Supplementary Experimental Procedures

Samples and iPSC reprogramming

The 18 iPSC lines analyzed in this study are part of the iPSCORE resource (Panopoulos et al., 2017). 273 individuals of diverse ethnicities and ages were recruited into the iPSCORE study and whole genome sequencing of their blood or skin fibroblast DNA (254 DNA samples isolated from blood and 19 DNA samples isolated from skin fibroblasts) was conducted as previously described (DeBoever et al., 2017; Jakubosky et al., In preparation). Written informed consent was obtained from all the individuals. The iPSCORE iPSC lines were systematically derived as described in Panopoulos et al. (Panopoulos et al., 2017). Briefly, cultures of primary skin fibroblast cells were expanded for approximately 3 passages, plated at a density of 2.5 x 10^5 cells/well of 6-well plate, and infected with the Cytotune Sendai virus (Life Technologies) per manufacturer's protocol. The Sendai infected cells were maintained with 10% FBS/DMEM (Invitrogen) for Days 4-7 until the cells recovered and repopulated the well. These cells were enzymatically dissociated using TrypLE (Life Technologies) and seeded onto a 10-cm dish pre-coated with mitotically inactive-mouse embryonic fibroblasts (MEFs) at a density of 5 x 10^5 per dish and maintained with hESC medium, as previously described(Ruiz et al., 2010). Emerging iPSC colonies were manually picked after Day 21 and maintained on Matrigel (BD Corning) with mTeSR1 medium (Stem Cell Technologies) as previously described (Panopoulos et al., 2012). Clones were cultured to passage 12-16 (typically passage 12). All iPSC lines have good genomic integrity and express pluripotency markers at high levels (Panopoulos et al., 2017).

The iPSCORE resource was established as part of the Next Generation Consortium of the National Heart, Lung and Blood Institute and the iPSC lines are available through the biorepository at WiCell Research Institute (www.wicell.org; NHLBI Next Gen Collection).

Whole genome sequencing

Genomic DNA was isolated from the 18 iPSC lines (AllPrep DNA/RNA Mini Kit, Qiagen) and WGS was performed as described in DeBoever et al. (DeBoever et al., 2017). The reads were aligned to human genome hg19 with decoy sequences (Genomes Project et al., 2015) using BWA-MEM with default parameters (Li and Durbin, 2009). Duplicate reads were marked using Biobambam2 (Tischler and Leonard, 2014), and reads were sorted by genomic coordinate using Sambamba (Tarasov et al., 2015) in BAM format. The 18 iPSC WGS data (and the previously generated matched blood WGS) were high quality, having 4-20% duplicates and a minimum of 700M reads after duplicate removal (Figure S3). Inherited variants in the 18 iPSC lines were called together with the 273 iPSCORE blood and fibroblast WGS (254 blood and 19 skin fibroblasts) using GATK best practices (DeBoever et al., 2017; Jakubosky et al., In preparation; McKenna et al., 2010), including indel-realignment, base-recalibration, genotyping using HaplotypeCaller, as well as joint genotyping using GenotypeGVCFs (DePristo et al., 2011; Van der Auwera et al., 2013).

Somatic SNVs and indel calling

We used Mutect (Cibulskis et al., 2013) to detect somatic SNVs and Strelka (Saunders et al., 2012) to detect SNVs and small indels that were present in DNA isolated from the 18 iPSC lines but not in the DNA isolated from matched blood. Results from the two variant callers were intersected and only SNVs called with both methods were considered as valid somatic mutations. DNVs were identified by merging two SNVs with distance = 1 bp between each other. For indels, it was not possible to accurately estimate allelic fraction from Strelka calls, so they were not included in the analysis of subclonal variants. Additional filters were added to exclude likely false positives. First, only variants with at least 14X coverage in iPSCs, 8X in the matched blood and with an allelic fraction higher than 10% in the iPSC line were retained. Second, since a visual inspection of the variants with low allelic fraction (<30%) showed that several occur next to homonucleotide stretches, all variants with allelic fraction <30% that were next to a sequence of five or more identical nucleotides were considered as false positives and eliminated from the analysis. We identified 603 SNVs and DNVs with allelic fraction > 80%, of which 513 are hemizygous and occurred on chromosome X or Y in iPSCs derived from males, 18 were in regions where the sister chromosome contained a large CNA deletion (Table S2), and 72 of which were likely false positives given the extremely low likelihood that two independent events would cause the same mutation on both alleles.

Somatic CNA detection

To investigate the effects of somatic CNAs on gene expression, we used the population level read-depth and split-read caller Genome STRiP (svtoolkit 2.00.1611) to discover and genotype CNAs (duplications, deletions and multi-allelic CNAs) in the 18 iPSCs and their matched blood genomes (Handsaker et al., 2015). We ran Genome STRiP using the suggested settings for high coverage genomes (window size: 1000 bp, window overlap: 500 bp, minimum refined length: 500 bp, boundary precision: 100 bp, reference gap length: 1000). We considered the CNAs as somatic mutations if they: 1) were present in the iPSC line but not the matched blood genome and; 2) they were singleton, i.e. present only in one iPSC genome and not present in any of the additional 256 genomes without matched iPSCs (Jakubosky et al., In preparation).

Detection of CNAs using HumanCoreExome array

The detection of CNAs in iPSC lines in the iPSCORE resource using the HumanCoreExome BeadChip has been previously described (Panopoulos et al., 2017). Briefly, genomic DNA from the iPSCs and from their matched blood samples, was normalized to 200 ng, hybridized in pairs to Illumina HumanCoreExome arrays (Illumina), and stained per Illumina's standard protocol. BeadChips were scanned on the Illumina HiScan and processed in GenomeStudio (v 1.9.4) using the supplied cluster files for SNP calling on the HumanCoreExome arrays (average call rate 0.99, GenCall threshold 0.15). Processed SNP array data for all 18 iPSCs underwent both computerized and manual analysis for CNA detection. Computerized analysis was performed using Nexus CN (version 7.5). We used the following Nexus files and settings: Systematic Correction File: Catlg_ILM_HumanCoreExome-12v1-1_B_20140311.bed_hg19_ilum_correction.txt (as supplied by Biodiscovery Inc), Recenter Probes to Median, Analysis performed with the SNPRank Segmentation algorithm. Significance threshold 5.0 x 10^{-9}, Min Number of probes per segment = 7, High Gain 0.75, Gain 0.22, Loss -0.2, Big Loss -1.1. CNAs shorter than 100 kb were removed. All Nexus calls underwent systematic manual inspection of B allele frequencies and log R rations. All Nexus calls that were not visually consistent with a CNA based on B allele frequencies and log R ratios were removed. Manual inspection of the entire genome was also performed for each iPSC line and its associated blood sample.

Somatic mutations in adult stem cells and cancer

Somatic mutations in adult stem cells (ASCs) from three different tissue types (liver, small intestine and colon) derived from 45 subjects were obtained from Blokzijl et al. (Blokzijl et al., 2016). Since the number of somatic mutations in ASCs depends on the subject's age, we divided ASCs in three categories: 1) ASCs from 11 young subjects (< 15 years old); ASCs from 19 subjects between 16 and 60 years old; and 3) ASCs from 15 subjects older than 60 years.

The number of somatic mutations in each subject for each tumor were obtained from three collections: 1) 507 tumors (four tumor types) from Alexandrov et al. (Alexandrov et al., 2013a); 2) 25 melanomas from Berger et al. (Berger et al., 2012); and 3) 3,011 tumors (11 tumor types) from the International Cancer Genomics Consortium (ICGC) (Alexandrov et al., 2013a; International Cancer Genome et al., 2010; Nik-Zainal et al., 2016). Tumors are divided into mutagens, adult solid tumors, liquid and pediatric following the same criteria as in Vogelstein et al.(Vogelstein et al., 2013).

Comparing the mutational landscape of iPSCs with 30 cancer mutational signatures

Somatic SNVs were divided into 96 substitution classes defined by the substitution type (C>A, C>G, C>T, T>A, T>C or T>G) and the sequence context immediately upstream and downstream of the mutated base. The total possible number of 96 substitution classes is given by the number of substitution types (6), multiplied by four possible upstream bases and four possible downstream bases. The distribution of mutations across the 96 substitution classes was compared with their associated distributions in 30 different mutational signatures included in COSMIC (Alexandrov et al., 2013a; Alexandrov et al., 2013b) (http://cancer.sanger.ac.uk/cosmic/signatures). Correlation was calculated between the mutational landscape of each of the 18 iPSC lines and each of the 30 mutational signatures.

Annotation of functional effects of somatic SNVs and DNVs

Mutations were annotated using SnpEff v. 4 (Cingolani et al., 2012). Gene and transcript data were derived from Gencode v. 19(Harrow et al., 2012), while transcription factor binding sites were derived from the default annotations included in SnpEff. Mutations were grouped according to their impact, as defined by SnpEff (Table S6), and lists of mutated genes group for each impact category were retrieved from SnpEff annotations (Table S7).

Cancer gene enrichment analysis

We downloaded a list of 248 known cancer genes from the Cancer Gene Census (frozen at April 21 2016), including 141 tumor suppressors and 110 oncogenes (Forbes et al., 2015; Futreal et al., 2004). Three genes are considered both as oncogenes and tumor suppressors and are included in both lists. These gene lists were intersected with the mutated genes in Table S7. GOseq v. 1.24.0 (Young et al., 2010) was used to examine if cancer genes were enriched for being mutated in the iPSC lines.

Detection of associations between mutations and chromatin states

The whole genome was divided into 200-bp bins and the bins with similar sequence characteristics (and thus similar mutation rates) were clustered together using four covariates (D'Antonio et al., 2017; Lawrence et al., 2013): 1) DNA replication timing derived from the ENCODE wgEncodeUwRepliSeq track on the UCSC Genome browser (Hansen et al., 2010; Thurman et al., 2007); 2) open vs. closed chromatin status as measured by Hi-C mapping (Lawrence et al., 2013; Lieberman-Aiden et al., 2009) in iPSCs (see ”Hi-C data processing” section below); 3) GC content; and 4) gene density, measured as the number of base-pairs that overlap Gencode V. 19 genes (Harrow et al., 2012) in the 500 kb surrounding each bin. The values of all covariates were normalized to have mean = 0 and standard deviation = 1. Normalized covariate values were used to cluster all 200-bp bins in each chromosome using k-means clustering, where k was selected to have on average 200 bins in each cluster. A BED file was created for each cluster. Each somatic variant (including SNVs and small indels) was assigned to its 200 bp bin and its position was permuted 100 times within all the sequences in its cluster. The number of variants associated with each chromatin mark was determined in all permutations, then mean and standard deviation were calculated. Enrichment for each mutation class in each chromatin state was determined on a per tissue basis as Z-scores against the 100 permutations by subtracting from the observed value its corresponding mean across the 100 permutations and dividing by the standard deviation.

Hi-C data processing for clustering of 200-bp bins

To use open vs. closed chromatin status as a covariate for clustering of the 200-bp bins, we processed Hi-C data as described by the MutSigCV method (Lawrence et al., 2013). Hi-C experiments were performed as described (Li et al., In preparation) in iPSCs from 7 individuals in Family 2 of the iPSCORE resource (Panopoulos et al., 2017). Sequencing data from the Hi-C experiments for all 7 individuals were combined for analysis and a total of ~3 billion raw reads were obtained. We applied the Juicer pipeline (Rao et al., 2014) to align and quality check (QC) the read pairs. After QC, ~1 billion intra-chromosome read pairs were kept for the analysis, resulting in map resolutions of 2kb, defined as > 80% of the bins with > 1000 contacts. Contact domains were determined using the Arrowhead algorithm in Juicer at 5 kb-resolution (Durand et al., 2016; Rao et al., 2014). Chromatin contact intensity for each contact domain was used as a measure of open vs. closed chromatin status, as defined by the MutSigCV method (Lawrence et al., 2013; Lieberman-Aiden et al., 2009).

RNA-seq data processing and gene expression analysis

To analyze the effects of somatic mutations on gene expression in the 18 iPSC lines, we relied on our previously published collection of RNA-seq data in iPSC lines at passage 12 (DeBoever et al., 2017). Briefly, we determined total RNA quality using an Agilent Tapestation, and samples determined to have an RNA Integrity Number (RIN) of 7 or greater were used to generate RNA libraries using Illumina's TruSeq Stranded Total RNA Sample Prep Kit. RNA libraries were multiplexed and sequenced with 125 bp paired end reads (PE100) to a depth of approximately 25 million reads per sample on an Illumina HiSeq2500. We aligned RNA-seq reads to the human genome (hg19) with STAR 2.4.0h (outFilterMultimapNmax 20, outFilterMismatchNmax 999, outFilterMismatchNoverLmax 0.04, outFilterIntronMotifs RemoveNoncanonicalUnannotated, outSJfilterOverhangMin 6 6 6 6, seedSearchStartLmax 20, alignSJDBoverhangMin 1) using a splice junction database constructed from Gencode v19 (Dobin et al., 2013; Harrow et al., 2012). Reads overlapping genes were counted using HTSeq-count (-s reverse -a 0 -t exon -i gene_id -m union) (Anders et al., 2015; Anders et al., 2012). For each gene, TPM values were normalized using the calcNormFactors function in the preprocessCore package in R, which resulted in having all genes with mean expression = 0 and standard deviation = 1. For 18,284 genes with mean TPM >0, we calculated the distance between their TSS and their closest upstream mutation. For each gene, we determine the normalized expression value in the mutated sample, expressed in Z-scores (defined as the number of standard deviations from the mean). To determine enrichment between different mutational classes, we calculated the fraction of genes with Z-score >2 or < -2 and compared it with clonal SNVs for all mutations within 500 kb from the TSS.

Somatic CNA impact on gene expression

To assess overlap between CNAs and genes, first, we intersected the coordinates of each CNA with the coordinates of GENCODE genes using Bedtools intersect and found 1,325 genes that overlapped the 255 CNAs, of which 1,049 overlapped the 50 Mb chromosome 1q duplication in iPSCORE_3_4 (Table S10). For each gene that overlapped a CNA, we compared its normalized expression level (calculated for each gene by subtracting its mean expression level and dividing by the standard deviation across all samples) in the iPSC line that harbored the CNA with respect to all other lines.

To analyze the large duplication on chromosome 1, allelic-bias was determined using WASP (van de Geijn et al., 2015) and allele specific expression was calculated using MBASED (Mayba et al., 2014) as described in DeBoever et al. (DeBoever et al., 2017) .

Assessing clonal evolution of iPSCs using RNA-seq

To assess evolution of somatic mutations in iPSCs, we used 71 independent RNA-seq data sets from three other iPSCORE studies (Benaglio et al., Submitted; DeBoever et al., 2017; Panopoulos et al., 2017). The RNA-seq data were generated from iPSCs at varying passages ~12-25, and iPSC-derived cardiomyocytes (iPSC-CMs) at differentiation day 15 for four subjects (iPSCORE_2_1, iPSCORE_2_2, iPSCORE_2_3 and iPSCORE_2_9) and a time course analysis of iPSC-CM differentiation (days 2, 5 and 9) in three subjects (iPSCORE_2_2, iPSCORE_2_3 and iPSCORE_2_9) (Table S11). In order to examine the evolution of subclonal mutations at different iPSC passages and during differentiation, we used the RNA-seq approach recently developed to study subclonal somatic mutations in embryonic stem cells (ESCs) (Merkle et al., 2017), on the basis of the observation that the allelic fraction of coding mutations from RNA-seq is highly correlated with the allelic fraction determined with WGS for the same mutations. For each of the 71 RNA-seq BAM files associated with the four subjects, we used Samtools mpileup (Li et al., 2009) at the positions of all somatic mutations to derive read count and allelic fraction (determined as the ration between the number of reads with non-reference nucleotides divided by the total number of reads at a mutated position). Only 146 mutations with read count > 10 were considered. Cochran-Armitage Test for trend (R package DescTools V. 0.99.23, https://CRAN.R-project.org/package=DescTools) was performed on reference and alternative read counts for each mutation to determine whether allelic frequency changed over time. Heatmaps in Figure 7 were made using the pheatmap package in R.

Supplementary Figures

Figure S1: Description of the individuals from which the whole-genome sequenced iPSCs were derived. Related to Figure 1.

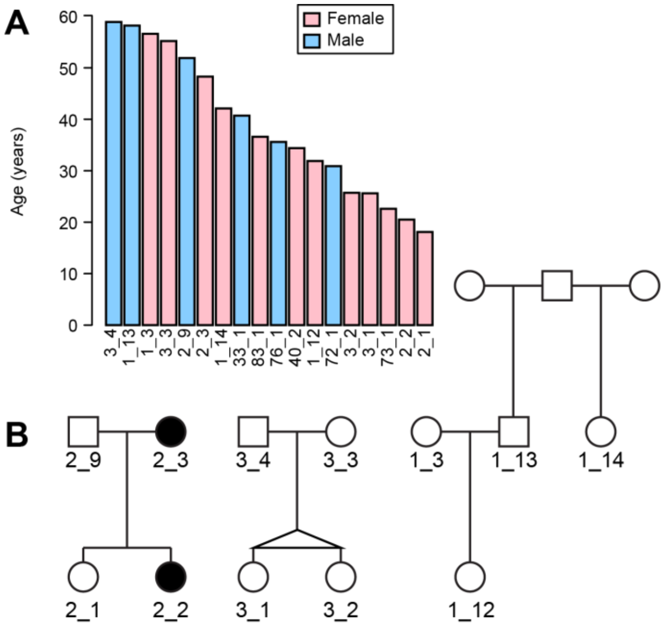

Figure S1: Description of the individuals from which the whole-genome sequenced iPSCs were derived. Related to Figure 1. (**A**) Age and gender of each subject. (**B**) Pedigrees showing the familial relationships among twelve of the 18 analyzed subjects. For the 18 subjects, the subject ID and family ID are shown (for example 3_4 = family 3, subject 4). The two siblings in family 3 are identical twins. Two subjects with Long QT Syndrome Type II (LQT2) in family 2 are colored in black.

Figure S2: Mutational signatures in the 18 iPSCs. Related to Figure 2

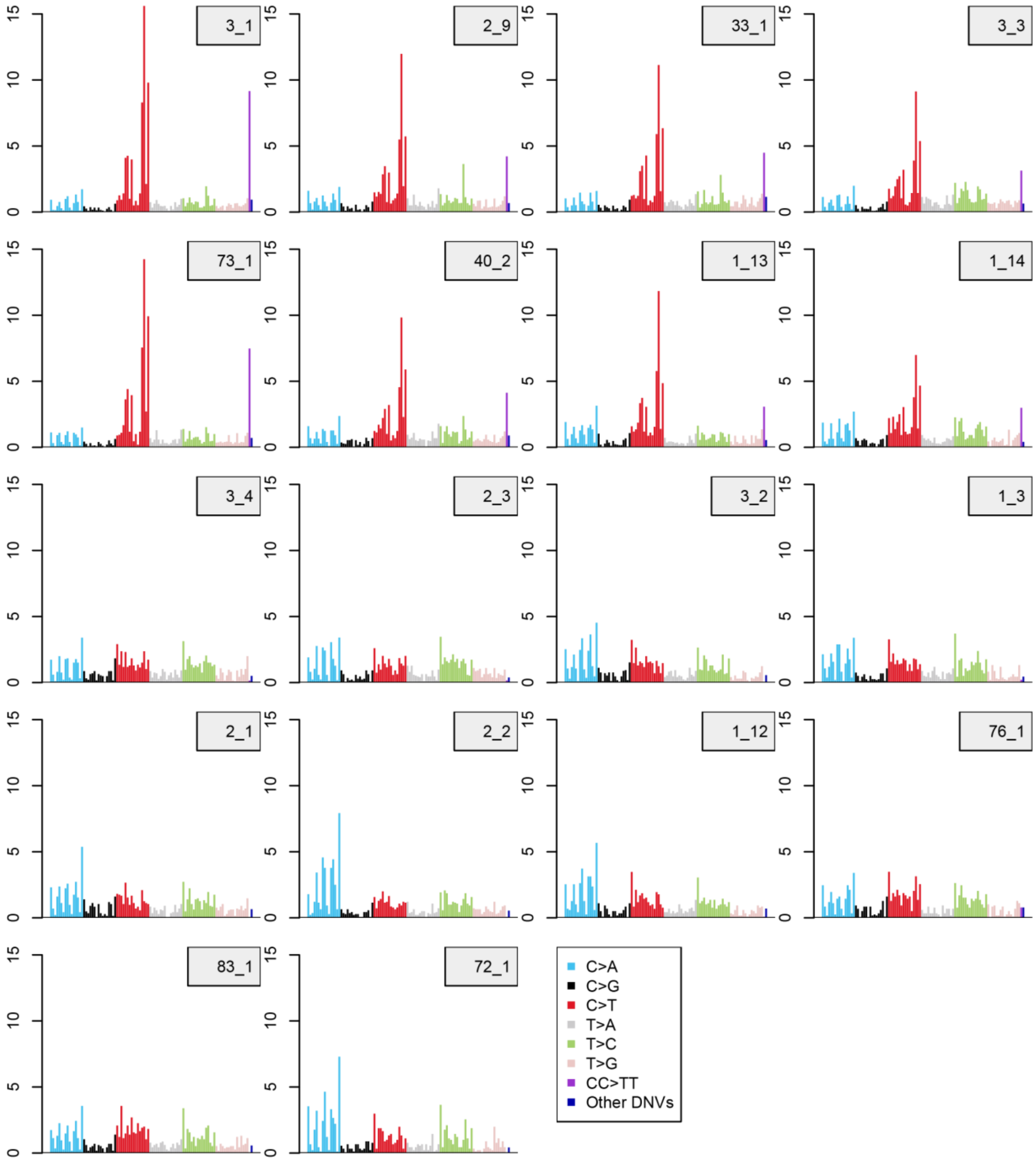

Figure S2: Mutational signatures in the 18 iPSC lines. Related to Figure 2. The percentage of mutations in each of the 96 substitution classes defined by the substitution type (C>A, C>G, C>T, T>A, T>C or T>G) and the sequence context immediately upstream and downstream of the mutated base. The X axis represents the somatic substitution type, divided by sequence context (one base pair upstream and one downstream) (Alexandrov et al., 2013b). The Y axis represents the percentage of all mutations associated with each substitution type. The last two columns represent CC>TT DNVs and all other DNVs (depicted in purple and dark blue, respectively).

Figure S3: Whole genome sequencing quality. Related to Figure 1.

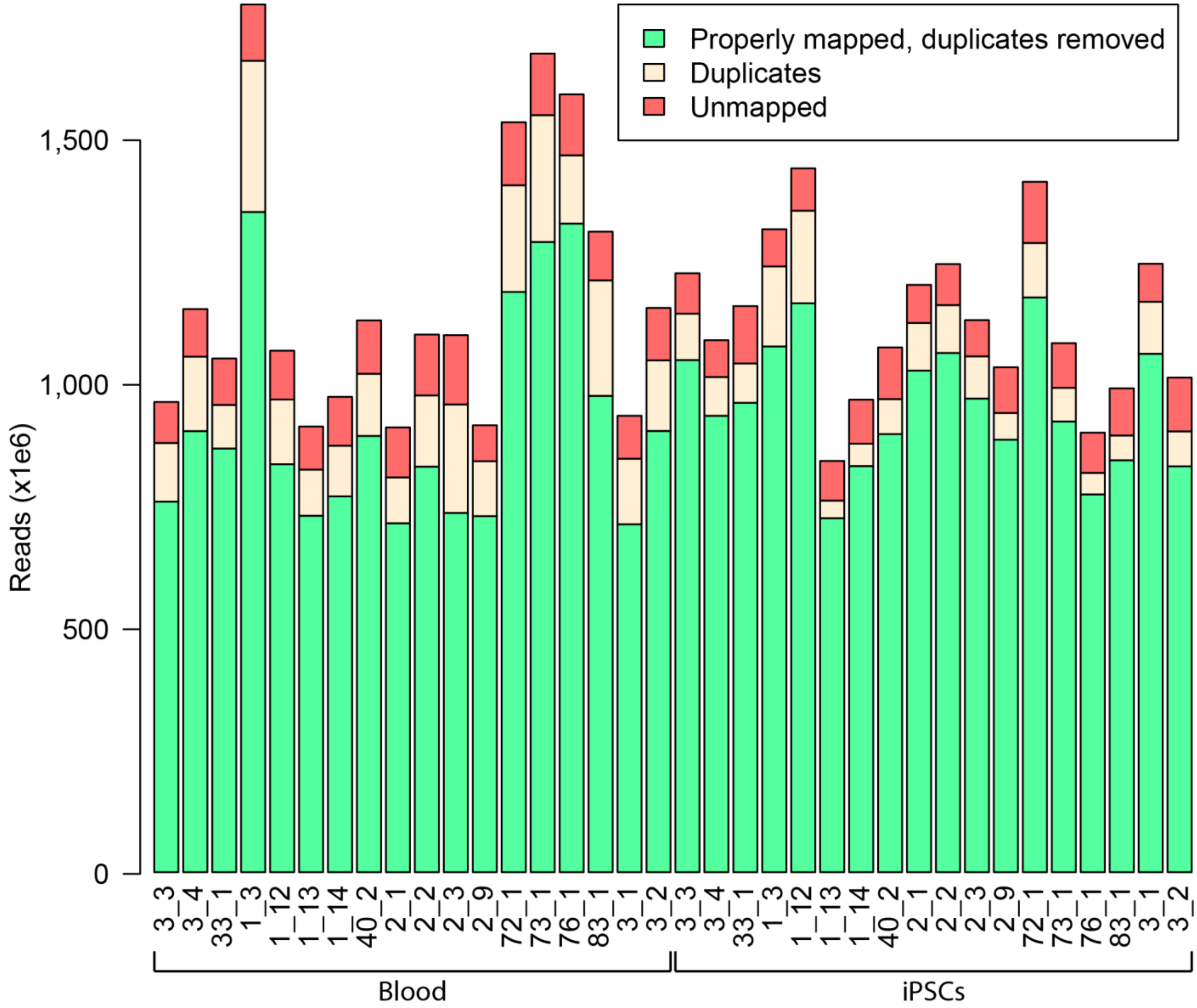

Figure S3: Whole genome sequencing quality. Related to Figure 1. The number of reads that result from WGS of the 36 samples: 18 from blood (left) and 18 from iPSCs (right) is shown. The number of high quality properly mapped with duplicates removed are shown in green.

Supplementary Tables

Table S1: Description of all subjects in the study. Related to Figure 1.

Subject UUID, subject ID, family ID and family UUID are provided (Panopoulos et al., 2017). Self-reported race was directly obtained by the subject or physician (denoted by asterisk) and was translated into one of seven groups (African American, Asian, European, Hispanic, Indian, Middle Eastern, and Multiple ethnicities reported) and defined as recorded ethnicity grouping (column G).

Table S2: Variant counts in iPSCs. Related to Figure 1.

Subject UUID, subject ID and iPSC ID are the same as in Panopoulos et al. (Panopoulos et al., 2017). The 18 iPSC lines are available through WiCell Research Institute and the corresponding WiCell iPSC ID is provided. The whole genome sequencing UUID for blood is the same as in DeBoever et al.(DeBoever et al., 2017). For some subjects, multiple iPSC clones were derived, so we indicate both the clone number and the passage at which it was sequenced.

Shown is the number of somatic variants per iPSC line. SNVs are divided in three classes, based on their allelic frequency: 1) clonal SNVs have allelic frequency between 30% and 80%; 2) subclonal SNVs have allelic frequency <30%; and 3) homozygous SNVs have allelic frequency >80%. Homozygous SNVs were subsequently divided into three categories, in order to identify possible false positives: 1) SNVs overlapping deletions are likely true positives and are seen as “homozygous” because one copy is lost; 2) SNVs not overlapping deletions in autosomes are likely false positives, and account for 72 mutations overall (0.16% of all SNVs); and 3) SNVs on sex chromosomes in males are likely true positives. Row 22 (the last row) shows the total number of variants of the given variant type over all 18 iPSC lines. Column T shows the total number of point and indel mutations (SNVs + DNVs +indels).

Table S3: Description of somatic CNAs. Related to Figure 1.

Shown are the ID of the iPSC line harboring the somatic CNA, the CNA coordinates, type and length.

Table S4: Agreement between HumanCoreExome array and Genome STRiP calls. Related to Figure 1.

Subject ID	CNA Type	Chromosome	CNA coordinates			
			SNP array		Genome STRiP	
			Start	End	Start	End
iPSCORE_1_13	Loss	chr4	143,094,811	143,212,870	143,109,554	143,130,189
iPSCORE_1_3	Loss	chr16	78,744,299	78,863,714		
iPSCORE_2_1	Loss	chr5	128,294,900	128,536,973	128,294,673	128,536,788
iPSCORE_2_9	Allelic Imbalance*	chr3	30,048,710	30,249,375	30,061,327	30,244,478
iPSCORE_3_3	Gain	chr10	135,656	1,948,065	92,349	1,961,954
iPSCORE_3_3	Trisomy Xp	chrX	181,779	54,864,853		

iPSCORE_3_4	Gain	chr1	201,105,488	249,212,430	201,175,747	249,195,464
iPSCORE_3_4	Loss	chr12	80,681,608	84,820,483	81,149,622	82,327,509
iPSCORE_3_4	Loss	chr2	156,854,409	156,965,011		
iPSCORE_40_2	Gain	chr15	32,912,411	33,188,632	32,899,012	33,152,083
iPSCORE_72_1	Loss	chr12	106,267,942	106,561,353	106,266,196	106,560,050
iPSCORE_72_1	Loss	chr14	89,839,550	90,037,290	89,853,329	89,981,640
iPSCORE_76_1	Loss	chr4	183,170,251	183,620,839	183,178,416	183,619,812
iPSCORE_83_1	Loss	chr16	6,488,959	6,817,144	6,484,637	6,799,804
iPSCORE_83_1	Loss	chr2	154,272,758	154,403,750	154,272,514	154,344,735
iPSCORE_83_1	Loss	chr8	121,199,648	121,335,994	121,200,297	121,341,337
iPSCORE_84_1	Loss	chr13	39,737,588	39,997,264		

* Genome STRiP defines this CNA as a duplication

Shown are the 17 CNAs detected by SNP arrays (Panopoulos et al., 2017) and their overlap with Genome STRiP calls.

Table S5: DNV subtypes. Related to Figure 2.

Shown are the number of DNVs divided by subtype for each of the 18 iPSC lines.

Table S6: Associations between mutations and impact. Related to Figure 4.

Mutation type	Impact	Clonal SNVs	Clonal C>T SNVs	Clonal CC>TT DNVs	Subclonal SNVs	Total
Intergenic	No	13947	10061	1136	2836	45108
Intronic	No	8857	5897	644	1730	
3' UTRs	Low	171	105	8	50	768
5' UTRs	Low	52	32	8	15	
TF binding sites	Low	9	21	4	5	
Non-coding exons	Low	77	66	12	22	
Synonymous	Low	29	44	8	30	
Splice sites	Moderate	34	18	5	11	357
Missense	Moderate	132	108	12	37	
Nonsense	High	10	6	3	2	21

Shown are the number of somatic SNVs and DNVs distributed by impact grouping and the four classes of somatic mutations: 1) clonal SNVs (C>G, C>A, T>G, T>C, T>A); 2) clonal C>T SNVs; 3) clonal CC>TT DNVs; and 4) subclonal SNVs. Colors correspond to the impact of mutations, as defined by SnpEff (Cingolani et al., 2012).

Table S7: Genes affected by somatic mutations. Related to Figure 4.

The table shows the 504 genes carrying a mutation, the classification of that mutation (clonal SNVs, clonal C>T SNVs, clonal CC>TT DNVs or subclonal SNVs, in column C, Figure 4), the impact of that mutation and the subjects that carry the mutation (Table S6).

Table S8: Enrichment of the four classes of point and indel mutations across 15 chromatin states. Related to Figure 5.

For each of the four mutational classes (clonal SNVs, clonal C>T SNVs, clonal CC>TT DNVs and subclonal SNVs, in column C), the table shows the enrichment, calculated as Z-scores, compared with 100 random permutations for the 15 chromatin states in each of the 127 human tissues (EID number given in column A) (Figure 5). Columns A-C show the tissue ID as derived from Roadmap, whether or not it is one of the 22 stem cell tissues, and the mutational class, respectively. The other columns show the Z-scores for each of the 15 chromatin states.

Table S9: Associations between gene expression and upstream mutations. Related to Figure 6.

The table shows 17,874 genes (including Gencode V.19 gene ID, symbol, gene type, coordinates and strand), their closest mutation up to 500 kb (including the sample harboring the mutation, position, mutation class and distance from the TSS) and gene expression levels (normalized expression in the mutated sample, TPM in the mutated sample and average TPM across all 215 iPSC lines).

Table S10: Overlap between genes and CNAs. Related to Figure 6.

Overlap between Gencode V.19 genes and CNAs was determined using BedTools intersect. Column G shows whether each gene is included in a deletion, duplication or the large duplication on chromosome 1. “Overlap” refers to whether a CNA fully encompasses a gene sequence (complete overlap) or only partially overlaps a gene.

Table S11: RNA-seq data to assess clonal evolution of iPSCs. Related to Figure 7.

The table shows the RNA-seq UUID for the 71 RNA-seq datasets generated for the iPSC lines from four subjects (iPSCORE_2_1, iPSCORE_2_2, iPSCORE_2_3 and iPSCORE_2_9). For each RNA-seq sample, cell type (either iPSC or iPSC-CM), passage (for iPSCs) and day of differentiation (for iPSC-CMs) is shown. These data have also been used in other iPSCORE studies (Benaglio et al., Submitted; DeBoever et al., 2017; Panopoulos et al., 2017).

References

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L.*, et al.* (2013a). Signatures of mutational processes in human cancer. Nature *500*, 415-421.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013b). Deciphering signatures of mutational processes operative in human cancer. Cell reports *3*, 246-259.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome research *22*, 2008-2017.

Benaglio, P., DeBoever, C., Li, H., D'Antonio, M., Drees, F., Singhal, S., Aguirre, A., Matsui, H., D'Antonio-Chronowska, A., Smith, E.N.*, et al.* (Submitted). iPSC-derived cardiomyocytes are a powerful model system for functionally characterizing human regulatory variants.

Berger, M.F., Hodis, E., Heffernan, T.P., Deribe, Y.L., Lawrence, M.S., Protopopov, A., Ivanova, E., Watson, I.R., Nickerson, E., Ghosh, P.*, et al.* (2012). Melanoma genome sequencing reveals frequent PREX2 mutations. Nature *485*, 502-506.

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P.*, et al.* (2016). Tissue-specific mutation accumulation in human adult stem cells during life. Nature.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology *31*, 213-219.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80-92.

D'Antonio, M., Weghorn, D., D'Antonio-Chronowska, A., Coulet, F., Olson, K.M., DeBoever, C., Drees, F., Arias, A., Alakus, H., Richardson, A.L.*, et al.* (2017). Identifying DNase I hypersensitive sites as driver distal regulatory elements in breast cancer. Nature communications *8*, 436.

DeBoever, C., Li, H., Jakubosky, D., Benaglio, P., Reyna, J., Olson, K.M., Huang, H., Biggs, W., Sandoval, E., D'Antonio, M.*, et al.* (2017). Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. Cell stem cell *20*, 533-546 e537.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M.*, et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics *43*, 491-498.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell systems *3*, 95-98.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S.*, et al.* (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic acids research *43*, D805-811.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nature reviews Cancer *4*, 177-183.

Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A.*, et al.* (2015). A global reference for human genetic variation. Nature *526*, 68-74.

Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M., and McCarroll, S.A. (2015). Large multiallelic copy number variations in humans. Nature genetics *47*, 296-303.

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proceedings of the National Academy of Sciences of the United States of America *107*, 139-144.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S.*, et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome research *22*, 1760-1774.

International Cancer Genome, C., Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I.*, et al.* (2010). International network of cancer genome projects. Nature *464*, 993-998.

Jakubosky, D., D’Antonio, M., DeBoever, C., Li, H., Matsui, H., Smith, E.N., Nariai, N., and Frazer, K.A. (In preparation). Discovery and functional characterization of high resolution copy number variants by deep whole-genome sequencing of 273 individuals

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A.*, et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Greenwald, W.W., Benaglio, P., Schmitt, A., Qiu, Y., Ren, B., Selvaraj, S., Jakubosky, D., D'Antonio, M., D'Antonio-Chronowska, A.*, et al.* (In preparation). Chromatin loop variability between cell types and genetic haplotypes in humans.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O.*, et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.

Mayba, O., Gilbert, H.N., Liu, J., Haverty, P.M., Jhunjhunwala, S., Jiang, Z., Watanabe, C., and Zhang, Z. (2014). MBASED: allele-specific expression detection in cancer tissues and cell lines. Genome biology *15*, 405.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.

Merkle, F.T., Ghosh, S., Kamitaki, N., Mitchell, J., Avior, Y., Mello, C., Kashin, S., Mekhoubad, S., Ilic, D., Charlton, M.*, et al.* (2017). Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. Nature.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C.*, et al.* (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47-54.

Panopoulos, A.D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S.I., