

ARTICLE

Received 6 Feb 2012 | Accepted 13 Dec 2012 | Published 22 Jan 2013

DOI: 10.1038/ncomms2381

Analysis of protein-coding mutations in hiPSCs and their possible role during somatic cell reprogramming

Sergio Ruiz^{1,*}, Athurva Gore^{2,*}, Zhe Li², Athanasia D. Panopoulos¹, Nuria Montserrat³, Ho-Lim Fung², Alessandra Giorgetti³, Josipa Bilic³, Erika M. Batchelder¹, Holm Zaehres⁴, Hans R. Schöler⁴, Kun Zhang² & Juan Carlos Izpisua Belmonte^{1,3}

Recent studies indicate that human-induced pluripotent stem cells contain genomic structural variations and point mutations in coding regions. However, these studies have focused on fibroblast-derived human induced pluripotent stem cells, and it is currently unknown whether the use of alternative somatic cell sources with varying reprogramming efficiencies would result in different levels of genetic alterations. Here we characterize the genomic integrity of eight human induced pluripotent stem cell lines derived from five different non-fibroblast somatic cell types. We show that protein-coding mutations are a general feature of the human induced pluripotent stem cell state and are independent of somatic cell source. Furthermore, we analyse a total of 17 point mutations found in human induced pluripotent stem cells and demonstrate that they do not generally facilitate the acquisition of pluripotency and thus are not likely to provide a selective advantage for reprogramming.

¹ Gene Expression Laboratory, Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. ² Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. ³ Center of Regenerative Medicine in Barcelona, Doctor Aiguader, 88, Barcelona 08003, Spain. ⁴ Department of Cell and Developmental Biology, Max Planck Institute for Molecular Biomedicine, Röntgenstrasse 20, 48149 Münster, Germany. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to K.Z. (email: kzhang@bioeng.ucsd.edu) or to J.C.I.B. (email: belmonte@salk.edu or izpisua@cmrb.eu).

he induction of pluripotency in human somatic cells by defined transcription factors represents a breakthrough in regenerative medicine^{1–5}. The generation of patient-specific human induced pluripotent stem cells (hiPSCs) and their autologous cell derivatives would help to overcome the problems of immune rejection and tissue availability. However, the applications of cell therapies in human patients are subject to very stringent safety requirements, and there is a general concern in the field about the safety of hiPSCs.

Successful generation of hiPSCs depends on the complete reprogramming of the somatic epigenome to a pluripotent state while the genome remains unchanged. Although initial reports demonstrated that human embryonic stem cells (hESCs) and hiPSCs were very similar, recent reports have uncovered striking genetic and epigenetic differences between these two pluripotent cell types⁶⁻¹¹. It has been shown that hiPSCs display proteincoding mutations, large-scale genomic rearrangements, persistent epigenetic marks from the somatic cell type of origin and aberrant methylation patterns^{6,9,11}. These findings indicated that hiPSCs contain genomic defects that could preclude their use in stem cell therapies. However, most of these studies focused on fibroblastderived hiPSCs, and a more comprehensive analysis is essential to determine whether there are specific somatic cell types that may reprogram into hiPSCs with fewer (or perhaps none) of these aberrations. Additionally, it is unclear whether the protein-coding mutations found in hiPSCs provide any functional advantage and, thus, are selected for during the process of reprogramming.

In this work, we characterize at single-nucleotide resolution the genomic integrity of eight hiPSC lines derived from five different non-fibroblast somatic cell types with varied reprogramming efficiencies. Moreover, we functionally characterize the role of 17 point mutations found in hiPSCs for their ability to increase reprogramming efficiency. We demonstrate that the majority of these mutations do not favour the reprogramming process and suggest that most of them originated randomly or were initially present in the somatic population of origin. Our observations of the genetic abnormalities of hiPSCs will contribute to a deeper understanding of the reprogramming process.

Results

hiPSC lines from varied cell types contain protein-coding mutations. We previously sequenced the protein-coding regions of 22 fibroblast-derived hiPSC lines and discovered that the hiPSCs analysed carried between 2 and 14 point mutations in protein-coding regions⁶. In this study, we sought to determine whether low reprogramming efficiency (and therefore a potentially higher level of selection pressure that could allow the fixation of advantageous mutations) or cell type of origin (as fibroblasts could possess a higher somatic mutation rate than other cell types) could contribute to the overall reprogrammingassociated mutational load. To this end, we performed targeted exome sequencing on eight non-fibroblast-derived hiPSC lines and their five somatic cell types of origin using an in-solution hybridization capture method (Supplementary Table S1). Somatic mutations in each hiPSC line were identified via pairwise comparison with the matched somatic cell of origin and independently confirmed with capillary Sanger sequencing. We identified a total of 40 point mutations throughout all the hiPSC lines analysed, leading to an average of five coding mutations per line (Table 1). As we identified \sim 89% of expected total singlenucleotide polymorphisms at high sequencing depth in proteincoding regions, this led to a projection of 45 total mutations in protein-coding regions, or \sim 6 coding mutations per cell line. The levels of mutational load from each individual somatic cell type were statistically indistinguishable, and within the range

previously observed for fibroblast-derived hiPSC lines⁶ (Table 1). These results indicate that hiPSC-associated mutations cannot be avoided by using younger or potentially more genetically protected somatic cell sources as progenitor cells. Moreover, we determined that reprogramming efficiency, which varies between 0.001 to 3% for these cell types, did not seem to have a measurable effect on the hiPSC mutational load. Thus, reprogramming-associated point mutations appear to be a general feature of hiPSCs.

