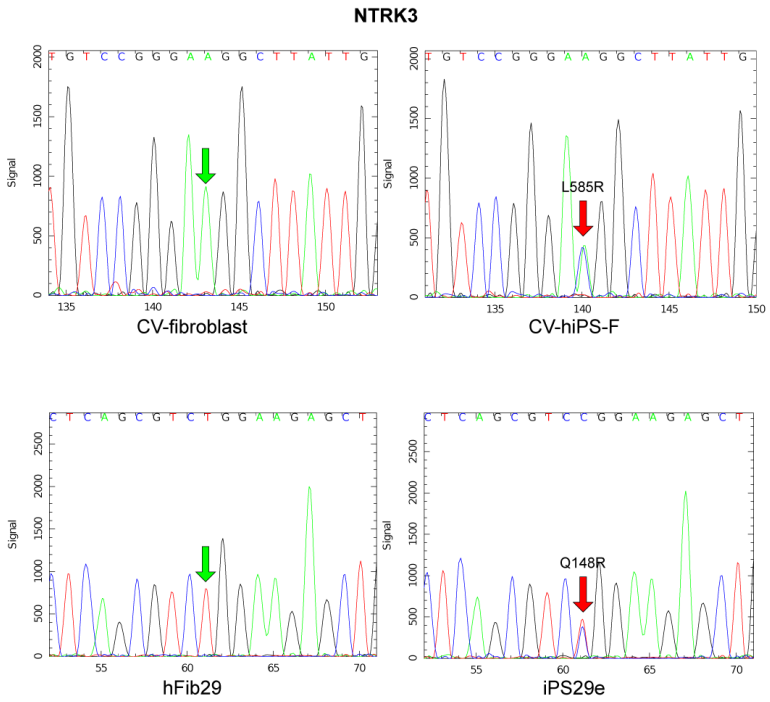

Figure 1. hiPS acquired protein-coding somatic mutations

Somatic mutations in the gene *NTRK3* were found in two independent hiPS lines but were not present in their fibroblast progenitors. Detailed information for all mutations is in the Supplementary Materials.

Table 1

Sequencing statistics for mutation discovery

Quality filtered sequence represents the total amount of sequencer data generated that passed the Illumina GA IIx quality filter. Number of high quality coding variants is the number of variants found with sequencing depth of at least 8 and consensus quality score of at least 30. dbSNP percentage represents the percent of identified variants present in the dbSNP database. Shared coding region is the portion of the genome, in base pairs, that was sequenced at high depth and quality in both the iPS 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 CCDS variants (out of ~17,000 expected variants)16 successfully identified in both hiPS and Fibroblast.

Cell Line	Exome Capture Method	Quality Filtered Sequence (bps)	# of High-Quality Coding Variants	dbSNP Percentage	Shared High-Quality Coding Region (bps)	# Coding Mutations – Observed (Projected)
CV-hiPS-F	Padlock+SeqCapEZ	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+SeqCapEZ	7,586,731,600	15,442	98%		
DF-6-9-9	Padlock+SeqCapEZ*	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+SeqCapEZ*	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)

Cell Line	Exome Capture Method	Quality Filtered Sequence (bps)	# of High-Quality Coding Variants	dbSNP Percentage	Shared High-Quality Coding Region (bps)	# Coding Mutations – Observed (Projected)
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	1 (2)
CF-RiPS1.9	SeqCap EZ	3,143,591,760	13,674	95%	21,165,013	3 (4)
CF Fibroblast	SeqCap EZ	3,204,874,880	11,855	96%		
FiPS3F1	SeqCap EZ	3,397,397,360	13,333	94%	20,723,620	3 (4)
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	4 (4)
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%		

*For DF-6-9-9 and FS, 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 vs. unbiased coverage); it appears these factors directly affect the ability to find separate mutations.

Table 2
List of genes found to be mutated in coding regions in hiPS cells

The full details of each mutation are in Supplementary Table 1.

Cell Line	Mutated Genes	Number of Non-Silent Mutations	Detectable at Low Frequency in Fibroblasts?
CF-RiPS1.4	*OR52E8, TEAD4*	1	N/A
CF-RiPS1.9	*OR52E8, FAM171A1, TMED9, TEAD4, RASEF*	3	N/A
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, RP4-788L13.1, KCNJ8, SETBP1, ZNF471, TMEM40*	5	N/A
DF6-9-9	*ZZZ3, AKR1C4, NEK5, DAPL1, ITCH, PPP1R2*	5	0/5
dH1CF16-iPS1	*IRGQ, TM9SF4*	1	N/A
dH1CF16-iPS4	*PKP1, MYOG, ABCA3, PTPRM, RANBP3L, CALN1*	4	N/A
dH1F-iPS8	*CABC1, C1orf100, OR5AN1, CACNG3, MYRIP, SLC1A3, DSP, KLRG2*	6	N/A
dH1F-iPS9	*SEMA6C, MYRIP, SLC1A3*	3	N/A
FiPS3F1	*SORCS3, GLRA3, CARM1, EPB41L1*	2	N/A
FiPS4F7	*GDF3, ZER1*	2	N/A
iPS11A	*GTF3C1, SAL1, SLC26A3, ZNF16*	3	1/1
iPS11B	*MARCKSL1, PRDM16, ATM, LRP4, TCF12, SH3PX3, OSBPL3*	5	0/1
iPS17A	*HK1, ANKRD12, SCN1A, IFNGR1*	4	N/A
iPS17B	*HK1, CCKBR, ANKRD12, SCN1A, IFT122*	5	1/1
iPS29A	*PRICKLE1, RFX6*	2	2/2
iPS29E	*C14orf174, NTRK3, VAC14, ASB3, STX7, POLR1C, LINGO2*	6	1/4
iPS4.7	*POLE, UBA2, L3MBTL2, C4orf41*	2	N/A
PGP1-iPS	*C11orf67, OSBPL8, NEK11*	1	1/3
FiPS4F2	*TMEM57, RANBP6, CTSL1, SAV1, KRT25, BCL2L12, LGALS1, TTYH2**, *COPA**, *ARSB**, *MT1B**	7	N/A
FiPS4F-shpRB4.5	*NTRK1, CD1B, LRCH3, SH3TC1, GPC2, CDK5RAP2, MYH4, TRMU*	5	N/A

[*] Mutation was observed at passage 40 but not at passage 9. FiPS4F2 was sequenced at both passage 9 and passage 40. Six mutations were present after reprogramming (FiPS4F2P9), while four more became fixed after extended culture (FiPS4F2P40). All six mutations found after reprogramming were also present after extended culture.