STEM CELLS®

EMBRYONIC STEM CELLS/INDUCED PLURIPOTENT STEM CELLS

# Elevated Coding Mutation Rate During the Reprogramming of Human Somatic Cells into Induced Pluripotent Stem Cells

Junfeng Ji[*1], Siemon H. Ng[*1], Vivek Sharma[1], Dante Neculai[1], Samer Hussein[2], Michelle Sam[1], Quang Trinh[1], George M. Church[3], John D. McPherson[1], Andras Nagy[2], and Nizar N. Batada[1#]

[1]Ontario Institute for Cancer Research, Cancer Genomics, 101 College Street, Toronto, Canada M5G 0A3; [2]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada M5G 0A3; [3] Harvard Medical School, Department of Genetics, Harvard Medical School, 77 Ave Louis Pasteur, Boston, MA 02115

**Key words.** iPSCs • stem cells • reprogramming • mutations • genome instability • exome sequencing

**ABSTRACT**

**Mutations in human induced pluripotent stem cells (iPSCs) pose a risk for their clinical use due to preferential reprogramming of mutated founder cell and selection of mutations during maintenance of iPSCs in cell culture. It is unknown, however, if mutations in iPSCs are due to stress associated with oncogene expression during reprogramming. We performed whole exome sequencing of human foreskin fibroblasts and their derived iPSCs at two different passages. We found that *in vitro* passaging contributed 7% to the iPSC coding point mutation load and ultra deep amplicon sequencing revealed that 19% of the mutations preexist as rare mutations in the parental fibroblasts suggesting that the remaining 74% of the mutations were acquired during cellular reprogramming. Simulation suggests that the mutation intensity during reprogramming is 9-fold higher than the background mutation rate in culture. Thus the factor induced reprogramming stress contributes to a significant proportion of the mutation load of iPSCs.**

## INTRODUCTION

Somatic cells can be reprogrammed to embryonic stem (ES) cell-like cells known as induced pluripotent stem cells (iPSCs) via forced expression of defined transcription factors (Takahashi, Tanabe et al. 2007; Yu, Vodyanik et al. 2007; Park, Zhao et al. 2008). However, reprogramming might be mutagenic as most of the reprogramming factors are known to be oncogenic (Rowland, Bernards et al. 2005; Bass, Watanabe et al. 2009; Viswanathan, Powers et al. 2009) and generate genotoxic stress (Banito, Rashid et al. 2009; Marion, Strati et al. 2009; Esteban, Wang et al. 2010) causing cell cycle arrest (Hong, Takahashi et al. 2009), cellular senescence (Banito, Rashid et al. 2009; Li, Collado et al. 2009) and apoptosis (Marion, Strati et al. 2009) in factor recipient fibroblasts.

Karyotypic (Taapken, Nisler et al. 2011) and meta-analysis of gene expression data (Mayshar, Ben-David et al. 2010) revealed aneuploidy and copy number analysis detected large scale sub-chromosomal aberrations (Laurent, Ulitsky et al.

Author contributions: J.J.: Conception and design, Provision of study material, Collection and/or assembly of data; S.N.: Conception and design, Provision of study material, Collection and/or assembly of data, Data analysis and interpretation; D.N.: Collection and/or assembly of data; V.S.: Collection and/or assembly of data; M.S.: Collection and/or assembly of the data; S.H.: Collection and/or assembly of data; G.C.: Collection and/or assembly of data; J.M.: Collection and/or assembly of data; A.N.: Collection and/or assembly of data; N.B.: Conception and design, Data analysis and interpretation, Financial support and Manuscript writing

**Corresponding author**: Nizar N. Batada, Ontario Institute for Cancer Research, 101 College Street, Toronto, Canada M5G 0A3, Phone: (416) 673-8594 Fax: (416) 977-1118 (fax), Email: nizar.batada@oicr.on.ca; *These authors made equal contribution; Received September 30, 2011; accepted for publication November 23, 2011.  1066-5099/2011/$30.00/0 doi: 10.1002/stem.1011