We next investigated whether mutations in hiPSCs were either enriched or depleted in protein-coding regions. To this end, we examined additional non-coding regions captured in our sequencing analysis, and found a similar mutation rate per base pair analysed for both coding and non-coding regions (Table 2). We also investigated whether point mutations in hiPSCs tended to occur in active/ubiquitous or silent/tissue-specific genes. Among a total of 132 mutated genes (from this study and ref. 6) annotated in the TiGER Database¹², 37% of these genes showed tissue-specific expression, which is very similar to the overall level of tissue specificity observed in the genes annotated in the database (34%; P = 0.4975), indicating that mutations are not preferentially occurring in silent genes. We additionally checked for any potential enrichment of mutations in active or inactive transcriptional regions of the genome¹³. We found that mutations were not significantly enriched in the active or inactive chromatin regions of fibroblasts (P = 0.79), hESCs (P = 0.29) or hiPSCs (P = 0.07). Furthermore, only one gene (NTRK3) was found mutated in more than one independent hiPSC line, and mutated genes did not cluster in a specific functional pathway (ref. 6 and results herein). These combined findings suggest that mutations in hiPSCs are spread throughout both transcriptionally active and silent regions of the genome.

hiPSC-point mutations do not favour the process of reprogramming. We previously showed that at least half of reprogramming-associated point mutations pre-exist in starting somatic cell populations at low frequency⁶. This leads to a hypothesis that a sub-population of somatic cells carrying certain mutations could be primed for reprogramming, which would be consistent with the elite model for reprogramming¹⁴. To investigate the functional potential of these mutations during reprogramming, we first assessed whether mutated alleles were expressed in the hiPSC lines. We isolated RNA from three hiPSC lines, reverse-transcribed it into cDNA and sequenced a total of six transcripts of randomly selected genes found mutated in these hiPSC lines. We detected heterozygous expression of both mutant (mut) and wild-type(wt) alleles in all cases (Fig. 1), indicating that mutated transcripts are expressed in hiPSCs.

We next sought to determine whether reprogrammingassociated mutations could contribute functionally in facilitating the acquisition of pluripotency during reprogramming. From a total of 164 different genes found mutated in hiPSC lines (ref. 6 and this study), we assayed the function of 17 candidate genes and their mutated forms during reprogramming (Supplementary Table S2). These candidate genes were selected based on the likelihood of the mutation to change protein function, the mutation type (only non-synonymous mutations were analysed) and whether the gene was known to be related to the maintenance and/or acquisition of pluripotency⁶ (Table 1; Supplementary Fig. S1; Supplementary Table S2). We also analysed the expression of these 17 genes in BJ fibroblasts, human umbilical vein endothelial cells (HUVEC), hESC and hiPSC lines to ensure gene expression in at least one of the somatic cell types used in this work (Supplementary Fig. S2). Owing to the difficulty in predicting the functional consequences of each specific mutation, we first

