HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2013 June 20.

Published in final edited form as:

Nature. 2012 December 20; 492(7429): 438–442. doi:10.1038/nature11629.

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

Alexej Abyzov[1,2,3], **Jessica Mariani**[1,4,*], **Dean Palejev**[1,4,*], **Ying Zhang**[1,6,*], **Michael Seamus Haney**[12,13,*], **Livia Tomasini**[1,4,*], **Anthony Ferrandino**[1,4,5], **Lior A. Rosenberg Belmaker**[1,4], **Anna Szekely**[1,6,7], **Michael Wilson**[1,2,3], **Arif Kocabas**[1,4], **Nathaniel E. Calixto**[1,4], **Elena L. Grigorenko**[1,4,8,9], **Anita Huttner**[1,11], **Katarzyna Chawarska**[1,4], **Sherman Weissman**[1,6], **Alexander Eckehart Urban**[1,12,13,#], **Mark Gerstein**[1,2,3,10,#], and **Flora M. Vaccarino**[1,4,5,#]

[1]Program in Neurodevelopment and Regeneration, Yale University, New Haven CT 06520

[2]Program in Computation Biology and Bioinformatics, Yale University, New Haven CT 06520

[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven CT 06520

[4]Child Study Center, Yale University, New Haven CT 06520

[5]Department of Neurobiology, Yale University, New Haven CT 06520

[6]Department of Genetics, Yale University, New Haven CT 06520

[7]Department of Neurology, Yale University, New Haven CT 06520

[8]Department of Psychology, Yale University, New Haven CT 06520

[9]Department of Epidemiology and Public Health, Yale University, New Haven CT 06520

[10]Department of Computer Science, Yale University, New Haven CT 06520

[11]Department of Pathology, Yale University, New Haven CT 06520

[12]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, USA

[13]Department of Genetics, School of Medicine, Stanford University, Stanford, USA

Abstract

Correspondence and requests for materials should be addressed to: flora.vaccarino@yale.edu; mark.gerstein@yale.edu; aeurban@stanford.edu. #corresponding author .
*these authors contributed equally to this work

The authors declare no competing financial interests

Author contribution The authors contributed 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 L.T., and Y.Z. Processing and analysis of RNAseq data: D.P. and A.A. Processing and analysis of DNAseq data: A.A. and M.W. qPCR validation: A.F. PCR validation: Y.Z. and A.A. aCGH hybridization 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., F.M.V., L.T., D.P., J.M., N.E.C., Y.Z. and M.S.H. Writing manuscript: A.A., F.M.V., and A.E.U. The following authors contributed equally to the study: J.M, D.P., Y.Z., M.S.H., L.T. All authors participated in discussion of results and manuscript editing.

Reprogramming human somatic cells into induced pluripotent stem cells (iPSCs) has been suspected of causing *de novo* copy number variations (CNVs)[1-4]. To explore this issue, we performed a whole-genome and transcriptome analysis of 20 human iPSC lines derived from primary skin fibroblasts of 7 individuals using next-generation sequencing. We find that, on average, an iPSC line manifests two CNVs not apparent in the fibroblasts from which the iPSC was derived. Using qPCR, PCR, and digital droplet PCR (ddPCR), we show that at least 50% of those CNVs are present as low frequency somatic genomic variants in parental fibroblasts (i.e. the fibroblasts from which each corresponding hiPSC line is derived) and are manifested in iPSC colonies due to the colonies' clonal origin. Hence, reprogramming does not necessarily lead to *de novo* CNVs in iPSC, since most of 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 of deriving iPSCs from somatic cells[5-8] has opened exciting new possibilities for the study of human development, human genetic variation and regenerative medicine[9-13]. 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. However, there are reports of genomic instability in stem and precursor cells, indicating that copy number variations/structural variations (CNVs/SVs) might arise in iPSCs, in addition to single base-pair changes[1-4,14-17]. 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 elements[18-21]. Such a phenomenon could have far-reaching physiological consequences yet is still poorly understood and very difficult to study[22-25]. The derivation of iPSCs offers the opportunity to analyze a single cell's genome 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 (RT-PCR, microarrays, complete transcriptome by RNAseq) and 4) demethylation of canonical pluripotency factor promoters (Supplementary Figs 2-3 and Supplementary Tables 1-2). This thorough evaluation (Supplementary information) revealed extensive similarity of our hiPSCs to hESCs and divergence of hiPSC 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 (PE) sequencing data on the ILLUMINA HiSeq platform for 20 hiPSC lines and predicted CNVs in hiPSC lines with CNVnator[26] (Supplementary Fig. 1B). CNVnator uses read depth (RD) analysis and was shown to have the highest sensitivity in confirming CNVs previously discovered with arrays and fosmid sequencing[27]. 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 parental fibroblasts (i.e. the fibroblast line of origin for each respective clonal hiPSC line) to identify the variants manifested only in hiPSCs, i.e. line-manifested CNVs (LM-CNV). We were able to discover CNVs as small as 2 kbp, but the highest sensitivity was for CNVs of at least 5 kbp in size (Supplementary Fig. 5). Using conservative criteria, we predicted a total of 74 LM-CNVs in all 20 lines (Supplementary Table 3), i.e. just a few LM-CNVs per line. Similar numbers of LM-CNVs per line were observed for few additional 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 initially discovered in hiPSC lines and performing RD analysis at higher coverage (~20X) 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 of an individual from the other family and observed roughly forty different CNVs per comparison (i.e., significantly more than LM-CNVs per hiPSC line, Fig. 1C), which is consistent with interindividual variations in similar size range as described previously[27].