2011; Martins-Taylor, Nisler et al. 2011) in iPSCs that arise upon prolonged passaging *in vitro*. Genome-wide copy number analysis of multiple iPSC lines found that regardless of the reprogramming factor combinations and gene delivery methods (retroviral vector and piggyBac transposon), iPSCs had many copy number variations (CNV) not present in the bulk parental cells (Hussein, Batada et al. 2011). Sequencing of iPSC lines made using both integrative and non-integrating methods (episomal and mRNA delivery) revealed an average of 6 nonsynonymous (i.e. protein sequence changing) point mutations per iPSC line (Gore, Li et al. 2011). About 60% of these mutations were present in the parental fibroblasts in very low frequency suggesting selection for mutated cells during reprogramming. Thus, so far, it is known that iPSCs harbor mutations despite absence of the MYC oncogene in the reprogramming factor cocktail and use of non-integrative reprogramming factor delivery methods and that some of the mutations are preexisting in the parental calls and some are acquired during passaging. However it remains unknown what proportion of the mutations in iPSCs are acquired due to the genotoxic stress associated with reprogramming.

In this study, we determined the mutation rate during iPSC passaging by whole exome sequencing of several iPSC lines at two different passages. We further estimated the proportion of iPSC mutations that preexist as rare mutations in the parental population using ultradeep amplicon sequencing. Despite being derived from a common parental source, these iPSCs had many unique non-silent coding mutations absent in the parental cells. We thus provide evidence that many of the coding mutations in iPSCs are incurred during the reprogramming phase.

## MATERIALS AND METHODS

***Cell culture***. Human neonatal foreskin fibroblasts (HFFs) (ATCC, Manassa, VA) were maintained in fibroblast medium consisting of DMEM (Invitrogen, Carlsbad, CA) supplemented with 10% fetal calf serum (FCS; Hyclone Laboratories, Mississauga, ON) and 1mM L-glutamine (Invitrogen, Carlsbad, CA). Human embryonic stem cells (hESCs), HES2 (WiCell, Madison, WI), and iPSCs were maintained on feeder-free Matrigel (BD Biosciences, Mississauga, ON)-coated plate in complete mTeSR medium (STEMCELL Technologies, Vancouver BC) as previously described1,2.

***Retrovirus production***. Four moloney-based retroviral vectors (pMXs) containing the human complimentary DNAs (cDNAs) of OCT4, SOX2, KLF4 and c-MYC (ref. 3) were obtained from Addgene (Addgene, Cambridge, MA). These plasmids were transfected into a previously established 293GPG packaging cell line that incorporated pMD.gagpol and tetracycline-inducible VSV-G plasmids to generate high titer retroviruses.4 Viral supernatant was collected 48, 72 and 96 h post-transfection and filtered by 0.45 um syringe filters.

***Generation of human induced pluripotent stem (iPS) cells.*** Approximately ~3 x $10^5$ HFFs were seeded in gelatin-coated 100 mm dishes in fibroblast medium and were infected twice by OCT4, SOX2, KLF4 and c-MYC trangene containing retroviruses during a 48 h period after seeding HFFs. Approximately 24h after second viral infection, cells were switched to hESC media consisting of Knockout DMEM supplemented with 20% knockout serum replacement, 1mM L-glutamine, 1% non essential amino acid, 0.1mM ß-mercaptoethanol and 10 ng/ml human basic fibroblast growth factor (bFGF; Invitrogen, Carlsbad, CA). Human iPS cell lines were established 3 to 4 weeks post-

infection by selecting newly formed colonies with hESC-like colony morphology.

