Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations

Florian T. Merkle*[,1-5], Sulagna Ghosh*[,1-4], Nolan Kamitaki[3,6,7], Jana Mitchell[1-4], Yishai Avior[8], Curtis Mello[3,6,7], Seva Kashin[3,6,7], Shila Mekhoubad[1,2,4,9], Dusko Ilic[10], Maura Charlton[1-4], Genevieve Saphier[1,3,4], Robert E. Handsaker[3,6,7], Giulio Genovese[3,6,7], Shiran Bar[8], Nissim Benvenisty[8], Steven A. McCarroll[3,6,7], Kevin Eggan[1-4]

1. Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA
2. Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA
3. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
4. Harvard Stem Cell Institute, Cambridge, MA 02138, USA
5. Current address and affiliations: Metabolic Research Laboratories and Medical Research Council Metabolic Diseases Unit, Wellcome Trust-Medical Research Council Institute of Metabolic Science, and Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge CB2 0QQ, UK
6. Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
7. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
8. The Azrieli Center for Stem Cells and Genetic Research, Institute of Life Sciences, Hebrew University of Jerusalem, Givat-Ram, Jerusalem 91904, Israel
9. Current address and affiliation: Stem Cell Research, Biogen, 115 Broadway, Cambridge, MA 02142
10. Stem Cell Laboratories, Guy's Assisted Conception Unit, Division of Women's Health, Faculty of Life Sciences and Medicine, King's College London, London, UK

* These authors contributed equally

Human pluripotent stem cells (hPSCs) can self-renew indefinitely, making them an attractive source for regenerative therapies. This expansion potential has been linked with acquisition of large copy number variants (CNVs) that provide mutant cells with a growth advantage in culture[1-3]. However, the nature, extent, and functional impact of other acquired genome sequence mutations in cultured hPSCs is not known. Here, we sequenced the protein-coding genes (exomes) of 140 independent human embryonic stem cell (hESC) lines, including 26 lines prepared for potential clinical use[4]. We then applied computational strategies for identifying mutations present in a subset of cells[5]. Though such mosaic mutations were generally rare, we identified five unrelated hESC lines that carried six mutations in the *TP53* gene that encodes the tumor suppressor P53. Notably, the *TP53* mutations we observed are dominant negative and are the mutations most commonly seen in human cancers. We used droplet digital PCR to demonstrate that the *TP53* mutant allelic fraction increased with passage number under standard culture conditions, suggesting that P53 mutation confers selective advantage. When we then mined published RNA sequencing data from 117 hPSC lines we observed another nine *TP53* mutations, all resulting in coding changes in the DNA binding domain of P53. Strikingly, in three lines, the allelic fraction exceeded 50%, suggesting additional selective advantage resulting from loss of heterozygosity at the *TP53* locus. As the acquisition and favored expansion of cancer-associated mutations in hPSCs may go unnoticed during most applications, we suggest that careful genetic characterization of hPSCs and their differentiated derivatives should be carried out prior to clinical use.

Somatic mutations that arise during cell proliferation and are then subject to positive selection are major causes of cancer and other diseases[6]**.** Acquired mutations are often present in a subset of cells in a sample, and can therefore be detected in next generation sequencing data from their presence at allelic fractions less than 50%[5,7]. We reasoned that similar analysis of sequencing data from a large number of hESCs might reveal previously unappreciated mosaic mutations and mutation-driven expansions acquired during hESC culture at single-nucleotide resolution. This approach would complement previous studies describing culture-derived chromosomal-scale aneuploidies and megabase-scale CNVs in hPSCs[1,8,9].

To this end, we sought to collect and perform whole exome sequencing (WES) of hESC lines that were derived under appropriate informed consent and were readily available for distribution (Fig. 1a). We therefore turned to the registry of hESC lines maintained by the US National Institutes of Health (NIH) (Fig. 1b) and were able to obtain, bank, and sequence 114 independent hESC lines with minimal restrictions on use (Fig. 1c-e). We selected cell lines at low to moderate passage numbers (mean P18, range P3-P37) and cultured them in a common set of growth conditions for an average of 2.7 ± 0.7 (± STD) passages (range 2-6 passages) prior to banking and sequencing (Fig. 1f,g). This minimal time in culture ensured that at the time of sequencing, cell lines were at similar densities and growth rates and were broadly representative of those we received. Since hESC-derived differentiated cells are currently being studied in clinical trials for their safety and utility in a range of diseases such as macular degeneration[10], we also sequenced genomic DNA samples isolated directly from an additional 26 independent hESC lines that had been prepared under good manufacturing practice (GMP) conditions for potential clinical use (Fig. 1c,e,g). We sequenced the whole exomes of these 140 hESC lines from 19 institutions to a mean read depth of 79.7 ± 0.1 (± SEM) (range 57 for UM4-6 to 115 for UM78-2) (Fig. 1h). Further details on cell lines acquisition and selection can be found in Table S1 and Supplementary Methods.

To identify potentially acquired mutations, we examined the number of

sequencing reads at high-quality and high-coverage heterozygous sites across the exome. We observed that the allelic fractions for most variants followed a binomial distribution, reflecting statistical sampling around the 50% level expected of inherited alleles (Fig. 2a), as well as a much smaller set of sites where variant alleles were present at lower fractions (Fig. 2a,b). We applied a statistical test to identify variants for which the observed allelic fractions were unlikely (P<0.01 by binomial test) to have been generated by random sampling of two equally-present alleles. We found 263 candidate mosaic variants, of which 28 were predicted to have a damaging or disruptive effect on gene function (Table S2).

The only gene affected by more than one such mutation was the tumor suppressor gene *TP53. TP53* was affected by six of the 28 mutations, found in five unrelated hESC lines (Table S3), and one GMP-prepared cell line (MShef10) which carried two distinct *TP53* variants (G245S and R248W). These six missense mutations, while rare (<0.01%) in the general population[11] (Fig. 2c), mapped precisely to the four codons most frequently disrupted in human cancer (Fig. 2d, Table S3)[12-14]. Since P53 is mutated in approximately 50% of tumors[15], coding mutations in these four residues are associated with a substantial fraction of human cancer disease burden. Each of the six mutations involved a cytosine residue of a CpG dinucleotide and may therefore involve a highly mutable site[16].

On a crystal structure of the human P53 protein complexed with DNA, each of the mutations mapped to the DNA-binding domain of P53[17] (Fig. 2e,f). Mutations at these positions are strongly associated with cancer and are known to act in a dominant negative fashion to substantially diminish P53's regulation of apoptosis, cell cycle progression, and genomic stability[18]. Individuals with germ-line mutations at these residues develop Li-Fraumeni syndrome, an autosomal dominant disease with a lifetime cancer risk of nearly 100%[19]. In these patients, tumors can arise at any age and can affect most tissues, including the brain, bones, lung, skin, soft tissues, adrenal gland, colon, stomach, and blood[20].

To independently test the hypothesis that the inactivating *TP53* mutations were acquired during cell culture, we developed droplet digital PCR (ddPCR) assays to digitally count the abundance of each allele at the four *TP53* mutation sites (Fig. 3a,b, Table S4). Analysis of genomic DNA derived from the 140 hESC lines confirmed that all six mutations identified by WES were indeed mosaic, with allelic factions ranging from 7-40%, suggesting their presence in 14-80% of cells in culture (Fig. 3c). We did not identify additional cell lines carrying mutations at these positions, suggesting they were either absent or present at allelic fractions below the sensitivity of the assay (approximately 0.1%)[21]. These findings demonstrate that each of the *TP53* mutations identified in hESCs was an acquired mutation and that cells with the mutation had come to represent a significant fraction of all cells.

We next asked whether the cells harboring these *TP53* mutations expanded their representation within the hESC population across passages. To this end, we re-obtained early passage vials for hESC lines that were mosaic for *TP53* mutations (CHB11 at P22, and WA26 at P13), thawed a fresh vial of ESI035 at P36, and analyzed the genomic DNA from the frozen vial and at each subsequent passage to test for changes in mutant allelic fraction. In each of the three hESC lines, *TP53* mutant alleles increased in representation at each analyzed passage (Fig. 3d) in all but one experiment, suggesting these mutations conferred a strong selective advantage (approximately 1.9-fold/passage) under standard hESC culture conditions (Fig. 3e, Fig. S1, Table S5). To confirm that this selective advantage was conferred by *TP53* mutations and not by CNVs at Chr20q11.21[1-3], we analyzed all 140 hESC lines by SNP arrays and found that none of the lines carrying *TP53* mutations also carried the Chr20q CNV (Table S6). Our

overall results are consistent with a model in which positive selective pressure for clonal expansion of mutations that inactivate P53 are present during the routine culture of hPSCs. Indeed, it has previously been reported that loss of P53 activity facilitates the reprogramming of somatic cells to pluripotency[22,23] and promotes hPSC survival and proliferation[24], suggesting a prominent role for P53 in regulating self-renewal in hPSCs (Fig. 3e).