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

To obtain an independent confirmation of our approach for LM-CNV detection we analyzed 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 hiPSC from these individuals, were also confirmed by aCGH (Supplementary Figs 45-54). However, no additional LM-CNV could be discovered using aCGH data, since the estimated FDR of the set of additional predictions was close to 100%, based on qPCR validation of a random subset (Supplementary Tables 4-5). 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, i.e., passage 17-52, in five hiPSC lines. We observed a strong correlation (Pearson's coefficient 0.96) between qPCR results

obtained in late versus early passage (Supplementary Fig. 6). Among sixteen LM-CNVs that were tested, 87.5% were validated in late passage (Table 1), suggesting long-term stability of the hiPSC genome.

We then analyzed the origin of LM-CNVs, i.e., whether they had arisen *de novo* in the hiPSC as a sequel to reprogramming or they 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 same individual's fibroblast culture (Table 1; Fig. 2a; 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 0.5, indicative of one haplotype deletion, using both RD analysis and their qPCR validation (Supplementary Fig. 6, Supplementary information).

To test for actual presence of somatic CNVs in the fibroblast cultures, we performed PCR amplification with diagnostic primers across CNV breakpoints in hiPSC and the corresponding donor fibroblasts for 20 LM-CNVs with good initial estimate of their breakpoints from PE analyses (Fig. 2b, Table 1, Supplementary Table 3). We observed expected bands in all cases when using hiPSC DNA and in 8 cases when using DNA from the corresponding fibroblast cultures (Table 1; see Fig. 2b,e,g for representative examples and Supplementary Figs 7-39). For 15 LM-CNVs we additionally 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 ratio between the target and the control regions. The frequency of the duplication in chromosome X in fibroblast cells was estimated to be 12.6% (Fig. 2d). Cell frequencies varied from 14.6% (Fig. 2f) to less than 1% (Fig. 2h) and are summarized in Table 1. In total, using PCR and ddPCR allowed us to establish the presence in the parental fibroblast culture of 10 out of 20 LM-CNVs, suggesting that fibroblast somatic genomic heterogeneity can explain at least 50% of the LM-CNVs in hiPSC (Supplementary Table 6).

Sanger capillary sequencing of PCR bands allowed us to determine breakpoints with base pair resolution for 18 non-redundant LM-CNVs (Supplementary alignment file). Analysis of sequences around breakpoints suggests non-homologous end joining (NHEJ) as a key mechanism in the creation of LM-CNVs. Finally, we examined whether LM-CNVs affect the expression of intersected genes. Statistical analysis, using Fischer's exact test, showed that with a p-value of 0.01 there was a direct association of gene expression with its copy number, i.e., duplications increased expression while deletions decreased it (Supplementary Fig. 56).

In summary, we report genomic stability of hiPSC lines and the presence of extensive somatic mosaicism for copy-number variation in the genome of human skin fibroblasts. This is the result of a systematic discovery and analysis of CNVs in 20 hiPSC lines relative to seven fibroblast cultures from which the hiPSC lines were derived. As hiPSCs are clonally

derived from a few or just one fibroblast cell, analysis of their 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/ddPCR across breakpoints to genotype CNVs in the parental fibroblasts and estimated that 50% of CNVs manifested in hiPSCs could be traced back to the original fibroblast population. We may be underestimating this phenomenon because very low allele frequency somatic CNVs might still escape confirmation by PCR/ddPCR in fibroblasts due 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 hiPSC manifest on average two validated CNVs larger than 10 kbp, which is considerably more than in two previous studies[1,28]. The difference is likely attributable to us using sequencing (generally a more sensitive approach, see Supplementary Discussion) as opposed to using SNP arrays[1]. Whereas Cheng et al.[28] also used sequencing, they analyzed only three hiPSC lines, thus, 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 by Cheng et al. from the former manifest fewer LM-CNVs than do our hiPSC derived from the latter.

It was previously hypothesized that CNVs might arise in hiPSC 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 LM-CNVs preexisted in parental fibroblast cells. We also found no significant difference in the number of LM-CNVs in relation to passage number. Thus, our analysis support neither the hypothesis[3] that hiPSC generally have a large rate of *de novo* mutations nor the observation that most LM-CNVs in hiPSC disappear in late passages[3]. Using different parental cells and applying different protocols for cell culturing could be the factors accounting for the difference in the results.

In 6 hiPSC we determined that at least one LM-CNV originated in parental fibroblast cells. Assuming that each hiPSC colony represents a single, clonally expanded cell, we estimate that 30% (=6/20) 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 percent, suggesting wide variability in the extent of fibroblast mosaicism. Although it is possible that some CNVs could have arisen during the fibroblast cell culture[29], we think this is unlikely given that they were passaged less than 5 times before proceeding with hiPSC generation.

It has been known for a while[22] that somatic variants can be responsible for various diseases, including cancer, and we have just provided evidence that the extent of somatic variation could have been drastically underestimated. If true, this needs to be taken into account when designing an hiPSC-based study. But more importantly, this finding may challenge widely adopted experimental designs for genetic analyses of diseases with complex inheritance where only the genomes of lymphoblastoid cells are being analyzed. 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 contribution in some of the complex and especially in neurodevelopmental diseases, for which determining the exact loci for genetic predisposition has proven difficult[30].

Methods

Induced pluripotent stem cells (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 IRB and YCCI of 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, SOX2, KLF4 and c-MYC) using an MOI of 5. After one month in culture, colonies with the typical hESC 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/mL BSA Fraction V (all from Invitrogen), 0.12 mM monothioglycerol (Sigma, M-6145), and supplemented with 80 ng/ml 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 were reverse-transcribed using SuperScript III Reverse Transcriptase and random hexamers. Primers for ES cell marker genes are described elsewhere[31]. Primers used for Oct4, c-Myc and Sox2 specifically detect the transcripts from endogenous genes. β-actin was used as a loading control.

