Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells

Alexej Abyzov1,2,3, Jessica Mariani1,4*, Dean Palejev1,4*, Ying Zhang1,5*, Michael Seamus Haney6,7*, Livia Tomasiñi1,4*, Anthony F. Ferrandino1,4*, Lior A. Rosenberg Belmaker1,4, Anna Szekely1,5,8, Michael Wilson1,2,4, Arif Kocabas1,4, Nathaniel E. Calixto1,4, Elena L. Grigorenko1,4,9,10, Anita Huttner1,11, Katarzyna Chawarska1,4, Sherman Weissman1,5, Alexander Eckehart Urban1,6,7, Mark Gerstein1,2,3,12 & Flora M. Vaccarino1,4,13

Reprogramming somatic cells into induced pluripotent stem cells (iPSCs) has been suspected of causing de novo copy number variation1–8. To explore this issue, here we perform a whole-genome and transcriptome analysis of 20 human iPSC lines derived from the primary skin fibroblasts of seven individuals using next-generation sequencing. We find that, on average, an iPSC line manifests two copy number variants (CNVs) not apparent in the fibroblasts from which the iPSC was derived. Using PCR and digital droplet PCR, we show that at least 50% of those CNVs are present as low-frequency somatic genetic variants in parental fibroblasts (that is, the fibroblasts from which each corresponding human iPSC line is derived), and are manifested in iPSC lines owing to their clonal origin. Hence, reprogramming does not necessarily lead to de novo CNVs in iPSCs, because most of the line-manifested CNVs reflect somatic mosaicism in the human skin. Moreover, our findings demonstrate that clonal expansion, and iPSC lines in particular, can be used as a discovery tool to reliably detect low-frequency CNVs in the tissue of origin. Overall, we estimate that approximately 30% of the fibroblast cells have somatic CNVs in their genomes, suggesting widespread somatic mosaicism in the human body. Our study paves the way to understanding the fundamental question of the extent to which cells of the human body normally acquire structural alterations in their DNA post-zygotically.

The ability to derive iPSCs from somatic cells5–8 has opened exciting new possibilities for the study of human development and regenerative medicine, and has enabled the effect of human genetic variation on developmental processes to be examined9–13. However, all of these applications require that iPSCs, clonal cell lines each derived from one or just a few somatic cells, stably maintain the genetic background of the individual from whom they are derived. Yet, there are reports of genomic instability in stem and precursor cells, indicating that copy number variation/structural variation might arise in iPSCs, in addition to single base-pair changes14–16. These variations could be caused by the de-differentiation procedures, result from extensive time in culture, or pre-exist in the somatic tissue of origin at low frequency. Emerging evidence suggests potentially widespread genomic mosaicism not only in cancer but also in somatic cell lineages, as a result of errors during DNA replication, DNA repair, mitosis and mobilization of transposable elements17–21. Such a phenomenon could have far-reaching physiological consequences yet is still poorly understood and very difficult to study22–24. The derivation of iPSCs offers the opportunity to analyse the genome of a single cell at high resolution and sensitivity.

Using the canonical retroviral method, we have produced 21 human iPSC (hiPSC) lines derived from skin fibroblasts collected from seven members of two families (Supplementary Fig. 1). The hiPSC lines were characterized by four sets of quality control criteria: (1) morphology; (2) expression of pluripotency factors at the protein level; (3) gene expression analyses (reverse transcription PCR (RT–PCR), microarrays, complete transcriptome by high-throughput RNA sequencing (RNA-seq)); and (4) demethylation of canonical pluripotency factor promoters (Supplementary Figs 2 and 3 and Supplementary Tables 1 and 2). This thorough evaluation (Supplementary Information) showed there was extensive similarity between our hiPSCs and human embryonic stem cells, and divergence of hiPSCs from the fibroblasts, indicating complete reprogramming. Finally, by using neuronal differentiation assays, we found that the hiPSCs exhibited comparable propensities for neural lineage differentiation (Supplementary Fig. 4).

We then generated one lane of whole-genome paired-end sequencing data on the Illumina HiSeq platform for 20 hiPSC lines, and predicted CNVs in hiPSC lines with CNVnator26 (Supplementary Fig. 1b). CNVnator uses read-depth analysis and was shown to have the highest sensitivity in confirming CNVs previously discovered with arrays and fosmid sequencing27. First, we discovered CNVs in fibroblast and hiPSC samples by comparison with the reference human genome, and then compared genotypes of each hiPSC line to their respective fibroblast cell population of origin to identify the variants manifested only in hiPSCs, that is, line-manifested CNVs (LM-CNVs). We were able to discover CNVs as small as 2 kilobases (kb), but the highest sensitivity was for CNVs of at least 5 kb in size (Supplementary Fig. 5). Using conservative criteria, we predicted a total of 74 LM-CNVs in all 20 lines (Supplementary Table 3), that is, just a few LM-CNVs per line. Similar numbers of LM-CNVs per line were observed for a few other hiPSC lines produced by the episomal method (Supplementary Information).

