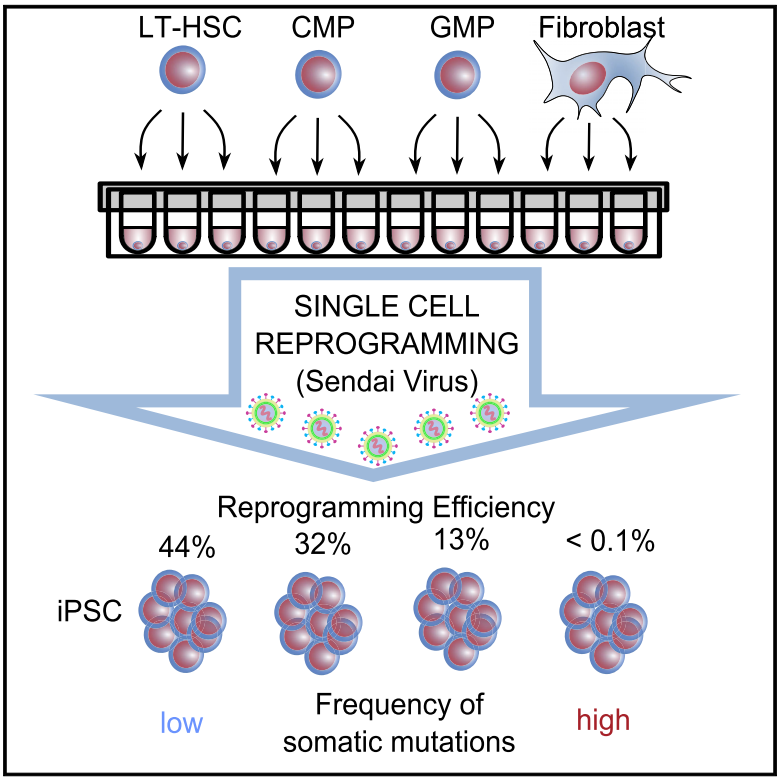
Article

Cell Reports

Ultra-High-Frequency Reprogramming of Individual Long-Term Hematopoietic Stem Cells Yields Low Somatic Variant Induced Pluripotent Stem Cells

Graphical Abstract



Authors

Kai Wang, Anthony K. Guzman, Zi Yan, ..., Kenny Ye, Boris Bartholdy, Eric E. Bouhassira

Correspondence

eric.bouhassira@einstein.yu.edu

In Brief

Wang et al. show that single adult human long-term hematopoietic stem cells can be reprogrammed into induced pluripotent stem cells at close to 50% efficiency and contain fewer somatic single-nucleotide variants and indels than skin fibroblasts. They may become the preferred source for the production of clinical-grade iPSCs.

Highlights

- Single adult human LT-HSCs can be reprogrammed into iPSCs at close to 50% efficiency.
- LT-HSCs contain less somatic variants than skin fibroblasts.
- LT-HSCs may become the preferred source for the production of clinical-grade iPSCs.

Wang et al., 2019, Cell Reports *26*, 2580–2592
March 5, 2019
https://doi.org/10.1016/j.celrep.2019.02.021

CellPress

OPEN ACCESS
CellPress

Cell Reports
Article

Ultra-High-Frequency Reprogramming of Individual Long-Term Hematopoietic Stem Cells Yields Low Somatic Variant Induced Pluripotent Stem Cells

Kai Wang,[1] Anthony K. Guzman,[2] Zi Yan,[1] Shouping Zhang,[1] Michael Y. Hu,[1] Mehdi B. Hamaneh,[3] Yi-Kuo Yu,[3] Seda Tolu,[2] Jinghang Zhang,[4] Holly E. Kanavy,[2] Kenny Ye,[5] Boris Bartholdy,[1] and Eric E. Bouhassira[1,6,*]

[1]Department of Cell Biology, Albert Einstein College of Medicine, Bronx, NY
[2]Department of Internal Medicine, Division of Dermatology, Albert Einstein College of Medicine, Bronx, NY
[3]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD
[4]Department of Microbiology & Immunology, Albert Einstein College of Medicine, Bronx, NY
[5]Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY
[6]Lead Contact
*Correspondence: eric.bouhassira@einstein.yu.edu
https://doi.org/10.1016/j.celrep.2019.02.021

SUMMARY

Efficiency of reprogramming of human cells into induced pluripotent stem cells (iPSCs) has remained low. We report that individual adult human CD49f$^+$ long-term hematopoietic stem cells (LT-HSCs) can be reprogrammed into iPSCs at close to 50% efficiency using Sendai virus transduction. This exquisite sensitivity to reprogramming is specific to LT-HSCs, since it progressively decreases in committed progenitors. LT-HSC reprogramming can follow multiple paths and is most efficient when transduction is performed after the cells have exited G_0. Sequencing of 75 paired skin fibroblasts/LT-HSC samples collected from nine individuals revealed that LT-HSCs contain a lower load of somatic single-nucleotide variants (SNVs) and indels than skin fibroblasts and accumulate about 12 SNVs/year. Mutation analysis revealed that LT-HSCs and fibroblasts have very different somatic mutation signatures and that somatic mutations in iPSCs generally exist prior to reprogramming. LT-HSCs may become the preferred cell source for the production of clinical-grade iPSCs.

INTRODUCTION

The first human induced pluripotent stem cells (iPSCs) were produced by reprogramming fibroblasts using OCT4, KLF4, SOX2, and MYC (Takahashi et al., 2007) or OCT4, SOX2, NANOG, and LIN28 (Yu et al., 2007). Since then, multiple studies have shown that iPSCs can also be produced with fewer than four factors in certain cell types (Hermann et al., 2016) and by substituting KLF4, SOX2, and MYC with related genes (Nakagawa et al., 2008), microRNAs (miRNAs), or small molecules (Hou et al., 2013; Miyoshi et al., 2011; Zhao et al., 2015). Reprogramming yields can be increased by knocking down the expression of p53, or by using a variety of genes or small molecules (Takahashi and Yamanaka, 2016). Eminli et al. (2009) achieved a reprogramming frequency of 28% and demonstrated that hematopoietic stem and progenitor cells (HSPCs) were more amenable to reprogramming than mature blood cells using mouse cells engineered to express inducible reprogramming factors. Merling et al. (2013) reprogrammed human peripheral blood (PB) CD34$^+$ cells obtained from a few milliliters of blood with either removable lentiviruses or Sendai viruses; however, the use of the latter method demonstrated a highly variable frequency of reprogramming. Despite this progress, the efficiency of reprogramming of human cells remains generally low at about 0.1% for fibroblasts and 1%–5% for CD34$^+$ hematopoietic cells (Schlaeger et al., 2015)

The mechanism of reprogramming is incompletely understood but has been shown to involve multiple steps. The low efficiency of reprogramming has been partially attributed to abortive reprogramming (Plath and Lowry, 2011) because many cells transiently expressing the four factors undergo dramatic morphological changes but die before completing the process. It has been proposed that these reprogramming factors act as pioneer factors that are able to bind and activate first enhancers and then promoters that are not in an open chromatin configuration (Soufi et al., 2012) and that the early steps of reprogramming are stochastic in nature (Buganim et al., 2012), possibly due to non-synchronous binding of the reprogramming factors to cellular enhancers and promoters that are not in favorable configurations. Once this initial wave of genes are activated, the process appears to be more predictable (Buganim et al., 2012). Rais et al. (2013) demonstrated in both mouse and human cells that knocking out methyl-CpG binding domain protein 3 (Mbd3) decreases the early barrier to fibroblast reprogramming and allows the production of iPSCs at a very high frequency in a more deterministic manner.

Analysis of over a thousand lines has revealed that iPSC are karyotypically stable (Taapken et al., 2011), although a few recurring rearrangements have been detected in a subset of iPSC lines (Peterson and Loring, 2014). Detailed exome and genome sequencing analyses have shown that iPSCs derived from skin

fibroblasts carry many somatic variants that are for the most part already present in the source cell (Abyzov et al., 2012; Gore et al., 2011; Young et al., 2012). A recent genome-wide study confirmed that iPSC lines derived from fibroblasts collected from a single patient carry between 200 and 700 somatic variants per genome (Bhutani et al., 2016). Although most of these variants are intergenic and are unlikely to have important functional significance, a few will likely prove detrimental. Minimizing the number of somatic variants in iPSC is therefore important, if these cells are to be used in large-scale clinical applications.

Hematopoietic stem cells (HSCs) have long been detectable in transplantation assays but the isolation of a pure population of human HSCs with long-term repopulating activity (LT-HSCs) has been difficult because of the lack of appropriate cell surface markers. Notta et al. (2011) reported a cell isolation strategy based on 15 markers, including the CD49f antigen, which allows for the isolation of a population of cells from cord blood (here referred to as 49f cells) and contains about one LT-HSCs in 10 cells. We report here that 49f cells isolated from human PB or bone marrow (BM) can be reprogrammed into iPSCs at an extremely high efficiency (>45%) and that they contain significantly fewer somatic variants than skin fibroblasts.

RESULTS

To measure the frequency of reprogramming of purified HSPCs, we sorted 49f cells, common myeloid progenitors (CMPs), and granulocyte macrophage progenitors (GMPs) using the markers described by Notta et al. (2011) and by Manz et al. (2002) (Figure 1A) and verified the quality of our PB 49f cells fraction by assessing their repopulating activity into irradiated immunodeficient mice (Figures S1A and S1B). Frequencies of reprogramming were then assessed by transducing individual 49f cells, CMPs, and GMPs sorted into the wells of a round-bottom 96-well plate and transduced prior to their first division with Sendai viruses encoding the reprogramming factors Oct4, KLF4, Myc, and Sox2 (Figure 1B) using a protocol derived from that of Merling et al. (2013). The efficiency of reprogramming was assessed by alkaline phosphatase staining and three to six iPSC clones from each individual tested were selected and characterized with several quality control assays (see below).

Reprogramming frequencies for 49f cells, CMPs, and GMPs isolated from seven different individuals were respectively 44.5 ± 4.1% (average ± SEM), 32.1 ± 3.2%, and 13.2 ± 4.0% (Figure 1C). Differences between LT-HSCs and CMPs, and between LT-HSCs and GMPs, were statistically significant (t test, p values equal to 0.008 and 0.001, respectively). 49f cells were respectively 100-fold and 10-fold more amenable to reprogramming than skin fibroblasts and bulk PB $CD34^+$ cells since, as expected, the former cell type yielded about 1 iPSC per 1,000 cells, and the latter about 4 iPSC per 100 cells (data not shown; Schlaeger et al., 2015).

Quality control assays revealed that all 49f-iPSC clones tested had normal karyotype, expressed SSEA4, TRA-1-60, and TRA1-80 homogeneously, and were able to differentiate into the three germ layers as determined by embryoid body and by teratoma formation (Figures S2A–S2E). By passage 15, all of the 49f-iPSCs tested were transgene-free (Figure S2F).

Abortive Reprogramming

Microscopic observation during the first week after transduction revealed that 49f cells that did not become iPSCs almost always underwent the typical morphological changes associated with iPSCs but died before completing the process (data not shown). Therefore, abortive reprogramming occurs in PB 49f cells, although at a much lower frequency than during fibroblast reprogramming.

In additional experiments, we let the PB 49f cells divide three to five times before the Sendai virus transduction. Under these conditions, close to 100% of the wells yielded an iPSC clone demonstrating that almost all PB 49f cells produce progeny that are reprogrammable (Figure 1D). Therefore abortive reprogramming was not caused by an intrinsic defect in the 49f cells, but either to differences in transduction efficiency, to the stochasticity of the reprogramming process, or to the phase of the cell cycle in which the cells were located at the time of the transduction.

To determine whether transduction efficiency was likely to be an important factor, we infected PB 49f cells with a Sendai virus encoding GFP. This revealed that PB 49f cells are exquisitely sensitive to Sendai virus infection since all of the cells were brightly green fluorescent as little as 12 h after transduction (not shown), suggesting that low efficiency of infection is likely not the cause of abortive reprogramming.

Reprogramming of BM 49f Cells

The 49f cells found in PB are a minority of all 49f cells in the body since most 49f cells are located in the BM. To determine whether BM 49f cells are also highly susceptible to reprogramming, we attempted to reprogram them using the protocol that we used for PB 49f cells, but the frequency of reprogramming was much lower (3 out 58 tested cells became iPSCs). Further analysis revealed that BM 49f cells take longer to re-enter the cell cycle than PB 49f cells when plated in week 1/STIF media since 23% and 66% of the PB 49f cells, respectively, divided within 24 and 48 h while only 2% and 31% of the BM 49f divided during the same time interval. Since almost all 49f cells from both PB and BM eventually divided, we also concluded that most 49f cells were viable (Figure 2A, top).