***Immunocytochemistry***. iPSCs were fixed with PBS containing 4% paraformaldehyde for 20 min at room temperature, washed with PBS. For NANOG and SOX17 intracellular staining, cells were permeabilized with 0.2% Triton X-100 for 10 min at room temperature. The cells were then blocked with 10% normal goat or mouse serum (Vector Labs) in PBS for 1 h and incubated overnight at 4 °C with one of the following primary antibodies: SSEA4 (1:50, Developmental Studies Hybridoma Bank), TRA-1-60 (1:50, Millipore), NANOG (1:20, R&D Systems, Minneapolis, MN), SOX17 (1:50, R&D Systems, Minneapolis, MN), A2B5 (1:50, R&D Systems, Minneapolis, MN). Cells were then washed and incubated for 1 h at room temperature with Alexa488 or Alexa594-conjugated secondary antibodies (1:250, Invitrogen, Carlsbad, CA).

***Flow cytometry***. Single cell suspension were prepared from day 15 embryoid bodies (EBs) by treatment with Collagenase B (Roche, Mississauga, ON) and cell dissociation buffer (Invitrogen, Carlsbad, CA) followed by staining with human CD31-PE and CD34-FITC antibodies (Miltenyi Biotec, Auburn, CA). The cells were then subjected to FACSCalibur (BD Biosciences, Mississauga, ON) for data acquisition and data were analysed by FlowJo software (www.flowjo.com, Tree Star, Ashland, OR). Flow gates were based on isotype controls.

***PCR and qPCR***. Total RNA was isolated using RNeasy (QIAgen, Valencia, CA) and treated with Turbo DNAase (Ambion, Carlsbad, CA) to remove genomic DNA contamination. DNAase-treated RNA was re-purified using ammonium acetate precipitation method after inactivation of DNAase by EDTA (Sigma-Aldrich, Oakville, ON). cDNA was generated from 1 ug of total purified RNA using Reverse Transcription System (Promega, Madison, WI) according to the manufacturer's instructions. All PCR was performed with High-Fidelity Taq DNA polymerase (Invitrogen, Carlsbad, CA). Quantitative PCR was performed with SYBR Green qPCR kit (New England Biolabs, Ipswich, MA) using 20ul of total reaction and analyzed on the 7900HT real-time PCR system (Applied Biosciences, Carlsbad, CA). Primer sequences are given in Supplementary Table.

***Bisulfite Conversion***. 1 ug of genomic DNA from human iPSCs, hESC line HES2, and parental HFFs were processed for bisulfite modification using EpiTech Bisulfite Kit (QIAgen, Valencia, CA) according to the manufacturer's instructions. The promoter regions of human OCT4 were amplified by PCR using previously reported primer sets, cloned into pCR2.1-TOPO vector using TOPO T/A cloning kit (Invitrogen, Carlsbad, CA) and sequenced using both forward and reverse primers.

***Next Generation Sequencing***. DNA was extracted using Blood & Cell Culture DNA Midi Kit (QIAgen, Valencia, CA). Illumina libraries were prepared according to manufacturer's protocols and exome was captured using Agilent's SureSelect Exon Capture (50 Mb target) according to manufacturer provided protocol. All sequencing was carried out on a Illumina Genome Analyzer IIx sequencer with 76 x 2 paired-end reads. One lane of sequencing was done for each sample.

***Read mapping***. The reference human genome used in these analysis was UCSC assembly hg18 (NCBI build 36.1) containing unordered sequences (i.e. sequences that are known to be in a particular chromosome, but could not be reliably ordered within the current sequence). The 76x2 paired-end reads were mapped on to the human reference genome using BWA (version 0.5.7) (Li and Durbin 2009). Quality scores were recalibrated using GATK (McKenna, Hanna et al. 2010) and PCR artifacts were discarded using Picard. Only uniquely mapping reads (BWA mapping quality >= 1) were retained for further analysis.