Bisulfite sequencing

200 ng of genomic DNA from fibroblast cells or hiPSCs was bisulfite converted using the MethylCode Bisulfite converstion kit (Life Technology, CA). Bisulfite converted DNA was amplified by PCR with the primer sets 7 for Oct4 [32] and sets 3 [33] for Nanog. PCR was performed with the following components: 200 μM dNTPs, 200 nM forward or reverse primer, and 2 Units of PfuTurboCx hotstart DNA polymerase (Agilent Technologies, CA), using the PCR conditions of 95° C for 5 minutes, 35 cycles of 95° C for 30 seconds, 58/55° C for 1 min 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 amplicon were selected for Sanger sequencing.

Neuronal Differentiation

Neuronal differentiation was done by slightly modifying a protocol already used in the hiPSC field[13,34]. 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 in serum-free medium containing recombinant

Noggin (200ng/mL). After two days, the resulting embryoid bodies (EBs) were transferred to a Petri dish, cultured in suspension for an additional two days, and then transferred to a Matrigel substrate in serum-free medium supplemented with Noggin (200ng/mL), FGF2 (20ng/mL) and Dkk1 (200ng/mL). After 24 hours the EBs generated neuro-epithelial structures known as rosettes. A monolayer of neural progenitor cells (NPCs) was obtained after manual dissection, dissociation and replating of the neural rosettes on poly-ornithine and laminin coated dishes in the presence of FGF2 and EGF (both at 10ng/mL) that allowed for the expansion (3 or 4 passages) of the proliferating neural progenitors.

Microarrays for gene expression analysis

Total RNA isolated as above was analyzed by HumanHT-12 v4 BEADCHIP Illumina microarrays. Values were analyzed by GenomeStudio using quantile normalization and background subtraction. Differential scores were compared to values obtained from the federally approved H1 human embryonic stem cell (hESC) line.

Library preparations for Paired-End (PE) RNA and DNA sequencing

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

To make DNA libraries, the Illumina protocol of PE DNA sample preparation was followed with minor modification. In short, gDNA was sonicated to generate fragments ranging from 200bp to 800 bp, which were end repaired, “A” attached at the end, ligated with Illumina PE adaptors, size selected (450bp – 550 bp) on 2% E-gel (Invitrogen, CA) and extracted from the gel. The final PCR step is the same as in RNA-seq library preparation but with 18 cycles.

Conservative prediction of line-manifested CNVs in hiPSC

Using BWA 0.5.9-r16[35] aligner with options ’-t 4 -q 15′ we have aligned genomic sequence reads to the human reference genome used by the 1000 Genome Project (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference), which is based on hGRC37 and included few extra 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. In order to predict CNVs, the bam files were processed by the CNVnator method[26,36] which is based on read depth analysis (see Mills et al.[27] for review). For analysis of genomes sequenced at low coverage we used 1000 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) copy-number (CN) of CNVs predicted in hiPSCs. In a normal cell, CN should

be a whole number (e.g., 0, 1, 2, etc.), however, if the population of cells used for analysis is not heterogeneous, then the CN can be a non-negative real number (e.g., 1.5). We declared CNV as a line manifested deletion candidate in hiPSCs compared to fibroblasts if i) $CN^i < 1.5$ & $CN^f > 1.5$ & $CN^f - CN^i > 0.5$; or ii) $CN^i < 0.5$ & $CN^f > 0.5$ & $CN^f - CN^i > 0.5$ for X and Y chromosomes in samples collected from males. Here, CN^i and CN^f stand for CN in iPSCs and fibroblast samples respectively. Similarly, we declared CNV as a line manifested duplication candidate if iii) $CN^i > 2.5$ & $CN^f < 2.5$ & $CN^i - CN^f > 0.5$; or iv) $CN^i > 1.5$ & $CN^f < 1.5$ & $CN^i - CN^f > 0.5$ for X and Y chromosomes in samples collected from males. In other words, we considered CNV with an estimated allele frequency in fibroblasts of at least 25% and difference in allele frequency when compared to hiPSC line of at least 25%. We then manually inspected the RD signal track to select the most confident line-manifested CNV (LM-CNV) candidates for validation. To select confident candidates, we relied on human expertise to visually evaluate the RD signal in the candidate regions, presence of discordant paired-end reads supporting a prediction (see below), as well as requiring very pronounced signals in regions of segmental duplications; we also took into account whether CNVs were previously discovered CNVs[27,37]. Two CNV boundaries were re-estimated. Selected confident LM-CNV candidates have been validated experimentally by qPCR, aCGH, PCR and ddPCR.

Sensitive prediction of line-manifested CNVs in hiPSC

To perform a more sensitive CNV calling with CNVnator, we used option '-relax', which allowed us to find CNVs with allele frequencies down to 12.5% as opposed to 25% with the default options. Of note, the heterozygous deletion/duplications on a diploid chromosome have a 50% allele frequency. Additionally, we relaxed the criteria on declaring a CNV as a LM-CNV. Specifically, we used the following criteria i) $CN^i < 1.7$ & $CN^f > 1.5$ & $CN^f - CN^i > 0.3$; and ii) $CN^i < 0.7$ & $CN^f > 0.5$ & $CN^f - CN^i > 0.3$ to call for line-manifested deletions on diploid and haploid chromosomes respectively. Similarly, we used iii) $CN^i > 2.3$ & $CN^f < 2.5$ & $CN^i - CN^f > 0.3$; and iv) $CN^i > 1.3$ & $CN^f < 1.5$ & $CN^i - CN^f > 0.3$ to call for line-manifested duplications on diploid and haploid chromosomes respectively. In other words, we considered CNVs with an estimated allele frequency in fibroblasts (down to 15%) and a difference in allele frequency (down to 15%) when compared to hiPSC lines.