To increase the rate of cell cycle re-entry of BM 49f cells after cell sorting, we tested several cytokines and small molecules. This yielded SR1+, a 49f cell expansion cocktail composed of stem cell factor (SCF), Tpo, IL6, FLT3 (the STIF cocktail) plus granulocyte-colony-stimulating factor (G-CSF), granulocyte-macrophage colony-stimulating factor (GM-CSF), and SR1 (Boitano et al., 2010). BM 49f cells plated in SR1+ re-entered the cell cycle more rapidly than in the week 1/STIF cocktail, although with slower kinetics than the PB 49f cells (Figure 2A, bottom).

Individual BM and PB 49f cells from two different individuals plated in the SR1+ cocktail could be reprogrammed at an average rate of 20.3 ± 1.9% and 24.5 ± 1.8%, respectively, in four independent experiments (Figure 2B). Together, these data suggest that BM 49f cells plated in SR1+ are as amenable to reprogramming as PB 49f cells plated in the same conditions.

To determine whether the cell cycle status of the cells at the time of Sendai virus infection affected reprogramming efficiency,

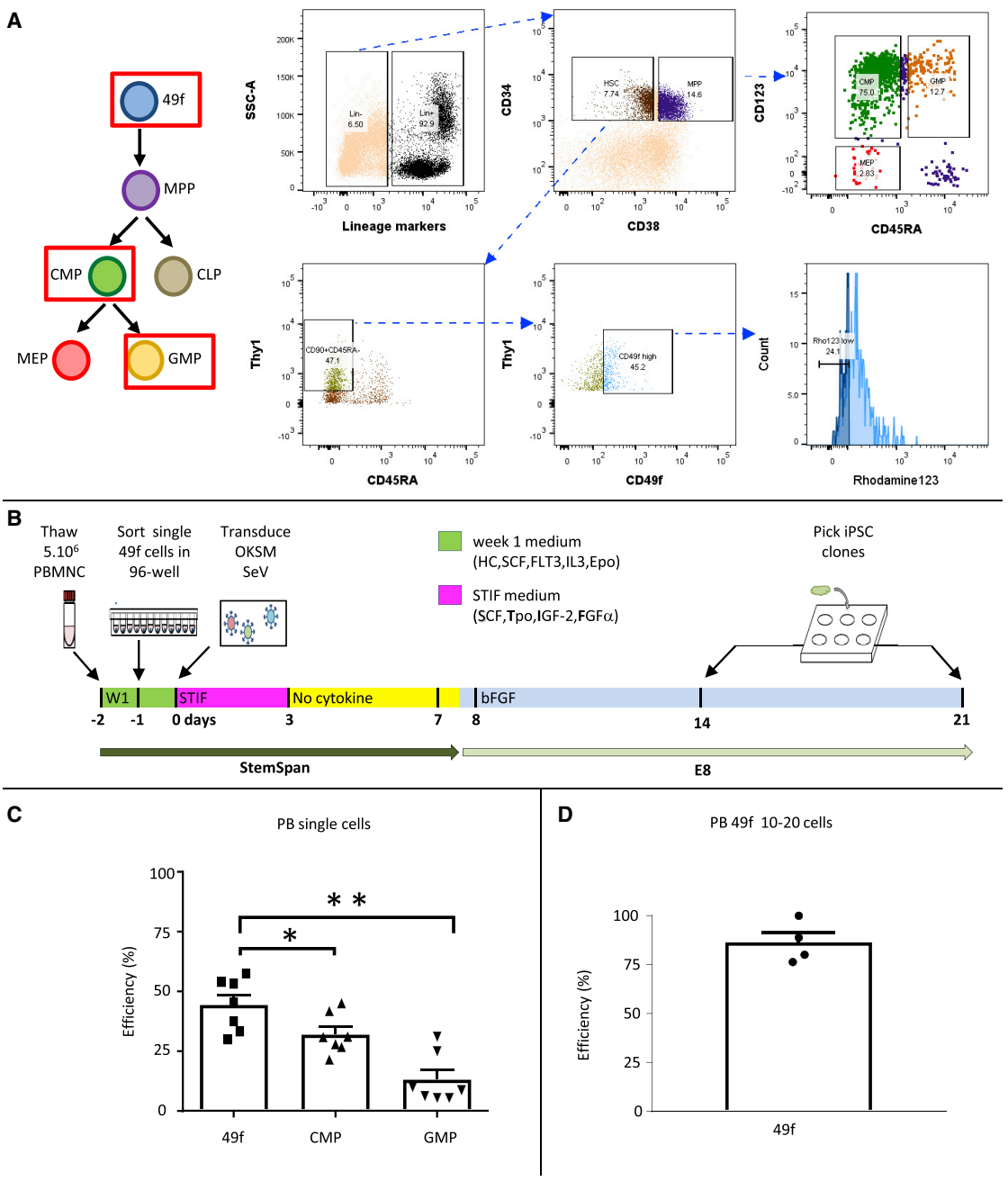


Figure 1. Reprogramming of PB 49f Cells

(A) Diagrams illustrating the sorting strategy to isolate PB 49f cells, CMPs, and GMPs. Mononuclear cells were stained with a cocktail of lineage antibodies, CD34, CD38, CD123, CD45RA, Thy1, and CD49f, and with Rhodamine 123 and sorted as single cells in round-bottom 96-well plates. Dead cells were gated out using DAPI staining.

(B) Individually sorted cells were transduced with Sendai virus on day 0, incubated with cytokines until day 3, without cytokines from day 3 to day 7, and switched to iPSC growing condition. OKSM, Oct4, KLF4, Sox2, Myc. MPP, multipotent progenitors. CLP, common lymphoid progenitor.

(C) Bar graph illustrating the efficiency of reprogramming of PB 49f cells, CMPs, and GMPs transduced as described in (B). The results of experiments in which 20–40 individual 49f cells, CMPs, or GMPs from seven different individuals were transduced are summarized. 49f cells can be reprogrammed more efficiently than CMPs and GMPs. The error bars represent the SEM. * t test, $p < 0.05$; **t test, $p < 0.01$. Efficiency of reprogramming was assessed 3 weeks after transduction using alkaline phosphatase staining and confirmed by multiple assays (see text).

(D) Individual PB 49f cells were sorted in individual wells and allowed to divide three to five times before being transduced as above. In these conditions nearly 100% of the wells yielded multiple colonies of iPSCs demonstrating that all the 49f cells that survive the sorting are competent to be reprogrammed. Two experiments in which either 10 or 20 49f cells were sorted are summarized.

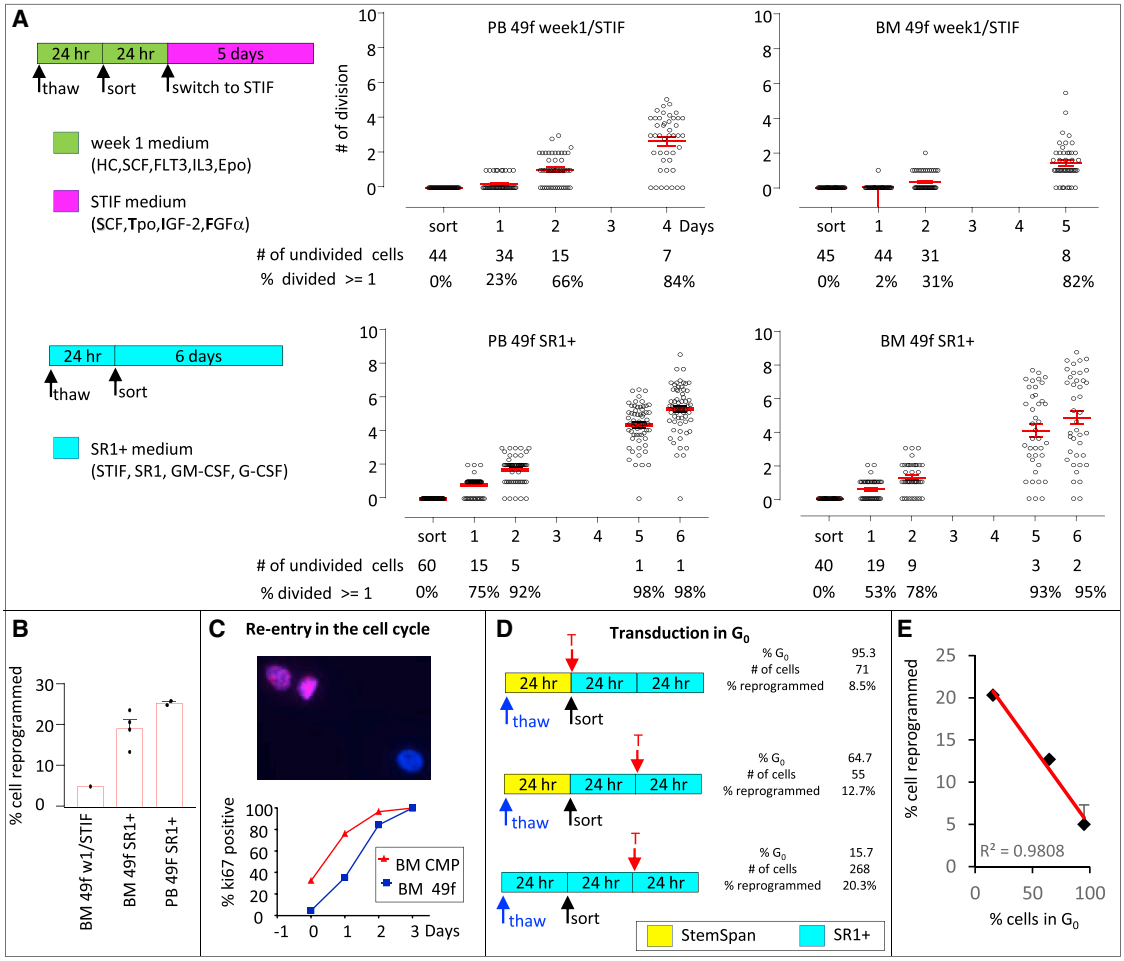


Figure 2. Reprogramming of BM 49f Cells

(A) Latency to first division is longer for BM than for PB 49f cells. PB and BM 49f cells were sorted in individual wells as described in Figure 1 and placed in either the week 1/STIF or SR1+ conditions (see STAR Methods), and cell division was monitored daily. Almost all 49f-positive cells eventually divided in either media but with different kinetics. PB 49f cells divided with a shorter latency than BM 49f cells. SR1+ medium reduced the latency period to first division for both BM and PB 49f cells. The diagrams on the left summarizes the culture conditions; the graphs on the right summarizes the results of two experiments. x axis, numbers of days after sort; y axis, number of cell divisions calculated as the squared root of the number of cells in the well. Each circle represents one well. The first row below the x axis represents the number of wells containing a single undivided cell; the row below, the percentage of wells in which the cells have divided at least once. The graph summarizes the results of two experiments performed on cells from two different individuals. Twenty to 40 cells were sorted in each experiment.

(B) Left bar illustrates the reprogramming efficiency of BM 49f cells transduced as described in Figures 1B and 1C (week 1 medium for 2 days followed by 3 days in STIF, with transduction at day −1). Middle and right bars illustrate reprogramming efficiency of BM (middle) and PB (right) 49f cells grown for 3 days in SR1+ with transduction on day 1. The graph summarizes the results of two experiments performed on cells from two different individuals. Twenty to 40 cells were tested in each condition.

(C) Top: micrograph illustrating 49f cells after Ki67 (pink) and DAPI staining (blue), 1 day after having been placed in SR1+. The two cells on the upper left are positive for ki67 and have therefore re-entered the cell cycle. The cell on the lower right is negative for Ki67 and is therefore in G_0. Bottom: scatter-plots illustrating the percentage of BM CMPs and 49f cells in G_0 as a function of time after plating in SR1+. Cells were stained with Ki67 and DAPI shortly after, or 1, 2, and 3 days after sorting. Between 50 and 100 cells from two different individuals were counted at each time point. BM 49f cells were almost all in G_0 at sorting time and re-entered the cell cycle slower than CMPs.

(D) BM 49f cell reprogramming efficiency and cell cycle re-entry. BM mononuclear cells were thawed, 49f cells were sorted and transduced with Sendai viruses encoding the reprogramming factors in three different conditions as shown in the graphs. The blue, black, and red arrows indicate thaw, sort, and transduction (T) times, respectively. The percentage of cells in G_0 at the time of transduction, the number of cells tested, and the percent reprogrammed cells (assessed by alkaline phosphatase staining) are indicated on the right. Cells from two individuals were tested in each conditions.