We observed positive yet non-significant correlations between the number of LM-CNVs and the passage number at which hiPSC lines were sequenced (Fig. 1a). Neither more relaxed CNV calling nor more sensitive criteria for LM-CNV identification made the correlation significant. LM-CNVs represent a small fraction of all CNVs that were discovered in hiPSC lines by comparison with the reference human genome and performing read-depth analysis at higher coverage (~20×) did not change the proportion of LM-CNVs versus the total number of CNVs (Fig. 1b). Even with sensitive criteria for LM-CNV prediction, their fraction did not exceed 17%. As a positive control and using the same approach, we compared an hiPSC line to the fibroblasts
Figure 1 | Characterization of candidate LM-CNVs with respect to passage number and total CNVs. a, The number of LM-CNVs does not show significant changes with respect to passage, irrespective of the sensitivity of our detection criterion. Throughout this paper, conservative criteria (blue symbols) were used unless noted. b, The percentage of LM-CNVs of all CNVs detected in hiPSCs by comparison with the reference human genome; square symbols represent data obtained at increased (20×) coverage. LM-CNVs represent a small fraction the total number of CNVs in a person. c, Counts of LM-CNVs in of an individual from the other family and observed roughly 40 different CNVs per hiPSC line, Fig. 1c), which is consistent with inter-individual variations in a similar size range, as described previously27.

Discordant paired-end reads analysis confirmed 22 LM-CNVs discovered by read-depth analysis (Supplementary information). For 39 of the most confident predictions, we performed quantitative PCR (qPCR) validation assays in early passage hiPSCs (passages 5–13), and also, when available, in late passage cells (passages 17–52) (see below). These analyses validated 33 LM-CNVs (Table 1, Supplementary Table 3 and Supplementary Figs 6–44). Validated LM-CNVs were present in 15 out of 20 (75%) hiPSC lines, with 9 (45%) hiPSC lines having more than one LM-CNV.

To test for the actual presence of somatic CNVs in the fibroblast cultures, we performed PCR amplification with diagnostic primers across CNV breakpoints in hiPSCs and the corresponding donor fibroblasts for 20 LM-CNVs with good initial estimate of their breakpoints from paired-end analyses (Fig. 2b, Table 1 and Supplementary Table 3). We observed expected bands in all cases when using hiPSC DNA, and in eight cases when using DNA from the corresponding fibroblast cultures (Supplementary Table 3; see Fig. 2a and Supplementary Fig. 55). Further evidence for genomic heterogeneity was the realization that for many CNVs, copy number ratios were deviating from 1.5, indicative of one haplotype duplication, or from 0.5, indicative of one haplotype deletion, using both read-depth analysis and their qPCR validation (Supplementary Fig. 6 and Supplementary Information).

To obtain an independent confirmation of our approach for LM-CNV detection, we analysed the hiPSC and fibroblast samples from the mother of family S1123 and the proband of family 03 by high-resolution array-based comparative genome hybridization (aCGH). All of the 10 LM-CNVs validated by qPCR (Table 1), which were found by sequencing in the hiPSCs from these individuals, were also confirmed by aCGH (Supplementary Figs 45–54). However, no further LM-CNVs could be discovered using aCGH data, as the estimated false discovery rate (FDR) of the set of other predictions was close to 100%, based on qPCR validation of a random subset (Supplementary Tables 4 and 5). These data suggest that analysis of sequencing data alone allows the discovery of all or almost all LM-CNVs. Finally, we tested by qPCR the presence of validated LM-CNVs at later passages (passages 17–52), in five hiPSC lines. We observed a strong correlation (Pearson’s coefficient 0.96) between qPCR results obtained in late versus early passages (Supplementary Fig. 6). Among 16 LM-CNVs that were tested, 87.5% were validated in late passage (Table 1), suggesting long-term stability of the hiPSC genome.

We then analysed the origin of LM-CNVs, that is, whether they had arisen de novo in the hiPSCs as a sequel to reprogramming or were present at low allele frequencies in the donor fibroblast population. The first indirect, but suggestive evidence for fibroblast somatic genomic heterogeneity was the observation of the same validated LM-CNVs (chrX:64962001–65029000) in two different hiPSC lines (3 and 4) derived from the fibroblast culture of the same individual (Table 1, Fig. 2a and Supplementary Fig. 55). Further evidence for genomic heterogeneity was the realization that for many CNVs, copy number ratios were deviating from 1.5, indicative of one haplotype duplication, or from 0.5, indicative of one haplotype deletion, using both read-depth analysis and their qPCR validation (Supplementary Fig. 6 and Supplementary Information).

