LETTER

doi:10.1038/nature22312

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

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

**Human pluripotent stem cells (hPS cells) can self-renew indefinitely, making them an attractive source for regenerative therapies. This expansion potential has been linked with the acquisition of large copy number variants that provide mutated cells with a growth advantage in culture[1–3]. The nature, extent and functional effects of other acquired genome sequence mutations in cultured hPS cells are not known. Here we sequence the protein-coding genes (exomes) of 140 independent human embryonic stem cell (hES cell) lines, including 26 lines prepared for potential clinical use[4]. We then apply computational strategies for identifying mutations present in a subset of cells in each hES cell line[5]. Although such mosaic mutations were generally rare, we identified five unrelated hES cell lines that carried six mutations in the *TP53* gene that encodes the tumour suppressor P53. The *TP53* mutations we observed are dominant negative and are the mutations most commonly seen in human cancers. We found that the *TP53* mutant allelic fraction increased with passage number under standard culture conditions, suggesting that the P53 mutations confer selective advantage. We then mined published RNA sequencing data from 117 hPS cell lines, and observed another nine *TP53* mutations, all resulting in coding changes in the DNA-binding domain of P53. In three lines, the allelic fraction exceeded 50%, suggesting additional selective advantage resulting from the loss of heterozygosity at the *TP53* locus. As the acquisition and expansion of cancer-associated mutations in hPS cells may go unnoticed during most applications, we suggest that careful genetic characterization of hPS cells and their differentiated derivatives be carried out before 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 just some of the 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 the analysis of sequencing data from hES cells might reveal previously unappreciated mosaic mutations and mutation-driven expansions acquired during hES cell culture. This approach would complement previous studies describing culture-derived chromosomal-scale aneuploidies and megabase-scale copy number variants (CNVs) in hPS cells[1,8,9].

To this end, we collected and performed whole-exome sequencing (WES) of hES cell lines that were listed on the registry of hES cell lines maintained by the US National Institutes of Health (NIH) (Fig. 1a, b) and were able to obtain, bank and sequence 114 independent hES cell lines (Fig. 1c–e). We selected cell lines at low to moderate passage (P) numbers (mean P18, range P3–P37) and cultured them in a common set of growth conditions for an average of $2.7 \pm 0.7$ ($\pm$ s.d.) passages (range 2–6 passages) before banking and sequencing (Fig. 1f, g). As hES-cell-derived differentiated cells are currently being evaluated in clinical trials for their safety and utility in a range of diseases such as macular degeneration[10], we also obtained genomic DNA from an additional 26 independent hES cell lines that had been prepared under good manufacturing practice (GMP) conditions for potential clinical use (Fig. 1a, c, e, g). We performed WES of these 140 hES cell lines from 19 institutions to a mean read depth of $79.7 \pm 0.1$ ($\pm$ s.e.m.) (range, 57 to 115 for UM4-6 to UM78-2) (Fig. 1h). Further details on cell line acquisition and selection are in Supplementary Table 1 and Methods.

To identify acquired mutations, we analysed the sequencing reads at high-quality, high-coverage heterozygous sites across the exome. To eliminate most inherited polymorphisms from consideration, we restricted this search to variants found in only 1–2 of the cell lines sequenced and in fewer than 0.01% of the individuals sampled by the Exome Aggregation Consortium (ExAC Database)[11]. The allelic fractions at which most remaining variants were represented among the sequence reads for any one hES cell DNA sample followed a binomial distribution, reflecting statistical sampling around the 50% level expected of inherited alleles (Fig. 2a); at a much smaller set of sites, 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. This search identified 263 candidate mosaic variants, of which 28 were predicted to have a damaging or disruptive effect on gene function (Supplementary Table 2).

The only gene affected by more than one such mutation was the tumour suppressor gene *TP53*. We identified six *TP53* mutations in five unrelated hES cell lines (Supplementary Table 3), including a GMP-prepared cell line (MShef10) that carried two distinct *TP53* variants (G245S and R248W). These six missense mutations, although rare (<0.01%) in the general population[11] (Fig. 2c), mapped to the four residues most frequently disrupted in human cancer[12–14] (Fig. 2d, Supplementary Table 3). As P53 is mutated in approximately 50% of tumours[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 in complex with DNA, each of the mutations mapped to the DNA-binding domain of

[1]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [4]Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA. [5]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [6]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [7]The Azrieli Center for Stem Cells and Genetic Research, Institute of Life Sciences, Hebrew University of Jerusalem, Givat-Ram, Jerusalem 91904, Israel. [8]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. †Present addresses: 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 (F.T.M.); Stem Cell Research, Biogen, 115 Broadway, Cambridge, Massachusetts 02142, USA (S.M.).
*These authors contributed equally to this work.