***Identification and annotation of mutations***. Only targeted genomic regions with at least 10x coverage and Phred-scaled base quality of 30 or higher were considered. Candidate iPSC mutations are defined as variants that are present in a given iPSC exome but not in the fibroblasts or in the other iPSC exomes. These candidate mutations were subjected through a series of filters: 1) candidate mutations were discarded if they were present exclusively in latter third of the reads; 2) we disregarded candidate mutations if the mutant allele was the flanking homopolymeric (defined as repeats of two or longer) base; 3) candidate mutations were discarded if BLAST alignment of reads containing them did not concur with BWA mapping. Functional annotation of SNPs into nonsynonymous and synonymous and prediction of damaging mutations were done using SeattleSeq Annotator online tool (http://gvs.gs.washington.edu/SeattleSeqAnnotation/).

***Simulation of mutations in iPS cells.*** To compute the expected number of mutations in iPSCs at passage 6 due to the number of cell doublings from the onset of the reprogramming till passage 6 if the mutation rate during reprogramming is not elevated, we simulated a random mutagenesis process and accounted for differences in cell cycle length and passaging interval of iPSC cells and fibroblasts using the following parameters: Cell doubling time of human foreskin fibroblasts in culture is taken to be 24 hours (Ho, Cheng et al. 2000) and doubling time of iPSCs is taken to be 44 hours (Takahashi, Tanabe et al. 2007). Reprogramming duration is taken to be 4 week (28 days) with a doubling time of 34 hours which represents the average doubling time of fibroblasts and iPSCs. Thus during the reprogramming phase, a reprogramming cell has undergone ~20 cell doublings. After picking the initial iPSC colonies, the duration of passaging was 1 week; therefore, during the 6 passages of the initial iPSC colony, ~23 cell doublings have taken place. Thus, a total of ~43 doublings have taken place since the reprogramming factors were delivered into the parental fibroblasts until the 6 passages of iPSCs. For the 32.4Mb targeted region that is presence above 10x in all the iPSCs and parental fibroblast exome, a background mutation rate of 0.02 coding mutations per cell division (which equates 6.7 x $10^{-10}$ per bp per cell division (Gore, Li et al. 2011)) leads to an expectation of 0.94 coding mutations per iPSC. Setting the background mutation rate to the mutation rate during iPSC passaging (0.035 coding mutations per cell division) leads to an expectation of 1.5 coding mutations in iPSCs at passage 6. We found an average of 12 mutations per iPSCs, out of which 7% were attributed to passaging and 19% were preexisting resulting in ~9 mutations that are not explained by iPSC passaging or from inheritance of mutations from parental cells. Thus the mutations rate during the reprogramming phase is 6 to 9.4 times higher than that expected for the background mutation rate associated with cell divisions.

## RESULTS

Human primary neonatal foreskin fibroblasts (ATTC, catalog # CRL-2429) (passage 14) were reprogrammed with retroviruses encoding *KLF4, MYC, OCT4* and *SOX2* transgenes (Takahashi, Tanabe et al. 2007). We sought to minimize technical variations by using the same gene delivery method, reprogramming factors, viral titer, culture conditions and passaging intervals. The randomly selected five iPSCs displayed all the hallmarks of pluripotent cells such as expression of pluripotency markers, demethylation of *OCT4* promoter, transgene silencing and potential to differentiate into derivatives from the three germ layers (**Supplementary Figure 1 and 2**). Each iPSC line was sequenced after 6 (p6-iPSC) and 12 (p12-iPSC) passages subsequent to picking the initial iPSC colony 28 days after reprogramming. We enriched for DNA encoding protein coding genes using the Agilent SureSelect Human All Exon kit and sequenced the captured DNA from the 11 samples (i.e. parental fibroblasts, five p6-

iPSC lines and five p12-iPSC lines) using the Illumina Genome Analyzer IIx (Bentley, Balasubramanian et al. 2008) with one sample per lane. After aligning the reads to the human reference genome, we obtained over 60 million uniquely aligning reads per sample (**Table 1a**) and from here on refer to the sequence data as the exome.

We developed a custom single nucleotide variant caller (**Supplementary Methods; Supplementary Figure 3**) to identify all the alleles in fibroblasts and iPSCs that are absent in the human reference NCBI build 36.1 (from here on referred to as variants) at targeted exons representing ~32.4Mb of the genome with a minimum of 10-fold (10x) coverage in fibroblasts and in all the iPSC lines (**Table 1a**). An iPSC variant is defined to be a mutation if it is present only in a single iPSC exome and absent in the parental fibroblasts exome (**Table 1b**). As the iPSCs were derived from a common batch of fibroblasts from a single individual, presence of mutations unique to each iPSCs (i.e. absent in parental cells and in other iPSCs derived from the same fibroblasts during the single experiment) provides a stringent test for excluding variants that did not arise during reprogramming. Thus, each iPSC candidate mutation was tested in 5 independent exomes (i.e. the parental fibroblasts and the four other p6-iPSC exomes) to discard preexisting parental mutations. We found 59 mutations (~12 mutations per iPSC line on average) in the p6-iPSC exomes (**Table 1b; Supplementary Table 1**). We randomly selected 30 candidate iPSC mutations and interrogated their presence in the parental fibroblasts and other iPSC lines via Sequenom MassArray SNP genotyping system which can detect alleles present in 10% frequency. We confirmed that all the candidate iPSC mutations found were present in their corresponding iPSC line and absent in the parental fibroblasts and other iPSC lines (**Supplementary Table 2**).

To estimate the proportion of coding mutations in p6-iPSCs that are likely acquired during passaging since the picking of the initial iPSC colony, we sequenced p12-iPSCs. We found a total of 60 mutations in p12-iPSCs. Coding point mutations largely persisted during passaging (56 of the 59 mutations in the p6-iPSC were present in the p12-iPS) (**Table 1b**). We designated all mutations identified in the p12-iPSC exome but absent in the p6-iPSC exome of the same iPSC line as passaging-induced mutations**.** We found a total of 4 passaging induced heterozygous mutations in the p12-iPSC lines (a mutation rate of 0.1333 coding point mutations per passage per iPSC line of 0.035 coding mutations per cell division) (**Table 1b; Supplementary Table 1**). Thus, assuming that the rate of mutations due to passaging is constant (as cells are treated identically during each passage), passaging induced mutations account for ~7% (4 out of 59) of the mutations in the p6-iPSC exome. As it is possible that some of these 4 mutations are also present in p6-iPSCs below detection limit, the estimated passaging induced mutation rate is likely an overestimate and consequently the proportion of mutations in p6-iPSCs that is non-passaging induced is likely an underestimate.

To estimate the proportion of mutations that might be preexisting mutations in rare fibroblast subpopulation, we performed deep amplicon sequencing of 46 randomly selected mutations out of the 59 mutations identified in the p6-iPSC exomes. The genomic regions spanning the candidate mutations were covered at ~3 million times on average allowing detection of rare mutant alleles in the fibroblast population; only 8 out of the 46 (or 17%) mutations were present in rare frequency (**Supplementary Table 3**).

As P53 is demonstrated to be required for maintaining genome integrity of iPSCs (Marion, Strati et al. 2009) and expected to prevent accumulation of reprogramming-induced mutations in iPSCs, we asked if any of these iPSC lines incurred mutations in *TP53* or were derived from founder cells with mutated *TP53* which may explain survival of iPSCs despite DNA damage. As the mutations inherited from the parental fibroblast or acquired during

reprogramming should be high in frequency, we used capillary sequencing to check if the *P53* gene incurs inactivating mutation during iPSC generation. None of the 11 exons of the *TP53* have nonsynonymous mutations in any of the iPSC lines (**Supplementary Table 4**). There were also no deleterious variants in *MDM2, CDKN2A, P21* and *BCL2* which genes are upstream or downstream of P53. Gene ontology analysis of mutated genes revealed no significant enrichment for membership in any particular biological process that suggests defects in checkpoint arrest or apoptosis. No homozygous nonsynonymous variant was found in known DNA repair genes (Wood, Mitchell et al. 2001). Thus the iPSC line founder fibroblast does not have obvious defects in genome maintenance which may make them prone to incur mutations during reprogramming.