To test the reproducibility of our observations and to explore the effects of P53 mutations in additional contexts, we screened for *TP53* mutations in publically available RNA sequencing data from 251 hPSC samples in 57 published studies, corresponding to 13 hESC and 104 hiPSC lines (Fig. 4a-c, Table S7). The relatively high expression of *TP53* in stem cells provided sufficient read depth for allelic counting and allowed us to identify nine instances of eight distinct point mutations in *TP53*, three of which we had independently seen by WES. Like the mutations ascertained by WES, each of these eight mutations led to missense substitutions in the DNA-binding domain of P53 (Fig. 4d-f, Table S3, Fig. S2). When we considered both WES and RNAseq data sets, we identified four codons that were recurrently mutated in hPSCs: R181, G245, R248, and R273. Notably, we identified four distinct *TP53* mutations in the commonly used WA09 (H9) hESC line grown in different laboratories (P151S, R181H, R248Q, and R267W) indicating that the mutations arose during cell culture (Fig. 4h).

Of the 15 instances of *TP53* mutations observed by either WES or RNAseq, the percentage of mutant reads suggested that 10 were mosaic and that three had reached fixation (50% of allelic fraction). Surprisingly, two cell lines, WA09 (R248Q) and WIBR3 (H193R) had 80% ± 3% and 100% mutant reads, respectively (Table S3). These findings were consistent with the excess allelic fraction observed during the culture of WA26 (Fig. 3d) and suggested the presence of additional mutational mechanisms affecting mutant *TP53* allelic fraction. Indeed, we observed loss of heterozygosity (LOH) of a large telomeric domain including the *TP53* locus along chromosome 17 (Fig. S3). that was essentially complete in a gene-targeted derivative of WIBR3 (Fig. 4i), and was partial in WA09, consistent with the observed high fraction (80%) of mutant *TP53* reads. These results suggest that follow-on LOH after an initial *TP53* point mutation likely confers additional selective advantage.

We next asked if *TP53* mutations might affect cell differentiation or might confer selective advantage or disadvantage in differentiated cells. To this end, we examined studies for which there was RNA sequencing data for both hESCs and their differentiated progeny. Cell lines with substantial factions of *TP53* mutant cells could readily form teratomas, gut epithelial cells[25] (Fig. 4i), neuroepithelial cells[26] (Fig. 4j), and pancreatic polyhormonal cells[27] (Fig. 4k). Notably, a hESC line harboring a mosaic G245C mutation expanded its mutant allelic fraction over the course of differentiation[27], suggesting a continued selective advantage in differentiating mutant cells.

Together, our analyses indicate that researchers have unknowingly and routinely used hPSCs that harbor cancer-related missense mutations in *TP53*, sometimes accompanied with LOH. These findings have practical implications for the use of hESCs in disease modeling and transplantation medicine. The fact that we observed *TP53* mutations among both hESC and hiPSCs cultured with a wide variety of media, substrates, and passaging methods (Fig. S4) suggests that new culture conditions should be explored to reduce the selective pressure for *TP53* mutations during cell culture. We also suggest that hPSCs should be regularly subjected to genetic testing, particularly before and after stressful interventions such as gene editing or single-cell cloning that force hPSC populations through bottlenecks (Fig. 4l,m). Our specific findings here suggest that the P53 pathway should be an immediate focus for these genetic tests. However, a comprehensive ascertainment of the culture-acquired mutations that

repeatedly occur due to selective pressures will require the analysis of still-larger collections of stem cell lines by both exome and whole genome sequencing.

Our findings also demonstrate that sequencing tests provide an opportunity to detect differentiated cell preparations produced from hPSCs that harbor mutations, which could prove harmful after transplantation. The implementation of these exome sequencing procedures during the production of cell therapies from hPSCs would be likely to increase their safety for therapeutic use in conditions ranging from diabetes to Parkinson's disease. In support of this notion, clinical trials with hPSC-derived materials have recently been halted due to the discovery of undisclosed mutations[28], but have since resumed. We suggest that hPSCs and their derivatives be subjected to genome-wide analyses at several key steps: during initial cell line selection, as part of the characterization of a master bank of hPSCs, and as an end-stage release criterion prior to the transplantation of the hPSC-derived cellular product. Importantly, although *TP53* mutations recurred at detectable fractions in several cell lines, most lines (~95%) were free of detectable *TP53* mutations despite having spent extensive time in culture. Regenerative medicine remains a viable and exciting goal that is more likely to succeed as potential pitfalls, like the one we report here, are identified and addressed.

Acknowledgements

We are grateful for the many institutions and investigators world-wide that generously provided their cell lines and supported the publication of the results. We are indebted to Diane Santos, Melissa Smith, Kristen Elwell, Mary Anna Yram, Stacey Ellender, Liz Bevilacqua, and Diane Gage for their assistance with the considerable regulatory and logistical efforts required to acquire and sequence hESC lines. We also thank Kiki Lilliehook for her helpful comments and Ilyas Yildirim for his assistance with the molecular modeling of P53 mutations. We regret the omission of any relevant references or discussions due to space considerations. The Genomics Platform at the Broad Institute performed sample preparation, sequencing, and data storage. YA is a Clore Fellow. NB is the Herbert Cohn Chair in Cancer Research and was partially supported by The Rosetrees Trust and The Azrieli Foundation. Costs associated with acquiring and sequencing hESC lines were supported by HHMI and the Stanley Center for Psychiatric Research. FTM, SAM, and KE were supported by grants from the NIH (HL109525, 5P01GM099117, 5K99NS08371). KE was supported by the Miller consortium of the HSCI and FTM is currently supported by funds from the Wellcome Trust, the Medical Research Council (MR/P501967/1), and the Academy of Medical Sciences (SBF001\1016).

Author contributions

FTM, SG, SAM, and KE conceived the project. FTM and KE acquired hESC lines with the assistance of MC and GS, who also helped with regulatory issues pertaining to sequencing and data distribution. FTM cultured and banked hESC lines, prepared them for sequencing, and coordinated efforts to interpret and visualize sequencing data with the assistance of SG. SG performed computational data analysis and visualization with the help of GG, BH and SK. YA preformed the analysis of *TP53* mutations in the RNA-seq database with the assistance of SB and NB. Data were interpreted by FTM, SG, NK, GG, YA, SB, NB, SAM, and KE. NK, JM and CM designed, performed and analyzed experiments to confirm mosaicism of *TP53* mutations and measure the competitive expansion of *TP53* mutations in culture. SM derived HUES 68, 69, 70, 74, 75, and DI provided the KCL series of hESC lines. FTM, SG, SAM, and KE prepared drafts of the manuscript text and figures with contributions and comments from all authors.

Conflict of Interest
Authors declare no competing financial interests.