(E) Graph illustrating the inverse correlation between the percentage of BM 49f cells in G_0 at the time of transduction and the reprogramming efficiency. x axis, percentage of cells in G_0; y axis, reprogramming efficiency.

we first measured the timing of the exit from G_0 of BM 49f cells using CMPs as controls. KI67 staining revealed that 95% of freshly sorted BM 49f cells, that had been thawed and cultured for 24 h in StemSpan prior to sorting were in G_0, and that the proportion of cells in G_0 decreased to 64.7% and 15.7% when the 49f cells were exposed to SR1+ for 24 and 48 h. By contrast,

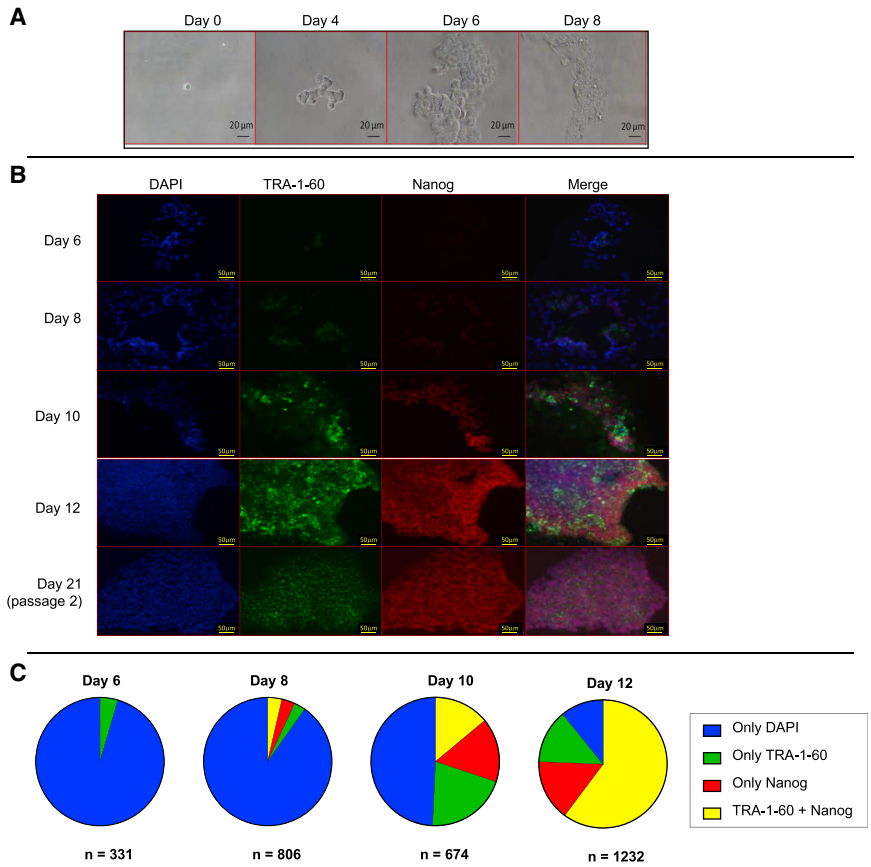


Figure 3. Morphology and Immunostaining of 49f Cells in the Process of Being Reprogrammed

(A) Morphology of 49f cells during reprogramming: PB 49f cells were transduced with Sendai viruses in a well of a 96-well plates and transferred to Matrigel-coated chamber slides at day 3. Bright-field micrograph illustrating 49f cells 4, 6, and 8 days after transduction. Colonies are detectable as soon as day 4, but adopt a typical iPSC phenotype at around days 14–21 (not shown).

(B) Kinetics of reprogramming: PB 49f cells treated as in Figure 3A were fixed and stained on day 6, 8, 10, 12, and 21 with DAPI, TRA-1-60-FITC, and Nanog-PE. The first three columns illustrate the individual stains, and the fourth column the merged images of the first three columns. TRA-1-60 and Nanog are detectable at day 6. By day 8 and until day 14, patches of cells express either TRA-1-60, Nanog, or both. Homogeneous staining pattern typical of fully reprogrammed iPSCs, in which almost every cell expressed both markers, was only detected after the first passage at day 21.

(C) Pie chart illustrating the number of cells expressing no marker, only TRA-1-60, only Nanog, or both markers. The number of cells counted at each time point is indicated below the chart, which represents the sum of two independent experiments performed with cells from two different individuals. In most colonies, the majority of cells express only one of the two markers until day 12, suggesting that the cells might not all follow the same path to de-differentiate from a 49f cell to an iPSC.

68%, 24%, and 5% of controls CMPs were in G_0 after 0, 24, and 48 h in the same conditions (Figures 2C and 2D).

To determine whether BM 49f cells in G_0 could be reprogrammed, we thawed the BM of two different individuals in either StemSpan without cytokines, or in SR1+ and transduced with Sendia viruses sorted 49f cells either unexposed, or exposed to SR1+ for 24 or 48 h (Figure 2D). This revealed a highly significant inverse correlation ($r^2 > 0.98$) between the proportion of cells in G_0 and the efficiency of reprogramming (Figure 2E), strongly suggesting that BM 49f cells are more efficiently reprogrammed when the transduction occurs after their exit from G_0.

Kinetics of Reprogramming of PB 49f Cells

Microscopic observation of PB 49f after transduction with Sendai viruses revealed that cells with an iPSC morphology typically appeared 8 days after transduction and that large iPSC colonies were present by day 14 to 21 (Figure 3A). To gain additional insight into the reprogramming process, we assessed the expression of transcription factor Nanog, and of TRA-1-60, a cell surface marker of pluripotent cells, by immunocytochemistry. This revealed that both proteins became detectable in a fraction of the cells at days 6–8 post-infection and that most cells contained either one or both proteins in highly irregular patterns by day 12. Regular staining patterns with all cells containing both proteins in similar amounts could only be observed at day 21 (Figures 3B and 3C). Therefore, the reprogramming of 49f cells is not homogeneous, since even within the same colony, some cells re-expressed TRA-1-60 before Nanog while other re-expressed Nanog before TRA-1-60.

Number of Somatic Single-Nucleotide Variants in 49f Cells as Compared to Fibroblasts

The data above suggest that 49f cells might be an excellent source of cells to produce clinical-grade iPSCs because a high frequency of reprogramming should be associated with more uniform iPSCs, but the high incidence of clonality in HSCs raised concerns about the number of somatic mutations in these cells (Genovese et al., 2014; Jaiswal et al., 2014). Since a critical consideration for clinical iPSCs is the number of somatic mutations they contain, and since there was no systematic data on the somatic mutation load in 49f cells as compared to skin fibroblasts, we decided to collect paired skin biopsies and PB samples from healthy volunteers and to directly measure the somatic mutation load in these two cell types by sequencing. Since PB can be obtained less invasively than BM, they are theoretically a preferable source of cells to produce iPSCs, particularly for individuals with pathologies that complicate BM aspiration. To compare the somatic variant burden of 49f cells obtained from these two sources, we also collected paired BM aspirates and PB samples from volunteer donors.

Table 1. Characteristics of the Healthy Volunteers Who Contributed Samples to the Study

Sample ID	Age	Sex	Ethnicity	Sample	sFib	PB 49f	BM 49f
					Number of Pairs Sequenced		
SB02	31	F	Asian	PB/sFib	5	3	–
SB06	32	M	Caucasian	PB/sFib	3	3	–
SB08	32	F	Hispanic	PB/sFib	3	3	–
SB10	31	F	Caucasian	PB/sFib	5	4	–
NY22	53	M	Caucasian	PB/sFib	3	3	–
L1	24	M	Black	PB/BM	–	3	3
L2	18	M	Black	PB/BM	–	3	4
L3	23	M	Caucasian	PB/BM	–	5	4
Total					19	27	11

sFib, skin Fibroblasts; PB, peripheral blood; BM, bone marrow.

Large-Fragment Reduced Representation Sequencing

The major difficulty in detecting somatic variants is their rarity (Hoang et al., 2016; Kinde et al., 2011; Quispe-Tintaya et al., 2016; Zong et al., 2012). To measure the burden of somatic variants in individual 49f cells, we devised a sib-pair approach similar to previously published methods based on sequencing of clonal populations (Abyzov et al., 2017; Behjati et al., 2014; Franco et al., 2018). This approach relies on creating two sibling clonal populations by sorting single cells in individual wells, allowing them to divide once, and splitting and culturing the two resulting cells in two different wells (Figure S3A). To reduce the cost of the approach, we sequenced about 10% of the genome of 57 sib-pairs (Tables 1 and S1) using our large-fragment reduced representation approach (Figure S3A), which maintains a balanced fractional representation of functional genomic annotations (Figure S3B).

Somatic variants unique to a sib-pair were identified as the variants that are present in the two members of the same sib-pair but absent in other sib-pairs derived from the same individuals. Averages of 33.4 ± 3.8, 41.1 ± 2.4, and 46.5 ± 4.07 somatic single-nucleotide variants (SNVs) were respectively found in BM 49f, PB 49f, and skin fibroblasts (Tables S2A–S2C). Nine out of nine randomly selected somatic SNVs (Figures S4A and S4B) were then successfully validated by sequencing of PCR fragments, demonstrating the acccuracy of our sib-pair approach.

Additional somatic SNVs present in more than one sib-pair were also identified in two of the five individuals tested (Table S2D). In SB02, 4 out of 5 fibroblasts share 29 somatic SNVs and 2 of these 4 fibroblasts shared 3 additional somatic SNVs. In SB10, all 5 fibroblasts studied shared 41 somatic SNVs that were absent in the blood cells from the same individual, and subgroups of fibroblasts shared additional SNVs. We interpreted these somatic variants common to more than one sib-pairs as variant that arose in a common ancestor of these fibroblasts. Figure S5 illustrates the putative clonal history of the skin fibroblasts in the biopsy of these two individuals. No clonal variant were found in any of the blood cells.

Number of Somatic SNVs Is Higher in Fibroblasts Than in 49f Cells

The number of unique and clonal somatic SNVs in each sib-pair were added together and extrapolated to the whole genome revealing that there were averages of 225.3 ± 23.4, 330.1 ± 32.45, and 517.6 ± 35.3 somatic SNVs/genome in the BM and PB 49f cells and in the skin fibroblasts, respectively (Figure 4; Table S2). There was no significant difference in the frequency of somatic SNVs between the BM and PB 49f cells, but the fibroblasts had more than twice as many somatic SNVs as the 49f cells ($p < 0.001$; Figure 4A). In addition to this variability across tissue, the number of somatic SNVs varied between 55 and 866 SNVs/genome in the 57 cells analyzed. This variability was present within and between each cell type (Figures 4A, S6A, and S6B; Table S2).

Number of Somatic SNVs Increases with Age

Although the age range of our donors was relatively narrow, analysis of the number of somatic SNVs in 49f cells as a function of the age of the donors revealed a strong positive correlation (Pearson $r^2 = 0.804$) (Figure 4B). Since the correlation was driven in part by the oldest individual, we confirmed this result by analyzing a smaller RNA-sequencing (RNA-seq) dataset that contains pairs of sib-clones derived from individual cells that we have generated for an expression study that will be published later. This analysis of 12 cells from three individuals 13, 53, and 55 years of age confirmed the observations that there are more somatic SNVs in hematopoietic cells from older than from younger individuals (Figures S6E and S6F; Tables S3A and S3B).

Modeling of the results of the combined RNA-seq and reduced representation genomic sequencing studies using multiple regression analysis or quasi-Poisson regression demonstrated that the correlation between age and number of somatic SNVs was significant and independent of the call rate and the fraction of genome sequenced ($p < 0.0001$). Together, these data demonstrate that 49f cells from healthy individuals acquire on average about 11.7 somatic SNVs/year (the slope of the regression curve).