## DISCUSSION

We partitioned iPSC mutations into three mutually exclusive classes: Class I represents mutations that preexisted in the parental cells, Class II represents mutations incurred during reprogramming, and Class III represents mutations acquired during iPSC maintenance. Two models may account for the number and the relative proportion of mutations in p6-iPSC that are Class I, Class II and Class III. According to the first model, which we refer to as the Constant Mutation Rate model, mutations in p6-iPSCs reflect accrual of mutations that occur at a background mutation rate during the numerous cell divisions that take place during reprogramming. According to the second model, which we refer to as the Reprogramming Stress model, the mutation rate during reprogramming is highly elevated due to the stress associated with cell fate alteration caused by the overexpression of oncogenic reprogramming factors. This model is based on the fact that the reprogramming factors have oncogenic potential (Rowland, Bernards et al. 2005; Bass, Watanabe et al. 2009; Viswanathan, Powers et al. 2009) and activate the DNA damage response (Banito, Rashid et al. 2009; Marion, Strati et al. 2009; Esteban, Wang et al. 2010) reflecting an increased rate of genome instability during reprogramming. Due to the low efficiency of reprogramming, empirical measurement of mutation rate in the subset of fibroblasts that undergo reprogramming is technically challenging. To determine which of these two models better explains the mutations seen in p6-iPSCs, we simulated the random mutational process associated with genome duplication according to the Constant Mutation Rate model. Substitution point mutations can accumulate during the reprogramming phase and during passaging of the initial iPSC colony for 6 passages (parameters are listed in the **Methods**). To match the number of parental cells seeded during reprogramming, we simulated the mutational process for 500,000 single cells. The distribution of the number of mutations simply due to background mutation rate (0.02 coding mutations per cell division (Gore, Li et al. 2011)) in an iPSC line at passage 6 gives a median of one coding point mutation (**Figure 1a**). Using the mutation rate during iPSC passaging as the background mutation rate increases the median number of coding point mutations in p6-iPSC to 2 mutations per line. To explain the observed number of coding mutations seen in the p6-iPSCs without the need for elevated mutation rate during reprogramming at a background mutation rate of 0.3 coding mutations per cell division is required. This mutation intensity is 9 fold higher than iPSC passaging mutation rate. Thus the Reprogramming Stress model better explains the observed mutation rate in p6-iPSCs. That our iPSCs were derived from the same batch of parental cells but harbored unique mutations not found in other iPSCs derived from the same parental cells further supports the reprogramming associated mutagenesis model.

## CONCLUSION

While the Gore et al study (Gore, Li et al. 2011) demonstrated that iPSCs had mutations and that many have originated from the parental founder cell, we have shown that mutations are also acquired during reprogramming and passaging. Furthermore, we find that less than ~20% of the mutations in iPSCs were preexisting mutations from the parental cells and that reprogramming contributes ~75% of the mutations found in our fibroblast derived iPSCs. This discrepancy cannot be explained by the fact that we used neonatal source while Gore et al used adult tissue as age was poorly correlated with the proportion of mutations found in iPSCs (Gore, Li et al. 2011). Our study provides strong evidence that coding point mutations are incurred during the reprogramming phase of iPSC generation and that passaging of iPSCs after the initial colony picking contributes to only a small proportion of the overall iPSC mutation load (**Figure 1b**). Furthermore, unlike in the case of copy number variations (Hussein, Batada et al. 2011), many of the coding point mutations in iPSCs persist during passaging. Our work highlights the need for identification of optimal conditions of reprogramming that reduce the mutations associated with iPSC generation.

**Data availability**

Upon acceptance, the raw exome sequencing data will be made available at http://batadalab.oicr.on.ca

## ACKNOWLEDGMENTS

This work was supported by funding to NNB who is the recipient of a New Investigator Award from the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research and Innovation.