Obtaining additional support for CNVs by paired-end analysis

To obtain additional support for a predicted CNV, we searched for abnormally mapped paired-ends (PE) in hiPSC lines for which CNVs were predicted and in parental fibroblasts[38]. For a deletion, the supporting PEs 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 PE smust 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 PEs with one end mapping close to predicted duplication boundaries and other ends clustering somewhere in genome. It is well known, see Lam et al.[39], that CNVs are enriched for repeats and homologous sequences around breakpoints, where read mapping is ambiguous. Thus, the absence of PE support for a predicted CNV does not invalidate the CNV. We considered a PE 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 CNVs guarantees that the span of supportive PEs is at least a few kbp, which is much larger than the span expected from the sequencing library preparation, i.e., 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). 2-4 kbp of DNA sequence near the center 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 65bp amplicon from the RPP30 gene (forward primer: AGATTTGGACCTGCGAGCG; reverse primer: GAGCGGCTGTCTCCACAAGT) and a 128bp amplicon from the ZNF423 gene (forward primer: AGATGATCGGAGATGGTTGTG; reverse primer: GATCTGCTCGTGCCTCTTCAA). These genes are known to be present as single copies in the haploid human genome[40,41]. 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 analyze the data for the CNVs in fibroblasts and iPSCs.

All reactions for each primer set were run in triplicate and were prepared from the same master mix containing 1× Power SYBR Green PCR Master Mix, 300nM CNV forward primer, 300nM CNV reverse primer and 10ng 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 seconds followed by a 60°C annealing/extension step for 60 seconds. 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 CNV assay.

RNA-Seq analyses and correlation with genomic CNVs

Tophat[42] was used to align the data against the human genome (hGRC37) and dynamically constructed exons and splice libraries. The Tophat output in BAM format was converted to SAM format using SAMtools[43] and then, using RSEQtools[44], to a standardized compact data format, Mapped Read Format (MRF). For each of the GENCODE[45] genes, RSEQtools was used to compute the normalized abundance levels of transcripts measured in RPKM, Reads Per Kb 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 hiPSC in the same triad, with and without LM-CNV.

PCR to detect heterogeneity in fibroblasts

To validate LM-CNVs candidates and to detect heterogeneity in fibroblasts, specific primers (Supplementary Table 3) were designed to target both sides of region adjacent to the deleted or the 5′ and 3′ end 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 (i.e., excess) of fibroblast gDNA, 500ng of gDNA from negative control, 200uM dNTPs, 200 nM of forward and reverse primers, 1.5 mM Mg^{2+}, and 4 units of Taq polymerase (Invitrogen, CA), using thermal cycling conditions consisting of 95 °C for 2 minutes, 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 minutes. For one event, a second round of PCR with 30 cycles was performed to further increase the signals. For CNVs with substantial yield of PCR product in the first run, an additional PCR run with 30 cycles was performed with the same conditions except reduced amounts of starting fibroblast gDNA to 10 ng (i.e., equal to the amount of gDNA from hiPSC). All specifically amplified PCR bands were run on a 2% E-gel (Invitrogen, CA), the gel was extracted by MinElute gel purification Kit (Qiagen, MD), and the extracted DNA was sequenced using both forward and reverse primers. The resulting bands were aligned to the reference genome using AGE[46] to derive the exact CNV breakpoints.

Digital PCR to estimate LM-CNV cell frequency in fibroblasts

Digital Droplet PCR (ddPCR)[47] was carried out using the Bio-Rad QX100 platform Quantalife system (Bio-Rad Laboratories Inc., Hercules CA). Following the manufacturer′s instructions, 20ul PCR reaction mixtures consisting of ddPCR mastermix and TaqMan reagents were partitioned into 15,000 to 20,000 water-in-oil droplets. Each chemically homogenous droplet supports PCR amplification in a thermal cycler. TaqMan reagents enable fluorescent labeling of amplified reference and target regions. PCR products are then inserted into an automated droplet flow cytometer, where single-file, simultaneous two-color detection of the droplets is measured. Given that the PCR mixture is randomly partitioned into 15,000 to 20,000 reactions vesicles, Poisson statistics can be applied to this process to yield target nucleic acid quantification of the sample.

In this instance, VIC fluorescent probes hybridizing to an amplicon targeting the RPP30 gene served as a reference region of which two copies should be present in each cell (probes and primers provided by BioRad). LM-CNV specific FAM probes were synthesized such that they would hybridize to amplicons targeting a given LM-CNV. Primers were designed to target LM-CNVs such that the amplicon would contain the breakpoint sequence and the FAM probe was designed to hybridize directly onto this breakpoint sequence, whenever possible (LM-CNV specific primers and probes from IDT, San Diego, CA). In the absence of the targeted LM-CNV in a given droplet, no PCR reaction would take place. Copy-numbers of target regions were then calculated in reference to the RPP30 event counts.

ddPCR measures allele counts of reference region and target CNV. Let *M* be the measurement (i.e. counts) of reference region and M_{CNV} be the measurement (i.e. counts) of the target CNV allele in hiPSC. Then assuming homogeneous population of cells in hiPSC, we expect that the estimated allele frequency of a target heterozygous CNV to be ~50% for LM-CNVs on diploid chromosomes (one haplotype has no LM-CNV) and ~100% for LM-CNVs on haploid chromosomes. That is

$$M_{CNV}/M=0.5 \quad \text{for diploid chromosomes}$$
$$2^*M_{CNV}/M=1.0 \quad \text{for haploid chromosomes}$$

(here we need to multiple by 2 to account for haploid chromosome, as the reference region is on diploid chromosome). Indeed we observed that measured values are very close to the mentioned expected ones, validating our assumptions that hiPSC cells are homogeneous and LM-CNVs are heterozygous.