To test for the actual presence of somatic CNVs in the fibroblast cultures, we performed PCR amplification with diagnostic primers across CNV breakpoints in hiPSCs and the corresponding donor fibroblasts for 20 LM-CNVs with good initial estimate of their breakpoints from paired-end analyses (Fig. 2b, Table 1 and Supplementary Table 3). We observed expected bands in all cases when using hiPSC DNA, and in eight cases when using DNA from the corresponding fibroblast cultures (Supplementary Table 3; see Fig. 2a and Supplementary Figs 7–39). For 15 LM-CNVs we also performed digital droplet PCR (ddPCR) (Fig. 2c), which allows not only the observation of low-frequency somatic CNVs but also an estimation of their allelic frequency in the somatic mosaic, with a sensitivity down to 0.1%. From the allele frequencies, cellular frequencies in the fibroblasts were calculated as explained in the Methods using the
genome allowed us to discover CNVs present in a subset of parental fibroblast cells, such that very low allele frequency variants in the original populations could be unmasked. We then used PCR and ddPCR across breakpoints to genotype CNVs in the parental fibroblasts, and estimated that 50% of the CNVs manifested in hiPSCs could be traced back to the original fibroblast population (Table 1). We may be underestimating this phenomenon because very low allele frequency somatic CNVs might still escape confirmation by PCR/ddPCR in fibroblasts owing to technical limitations. Despite this, conceptually, our approach can be used for comparison of any clonal (not only iPSC) and parental cell populations with the aim of studying somatic variation.

Overall, we found that hiPSCs manifest on average two validated CNVs larger than 10 kb, which is considerably more than in two previous studies\(^1,28\). The difference is probably attributable to us using sequencing as a more sensitive approach, see Supplementary Discussion) as opposed to single nucleotide polymorphism arrays\(^1\). One of the aforementioned studies\(^28\) also used sequencing but analysed only three hiPSC lines derived from fibroblasts, such that very low allele frequency variants in the original populations could be unmasked. We then used PCR and ddPCR across breakpoints to genotype CNVs in the parental fibroblasts, and estimated that 50% of the CNVs manifested in hiPSCs could be traced back to the original fibroblast population (Table 1). We may be underestimating this phenomenon because very low allele frequency somatic CNVs might still escape confirmation by PCR/ddPCR in fibroblasts owing to technical limitations. Despite this, conceptually, our approach can be used for comparison of any clonal (not only iPSC) and parental cell populations with the aim of studying somatic variation.

Overall, we found that hiPSCs manifest on average two validated CNVs larger than 10 kb, which is considerably more than in two previous studies\(^1,28\). The difference is probably attributable to us using sequencing (generally a more sensitive approach, see Supplementary Discussion) as opposed to single nucleotide polymorphism arrays\(^1\). One of the aforementioned studies\(^28\) also used sequencing but analysed only three hiPSC lines, therefore by extrapolating to a larger number their results could still be consistent with ours. Alternatively, bone marrow mononuclear cells may have fewer somatic variations than fibroblast cells, explaining why hiPSC lines derived from mononuclear cells in the previous study\(^28\) manifested fewer LM-CNVs than do our hiPSC lines derived from fibroblasts.

It was previously proposed that CNVs might arise in hiPSCs as a consequence of DNA damage or impaired DNA repair during reprogramming. Although we acknowledge that some CNVs might arise...
during reprogramming in some hiPSC lines, our data suggest that reprogramming per se does not obligatorily induce de novo mutations, as at least half of the LM-CNVs preexisted in parental fibroblast cells (Table 1). We also found no significant difference in the number of LM-CNVs in relation to passage number. Thus, our analyses support neither the hypothesis that hiPSCs generally have a large rate of de novo mutations nor the observation that most LM-CNVs in hiPSCs disappear in late passages. Using different parental cells and applying novo mutations, nor the observation that most LM-CNVs in hiPSCs represents a single, clonally expanded cell, we estimate that 30% (6 out of 20) of skin fibroblast cells carry large somatic CNVs. To our knowledge, this is the first such estimate. Furthermore, with ddPCR, we estimated cell frequency as high as 15% and as low as a fraction of a per cent, suggesting wide variability in the extent of fibroblast mosaicism. Although it is possible that some CNVs could have arisen during the fibroblast cell culture, as at least half of the LM-CNVs preexisted in parental fibroblast cells, we think this is unlikely given that they were passaged less than five times (and in most cases only three times) before proceeding with hiPSC generation.

It has been known for a while that somatic variants can be respon
ding an hiPSC-based study. But more importantly, this finding
may challenge widely adopted experimental designs for genetic analyses of diseases with complex inheritance in which only the genomes of lymphoblastoid cells are being analysed. By influencing the phenotype in unexpected ways, somatically acquired CNVs might represent at least part of the explanation for the challenges in identifying the genetic predisposition in some of the complex and especially in neurodevelopmental diseases, for which determining the exact loci for genetic predisposition has proven difficult.