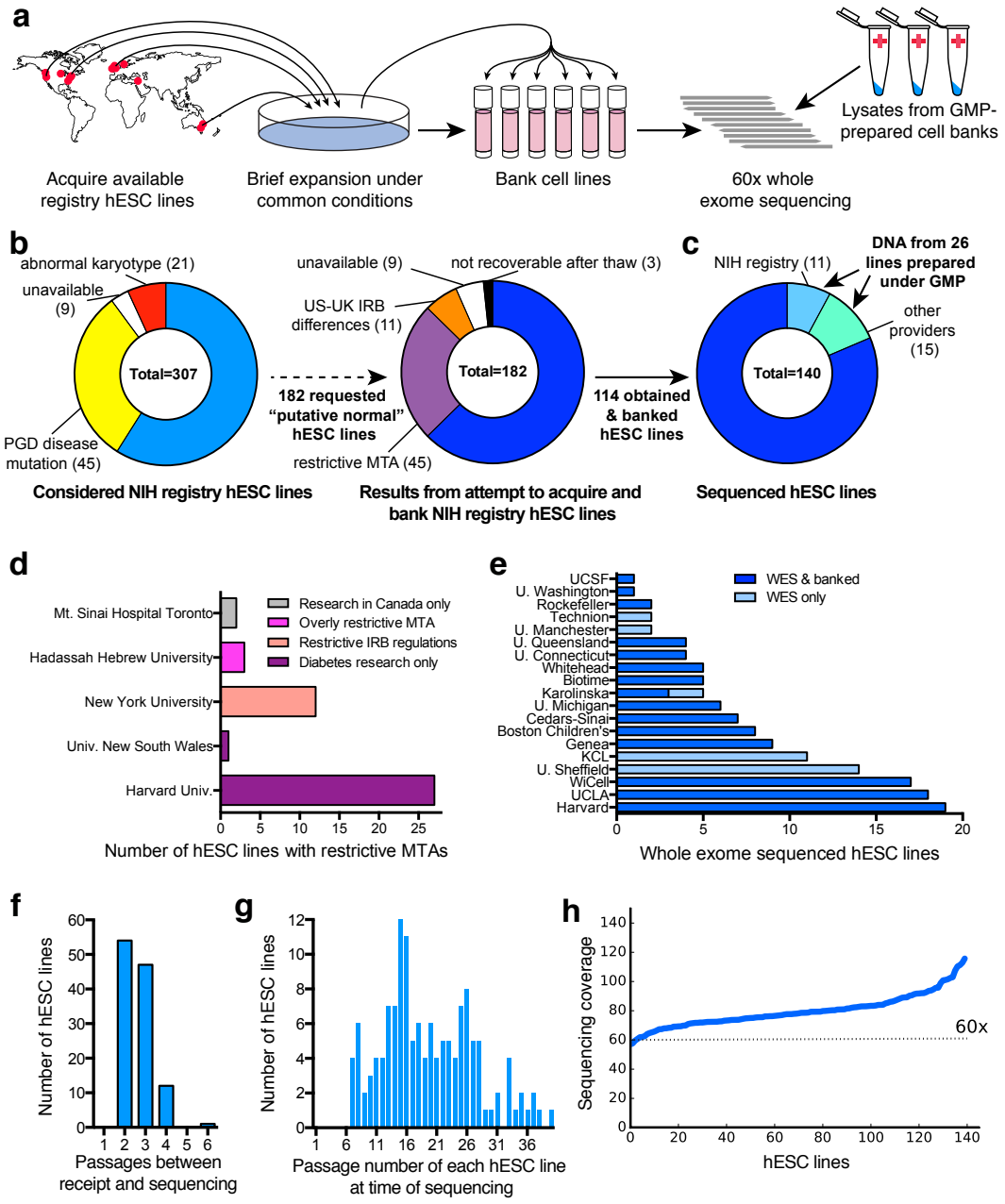

Figure 1. Acquisition and WES of 140 hESC lines.
a, Schematic workflow for hESC line acquisition and sequencing. **b,c,** Of the 307 hESC lines listed on the NIH registry at the time of this study, 182 were requested and 114 could be obtained, banked (b), and analyzed by WES along with 26 GMP-prepared cell lines (c). **d,** 45 hESC lines were excluded due to significant restrictions on their use. **e,** 140 hESC lines from 19 institutions were banked and/or sequenced. Please see Table S1 and supplementary methods for further details. **f,** HESCs were minimally cultured before banking and sequencing. **g,** Cumulative passage number of hESCs was moderate. **h,** Mean WES coverage exceeded 60x for almost all sequenced hESC lines. IRB, institutional review board; MTA, material transfer agreement; PGD, pre-implantation genetic diagnosis.

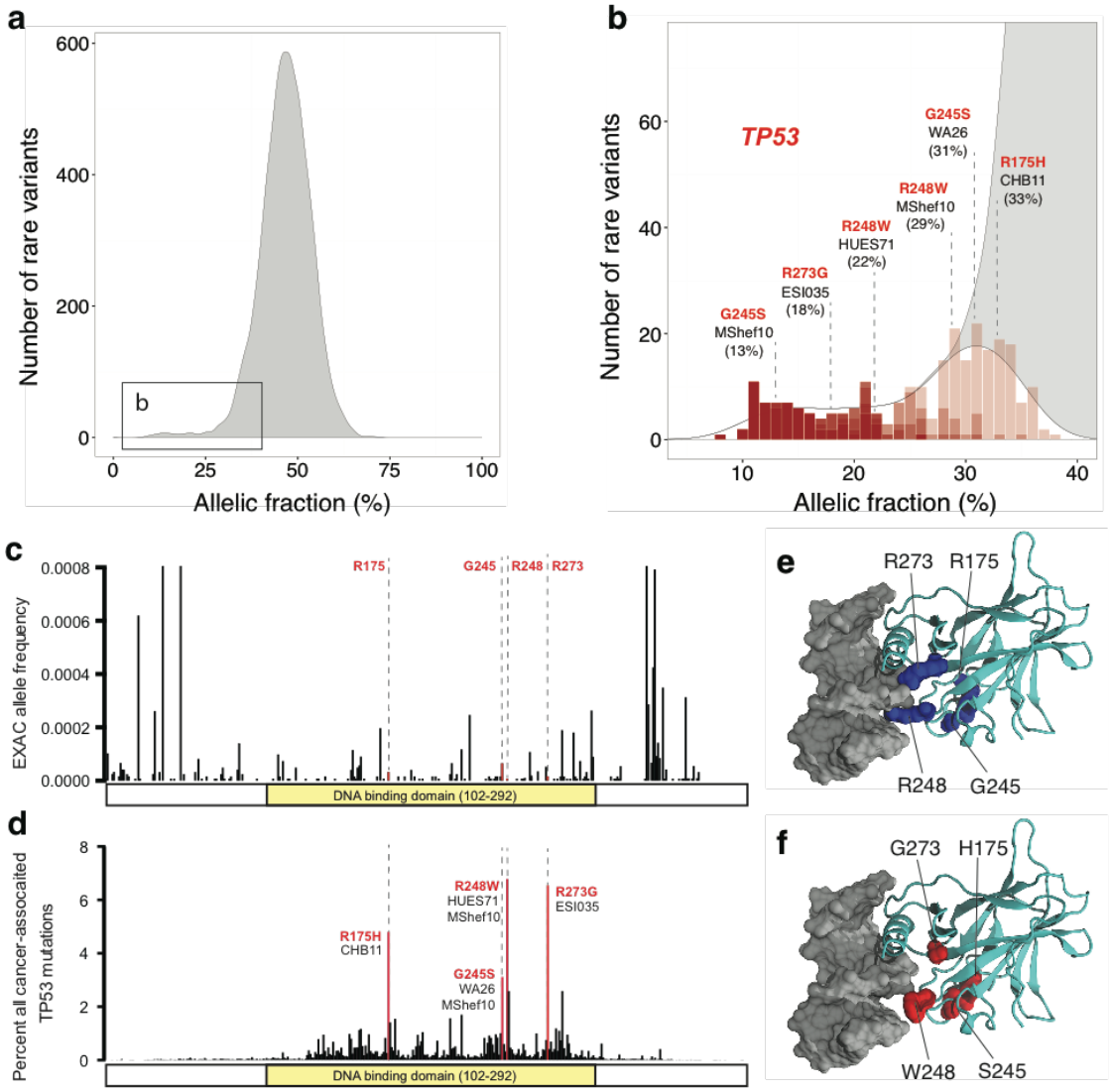

Figure 2. Identification of recurrent, cancer-associated *TP53* mutations in hESCs.
a, Most rare heterozygous genetic variants identified by WES in hESCs variants are present at approximately balanced (50/50) allelic proportions, but a small set of variants is present at lower allelic fractions (boxed left). **b,** A binomial test used to identify variants likely to be mosaic (P<0.01 for lightest red shading), of which six encode missense mutations in the tumor suppressor gene *TP53* (Table S3). **c**, Genetic variants in the four P53 codons found mutated in hESCs are rare in ExAC (<0.0001). **d,** The four P53 residues found mutated in hESCs are the four residues most commonly mutated in human tumors. **e,** On a crystal structure of P53 bound to DNA, the affected residues map to the DNA binding domain, including to arginine residues that directly interact with DNA. **f)** The residues mutated in hESCs disrupt DNA binding by P53.

