Title

Genomic integrity of human induced pluripotent stem cells across nine studies in the NHLBI NextGen program

Authors

Kanchan, Kanika

Iyer, Kruthika

Yanek, Lisa R

et al.

Publication Date

2020-07-01

DOI

10.1016/j.scr.2020.101803

HHS Public Access

Author manuscript

Stem Cell Res. Author manuscript; available in PMC 2021 July 01.

Published in final edited form as:

Stem Cell Res. 2020 July ; 46: 101803. doi:10.1016/j.scr.2020.101803.

Genomic integrity of human induced pluripotent stem cells across nine studies in the NHLBI NextGen program

Kanika Kanchan[a], **Kruthika Iyer**[a], **Lisa R Yanek**[a], **Ivan Carcamo-Orive**[b], **Margaret A Taub**[c], **Claire Malley**[a], **Kristin Baldwin**[d], **Lewis C Becker**[a], **Ulrich Broeckel**[e], **Linzhao Cheng**[a], **Chad Cowan**[f], **Matteo D'Antonio**[g], **Kelly A Frazer**[g], **Thomas Quertermous**[b], **Gustavo Mostoslavsky**[h], **George Murphy**[h], **Marlene Rabinovitch**[b], **Daniel J Rader**[i], **Martin H Steinberg**[j], **Eric Topol**[k], **Wenli Yang**[l], **Joshua W Knowles**[b], **Cashell E Jaquish**[m], **Ingo Ruczinski**[c], **Rasika A Mathias**[a,*]

[a]Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

[b]Department of Medicine, Cardiovascular Institute and Diabetes Research Center, Stanford University, School of Medicine, Stanford, CA, USA

[c]Department of Biostatistics, Bloomberg School of Public health, Johns Hopkins University, Baltimore, MD, USA

[d]Department of Molecular and Cellular Neuroscience, Dorris Neuroscience Center, The Scripps Research Institute, La Jolla, CA, USA

[e]Department of Pediatrics, Medical College of Wisconsin, Milwaukee, WI, USA

[f]Division of Cardiovascular Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

[g]Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA

[h]The Center for Regenerative Medicine, Boston Medical Center, Boston University School of Medicine, Boston, MA, USA

[*]Corresponding author.

Author contributions

KK, KI, LRY, MAT, CM, IR and RAM analyzed data, KK, KI, LRY, MAT, IR, RAM, KCB, KAF, JWK, CEJ helped with the interpretation of the data and critically drafted the manuscript, RAM, LCB, KAF, JWK, KB, UB, LC, CC, MD, IC-O, TQ, GM, GM, MR, DJR, MHS, ET, and WY all helped in the generation of the iPSC lines, and participated in the interpretation of the results. All authors reviewed and approved the final manuscript.

Credit author statement

Rasika Ann Mathias, Lewis C Becker, Kelly A Frazer, Joshua W Knowles, Kristin Baldwin, Ulrich Broeckel, Linzhao Cheng, Chad Cowan, Matteo D'Antonio, Ivan Carcamo-Orive, Thomas Quertermous, George Murphy, Gustavo Mostoslavsky, Marlene Rabinovitch, Daniel J Rader, Martin H Steinberg, Eric Topol, and Wenli Yang: Conceptualization, Resources and Methodology. Kanika Kanchan, Kruthika Iyer, Lisa R Yanek, Margaret A Taub, Claire Malley, Ingo Ruczinski and Rasika Ann Mathias: Formal Analysis. Kanika Kanchan, Lisa R Yanek, Ivan Carcamo-Orive, Margaret A Taub, Ingo Ruczinski, Rasika Ann Mathias, Kristin Baldwin, Kelly A Frazer, Joshua W Knowles, Cashell E Jaquish: Data interpretation and Writing-Original draft preparation. Kanika Kanchan, Ivan Carcamo-Orive, Joshua W Knowles, Rasika Ann Mathias and Ingo Ruczinski: Writing-Review and Editing.

Accession numbers

Access to data from the NextGen Consortium is available through individual dbGAP accession for the studies: phs001341.v1.p1, phs001074.v1.p1, phs001139.v1.p1, phs001212.v1.p1, phs000998.v2.p1, phs000924.v4, phs001325.v3, and phs000827.v3.p1.

Declarations of Competing Interest

None.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.scr.2020.101803.

[i]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[j]Department of Medicine, Section of Hematology-Oncology, Boston University School of Medicine, Boston, MA, USA

[k]Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA, USA

[l]Penn Center for Pulmonary Biology and Institute for Regenerative Medicine, University of Pennsylvania, Philadelphia, PA, USA

[m]National Heart, Lung and Blood Institute, NIH, Bethesda, MD, USA

Abstract

Human induced pluripotent stem cell (hiPSC) lines have previously been generated through the NHLBI sponsored NextGen program at nine individual study sites. Here, we examined the structural integrity of 506 hiPSC lines as determined by copy number variations (CNVs). We observed that 149 hiPSC lines acquired 258 CNVs relative to donor DNA. We identified six recurrent regions of CNVs on chromosomes 1, 2, 3, 16 and 20 that overlapped with cancer associated genes. Furthermore, the genes mapping to regions of acquired CNVs show an enrichment in cancer related biological processes (IL6 production) and signaling cascades (JNK cascade & NFκB cascade). The genomic region of instability on chr20 (chr20q11.2) includes transcriptomic signatures for cancer associated genes such as *ID1, BCL2L1, TPX2, PDRG1* and *HCK*. Of these *HCK* shows statistically significant differential expression between carrier and non-carrier hiPSC lines. Overall, while a low level of genomic instability was observed in the NextGen generated hiPSC lines, the observation of structural instability in regions with known cancer associated genes substantiates the importance of systematic evaluation of genetic variations in hiPSCs before using them as disease/research models.

Keywords

hiPSCs; Structural integrity; GWAS, VanillaICE; Oncogenes; Tumor suppressor genes

1. Introduction

Stem cells are primal cells with the capability of unlimited self-renewal and potential to differentiate into specific cell types. Two major types of human pluripotent stem cells are human embryonic stem cells (hESCs) and human induced pluripotent stem cells (hiPSCs); hESCs are derived from the undifferentiated inner cell mass of an embryo, whereas hiPSCs are reprogrammed from somatic cells by introducing defined pluripotency factors (Takahashi et al., 2007). Recently, hiPSCs have been evolved as a tool for disease modelling, drug development and cytotoxicity studies and individualized regenerative cell therapy (Shi et al., 2017). hiPSC lines reprogrammed from selected donors represent their genetic background and can provide a strong prototype to study the functional link between genetic variants and disease phenotype (Chamberlain, 2016; Warren and Cowan, 2018).

Numerous genetic variants associated with diseases have been identified by genome-wide association studies (GWAS)(Mahajan et al., 2014; Nikpay et al., 2015; Visscher et al., 2017; Tam et al., 2019). However, the functional validation of identified variants remains a challenge. Thus, the National Heart, Lung, Blood Institute (NHLBI) launched the Next Generation Genetic Association (NextGen) program to understand the causal effects of human genetic variations on heart, lung and blood disease phenotypes using hiPSC models (Warren et al., 2017). The NextGen program used iPSC technology to generate cell types relevant to heart, lung and blood disease processes from a large number of donors ascertained on the basis of heart, lung and blood phenotypes. The nine individual studies of the NextGen program integrate the results of genetic association studies (mainly GWAS) with hiPSC based functional studies to interrogate NHLBI-relevant phenotypes (see Supp. Table S1). The cell lines derived within the NextGen program are publicly available through WiCell, stem cell bank (www.wicell.org; NHLBI NextGen collection). The unique features of the NextGen generated hiPSCs include that the lines are derived from different cell types and from donors of several race/ethnicities using several reprogramming methods and then differentiated into various functional cell types.

Prior studies have documented genomic instabilities such as karyotypic abnormalities, single nucleotide variations (SNVs), copy number variations (CNVs) and insertions/deletions (INDELs) in hiPSCs (Mayshar et al., 2010). The abnormalities may occur via three different routes: (i) pre-existing variations in donor cells may be inherited by hiPSCs, (ii) the reprogramming procedure induced mutations, and (iii) passaging induced mutations. Studies suggest that genetic abnormalities in hiPSCs affect their clinical applicability and reliability as research models (D'Antonio et al., 2018; Martins-Taylor and Xu, 2012; Oliveira et al., 2014; Peterson and Loring, 2014; Yoshihara et al., 2017). Earlier genomic characterization studies using a small number of hiPSC lines (range:22–66 cell lines) have reported high mutational loads on the pluripotent cells (Gore et al., 2011; Hussein et al., 2011; Laurent et al., 2011; Mayshar et al., 2010). A recent large scale study of genome-wide profiling of hiPSC lines (N = 711) generated within Human Induced Pluripotent Stem Cells Initiative (HipSci) also reported that 41% of lines had one or more CNVs (Kilpinen et al., 2017).