Skin Fibroblasts Contain About 1.6-fold More Somatic Indels Than 49f Cells

A similar analysis revealed that the number of indels was significantly higher in skin fibroblasts than in their paired PB 49f cell counterparts (t test $p = 0.046$) or than in all 49f cells analyzed (t test $p = 0.002$) and averaged 17.5 ± 3.9, 17.1 ± 2.4, and

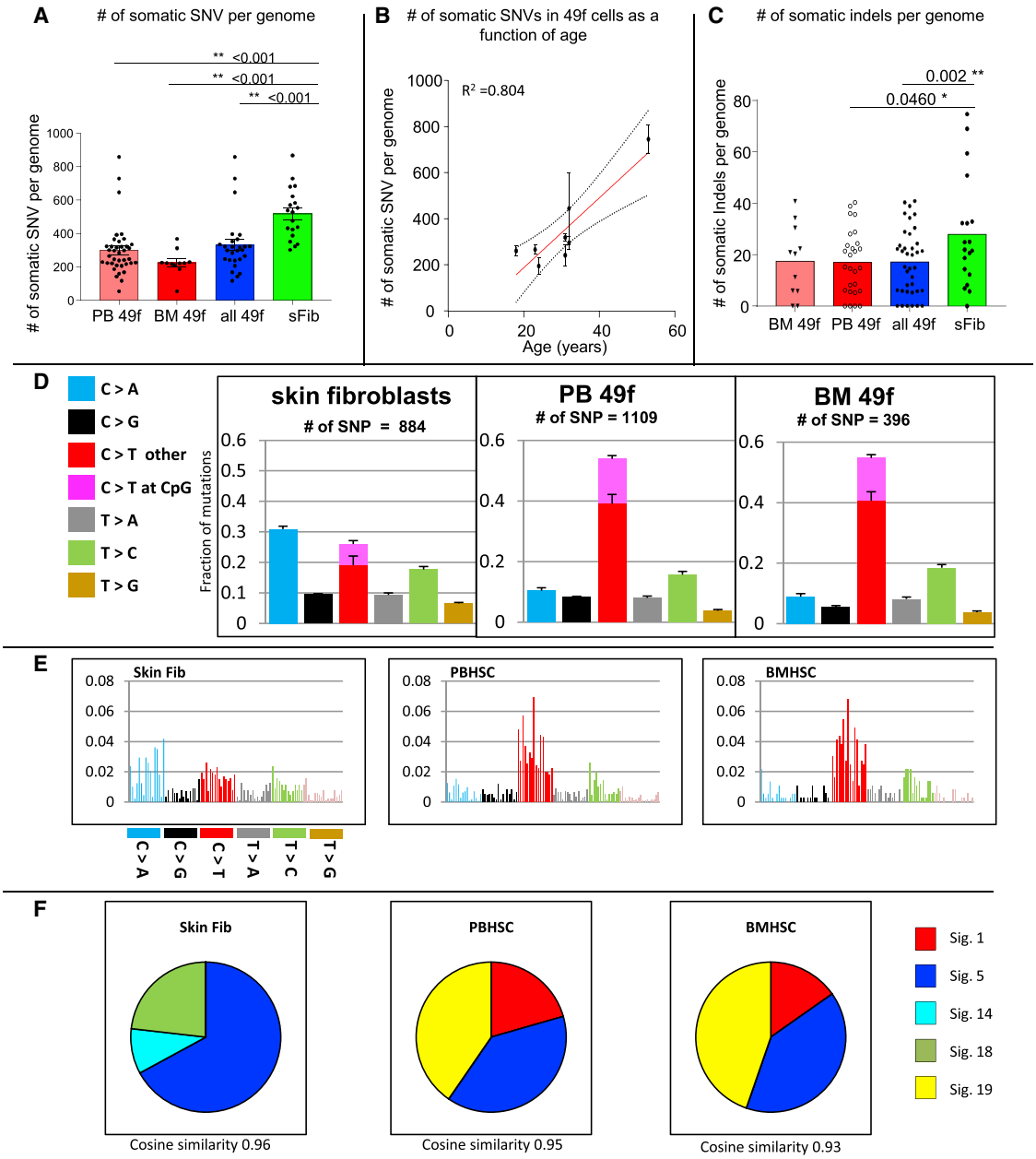


Figure 4. Number of Somatic Variants in 49f Cells and in Skin Fibroblasts

(A) Bar graph illustrating the number of somatic SNVs per genome in BM 49f cells, PB 49f cells, all 49f cells, and skin fibroblasts. The bars represent the average number of SNVs per genome; the error bars, the SEM. The inverted triangles, the open and closed circles, and the diamonds represent individual cells. The number of somatic SNVs observed in PB, BM 49f, or all 49f cells is significantly higher than in fibroblasts (t test, $p < 0.001$ in all cases). y axis, number of SNV per genome. The number of cells in each category is depicted in Table 1.

(B) Correlation between the number of somatic SNVs/genome in 49f cells and the age of the donors. x axis, age of each individual. y axis, average (±SEM) of the number of somatic SNVs/genome observed in all the 49f cells analyzed for each individual. Red line and dotted lines: best fit and 95% confidence intervals of the best-fit line.

(C) Bar graph illustrating the number of somatic indels per genome in BM 49f cells, PB 49f cells, all 49f cells, and skin fibroblasts. Graph is organized as in (A). Results suggest that there are significantly more indels in fibroblasts than in 49f cells.

(D) Unique somatic variant spectra in skin fibroblasts, PB, and BM 49f cells. Spectra of the PB 49f and BM 49f cells are very similar to each other but very different from that of the skin fibroblasts. The y axis represents the proportion of each mutation type. The proportion of C-to-T mutations at CpG and at other dinucleotides are shown separately.

(legend continued on next page)

28.4 ± 3.9 per genome for BM 49f, PB 49f, and skin fibroblasts (Figures 4C, S6C, and S6D; Tables S4A–S4C). Eight out of eight indels tested could be successfully validated by PCR fragment sequencing (Figures S4C and S4D). No clonal somatic indels could be detected.

Different Variant Spectra in 49f Cells and Skin Fibroblasts

Analysis of the SNVs unique to each cells with the R/Bioconductor package MutationalPatterns revealed that the PB and BM 49f cells had almost identical mutation spectra, which were very different from the skin fibroblast spectrum, as the skin fibroblasts had more than three times as many C>A transversions than the 49f cells, while there were more C>T transitions in the 49f cells (Figure 4D). Finer analysis at the individual and at the clone levels revealed that the mutation spectra of all blood cells and of all fibroblasts were respectively similar to the average patterns for each tissue (Figures S7A and S7B), suggesting that all cells from a given tissue acquire somatic mutations by similar mechanisms. In all three cell types about 26% of the C>T transversions occurred at CpG dinucleotides, suggesting that DNA methylation does not explain the difference in the variant spectrum between 49f cells and skin fibroblasts.

Variant Signature Analysis

To gain further insight into the mutational processes that are responsible for the accumulation of somatic SNVs in both skin fibroblasts and 49f cells, we decomposed these variants spectra further by analyzing the nucleotides on each side of the mutated bases, and comparing these variant profiles to the 30 published cancer signatures (Alexandrov et al., 2013a, 2013b). Initial analysis with all signatures revealed that seven signatures could account for all three variant profiles. To refine the analysis, we performed a second pass in which only these seven signatures were provided to the non-negative matrix factorization (NNMF) algorithm. This revealed that the 49f cells profiles could be decomposed into signatures 1, 5, and 19, and the skin fibroblast profiles into signatures 5, 14, and 18 (Figures 4E and 4F).

Structural Variants

Multiple somatic structural variants (SVs) were detected in the sib-pairs using DELLY (Rausch et al., 2012) (Table S5), but none could not be validated by PCR (Figure S4E). They were therefore not analyzed in detail.

Distribution and Impact of the Somatic Variants

SnpEff (De Baets et al., 2012; Reumers et al., 2006) analysis classified 0.78% and 0.60% of the somatic SNVs and 2.52% and 4.63% of the somatic indels that we detected in 49f and skin fibroblasts, respectively, as having a moderate or high impact on gene function (Table 2). Extrapolation of these numbers to the whole genome (see STAR Methods) suggests that iPSCs generated from PB 49f cells obtained from a 20-year-old individual would contain an average of 1.4 potentially deleterious variants (1.2 SNVs and 0.2 indels) that pre-existed in the donor cells, prior to the reprogramming. Since, in these cells, the number of SNVs increase by about 11.7 per genome and per year, the number of potential deleterious variants would increase by about 1.2 variants per genome per decade of age, yielding about 6.2 potentially deleterious variants in the genome of a 60 year old. IPSCs generated from skin fibroblasts obtained from a 20-year-old individual would contain an average of 2.9 potentially deleterious pre-existing variants (2.0 SNVs and 0.9 indels). Assuming the same rate of variants accumulation with age, that number would increase to about 6.9 potentially deleterious variants/genome in a 60 year old.

Variants Caused by Reprogramming

To assess the number of variants acquired during reprogramming, we then measured the number of somatic variants in three pairs of sib-iPSCs derived from HSPCs ($CD34^+CD38^-$) and from low-passage fibroblasts from individual NY22 (Figure 5A). The newly generated fib-iPSCs exhibited an average of 912.8 ± 114 somatic SNVs/genome and 38.6 ± 10.3 somatic indels/genome (Table S6; Figure 5B), which was about 2-fold higher (t test $p = 0.0192$) than the 451.3 ± 41.1 SNVs/genome and the 28.4 ± 3.9 indels/genome found in un-reprogrammed fibroblasts from the same individual at the same passage number (about 4). The number of SNVs and indels in HSPC-iPSCs was not significantly different from the number of SNVs in PB 49f cells from the same individual (804.1 ± 10.7 versus 745 ± 61 for SNVs and 5.6 ± 5.6 versus 16.4 ± 10.4 for indels) (Table S6; Figure 5B).

Together, these data suggest that Sendai virus reprogramming per se is not highly mutagenic in either 49f cells or skin fibroblasts, although it led to a statistically significant 2-fold increase in the number of somatic SNVs found fib-iPSCs as compared to the parent fibroblasts, perhaps due to the longer reprograming process for this cell type.

Variant spectrum analysis revealed that low-passage iPSCs had variant spectra similar to their cell of origin (Figure 5C). Together with the observation that the number of somatic variants in iPSCs was similar to that of the donor cells, this suggested that most of the somatic variants in iPSCs were already present in the donor cells, prior to reprogramming.

DISCUSSION

We report here that single 49f cells can be reprogrammed into iPSCs at a frequency close to 50% and that these extremely high frequencies were specific to 49f cells since CMPs, GMPs, and bulk $CD34^+$ cells exhibit progressively lower reprogramming efficiencies. 49f cells, which differ from progenitors by their

(E) Somatic mutation profile: each signature is displayed according to the 96 substitution classification defined by the substitution class, and by the sequence context immediately 3′ and 5′ to the mutated base. The mutation profiles of PB 49f and BM 49f cells are very similar to each other but very different from that of skin fibroblasts.

(F) Pie chart illustrating the decomposition of the somatic mutation profiles of skin fibroblast, PB, and BM 49f cells into mutational signatures. Skin fibroblasts can be decomposed into signatures 5, 14, and 18, while PB and BM 49f cells can be decomposed into signatures 1, 5, and 19. The cosine similarity of the reconstituted profile is indicated below the graph.

Table 2. Somatic Variant Impact

	SNP		Indels	
	49f	Fib	49f	Fib
Number of variants analyzed	1,476	884	96 (INS:16; DEL:71)	64 (INS:19;DEL:45)
Number of Effects by Impact				
High	3 (0.06%)	0 (0%)	8 (2.524%)	6 (2.317%)
Moderate	34 (0.71%)	17 (0.60%)	0 (0%)	6 (2.317%)
Low	61 (1.28%)	56 (1.96%)	1 (0.31%)	1 (0.39%)
Modifier	4,661 (97.94%)	2,777 (97.44%)	308 (97.16%)	246 (94.98%)
Summary				
Number of effects	4,759	2,850	317	259
% High or moderate impact	0.78%	0.60%	2.52%	4.63%
Number of Effects by Type and Region				
3′-UTR variant	38 (0.8%)	19 (0.66%)	11 (3.47%)	1 (0.39%)
5′-UTR premature start codon gain	2 (0.04%)	0 (0%)	0 (0%)	0 (0%)
5′-UTR variant	10 (0.21%)	29 (1.00%)	5 (1.58%)	38 (14.67%)
TF binding site variant	5 (0.10%)	0 (0%)	0 (0%)	0 (0%)
Downstream gene variant	485 (10.19%)	302 (10.43%)	28 (8.83%)	0 (0%)
Intergenic region	732 (15.38%)	406 (14.03%)	37 (11.67%)	34 (13.13%)
Frameshift variant	0 (0%)	0 (0%)	8 (2.52%)	6 (2.32%)
Conservative in frame insertion	NA	NA	0 (0%)	6 (2.32%)
Intron variant	2,716 (57.06%)	1,787 (61.75%)	179 (56.47%)	130 (50.19%)
Missense variant	32 (0.67%)	17 (0.59%)	0 (0%)	0 (0%)
Nc transcript exon variant	67 (1.41%)	25 (0.86%)	1 (0.31%)	4 (1.54%)
Sequence feature	19 (0.4%)	11 (0.38%)	0 (0%)	1 (0.39%)
Upstream gene variant	611 (12.84%)	246 (8.50%)	47 (14.83%)	39 (15.06%)
Splice region variant	1 (0.02%)	44 (1.52%)	1 (0.31%)	0 (0%)
Start lost	1 (0.02%)	0 (0%)	0 (0%)	0 (0%)
Stop gained	2 (0.042%)	0 (0%)	0 (0%)	0 (0%)
Synonymous variant	38 (0.8%)	8 (0.276%)	0 (0%)	0 (0%)
Coding Region Variants				
Missense	33	17	NA	NA
Nonsense	2	0	NA	NA
Synonymous	38	8	NA	0
Frameshift variant	NA	NA	8	6
Conservative in-frame insertion	NA	NA	0	6

High-impact variants: SNPs or indels that cause nonsense, missense in rare amino acid, frameshift, loss of start/stop, or splice donor/acceptor variants. Moderate-impact SNPs or indels: variants that are within splice regions or that cause missense in common amino acid or in-frame codon insertions/deletion.

potency, their self-renewal capacity, and their higher level of quiescence (Laurenti et al., 2015), could only be reprogrammed efficiently after they had exited G_0. They are therefore more efficiently reprogrammed than progenitors, despite the fact that they are more quiescent. This suggests that the reprogramming susceptibility of the 49f cells is an intrinsic property of these cells associated with their greater self-renewal and/or differentiation potential.