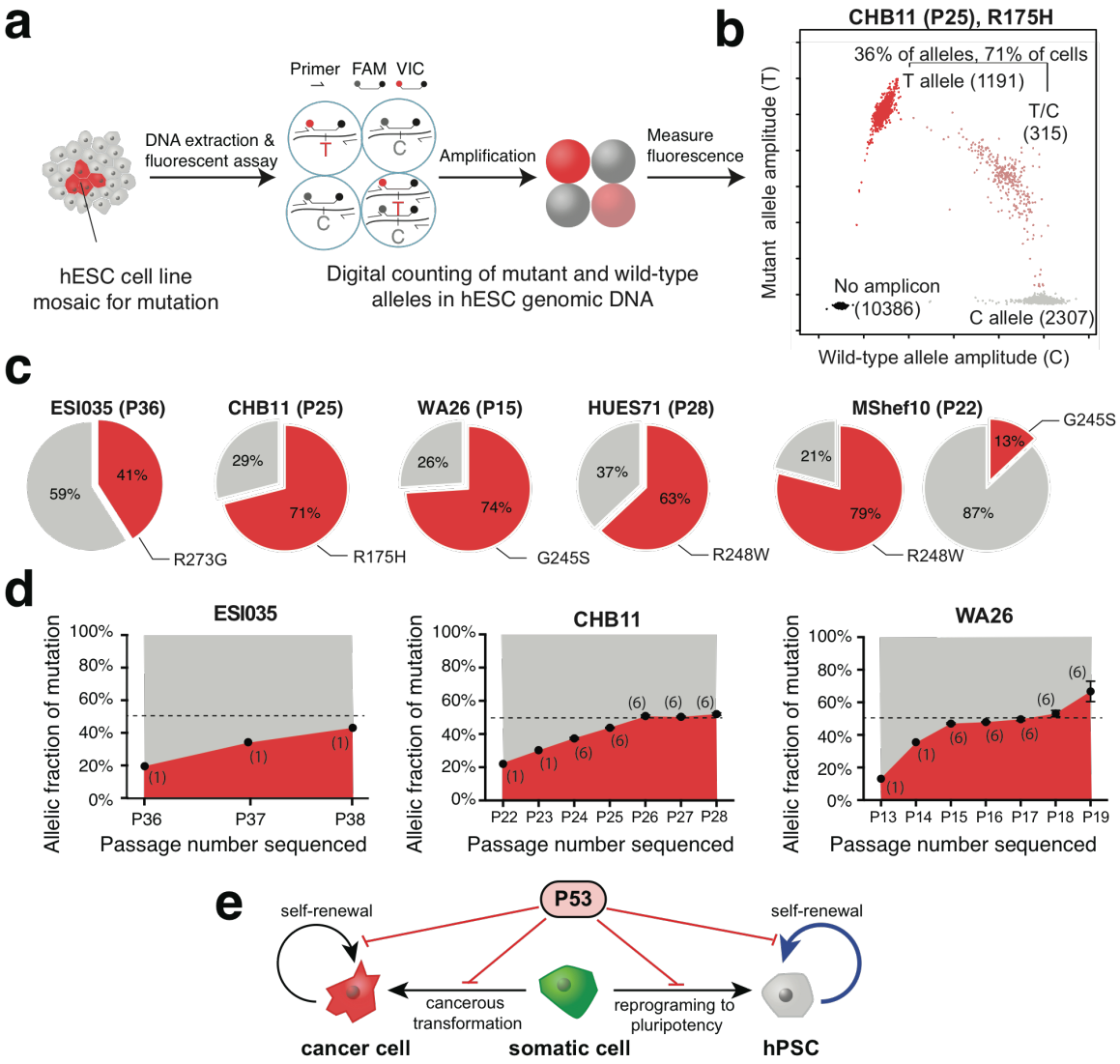

Figure 3. *TP53* mutations in hESCs are mosaic and confer strong selective advantage.
a, Droplet digital PCR (ddPCR) assay schematic. **b,** A representative assay exhibits four well-separated clusters of droplets that contain the reference allele (C, gray), mutant allele (T, red), both alleles (T/C, pink), or neither allele ("no amplicon", black). The fraction of hESCs carrying the *TP53* variant is estimated by digitally counting the four species of droplets. **c,** Estimated fraction of mutant cells (red) in affected hESC lines. **d,** The pronounced increase in the fraction of *TP53* mutant alleles during standard hESC culture suggests that *TP53* mutation confers a selective advantage. Error bars depict SEM and numbers indicate replicate wells. Note the second wave of allele-fraction expansion (after P17) for WA26, likely involving LOH. **e,** Model of P53's role in inhibiting both cancer formation and hPSC reprogramming, as well as self-renewal in both cell types.

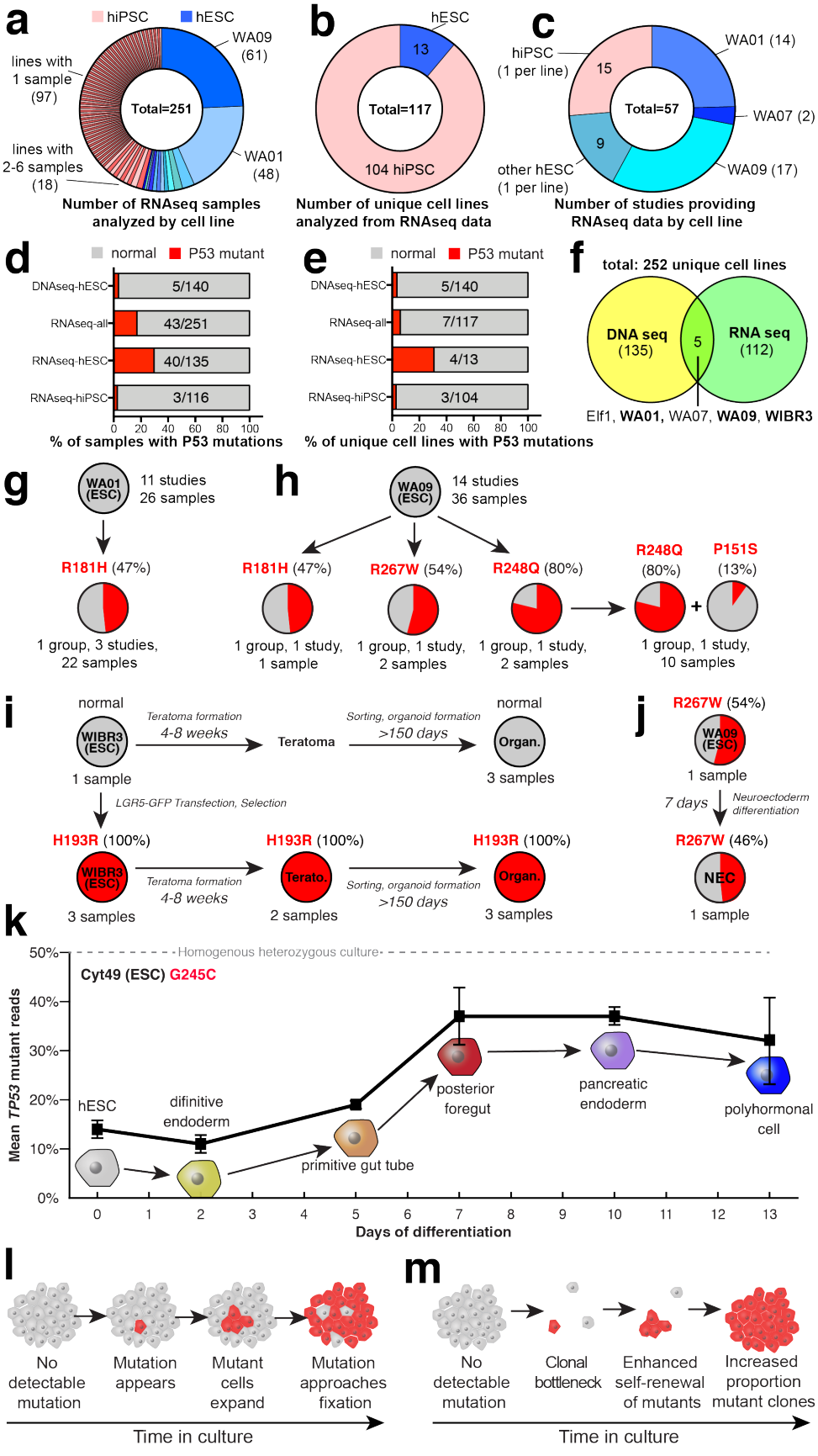
a
hiPSC
hESC
WA09 (61)
lines with 1 sample (97)
Total=251
lines with 2-6 samples (18)
WA01 (48)
Number of RNAseq samples analyzed by cell line
b
hESC
13
Total=117
104 hiPSC
Number of unique cell lines analyzed from RNAseq data
c
hiPSC (1 per line)
15
WA01 (14)
Total=57
WA07 (2)
9
other hESC (1 per line)
WA09 (17)
Number of studies providing RNAseq data by cell line
d
normal
P53 mutant
DNAseq-hESC
5/140
RNAseq-all
43/251
RNAseq-hESC
40/135
RNAseq-hiPSC
3/116
0 20 40 60 80 100
% of samples with P53 mutations
e
normal
P53 mutant
DNAseq-hESC
5/140
RNAseq-all
7/117
RNAseq-hESC
4/13
RNAseq-hiPSC
3/104
0 20 40 60 80 100
% of unique cell lines with P53 mutations
f
total: 252 unique cell lines
DNA seq (135)
5
RNA seq (112)
Elf1, WA01, WA07, WA09, WIBR3
g
WA01 (ESC)
11 studies
26 samples
R181H (47%)
1 group, 3 studies, 22 samples
h
WA09 (ESC)
14 studies
36 samples
R181H (47%)
1 group, 1 study, 1 sample
R267W (54%)
1 group, 1 study, 2 samples
R248Q (80%)
1 group, 1 study, 2 samples
R248Q (80%)
P151S (13%)
1 group, 1 study, 10 samples
i
normal
WIBR3 (ESC)
1 sample
Teratoma formation
4-8 weeks
Teratoma
Sorting, organoid formation
>150 days
normal
Organ.
3 samples
LGR5-GFP Transfection, Selection
H193R (100%)
WIBR3 (ESC)
3 samples
Teratoma formation
4-8 weeks
H193R (100%)
Terato.
2 samples
Sorting, organoid formation
>150 days
H193R (100%)
Organ.
3 samples
j
R267W (54%)
WA09 (ESC)
1 sample
7 days
Neuroectoderm differentiation
R267W (46%)
NEC
1 sample
k
Homogenous heterozygous culture
Cyt49 (ESC) G245C
Mean TP53 mutant reads
50%
40%
30%
20%
10%
0%
hESC
difinitive endoderm
primitive gut tube
posterior foregut
pancreatic endoderm
polyhormonal cell
0 1 2 3 4 5 6 7 8 9 10 11 12 13
Days of differentiation
l
No detectable mutation
Mutation appears
Mutant cells expand
Mutation approaches fixation
Time in culture
m
No detectable mutation
Clonal bottleneck
Enhanced self-renewal of mutants
Increased proportion mutant clones
Time in culture

Figure 4. A substantial fraction of hPSCs in published RNA sequencing studies harbor *TP53* mutations.

a,b,c, Published RNA sequencing (RNAseq) datasets from 251 hESC and hiPSC samples (a), contain 13 unique hESC and 104 unique hiPSC lines (b,c). **d,e,** *TP53* mutations are observed in many hESC samples (40/135, 30%) and several hiPSC samples (3/116, 3%) (d), together affecting 7/117 (6%) of sequenced hPSC lines (e). **f,** Combined RNAseq and WES analysis reveals 12/252 distinct cell lines (5%) affected by 15 *TP53* mutations (Table S3), three of which (WA01, WA09, WIBR3) showed *TP53* mutations only certain RNAseq studies. **g,** 22 samples of WA01 from three studies carried out by one laboratory carried an R181H mutation (allelic fraction in parentheses). **h,** The commonly used WA09 cell lines carried four distinct *TP53* mutations observed in three different groups, indicating that these mutations arose independently in culture. **i**, WIBR3 with wild-type *TP53* underwent gene editing and acquired a *TP53* mutation present in 100% of alleles, indicating likely fixation on a mutation on one allele and a LOH on the other allele; these mutations did prevent teratoma formation or gut differentiation. **j**, WA09 carrying a fixed *TP53* mutation could be differentiated to neuroectodermal cells. **k**, The *TP53* mutant allelic fraction increased over the course of Cyt49 differentiation to pancreatic polyhormonal cells, suggesting a persistent selective advantage in differentiating cells. **l,m**, Model of *TP53* mutation enrichment by selective advantage during hPSC maintenance (l) or during clonal bottlenecks that occur during gene editing or inefficient passaging (m).