Due to experimental variability (e.g. primer efficiency), those two ratios are slightly different from 0.5 or 1.0. Introducing as an experimental bias *b* accounting for the difference, then in hiPSC

$$M_{CNV}/M^*b=0.5 \quad \text{for diploid chromosomes}$$
$$2^*M_{CNV}/M^*b=1.0 \quad \text{for haploid chromosomes}$$

Giving us $b=0.5*M/M_{CNV}$ for either diploid of haploid chromosome.

Using the same logic we can now derive an estimation of LM-CNV allele frequency in the fibroblasts. Let *F* be the measurement (i.e. counts) of the reference allele and F_{CNV} be the measurement (i.e. counts) of the target CNV allele in fibroblasts. Allele frequency can be estimated as follows

$$\text{Allele frequency}=F_{CNV}/F^*b \quad \text{for diploid chromosomes}$$
$$\text{Allele frequency}=2^*F_{CNV}/F^*b \quad \text{for haploid chromosomes}$$

b is estimated from analysis of data for hiPSC and typically is close to 1. The *Cell CNV frequency*, i.e., number of cell carrying the CNV, can be estimated as

$$\text{Cell CNV frequency}=2^*F_{CNV}/F^*b \quad \text{for} \quad LM-CNV \quad \text{on either}$$

haploid or diploid chromosome.

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 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 expected 1:2 (one diploid chromosomes) or 1:1 (on haploid

chromosomes). We thus estimate a correction factor *b* of less than 1.16, giving us an estimation of background noise of 2*1/6146*1.16=0.038%. Therefore, an estimation of allele frequency of 0.1% is at least 1.63 standard deviations away (assuming a Poisson nature of noise counts) from background noise.

Array CGH

Each sample was hybridized on a NimbleGen 4.2M whole-genome CNV array[48] 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 S1123-02. 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. Following 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 analyzed 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 $\log_2$ 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 CN gain were set at a $\log_2$ value of 0.37 and −0.5 for a CN loss (which roughly matches the criteria of conservative calling using sequencing). Log_2 thresholds for high gains (1 or more copies) and high loss were set at 1.0 and −1.1 respectively.

Calls for proband 03-03 are candidate LM-CNV by definition, as we hybridized hiPSC DNA against fibroblast DNA. For proband S1123-02, we selected LM-CNV candidate as calls in hiPSC that do not overlap with any call in the corresponding fibroblasts. For this person, we further filtered out calls that are likely to be noise, i.e. calls smaller than 6 kbp and in centromeres and telomeres.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We acknowledge support from the NIH and from the AL Williams Professorship fund and the Harris Professorship fund. We also acknowledge the Yale University Biomedical High Performance Computing Center; its support team (in particular, Robert Bjornson and Nicholas Carriero). We thank Dr. Ami Klin for essential help with family recruitment. We thank Dr. Maria Vittoria Simonini for technical help, Dr. In-Hyun Park for advice in the characterization of iPSC lines and the gift of the iPS PGP1-1 and Dr. Stephen A. Duncan for the gift of the iPS K3 iPSC line. We acknowledge the following grant support: NIMH MH089176 and MH087879, the Simons Foundation 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 Dr. John Overton at the Yale Center for Genome Analysis for advice in carrying out DNA and RNA sequencing. Finally, we thank Ms. Maeve O'Huallachain and Dr. Jennifer Li-Pook-Than of Stanford University for their advice on planning, carrying out and analyzing the ddPCR experiments.