Sample	Chromosome	Position	Gene	Alleles	Protein change	Mutation type	SIFT functional prediction	Mutated in cancer?
ASThiPS4F4	6	31783527	LY6G6F	GAC-GAt	D122D	Synonymous	NA	Yes
ASThiPS4F4	8	68087821	LRRC67	CTT-aTT	L121I	Non-synonymous	TOLERATED	No
ASThiPS4F5	11	54891946	OR4A15	CTG-CcG	L4P	Non-synonymous	DAMAGING	No
ASThiPS4F5	13	110343392	ANKRD10	AAG-AAt	K225N	Non-synonymous	TOLERATED	Yes
ChiPS4F8	1	205153901	FAIM3	TTC-aTC	F67I	Non-synonymous	DAMAGING	Yes
KhiPS4F8	5	121215932	FTMT	CAC-CAt	H125H	Synonymous	NA	Yes
ChiPS4F8	14	62486817	KCNH5	GAC-aAC	D386N	Non-synonymous	TOLERATED	No
NSChiPS2F	5	79774746	ZFYVE16	TCT-TaT	S823Y	Non-synonymous	DAMAGING	No
NSChiPS2F	12	54853783	SMARCC2	CCA-CCg	P538P	Synonymous	NA	Yes
HUVhiPS4F1	2	169809670	LRP2	TCG-TtG	S1070L	Non-synonymous	TOLERATED	Yes
HUVhiPS4F1	10	33542444	NRP1	GGC-GaC	G497D	Non-synonymous	DAMAGING	No
HUVhiPS4F1	16	17139792	XYLT1	AAG-AgG	K562R	Non-synonymous	TOLERATED	No
HUVhiPS4F1	4	155376303	DCHS2	GGA-GtA	G2529V	Non-synonymous	DAMAGING	No
HUVhiPS4F3	6	155183150	RBM16	GTA-cTA	V595L	Non-synonymous	TOLERATED	No
HUVhiPS4F3	9	394921	DOCK8	TCA-aCA	S1012T	Non-synonymous	TOLERATED	Yes
HUVhiPS4F3	X	109889590	CHRDL1	CTT-aTT	L86I	Non-synonymous	TOLERATED	No
HUVhiPS4F3	19	7475243	C19orf45	TCA-TaA	S229*	Nonsense	NA	Yes
MSChiPS4F4	1	35998698	CLSPN	GTG-tTG	V471L	Non-synonymous	TOLERATED	Yes
MSChiPS4F4	1	153185686	PBXIP1	GAC-GgC	D363G	Non-synonymous	TOLERATED	Yes
MSChiPS4F4	2	154960801	GALNT13	GAA-GAg	E403E	Synonymous	NA	Yes
MSChiPS4F4	5	126704124	MEGF10	GTC-GTg	V74V	Synonymous	NA	Yes
MSChiPS4F4	6	130572400	SAMD3	ATG-tTG	M106L	Non-synonymous	DAMAGING	Yes
MSChiPS4F4	10	42974297	CSGALNACT2		M264V	Non-synonymous	TOLERATED	Yes
MSChiPS4F4	11	36473107	TRAF6	GAA-aAA	E225K	Non-synonymous	DAMAGING	Yes
MSChiPS4F4	17	50475673	STXBP4	GTA-GTg	V236V	Synonymous	NA	Yes
MSChiPS4F4	19	40938581	HSPB6	TCGCCG-	S84S	Synonymous	NA	No
viseim s ii i	17	10730301	1131 50	TCatCG	P85S	Non-synonymous	DAMAGING	110
MSChiPS4F4	20	46706999	PREX1	GCC-GCt	A703A	Synonymous	NA	Yes
MSChiPS4F4	21	39493296	BRWD1	AAA-AtA	K1639I	Non-synonymous	DAMAGING	Yes
MSChiPS4F4		11688927	MSL3	TCT-TtT	S111F	Non-synonymous	DAMAGING	Yes
MSChiPS4F8	1	39703363	MACF1	GGC-tGC	G5698C	Non-synonymous	DAMAGING	Yes
MSChiPS4F8	1	158594563	NCSTN	TTG-cTG	L670L	Synonymous	NA	No
MSChiPS4F8	1	231873806	KCNK1	GAC-GAt	D224D	Synonymous	NA	Yes
MSChiPS4F8	2	21087987	APOB	CAC-CgC	H1753R	Non-synonymous	DAMAGING	Yes
MSChiPS4F8	2	234287122	UGT1A8	GTC-GaC	V249D	Non-synonymous	NA	Yes
MSChiPS4F8	5	128390915	SLC27A6	GAC-GAa	D482E	Non-synonymous	DAMAGING	No
MSChiPS4F8	8	16079769	MSR1	CCG-tCG	P34S	Non-synonymous	TOLERATED	Yes
MSChiPS4F8	9	125834763	LHX2	GAG-tAG	E393*	Non-synonymous	NA	No
MSChiPS4F8	14	46496570	MDGA2	TTG-aTG	L318M	Non-synonymous	TOLERATED	Yes
VISCHIPS4F8 MSChiPS4F8	15	37669438	THBS1	TGC-TGt	C689C	Synonymous	NA	Yes
MSChiPS4F8	15 X	152498688	ATP2B3	TCC-TaC	S1134Y	Non-synonymous	DAMAGING	Yes

*Stop codon.

performed 'loss-of-function' reprogramming experiments to mimic a possible diminished activity or protein instability of the mutated form. To this end, we designed a panel of lentiviruses encoding short hairpin RNAs (shRNAs) against the selected genes (Supplementary Fig. S3a), and coinfected each separately with retroviruses expressing OCT4, SOX2, KLF4 and cMyc (OSKC) in BJ fibroblasts (Fig. 2a). Moreover, to determine whether these effects were cell-type specific, we performed similar reprogramming experiments in HUVEC (Supplementary Fig. S4a). If a genetic mutation was selected for its ability to facilitate reprogramming due to a loss of protein function, it would be expected that downregulation of the mutated gene would increase reprogramming efficiency. A decrease in reprogramming efficiency was detected after downregulation of FAIM3, SAMD3, ZNF16, MARCKSL1, NRP1, TRAF6, GSG1 and HK1, whereas no significant changes were detected after the downregulation of all but one of the assayed genes, POLR1C (Fig. 2a, Supplementary Fig. 4a, Supplementary Fig. S4b). Interestingly, we observed that downregulation of POLR1C in BJ fibroblasts, but not in HUVEC, resulted in an increased

reprogramming efficiency. However, it is unclear whether the specific reprogramming-associated mutation in POLR1C would result in the same phenotype. Overall, our data suggest that protein-coding point mutations generally do not prime rare cells for reprogramming through the loss-of-function mechanism.