MATERIALS AND METHODS

hESC acquisition

As a source of hESCs for this study, we focused on those that had been voluntarily listed by research institutions on the registry of hESC lines maintained by the US National Institutes of Health (NIH) (http://grants.nih.gov/stem_cells/registry/current.htm). As of July 8, 2015, a total of 307 hESC lines were listed on this registry. Of these, we requested viable frozen stocks of the 182 lines annotated to be available for distribution and to lack known karyotypic abnormalities or disease-causing mutations. During our effort to obtain these cell lines, we found that 45 were subject to overly restrictive material transfer agreements that precluded their use in our studies and 11 could not be readily obtained as frozen stocks due to differences in human subjects research regulations between the U.S. and the U.K. Nine cell lines were unavailable upon request or were overly difficult to import, and three could not be cultured despite repeated attempts. Further details on the availability of cell lines can be found in Table S1.

hESC culture

Different laboratories employ different methods to culture hESCs, raising the question of how best to thaw and culture the cell lines we obtained from multiple sources. Traditionally, hESCs are maintained on gelatinized plates and co-cultured with replication incompetent mouse embryonic fibroblast (MEF) feeder cells in tissue culture medium containing knockout serum replacement (KOSR). More recently, hESCs have been shown to grow well on a substrate of cell line-derived basal membrane proteins known by the trade names of Matrigel (BD Biosciences) or Geltrex (Life Technologies), in mTeSR1[29], E8[30], or similar in the absence of feeder cells. In previous work, we found that a medium containing an equal volume of KOSR-based hESC medium (KSR) and mTeSR1 (STEMCELL Technologies) (KSR:mTeSR1) robustly supports the pluripotency of hESCs undergoing antibiotic selection during the course of gene targeting experiments under feeder-free conditions[31]. To minimize stress to hESCs previously cultured and frozen under diverse conditions, cell lines were thawed in the presence of 10 μM Y-27632 (DNSK International) into two wells of a 6-well plate, one of which contained KSR:mTeSR1 on a substrate of Matrigel, and the other containing KOSR-based hESC medium on a monolayer of irradiated MEFs. After 24 hours, Y-27632 was removed and cells were fed daily with the aforementioned media in the absence of any antibiotics. All cultures were tested for the presence of mycoplasma and cultured in a humidified 37°C tissue culture incubator in the presence of 5% CO_2 and 20% O_2.

Colonies of cells with hESC morphology and with a diameter of approximately 400-1000 micrometers were transferred into MEFcm:mTeSR medium containing 10 μM Y-27632 on a substrate of Matrigel by manual picking under a dissecting microscope. Cells with differentiated morphology were removed from plates by aspiration during feeding. Once cultures consisting of cells with homogeneous pluripotent stem cell morphology had been established, they were passaged by brief (2-10 minute) incubation in 0.5 mM EDTA in PBS followed by gentle trituration in MEFcm:mTeSR medium containing 10 μM Y-27632 and re-plating. Once cultures had reached approximately 90% confluence in one well of a 6 well plate, they were passaged with ETDA onto a Matrigel-coated 10 cm plate. Upon reaching approximately 90% confluence, cell lines were dissociated with EDTA as described above and banked for later use in identical cryovials containing approximately 3 million cells each in cryoprotective medium containing 50% MEFcm:mTeSR, 10 μM Y-27632, 10% DMSO, and 40% fetal bovine

serum (HyClone). A subset of hESC lines (Table S1) were passaged enzymatically with TrypLE Express (Life Technologies), expanded onto two 15 cm plates, and frozen down in 25 cryovials.

Whole exome sequencing and genotyping

Cell pellets of approximately 1-5 million cells were generated from banked cryovials of research-grade hESC lines, or were obtained directly from institutions providing GMP-grade hESC lines. Cell pellets were digested overnight at 50°C in 500 μl lysis buffer containing 100 μg/ml proteinase K (Roche), 10 mM Tris (pH 8.0), 200 mM NaCl, 5% w/v SDS, 10 mM EDTA, followed by Phenol:Chloroform precipitation, ethanol washes, and resuspension in 10 mM Tris buffer (pH 8.0). Genomic DNA was then transferred to the Genomics Platform at the Broad Institute of MIT and Harvard for Illumina Nextera library preparation, quality control, and sequencing on the Illumina HiSeq 2500 platform. Sequencing reads (150 bp, paired-end) were aligned to the hg19 reference genome using the BWA alignment program. Genotypes from WES data for the cell lines were computed using best practices from GATK software[32] compiled July 31, 2015. Sequencing quality and coverage were analyzed using Picard tool metrics. Cross sample contamination was estimated using VerifyBamID (v1.1.2)[33]. Data from each cell line was independently processed with the HaplotypeCaller walker and further aggregated with the CombineGVCFs and GenotypeGVCFs walkers to generate a combined variant call format (VCF) file. Genotyped sites were finally filtered using the ApplyRecalibration walker.

To determine whether lines with or without acquired *TP53* mutations show other chromosomal aberrations or smaller regional changes in copy number, additional genotyping of the 140 hESC lines was performed using a custom high density SNP array ("Human Psych array") that contains more than half a million SNPs across the genome. CNVs larger than 500 kb were identified using the PennCNV (v1.0.0)[34] tool (penncnv.openbioinformatics.org). All CNVs were manually reviewed and are shown in Table S6.

Mosaic variant analysis

To identify candidate mosaic variants, a table of heterozygous variants was generated from the VCF (Table S2). To limit the frequency of false positive calls due to sequencing artifacts and PCR errors, variants were included if they had a variant read depth (DP) of at least 10, if they were either flagged as a "PASS" site or were not reported in the Exome Aggregation Consortium's (ExAC) database[11], and if they were not located in regions of the genome with low sequence complexity, common large insertions and segmental duplications, as described by Genovese and colleagues[5]. Multiallelic sites were split, left-aligned, and normalized. The resulting list of 2.1 million "high-quality heterozygous variants" was further refined to include sites that were covered by at least 60 unique reads and had a high confidence variant score ("PASS") as ascertained by GATK's Variant Quality Score Recalibration software (840,222 variants). To exclude common inherited variants we selected variants present in less than 0.01% of the (ExAC) control population and restricted our analysis to only singleton or doubleton variants (9490 variants present in 1-2 of the 140 samples). Coverage was calculated by summing reference and alternate allele counts for each variant. Allelic fraction (AF) was calculated by dividing the alternate allele count by the total read coverage (both alleles) of the site.

Although the allelic fraction of inherited heterozygous variants is expected to be 50%, reference capture bias (a tendency of hybrid selection to capture the reference allele more efficiently than alternative alleles) causes the actual expected allele fraction

for SNPs and indels to be closer to 45% and 35% respectively[5]. To account for these technical biases, we used a binomial test with a null model centered at 45% allelic fraction for inherited SNPs and 35% for inherited indels. Variants for which this binomial test was nominally significant ($P < 0.01$) were deemed to be candidate mosaic variants. The nominal P-value threshold of 0.01 was chosen as an inclusive threshold in order to screen sensitively for potentially mosaic variants, at the expense of also capturing false positives for which low allelic fractions represented statistical sampling fluctuations. For this reason, we considered it important to further evaluate putative mosaic variants by independent molecular methods that deeply sample alleles at the nominated sites (Figure 3). A much more stringent computational screen based on a P-value threshold of 10^{-7} identified three of the six *TP53* variants, and *TP53* was also the only gene with multiple putatively mosaic variants in this screen.

We also identified all high quality heterozygous variants that passed the inclusive statistical threshold of ($P<0.01$) in our binomial test and could potentially be mosaic (n=36,396). These data are included in Table S2.

Variant annotation was performed using SnpEff with GRCh37.75 Ensembl gene models. Variants with moderate impact were classified as damaging by a consensus model based on seven *in silico* prediction algorithms[35].

Assessment of *TP53* mutation frequency in cancer

We turned to the ExAC database[11] that complies the whole exome sequences of over 60,000 individuals to assess the frequency at which the amino acid residues we observed to be mutated in some hESCs were affected in the general population. We then consulted the COSMIC (http://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=TP53)[12], ICGC (https://cbioportal.gdc.cancer.gov/cbioportal/)[13], and IARC P53 (http://p53.iarc.fr/TP53SomaticMutations.aspx)[14] databases and plotted the percent of tumors carrying a mutation in each codon (Fig. 2d, Fig. S2b).