METHODS SUMMARY

hiPSC lines were generated with Yamanaka’s four retroviral vectors. One hundred nanograms of total RNA extracted from hiPSC lines was reverse-transcribed using SuperScript III Reverse Transcriptase and random hexamers. Primers used for OCT4 (also known as POU5F1), MYC (also known as c-Myc) and SOX2 specifically detect the transcripts from the endogenous genes. Neuronal differentiation was done by slight modification of an established protocol. For microarray expression analysis, total RNA was analysed by HumanHT-12 v4 BEADCHIP Illumina microarrays (384 well design) and by RNAseq analysis by GenomeStudio using quantile normalization and background subtraction. RNA and DNA single-end and paired-end libraries for sequencing were prepared using standard Illumina protocols with minor modification. Whole genome sequenced reads were aligned with the reference genome with Burrows–Wheeler aligner (BWA). Primer pairs for qPCR were designed using ProbeFinder software from Roche Applied Science. Real-time qPCR was run using the Applied Biosystems StepOne Real-Time PCR System (ABI), with SYBR Green chemistry. RNA-seq reads were aligned to the reference genome with TopHat and processed with RSeQtools. Primers for PCR/ddPCR validation experiments were designed with the Primer3 software. ddPCR was performed using QuantiFluor machines following manufacturer’s instructions. For aCGH analysis, each sample was hybridized on a NimbleGen 4.2M CNV array under standard conditions as recommended by the manufacturer. Data from the arrays were analysed with Nexus Copy Number.

Full Methods and any associated references are available in the online version of the paper.

Received 16 January; accepted 28 September 2012.

Published online 18 November 2012.

1. Laurent, L. C. et al. Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ES and iPS during reprogramming and time in culture. Cell Stem Cell 8, 106–118 (2011).
2. Quinlan, A. R. et al. Genome sequencing of mouse induced pluripotent stem cells reveals retroelement stability and infrequent DNA rearrangement during reprogramming. Cell Stem Cell 9, 366–373 (2011).
3. Hussein, S. M. Published online 18 November 2012. Received 16 January; accepted 28 September 2012.
4. Li, Y., Skarnes, W. C. & Bradley, A. Extensive genomic copy number variation in embryonic stem cells. Proc. Natl Acad. Sci. USA 105, 17453–17456 (2008).
5. Wu, H. et al. Copy number variant analysis of human embryonic stem cells. Stem Cells 26, 1484–1489 (2008).
6. Elliott, A. M., Elliott, K. A. & Kammesheidt, A. High resolution array-CGH characterization of human stem cells using a stem cell focused microarray. Mol. Biotechnol. 46, 234–242 (2010).
7. O’Hallachain, M., Karczewski, K. J., Weissman, S. M., Urban, A. E. & Snyder, M. P. Extensive genetic variation in somatic human tissues. Proc. Natl Acad. Sci. USA http://dx.doi.org/10.1073/pnas.1213736109 (5 October 2012).
8. De, S. Somatic mosaicism in healthy human tissues. Trends Genet. 27, 217–223 (2011).
9. Bailleau, J. K. et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479, 534–537 (2011).
10. Coutal, N. G. et al. L1 retrotransposition in human progenitor cells. Nature 460, 1127–1131 (2009).
11. Rehen, S. K. et al. Constitutional aneuploidy in the normal human brain. J. Neurosci. 25, 2176–2180 (2005).
12. Yousoufian, H. & Pyeert, R. E. Mechanisms and consequences of somatic mosaicism in humans. Nature Rev. Genet. 3, 748–758 (2002).
13. Piotrowski, A. et al. Somatic mosaicism for copy number variation in differentiated human tissues. Hum. Mutat. 29, 1118–1204 (2008).
14. Mikrтыchen, H. et al. Early embryonic chromosome instability results in stable mosaicism in human tissues. Genome Res. 21, 974–984 (2011).
15. Poduri, A. et al. Somatic activation of AKT3 causes hemispheric developmental brain malformations. Neuron 74, 41–48 (2012).
16. Abzyov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing data. Genome Res. 21, 974–984 (2011).
17. Mills, R. E. et al. Mapping copy number variation by population-scale genome sequencing. Nature 470, 59–65 (2011).
18. Cheng, L. et al. Low incidence of DNA sequence variation in human induced pluripotent stem cells generated by nonintegrating plasmid expression. Cell Stem Cell 10, 337–344 (2012).
19. Artl, M. F., Ozdemir, A. C., Birkeland, S. R., Wilson, T. E. & Glover, T. W. Hydroxyurea induces de novo copy number variants in human cells. Proc. Natl Acad. Sci. USA 108, 17360–17365 (2011).
20. Eicher, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Rev. Genet. 11, 446–450 (2010).

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

Acknowledgements We acknowledge support from the National Institutes of Health (NIH) and from the AL Williams Professorship fund and the Harris Professorship fund. We also acknowledge the Yale University Biomedical High Performance Computing Center and its support team (in particular, R. Bjornson and N. Carriero). We thank A. Klim for help with family recruitment. We thank M. V. Simontini for technical help, I.-H. Park for advice in the characterization of iPSC lines and the gift of the iPSC PGP1-1, and S. A. Duncan for the gift of the RH35-iPSC line. We acknowledge the following grant support: NIMH MH089176 and MH087879, the Simons Foundation (SFARI 137055 F.V.) and the State of Connecticut, which funded the hiPSC generation and characterization; and NIH grant RR19895, which funded the instrumentation. We acknowledge the Yale Center for Clinical Investigation for clinical support in obtaining the biopsy specimens. We thank J. Overton for advice in carrying out DNA and RNA sequencing. Finally, we thank M. O’Hallachain and J. Li-Pook-Than for their advice on planning, carrying out and analysing the ddPCR experiments.