Next, we performed 'gain-of-function' reprogramming experiments to determine whether expression of the mutated form facilitated cell reprogramming. To this end, we designed a panel of retroviruses encoding both the wt form and the corresponding mutated form found in hiPSCs of each specific gene (see specific mutations in Supplementary Table S2; Supplementary Fig. S3b), and coexpressed them with OSKC in BJ fibroblasts and HUVECs (Fig. 2b, Supplementary Fig. S4c). If a mutation were selected during reprogramming due to a gain-of-function, it would be expected that expression of the mutated form would increase the reprogramming efficiency. We observed that only the expression of *HK1* slightly increased reprogramming efficiency (Fig. 2b and Supplementary Fig. S4c). Importantly, we did not observe significant differences in reprogramming efficiency between cells overexpressing the

Table 2 List of lines.	candidate	non-coding	mutations	in hiPSC

Cell line	Non-co	oding mutatio	Exon mutation rate (per bp)	Non-exon mutation rate (per bp)	
ASThiPS4F4	9 11 13	Position 111225067 64089233 38444609	$\begin{array}{c} \textbf{Mutation} \\ \textbf{C} & -> \textbf{T} \\ \textbf{G} & -> \textbf{T} \end{array}$	8.0E – 08	6.2E – 08
ASThiPS4F5	2 12 16 17		$\begin{array}{ccc} G & - > T \\ G & - > T \end{array}$	8.0E – 08	1.0E — 07
FiPS3F1	5 9 10 11 19	149190453 5175241 45274877 85134161 48465587	$\begin{array}{c} C \ -> T \\ G \ -> T \\ T \ -> C \end{array}$	1.6E – 07	1.6E – 07
FiPS4F7	1 2 2 9	171784008 116251932 189575154 98839743	C -> A C -> A	1.2E — 07	1.3E — 07
HUVhiPS4F1	11 19		$\begin{array}{l} G \; - > A \\ G \; - > A \end{array}$	1.6E — 07	7.2E – 08
HUVhiPS4F3	2 11 15 22	25902050	G -> T	1.6E — 07	1.1E — 07
KhiPS4F8	5 9 10 17	122778753 85962168		1.2E — 07	1.8E — 07
NSChiPS2F	5 9 9 11 19	110120401 115077723 127398270 9457743 46045755	$\begin{array}{c} C & - > T \\ C & - > T \\ T & - > A \end{array}$	8.0E – 08	1.2E — 07
MSChiPS4F4	1 2 2 3 4 4 5 6 6 8 10 10 11 15 18 X	88885962 230820856 51088015 67994660 156930655 156683693 73887041 129865735 24379410 94807938 100179495 104534424 61826051	$\begin{array}{l} T \; - > C \\ G \; - > A \\ G \; - > A \\ G \; - > A \\ C \; - > T \\ T \; - > C \end{array}$	4.8E – 07	4.2E – 07
MSChiPS4F8	1 1 1 2 2 3 4 4 4 4 6 6 6 7 7 7 9 9 14 14 16 16 17 17 18 19 19 19 19 19 19 19 19 19 19 19 19 19	85589993 88885962 128062632 12183849	$\begin{array}{llllllllllllllllllllllllllllllllllll$	4.4E – 07	6.5E — 07
	X	1304/0822	∪ -> A	1.9E — 07	2.0E — 07

The mutation rate per base pair was similar for exonic and non-exonic regions.

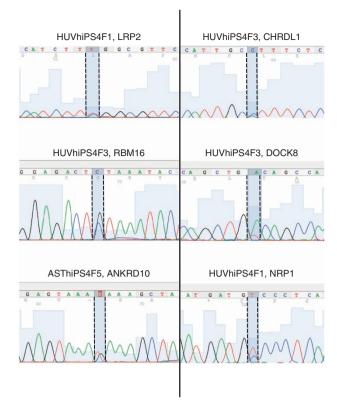


Figure 1 | Mutated alleles are expressed in hiPSC lines. Sanger chromatograms showing the results of RNA Sequencing analysis performed on the indicated genes found mutated in the indicated hiPSC lines. Dashed lines highlight the point-mutated nucleotide. Note the expression of both reference and mutated alleles in all cases analyzed.

mutated forms and cells overexpressing their respective wt forms (Fig. 2b), indicating that the presence of the mutated protein does not increase reprogramming efficiency.

We have previously shown that both the mut allele and the wt allele are expressed in hiPSCs (Fig. 1). However, it is possible that a similar level of expression of the wt and mut protein forms is necessary in order for the mutation to influence reprogramming efficiency in a gain-of-function manner. To clarify this, we performed a reprogramming experiment where OSKC were coexpressed together with a similar total amount of retrovirus encoding either only the wild type form or both the wt and mut forms of a mutated gene in an equal ratio (1:1). Using this strategy, we were able to compare the reprogramming efficiency of cells overexpressing wt and mutated protein (wt/mut) in equal amounts with that of cells overexpressing wt protein alone (wt/wt). Interestingly, we did not observe any difference in reprogramming efficiency between cells overexpressing the wt/wt and wt/mut proteins (Fig. 3a). Finally, we investigated whether silencing of retroviral transgenes during reprogramming could mask a gain-of-function effect of the mutated genes at a later stage of reprogramming. We thus analysed the reprogramming efficiency of cells infected with retroviruses expressing OSKC, the wt or mutated forms of the genes evaluated in this study, and a red fluorescent protein (RFP) reporter gene to monitor transgene silencing. Reprogramming efficiency was evaluated based on the number of Tra-1-60 + /RFP + colonies present at day 14. These colonies represent putative bona-fide hiPSC colonies, as they express the stem cell marker Tra-1-60 but lack silencing of the exogenous transgenes. Thus, we only considered reprogramming events where transgene expression was still active. Importantly, we did not observe differences in reprogramming efficiency