Molecular modeling of P53 protein

To visualize the spatial location of the amino acid residues affected by *TP53* mutations observed in hESCs by WES on the P53 protein, we downloaded the 1.9 Angstrom X-ray diffraction based structure file from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (file 3KMD) and visualised the secondary structure of a P53 monomer complexed to DNA as a ribbon diagram. DNA was illustrated as a space-filling model. Affected residues were indicated as space-filling model superimposed on the ribbon diagram of P53 and highlighted in blue (wild-type) or red (mutated) without consideration of how the mutations might affect the secondary or tertiary structure of the protein.

Measurement of *TP53* variant allele fraction by ddPCR

We assayed the allelic fraction of the four distinct *TP53* mutations identified by WES (Table S3) in the 140 hESC lines by droplet digital PCR (ddPCR). Each ddPCR analysis incorporated a custom TaqMan assay (IDT). Assays were designed with Primer3Plus and consisted of a primer pair and a 5' fluorescently labeled probe (HEX or FAM) with 3' quencher (Iowa Black with Zen) for either the control (reference) or mutant (alternative) base for each identified P53 variant. Primer and probe sequences used in this assay are given in Table S4. Genomic DNA from each hESC line was analyzed by ddPCR according to the manufacturer's protocol (BioRad). The frequency of each allele for a given sample was estimated first by Poisson correction of the endpoint

fluorescence reads[21]. These corrected counts were then converted to fractional abundance estimates of the mutant allele and multiplied by two to determine the fraction of cells carrying the variant allele.

Longitudinal hESC culture and calculation of *TP53* mutation expansion

To assess how the allelic fraction of *TP53* mutations might change over time in culture, hESC lines CHB11 (passage 22 or 25), WA26 (passage 13 or 15), and ESI035 (passage 36 in two separate experiments) were serially passaged in mTeSR1 media (STEMCELL Technologies) at a density of approximately 30,000 cells/cm^2 in the presence of 10 μM Y-27632. Cells were fed daily with mTeSR1 and passaged with Accutase (Innovative Cell Technologies Inc.) at approximately 90% confluence. To monitor changes in allelic fractions, genomic DNA from cells at the indicated passages were analyzed by ddPCR. To calculate the relative expansion rate of mutant relative to wild-type cells, we applied the following formula:

$$g = \frac{\ln R_2 - \ln R_1}{T_2 - T_1}$$

where R_0 is defined as the ratio of (variant positive cells)/(variant negative cells) after some number of starting passages and R_1 and R_2 represent the aforementioned ratios measured on the same sample at T_1 and T_2>T_1 passages respectively. From this equation, the estimation of variant positive cells after t passages from starting ratio R_0 can be defined as

$$R_0 e^{gt}$$

Note that this equation estimates the relative growth rate of cells carrying the variant allele with a round of passaging as unit of time, with both relative survival and growth being incorporated. These data are included in Table S5. For the subsequent calculation of the earliest passage at which these mutations might have become detectable, the detection thresholds (R_0) for WES and ddPCR was assumed to be 0.1 (10/100 reads) and 0.001 (1 per 1000 droplets), respectively.

RNA sequencing analysis

In order to identify *P53* mutations in hPSCs, we analyzed 256 publicly available high-throughput RNA sequencing samples of hPSCs from the SRA database (http://www.ncbi.nlm.nih.gov/sra)[36]. Data accession numbers for SRA (and GEO, where applicable) are provided in Table S7. Five of these 256 samples were not considered further as they were from single cells rather than cell lines. Following sequence alignment to the hg19 reference genome with Tophat2[37], single nucleotides divergent from the reference genome were identified using GATK HaplotypeCaller[32]. As sufficient sequencing-depth is required to deduce sequence mutation, a threshold of 25 reads per nucleotide was set. Under this criterion, 43 samples (40 hESC and 3 hiPSC) had a missense mutation in *TP53*. 10 of the 40 hESC samples (WA09) carried two separate mutations (Table S7). Upon the identification of cell lines carrying mutant reads, RNA sequencing data from studies containing differentiated samples were included for analysis.

Loss of heterozygosity analysis

In order to evaluate *TP53* alleles, we assessed the level of polymorphism by calculating the ratio between the minor and major alleles across chromosome 17. In order to minimize sequencing noise and errors, we included SNPs covered by more than 10 reads and that are located in the dbSNP build 142 database[38]. The resulted wig files were then plotted using Integrative Genomics Viewer (IGV)[39] (Fig. S4). In order to quantify the difference in polymorphism between samples, we converted the wig files to BigWig using UCSC Genome Browser utilities[40] and summed the allelic ratios between the distal part of the short arm of chromosome 17 (17p), the proximal side of this arm and the long arm of chromosome 17 (17q). The allelic ratio sum was then divided by the region's length (bp), which resulted in the proportion of SNPs, followed by one-sided Z score test for two population proportion to compare between the chromosome 17 areas within each sample. While most samples with mutations in *TP53* showed a comparable, non-significant rate of polymorphic sites along the chromosome, WIBR3 samples with H193R mutations and WA09 samples with both P151S and R248Q mutations had a significantly different proportion ($p<0.001$) of polymorphic sites, in the distal part of the short arm of the chromosome (first $16x10^6$ base pairs), including *TP53* site. Unlike the three mutant WIBR3 samples, the wild-type WIBR3 sample had a normal distribution of polymorphic sites with no significant difference between the short and long arms.

SUPPLEMENTARY FIGURES AND LEGENDS

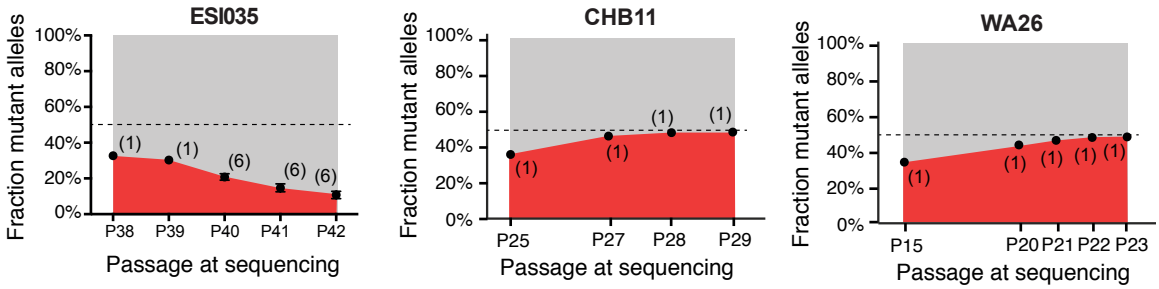

Figure S1. Replicates of cell competition assays carried out at earlier starting passages. Note that while the mutant allelic fractions for lines CHB11 and WA26 approach fixation, that the fraction of mutant cells unexpectedly decreases for ESI035 over several passages, indicating a potential selective disadvantage that co-segregates with the *TP53* mutation in this experiment. The number of replicate wells is indicated in each graph. Error bars depict SEM.

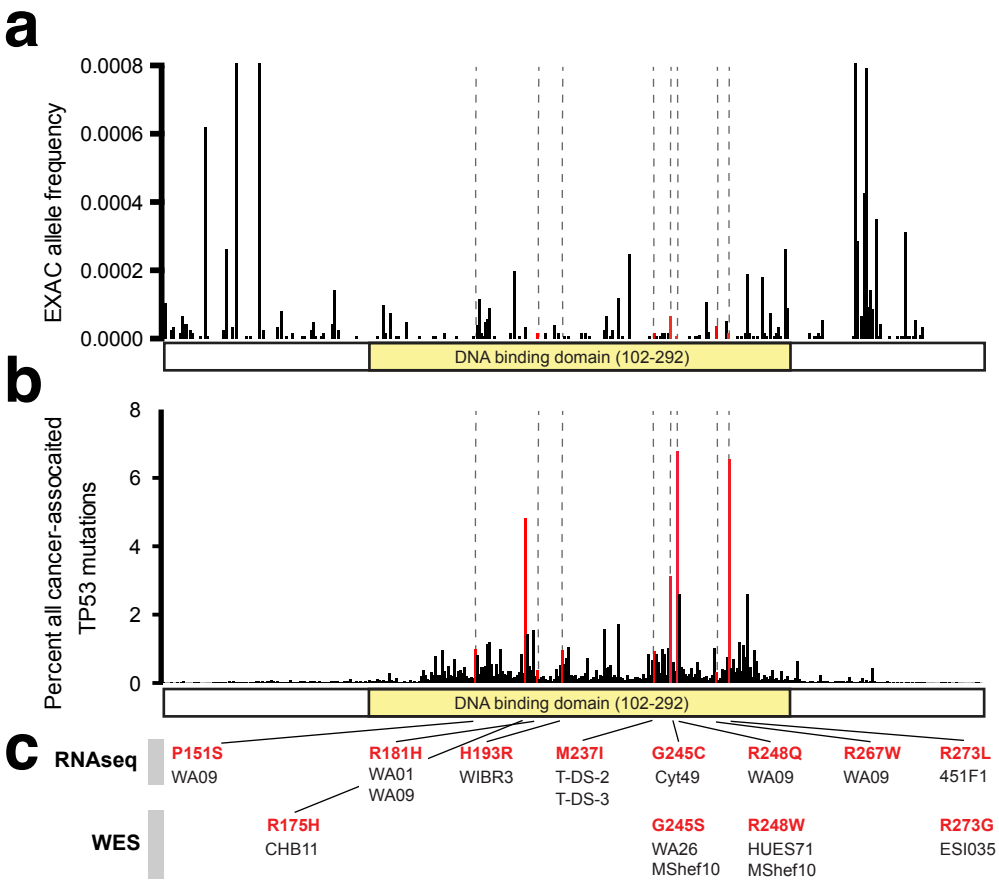

Figure S2. Summary of all observed P53 mutations.
a-b, Graphical representation of each of the 9 mutated bases in P53 observed across the 252 whole exome sequenced (WES) and RNA sequenced (RNAseq) hPSC lines depicting their allele frequency in ExAC (a) and the incidence with which the relevant codons are mutated in human cancer (b). **c**, The 15 instances of these mutations in 12 distinct cell lines is represented along with whether the mutation was seen by WES or RNAseq. Although the M237I event is seen in two distinct iPSC lines, it is conservatively counted here as a single event since the hiPSC clones may be related.