Reprogramming of primary human fibroblasts has clearly been shown to be stochastic. Reprogramming of 49f cells is much more efficient than that of skin fibroblasts but is not completely deterministic because we observed some cases of abortive reprogramming and because the re-expression of Nanog and TRA-1-60 was heterogeneous within developing iPSC colonies that had been initiated from single cells. This latter observation strongly suggests that a 49f cell can follow different routes to reach the stable iPSC pluripotent state. The high-efficiency reprogramming of 49f cells described here might become a powerful experimental tool to analyze the reprogramming process at the single-cell level.

Comparison of paired blood samples and skin biopsies demonstrated that 49f cells contain on average about 2-fold fewer somatic SNVs and indels than skin fibroblasts, but the number of somatic SNVs per cell varied more than 15-fold within

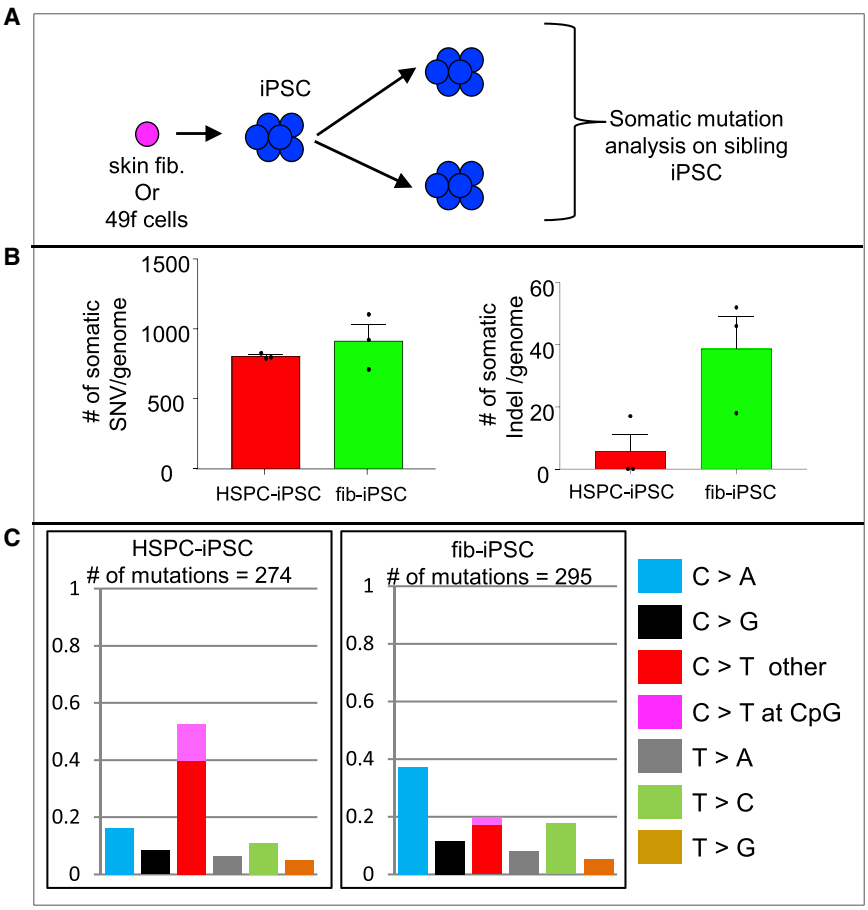


Figure 5. Number of Somatic SNVs in HSPC-iPSCs and in fib-HSPCs

(A) Experimental strategy to produce sibling pairs of iPSCs from skin fibroblasts or HSPCs. iPSCs were generated from skin fibroblasts or 49f cells, and single clones were isolated at passage 2 and split into multiple sib-clones. Two of these sib-clones were then expanded and sequenced as described in the text.

(B) Bar graphs illustrating the number of somatic SNVs (left) and indels (right) in three sib-pairs of HSPC-iPSCs and three sib-pairs of fib-iPSCs from the same individual.

(C) Variant spectra of HSPC-iPSCs and fib-HSPCs are very different from each other but similar to those of their respective donor cells (shown in Figure 4D). The y axis represents the proportion of each mutation type. The proportion of C-to-T mutations at CpG and at other dinucleotides are shown separately.

each cell type, even in the same individual. We therefore conclude that the tissue of origin contributes to the number of somatic mutations but less than the individual history of each cell.

We detected 84 somatic mutations present in multiple sib-pair of fibroblasts in two 31-year old individuals but none in paired blood cells (Table S2D). Our failure to detect clonality in blood cells is not surprising because clonality in the hematopoietic system is generality found in people older than our cohort. The high observed rate of clonality in skin fibroblasts might reflect the greater mutation rate in this cell type, but might also reflect low skin fibroblast tissue mobility since our observations suggest that multiple ancestors and their progeny carrying increasing number of somatic mutations resided, presumably for many years, in proximity to each other at the site of the 3-mm skin biopsies that were performed to collect the cells.

We observed a strong correlation between the number of somatic variants and age and calculated a rate of SNV acquisition of 11.7 somatic SNVs/year genome-wide for 49f cells which is in agreement with the results of Welch et al. (2012) in hematopoietic cells. This rate of somatic SNV acquisition is similar to the 13 SNV/year recently reported in muscle satellite cells (Franco et al., 2018) but about 3.5 times lower than what was reported for adult liver, colon, and intestinal stem cells (Blokzijl et al., 2016). Although these rates are difficult to compare because they were obtained with different approaches, they suggest that 49f cells contain fewer somatic mutations than other adult stem cells. Importantly, the rate of SNV acquisition/year in adult 49f cells is about an order of magnitude higher than in the germline (Campbell and Eichler, 2013; Conrad et al., 2011; Kong et al., 2012; Milholland et al., 2017; Rahbari et al., 2016; Wang et al., 2012), suggesting that 49f cells are devoid of some of the mechanisms that protect germ cells from somatic mutations.

Signatures 1 and 5 are associated with aging, accounted for about 60% of the SNVs in 49f cells, and are also major signatures in the germline and in adult somatic stem cells from muscle, colon, and intestine (Alexandrov et al., 2015; Blokzijl et al., 2016; Franco et al., 2018), suggesting that similar mutational processes operate in most stem cells. The remaining 40% of the SNV in 49f cells were associated with signature 19, which is not found in other adult stem cells, suggesting that there might be mutational processes specific to the HSC niche. The biological process underlying signature 19 is not known.

The skin fibroblasts mutation profiles were associated with signatures 5, 14, and 18 in agreement with a previous study (Abyzov et al., 2017). Signature 18 has been found in liver organoid derived from adult stem cells but has also been associated with cell culture (Blokzijl et al., 2016). Therefore, a fraction of the SNVs that we detected in fibroblasts might have been acquired during the short culture period that was technically necessary to obtain pure skin fibroblasts before we generated the clones. The lack of detection of a UV signature in the skin fibroblasts that we analyzed likely reflects the fact that we collected them from the lower back, a body site that is not generally exposed to the sun.

Analysis of somatic variants in fresh iPSCs shortly after reprogramming demonstrated that the reprogramming process per se is not highly mutagenic. Combined with the observation that the variant spectra in low-passage iPSCs are very similar to that of the donor cells, this suggests that the vast majority of variants in low-passage iPSCs pre-exist in the donor cells, confirming previous reports (Young et al., 2012).

We propose that PB 49f cells are ideal candidates for the production of clinical-grade iPSCs. PB seems preferable to BM because it is easier to collect and because 49f cells from both tissue have similar variant burden. There are very few 49f cells in PB, but we were nevertheless able to generate iPSCs from this cell source in more than 90% of 18 independent attempts. CMPs, which are less efficiently reprogrammed but more frequent in the PB, are also potentially good donor cells.

We detected potentially deleterious pre-existing variants in 49f cells but in reasonably small numbers. These variants are therefore unlikely to be an obstacle to the use of 49f-iPSCs for regenerative medicine purposes, particularly if the donor cells are collected from young individuals. iPSC sequencing could prove useful to identify lines with particularly low mutation burden since the number of somatic variant varies dramatically between cells of the same tissue origin.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human subjects
- METHOD DETAILS
  - Sendai virus-mediated single cell reprogramming: PB HSPC reprogramming
  - Ki-67 staining
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Alignment
  - Somatic variant identification and normalization
  - Germline SNVs
  - CR estimation
  - Identification of somatic SNVs unique to a single sib-pair
  - Extrapolation to the whole genome
  - Indel calling
  - Structural variants
  - Variant spectrum and variant signature analysis
  - Identification of somatic SNVs common to multiple cells (clonality analysis)
  - impact of the somatic variants
  - Statistics
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and six tables and can be found with this article online at https://doi.org/10.1016/j.celrep.2019.02.021.

ACKNOWLEDGMENTS

E.E.B., K.W., Z.Y., S.Z., B.B., and E.E.B. were supported by NYSTEM grants C030135 and C029154; NIH grant HL130764; and Doris Duke Foundation grant 2017087. M.B.H. and Y.-K.Y. are supported by the Intramural Research Program of the NIH, National Library of Medicine. We thank Daqian Sun and Swathi-Rao Narayanagari from the Stem Cell Flow Cytometry and Xenotransplantation Facility for expert help with flow sorting and xenotransplantation. We thank the Einstein Flow Cytometry Core Facility for expert help on flow sorting. We thank Rob Durbin from the Einstein Epigenomic Core for help with bioinformatics.

AUTHOR CONTRIBUTIONS

K.W., S.Z., Z.Y., and J.Z. performed the experiments and contributed to experimental design. A.K.G., S.T., and H.E.K. provided samples and contributed to manuscript writing. M.B.H., Y.-K.Y., K.Y., and M.Y.H. contributed to data analysis. B.B. contributed to data analysis and manuscript writing. E.E.B. supervised the project and contributed to experimental design, data analysis, and manuscript writing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 22, 2018
Revised: December 18, 2018
Accepted: February 6, 2019
Published: March 5, 2019

REFERENCES

Abyzov, A., Mariani, J., Palejev, D., Zhang, Y., Haney, M.S., Tomasini, L., Ferrandino, A.F., Rosenberg Belmaker, L.A., Szekely, A., Wilson, M., et al. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature *492*, 438–442.

Abyzov, A., Tomasini, L., Zhou, B., Vasmatzis, N., Coppola, G., Amenduni, M., Pattni, R., Wilson, M., Gerstein, M., Weissman, S., et al. (2017). One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. Genome Res. *27*, 512–523.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013a). Deciphering signatures of mutational processes operative in human cancer. Cell Rep. *3*, 246–259.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013b). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407.

Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D.C., Tamuri, A.U., Martincorena, I., Petljak, M., Alexandrov, L.B., Gundem, G., et al. (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature *513*, 422–425.

Bhutani, K., Nazor, K.L., Williams, R., Tran, H., Dai, H., Džakula, Ž., Cho, E.H., Pang, A.W.C., Rao, M., Cao, H., et al. (2016). Whole-genome mutational burden analysis of three pluripotency induction methods. Nat. Commun. *7*, 10536.

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., et al. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. Nature *538*, 260–264.

Boitano, A.E., Wang, J., Romeo, R., Bouchez, L.C., Parker, A.E., Sutton, S.E., Walker, J.R., Flaveny, C.A., Perdew, G.H., Denison, M.S., et al. (2010). Aryl

hydrocarbon receptor antagonists promote the expansion of human hematopoietic stem cells. Science *329*, 1345–1348.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell *150*, 1209–1222.

Campbell, C.D., and Eichler, E.E. (2013). Properties and rates of germline mutations in humans. Trends Genet. *29*, 575–584.

Chen, G., Gulbranson, D.R., Hou, Z., Bolin, J.M., Ruotti, V., Probasco, M.D., Smuga-Otto, K., Howden, S.E., Diol, N.R., Propson, N.E., et al. (2011). Chemically defined conditions for human iPSC derivation and culture. Nat. Methods *8*, 424–429.

Conrad, D.F., Keebler, J.E.M., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., et al.; 1000 Genomes Project (2011). Variation in genome-wide mutation rates within and between human families. Nat. Genet. *43*, 712–714.

De Baets, G., Van Durme, J., Reumers, J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., Schymkowitz, J., and Rousseau, F. (2012). SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res. *40*, D935–D939.