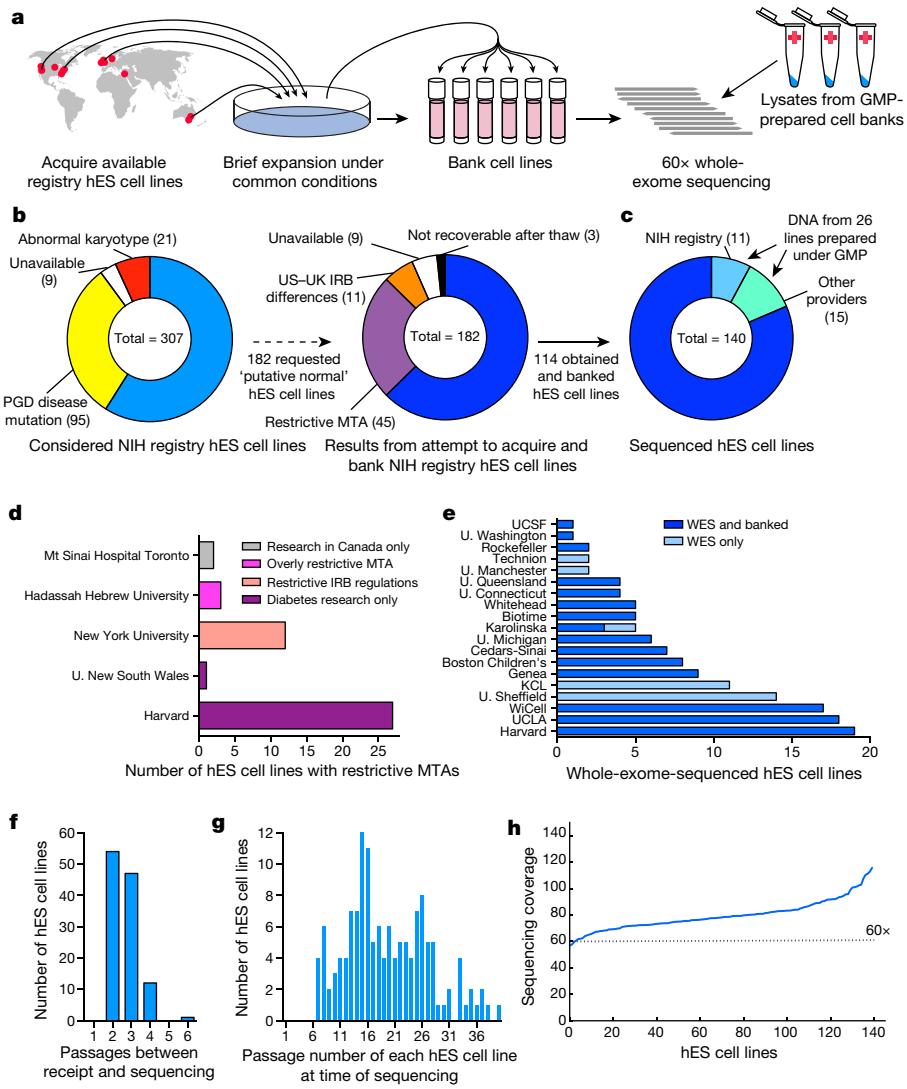



**Figure 1 | Acquisition and WES of 140 hES cell lines.** **a**, Schematic workflow for hES cell line acquisition and sequencing. **b**, **c**, 114 hES cell lines were obtained, banked (**b**), and analysed by WES along with 26 GMP-prepared cell lines (**c**). **d**, 45 hES cell lines were excluded owing to use restrictions. **e**, 140 hES cell lines were banked and/or sequenced (see also Supplementary Table 1 and Methods). **f**, HES cells were minimally cultured before banking and sequencing. **g**, Cumulative passage number of hES cells was moderate. **h**, WES coverage for sequenced hES cell lines. IRB, institutional review board; MTA, material transfer agreement; PGD, pre-implantation genetic diagnosis.

P53 (ref. 17) (Fig. 2e, f). Mutations at these positions are associated with cancer and act in a dominant negative fashion to diminish P53-mediated 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, tumours 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, we developed droplet digital PCR (ddPCR) assays to count the abundance of each allele at the four *TP53* mutation sites (Fig. 3a, b, Supplementary Table 4). Analysis of genomic DNA derived from the 140 hES cell lines confirmed that all six mutations identified by WES were indeed mosaic, with allelic fractions 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 that such mutations 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 hES cells was an acquired mutation and that cells with the mutation had come to represent a significant fraction of cells in affected lines.

We next asked whether the cells harbouring these *TP53* mutations expanded their representation within the hES cell population across passages. We re-obtained early passage vials for hES cell lines that were mosaic for *TP53* mutations (CHB11 at P22, and WA26 at P13), thawed a fresh vial of ESI035 at P36, and analysed 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 hES cell lines, *TP53* mutant alleles increased in representation over passages (Fig. 3d) in all but one experiment, suggesting that *TP53* mutations conferred a strong selective advantage (approximately 1.9-fold per passage) under routine culture conditions (Fig. 3e, Extended Data Fig. 1, Supplementary Table 5). To confirm that this selective advantage was conferred by *TP53* mutations and not by CNVs at chr20q11.21 (refs 1–3), we analysed all 140 hES cell lines using single nucleotide polymorphism (SNP) arrays and found that none of the lines carrying *TP53* mutations also carried the chr20q CNV (Supplementary Table 6). Our results are consistent with a model in which the routine culture of hPS cells selects for mutations that inactivate P53, resulting in the rapid clonal expansion of such mutations when they occur. Indeed, it has previously been reported that loss of P53 activity facilitates the reprogramming of somatic cells to pluripotency[22,23] and promotes hPS cell survival and proliferation[24], suggesting a prominent role for P53 in regulating self-renewal in hPS cells (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 (RNA-seq) data from 251 hPS cell samples in 57 published studies, corresponding to 13 hES cell and 104 human induced PS cell (hiPS cell) lines (Fig. 4a–c, Supplementary Table 7). The relatively high expression of *TP53* in hPS cells provided sufficient read depth for allelic counting and allowed us

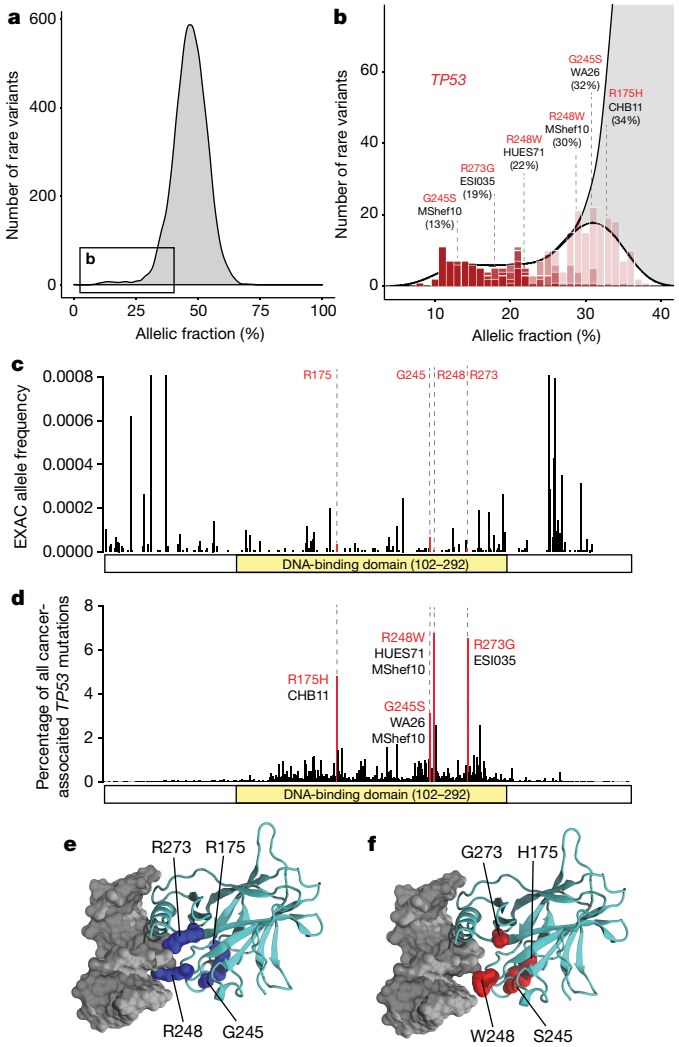



**Figure 2 | Identification of recurrent, cancer-associated *TP53* mutations in hES cells. a**, Some heterozygous variants are present at low allelic fractions (boxed left) in hES cells. **b**, **c**, Likely mosaic variants ($P < 0.01$, binomial test, red shading), include six mutations in *TP53* (**b**, Supplementary Table 3) that are rare in ExAC (<0.0001) (**c**). **d**, The four affected P53 residues are commonly mutated in human tumours. **e**, On a crystal structure of P53 bound to DNA, the affected residues map to the DNA binding domain and include arginine (R) residues that directly interact with DNA. **f**, The residues mutated in hES cells disrupt DNA binding by P53.

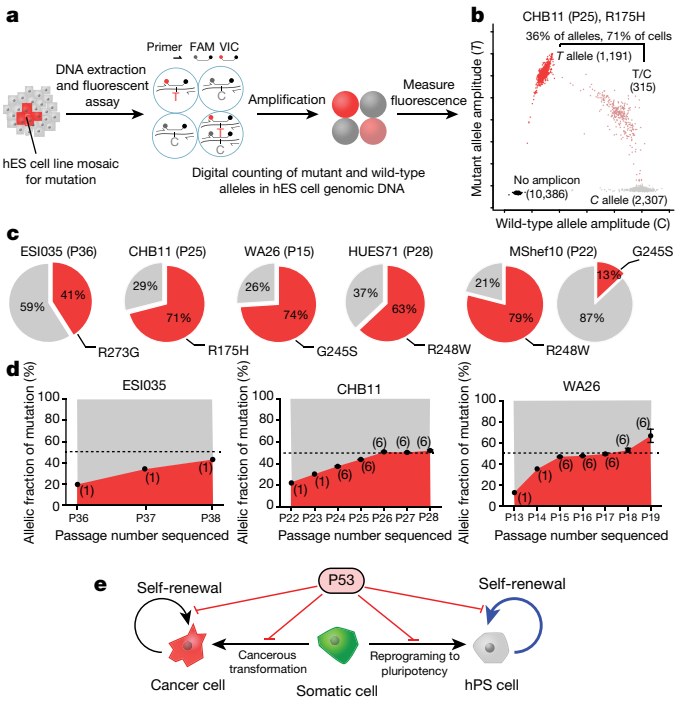



**Figure 3 | *TP53* mutations in hES cells are mosaic and confer strong selective advantage. a**, ddPCR assay schematic. **b**, Representative ddPCR data showing droplets containing the reference allele (grey), mutant allele (red), both alleles (pink), or neither allele (black). **c**, Estimated fraction of mutant cells (red) in affected hES cell lines. **d**, Mutant allelic fraction rapidly increases during standard hES cell culture. Error bars depict s.e.m. and numbers indicate replicate wells. Similar results were observed in replicate experiments (Extended Data Fig. 1). Note further allele-fraction expansion (after P17) for WA26, probably involving LOH. **e**, Model of the role of P53 in both cancer and stem cell biology.

to identify nine instances of eight disparate point mutations in *TP53*. These eight variants were all distinct at the nucleotide level from those we had previously seen by WES. However, like the mutations ascertained by WES, each of these variants led to missense substitutions in the DNA-binding domain of P53 (Fig. 4d–f, Supplementary Table 3, Extended Data Fig. 2). When we considered both WES and RNA-seq datasets, we identified four codons that were recurrently mutated in hPS cells: R181, G245, R248, and R273. Notably, the commonly used WA09 (H9) hES cell line manifested four distinct *TP53* mutations (P151S, R181H, R248Q, and R267W) in different laboratories, further demonstrating that the mutations arose during cell culture (Fig. 4h).

Of the 15 instances of *TP53* mutations observed by either WES or RNA-seq, the percentage of mutant reads suggested that 10 were mosaic and that 3 had reached fixation (allelic fraction of 50%). Surprisingly, *TP53* mutations in two cell lines, WA09 (R248Q) and WIBR3 (H193R), were present in 80% ± 3% and 100% of reads, respectively (Supplementary Table 3). 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 along chromosome 17 including the *TP53* locus (Extended Data Fig. 3) that was almost 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 confers additional selective advantage.