References

1. Laurent LC, et al. Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. Cell Stem Cell. 2011; 8:106–118. [PubMed: 21211785]
2. Quinlan AR, et al. Genome Sequencing of Mouse Induced Pluripotent Stem Cells Reveals Retroelement Stability and Infrequent DNA Rearrangement during Reprogramming. Cell Stem Cell. 2011; 9:366–373. [PubMed: 21982236]
3. Hussein SM, et al. Copy number variation and selection during reprogramming to pluripotency. Nature. 2011; 471:58–62. [PubMed: 21368824]
4. Mayshar Y, et al. Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. Cell Stem Cell. 2010; 7:521–531. [PubMed: 20887957]
5. Takahashi K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007; 131:861–872. [PubMed: 18035408]
6. Yu J, et al. Induced pluripotent stem cell lines derived from human somatic cells. Science. 2007; 318:1917–1920. [PubMed: 18029452]
7. Wernig M, et al. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature. 2007; 448:318–324. [PubMed: 17554336]
8. Lowry WE, et al. Generation of human induced pluripotent stem cells from dermal fibroblasts. Proc Natl Acad Sci U S A. 2008; 105:2883–2888. [PubMed: 18287077]
9. Vaccarino FM, et al. Annual Research Review: The promise of stem cell research for neuropsychiatric disorders. J Child Psychol Psychiatry. 2011; 52:504–516. [PubMed: 21204834]
10. Park IH, et al. Disease-specific induced pluripotent stem cells. Cell. 2008; 134:877–886. [PubMed: 18691744]
11. Lee G, et al. Modelling pathogenesis and treatment of familial dysautonomia using patient-specific iPSCs. Nature. 2009; 461:402–406. [PubMed: 19693009]
12. Hargus G, et al. Differentiated Parkinson patient-derived induced pluripotent stem cells grow in the adult rodent brain and reduce motor asymmetry in Parkinsonian rats. Proc Natl Acad Sci U S A. 2010; 107:15921–15926. [PubMed: 20798034]
13. Brennand KJ, Gage FH. Concise review: the promise of human induced pluripotent stem cell-based studies of schizophrenia. Stem Cells. 2011; 29:1915–1922. [PubMed: 22009633]
14. Liang Q, Conte N, Skarnes WC, Bradley A. Extensive genomic copy number variation in embryonic stem cells. Proc Natl Acad Sci U S A. 2008; 105:17453–17456. [PubMed: 18988746]
15. Wu H, et al. Copy number variant analysis of human embryonic stem cells. Stem Cells. 2008; 26:1484–1489. [PubMed: 18369100]
16. Elliott AM, Elliott KA, Kammesheidt A. High resolution array-CGH characterization of human stem cells using a stem cell focused microarray. Mol Biotechnol. 2010; 46:234–242. [PubMed: 20524159]
17. Howden SE, et al. Genetic correction and analysis of induced pluripotent stem cells from a patient with gyrate atrophy. Proc Natl Acad Sci U S A. 2011; 108:6537–6542. [PubMed: 21464322]
18. De S. Somatic mosaicism in healthy human tissues. Trends in genetics : TIG. 2011; 27:217–223. [PubMed: 21496937]
19. Baillie JK, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011
20. Coufal NG, et al. L1 retrotransposition in human neural progenitor cells. Nature. 2009; 460:1127–1131. [PubMed: 19657334]
21. Rehen SK, et al. Constitutional aneuploidy in the normal human brain. J Neurosci. 2005; 25:2176–2180. [PubMed: 15745943]
22. Youssoufian H, Pyeritz RE. Mechanisms and consequences of somatic mosaicism in humans. Nat Rev Genet. 2002; 3:748–758. [PubMed: 12360233]
23. Piotrowski A, et al. Somatic mosaicism for copy number variation in differentiated human tissues. Human mutation. 2008; 29:1118–1124. [PubMed: 18570184]

24. Mkrtchyan H, et al. Early embryonic chromosome instability results in stable mosaic pattern in human tissues. PLoS One. 2010; 5:e9591. [PubMed: 20231887]
25. Poduri A, et al. Homozygous PLCB1 deletion associated with malignant migrating partial seizures in infancy. Epilepsia. 2012
26. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011; 21:974–984. [PubMed: 21324876]
27. Mills RE, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011; 470:59–65. [PubMed: 21293372]
28. Cheng L, et al. Low incidence of DNA sequence variation in human induced pluripotent stem cells generated by nonintegrating plasmid expression. Cell stem cell. 2012; 10:337–344. [PubMed: 22385660]
29. Arlt MF, Ozdemir AC, Birkeland SR, Wilson TE, Glover TW. Hydroxyurea induces de novo copy number variants in human cells. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:17360–17365. [PubMed: 21987784]
30. Eichler EE, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11:446–450. [PubMed: 20479774]

References

31. Chan EM, et al. Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. Nat Biotechnol. 2009; 27:1033–1037. [PubMed: 19826408]
32. Deb-Rinker P, Ly D, Jezierski A, Sikorska M, Walker PR. Sequential DNA methylation of the Nanog and Oct-4 upstream regions in human NT2 cells during neuronal differentiation. J Biol Chem. 2005; 280:6257–6260. [PubMed: 15615706]
33. Freberg CT, Dahl JA, Timoskainen S, Collas P. Epigenetic reprogramming of OCT4 and NANOG regulatory regions by embryonal carcinoma cell extract. Molecular biology of the cell. 2007; 18:1543–1553. [PubMed: 17314394]
34. Kim JE, et al. Investigating synapse formation and function using human pluripotent stem cell-derived neurons. Proc Natl Acad Sci U S A. 2011; 108:3005–3010. [PubMed: 21278334]
35. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26:589–595. [PubMed: 20080505]
36. Wang LY, Abyzov A, Korbel JO, Snyder M, Gerstein M. MSB: a mean-shift-based approach for the analysis of structural variation in the genome. Genome research. 2009; 19:106–117. [PubMed: 19037015]
37. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res. 2006; 115:205–214. [PubMed: 17124402]
38. Korbel JO, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. 2009; 10:R23. [PubMed: 19236709]
39. Lam HY, et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol. 2010; 28:47–55. [PubMed: 20037582]
40. Sanders SJ, et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron. 2011; 70:863–885. [PubMed: 21658581]
41. Qin J, Jones RC, Ramakrishnan R. Studying copy number variations using a nanofluidic platform. Nucleic Acids Res. 2008; 36:e116. [PubMed: 18710881]
42. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
43. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
44. Habegger L, et al. RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics. 2010; 27:281–283. [PubMed: 21134889]

45. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

46. Abyzov A, Gerstein M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. Bioinformatics. 2011; 27:595–603. [PubMed: 21233167]

47. Hindson BJ, et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem. 2011; 83:8604–8610. [PubMed: 22035192]

48. Haraksingh RR, Abyzov A, Gerstein M, Urban AE, Snyder M. Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms. PLoS One. 2011; 6:e27859. [PubMed: 22140474]