Eminli, S., Foudi, A., Stadtfeld, M., Maherali, N., Ahfeldt, T., Mostoslavsky, G., Hock, H., and Hochedlinger, K. (2009). Differentiation stage determines potential of hematopoietic cells for reprogramming into induced pluripotent stem cells. Nat. Genet. *41*, 968–976.

Franco, I., Johansson, A., Olsson, K., Vrtačnik, P., Lundin, P., Helgadottir, H.T., Larsson, M., Revêchon, G., Bosia, C., Pagnani, A., et al. (2018). Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. Nat. Commun. *9*, 800.

Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487.

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

Hermann, A., Kim, J.B., Srimasorn, S., Zaehres, H., Reinhardt, P., Schöler, H.R., and Storch, A. (2016). Factor-reduced human induced pluripotent stem cells efficiently differentiate into neurons independent of the number of reprogramming factors. Stem Cells Int. *2016*, 4736159.

Hoang, M.L., Kinde, I., Tomasetti, C., McMahon, K.W., Rosenquist, T.A., Grollman, A.P., Kinzler, K.W., Vogelstein, B., and Papadopoulos, N. (2016). Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. Proc. Natl. Acad. Sci. USA *113*, 9846–9851.

Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., Zhao, T., Ye, J., Yang, W., Liu, K., et al. (2013). Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. Science *341*, 651–654.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498.

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

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature *488*, 471–475.

Laurenti, E., Frelin, C., Xie, S., Ferrari, R., Dunant, C.F., Zandi, S., Neumann, A., Plumb, I., Doulatov, S., Chen, J., et al. (2015). CDK6 levels regulate quiescence exit in human hematopoietic stem cells. Cell Stem Cell *16*, 302–313.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Manz, M.G., Miyamoto, T., Akashi, K., and Weissman, I.L. (2002). Prospective isolation of human clonogenic common myeloid progenitors. Proc. Natl. Acad. Sci. USA *99*, 11872–11877.

Merling, R.K., Sweeney, C.L., Choi, U., De Ravin, S.S., Myers, T.G., Otaizo-Carrasquero, F., Pan, J., Linton, G., Chen, L., Koontz, S., et al. (2013). Transgene-free iPSCs generated from small volume peripheral blood nonmobilized $CD34^+$ cells. Blood *121*, e98–e107.

Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., and Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. Nat. Commun. *8*, 15183.

Miyoshi, N., Ishii, H., Nagano, H., Haraguchi, N., Dewi, D.L., Kano, Y., Nishikawa, S., Tanemura, M., Mimori, K., Tanaka, F., et al. (2011). Reprogramming of mouse and human cells to pluripotency using mature microRNAs. Cell Stem Cell *8*, 633–638.

Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat. Biotechnol. *26*, 101–106.

Notta, F., Doulatov, S., Laurenti, E., Poeppl, A., Jurisica, I., and Dick, J.E. (2011). Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. Science *333*, 218–221.

Olivier, E., Qiu, C., and Bouhassira, E.E. (2012). Novel, high-yield red blood cell production methods from CD34-positive cells derived from human embryonic stem, yolk sac, fetal liver, cord blood, and peripheral blood. Stem Cells Transl. Med. *1*, 604–614.

Olivier, E.N., Marenah, L., McCahill, A., Condie, A., Cowan, S., and Mountford, J.C. (2016). High-efficiency serum-free feeder-free erythroid differentiation of human pluripotent stem cells using small molecules. Stem Cells Transl. Med. *5*, 1394–1405.

Peterson, S.E., and Loring, J.F. (2014). Genomic instability in pluripotent stem cells: implications for clinical applications. J. Biol. Chem. *289*, 4578–4584.

Plath, K., and Lowry, W.E. (2011). Progress in understanding reprogramming to the induced pluripotent state. Nat. Rev. Genet. *12*, 253–265.

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. https://doi.org/10.1101/201178.

Quispe-Tintaya, W., Gorbacheva, T., Lee, M., Makhortov, S., Popov, V.N., Vijg, J., and Maslov, A.Y. (2016). Quantitative detection of low-abundance somatic structural variants in normal cells by high-throughput sequencing. Nat. Methods *13*, 584–586.

Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Turki, S.A., Dominiczak, A., Morris, A., Porteous, D., Smith, B., et al.; UK10K Consortium (2016). Timing, rates and spectra of human germline mutation. Nat. Genet. *48*, 126–133.

Rais, Y., Zviran, A., Geula, S., Gafni, O., Chomsky, E., Viukov, S., Mansour, A.A., Caspi, I., Krupalnik, V., Zerbib, M., et al. (2013). Deterministic direct reprogramming of somatic cells to pluripotency. Nature *502*, 65–70.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.

Reumers, J., Maurer-Stroh, S., Schymkowitz, J., and Rousseau, F. (2006). SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. Bioinformatics *22*, 2183–2185.

Schlaeger, T.M., Daheron, L., Brickler, T.R., Entwisle, S., Chan, K., Cianci, A., DeVine, A., Ettenger, A., Fitzgerald, K., Godfrey, M., et al. (2015). A comparison of non-integrating reprogramming methods. Nat. Biotechnol. *33*, 58–63.

Soufi, A., Donahue, G., and Zaret, K.S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell *151*, 994–1004.

Taapken, S.M., Nisler, B.S., Newton, M.A., Sampsell-Barron, T.L., Leonhard, K.A., McIntire, E.M., and Montgomery, K.D. (2011). Karotypic abnormalities

in human induced pluripotent stem cells and embryonic stem cells. Nat. Biotechnol. *29*, 313–314.

Takahashi, K., and Yamanaka, S. (2016). A decade of transcription factor-mediated reprogramming to pluripotency. Nat. Rev. Mol. Cell Biol. *17*, 183–193.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

Wang, J., Fan, H.C., Behr, B., and Quake, S.R. (2012). Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell *150*, 402–412.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell *150*, 264–278.

Young, M.A., Larson, D.E., Sun, C.-W., George, D.R., Ding, L., Miller, C.A., Lin, L., Pawlik, K.M., Chen, K., Fan, X., et al. (2012). Background mutations in parental cells account for most of the genetic heterogeneity of induced pluripotent stem cells. Cell Stem Cell *10*, 570–582.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., et al. (2007). Induced pluripotent stem cell lines derived from human somatic cells. Science *318*, 1917–1920.

Zhang, C.C., Kaba, M., Ge, G., Xie, K., Tong, W., Hug, C., and Lodish, H.F. (2006). Angiopoietin-like proteins stimulate ex vivo expansion of hematopoietic stem cells. Nat. Med. *12*, 240–245.

Zhang, M., Zhang, Y., Scheuring, C.F., Wu, C.-C., Dong, J.J., and Zhang, H.-B. (2012). Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. Nat. Protoc. *7*, 467–478.

Zhao, Y., Zhao, T., Guan, J., Zhang, X., Fu, Y., Ye, J., Zhu, J., Meng, G., Ge, J., Yang, S., et al. (2015). A XEN-like state bridges somatic cells to pluripotency during chemical reprogramming. Cell *163*, 1678–1691.

Zong, C., Lu, S., Chapman, A.R., and Xie, X.S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science *338*, 1622–1626.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD2 Monoclonal Antibody (RPA-2.10), PE-Cyanine5	ThermoFisher	Cat# 15-0029-42; RRID: AB_10736743
CD3 Monoclonal Antibody (UCHT1), PE-Cyanine5	ThermoFisher	Cat# 15-0038-42; RRID: AB_10598354
CD4 Monoclonal Antibody (S3.5), TRI-COLOR	ThermoFisher	Cat# MHCD0406; RRID: AB_10392548
CD7 Monoclonal Antibody (CD7-6B7), TRI-COLOR	ThermoFisher	Cat# MHCD0706; RRID: AB_10373996
CD8 Monoclonal Antibody (3B5), TRI-COLOR	ThermoFisher	Cat# MHCD0806; RRID: AB_10372207
CD10 Monoclonal Antibody (eBioCB-CALLA (CB-CALLA)), PE-Cyanine5	ThermoFisher	Cat# 15-0106-42; RRID: AB_10596518
CD14 Monoclonal Antibody (TuK4), TRI-COLOR	ThermoFisher	Cat# MHCD1406; RRID: AB_10373566
CD19 Monoclonal Antibody (HIB19), PE-Cyanine5	ThermoFisher	Cat# 15-0199-42; RRID: AB_10853658
CD20 Monoclonal Antibody (2H7)	ThermoFisher	Cat# 15-0209-42; RRID:AB_10548510
CD56 Monoclonal Antibody (MEM-188), TRI-COLOR	ThermoFisher	Cat# MHCD5606; RRID: AB_10372520
CD38 Monoclonal Antibody (HIT2), PE-Cyanine7	ThermoFisher	Cat# 25-0389-42; RRID: AB_1724057
CD90 (Thy-1) Monoclonal Antibody (eBio5E10 (5E10)), Biotin	ThermoFisher	Cat# 13-0909-82; RRID: AB_763525
CD45RA Monoclonal Antibody (HI100), Super Bright 600	ThermoFisher	Cat# 63-0458-42; RRID: AB_2688186
CD49f Monoclonal Antibody (eBioGoH3 (GoH3)), PE	ThermoFisher	Cat# 12-0495-82; RRID: AB_891474
CD45 Monoclonal Antibody (HI30), APC	ThermoFisher	Cat#17-0459-42; RRID: AB_10667894
CD45.1 Monoclonal Antibody (A20), eFluor 450	ThermoFisher	Cat# 48-0453-82; RRID: AB_1272189
CD33 Monoclonal Antibody (WM-53 (WM53)), PE	ThermoFisher	Cat# 12-0338-42; RRID: AB_10855036
CD19 Monoclonal Antibody (HIB19), FITC	ThermoFisher	Cat# 11-0199-42; RRID: AB_10669461
TRA-1-81 (Podocalyxin) Monoclonal Antibody (TRA-1-81), PE	ThermoFisher	Cat# 12-8883-82; RRID:AB_891606
TRA-1-60 (Podocalyxin) Monoclonal Antibody (TRA-1-60), PE	ThermoFisher	Cat#12-8863-82; RRID: AB_891602
Anti-Ki67 antibody	Abcam	Cat# ab15580; RRID: AB_443209
PE-Cy5 Mouse Anti-Human CD235a	BD Biosciences	Cat# 561776; RRID: AB_10894595
APC Mouse Anti-Human CD34	BD Biosciences	Cat# 555824; RRID: AB_398614
BV421 Streptavidin	BD Biosciences	Cat# 563259
PE Mouse Anti-Human CD123	R&D Systems	Cat# 555644
PE Mouse anti-SSEA-4	R&D Systems	Cat# 560128
PE Rat anti-SSEA-3	R&D Systems	Cat# 560237
Anti-TRA-1-60 Antibody	Millipore	Cat# MAB4360
Nanog (D73G4) XP Rabbit mAb	Cell Signaling Technology	Cat# 4903
Human alpha -Fetoprotein/AFP Antibody	Cell Signaling Technology	Cat# MAB1369
Human/Mouse/Rat alpha -Smooth Muscle Actin Antibody	Cell Signaling Technology	Cat# MAB1420
Neuron-specific beta -III Tubulin Antibody	Cell Signaling Technology	Cat# MAB1195
Bacterial and Virus Strains		
CytoTune-iPS Sendai Reprogramming	ThermoFisher	Cat# A16517
Biological Samples		
Unprocessed Human BM	Lonza	Cat#: 1M-105
Unprocessed Human Peripheral Blood	Lonza	Cat#: 1W-500
Chemicals, Peptides, and Recombinant Proteins		
StemSpan SFEM II	Stem Cell Technologies	Cat#: 09655
DMEM/F12	ThermoFisher	Cat#: 11320033
Matrigel	Corning	Cat#: 354234
DPBS	ThermoFisher	Cat#: 14190250
hydrocortisone	Sigma-Aldrich	Cat#: H4001