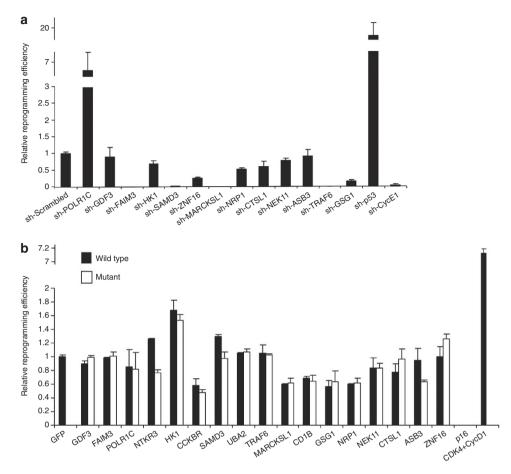


Figure 2 | Evaluation of the functional effect of hiPSC mutations on reprogramming efficiency. (**a,b**) Human BJ fibroblasts were infected with retroviruses encoding OSKC, and either lentiviruses encoding shRNAs against the indicated proteins (**a**) or retroviruses encoding the wild type or mutated proteins (**b**). Relative reprogramming efficiencies (evaluated as percentage of Nanog ⁺ colonies) are shown as fold change normalized to the averaged efficiency observed in (**a**) pLVTHM or (**b**) pMX-GFP-infected fibroblasts. In **a**, lentiviruses encoding shRNAs against *CycE1* or *p53* were used as controls of reduced or increased reprogramming efficiency, respectively. In **b**, retroviruses encoding *p16* or the pair CDK4/CycD1 were used as controls of reduced or increased reprogramming efficiency, respectively. For **a**, 20,000 infected cells were plated when shRNAs against *POLR1C* and *p53* were used, and 70,000 infected cells were plated under all other conditions. For **b**, a total of 25,000 infected cells were plated under all conditions. Two independent experiments with two biological replicates were carried out. All error bars depict the s.d.

between cells overexpressing the mutated forms and cells overexpressing their respective wt forms (Fig. 3b). Furthermore, we also evaluated reprogramming efficiency in the same experiment at day 14 by analysing the number of Tra-1-60 $^+/$ RFP $^-$ colonies (evaluating putative bona-fide hiPSC colonies where transgene silencing occurred), and obtained a similar result (data not shown). Overall, these data suggest that most of these mutated genes do not facilitate reprogramming through a gain-of-function or loss-of-function mechanism.

Discussion

Our work demonstrates that hiPSCs contain protein-coding mutations independent of the cell type of origin (as we analysed hiPSC lines derived from five tissue types). Moreover, we determined that reprogramming efficiency, and therefore the level of selection pressure which could allow the fixation of advantageous mutations, did not to have a measurable effect on the hiPSC mutational load. Although the functional consequences of individual protein-coding mutations detected in hiPSCs remain to be characterized, these alterations could potentially contribute to the functional differences observed between hiPSC lines^{15–17}.

Two independent groups have recently reported the whole-genome sequencing of human and murine iPSC lines and their corresponding somatic cell lines^{18,19}. They identified hundreds of single-nucleotide variants (SNVs) in non-coding regions and an average of 6–12 SNVs in coding regions^{18,19}, which is consistent with our results⁶. Importantly, their data suggest that much of the genetic variation in iPSC clones pre-exists in the somatic population of origin and is fixed as a consequence of cloning individual cells during iPSC generation^{18,19}. Although these reports supported previous observations⁶, they did not investigate whether identified mutations contribute functionally to facilitate the acquisition of pluripotency during reprogramming.

In this work, we show evidence suggesting that most reprogramming-associated point mutations do not provide a detectible selective advantage towards a reprogrammed state. As inhibiting wt POLR1C expression had a positive impact on reprogramming efficiency, we cannot rule out a potential role of the mutation found in POLR1C in facilitating reprogramming. If this is the case, the fact that downregulation of POLR1C increases reprogramming efficiency in fibroblasts, but not in HUVECs, could indicate the existence of tissue-specific mutations affecting reprogramming efficiency, as POLR1C P^{278R} was found in one hiPSC line derived from human fibroblasts. Although it remains

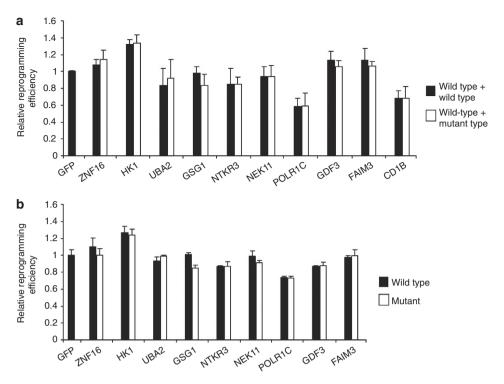


Figure 3 | Retroviral silencing or wild-type/mutant gene ratio do not alter reprogramming efficiency. (a) HUVEC cells were infected with retroviruses encoding OSKC, and a similar total amount of retroviruses encoding only the wild-type form or both, the wild-type and mutant forms of the protein in an equal proportion. (b) HUVEC cells were infected with retroviruses encoding OSKC, RFP and the wild-type or mutated forms of the genes indicated. Relative reprogramming efficiencies (evaluated as percentage of Tra-1-60 + colonies) are shown as fold change normalized to the averaged efficiency observed in green fluorescent protein-infected HUVECs. Ten thousand infected cells were plated under all the conditions. Two independent experiments with two biological replicates were carried out. All error bars depict the s.d.