Author Contributions The authors contributed to this study at different levels, as described in the following. Study conception and design: F.M.V., A.A. and A.E.U. Family selection: E.L.G. Skin biopsy: A.S. Fibroblast culture: A.H. hiPSC generation and characterization: L.A.R.B., J.M. and L.T. Virus production: A.K. Microarrays data analysis: L.T. Neuronal differentiation: L.A.R.B., N.E.C. and L.T. Process validation: Y.Z. and A.A. aCGH validation and analysis: M.S.H. ddPCR experiments and analysis: M.S.H. and A.A. Human subjects: K.C. Coordination of analyses: F.M.V., S.W., A.E.U. and M.G. Display item preparation: A.A., M.V., L.T., D.P., J.M., N.E.C., Y.Z. and A.A. Manuscript writing: A.A., F.M.V. and A.E.U. The following authors contributed equally to the study: J.M., D.P., Y.Z. and M.S.H. and L.T. All authors participated in discussion of results and manuscript editing.

Author Information The CNV array and sequencing data are available from Gene Expression Omnibus under accessions GSE41716 and GSE41563, and from https://ncbi.nlm.nih.gov/ncbirndatad289. Reprints and permissions information is available at our website: www.nature.com/ reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.M.V. (flora.vaccarino@yale.edu), M.G. (mark.gerstein@yale.edu) or A.E.U. (aeurban@stanford.edu).
METHODS

iPSC generation. A skin biopsy was obtained from the inner area of the upper arm from each member of the two families using standard techniques. Informed consent was obtained from each subject enrolled in the study according to the regulations of the Institutional Review Board and Yale Center for Clinical Investigation at Yale University. Primary cultures of fibroblasts were derived using standard procedures and infected at passage 3 with Yamanaka’s four retroviral vectors, encoding for the canonical reprogramming factors (OCT4 (also known as POU5F1), SOX2, KLF4 and MYC) using a multiplicity of infection (m.o.i.) of 5. After 1 month in culture, colonies with the typical human embryonic stem-cell morphology were picked, expanded on Matrigel substrate in DMEM/F12 containing 1% N2 supplement, 2% B27 supplement, 2 mM l-glutamine, 0.1 mM non-essential amino acids, 1% penicillin/streptomycin, 0.5 mg m⁻¹ BSA fraction V (all from Invitrogen), 0.12 mM monothioglycerol (Sigma), and supplemented with 80 ng ml⁻¹ recombinant human basic fibroblast growth factor (Millipore). Colonies were characterized by immunofluorescence, RT–PCR and gene expression (see below).

RT–PCR. Total RNA was purified from hiPSC clones at passages between 5 and 13 using PicoPure RNA isolation kit (Arcturus). One hundred nanograms of total RNA extracted from hiPSC lines was reverse-transcribed using SuperScript III Reverse Transcriptase and random hexamers. Primers for embryonic stem-cell marker genes are described elsewhere. Primers used for OCT4 (also known as POU5F1), OCT3 and SOX2 specifically detect the transcripts from endogenous genes. β-actin was used as a loading control.

Bisulphite sequencing. Two-hundred nanograms of genomic DNA from fibroblast cells or hiPSCs was bisulphite-converted using the MethylCode Bisulphite conversion kit (Life Technology). Bisulphite-converted DNA was amplified by PCR with the primer sets human OCT4 (ref. 32) and OCT3 (ref. 33) for human hiPSC lines. PCR was performed under the following conditions: 200 μM dNTPs, 400 μM forward or reverse primer, and 2 U of PfTurboCx hotstart DNA polymerase (Agilent Technologies), using the PCR conditions of 95 °C for 30 s, 95 °C for 20 cycles of 95 °C for 20 s, annealing temperature for 20 s, 72 °C for 20 s, and 72 °C for 1 min, followed by extension for 10 min at 72 °C. PCR products were then cloned and 7–8 colonies for each ampiclon were selected for Sanger sequencing.

Neuronal differentiation. Neuronal differentiation was done by slightly modifying a protocol already used in the hiPSC field. Undifferentiated hiPSC colonies maintained on Matrigel were pre-incubated with the ROCK inhibitor (Y-27632), dissociated to single cells and then re-aggregated using V-bottom Aggrewell plates maintained on Matrigel were pre-incubated with the ROCK inhibitor (Y-27632), dissociated to single cells and then re-aggregated using V-bottom Aggrewell plates and cultured in suspension for a further 2 days, and then transferred to a Matrigel substrate in serum-free medium supplemented with noggin (200 ng ml⁻¹). After 2 days, the resulting embryoid bodies were transferred to a Petri dish, cultured in suspension for a further 2 days, and then transferred to a Matrigel substrate in serum-free medium supplemented with noggin (200 ng ml⁻¹), FGF2 (20 ng ml⁻¹) and DKK1 (200 ng ml⁻¹). After 24 h the embryoid bodies generated neuro-epithelial structures known as rosettes. A monolayer of neural progenitor cells was obtained after manual dissection, dissociation and replating of the neural rosettes on polyornithine- and laminin-coated dishes in the presence of FGF2 and EGF (both at 10 ng ml⁻¹) that allowed for the expansion (three or four passages) of the proliferating neuro-epigenetic progenitors.

Microarrays for gene expression analysis. Total RNA isolated as above was analysed by HumanHT-12 v4 BEADCHIPS Illumina microarrays. Values were analysed by GenomeStudio using quantile normalization and background subtraction. Differential scores were compared to values obtained from the federally approved H1 human embryonic stem-cell line.

Library preparations for RNA and DNA sequencing. For RNA-seq libraries, polyadenylated RNA fragments were purified by a Dynabeads mRNA Purification Kit (Invitrogen), fragmented (RNA fragmentation buffer, Ambion), and reverse transcribed into first-strand complementary DNA using random hexamer and superscript II (Invitrogen), followed by second-strand cDNA synthesis using RNaseH and DNA polymerase I (Invitrogen). The cDNA was end-repaired, added single ‘A’ at the 3’ ends before ligation with Illumina adaptors. After running on a gel, DNA fragments from 250 to 350 bp were cut out and extracted using MinElute gel purification kit (Qiagen), and PCR-amplified using Phusion High-Fidelity master mix and Illumina primers with the condition of 98 °C for 30 s, 15 cycles of 98 °C for 10 s, 65 °C for 30 s, and 72 °C for 30 s, and concluding with 72 °C for 5 min.

To make DNA libraries, the Illumina protocol of paired-end DNA sample preparation was followed with minor modification. In short, genomic DNA was sonicated to generate fragments ranging from 200 to 800 bp, which were end-repaired, ‘A’ attached at the end, ligated with Illumina paired-end adaptors, size selected (150–250 bp), and purified and extracted using MinElute gel purification kit (Qiagen), and PCR-amplified using Phusion High-Fidelity master mix and Illumina primers with the condition of 98 °C for 30 s, 15 cycles of 98 °C for 10 s, 65 °C for 30 s, and 72 °C for 30 s, and concluding with 72 °C for 5 min.

Conservative prediction of LM-CNVs in hiPSC. Using Burrows–Wheeler aligner (BWA) 0.5.9-r16 (ref. 35) aligner with options ‘-t 4 -q 15’ we have aligned genomic sequence reads to the human reference genome used by the 1000 Genomes Project (ftp://ftp-trace.ncbi.nih.gov/1000genomes ftp//technical/reference), which is based on hgRc37 and included unplaced contigs. Aligned reads were paired, mapped and sorted by BWA invoked with the following options ‘-a 1000 -n 1 -N 1’. As a result, for each sequenced sample we obtained a file with mapped reads in BAM format. To predict CNVs, the BAM files were processed by the CNVnator method, which is based on read-depth analysis (see ref. 27 for review). For analysis of sequences found at low coverage we used 1,000-bp bins. For analysis of two genomes sequenced at high coverage we used 400-bp bins. Then, in hiPSCs and corresponding fibroblasts, we estimated/genotyped and compared (by CNVnator) the copy number of CNVs predicted in hiPSCs to normal cell, the copy number should be a whole number (for example, 0, 1 or 2); however, the population of cells used for analysis is not heterogeneous, then the copy number can be a non-negative real number (for example, 1.5). We declared copy number variation as a line-manifested deletion candidate in hiPSCs compared to fibroblasts if (1) CNi > 1.5 and CNf > 1.5 and CNi – CNf > 0.5; or (2) CN < 0.5 and CNi > 0.5 and CNi – CNf > 0.5 for X and Y chromosomes in samples collected from males, in which CNi and CNf denote copy number in iPSC and fibroblast samples, respectively. Similarly, we declared copy number variation as a line-manifested duplication candidate if (3) CNi > 2.5 and CNf < 2.5 and CNi – CNf > 0.5; or (4) CNi > 1.5 and CNf > 1.5 and CNi – CNf > 0.5 for X and Y chromosomes in samples collected from males. In other words, we considered copy number variation with an estimated allele frequency (down to 1%) when the copy number variation as a line manifestation when compared to hiPSC lines of at least 25%. We then manually inspected the read-depth signal track to select the most confident LM-CNV candidates for validation. To select confident candidates, we relied on human expertise to evaluate visually the read-depth signal in the candidate regions, presence of discordant paired-end reads supporting a prediction (see below), as well as requiring any pronounced signals in regions of segmental duplications; we also took into account whether CNVs were previously discovered CNVs. To obtain further support for CNVs by paired-end analysis. To obtain further support for a predicted CNV, we searched for abnormally mapped paired-ends in hiPSC lines for which CNVs were predicted and in parental fibroblasts. For a deletion, the supporting paired-ends must map with expected orientation but should have a larger span compared to the expected one from the sequencing library preparation. For a tandem duplication, the supporting paired-ends must map with an orientation different from the expected and also have a larger span (Supplementary Fig. 57). Predicted duplications may be tandem or dispersed. For dispersed duplication we searched for clusters of paired-ends with one end mapping close to predicted duplication boundaries and the other ends clustering somewhere in genome. It is well known that CNVs are enriched for repeats and for sequences around repeat borders at least 25% and a nucleotide sequence that is dispersed outside the boundaries of the predicted duplication. Thus, the absence of paired-end support for a predicted CNV does not invalidate the CNV. We considered a paired-end to support a deletion/duplication if it has a proper (for the type of CNV) pattern of read mapping, and its span and predicted CNV size has at least 80% mutual overlap. This condition and kilobase size of predicted CNV’s guarantees that the span of supportive paired-ends is at least a few kilobases, which is much larger than the span expected from the sequencing library preparation, that is, 300–800 bp. Finally, although we did not require any particular read mapping quality, it was no less than 25 (meaning less than a 0.003 chance of incorrect mapping according to the mapper) for each supportive read. As only around 100 supportive reads were found, we do not expect any single one of them to be mapped incorrectly.