To minimize the genomic instability in NextGen hiPSCs, all nine individual studies used non-integrative reprogramming methods and cell line maintenance under feeder free conditions (Warren et al., 2017). Since NextGen is a substantial resource of hiPSCs publicly available to the greater scientific community through WiCell, we performed a thorough analysis for genomic instabilities. The objective of our study was to identify newly acquired CNVs (CNVs that likely occur during reprogramming/passaging or that are sub-clonal in the cell of origin) in the NextGen hiPSCs. In this work, we extend our understanding of the genomic structural integrity of hiPSC lines across a wide range of donor DNA, human diseases and cell engineering techniques. The integrity of hiPSC lines was assessed by automated CNV calling followed with subsequent visual inspection of CNV signals. In this work we provide an in-depth understanding of the genomic stability of NextGen hiPSC lines. In future work, these hiPSCs may serve as surrogates for genetics and functional studies for several different disease phenotypes.

2. Methods

2.1. Generation of hiPSCs

The hiPSC lines within the NextGen program were reprogrammed at nine study sites (Warren et al., 2017) from various cell types including PBMCs, fibroblasts and erythroblasts, and using different cell engineering techniques including nonintegrating episomal vectors, lentivirus and Sendai virus. The culture medium, matrix/feeders and passaging methods used at each study sites are shown in Supp. Table S1. Reprogramming methods and maintenance of the NextGen hiPSC lines have been previously published (Warren et al., 2017 and reference therein) and also available on the Wicell website (www.wicell.org; NHLBI NextGen collection). The subjects include diseased individuals (sickle cell disease, platelet aggregation, coronary disease, pulmonary hypertension, arrhythmias, lipid disorders, left ventricular dynamics and hypertrophy) as well as healthy controls. At some study sites (Table 1) multiple passages were available per donor subject offering up the opportunity to look for acquisition of CNVs between differing passages of the same hiPSC.

2.2. Genotyping, genotype calling and quality control of the genotype data

Genotyping was performed through the NHLBI DNA Resequencing and Genotyping (RS&G) Service Center using the Illumina Infinium MEGA Chip. Genotype calls were made using GenomeStudio (version 2011.1) and Genotyping Module (version 1.9.4) and a total of 2,036,060 SNPs for 2,585 samples were released. The numbers of the genotyped samples and the entire workflow of the analyses are shown in Supp. Fig 1A. We applied a set of quality control (QC) steps to the genotype data and excluded SNPs as follows: 501,246 monomorphic SNPs; SNPs with > = 10% missing data; and SNPs showing deviations from Hardy–Weinberg equilibrium at a P value < 10^{-6}. This resulted in 1,321,228 retained SNPs. Furthermore, SNPs with a MAF of <5% and SNPs in high linkage disequilibrium (LD) were also removed from the dataset to achieve 202,042 LD pruned SNPs (Supp. Fig 1B). It should be noted that the SNPs were dropped only for QC steps and not from the CNV analysis where all SNPs were included. Using a set of N = 202,042 LD pruned markers we estimated the identity by state (IBS) for all 2,585 samples by PLINK (Purcell et al., 2007); a total of 3,339,821 pairs for 2,585 samples were generated.

Of the 2,585 samples, we excluded 510 samples as follows: 428 samples with no pair for comparison, 52 pairs with sample swaps at plating where the proportion IBD (PI_HAT) = < 0.6 and 13 pairs that were technical replicates (Supp. Fig 1A). The final pairwise comparisons for IBS were then limited to 1,520 pairs including 1,114 pairs of donor DNA-hiPSC lines and 406 control pairs (Supp. Fig 1A). The IBS plots for all donor and hiPSC line pairs across 1,114 subjects and 406 HapMap controls indicated that all pairs passing IBS QC have Z2 = 1 and Z1 = Z0 = 0 where Z2, Z1 and Z0 are the probabilities of sharing 2, 1, and 0 alleles identical by state between the pair, respectively (Supp. Fig 1C). From the 1,114 donors and hiPSC pairs, 506 pairs were included in the structural integrity analysis because of the plating approach during genotyping, and other QC reasons that are described in detail in the next section (Supp. Fig 1A).

2.3. CNV calling from genotype data and QC measures

GenomeStudio software (Illumina inc.) provided the Log R ratios (LRRs: a normalized measure of the total signal intensity for two alleles of the SNP) and B-allele frequencies (BAFs: a normalized measure of the allelic intensity ratio of two alleles) on all the 2,036,060 genotyped SNPs. We did not exclude any monomorphic SNPs or low MAF SNPs from CNV calling as they provide information about copy number by total intensity. For CNV calling we applied a Hidden Markov Model (HMM) using a threshold of 10 supporting probes on the genotype data to identify CNVs in the samples using R package 'VanillaICE' (Scharpf et al., 2008). We used the hg19/GRCh37 reference version of the human genome in the CNV calling pipeline. The X and Y chromosomes were removed due to the complexity of reliably determining copy number in these copy variable and highly repetitive chromosomes.

Given that copy number estimation is very sensitive to technological artifacts and differences in the preprocessing and normalization steps (Halper-Stromberg et al., 2011; Scharpf et al., 2011a; Scharpf et al., 2011b) we delineated signal intensities, and subsequent CNV calling was carried out separately for each plate. The main reason behind this sensitivity is that copy numbers are based on total probe A and B intensities (as compared to genotypes, which are based on relative probe intensity). Moreover, the number of CNVs called depends on the quality of the data, which differs between plates. Specifically, the larger the variability in the LRR of a sample (i.e. the lower the quality), the more CNVs are called on average. Since accuracy of CNV calls does not improve with decrease in quality of the genomic data, the increase in CNV calls represents an increase in false positive calls (see Supp. Fig 2A). The top panel shows the median absolute deviation (MAD) of the LRR per sample (x-axis) versus the number of CNVs called in the sample (y-axis). The red line is a loess smoother based on all samples, the blue line is a loess smoother based only on the samples with fewer than 100 CNV calls (which still shows a doubling on average of the number of CNVs called, from 9.8 for the best quality calls, to 18.2 for those with LRR MAD near 0.25). The bottom panel shows boxplots of the MAD LRRs for the samples per plate, showing clear between plate differences.

Furthermore, technical artifacts across samples within a plate tend to be more alike than technical artifacts across plates. However, false positive calls due to technical artifacts can be greatly reduced (at no change of sensitivity) if samples to be compared are plated together (Scharpf et al., 2012). Thus, we have restricted our structural integrity analysis to the pairs that were plated together. In addition, we also compared MAD of LRR and total number of called CNVs within each sample. The plot suggests that any sample with more than 50 CNVs should not be included in the analysis (Supp. Fig 2B). Based on the aforementioned reasons we dropped 524 pairs that were not plated together and 84 pairs where either donor or hiPSC had >50 CNVs. As a final dataset, 506 hiPSC-donor DNA pairs were analyzed for their structural integrity (Supp. Fig 1A).

2.4. Structural integrity analysis

Given sensitivities in calling CNVs using an automated pipeline (Halper-Stromberg et al., 2011), we attempted to minimize false positives by generating sub-chromosomal plots from the LRRs and BAFs for manual inspection. Since we were specifically interested in CNVs

that were acquired in hiPSC lines relative to the donor DNA, we contrasted the CNVs in donor DNA and hiPSC lines to identify the CNVs that were called in the hiPSC line but not the donor DNA, or those that were called in both but were of differing lengths between the hiPSC line and donor DNA. Then we manually inspected the LRR and BAF plots for each genomic region harboring the CNV, to examine the structural integrity of our 506 hiPSC lines (Supp. Fig 1C). Only CNVs that passed dual manual inspection (by two independent readers) were confirmed as the newly acquired CNVs.

2.5. Gene ontology and gene set enrichment analysis (GSEA)

The Gene List Analysis tool of PANTHER (Protein ANalysis THrough Evolutionary Relationships) classification system (http://www.pantherdb.org) was used for gene ontology annotation, functional classification and GSEA. Briefly, genes mapping to the regions of acquired CNVs were uploaded to PANTHER user interface, and functional classification and statistical over-representation analyses were performed. Both functional classification and statistical over-representation analyses were done using the default Homo sapiens reference list (20,996 genes) in the database. Statistical overrepresentation test finds functional classes that are statistically over- (or under-) represented in the input list, compared to the default (reference) list. We used the PANTHER GO-Slim annotation dataset for analysis. Significance was determined with the Fisher's exact test and the False Discovery Rate (FDR) was calculated using the Benjamini-Hochberg procedure. We performed the statistical over-representation test in a sequential manner based on the number of acquired CNVs per line: (i) Input set1: genes mapping to 258 CNVs (149 hiPSC lines that acquired ≥1 CNV); (ii) Input set2: genes mapping to 165 CNVs (56 hiPSC lines that acquired ≥2 CNVs); (iii) Input set3: genes mapping to 107 CNVs (27 hiPSC lines that acquired ≥3 CNVs); (iv) Input set4: genes mapping to 59 CNVs (11 hiPSC lines that acquired ≥4 CNVs) and (v) Input set5: genes mapping to 31 CNVs (4 hiPSC lines that acquired ≥5 CNVs). The sequential GSEA was similarly done stratified for acquired amplifications (Amps) and deletions (Dels) separately.

2.6. Transcript expression analysis

Among the 15 hiPSC lines that were found to harbor a chr20q11.2 CNV, 9 hiPSC lines were reprogrammed at one study site (site5). We relied on RNA sequencing (RNA-seq) previously generated to validate the CNVs testing for differential expression at genes mapping to this region for these lines. The detailed protocol for RNA-seq and pre-processing is published by Carcamo-Oribe et al., (Carcamo-Orive et al., 2017). Briefly, we used log2 counts per million (CPM) following TMM normalization (Robinson and Oshlack, 2010) implemented in edgeR (Robinson et al., 2010) RNA-seq data. Data were adjusted for the effects of 8 sequencing batches and 2 RNA preparation kits and genes with CPM ≥ 1 for at least 30% of the samples were retained.

The hiPSC lines harboring chr20q11.2 CNV were termed as carrier hiPSC lines (N = 8) and not harboring the CNV were termed as non-carrier hiPSC lines (N = 8). The core overlap region of Amps among eight carrier hiPSC lines is 0.87 Mb (chr20:29810887-30682979). Mapping transcripts were ascertained by overlap between the CNV region and transcript's start site (i.e. if the start site of the transcript lies within CNV region, they were considered

as mapped transcripts). We calculated fold changes and *p*-values for differential expression using moderated t-statistics using the BioConductor package limma, and calculated *p*-values using the Bioconductor q-value package. Confidence intervals were calculated in R based on log fold changes and moderated standard errors.

2.7. Statistical analysis

Linear regression models were used to determine the effects of cellular reprogramming approaches and passaging on the structural variations of hiPSC lines. Briefly, we generated a matrix of hiPSC with newly acquired CNVs and their phenotypic characteristics and performed regression analysis in R (version 3.3.3). We estimated the significance using three different linear models, first for all CNVs combined, second for only amplifications and third for only deletions. The beta estimates were used to interpret the results and the p-values from the regression models are reported.

Statistical enrichment of oncogenes in the acquired Amps and tumor suppressor genes (TSG) in the acquired Dels region was determined using Fisher's exact test. We obtained the list of cancer related genes from the COSMIC (https://cancer.sanger.ac.uk/cosmic) database. A 2*2 contingency table was set up for amplifications in oncogenes and deletions in TSG. Fisher's exact test was done in R (version 3.3.3) for both tables, and ORs and *P*-values from the test are given.

3. Results

3.1. Genomic structural integrity of NextGen hiPSC lines

Out of 506 hiPSC lines compared to donor DNA, 357 lines showed no detectable difference in CNV calls compared to the paired donor DNA. We observed differences in CNVs between donor and hiPSC in 149 lines: 93 lines had one CNV difference, 29 lines had two and 27 lines had three or more CNVs different from the paired donor DNA. One hiPSC line had 15 CNVs, while the remaining 148 lines had six or fewer (Fig 1A). We did not observe any instance of aneuploidy in our hiPSC collection. The summary of the hiPSC lines examined and newly acquired CNVs are shown in Table 1. Overall, we identified 258 newly acquired CNVs (Supp. Fig 3) across the 149 hiPSC lines. The newly acquired CNVs include 111 Amps and 147 Dels. Size of the acquired CNVs ranged from 0.02 Mb to 26.77 Mb (mean = 0.98 Mb, median = 0.28 Mb). For each of the 149 lines harboring one or more CNVs, we calculated the cumulative size of the gained CNVs (mean = 1.09 Mb; median = 0.26 Mb) and the cumulative impact per hiPSC line was small. The majority of hiPSC lines (N = 132, 85%) had cumulative CNV coverage of less than 2 Mb. The cumulative size of CNVs in the remaining 17 lines, ranged from 2 to 4 Mb in fourteen lines, a deletion of 12 Mb in one line, and Amps of 27 Mb in two lines (Fig 1B). The distribution of acquired CNVs was relatively even across all the chromosomes, with a few exceptions (Fig 1C). First, we did not observe any CNVs on chr21. Second, we observed a large region of Amp on chr17 in two hiPSC lines that overlaps 240 protein coding genes. Third, we observed a set of six regions on five chromosomes (chr1, chr2, chr3, chr16 and chr20) with recurrent CNVs (defined as more than one line with a CNV mapping to the region), affecting 30 hiPSC lines (Fig 1C).

3.2. Functional enrichment of genes mapping to acquired CNVs

A total of 1,649 genes mapped to the regions of acquired CNVs and we were able to map 1,395 of these genes to PANTHER IDs in the PANTHER database (Mi et al., 2019). On stratifying the acquired CNVs into Amps and Dels 1,010 genes mapped to the regions of acquired Amps and 700 genes mapped to the regions of acquired Dels. Of these 847 genes in Amps category and 585 genes in Dels category were mapped to PANTHER IDs. Using the biological process GO annotation dataset, we identified functional classification of the mapped genes (genes mapping to PANTHER IDs) under 16 GO terms. A descriptive overview of all GO terms, and gene sets within each GO term are represented in Supp. Table S2.

We detected enrichment in 22 biological processes gene sets at an FDR of 5% ($q < 0.05$) including a total of 470 genes (Fig 2A). To further understand if there were differences in regions with repetitive CNV acquisition over the 149 lines, we performed GSEA sequentially: Input set1/Input set2/Input set3/Input set4/and Input set5 included genes mapping to regions of acquired CNVs in 149(≥1CNVs), 56(≥2CNVs), 27(≥3CNVs), 11(≥4CNVs) and 4(≥5CNVs) lines. This consisted 1,649/1,016/630/308/and 202 input genes, corresponding to 1,395/866/532/246/and 163 mapped genes, respectively. There were total 55 gene sets with a significant enrichment ($q < 0.05$) for at least one of the five input sets of genes Supp. Table S3. Interestingly, the fold enrichment (FE) of the identified gene sets increased markedly as we went from Input set 1 through to Input set 5, suggesting stronger enrichment in pathways for the more recurrent regions of acquired structural variation.

Of these, we highlight two cancer related signaling cascades that were over-represented at an FDR of 5% in all five input sets: “I-kappaB kinase/NF-kappaB signaling cascade” (GO:0007249) and “JNK cascade” (GO:0007254). For the I-kappaB kinase/NF-kappaB cascade FE was 3.10 ($q = 4.32E-02$) for the genes in Input set1, 4.64 ($q = 2.51E-03$) for the genes in Input set2, 6.96 (q = 9.7E-05) for the genes in Input set3, 13.81 ($q = 3.26E-07$) for the genes in Input set4 and 18.94 ($q = 6.18E-08$) for Input set5. For the JNK cascade, FE increased from 4.10 ($q = 1.68E-02$) to 5.51 ($q = 5.31E-03$), 8.07 ($q = 6.07E-04$), 17.46 ($q = 1.18E-06$), and 26.35 ($q = 4.56E-08$), respectively. In addition, we also identified the over-representation of “Cytokine production” (GO:0001816) among five input sets of genes with a FE of 3.50 ($q = 9.12E-02$) in Input set1, 5.64 ($q = 4.81E-03$) in Input set2, 8.26 ($q = 5.5E-04$) in Input set3, 17.86 ($q = 1.07E-06$) in Input set4 and 23.96 ($q = 5.3E-07$) in Input set5. Specifically, GO-term defining “IL-6 production” (GO:0032635), a pro-tumorigenic cytokine was identified to be overrepresented with FE raising from 8.03 ($q = 6.48E-03$) to 12.93 ($q = 3.8E-04$) to 21.05 ($q = 1.05E-05$), 45.52 ($q = 2.74E-08$) and 68.72 ($q = 1.39E-09$) from Input set 1 through to Input set 5 (Fig 2B).

To understand if the observed enrichment is driven either by acquired Amps or Dels, we performed sequential GSEA for gene sets mapping to the regions of acquired Amps and Dels separately. At an FDR of 5%, the FE for both signaling cascades, cytokine and IL6 production was higher in the gene sets mapping to regions of acquired Dels than gene sets mapping to all CNVs combined. There was only 1 hiPSC line that had ≥5 Dels, therefore here we had four input sets including 700/428/350/andl39 input genes corresponding to

585/352/285/andl07 mapped genes. The FE for GO:0007249 was 6.78 (q = 4.83E-05) for the genes in Input set1, 10.4 (q = 1.77E-06) for the genes in Input set2, 12.73 (q = 1.78E-07) for the genes in Input set3, and 25.49 (q = 2.25E-08) for the genes in Input set4. For GO:0007254 FE increased from 8.06 (q = 2.31E-04) to 12 (q = 2.62E-05), 14.76 (q = 4.62E-06), and 39.4 (q = 1.23E-09), respectively. GO:C0001816 also represented the similar pattern as the FE went from 6.6 (q = 6.49E-03) in Input set1, to 10.9 (q = 1.76E-04) in Input set2, 13.42 (q = 3.80E-05) in Input set3, and 35.84 (q = 2.18E-08) in Input set4. The most dramatic effect was observed in over-representation of GO:0032635 with FE raising from 18.92 (q = 2.08E-05) to 31.3 (q = 5.83E-07) to 38.48 (q = 1.02E-07), and >100 (q = 5.36E-11) from Input set 1 through to Input set 4 (Fig 2C). Of note, all these GO-terms were the same as those seen in the combined analysis above; they were found to be statistically significant in the gene sets mapping to regions of acquired Dels and not Amps from the stratified analysis.

3.3. Enrichment of cancer related genes in the repetitive regions of CNVs in hiPSCs

Given the increasing FE of biological process in regions with repetitive CNV acquisition we performed an in-depth exploration of the genomic regions with instability. The sub-chromosomal region chr20q11.2 was the most recurrent region. Acquired CNVs in this region were observed in 16 hiPSC lines (15 Amps and 1Del) and the core overlap among the lines harboring amplifications was 0.60 Mb. This recurrent region of Amp encompasses 15 genes including known cancer associated loci *ID1, BCL2L1* and *TPX2* (Fig 3A). In addition to chr20q11.2, we identified an additional region on the p arm of the chr20. This region affects *MACROD2* (MACRO Domain Containing 2) which is expressed at low levels in hiPSCs (Fig 3B). A region on chr16 includes two acquired Dels and one Amp. This region overlaps with the tumor suppressor gene *WWOX* (WW domain containing oxidoreductase) (Fig 3C). Three recurrent regions of Dels were observed on chr3, chr2 and chr1. The chr3 region overlaps *FHIT* (Fragile histidine triad) gene, that acts as a tumor suppressor and plays a role in apoptosis (Fig 3D). The region on chr2 overlaps a tumor suppressor gene RPRM (Reprimo, TP53 dependent G2 arrest mediator candidate) that is involved in the regulation of p53-dependent cell-cycle arrest (Xu et al., 2012) and the p53 pathway is linked to reprogramming (Krizhanovsky and Lowe, 2009)(Fig 3E). The chr1 region affects 10 genes including three cancer associated genes *ITGB3BP, ROR1* and *EFCAB7*. This region also encompasses the *FOXD3* gene, required for maintenance of pluripotent cells in the pre-implantation and peri-implantation stages of embryogenesis (Fig 3F).

The genomic positions, sizes, recurring scores (number of lines with the acquired CNV) and the genes affected by the recurrent regions of CNVs are given in Supp Table S4. We queried these recurrent CNVs against the Database of Genomic Variants (dgv.tcag.ca) (MacDonald et al., 2014) and the DGV structural variants overlapping with CNV regions are shown in Supp. Fig 4. The average size of the DGV structural variants was greater than our recurrent CNVs, except for the chr1 and chr20q arm CNVs.

3.4. Transcriptomic integrity of hiPSC lines harboring chr20 amplification

We had a total of 1,280 assembled transcripts on chr20 for 16 hiPSC lines (including both carriers and non-carriers) in our RNA-seq dataset. Among those, 631 transcripts passed the

CPM QC threshold (CPM > 1 in more than 30% of all samples) and 18 mapped to the core Amp region (the intersection of the chr20 Amp region among all eight carriers). We observed increased expression of four genes clustering at one end of the core amplification region: *XKR7, CCM2L, RP1-310O13.7* and *HCK*. Fig 4 shows the fold changes (FCs) versus the genomic location of all 18 mapped genes on chr20, and the genes with significant up-regulation are shown in red. The FCs for the up-regulated genes are: *XKR7* (FC = 2.13, P = 8.80E-04, CI = 1.44–3.16), *CCM2L* (FC = 1.97, P = 3.86E-03, CI = 1.29-3.00), *RP1-310O13.7* (FC = 1.76, P = 3.36E-02, CI = 1.05–2.95), and HOC (FC = 2.16, P = 4.84E-02, CI = 1.01–4.62). Among the significantly up-regulated genes *HCK* gene is a known cancer associated gene. The FCs for other known cancer associated genes in the region are *ID1* (0.89), *BCL2L1* (1.17), *TPX2* (1.16) and *PDRG1* (1.21), but these where not statistically significant.

3.5. CNV patterns in sibling hiPSC lines

In our hiPSC collection there were 106 (21%) sibling hiPSC lines (two hiPSC lines reprogrammed from the same donor that have different passage numbers) reprogrammed from 53 donors. Among sibling hiPSCs, 81 lines showed no detectable difference in CNV calls compared to the paired donor DNA. We observed differences in CNVs between donor and hiPSC in 25 lines: 17 lines had one CNV difference, 6 lines had two and 2 lines had three CNVs different from the paired donor DNA (Table 2). In those with acquired CNVs, we compared the LRR and BAF plots between parent DNA and both early and late passage hiPSC lines. The CNV patterns were identical for both early and late passage hiPSCs (see Supp. Fig. 5) for 56% of the lines where we noted an acquired CNV. The passage numbers and CNVs acquired by each sibling hiPSC line are given in Supp. Table S5.

3.6. Effect of reprogramming method, and donor phenotypes on structural instability

To test the hypothesis that the choice of cell reprogramming and passaging might induce instability in the hiPSC lines we evaluated structural integrity of the different hiPSC lines among the cell reprograming approaches and passage number. We also looked at the effects of donor cell source, sex, and age-related effects on structural integrity in the hiPSC lines. A detailed overview of all the different hiPSC lines, donor cell type, reprograming method, passage number, genomic position of CNV, and CNV status is given in Supp Table S6. Given that our hiPSC lines are derived using non-integrating methods we did not observe any effect of cell reprogramming method on the stability of the lines. However, we observed a significant association between the structural variations of different hiPSC lines as it relates to passage number ($p = 0.01$). Notably, the hiPSC lines with higher passage number acquired more amplifications ($p = 0.001$). We also observed the highest number of acquired CNVs among the lines that were reprogrammed from fibroblast ($p = 0.01$) as compared to PBMCs. On stratifying the CNVs into Amps and Dels, the significant relationship was observed with Dels only ($p = 0.04$). In context of donor phenotypes, we did not observe any relationship based on the donor sex, but we observed age-related effects ($p = 0.004$). Lines derived from older individuals gained more Dels ($p = 0.02$) as compared to Amps.

4. Discussion

Genomic instability in hiPSC lines has been well recognized. To date there is only one large scale study that has assessed the phenotypic consequences of recurrent genomic CNVs from various standpoints including epigenome, transcriptome and proteome to cell differentiation and morphology (Kilpinen et al., 2017). This large scale genomic characterization study (Kilpinen et al., 2017) including 711 iPSC lines derived from 301 healthy individuals in the HipSci consortium used genotyping arrays to detect CNVs. The hiPSC lines in the HipSci collection (http://www.hipsci.org) are derived mostly from fibroblasts using Sendai virus. The authors identified trisomies in 4% of cell lines and 41% of lines harbored at least one or more CNVs of 7.15 Mb on average. They observed 35 recurrent regions of CNVs including 20 Amps, 11 Dels, three regions with both gain and loss, and one full chromosome duplication. Three most frequent CNVs identified in their study are X trisomy, chromosome 17 and 20. Genomic instability in our NextGen hiPSC lines was 29% with average size of 1.09 Mb, which is lower than the genomic instability reported in the HiPSci collection. The hiPSC lines in our collection weren't affected from any aneuploidies, and we observed 6 recurrent regions of CNVs on chromosomes 1, 2, 3, 16 and 20 (Fig 3). The Amp on chr20 and Del on chr1 have been previously observed in pluripotent cells but others have not been identified previously, to our knowledge. In addition to hiPSCi study, there are two other studies that evaluated the integrity of hiPSCs (Panopoulos et al., 2017) and hiPSC derived megakaryocyte cells (Kammers et al., 2017) from a genetic and transcriptomic perspective. However, the cell lines analyzed in both these studies are N = 222 and N = 14, respectively and is smaller than our NextGen collection.

Earlier reports included small-scale studies that evaluated chromosomal (Mayshar et al., 2010) and sub-chromosomal integrity (Gore et al., 2011; Hussein et al., 2011; Laurent et al., 2011). The first comprehensive evaluation of the chromosomal integrity or aneuploidy of hiPSC lines was performed through transcriptional profiling and reported the identification of a substantial number of abnormal cell-lines (Mayshar et al., 2010). Laurent et al., 2011 compared CNVs among normal somatic cell lines and hiPSC lines and reported that hiPSC lines contain more deletions than normal somatic cell lines. This work also found that the reprogramming process is associated with selection for deletions that affect tumor-suppressor genes, whereas maintenance of the cell lines is associated with selection for duplications in oncogenic genes (Laurent et al., 2011). Later, Hussein et al., 2011 observed a higher frequency of CNVs compared to the donor fibroblast cells. However, the number of CNVs in hiPSCs decreased with passaging (Hussein et al., 2011). They also concluded that the reprogramming process is associated with high mutation rates, causing increased levels of CNVs that lead to genetic mosaicism in early hiPSC lines (Hussein et al., 2011). In the aforementioned studies, the numbers of hiPSC lines were small (22–66 hiPSC lines) and the cells were reprogrammed using both integrating methods and non-integrating method (Mayshar et al., 2010; Hussein et al., 2011; Laurent et al., 2011). Our NextGen collection is very diverse in terms of source cells including PBMCs, fibroblasts and erythroblasts, and the hiPSC lines were reprogrammed using non-integrating methods only.

Another aspect of genomic instability was studied by Gore and colleagues (Gore et al., 2011). They explored the point mutations in hiPSCs using exome sequencing and validated

the ones that were fixed in the hiPSCs. Their study identified many missense mutations in the hiPSC lines that are predicted to alter protein function, and point mutations were enriched in cancer associated genes. We used SNP arrays to study structural instabilities in our NextGen hiPSC collection. Given that SNP arrays have limited resolution we were able to detect the larger CNVs but not point mutations in the NextGen hiPSC lines. It should be noted that every procedure that requires a "call" needs to consider the trade-off between sensitivity and specificity. In a typical CNV analysis, the false positive CNV calls substantially outnumber the true CNV calls depending on various factors such as the size of the CNV call and the data quality in that sample. Thus, it is imperative to use stringent requirements for data quality such as the MAD for the LRRs and a minimum number of probes supporting a CNV in an array-based approach. Obviously, this comes at the expense of possibly missing some true CNVs, specifically smaller CNVs. Compared to sequencing, array-based CNV calling has some advantages such as its high-throughput capacity, but also disadvantages such as the CNV resolution and the size detection limits. In our study, we identified 258 newly acquired CNVs across the 149 hiPSC lines, ranging from 0.02 Mb to 26.77 Mb. There could be CNVs smaller than 20 kB in truth, but these would be extremely hard, if not impossible, to detect with array-based technology. We recognize these limitations with respect to the CNV calling from the array data. In contrast, CNV calling based on short reads from sequencing has the capability to detect much shorter CNVs. In addition, sequencing also has a much higher resolution for break point detection, and sometimes even yields break points at base pair resolution.

Our study outlines the systematic genetic characterization of hiPSC lines at a large scale (>500 hiPSC lines). We observed high genomic structural integrity of the hiPSC lines, with 71% of the hiPSCs in the NextGen consortium showing no detectable CNVs. Among the 29% (N = 149) hiPSC lines that acquired CNVs: 49 lines acquired only Amps, 68 lines acquired only Dels and 32 lines acquired both Amps and Dels. The frequency of acquired Amps (43%) was lower than the frequency of Dels (57%) but the average size of Amps (1.28 Mb) was larger than Dels (0.76 Mb). In addition, to the recurrent CNVs on chromosomes 1, 2, 3, 16 and 20, one of the previously reported recurrent CNV on chr17 (chr17q11.2) was detected in one hiPSC line in our collection. We acknowledge that omission of chrX from our study is a limitation of our study. Our in-house CNV caller (VanillaICE) was developed only for the autosomes, implementing a likelihood in the hidden Markov model that reflects the data structures of LRRs and BAFs in autosomes. We understand that studying the structural instabilities on sex chromosomes will provide a complete structural landscape of all chromosomes of NextGen hiPSC lines, and additional work will be needed to examine sex chromosomal instabilities in these data. Generally, in our collection, hiPSCs acquired Dels included TSG whereas Amps included oncogenes, suggesting that cancer related genes provide additional selective advantage to the hiPSC line. On performing the Fisher's exact test, we observed a statistically significant enrichment of oncogenes in the Amp regions (OR = 1.80; $p = 0.02$) and TSG in the Del regions (OR = 2.11; $p = 0.01$).

We identified enrichment of cancer related signaling cascades and biological process among the genes mapping to the genomic regions of acquired CNVs. The FE for both I-kappaB kinase/NF-kappaB and JNK signaling cascades increased sequentially with repetitive CNV acquisition in our collection. However, it should be kept in mind that inflammatory pathways

are complicated, involved in many cellular processes, and are detected in almost every perturbation being analyzed. Cytokine production is a crucial biological process in context of cancer as cytokines exert both pro- and anti- cancer effects. Cytokines bind to their cognate receptor on the effector cells and promote or inhibit tumor development and progression in autocrine and paracrine manners (Van Gorp and Lamkanfi, 2019). A number of cytokines including interleukin-1 (IL-1), interleukin-6 (IL-6), interleukin-15 (IL-15), tumor necrosis factor (TNF), transforming growth factor-β (TGF-β) and vascular endothelial growth factor (VEGF) are implicated in tumorigenesis (Dranoff, 2004; Van Gorp and Lamkanfi, 2019). Among these cytokines, IL-6 is highly upregulated in many cancers and promotes several tumorigenic activities such as enhanced cell proliferation, survival, invasion and angiogenesis (Taniguchi and Karin, 2014). Elevated levels of IL-6 stimulate hyperactivation of IL-6/JAK/STAT3 signaling pathway that promotes tumor growth and progression and hinder antitumor immunity (Johnson et al., 2018).

Querying our recurrent CNVs against structural variants in the DGV suggested that these CNVs have been reported in human populations previously. This highlights that CNVs identified in our hiPSC lines are recurrent in the context of comparison to prior publications. Numerous studies have reported the chromosome 20 (20q11.21) Amp as a recurrent mutation in hESCs (Avery et al., 2013; International Stem Cell et al., 2011; Lefort et al., 2008; Narva et al., 2010; Nguyen et al., 2014; Spits et al., 2008; Wu et al., 2008) and in hiPSCs (Kilpinen et al., 2017; Laurent et al., 2011; Martins-Taylor et al., 2011). From our study, we conclude that only 2.96% (15 of 506) of our hiPSC lines were affected by this amplification. This is well below the previously published studies where the percentage of affected hiPSC lines ranged from 13 to 50% (Elliott et al., 2010; Laurent et al., 2011; Martins-Taylor et al., 2011). Previous studies in hESCs and hiPSCs reported that this Amp region contains several genes, including the known pluripotency-associated gene *DNMT3B* (Laurent et al., 2011), and *BCL2L1*, which encodes the anti-apoptotic protein BCL-X (International Stem Cell et al., 2011; Laurent et al., 2011). Our data confirmed that the duplicated region of chr20 (20q11.21) in all 15 hiPSC lines harbors three carcinoma associated genes; namely *ID1, BCL2L1* and *TPX2*. *ID1* plays role in cell growth, senescence, and differentiation and is associated with the salivary gland (Hu et al., 2017) and breast cancer (Gumireddy et al., 2014). *BCL2L1* acts as an apoptotic inhibitor and is related to many types of cancers such as melanoma, glioblastoma (Trisciuoglio et al., 2017), breast carcinoma (Choi et al., 2016) and colorectal carcinoma (Koehler et al., 2013). *TPX2* participates in cell cycle progression and this gene is linked to hepatocellular carcinoma (Hsu et al., 2017).

We extended the investigation of the genomic region with instability on chr20 of our hiPSC lines to the transcript level. A total of 18 transcripts mapped to the core overlap region of Amp, of which 13 are protein coding. In particular, there were four transcripts that were significantly different between the carrier and non-carrier Amp lines. Among the differentially expressed transcripts, transcript of *HCK* gene showed highest log2 FC. *HCK* is a member of the Src family of tyrosine kinases and is associated with tumor progression in colorectal cancer (Roseweir et al., 2019). The Amp region consists of transcripts for five cancer associated genes including *ID1, BCL2L1, TPX2, PDRG1* and *HCK*. Although we didn't observe the significant p-values for other cancer associated genes except *HCK* the FCs

were relatively high, and our limited sample size may impact the power to find these differences statistically meaningful. Confirming that the presence of oncogenes in the genomic regions of instability provide additional selective advantage to the hiPSC lines.

Our study design offered us an opportunity to study the CNV patterns of sibling hiPSC lines and compare with parent cell. The patterns were almost identical for both early and late passage hiPSCs for all instances where we noted an acquired CNV. Our work also demonstrates that using non-integrating reprogramming method can limit the occurrence of CNVs in the hiPSCs. However, the choice of source cell may affect the occurrence of the structural variations in the hiPSC lines. Among the hiPSC lines that acquired CNVs, the highest number of CNVs was found in lines generated from fibroblasts. It should be noted that these lines were compared to genomic DNA from blood/PBMC sources and not to donor fibroblasts and it is highly plausible that a meaningful number of the apparently 'acquired' CNVs might have originated from the source fibroblast.

5. Conclusions

In conclusion, our data shows the overall high quality of the hiPSC lines generated through the NextGen program. Our results also confirm that some sub-chromosomal regions are particularly susceptible to genomic instability and that in this large set of hiPSCs from NextGen, the acquired CNVs show a specific enrichment for cancer related biological process and signaling cascades. Additionally, the recurrent regions of acquired CNVs are enriched for cancer related genes. Given this cancer related enrichment, the estimation of genomic instability of hiPSCs is necessary for both basic research and hiPSC-based therapies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NHLBI projects U01 HL72518-05, U01 HL107446, U01 HL107388, U01 HL107393, U01 HL107436, U01 HL107442, U01 HL107437, U01 HL107440, U01 HL107443, NHGRI projects U01 HG006398, R01 HL112064 and NIDDK project P30DK116074. Genotyping was performed through the RS&G Service by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under U.S. Federal Government contract number HHSN268201100037C from the NHLBI.

References

Avery S, Hirst AJ, Baker D, Lim CY, Alagaratnam S, Skotheim RI, Lothe RA, Pera MF, Colman A, Robson P, et al., 2013 BCL-XL mediates the strong selective advantage of a 20q11.21 amplification commonly found in human embryonic stem cell cultures. Stem Cell Rep. 1, 379–386.

Carcamo-Orive I, Hoffman GE, Cundiff P, Beckmann ND, D'Souza SL, Knowles JW, Patel A, Papatsenko D, Abbasi F, Reaven GM, et al., 2017 Analysis of transcriptional variability in a large human iPSC library reveals genetic and non-genetic determinants of heterogeneity. Cell Stem Cell 20, 518–532 e519. [PubMed: 28017796]

Chamberlain SJ, 2016 Disease modelling using human iPSCs. Hum. Mol. Genet 25, R173–R181. [PubMed: 27493026]

Choi S, Chen Z, Tang LH, Fang Y, Shin SJ, Panarelli NC, Chen YT, Li Y, Jiang X, Du YC, 2016 Bcl-xL promotes metastasis independent of its anti-apoptotic activity. Nat. Commun 7, 10384. [PubMed: 26785948]

D'Antonio M, Benaglio P, Jakubosky D, Greenwald WW, Matsui H, Donovan MKR, Li H, Smith EN, D'Antonio-Chronowska A, Frazer KA, 2018 Insights into the mutational burden of human induced pluripotent stem cells from an integrative multi-omics approach. Cell Rep. 24, 883–894. [PubMed: 30044985]

Dranoff G, 2004 Cytokines in cancer pathogenesis and cancer therapy. Nat. Rev. Cancer 4, 11–22. [PubMed: 14708024]

Elliott AM, Elliott KA, Kammesheidt A, 2010 High resolution array-CGH characterization of human stem cells using a stem cell focused microarray. Mol. Biotechnol 46, 234–242. [PubMed: 20524159]

Gore A, Li Z, Fung HL, Young JE, Agarwal S, Antosiewicz-Bourget J, Canto I, Giorgetti A, Israel MA, Kiskinis E, et al., 2011 Somatic coding mutations in human induced pluripotent stem cells. Nature 471, 63–67. [PubMed: 21368825]

Gumireddy K, Li A, Kossenkov AV, Cai KQ, Liu Q, Yan J, Xu H, Showe L, Zhang L, Huang Q, 2014 ID1 promotes breast cancer metastasis by S100A9 regulation. Mol. Cancer Res 12, 1334–1343. [PubMed: 24948111]

Halper-Stromberg E, Frelin L, Ruczinski I, Scharpf R, Jie C, Carvalho B, Hao H, Hetrick K, Jedlicka A, Dziedzic A, et al., 2011 Performance assessment of copy number microarray platforms using a spike-in experiment. Bioinformatics 27, 1052–1060. [PubMed: 21478196]

Hsu CW, Chen YC, Su HH, Huang GJ, Shu CW, Wu TT, Pan HW, 2017 Targeting TPX2 suppresses the tumorigenesis of hepatocellular carcinoma cells resulting in arrested mitotic phase progression and increased genomic instability. J. Cancer 8, 1378–1394. [PubMed: 28638452]

Hu XM, Lin T, Huang XY, Gan RH, Zhao Y, Feng Y, Ding LC, Su BH, Zheng DL, Lu YG, 2017 ID1 contributes to cell growth invasion and migration in salivary adenoid cystic carcinoma. Mol. Med. Rep 16, 8907–8915. [PubMed: 29039489]

Hussein SM, Batada NN, Vuoristo S, Ching RW, Autio R, Narva E, Ng S, Sourour M, Hamalainen R, Olsson C, et al., 2011 Copy number variation and selection during reprogramming to pluripotency. Nature 471, 58–62. [PubMed: 21368824]

International Stem Cell,Amps I, Andrews K, Anyfantis PW, Armstrong G, Avery L, Baharvand S, Baker H, Baker J, Munoz D, M.B., et al., 2011 Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. Nat. Biotechnol 29, 1132–1144. [PubMed: 22119741]

Johnson DE, O'Keefe RA, Grandis JR, 2018 Targeting the IL-6/JAK/STAT3 signalling axis in cancer. Nat. Rev. Clin. Oncol 15, 234–248. [PubMed: 29405201]

Kammers K, Taub MA, Ruczinski I, Martin J, Yanek LR, Frazee A, Gao Y, Hoyle D, Faraday N, Becker DM, et al., 2017 Integrity of induced pluripotent stem cell (iPSC) derived megakaryocytes as assessed by genetic and transcriptomic analysis. PLoS One 12, e0167794. [PubMed: 28107356]

Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, Bala S, Bensaddek D, Casale FP, Culley OJ, et al., 2017 Corrigendum: common genetic variation drives molecular heterogeneity in human iPSCs. Nature 546, 686.

Koehler BC, Scherr AL, Lorenz S, Urbanik T, Kautz N, Elssner C, Welte S, Bermejo JL, Jager D, Schulze-Bergkamen H, 2013 Beyond cell death - anti-apoptotic Bcl-2 proteins regulate migration and invasion of colorectal cancer cells in vitro. PLoS One 8, e76446. [PubMed: 24098503]

Krizhanovsky V, Lowe SW, 2009 Stem cells: the promises and perils of p53. Nature 460, 1085–1086. [PubMed: 19713919]

Laurent LC, Ulitsky I, Slavin I, Tran H, Schork A, Morey R, Lynch C, Harness JV, Lee S, Barrero MJ, et al., 2011 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 8, 106–118. [PubMed: 21211785]

Lefort N, Feyeux M, Bas C, Feraud O, Bennaceur-Griscelli A, Tachdjian G, Peschanski M, Perrier AL, 2008 Human embryonic stem cells reveal recurrent genomic instability at 20q11.21. Nat. Biotechnol 26, 1364–1366. [PubMed: 19029913]

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW, 2014 The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic. Acids. Res 42, D986–D992. [PubMed: 24174537]

Martins-Taylor K, Nisler BS, Taapken SM, Compton T, Crandall L, Montgomery KD, Lalande M, Xu RH, 2011 Recurrent copy number variations in human induced pluripotent stem cells. Nat. Biotechnol 29, 488–491. [PubMed: 21654665]

Martins-Taylor K, Xu RH, 2012 Concise review: genomic stability of human induced pluripotent stem cells. Stem Cells 30, 22–27. [PubMed: 21823210]

Mayshar Y, Ben-David U, Lavon N, Biancotti JC, Yakir B, Clark AT, Plath K, Lowry WE, Benvenisty N, 2010 Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. Cell Stem Cell 7, 521–531. [PubMed: 20887957]

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD, 2019 PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic. Acids. Res 47, D419–D426. [PubMed: 30407594]

Narva E, Autio R, Rahkonen N, Kong L, Harrison N, Kitsberg D, Borghese L, Itskovitz-Eldor J, Rasool O, Dvorak P, et al., 2010 High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. Nat. Biotechnol 28, 371–377. [PubMed: 20351689]

Nguyen HT, Geens M, Mertzanidou A, Jacobs K, Heirman C, Breckpot K, Spits C, 2014 Gain of 20q11.21 in human embryonic stem cells improves cell survival by increased expression of Bcl-xL. Mol. Hum. Reprod 20, 168–177. [PubMed: 24217388]

Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, et al., 2015 A comprehensive 1,000 genomes-based genome-wide association meta-analysis of coronary artery disease. Nat. Genet 47, 1121–1130. [PubMed: 26343387]

Oliveira PH, da Silva CL, Cabral JM, 2014 Concise review: genomic instability in human stem cells: current status and future challenges. Stem Cells 32, 2824–2832. [PubMed: 25078438]

Panopoulos AD, D'Antonio M, Benaglio P, Williams R, Hashem SI, Schuldt BM, DeBoever C, Arias AD, Garcia M, Nelson BC, et al., 2017 iPSCORE: a resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types. Stem Cell Rep. 8, 1086–1100.

Peterson SE, Loring JF, 2014 Genomic instability in pluripotent stem cells: implications for clinical applications. J. Biol. Chem 289, 4578–4584. [PubMed: 24362040]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al., 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet 81, 559–575. [PubMed: 17701901]

DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium; Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; Mexican American Type 2 Diabetes (MAT2D) Consortium; Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium,Mahajan A, Go MJ, Zhang W, Below JE, et al., 2014 Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat. Genet 46, 234–244. [PubMed: 24509480]

Robinson MD, McCarthy DJ, Smyth GK, 2010 edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. [PubMed: 19910308]

Robinson MD, Oshlack A, 2010 A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25. [PubMed: 20196867]

Roseweir AK, Powell A, Horstman SL, Inthagard J, Park JH, McMillan DC, Horgan PG, Edwards J, 2019 Src family kinases, HCK and FGR, associate with local inflammation and tumour progression in colorectal cancer. Cell. Signal 56, 15–22. [PubMed: 30684564]

Scharpf RB, Beaty TH, Schwender H, Younkin SG, Scott AF, Ruczinski I, 2012 Fast detection of de novo copy number variants from SNP arrays for case-parent trios. BMC Bioinform. 13, 330.

Scharpf RB, Irizarry RA, Ritchie ME, Carvalho B, Ruczinski I, 2011a Using the R Package crlmm for genotyping and copy number estimation. J. Stat. Softw 40, 1–32.

Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I, 2008 Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. Ann. Appl. Stat 2, 687–713. [PubMed: 19609370]

Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA, 2011b A multilevel model to address batch effects in copy number estimation using SNP arrays. Biostatistics 12, 33–50. [PubMed: 20625178]

Shi Y, Inoue H, Wu JC, Yamanaka S, 2017 Induced pluripotent stem cell technology: a decade of progress. Nat. Rev. Drug Discov 16, 115–130. [PubMed: 27980341]

Spits C, Mateizel I, Geens M, Mertzanidou A, Staessen C, Vandeskelde Y, Van der Elst J, Liebaers I, Sermon K, 2008 Recurrent chromosomal abnormalities in human embryonic stem cells. Nat. Biotechnol 26, 1361–1363. [PubMed: 19029912]

Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S, 2007 Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131 (5), 861–872. [PubMed: 18035408]

Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D, 2019 Benefits and limitations of genome-wide association studies. Nat Rev Genet.

Taniguchi K, Karin M, 2014 IL-6 and related cytokines as the critical lynchpins between inflammation and cancer. Semin. Immunol 26, 54–74. [PubMed: 24552665]

Trisciuoglio D, Tupone MG, Desideri M, Di Martile M, Gabellini C, Buglioni S, Pallocca M, Alessandrini G, D'Aguanno S, Del Bufalo D, 2017 BCL-XL overexpression promotes tumor progression-associated properties. Cell Death. Dis. 8, 3216. [PubMed: 29238043]

Van Gorp H, Lamkanfi M, 2019 The emerging roles of inflammasome-dependent cytokines in cancer development. EMBO Rep. 20.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J, 2017 10 Years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet 101, 5–22. [PubMed: 28686856]

Warren CR, Cowan CA, 2018 Humanity in a dish: population genetics with iPSCs. Trends Cell Biol. 28, 46–57. [PubMed: 29054332]

Warren CR, Jaquish CE, Cowan CA, 2017 The NextGen genetic association studies consortium: a foray into in vitro population genetics. Cell Stem Cell 20, 431–433. [PubMed: 28388427]

Wu H, Kim KJ, Mehta K, Paxia S, Sundstrom A, Anantharaman T, Kuraishy AI, Doan T, Ghosh J, Pyle AD, et al., 2008 Copy number variant analysis of human embryonic stem cells. Stem Cells 26, 1484–1489. [PubMed: 18369100]

Xu M, Knox AJ, Michaelis KA, Kiseljak-Vassiliades K, Kleinschmidt-DeMasters BK, Lillehei KO, Wierman ME, 2012 Reprimo (RPRM) is a novel tumor suppressor in pituitary tumors and regulates survival, proliferation, and tumorigenicity. Endocrinology 153, 2963–2973. [PubMed: 22562171]

Yoshihara M, Hayashizaki Y, Murakawa Y, 2017 Genomic Instability of iPSCs: challenges towards their clinical applications. Stem Cell Rev 13, 7–16.

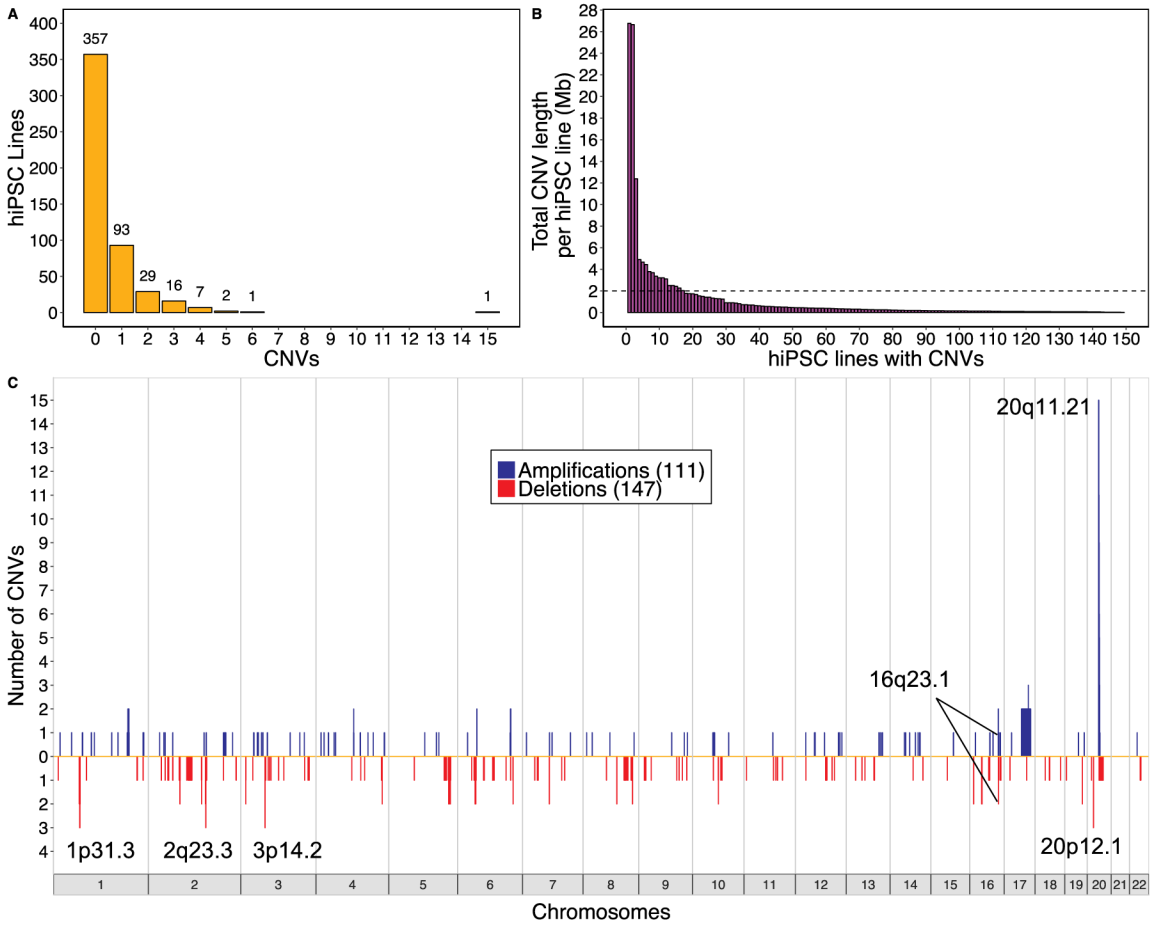

Fig. 1.

Characteristics of the Copy Number Variants in NextGen hiPSC lines. [A] Histogram showing the number of hiPSC lines with number of detected CNVs. [B] Histogram showing the cumulative size of CNVs in mega base pairs (Mb) per hiPSC line. Black dashed line shows that cumulative CNV coverage for 85% hiPSC lines was less than 2 Mb. [C] Distribution of 258 CNVs acquired by hiPSC lines across the genome. The chromosomal regions harboring cluster of CNVs are indicated.

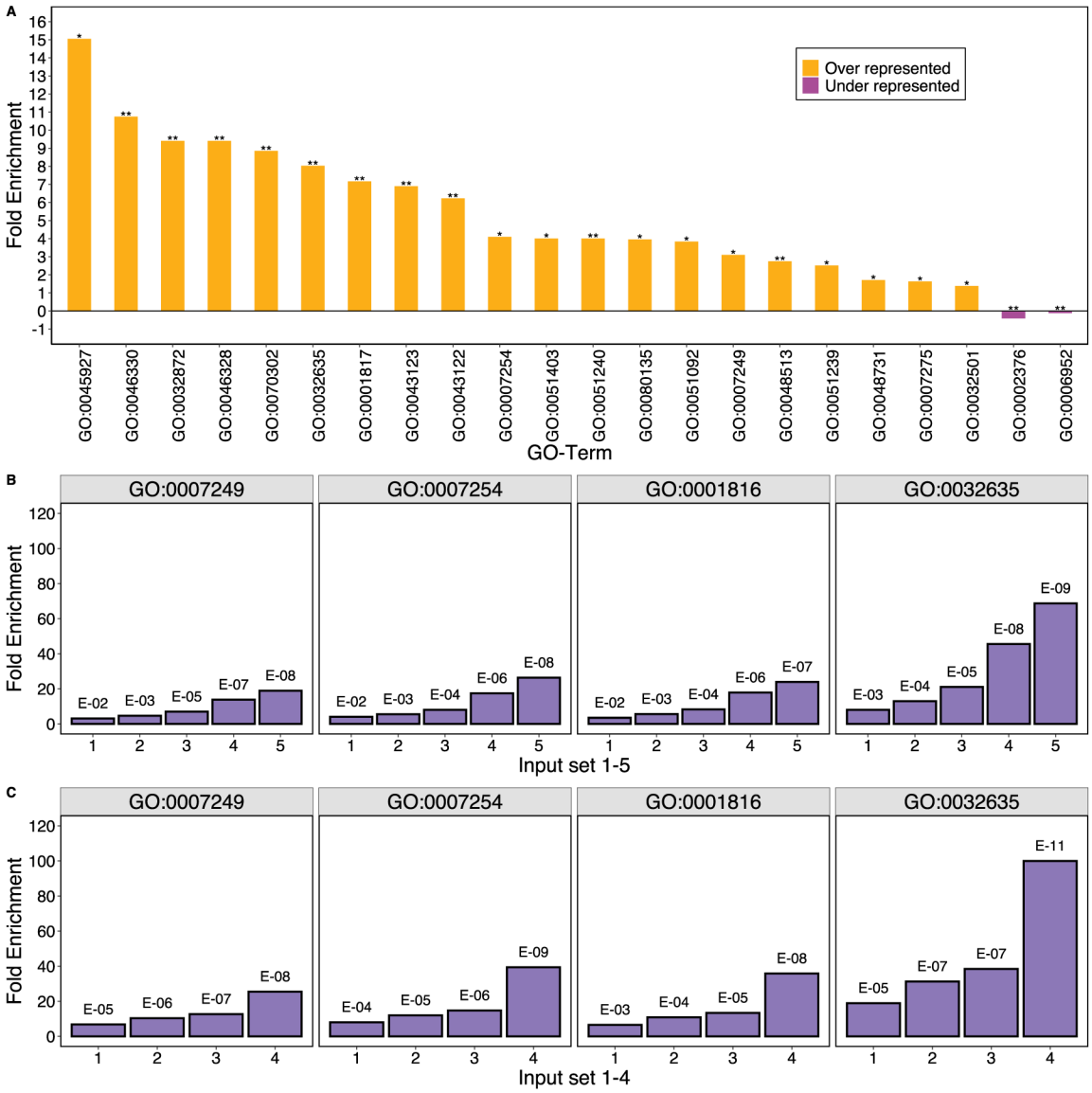

Fig. 2.

Gene set enrichment analysis for the set of 1,395 genes mapping to 258 CNVs from 149 hiPSC lines for enrichment with regards to biological processes. [A] Fold enrichment of the 22 GO Terms with an FDR of <5% ($q < 0.05$) are displayed, q values <0.01 and <0.001 are shown as "*" and "**", respectively. [B] Sequential Gene Set Enrichment of 4 biological process: I-kappaB kinase/NF-kappaB signaling cascade (GO:0007249), JNK cascade (GO:0007254), Cytokine production (GO:0001816) and IL-6 production (GO:0032635). Input set 1/2/3/4 and 5 include gene sets mapping to regions ≥1, ≥2, ≥3, ≥4 and ≥5 acquired CNVs in lines. The FE and FDR-*P* values increase markedly from input set1 through to input set5. [C] Sequential GSEA for gene sets mapping to regions of acquired Dels. Input set 1/2/3 and 4 include gene sets mapping to regions of ≥1, ≥2, ≥3 and ≥4 acquired Dels in hiPSCs. The FE and FDR-*P* values are higher for gene sets mapping to regions of Dels than all CNVs combined.

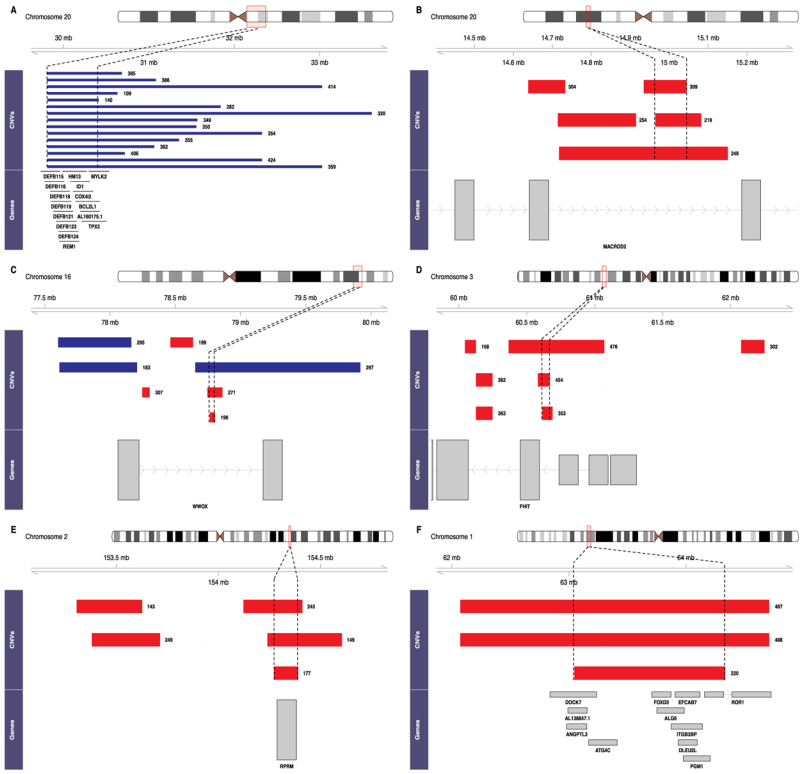

Fig. 3.

Chromosomal regions harboring clusters of CNVs. The chromosomal position of each CNV cluster is shown in a red box over the ideogram. The amplifications are represented as blue bars and deletions are shown in red bars. Dashed vertical lines delineate the core overlap regions, and in some cases nearby CNVs within the same genomic locus are also shown. The bottom panel shows the genes lying in the overlap region that could be potentially affected by the CNV in the hiPSC lines.

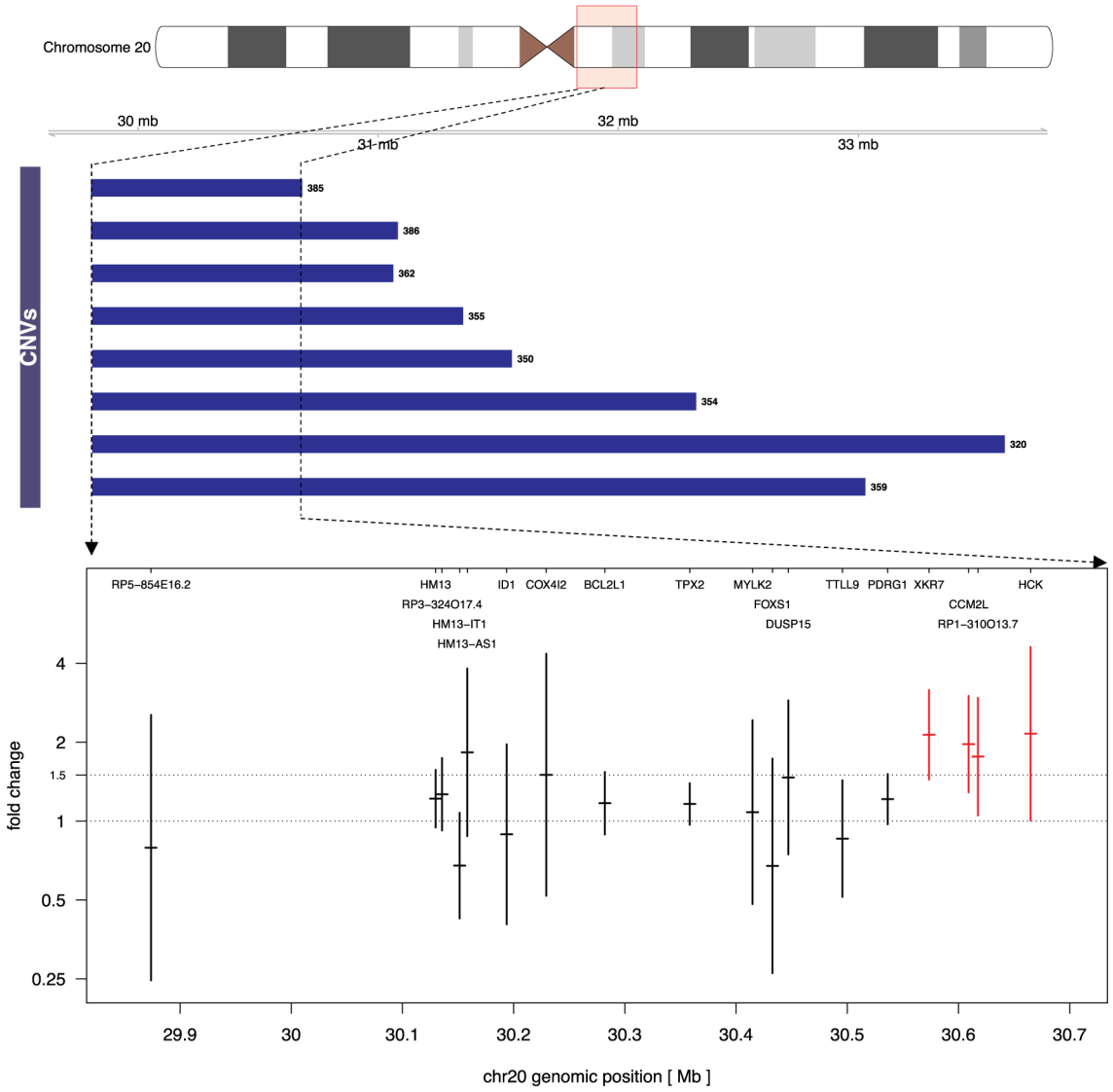

Fig. 4.

Expression levels of the transcripts mapping to the core overlap region of chr20 amplification. The chromosomal position of the amplification is shown in a red box over the ideogram. hiPSC lines harboring chr20 amplification are shown as blue bars and dashed vertical lines delineate the core overlap region. The bottom panel shows the fold changes comparing carriers to non-carries of the genes mapping to the overlap region, and their respective 95% confidence interval. Statistically significant up-regulation is observed for the genes marked in red. The expression of each gene is drawn positionally to the center of the gene.

Table 1

Genomic structural integrity of the NextGen hiPSC lines. Out of 506 hiPSC lines, 149 lines acquired CNVs during reprogramming/passaging.

Site	Donor DNA (N)	hiPSC Line (N)	Pairs (N)	hiPSC lines with acquired CNV(s) (N)	Acquired CNVs (N)
Site1	132	132	132	2	2
Site2	7	7	7	0	0
Site 3	1	1	1	1	1
Site4	171	171	171	104	203
Site5	46	83 *	83	21	26
Site6	9	9	9	2	2
Site7	49	60 *	60	7	7
Site8	0	0	0	0	0
Site9	38	43 *	43	12	17
Total	453	506	506	149	258

* >1 hiPSC line for some donor DNA.

Table 2

Details of sibling hiPSC lines. 25 hiPSC lines acquired CNVs as compared to their corresponding donor cell.

Site	Donor DNA (N)	iPSC Line (N)	Pairs (N)	hiPSC lines with acquired CNV(s) (N)	Acquired CNVs
Site5	37	74	74	17	22
Site7	11	22	22	2	2
Site9	5	10	10	6	11
Total	53	106	106	25	35