possible that untested mutated genes or a combination of mutations in a certain cellular context could have a role, the findings that only one gene (NTRK3) was found mutated in 2 out of 30 independent hiPSC lines, that mutated genes do not cluster in a specific functional pathway that could explain their selection during the reprogramming process, and that non-coding regions showed a similar mutational load, indicate that reprogrammingassociated mutations seem to occur through a random process without selection and/or are initially present in the somatic population of origin^{18,19}. It has been suggested that genomic alterations (that is, duplications, deletions and mutations) are selected for during reprogramming, yet this has not been demonstrated⁶⁻¹¹. In contrast to well-established recurrent genomic aberrations (for example, chromosome 12 duplications) present in hESC or hiPSC lines that are functionally selected upon prolonged culture⁸, our results suggest that reprogrammingpoint mutations generally do not reprogramming efficiency although there could be exceptions. To our knowledge, the data provided herein provides for the first time a functional analysis of the role of specific genomic alterations (that is, point mutations in coding regions) on the reprogramming process and have potential implications for the future of the hiPSC field in regenerative medicine.

Methods

Cell culture. The hiPSC lines ASThiPS4F4, ASThiPS4F5, HUVhiPS4F1, HUV-hiPS4F3, FhiPS4F7, NSChiPS2F and FhiPS3F1 were already described^{6,20–22}, and obtained from existing cultures. The hiPSC lines MSChiPS4F4, MSChiPS4F8 and KhiPS4F8 show all the requirements (morphology, pluripotent gene expression normal karyotype and *in vivo* differentiation by teratoma formation) to define them as hiPSC cell lines. Derived hiPSCs were cultured as described²³. 293T cells and BJ human fibroblasts (ATCC, CRL-2522) were cultured in DMEM (Invitrogen) supplemented with 10% FBS and 0.1 mM non-essential amino acids.

HUVEC cells were obtained from Lonza (C-2519A) and grown with EGM-2 media (Lonza) as recommended. MSCs were kindly provided by Cécile Volle (Sanofi-Aventis) and grown in α -MEM (Invitrogen) containing 10% FBS (Hyclone), penicillin/streptomycin, sodium pyruvate, non-essential amino acids, and L-glutamine (all from Invitrogen). Human keratinocytes were obtained and cultured as previously described 24 .

hiPSC generation. To generate hiPSCs (KhiPS4F8, MSChiPS4F4 and MSChiPS4) or to evaluate reprogramming efficiency, experiments were performed as described with minor modifications²³. Briefly, BJ fibroblasts, keratinocytes, MSCs or HUVEC cells were infected with an equal ratio of retroviruses or retroviruses plus lentiviruses by spinfection of the cells at 1850 r.p.m. for 1 h at room temperature in the presence of polybrene (4 μ g ml⁻¹). After one (in case of the HUVEC cells), two (in case of the BJs fibroblasts or keratinocytes) or three (in case of the MSCs) viral infections viral infections, cells were trypsinized and transferred onto fresh irradiated mouse embryonic or human fibroblasts where correspond. One day after, cells were switched to hES cell medium (DMEM/F12 or KO-DMEM (Invitrogen) supplemented with 20% knockout serum replacement (Invitrogen), 1 mM L-glutamine, 0.1 mM non-essential amino acids, 55 μM β-mercaptoethanol and 10 ng ml - 1 bFGF (Joint Protein Central)). Depending on the cell type of origin, colonies were stained for Nanog expression at day 18 (in the case of HUVEC-derived hiPS cells) or 24 (in the case of BJ fibroblasts-derived hiPS cells) or isolated to establish cell lines. To calculate the efficiency of reprogramming, we plated the same number of infected HUVEC or BJ fibroblasts cells on irradiated mouse embryonic fibroblasts after the infection and the relative percentage of Nanog + colonies to the value of the number of colonies generated with HUVEC or BJ fibroblasts cells infected with pLVTHM lentiviruses or green fluorescent protein-expressing retroviruses correspondingly is shown.

Plasmid construction. The reprogramming plasmids pMX-OCT4, pMX-SOX2, pMX-KLF4, pMX-cMyc together with pLVTHM were obtained from Addgene (plasmids 17217, 17218, 17219, 17220 and 12247, respectively). For the construction of pMX-NTRK3, pMX-FAIM3, pMX-POLR1C, pMX-GDF3 and pMX-HK1 (fragment corresponding to the nucleotides 277-2753), specific coding region sequences were amplified by PCR from Human ORFeome library plasmids containing the corresponding cDNAs. cDNA fragments were digested with adequate restriction enzymes, purified and subcloned into linearized pMX plasmid. For the construction of pMX-CCKBR, pMX-SAMD3, pMX-UBA2, pMX-TRAF6,

pMX-MARCKSL1, pMX-CD1B, pMX-GSG1, pMX-NRP1, pMX-NEK11, pMX-CTSL1, pMX-ASB3 and pMX-ZNF16, specific pDONR223 plasmids from Human ORFeome library containing the corresponding cDNAs were used to transfer the cDNAs to the vector pMX-GW (Addgene, 18656). The transfer was achieved by using the Gateway LR Clonase enzyme mix (Invitrogen). The plasmids pMX-p16, pMX-CDK4, pMX-CycD1, pLVTHM-CycE and pLVTHM-p53 were generated as described^{23,25}. The plasmid pMX-RFP was kindly provided by Dr Guanghui Liu (Gene Expression Laboratory, The SALK Insitute, La Jolla, CA). For the introduction of specific point mutations in the coding sequences of the above genes (see Supplementary Table S2 for specific mutations) the QuickChange Site-Directed Mutagenesis kit was used (Stratagene; see Supplementary Table S3 for specific primers). For the generation of plasmids encoding shRNAs against the genes used in this study, specific oligos (see Supplementary Table S3 for specific primers) were annealed, phosphorylated with T4 kinase and ligated into MluI/ClaIlinearized pLTVHM plasmid. The design of three different pairs of shRNAs was carried out using the SFold software (http://sfold.wadsworth.org/), and knockdown efficiency was assayed in 293T cells. The most efficient pairs of shRNAs were assayed in HUVEC or BJ fibroblasts cell (Supplementary Fig. S1a) and used in the corresponding experiments. All constructs generated were subjected to direct sequencing to rule out the presence of mutations.

Retroviral and lentiviral production. Moloney-based retroviral vectors (pMX and derived) and second-generation lentiviral vectors (pLVTHM and derived) were cotransfected with packaging plasmids to generate viral particles in 293T cells using Lipofectamine (Invitrogen) as previously described²³.

Immunostaining. Imnunofluorescence analysis for the detection of pluripotent markers in hiPSCs or for the detection of differentiation-associated markers in teratomas were performed as described²². Immunohistochemical/immunoflorescence detection of Nanog or Tra-1-60 was performed as described²³.

RNA isolation and real-time PCR analysis. Total RNA was isolated using Trizol Reagent (Invitrogen) according to the manufacturer's recommendations. cDNA was synthesized using the SuperScript II Reverse Transcriptase kit for RT-PCR (Invitrogen) or the RT Supermix M-MuLV kit (BioPioneer). Real-time PCR was performed using the SYBR-Green PCR Master mix (Applied Biosystems) in the ViiA 7 Real Time PCR System (Applied Biosystems). Glyceraldehyde 3-phosphate dehydrogenase expression was used to normalize values of gene expression and data is shown as fold change relative to the value of the sample control. All the samples were done in triplicate. Primers used for real-time PCR experiments are listed in Supplementary Table S3.

Whole-genome library construction. Library construction was performed as previously described^{6,26}. Briefly, for each sample, roughly 1.5–3 μg of genomic DNA (in 100 μl volumes) was sheared with a Covaris AFA. The fragmented genomic DNA was end repaired, A-tailed and ligated to sequencing adaptors, with a purification step between each process. The purified ligated products were then amplified by PCR to generate whole-genome libraries.

In-solution hybridization capture with DNA baits. Liquid exome capture was performed as previously described⁶.

Consensus sequence generation and variant calling. Variant calling was performed as previously described⁶. Briefly, reads obtained from the Illumina Genome Analyzer were post-processed and quality filtered using GERALD, mapped to the genome using BWA, downsampled using Picard and used to generate a consensus sequence for each sample using GATK. The consensus sequences were then compared with find candidate novel mutations in hiPSCs⁶. Sites where each hiPSC line showed heterozygous SNPs not observed in the progenitor line were considered as candidate mutations if no allelic content was present in the somatic progenitor and if the candidate mutation had not previously been observed in other samples or the dbSNP database.

Sanger validation of candidate mutations. Genomic DNA of both the hiPSC line and its somatic progenitor (6 ng each) was amplified in separate $50\,\mu l$ PCR reactions with $100\,nM$ of specifically designed forward and reverse primers around the mutation site (primers available under request) and $25\,\mu l$ of Taq $2\times$ master mix (NEB) at 94 °C for 2 min, followed by 35 cycles of 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 25 pmol of the forward primer for Sanger sequencing at Genewiz Inc.

Statistical analysis/TiGER database. To check for enrichment of reprogramming-associated mutations in genes that are expressed in a tissue-specific manner,

the fraction of UniGene IDs corresponding to mutated genes called as 'tissue-specific' in the TiGER database was identified as 49/132 (37%). As 6,699/19,526 (34%) of the genes annotated in the TiGER database are considered to be tissue specific, a χ^2 -test with one degree of freedom can be used to test for equivalency of distribution. The obtained χ^2 value is 0.460, indicating that the fraction of mutated hiPSC genes that are tissue specific is not significantly different than that found in a random sample of genes (P=0.4975). Reprogramming-associated mutations therefore do not appear to be enriched in tissue-specific genes.

Statistical analysis/active and inactive chromatin states. To check for enrichment of reprogramming-associated mutations in active or inactive chromatin, we utilized a χ^2 -test with three degrees of freedom to test for equivalency of distribution. We identified the chromatin state of each mutated gene using previously published data¹³. These data divided each gene into one of four categories: no trimethylation, H3K4 trimethylation, H3K27 trimethylation, or both. We compared the distribution of mutated genes across each of these four categories with the expected distribution for all genes in three cell types: fibroblasts, ESCs and iPSCs¹³. The obtained χ^2 values were 1.03 (P=0.79), 3.78 (P=0.29) and 6.97 (P=0.07), respectively, indicating that the distribution of mutated hiPSC genes in each chromatin region is not significantly different than expected by random chance (α =0.01). Reprogramming-associated mutations therefore do not appear to be enriched in active or inactive chromatin states.