qPCR for LM-CNV call validation. Primer pairs were designed using ProbeFinder software from Roche Applied Science (https://www.roche-applied-science.com/sis/rtpcr/upl/index.jsp). From 2 to 4 kb of DNA near the centre of the presumed CNV was scanned by ProbeFinder and the primer pair design was confirmed by UCSC In-Silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr) and
Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast) for uniqueness and chromosomal location, only a single product and amplicon size.

The control primers to be used in reference target assays yielded a 65-bp amplicon from the RPP30 gene (forward primer: 5′-AGATTTGAGCTACCGAGCG-3′; reverse primer: 5′-GAGCCGGTCTCCCAAGGTT-3′) and a 128-bp amplicon from the ZNF42 genes (forward primer: 5′-ATGATAGCGAGATGTTGCG-3′; reverse primer: 5′-GATGTTGCTCTGAGCTTCAA-3′). These genes are known to be present as single copies in the haploid human genome and they have been previously used as reference in genomic qPCR assays49. Real-time quantitative PCR was run using the Applied Biosystems StepOne Real-Time PCR System (ABI), with SYBR Green chemistry. The experimental data were processed with the StepOne Software v2.1. The comparative Ct method was used to analyse the data for the CNVs in fibroblasts and iPSCs.

All reactions for each primer set were run in triplicate and prepared from the same master mix containing 1× Power SYBR Green PCR Master Mix, 300 nM CNV forward primer, 300 nM CNV reverse primer and 10 ng genomic DNA. The thermal cycling conditions consisted of a pre-run at 95 °C for 10 min and 40 cycles with a 95 °C denaturation step for 15 s, followed by a 60 °C annealing/extension step for 60 s. The fibroblast calibrator was amplified in each run in parallel with the iPSC samples for each CNV. A no-template negative control run in duplicate was also included for each test.

RNA-seq analyses and correlation with genomic CNVs. TopHat2 was used to align the data against the human genome (hg19)27 and dynamically constructed exons and splice libraries. The TopHat output in BAM format was converted to SAM format using SAMTools43 and then, using RSeqTools44, to a standardized compact data format, mapped read format. For each of the GENCODE45 genes, RSeqTools was used to compute the normalized abundance levels of transcripts measured in reads per kilobase per million mapped reads.

For each triad of hiPSCs derived from the same person, we have selected genes intersecting LM-CNVs in at least one hiPSC in the triad and having different (conservatively, more than 5 standard deviations away) from zero expression in at least one hiPSC. Then, the expression values for selected genes were compared between hiPSCs in the same triad, with and without LM-CNV.

PCR to detect heterogeneity in fibroblasts. To validate LM-CNV candidates and detect heterogeneity in fibroblasts, specific primers (Supplementary Table 3) were designed to target both sides of the region adjacent to the deleted region, or the 5′ and 3′ ends of the duplicated region. In this way, specific products were amplified only when deletions or duplications were present. Genomic DNA from the HapMap cell line GM12878 was used as negative control. PCR was conducted with 10 ng of iPSC gDNA, 500 ng (that is, excess) of fibroblast gDNA, 500 ng of gDNA from negative control, 200 μM dNTPs, 200 nM of forward and reverse primers, 2 μM Mg2+, and 4 μl of Taq polymerase (Invitrogen), using thermal cycling conditions consisting of 95 °C for 2 min, 35 cycles of 95 °C for 30 s, 56 °C for 30 s, and 72 °C for 30 s, and a final extension of 72 °C for 5 min. For one event, a second round of PCR with 30 cycles was performed to increase the signals further.

CNV array. For each triad of hiPSCs derived from the same person, we have selected genes intersecting LM-CNVs in at least one hiPSC in the triad and having different (conservatively, more than 5 standard deviations away) from zero expression in at least one hiPSC. Then, the expression values for selected genes were compared between hiPSCs in the same triad, with and without LM-CNV.

PCR to estimate the sensitivity of the approach we performed a negative control experiment by applying primers for a LM-CNV confirmed in family S1123 to a sample from family 03, which does not have this specific LM-CNV. For 6,146 counts of reference allele in three replicas we observed only one spurious count of a LM-CNV allele. For all primers that we designed and used following the manufacturer’s instruction allele ratios in hiPSC did not exceed by 16% from the expected 1:2 (one diploid chromosomes) or 1:1 (on haploid chromosomes) ratios. We thus estimate a correction factor b of less than 1.16, giving us an estimation of background noise of (2 × 1.16)1.16 × 116 = 0.038%. Therefore, an estimation of cell frequency of 0.1% is at least 1.63 standard deviations away (assuming a Poisson noise of count numbers) from background noise.

Array CGH. Each sample was hybridized on a NimbleGen 42.4 M whole-genome CNV array47 under standard conditions as recommended by the manufacturer. Female-pooled DNA from Promega was used as the reference genome in each hybridization of the DNA samples derived from proband S1123. For the DNA samples derived from proband 03-03, each iPSC DNA sample was hybridized against the corresponding fibroblast DNA sample, onto the same array. After hybridization, each array was scanned on a NimbleGen MS200 Microarray scanner and the resulting images were pre-processed using NimbleScan 2.6 software. Data from the arrays were analysed further and visualized using Nexus Copy Number version 6.

Array analysis was performed in Nexus Copy Number 6 by implementing the Fast Adaptive States Segmentation Technique (FASST2) using raw probe intensity data generated by NimbleScan 2.6. This segmentation algorithm relates log, ratios of adjacent probes across the genome to estimate CNV events. The minimum number of probes per segment was set to 3, as this is standard for this segmentation algorithm. Thresholds for calling a copy number gain were set at a log2 value of 0.5 for a copy number loss (which roughly matches the criteria of expected 1:2 (one diploid chromosomes) or 1:1 (on haploid chromosomes) ratios. We thus estimate a correction factor b of less than 1.16, giving us an estimation of background noise of (2 × 1.16)1.16 × 116 = 0.038%. Therefore, an estimation of cell frequency of 0.1% is at least 1.63 standard deviations away (assuming a Poisson noise of count numbers) from background noise.

References
31. Chan, E. M. et al. Live cell imaging distinguishes bona fide human iPSC cells from partially reprogrammed cells. Nature Biotechnol. 27, 1033–1037 (2009).
32. Deb-Rinker, P., Ly, D., Jeziierski, A., Sikorska, M. & Walker, P. R. Serial DNA methylation of the Nanog and Oct-4 upstream regions in human NT2 cells during neuronal differentiation. J. Biol. Chem. 280, 6257–6260 (2005).
33. Freberg, C. T., Dahl, J. A., Timoskainen, S. & Collas, P. Epigenetic reprogramming of OCT4 and NANOG regulatory regions by embryonal carcinoma cell extract. Mol. Cell. Biol. 18, 1543–1553 (2007).
34. Kim, J. E. et al. Investigating synapse formation and function using human pluripotent stem cell-derived neurons. Proc. Natl Acad. Sci. USA 108, 3005–3010 (2011).
35. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595 (2010).
36. Wang, L. Y., Abyzov, A., Korbel, J. O., Snyder, M. & Gerstein, M. MSB: a mean-shift-based approach for the analysis of structural variation in the genome. Genome Res. 19, 106–117 (2009).
37. Zhang, J., Feuk, L., Duggan, G. E., Khaja, R. & Scherer, S. W. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet. Genome Res. 115, 205–214 (2006).
38. Korbel, J. O. et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. 10, R23 (2009).
39. Lam, H. Y. et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nature Biotechnol. 28, 47–55 (2010).
40. Sanders, S. J. et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70, 863–885 (2011).
41. Qin, J., Jones, R. C. & Ramakrishnan, R. Studying copy number variations using a nanofluidic platform. Nucleic Acids Res. 36, e116 (2008).
42. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111 (2009).
43. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
44. Habegger, L. et al. RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics 27, 281–283 (2011).
45. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
46. Abyzov, A. & Gerstein, M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. Bioinformatics 27, 595–603 (2011).
47. Hindson, B. J. et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal. Chem. 83, 8604–8610 (2011).
48. Haraksingh, R. R., Abyzov, A., Gerstein, M., Urban, A. E. & Snyder, M. Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms. PLoS ONE 6, e27859 (2011).