We next determined whether *TP53* mutations affect cell differentiation or affect the survival of differentiated cells. To this end, we examined studies in which there was RNA-seq data for both hES cells and their differentiated progeny. Cell lines with substantial fractions 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 mosaic G245C mutation in *TP53* expanded in 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 hPS cells that harbour cancer-related missense mutations in *TP53*, sometimes accompanied by LOH. These findings have practical implications for the use of hES cells in disease modelling and transplantation medicine. The fact that we observed *TP53* mutations among both hES cells and hiPS cells cultured with a wide variety of media, substrates, and passaging methods (Extended Data Fig. 4) suggests that new culture conditions should be explored to reduce the selective pressure for *TP53* mutations. We also suggest regular genetic testing of hPS cells, particularly before and after stressful interventions such as gene editing or single-cell cloning that force hPS cell populations through bottlenecks (Fig. 4l, m). Our specific findings here suggest that the P53 pathway could be an immediate focus for these genetic tests. A comprehensive ascertainment of recurrent culture-acquired

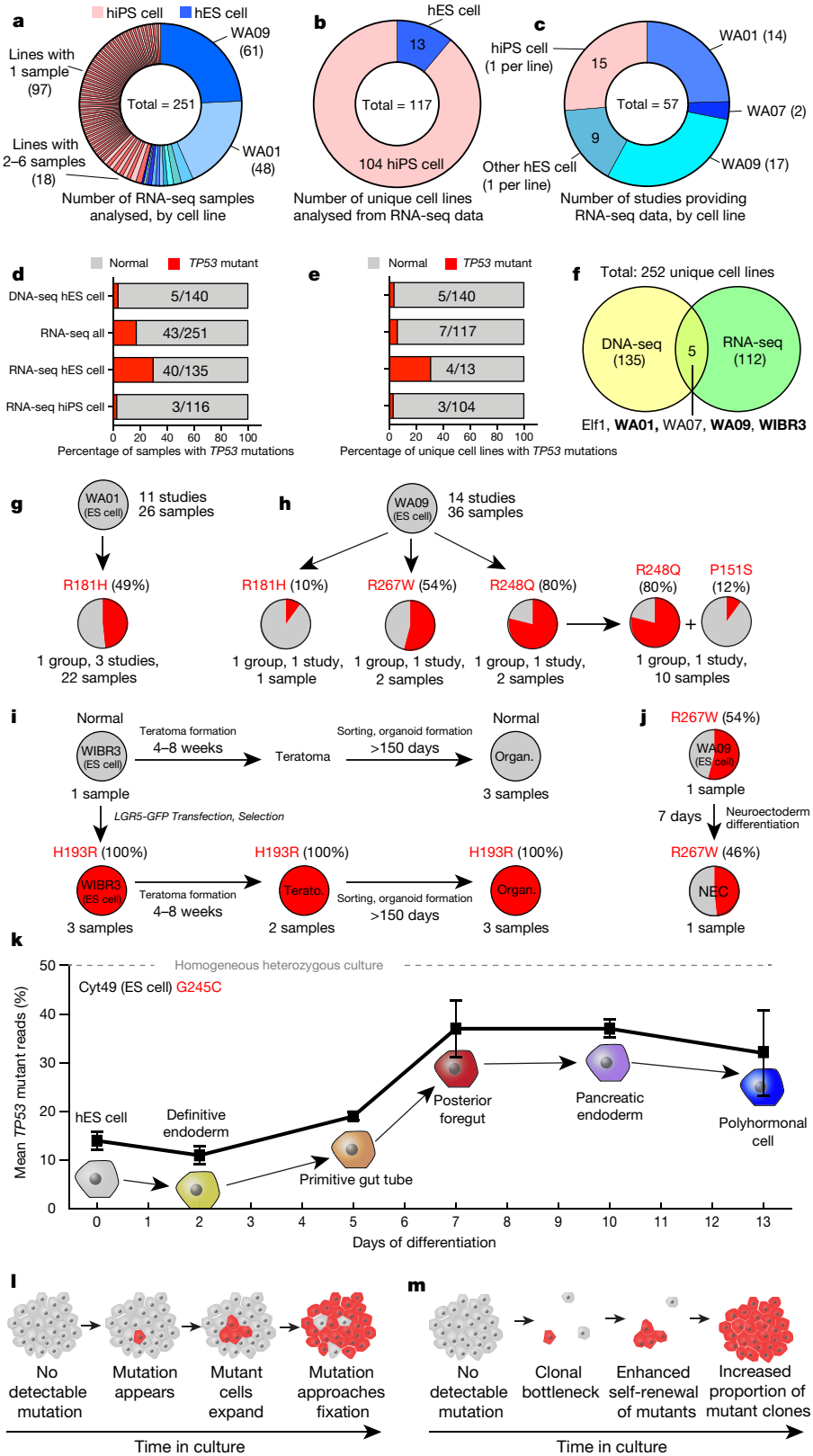



**Figure 4 | A substantial fraction of hPS cells in published studies harbour *TP53* mutations.** **a**–**e**, Published RNA-seq data show that 7 out of 117 (6%) unique hPS cell lines harbour P53 mutations. **f**, Combined DNA-seq and RNA-seq analysis reveals 12 out of 252 (5%) distinct cell lines affected by 15 *TP53* mutations (Supplementary Table 3). **g**–**i**, P53 mutant WA01 was seen in three studies (**g**), WA09 acquired four distinct *TP53* mutations in three groups (**h**), and WIBR3 lost all normal copies of *TP53* after gene editing (**i**). **j**, **k**, *TP53* mutant cells could be differentiated (**j**, **k**), and expanded relative to wild-type cells (**k**). Values are the mean of 2–3 replicate samples, and error bars depict s.e.m. **l**, **m**, Model of *TP53* mutation enrichment during hPS cell culture (**l**) or during clonal bottlenecks (**m**). Organ., gut organoid; NEC, neuroectodermal cell; Terato., Teratoma.

mutations 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 can detect potentially harmful mutations in differentiated cell preparations derived from hPS cells, providing an opportunity to increase the safety of cell replacement therapies for conditions ranging from diabetes to Parkinson's disease. Clinical trials with hPS-cell-derived materials had recently been halted owing to the discovery of undisclosed mutations[28], though such trials have since resumed. We suggest that hPS cells 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 hPS cells; and as an end-stage release criterion before the transplantation of the hPS cell-derived cellular product. Importantly, although *TP53* mutations recurred at detectable fractions in several cell lines, most lines (around 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.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 30 August 2016; accepted 31 March 2017.**
**Published online 26 April 2017.**