Non-coding versus coding mutations. To compare the mutation rates per base pair in coding and non-coding regions of the genome, variant calling was performed as above on non-coding regions of the genome surviving library enrichment in eight hiPSC lines and their progenitor lines. The mutation rate per base pair was then estimated by dividing the number of candidate coding and non-coding mutations by the number of exomic and non-coding base pairs covered. The average coding and non-coding mutation rates were compared.

References

- Takahashi, K. et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131, 861–872 (2007).
- Yu, J. et al. Induced pluripotent stem cell lines derived from human somatic cells. Science 318, 1917–1920 (2007).
- Lowry, W. E. et al. Generation of human induced pluripotent stem cells from dermal fibroblasts. Proc. Natl Acad. Sci. USA 105, 2883–2888 (2008).
- Meissner, A., Wernig, M. & Jaenisch, R. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat. Biotech.* 25, 1177–1181 (2007).
- Park, I. H. et al. Reprogramming of human somatic cells to pluripotency with defined factors. Nature 451, 141–146 (2008).
- Gore, A. et al. Somatic coding mutations in induced pluripotent stem cells. Nature 471, 63–67 (2011).
- Hussein, S. M. I. et al. Copy number variation and selection during reprogramming to pluripotency. Nature 471, 58–62 (2011).
- Mayshar, Y. et al. Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. Cell Stem Cell 7, 521–531 (2010).
- Lister, R. et al. Hotspots of aberrant epigenetic reprogramming in human induced pluripotent stem cells. Nature 471, 86–73 (2011).
- Laurent, L. C. et al. Dynamic changes in the copy number of pluripotency and cell proliferation genes in ESCs and iPSCs during reprogramming and time in culture. Cell Stem Cell 8, 106–118 (2011).
- Ohi, Y. et al. Incomplete DNA methylation underlies a transcriptional memory of somatic cells in human iPS cells. Nat. Cell Biol. 13, 541–549 (2011).
- Liu, X., Yu, X., Zack, D. J., Zhu, H. & Qian, J. TiGER: a database for tissuespecific gene expression and regulation. BMC Bioinformatics 9, 271 (2008).
- Guenther, M. G. et al. Chromatin Structure and gene expression programs of human embryonic and induced pluripotent stem cells. Cell Stem Cell 7, 249–257 (2010).
- Yamanaka, S. Elite and stochastic models for induced pluripotent stem cell generation. Nature 460, 49–52 (2009).
- Feng, Q. et al. Hemangioblastic derivatives from human induced pluripotent stem cells exhibit limited expansion and early senescence. Stem Cells 28, 704–712 (2010).
- Hu, B. Y. et al. Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. Proc. Natl Acad. Sci. USA 107, 4335–4340 (2010).
- 17. Miura, K. *et al.* Variation in the safety of induced pluripotent stem cell lines. *Nat. Biotech.* **27**, 743–745 (2009).
- Cheng, L. et al. Low incidence of DNA sequence variation in human induced pluripotent stem cells generated by nonintegrating plasmid expression. Cell Stem Cell 10, 337–344 (2012).
- Young, M. A. et al. Background mutations in parental cells account for most of the genetic heterogeneity of induced pluripotent stem cells. Cell Stem Cell 10, 570–582 (2012).

7

- Kim, J. B. et al. Direct reprogramming of human neural stem cells by OCT4. Nature 461, 649–643 (2009).
- Panopoulos, A. D. *et al.* Rapid and highly efficient generation of induced pluripotent stem cells from human umbilical vein endothelial cells. *PLoS ONE* 6, e19743 (2011).
- 22. Ruiz, S. et al. High efficient generation of induced pluripotent stem cells from astrocytes. PLoS ONE 5, e15526 (2010).
- 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 (2010).
- Aasen, T. et al. Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. Nat. Biotech. 26, 1276–1284 (2008).
- Kawamura, T. et al. Linking the p53 tumour suppressor pathway to somatic cell reprogramming. Nature 460, 1140–1144 (2009).
- Quail, M. A., Swerdlow, H. & Turner, D. J. Improved protocols for the illumina genome analyzer sequencing system. *Curr Protoc Hum Genet*. Ch. 18, Unit 18 12 (2009).

Acknowledgements

We express our gratitude to Travis Berggren, Margaret Lutz and Veronica Modesto for their support at the Salk Institute-Stem Cell Core, to Joaquin Sebastian for critically reading the manuscript, to Guanghui Liu for sharing reagents and to the rest of the Belmonte lab. A.G. was supported by the Focht-Powell Fellowship and a CIRM predoctoral fellowship. Work in this manuscript was supported by grants from Fundacion Cellex, TERCEL-ISCIII-MINECO, Sanofi, National Institutes of Health and the G. Harold and Leila Y. Mathers Charitable Foundation.

Author contributions

S.R., A.G., K.Z. and J.C.I.B. designed the experiments. S.R., A.G., N.M., A.Gi., J.B., A.D.P., Z.L., H.F. and E.M.B. carried out experiments. H.Z. and H.S. helped with procedures. S.R., A.G., A.D.P., K.Z. and J.C.I.B. wrote the manuscript. K.Z. and J.C.B.I. directed the project and revised the manuscript. All authors provided detailed comments.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Ruiz, S. et al. Analysis of protein-coding mutations in hiPSCs and their possible role during somatic cell reprogramming. Nat. Commun. 4:1382 doi: 10.1038/ncomms2381 (2013).