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SCF	PeproTech	Cat#: 300-07
FLT3	PeproTech	Cat#: 300-19
IL3	PeproTech	Cat#: 200-03
TPO	PeproTech	Cat#: 300-18
IL6	PeproTech	Cat#: 200-06
StemRegenin 1 (SR1)	Cayman	Cat#:10625
GM-CSF	PeproTech	Cat#: 300-03
G-CSF	PeproTech	Cat#: 300-23
IGF-2	Alfa Aesar	Cat#:J65170
FGF-acidic	PeproTech	Cat#:100-17A
EPO	AMGEN	Cat#: 55513-0126-10
L-Ascorbic acid	Sigma-Aldrich	Cat#: A4403
Sodium Selenite	Sigma-Aldrich	Cat#: S9133
Sodium Chloride	Sigma-Aldrich	Cat#: S5886
DAPI	ThermoFisher	Cat#: D1306
Fetal Bovine Serum	ThermoFisher	Cat#: 10437028
Holo-transferrin	Sigma-Aldrich	Cat#: T066
Critical Commercial Assays		
VECTOR Red Alkaline Phosphatase (Red AP) Substrate Kit	Vector Laboratories	Cat#: SK-5100
AP live stain	ThermoFisher	Cat#: A14353
NEBNext Ultra II DNA Library preparation kit	New England Biolabs	Cat#: E7845L
Deposited Data		
300 fastq and 27 *vcf.* have been deposited to SRA	PRJNA511439	NA
Experimental Models: Organisms/Strains		
NOD.Cg-Prkdc^{scid}.Il2rg^{tm1Wjl}/SzJ	Jackson Lab	005557 | NSG
Software and Algorithms		
Bwa (version 0.7.10, MEM algorithm)	SourceForge	http://bio-bwa.sourceforge.net/bwa.shtml
GATK (version 3.8)	GATKForum	https://gatkforums.broadinstitute.org/gatk/discussion/11188/gatk-version-3-8-download
DELLY (version 0.7.5)	GitHub	https://github.com/brewsci/homebrew-science/blob/master/Formula/delly.rb
SnpEff (version 2.0)	SourceForge	http://snpeff.sourceforge.net/features.html
MATLAB package SigProfiler	MathWorks	https://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler
relaimpo R package	CRAN R-Project	https://cran.r-project.org/web/packages/relaimpo/index.html
R function glm in R package lme4	CRAN R-Project	R function glm in R package lme4
MutationalPatterns	Bioconductor	http://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Eric Bouhassira (eric.bouhassira@einstein.yu.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

Skin specimens were acquired by punch biopsy. All participants agreed to the procedure with written informed consent. An area of gluteal skin (lower back which is generally not exposed to sun) with no dermatologic disease was identified, sterilized and locally

anesthetized. A circular blade was rotated downward through the epidermis and dermis, and into the subcutaneous fat, yielding a 4-mm cylindrical core of tissue sample. Hemostasis was achieved with the placement of two 4-0 nylon sutures. The patient was instructed to return to the clinic for suture removal in 7 days. The biopsies were placed in culture, expanded for three passages and multiple aliquots were frozen at passages 3. Human samples were collected under Eisntein IRB approved protocol 2008-201

Freshly collected BM aspirates and PB samples from healthy individuals were purchased from Lonza (Basel, Switzerland) and were shipped at 4°C overnight. PB and CB blood samples were also obtained in-house under IRB-approved protocols. Mononuclear cells were prepared from the PB and BM samples using Histopaque as recommended by the manufacturer (Sigma-Aldrich, St Louis, MO) and residual red blood cells were lysed by incubation in ice-cold buffer containing 790mg/L of ammonium bicarbonate and 7.7g/L ammonium chloride. Multiple aliquots were frozen. Samples were processed no later than 36 hours after collection.

The age and gender of all subject are reported in Table 1 of the main text.

Mouse

Twenty-four (12 males and 12 females) 8-12 week old NSG mice (NOD.Cg-Prkdcscid.Il2rgtm1Wjl/SzJ) mice acquired from the Jackson Lab (cat # 005557) were used. Mouse experiments were performed under animal protocol # 20141205 approved by the Einstein Animal Institute.

METHOD DETAILS

Sendai virus-mediated single cell reprogramming: PB HSPC reprogramming

The workflow of single cell reprogramming is illustrated in Figures 1A and 1B. Frozen mononuclear cells were thawed the day before sorting (day −2) and allowed to recover overnight in Week1 medium (StemSpan SFEM (Stem Cell Technologies, Vancouver, Canada), hydrocortisone 1uM, SCF, 50 ng/mL, FLT3 16.6 ng/mL, IL3 6.6 ng/mL, Epo 1.3U/mL), a culture medium that is based on the Week1/STIF pulse protocol that we described previously (Olivier et al., 2016). About 18 to 24 hours after thawing, individual 49f cells, CMP or GMP were then flow-sorted using a FACSAria II (BD) at 1 cell/well into 96-well round bottom plates (Nunc) in 80 μL of Week 1 medium (day −1). The next day (day 0), sorted cells were examined under the microscope, and wells containing single live cells were selected for reprogramming using CytoTune-iPS 2.0 Sendai Reprogramming Kit (Invitrogen). Transductions were conducted by adding $5X10^3$ CIU of hKOS, $5X10^3$ CIU of hc-MYC, $3X10^3$ CIU of hKLF4 and 4 μg/mL of Polybrene (Millipore) into each selected well in 30 μL of STIF medium (Zhang et al., 2006) (StemSpan SFEM, 5ng/ml SCF, 10ng/mL Tpo, 10 ng/mL IGF-2, 5 ng/mL FGFα). The plates were then centrifuged at 2250 rpm for 90 minutes at room temperature, and cultured overnight after addition of additional 30 μL of STIF medium. On day 1, the medium with the viruses was replaced by fresh STIF medium. On day 3, the transduced cells were transferred into 24-well plates coated with Matrigel in 0.5 mL of StemSpan SFEM medium without cytokines and cultured for 4 days with a medium change on day 5. On day 7 of transduction, the culture medium was changed to half StemSpan SFEM medium (without cytokines) and half Essential 8 iPSC culture medium. Cells were then transitioned to complete Essential 8 iPSC culture medium on day 8. From day 9 to day 21, the transduced cells were checked daily for the emergence of iPSC-like colonies. Essential 8 iPSC culture medium was changed daily. The iPSC-like colonies positive for Alkaline Phosphatase live staining were passaged mechanically once the size reached 5 mm diameter and then passaged with 0.5 mM EDTA as reported previously (Chen et al., 2011). Twenty to forty wells each containing one cells were tested in each experiment.

BM 49f reprogramming

The protocol was as described above for PB HSPCs except that cells were thawed at day −2, sorted at day −1, infected on day 0, and culture until day 3 in SR1+ media (StemSpan SFEM, SR1 0.75 μM, TPO 10 ng/mL, SCF 50 ng/mL, FLT3 50 ng/mL, IL6 20 ng/mL, GM-CSF 20 ng/mL, G-CSF 20 ng/mL) instead of the Week1/STIF media.

Fibroblast reprogramming

was performed as recommended by the manufacturer of the Cytotune 2.0 kit. Single cell reprogramming was unsuccessful because the frequency was too low. Efficiency of reprogramming was therefore measured by infecting 1000,000 cells. Frozen vials at passages 3-4 were thawed and allowed to grow for a few days before the reprogramming experiments.

Isolation of 49f cells, CMP and GMP

PBMCs were stained with Rhodamine123 (100ng/mL), washed, resuspended in PBS with 2% FBS and subsequently stained with a panel of fluorophore-conjugated monoclonal antibodies (See Key Resource Table). 49f cells are defined as (Lin-CD34+CD38-Thy1+CD45RA-CD49f+Rhodamine123low) (Notta et al., 2011), CMPs as (Lin-CD34+CD38+CD45RA-CD123+) (Manz et al., 2002), and GMP as (Lin-CD34+CD38+CD45RA+CD123+). An average of 86 ± 21 (n = 16) 49f cells, 1769 ± 606 CMPs (n = 7) and 237 ± 67 GMPs (n = 7) could be recovered per 10 mL of blood.

Alkaline Phosphatase staining

Alkaline Phosphatase (AP) staining was used to identify the iPSC clones 14 to 21 days after Sendai virus-mediated reprogramming. The AP live stain (Molecular Probes, Eugene, OR) was conducted on clones selected for further propagation and characterization according to the manufacturer's protocol. The Vector Red Alkaline Phosphatase substrate (Vector Laboratories, Burlingame, CA) was used to stain all other putative iPSC clones to calculate the reprogramming efficiency. In the latter case, cells were fixed with 4% paraformaldehyde for 5 minutes and then incubated with the substrate working solution prepared in 100 mM Tris-HCl at pH 8.5 for 20 minutes at room temperature. The result was checked under microscope after removing the substrate and washing cells with PBS. Colonies with strong red stain in more than 50% of cells were counted as healthy iPSC clones.

49f xenotransplantation and assessment

Single cord blood (CB) 49f cells or pools of cells were sorted into a 6x12 wells Terasaki plate in 5 μL StemSpan medium. The plate was spun down at 1000 rpm for 5 minutes and checked under the microscope to confirm the presence of single live cells in each well. The cells were then resuspended by gentle pipetting and loaded into a 27xxx G 0.5 mL tuberculin syringe preloaded with 5μL StemSpan medium. One to eighty 49f cells were injected intra-femorally on both sides of irradiated 6-week old NSG mice under anesthesia (NOD.Cg-Prkdc^{scid}.Il2rg^{tm1Wjl}/SzJ, Jackson Lab). The 49f xenograft was examined at 8, 16 and 24 weeks after injection by flow cytometry, by aspirating about 5 μL BM from each side of the femur of the transplanted mice.

FACS analysis of iPSCs and Xenotransplants

To examine the expression of pluripotent cell surface markers, cells were dissociated with 0.05% trypsin/EDTA, washed, and resuspended in PBS with 2% FBS. The cells (1×10^5/ml) were then stained for 30 minutes at 4°C with saturating amounts of PE-conjugated monoclonal antibodies against human, SSEA-4(BD), TRA-1-60 and TRA-1-81 (eBioscience).

To examine the 49f cell xenograft, the red blood cells from the BM from the recipient NSG mouse were lysed and the white blood cells were stained for 30 minutes at 4°C with saturating amounts of monoclonal antibodies recognizing human CD45, human CD19, human CD33 and mouse CD45.1 (eBioscience). All antibodies are described in the Key Resource Table. All the analyses were done on a Becton Dickinson FACSCalibur laser flow cytometric system (BD Biosciences).

Immunohistochemistry

For the detection of pluripotent markers, transduced LT-HSCs were transferred into matrigel (Corning)-coated 8-well chamber slides on day 3 of transduction, and cultured in the medium according to the reprogramming protocol up to the desired time points. For the three germ-layer differentiation detection, embryoid bodies (EBs) were formed using the hanging drop method and spontaneously differentiated in DMEM supplemented with 20% FBS and L-glutamine in a chamber slide for 10 days. Cells were subsequently fixed, permeabilized and stained. All primary and secondary antibodies are described in the Key Resource Table. Labeled slides were counter-stained with DAPI and mounted in ProLong Gold Antifade (Thermo Fisher, Waltham, MA) prior to visualization under a fluorescence microscope (ZEISS, AxioVert 200M).

Teratoma formation

80% confluent iPS cells from a well of a six-well plate were dissociated with 0.5 mM EDTA at room temperature and re-suspended in 100 μL iPSC medium. The cell suspension was mixed with equal volume of matrigel (BD Biosciences Franklin Lakes, NJ) and 10^6 cells were injected intramuscularly into the hind leg of a 6–8 week old NSG (NOD.Cg-Prkdc^{scid}.Il2rg^{tm1Wjl}/SzJ) mouse. After 6–12 weeks, the visible tumors were dissected and fixed in 10% formalin solution overnight. The tissues were paraffin-embedded, sectioned and stained with hematoxylin/eosin using standard procedures.

Generation of sibling clones

49f cells were sorted individually in 100 μL of week1 medium (or SR1+ medium, see above) in round-bottom 96-well plates and examined microscopically twice daily. When the cells had divided once, the cultures were split equally in 4 wells, spun and each well was examined microscopically to identify the two wells that contained a single cell. These two wells were then used to initiate the sib clonal culture using the following 3-week erythroid differentiation protocol. Cells were first incubated either in week1 medium for 48 hours (since sorting time) and then in STIF medium for five days (week1/STIF condition) or in SR1+ medium for seven days followed by a week in week1 and a week in week2 media (StemSpan SFEM (Stem Cell Technologies, Vancouver, Canada), hydrocortisone 1uM, SCF, 20 ng/mL, IL3 6.6 ng/mL, Epo 2U/mL) as described (Olivier et al., 2012). Cell culture volume was adjusted to keep the cell concentration below 10^6cells/ml at all times. In most cases, this resulted in a pair of sibling clonal populations containing each at least 100,000 basophilic erythroblasts. Cases in which one or both of the sibling populations expanded poorly were discarded.

Clonal sibling populations from skin fibroblasts were generated using a similar protocol but the cells were grown in RPMI 1640 containing 10% fetal bovine serum.

Ki-67 staining

Frozen BM mononuclear cells were thawed and allowed to recover overnight in StemSpan medium without any cytokines. Aliquots of 300 BM 49f cells and 500 CMPs were then sorted (in bulk) into wells of 96 well-plates and Ki67 expression was assessed by immunofluorescence on cells that were either freshly sorted, or incubated for various amounts of time in SR1+.