1. The International Stem Cell Initiative. ISCI. Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. *Nat. Biotechnol*. **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 Rep.* **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. *Hum. Mol. Genet.* **17,** R48–R53 (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. *Nat. Biotechnol.* **25,** 803–816 (2007).
9. Baker, D. *et al.* Detecting genetic mosaicism in cultures of human pluripotent stem cells. *Stem Cell Rep.* **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 Res.* **43,** D805–D811 (2015).
13. Zhang, J. *et al.* International Cancer Genome Consortium data portal—a one-stop shop for cancer genomics data. *Database* **2011,** 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. III, 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. Marión, 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).
25. Forster, R. *et al.* Human intestinal tissue with adult stem cell properties derived from pluripotent stem cells. *Stem Cell Rep.* **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. *Cell Stem Cell* **12,** 224–237 (2013).
28. Garber, K. RIKEN suspends first clinical trial involving induced pluripotent stem cells. *Nat. Biotechnol.* **33,** 890–891 (2015).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank the many institutions and investigators worldwide that provided their cell lines and supported the publication of the results. We are indebted to D. Santos, M. Smith, K. Elwell, M. A. Yram, S. Ellender, L. Bevilacqua, and D. Gage for their assistance with the regulatory and logistical efforts required to acquire and sequence hES cell lines. We also thank K. Lilliehook for her comments, I. Yildirim for his assistance with the molecular modelling of P53 mutations, and C. Usher for help with figure schematics. We regret the omission of any relevant references or discussion due to space limitations. The Genomics Platform at the Broad Institute performed sample preparation, sequencing, and data storage. Y.A. is a Clore Fellow. N.B. 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 hES cell lines were supported by HHMI and the Stanley Center for Psychiatric Research. F.T.M., S.A.M., and K.E. were supported by grants from the NIH (5P01GM099117, 5K99NS08371). K.E. was supported by the Miller consortium of the HSCI, and F.T.M. 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** F.T.M., S.G., S.A.M., and K.E. conceived the project. F.T.M. and K.E. acquired hES cell lines with the assistance of M.C. and G.S., who also assisted with regulatory issues pertaining to sequencing and data distribution. F.T.M. cultured and banked hES cell lines, prepared them for sequencing, and coordinated efforts to interpret and visualize sequencing data with the assistance of S.G. S.G. performed computational data analysis and visualization with the help of G.G., R.E.H., and S.K. Y.A. preformed the analysis of *TP53* mutations in the RNA-seq database with the assistance of S.B. and N.B. Data were interpreted by F.T.M., S.G., N.K., G.G., Y.A., S.B., N.B., S.A.M., and K.E. N.K., J.M., and C.M. designed, performed and analysed experiments to measure the mosaic nature and competitive expansion of *TP53* mutations. S.M. derived HUES 68, 69, 70, 74, 75, and D.I. provided the KCL lines. F.T.M., S.G., S.A.M., and K.E. prepared drafts of the manuscript text and figures with contributions and comments from all authors.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to K.E. (eggan@mcb.harvard.edu) or S.A.M. (mccarroll@genetics.med.harvard.edu).

## METHODS

**Data reporting.** No statistical methods were used to predetermine sample size. All cells were tested for mycoplasma and included for analysis only upon testing negative. The identity of all cell lines was confirmed by whole-exome sequencing and SNP array analysis. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**hES cell acquisition.** As a source of hES cells for this study, we focused on those that had been voluntarily listed by research institutions on the registry of hES cell lines maintained by the US National Institutes of Health (NIH) (http://grants.nih.gov/stem_cells/registry/current.htm). As of 8 July 2015, a total of 307 hES cell 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 owing to differences in human subjects research regulations between the US and the UK. 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 Supplementary Table 1.

The generation of hES cells used in this study was previously approved by the institutional review boards (IRBs) of all providing institutions. Use of the hES cells for sequencing at Harvard was further approved and determined not to constitute Human Subjects Research by the Committee on the Use of Human Subjects in Research at Harvard University.

**hES cell culture.** In a separate document, we describe in detail a protocol for the adaptation of hES cell lines from diverse culture conditions (F. Merkle and K. Eggan, unpublished work). In brief, we considered that different laboratories employ different methods to culture hES cells, raising the question of how best to thaw and culture the cell lines we obtained from multiple sources. Traditionally, hES cells 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, hES cells have been cultured 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 (ref. 29), E8 (ref. 30) or similar in the absence of feeder cells. In previous work, we found that a medium containing an equal volume of KOSR-based hES cell medium (KSR) and mTeSR1 (STEMCELL Technologies) (KSR–mTeSR1) robustly supports the pluripotency of hES cells undergoing antibiotic selection during the course of gene-targeting experiments under feeder-free conditions[31]. To minimize stress to hES cells 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 hES cell medium on a monolayer of irradiated MEFs. After 24 h, 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 hES cell morphology and with a diameter of approximately 400 μm were transferred into KSR–mTeSR1 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 min) incubation in 0.5 mM EDTA in PBS followed by gentle trituration in KSR–mTeSR1 medium containing 10 μM Y-27632 and re-plating. Once cultures had reached approximately 90% confluence in one well of a six-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 cryoprotective medium containing 50% KSR–mTeSR1, 10 μM Y-27632, 10% DMSO, and 40% fetal bovine serum (HyClone). A subset of hES cell lines (Supplementary Table 1) 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 hES cell lines, or were obtained directly from institutions providing GMP-grade hES cell lines. Cell pellets were digested overnight at 50 °C in 500 μl lysis buffer containing 100 μg $ml^{-1}$ 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 X10 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 on 31 July 2015. Sequencing quality and coverage were analysed using Picard tool metrics. Cross sample contamination was estimated using VerifyBamID (v1.1.2)[33], and none was detected. Data from each cell line were 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 showed other chromosomal aberrations or smaller regional changes in copy number, additional genotyping of the 140 hES cell 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 (http://penncnv.openbioinformatics.org). All CNVs were manually reviewed and are shown in Supplementary Table 6.

**Mosaic variant analysis.** To identify candidate mosaic variants, a table of heterozygous variants was generated from the VCF (Supplementary Table 2). To limit the frequency of false positive calls due to sequencing artefacts and PCR errors, variants were included if they had a variant read depth of at least 10, if they were either flagged as a 'PASS' site or were not reported in the Exome Aggregation Consortium (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 (9,490 variants present in 1–2 of the 140 samples). Coverage was calculated by summing reference and alternate allele counts for each variant. Allelic fraction 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 centred 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 (Fig. 3). A more stringent computational screen based on a *P*-value threshold of $1 \times 10^{-7}$ 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 Supplementary Table 2.