## REFERENCES

1. Banito, A., S. T. Rashid, et al. (2009). "Senescence impairs successful reprogramming to pluripotent stem cells." Genes Dev 23(18): 2134-9.
2. Bass, A. J., H. Watanabe, et al. (2009). "SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas." Nat Genet 41(11): 1238-42.
3. Bentley, D. R., S. Balasubramanian, et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature 456(7218): 53-9.
4. Esteban, M. A., T. Wang, et al. (2010). "Vitamin C enhances the generation of mouse and human induced pluripotent stem cells." Cell Stem Cell 6(1): 71-9.
5. Gore, A., Z. Li, et al. (2011). "Somatic coding mutations in human induced pluripotent stem cells." Nature 471(7336): 63-7.
6. Ho, H. Y., M. L. Cheng, et al. (2000). "Enhanced oxidative stress and accelerated cellular senescence in glucose-6-phosphate dehydrogenase (G6PD)-deficient human fibroblasts." Free Radic Biol Med 29(2): 156-69.
7. Hong, H., K. Takahashi, et al. (2009). "Suppression of induced pluripotent stem cell generation by the p53-p21 pathway." Nature 460(7259): 1132-5.
8. Hussein, S. M., N. N. Batada, et al. (2011). "Copy number variation and selection during reprogramming to pluripotency." Nature 471(7336): 58-62.
9. Laurent, L. C., I. Ulitsky, et al. (2011). "Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture." Cell Stem Cell 8(1): 106-18.
10. Li, H., M. Collado, et al. (2009). "The Ink4/Arf locus is a barrier for iPS cell reprogramming." Nature 460(7259): 1136-9.
11. Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics 25(14): 1754-60.
12. Marion, R. M., K. Strati, et al. (2009). "A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity." Nature 460(7259): 1149-53.
13. Martins-Taylor, K., B. S. Nisler, et al. (2011). "Recurrent copy number variations in human induced pluripotent stem cells." Nat Biotechnol 29(6): 488-91.
14. Mayshar, Y., U. Ben-David, et al. (2010). "Identification and classification of chromosomal aberrations in human induced pluripotent stem cells." Cell Stem Cell 7(4): 521-31.
15. McKenna, A., M. Hanna, et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res 20(9): 1297-303.

16. Park, I. H., R. Zhao, et al. (2008). "Reprogramming of human somatic cells to pluripotency with defined factors." Nature 451(7175): 141-6.
17. Rowland, B. D., R. Bernards, et al. (2005). "The KLF4 tumour suppressor is a transcriptional repressor of p53 that acts as a context-dependent oncogene." Nat Cell Biol 7(11): 1074-82.
18. Taapken, S. M., B. S. Nisler, et al. (2011). "Karotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells." Nat Biotechnol 29(4): 313-4.
19. Takahashi, K., K. Tanabe, et al. (2007). "Induction of pluripotent stem cells from adult human fibroblasts by defined factors." Cell 131(5): 861-72.
20. Viswanathan, S. R., J. T. Powers, et al. (2009). "Lin28 promotes transformation and is associated with advanced human malignancies." Nat Genet 41(7): 843-8.
21. Wood, R. D., M. Mitchell, et al. (2001). "Human DNA repair genes." Science 291(5507): 1284-9.
22. Yu, J., M. A. Vodyanik, et al. (2007). "Induced pluripotent stem cell lines derived from human somatic cells." Science 318(5858): 1917-20.

See www.StemCells.com for supporting information available online.

**Figure 1.** The contribution of passaging, reprogramming stress and inheritance of rare preexisting mutations from parental cells to the mutation load of iPSCs. a) Figure shows the expected number of coding mutations per iPSC line simply due to background mutation rate during the numerous cell divisions that take place during reprogramming and passaging to passage 6. Shown is the histogram of 500,000 simulations to match the number of parental fibroblasts that were plated. Parameters are given in Supplementary Methods. b) Illustration of the contribution of Class I, Class II and Class III to the overall coding mutation load in iPSCs. P0 represents the initial colony.