Cells were attached to glass slides using Smear Gell (SG-01, Diagnocine) and fixed with 4% paraformaldehyde according to manufacturer's protocol. After fixation, cells were permeabilized and blocked by 0.5% Triton X-100 and 6% donkey serum in PBS for 30 minutes and then stained with anti-Ki-67 rabbit polyclonal antibodies (abcam, ab15580, 1:200) in 0.1% Triton X-100 and 6% donkey serum in PBS at 4°C overnight. Secondary goat anti-rabbit IgG antibodies conjugated with Alexa Fluor 555 (Cell Signaling Technology, 4413S, 1:500) were then used to detect Ki-67. DNA was counterstained with DAPI and mounted in antifade solution. Cells were scored as KI-67 negative or positive. The number of cells available did not allow us to further sub-classify the Ki67 positive cells according to their position in the cell cycle based on the staining patterns.

Reduced representation sequencing

About 100,00 cells were encapsulated in agarose, genomic DNA was extracted, digested with PacI and DNA fragments 50-90 kb in size were isolated by pulse field electrophoresis as described in Zhang et al. (2012). Sequencing libraries with 350 bases inserts were then prepared using the low-input NEBNext® Ultra II DNA Library preparation kit and 150 base pair pair-end sequencing was performed on the Illumina platform to a depth of at least 12 Gb per library.

QUANTIFICATION AND STATISTICAL ANALYSIS

Alignment

Adaptor-trimmed reads were then aligned to the hg19 genome using Bwa (version 0.7.10, MEM algorithm) (Li and Durbin, 2010) and SNVs and indels were called using the GATK (version 3.8) following the recommended best practices (Poplin et al., 2017). To increase the precision of the calls, aligned reads from all the sub-clones derived from the same individual were grouped in a single BAM file and processed together. This resulted in the sequencing of an average of 3,495 ± 59 fragments sized between 50-90 kb per clone. These fragments covered an average (±SEM) of 10.69 ± 0.13% of the genome at a read depth of at least 10x and 6.09 ± 0.16% at a read depth of at least 20x (Table S1). Structural variants (SVs) were called using DELLY (version 0.7.5) (Rausch et al., 2012). To provide additional controls, genomic DNA from four individuals, extracted from 100,000 mononuclear cells was sequenced either completely or using the reduced-representation approach. In all cases, > 99% of the SNVs called as germline as described in methods, were validated in the high read depth regions of these controls, while none of the somatic SNV that we called could be found.

Somatic variant identification and normalization

Our workflow involved partially sequencing the genome of each individual studied 10 to 18 independent times (since we analyzed 5 to 9 cells per individual and since two sib-clones were generated for each cell). We took advantage of this high level of redundancy to identify somatic variants with a very high degree of precision, and to calculate a Call Rate (CR) based on the rate of detection of germline SNVs. The CR was used to correct for variation in read depth in each sequencing reaction.

Germline SNVs

The vast majority of the SNVs called by the GATK were expected to be present in the germline and therefore to be found in all the clones derived from the same individual. Germline SNVs can be heterozygous (GT = 0/1) if one chromosome carries the reference allele (REF) and one alternate allele (ALT), or homozygous (1/1) if both alleles were different from the hg19 assembly that we used as a reference. Rare cases where two different alternate alleles were detected were discarded.

Analysis of the results revealed that there were 632,139 ± 33,100 genomic positions where the GATK had called at least one SNV in one sib-clone and where the read depth was at least 5 for all single sib-clones of an individual. 177,349 ± 30,108 of those genomic positions had a read depth of at least 20 for all sib-clones (Table S1B). At 5x read depth, an average of 24.2% of the SNVs were universally called as 0/1 and 24.6% as 1/1; at 20x, 42.9% of the SNVs were universally called as 0/1 and 30.0% as 1/1 (Table S1B). Therefore and as expected, the number of universally called SNVs increased with the read depth, but even at 20x coverage for all sib-clones only 72.9% of all SNVs was uniformly called in all sib-clones. Further analysis revealed that at 20x coverage there were an additional 11.02 ± 1.52% of the positions where all but one sib-clone were called as heterozygous 0/1 and 2.16% where all but one clone was called as 1/1, bringing the total of likely germline SNVs to an average of at least 86.07% (Table S1B).

Close examination of the data revealed that genomic positions where the GATK had missed an allele in one of the sib-clones were most often regions of locally low read depth that originated in technical variations in the size of the PacI fragments purified by PFGE or in partial PacI digest in some samples. These low read depth regions resulted in false negative calls by the GATK for either the reference (leading to a 1/1 call), the alternate (leading to a 0/0 call), or both alleles (leading to a ./. call). We used these germline SNVs to estimate a Call Rate (CR) for each sib clone.

CR estimation

Since these regions of locally low read depth were clearly false negatives, we developed a method to mitigate their impact on our results without using overly stringent read depth criteria.

Each sib-clone X was assigned a CR defined as:

$$1 - \frac{\#\text{ of SNVs called as } 0/1 \text{ for all sib-clones except sib-clone X}}{\#\text{ of SNVs called } 0/1 \text{ for all sib-clones.}}$$

CR was then used to generate a corrected number of somatic variants for each sib-pair, by dividing the observed number of somatic variants (obtained as described below) by the product of CR_A and CR_B, the CR for the A and B clones of each sib-pair.

The CR coefficients for all sib-clones are shown in Table S2B. They were generally close to 1 between 0.8 and 0.99) because the PFGE procedure yielded fairly homogeneous fractions (see GenPlay project) and because the sequencing read depths were relatively similar.

Identification of somatic SNVs unique to a single sib-pair

To identify somatic SNVs, the v*cf.* files generated by the GATK were processed though the following pipeline:

1) Elimination of the low read depth SNVs (at least one sib-clone with more than 10 reads) and of SNVs that were in dbSNP (since somatic SNVs are not expected to be in dbSNP)
2) Elimination of SNVs with a GATK quality score < 150
3) Identification of putative somatic variants as SNVs called as 0/1 in both members of a sib-pair and as 0/0 in all other sib-clones (allowing for one uncalled sib-clone).
4) Elimination of the putative somatic variants for which the number of reads in either members of the mutated sib-pair was less than 5.
5) Elimination of the putative somatic variants for which the p value of a Fisher's exact test between the REF/ALT ratio of the putatively mutated sib-pair versus the aggregated REF/ALT ratio for all the other clones was > 0.01.
6) Elimination of the putative somatic variants in which at least one read containing the putative somatic variant was found in any of the non-mutated sub-clones.

This yielded a list of raw somatic variants for each pair of sib clones. The threshold for the quality score (150) the number of alternate reads (5) and the p value for the Fisher's exact test (0.01) were determined by minimizing the ratio of the number of reference versus alternate reads (REF/ALT). The REF/ALT ratio for the germline SNVs averaged 1.04 and was about 1.10 for the somatic variants (Table S1C).

Extrapolation to the whole genome

To estimate the number of somatic variants for the entire genome from this reduced-representation data, we then estimated the fraction of the genome that had been sequenced to a sufficient depth to call SNVs in both sib-clones for every pair of sib-clones. This fraction was estimated by dividing the genome into 5 kb windows and by selecting all windows in which a genotype call with a minimal REF and ALT read depth of at least 5 in each sib-clone of the pair had been made. Windows separated by less than 5 kb were then merged, and the merged windows smaller than 40 kb were eliminated (since we had isolated PacI fragments 50-90 kb in size). The cumulated length of these windows was termed the Genome Covered. The Genome Covered divided by the length of the genome (2.88×10^9 bases) was termed the Genome Fraction Sequenced (GFS) and was used to extrapolate the number of somatic SNVs per genome for each sib-pair by dividing the CR normalized number of SVs by GFS. The GFS for each sib-clones is shown in Table S2B.

Indel calling

Indels were processed separately from the SNVs but the same procedures were used. The ratio of reference/alternate reads averaged 1.21, slightly higher than for the somatic SNVs but still close to one (Table S1D).

Structural variants

Structural variants (SVs) were called using DELLY with default parameters. Putative variants were identified as SVs that were called as 0/1 in both members of a sib-clones pair and as 0/0 in all other sib-clones (allowing for two uncalled sib-clones). Putative somatic variants were then filtered by eliminating all variants that did not PASS the QUAL filter in both members of the sib-pair and that had a high-quality variant pairs (DV) + high-quality variant junction reads (RV) scores < 5 in both sib-pairs. The CR was calculated as for the SNVs. The GFS coefficients calculated for the SNVs were used to generalize the number of SVs to the whole genome.

Variant spectrum and variant signature analysis

Variant spectrum analysis was performed using R package ''MutationalPatterns.'' Variant signatures analysis was performed using the MATLAB package SigProfiler provided by the Alexandrov lab (Alexandrov et al., 2013a).

Identification of somatic SNVs common to multiple cells (clonality analysis)

In the algorithm above, SNVs are classified as germline if they are detected in all, or all-but-one sib-clone, and as somatic if they are found in the two sib-clones forming a sib-pair but in no other clones. SNVs that are not defined as germline or somatic are not classified at all. To detect SNVs common to more than one pair, which are indicative of clonality, we also searched for somatic SNVs present in two or more sib-pairs.

As above, we filtered out the very low read depth SNVs, the SNVs with a low quality score and the SNVs that were in DbSNP, and then searched for pairs, triplets, quadruplets and quintuplets of sib-pairs that were called as 0/1 while all other clones were called as 0/0, 0/. or ./. (allowing for one 0/1 call). The SNVs obtained were then further narrowed down by eliminating putative clonal SNVs that clustered in small hard to sequence regions, and those that exhibited either a low read depth specifically in all the non-mutated pairs but not in the mutated pairs, or that had a very skewed REF/ALT ratio (> than 1.4).

This algorithm yielded 20 to 200 putative clonal somatic SNVs per individual. To assess the significance of these putative clonal somatic SNVs, we created all possible combinations of pairs of sib-clones, and ran the algorithm described above several hundred fold for each individual. This demonstrated that the probability of observing one or two putative ''clonal somatic SNV'' by chance in

two or more pairs of random sib-pairs was > 0.1 in all cases but that groups of three or more putative clonal SNVs were almost never found by chance in random pairs ($p < 0.001$ in most individuals). We therefore defined clonal somatic SNVs as groups of three or more variants that fulfilled the above criteria and that were found in at least two sib-pairs. Clonal somatic variants were found exclusively in the fibroblasts of two of the individuals tested.

impact of the somatic variants

The somatic variants were analyzed using SnpEff (De Baets et al., 2012; Reumers et al., 2006) and default parameters.

Variants classified as high or moderate impact by SnpEff was considered potentially deleterious. The results were extrapolated to the whole genome for PB 49f cells and skin fibroblasts of different ages as follows. The average number of deleterious SNVs was estimated by multiplying the observed number of SNVs (330/genome for PB 49f cells and 517/genome for skin fibroblasts) by the proportion of deleterious SNVs (0.0078 in PB 49f cells or 0.0060 in skin fibroblasts).

Since the average age of the 19 cells sequenced from the 5 individuals from which we had both 49f cells and skin fibroblasts was 35 year old, we extrapolated the average number of deleterious SNVs for 20 year-old PB 49f as (330 –(35-20)*11.7))*0.0078 = 1.21 (using a rate of *de novo* somatic SNV acquisition of 11.7/year)

The number of deleterious indels was estimated by multiplying the number of observed indels (17.1 for PB 49f cells and 28.4 for skin fibroblasts) by the proportion of deleterious indels (0.0252 in PB49f cells or 0.0463 in skin fibroblasts 4.63).To extrapolate the results for 20 year old cells, we calculated the observed indel/SNP ratio (0.052 for 49f and 0.055 for skin fibroblasts) and multiplied that ratio by the predicted number of SNVs for a 20 year old. For instance the deleterious indels in PB 49f cells was calculated as (330-((35-20)*11.7))*0.052*0.0252 = 0.20

The total number of potentially deleterious variants was calculated by adding the numbers for SNPs and indels (1.21 + 0.20 = 1.41 in the case of 20-year old PB 49f cells). Similar calculations were performed for 60 year-old cells.

Statistics

Statistical calculations were generally performed in R. To analyze the association between the number of somatic mutations with age. We used quasi-Poisson regression (R function glm in R package lme4). The p value was calculated for the total number of somatic mutations in all cells from the same individual offset by the total genomic coverage multiplied by the CR of all cells from each individual.

To compare cell types, we applied mixed effect Poisson regression models using function glmer in R package lme4, in which subjects are treated as random effects. The response is the total number of somatic mutation of each cell, offset of its sequencing coverage multiplied by the CR. Models were fit to the SNVs and indels separately.

The relationship between somatic mutation rates and age was also analyzed using the relaimpo R package.

DATA AND SOFTWARE AVAILABILITY

300 Fastq and 26 v*cf.* files for all the experiments described are available at Short Read Archive: PRJNA511439.