Variant annotation was performed using SnpEff with GRCh37.75 Ensembl gene models. Variants with moderate effect 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 compiles 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 hES cells were affected in the general population. We then consulted the COSMIC[12] (http://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=TP53), ICGC[13] (https://dcc.icgc.org/), and IARC P53 (ref. 14) (http://p53.iarc.fr/TP53SomaticMutations.aspx) databases and plotted the percentage of tumours carrying a mutation in each codon (Fig. 2d, Extended Data Fig. 2b).

**Molecular modelling of P53 protein.** To visualize the spatial location of the amino acid residues affected by *TP53* mutations observed in hES cells by WES on the P53 protein, we downloaded the 1.85 Angstrom X-ray diffraction-based structure file from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (file 2AHI) and built the model protein/DNA system (chain IDs D, G, and H) to visualize the secondary structure of a P53 monomer complexed to DNA as a

ribbon diagram. DNA was illustrated as a space-filling model. Water molecules were discarded when building the wild-type model and minimized in two steps using the AMBER 16 package[36]. 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 (Supplementary Table 3) in the 140 hES cell 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 labelled 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 (Supplementary Table 4). Genomic DNA from each hES cell line was analysed 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 hES cell culture and calculation of *TP53* mutation expansion.** To assess how the allelic fraction of *TP53* mutations might change over time in culture, hES cell 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 on the day of passaging. 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 analysed by ddPCR. To calculate the relative expansion rate of mutant relative to wild-type cells, we applied the following formula:

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

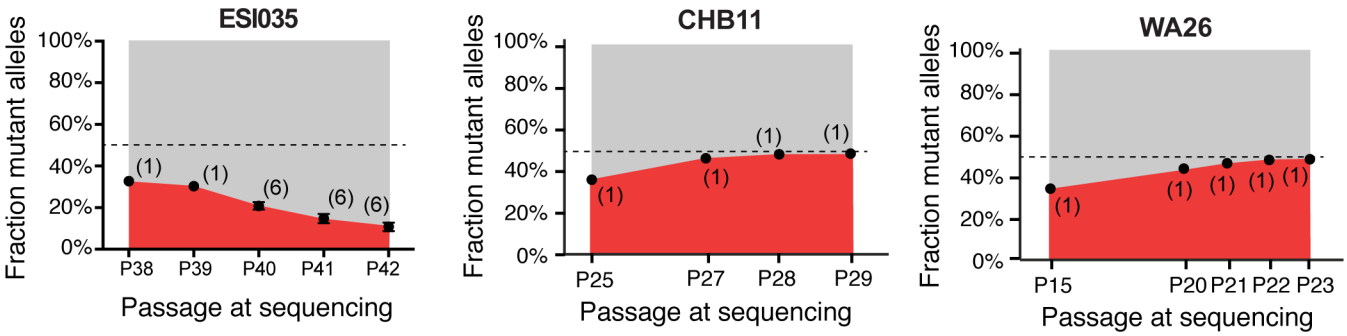
where $R_0$ is defined as the ratio of (variant positive cells)/(variant negative cells) after some number of starting passages and $R_1$ and $R_2$ represent the aforementioned ratios measured on the same sample at $T_1$ and $T_2 > T_1$ passages respectively. From this equation, the estimation of variant 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 Supplementary Table 5. 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 1,000 droplets), respectively.

**RNA sequencing analysis.** In order to identify *TP53* mutations in hPS cells, we analysed 256 publicly available high-throughput RNA sequencing samples of hPS cells from the SRA database[37] (http://www.ncbi.nlm.nih.gov/sra). Data accession numbers for SRA (and GEO, where applicable) are provided in Supplementary Table 7. 5 of these 256 samples were not considered further as they were from single cells rather than cell lines. Following sequence alignment to the hg19 human reference genome with Tophat2 (ref. 38), 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 hES cell lines and 3 hiPS cell lines) had a missense mutation in *TP53*. 10 of the 40 hES cell samples (WA09) carried two separate mutations (Supplementary Table 7). Upon the identification of cell lines carrying mutant reads, RNA sequencing data from studies containing differentiated samples were included for analysis.