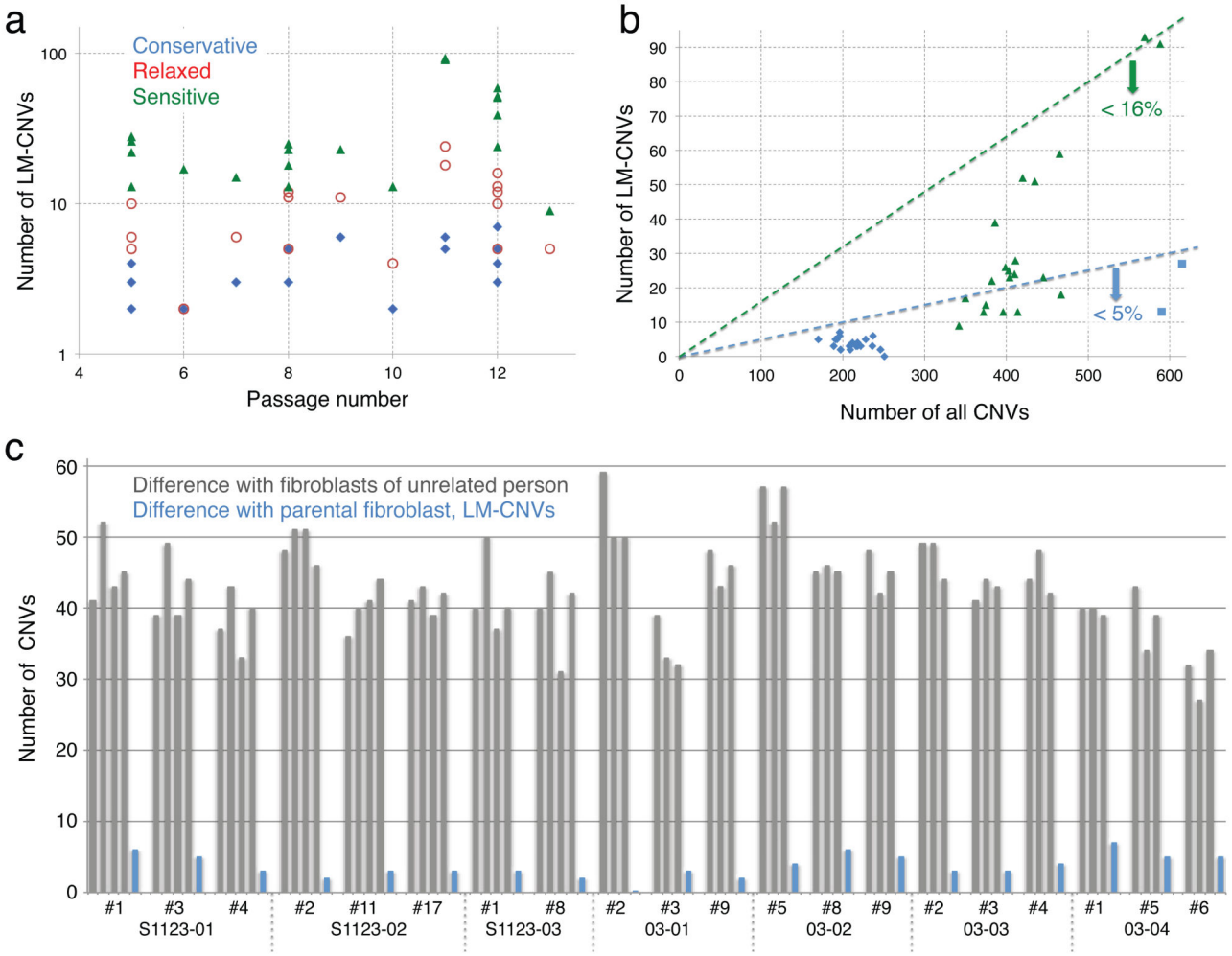

Figure 1. Characterization of candidate line manifested CNVs (LM-CNVs) with respect to passage number and total CNVs

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**, Percentage of LM-CNVs of all CNVs detected in hiPSCs by comparison with the reference human genome; square symbols represent data obtained at increased (20X) coverage. LM-CNVs represent a small fraction of all CNVs in a person. **c**, Counts of LM-CNVs in hiPSC using fibroblasts from different individuals as a baseline. Genomes of hiPSC are different in roughly 40 CNVs (gray bars) when compared to fibroblasts from unrelated persons, that is, individuals from the other family. In contrast, genomes of hiPSC differ by less than 10 CNVs as compared to their fibroblasts of origin (blue bars). LM-CNVs in hiPSC as compared to fibroblasts represent a small increment to the already existing genetic diversity in human population.

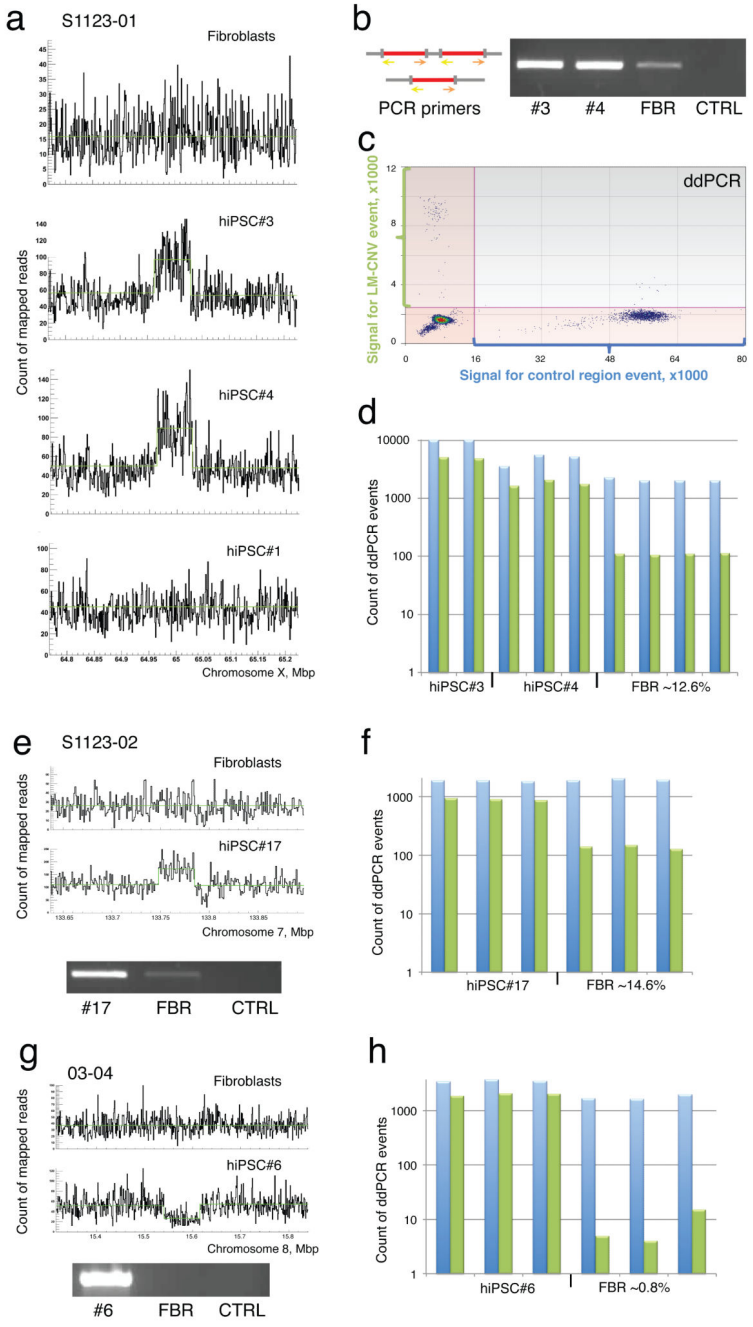

Figure 2. Validation and estimation of cell frequency of representative somatic CNVs in fibroblasts

a, Two of the three hiPSC lines obtained from fibroblast sample S1123-01 had the same duplication on chromosome X not detected in parental fibroblasts. **b**, PCR amplification across CNV breakpoints revealed that the duplication was present in parental fibroblasts at lower frequency (FBR=fibroblasts; CTRL=negative control) **c**, Scatter plot showing signal intensities associated with the PCR amplification across the breakpoints of the LM-CNV (Y axis, green). Signal for parallel amplification of a control region is shown on the X axis (blue). Each dot represents a single PCR event. There are significantly fewer dots for PCR in

CNV regions rather than for PCR in control region. **d**, The frequency of cells harboring the LM-CNV in fibroblasts is calculated assuming that frequency of such cells in hiPSCs is 100%, after normalizing event numbers for LM-CNVs by the control region. Counts of ddPCR events for the LM-CNV (green bars) and the control region (blue bars) allows estimating cell frequency in fibroblast of 12.6%. **e**, Duplication on chromosome 7 that was undetectable in parental fibroblasts by RD but detected as a faint band by PCR. **f**, This event had an estimated cell frequency in fibroblasts of 14.6% by ddPCR. **g**, Deletion on chromosome 8 that was undetectable in parental fibroblasts both by RD and PCR. **h**, This event had an estimated cell frequency in fibroblasts of 0.8% by ddPCR.

Table 1

Summary of validated line-manifested CNVs with additional experimental support obtained by PE analyses in hiPSC. For each CNV, no PE support was detected in the fibroblast sample. NA stands for events for which no successful ddPCR in both fibroblast and hiPSC could be conducted after tree attempts. Zero frequency suggests that a CNV is either not somatic or its frequency is beyond the detection limit of 0.1%.

Person	iPS	LM-CNV region chrom:start, type	Size, kb	Support, # of PE	Fraction in fibroblasts
Father S1123-01	#1	22:38755001, dup−	34	-	-
	#3	3:175005001, dup	59	3	NA
		5:168431001, dup	288	6	NA
		20:14809001, del	75	1	NA
		X:64962001, dup	67	3	12.6%
	#4	X:64963001, dup	65	5	12.6%
Mother S1123-02	#2	12:66253001, del	72	5	~0%
		13:111112001, del	48	3	-
	#11	4:130288001, del+	330	1	NA
	#17	7:133748001, dup	37	4	14.6%
		11:84329001, del	211	-	-
		20:15010001, del	182	5	NA
Proband S1123-03	#1	None	-	-	-
	#8	None	-	-	-
	#9	-	-	-	-
Father 03-01	#2	None	-	-	-
	#3	8:124671001, dup	33	-	-
		22:38753001, dup	36	-	-
	#9	None	-	-	-
Mother 03-02	#5	X:90672001, del	17	2	-
	#8	None	-	-	-
	#9	1:162043001, dup	65	-	-
		12:37961001, del	426	-	-
		18:70516001, del	27	-	-

Person	iPS	LM-CNV region chrom:start, type	Size, kb	Support, # of PE	Fraction in fibroblasts
		X:141153001, del	38	-	-
Proband 03-03	#2	14:76667001, del+	111	2	1.9%
		22:28832001, del+	47	2	~0%
	#3	5:263001, del +	134	4	-
	#4	11:84581001, dup+	107	6	~0%
Sibling 03-04	#1	1:243008001, del+	525	-	-
		7:2400001, dup+	400	7	-
		8:3558001, del+	127	3	-
		12:37993001, del+	429	-	-
	#5	1:234023001, del+	378	-	-
		8:43563001, del−	230	-	-
	#6	3:143236001, dup+	631	1	0.3%
		8:15540001, del+	75	1	0.8%
		10:70514001, dup+	622	3	0.4%
		10:74033001, dup+*	617	3	-

*predicted dispersed duplication from PE analysis. +CNV validated in late passages. −CNV not validated in late passages.