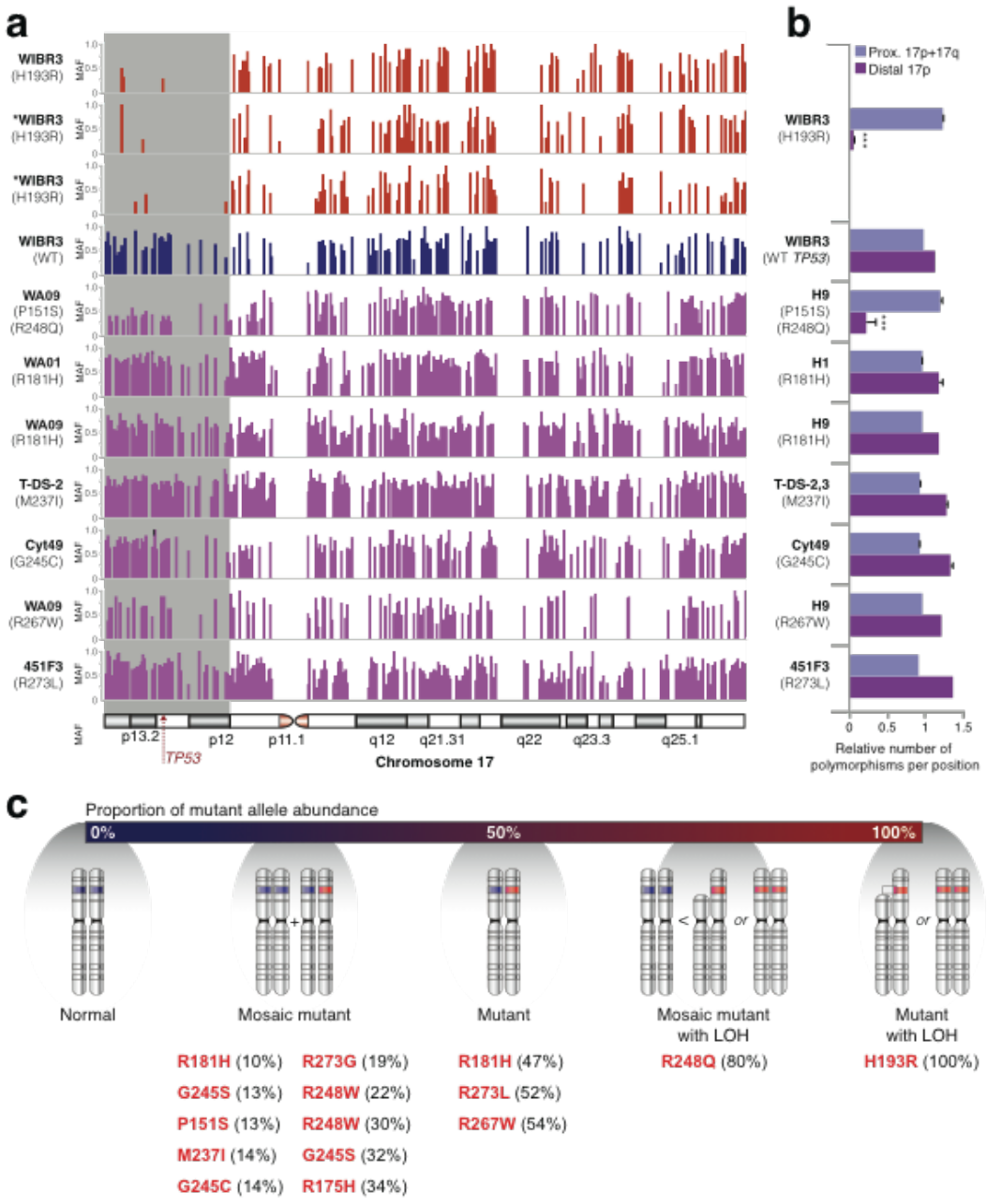

Figure S3. Analysis of loss of heterozygosity in RNA sequencing samples.
a, Polymorphic sites on chromosome 17 in different hPSCs with mutations in *TP53*. WIBR3 cells with H193R mutation and H9 cells with both P151S and R248Q mutations show less polymorphism in the distal part of chromosome 17p compared to the proximal part of 17p and 17q. *samples with less than 25 reads. **b**, Summation of the polymorphic sites in the distal part of chromosome 17p compared to the proximal part of 17p and 17q, divided by the overall frequency of polymorphic sites along chromosome 17. WIBR3 cells with H193R mutation and H9 cells with both P151S and R248Q mutations have a significantly different proportion between the two parts of the chromosome, implying loss of heterozygosity (LOH). ***$p < 0.001$. **c**, A schematic representation of possible allele states of *TP53* in cultured hPSCs with all observed mutations depicted. Depending on the percentage of mutant reads in a culture, one can deduce if the culture is homogenous or mosaic for a mutation, and whether, in addition to a point mutation, LOH has occurred in the *TP53* locus.

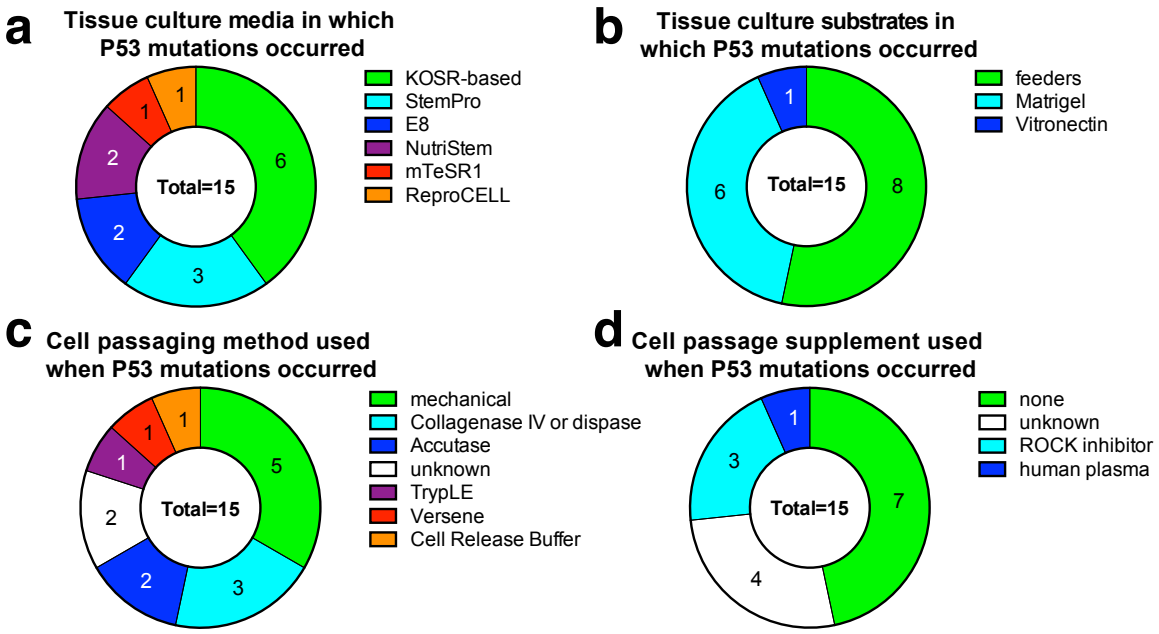

Figure S4. Culture and passaging method employed for samples bearing P53 mutations.
a, P53 mutations were observed in hPSCs grown in a broad array of culture media including home-made medium supplemented with knockout serum replacement (KOSR), and defined, commercial media such as E8. **b**, Similar numbers of P53 mutations were observed from cells grown with feeder cells or under feeder-free conditions. **c**, Since passaging hPSCs can introduce stresses or clonal bottlenecks, we examined whether P53 mutations were consistently seen when a particular passaging method was used, but we observed a wide variety of passaging methods associated with these mutations. Note that the interpretation of these data are complicated by the fact that the culture methods employed in the final published study may not reflect the previous culture history of that cell line, which may have previously passed through multiple laboratories, as well as by the lack of detail about culture methods present in some published studies. **d**, The addition of supplements such as rock inhibitor at passages does not appear to be sufficient to prevent P53 mutations in hPSCs.

SUPPLEMENTARY TABLE LEGENDS

Table S1. Considered and whole exome sequenced hESC lines.
Tab 1. We considered hESC lines for WES if they were listed on the NIH Human Embryonic Stem Cell Registry (http://grants.nih.gov/stem_cells/registry/current.htm) or if they were prepared under GMP conditions. Cell lines were typically excluded from consideration if they were unavailable for distribution or contained known karyotypic abnormalities in more than 10% of analyzed cells or disease-causing mutations identified by PGD. Cell lines with MTAs that restricted our ability to work with the cell lines, that could not be recovered upon thawing, or proved to be unavailable upon request were also excluded. Passage number at the time of request, the number of passages and time in culture from thaw to passaging, and the passaging method, media, and substrate, are provided, as is mean sequencing coverage and % cross sample contaminated per cell line. GMP, good manufacturing practice; MTA, material transfer agreement; PGD, pre-implantation genetic diagnosis; WES, whole exome sequencing. **Tab 2.** Summary of number of cells considered and sequenced, including reasons for exclusion. These data are presented graphically in Figure 1b-e.

Table S2. Identification of candidate mosaic variants present in sequenced hESCs.
Tab 1. Filters used to identify likely mosaic variants among all heterozygous variants present among the sequenced exomes of 140 hESCs. **Tab 2.** List of 263 candidate mosaic variants passing quality control filters and present no more than two times among the 140 sequenced hESC lines. Variants are arranged by chromosome position and annotated by likely functional impact and frequency in the general population (ExAC AC). **Tab 3.** Variants from the list in Tab 2 predicted to have either a high or damaging impact on gene function based on a consensus of 7 bioinformatic algorithms. See Materials and Methods for further details. **Tab 4.** In addition to mosaic variants identified using these stringent filters, we provide an inclusive list of all high confidence somatic variants (n=36,396) that pass the binomial test with a P value of <0.01. SNP, single nucleotide polymorphism; CHROM, chromosome number; POS, genomic position (hg19); REF, reference allele; ALT, alternate allele; HESC, human embryonic stem cell line; REFC, count of reference alleles; ALTC, count of alternate alleles; FILTER, high confidence variant score; EXACAC, allele count in the Exome Aggregation Consortium (ExAC) database; IMPACT, predicted effect of mutation; HESCAC, allele count in hESCs; TOTALC, REFC+ALTC; AF, allelic faction (ALTC/TOTALC); P, P value for binomial test on allelic fraction.

Table S3. Characteristics of *TP53* mutations identified in hESCs by WES and RNAseq.
Tab 1. Summary of all 15 instances of *TP53* mutations observed by WES and RNAseq with details of read depth, allelic fraction, P value, reference, and culture method. Note that all observed mutations are frequently seen in human cancer, and that most mutations have evidence of mosaicism, indicating that they were likely culture-derived. bFGF, basic fibroblast growth factor (FGF2); COSMIC, Catalogue of Somatic Mutations in Cancer (http://cancer.sanger.ac.uk/cosmic); ExAC, Exome Aggregation Consortium (http://exac.broadinstitute.org/); Freq., frequency; GMP, good manufacturing practice, IARC, International Agency for Research on Cancer (http://p53.iarc.fr/); ICGC, International Cancer Genome Consortium (http://icgc.org/); MEF, mouse embryonic fibroblast; Seq., sequencing; SNL, SNL mouse fibroblast feeder cell line; WES, whole exome sequencing. Errors denote SEM. **Tab 2.** Breakdown of the incidence of P53 mutations by culture media, substrate, and passaging method.

Table S4. Primer and probe sequences used for ddPCR-based determination of *P53* variant allele frequency.

Table S5. Calculation of selective advantage conferred by three distinct *TP53* variants.
The allelic fraction of *TP53* variants was measured at several passages by ddPCR in hESCs cultured under standard conditions. Replicate experiments per passage are shown in grey, and average values are shown in black. The observed increase in allelic frequency of each of the variants across time in culture is consistent with a substantial growth or survival advantage in all but one instance. See Materials and Methods for details on ddPCR and the calculation of the effect per passage.

Table S6. Large copy number variants in hESCs identified by the Human Psych array.
Tab 1. Summary of hESC lines with large copy number variants (>500kb) as ascertained by SNP array analysis. Two of the five cell lines with acquired *TP53* mutations harbored large structural alternations (HUES71 and MShef10). Five separate cell lines (CSES25, ESI051, MShef3, UM78-1 and WA21) had an amplification at the pericentromeric region of chromosome 20 (Chr20q11.21)**. Tab 2.** Complete list of large deletions or duplications (>500kb) identified across the 140 hESC lines.

Table S7. Identification of *TP53* mutations in hPSCs by RNA sequencing and WES
Tab 1. List of all RNA sequenced samples from hPSCs. Five of these samples (cell2-7) were removed since they were from single stem cells rather than cell lines. **Tab 2.** Summary of the number of samples and studies generated from each cell line. **Tab 3.** List of all samples harboring *TP53* mutations, their chromosomal location, and the relevant study. **Tab 4.** Summary of all affected cell lines and studies. **Tab 5.** Summary of affected samples, cell lines, and number of mutations seen in hESCs and hiPSCs by WES and RNAseq.

REFERENCES

1. ISCI. screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. *Nature Biotechnology* **29,** 1132–1144 (2011).
2. Avery, S. *et al.* BCL-XL mediates the strong selective advantage of a 20q11.21 amplification commonly found in human embryonic stem cell cultures. *Stem Cell Reports* **1,** 379–386 (2013).
3. Nguyen, H. T. *et al.* Gain of 20q11.21 in human embryonic stem cells improves cell survival by increased expression of Bcl-xL. *Mol. Hum. Reprod.* **20,** 168–177 (2014).
4. Unger, C., Skottman, H., Blomberg, P., Dilber, M. S. & Hovatta, O. Good manufacturing practice and clinical-grade human embryonic stem cell lines. *Human Molecular Genetics* **17,** R48–53 (2008).
5. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371,** 2477–2487 (2014).
6. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349,** 1483–1489 (2015).
7. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N Engl J Med* **371,** 2488–2498 (2014).
8. Adewumi, O. *et al.* Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. *Nature Biotechnology* **25,** 803–816 (2007).
9. Baker, D. *et al.* Detecting Genetic Mosaicism in Cultures of Human Pluripotent Stem Cells. *Stem Cell Reports* **7,** 998–1012 (2016).
10. Schwartz, S. D. *et al.* Human embryonic stem cell-derived retinal pigment epithelium in patients with age-related macular degeneration and Stargardt's macular dystrophy: follow-up of two open-label phase 1/2 studies. *Lancet* **385,** 509–516 (2015).
11. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
12. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* **43,** D805–11 (2015).
13. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database* **2011,** bar026–bar026 (2011).
14. Bouaoun, L. *et al.* TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. *Hum. Mutat.* **37,** 865–876 (2016).
15. Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408,** 307–310 (2000).
16. Rideout, W. M., Coetzee, G. A., Olumi, A. F. & Jones, P. A. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science* **249,** 1288–1290 (1990).
17. Cho, Y., Gorina, S., Jeffrey, P. D. & Pavletich, N. P. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* **265,** 346–355 (1994).
18. Willis, A., Jung, E. J., Wakefield, T. & Chen, X. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene* **23,** 2330–2338 (2004).
19. Malkin, D. Li-fraumeni syndrome. *Genes & Cancer* **2,** 475–484 (2011).
20. Xu, J. *et al.* Heterogeneity of Li-Fraumeni syndrome links to unequal gain-of-function effects of p53 mutations. *Sci Rep* **4,** 4223 (2014).
21. Hindson, B. J. *et al.* High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **83,** 8604–8610 (2011).

22. Marion, R. M. *et al.* A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity. *Nature* **460,** 1149–1153 (2009).
23. Zhao, Y. *et al.* Two supporting factors greatly improve the efficiency of human iPSC generation. *Cell Stem Cell* **3,** 475–479 (2008).
24. Amir, H. *et al.* Spontaneous Single-Copy Loss of TP53 in Human Embryonic Stem Cells Markedly Increases Cell Proliferation and Survival. *STEM CELLS* (2016). doi:10.1002/stem.2550
25. Forster, R. *et al.* Human intestinal tissue with adult stem cell properties derived from pluripotent stem cells. *Stem Cell Reports* **2,** 838–852 (2014).
26. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).
27. Xie, R. *et al.* Dynamic chromatin remodeling mediated by polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. *Stem Cell* **12,** 224–237 (2013).
28. Garber, K. RIKEN suspends first clinical trial involving induced pluripotent stem cells. *Nature Biotechnology* **33,** 890–891 (2015).
29. Ludwig, T. E. *et al.* Derivation of human embryonic stem cells in defined conditions. *Nature Biotechnology* **24,** 185–187 (2006).
30. Chen, G. *et al.* Chemically defined conditions for human iPSC derivation and culture. *Nat. Methods* **8,** 424–429 (2011).
31. Merkle, F. T. *et al.* Efficient CRISPR-Cas9-mediated generation of knockin human pluripotent stem cells lacking undesired mutations at the targeted locus. *Cell Rep* **11,** 875–883 (2015).
32. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).
33. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91,** 839–848 (2012).
34. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17,** 1665–1674 (2007).
35. Ganna, A. *et al. Ultra-rare disruptive and damaging mutations influence educational attainment in the general population*. (2016). doi:10.1101/050195
36. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
37. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
38. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29,** 308–311 (2001).
39. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29,** 24–26 (2011).
40. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. **26,** 2204–2207 (2010).