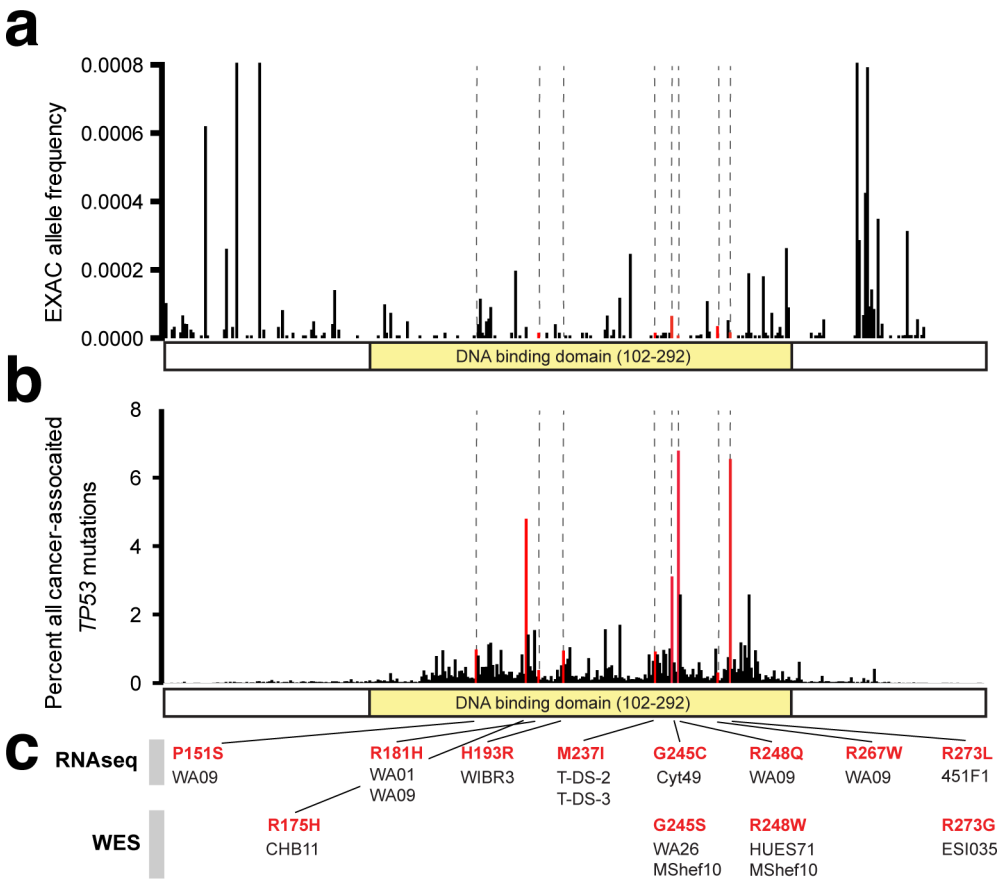
**Loss of heterozygosity analysis.** In order to evaluate *TP53* alleles, we assessed the level of polymorphism by calculating the ratio between the minor and major alleles across chromosome 17. So as to minimize sequencing noise and errors, we included SNPs covered by more than 10 reads and that are located in the dbSNP build 142 database[39]. The resulted wig files were then plotted using Integrative Genomics Viewer (IGV)[40] (Extended Data Fig. 4). In order to quantify the difference in polymorphism between samples, we converted the wig files to BigWig using UCSC Genome Browser utilities[41] 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. Whereas 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 $16 \times 10^6$ base pairs), including the *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.

**Data availability.** Sequence data from cell lines listed on the NIH hES cell registry have been deposited in the NCBI database of Genotypes and Phenotypes (dbGaP) under accession number phys001343.v1.p1 (at https://www.ncbi.nlm.nih.gov/gap/?term=phys001343.v1.p1). Sequence data from the remaining cell lines reported in our study have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001002400 (at https://www.ebi.ac.uk/ega/search/site/EGAS00001002400).

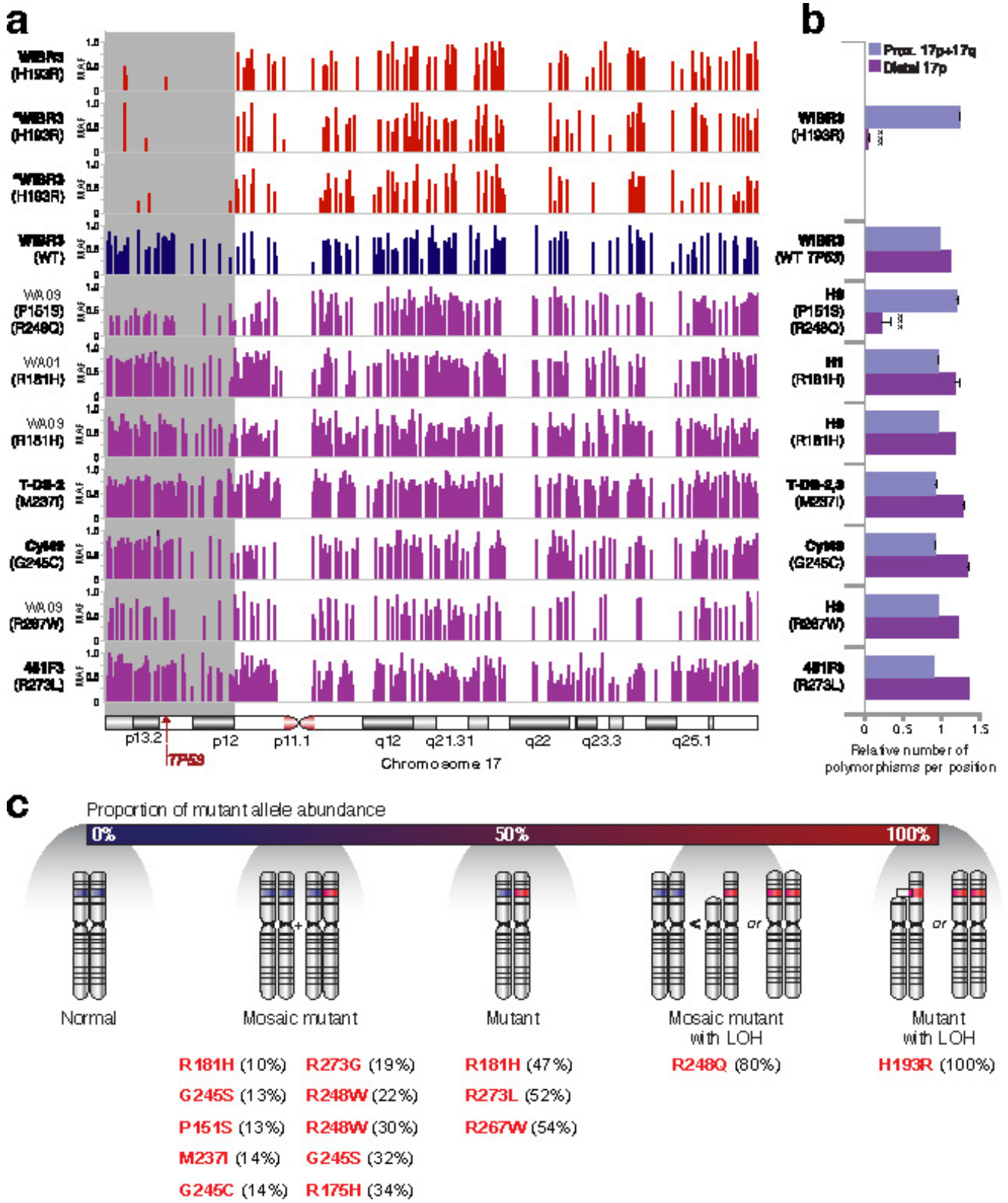

29. Ludwig, T. E. *et al.* Derivation of human embryonic stem cells in defined conditions. *Nat. Biotechnol.* **24,** 185–187 (2006).
30. Chen, G. *et al.* Chemically defined conditions for human iPSC derivation and culture. *Nat. Methods* **8,** 424–429 (2011).
31. Merkle, F. T. *et al.* Efficient CRISPR–Cas9-mediated generation of knockin human pluripotent stem cells lacking undesired mutations at the targeted locus. *Cell Reports* **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. **19,** 1563–1565 (2016).
36. Case, D. A. *et al.* AMBER 2016.
37. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
38. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
39. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
40. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).
41. 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).



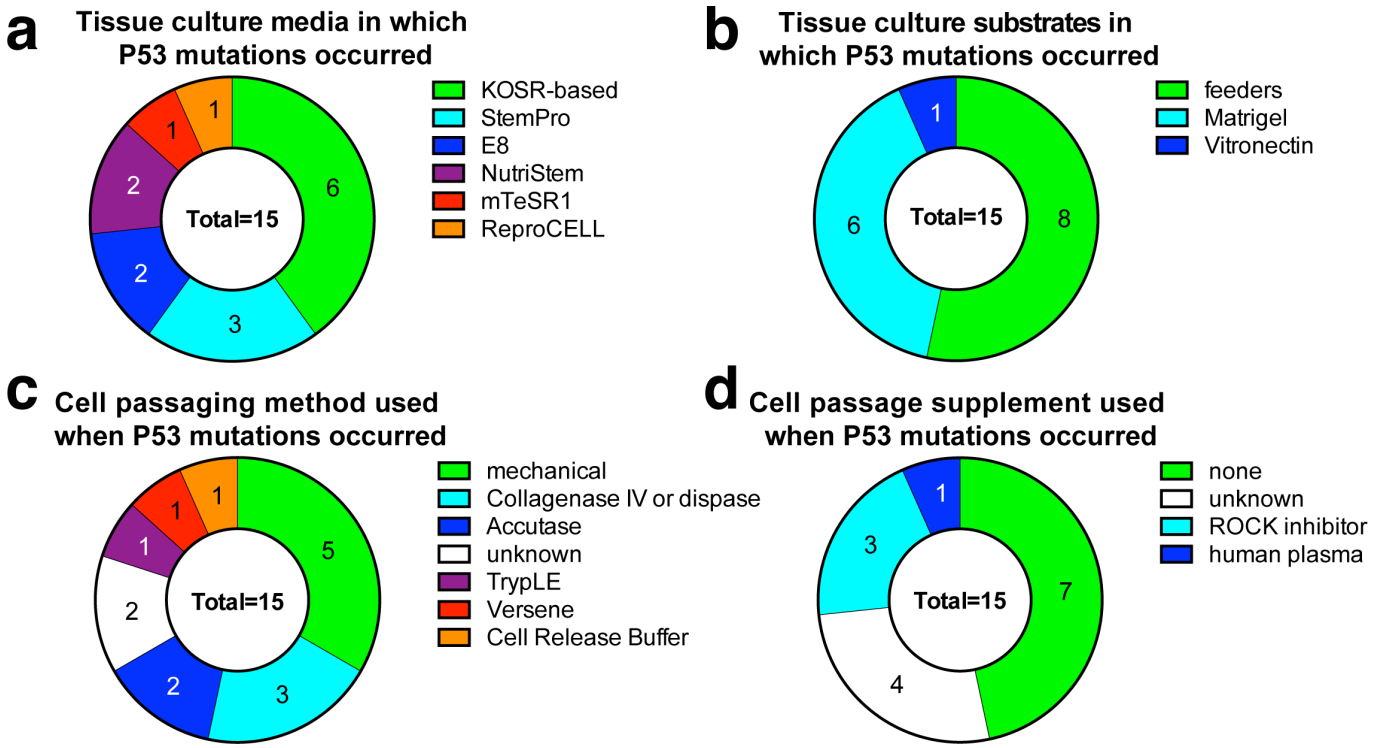

**Extended Data Figure 1 | Replicates of cell competition assays carried out at earlier starting passages.** Note that while the mutant allelic fractions for lines CHB11 and WA26 approach fixation, that the fraction of mutant cells unexpectedly decreases for ESI035 over several passages, indicating a potential selective disadvantage that co-segregates with the *TP53* mutation in this experiment. The number of replicate wells is indicated in each graph. Values depict the mean and error bars depict s.e.m.




**Extended Data Figure 2 | 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-seq hPS cell 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 and the method used to detect them are pictured. Although the M237I event is seen in two distinct hiPS cell lines, it is conservatively counted as a single event as the two affected clones may be derived from a common reprogrammed progenitor.




**Extended Data Figure 3 | Analysis of loss of heterozygosity in RNA sequencing samples. a**, Polymorphic sites on chromosome 17 in different hPS cells 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. Asterisk indicates samples with less than 25 reads. **b**, Ratio between the fraction of polymorphic alleles in the distal part of chromosome 17p or the reminder of chromosome 17 (proximal 17p + 17q) compared to that fraction for the entire chromosome 17. Values shown depict mean. Where present, error bars depict s.e.m. for 2–22 replicate samples. $***P < 0.001$, one-sided *Z*-score test for the two population proportion. 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 LOH. **c**, A schematic representation of possible allele states of *TP53* in cultured hPS cells 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. MAF, minor allele frequency.




**Extended Data Figure 4 | Culture and passaging method employed for samples bearing P53 mutations. a**, P53 mutations were observed in hPS cells grown in a broad array of culture media including media supplemented with knockout serum replacement (KOSR), and defined, commercial media such as E8. **b**, P53 mutations were observed from cells grown with feeder cells or under feeder-free conditions. **c**, As passaging hPS cells can introduce stresses or clonal bottlenecks, we examined whether P53 mutations were consistently seen when a particular passaging method was used and 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**, P53 mutations are seen in studies that either include or exclude supplements such as the rock inhibitor Y-27632 (10 μM) at the time of passaging.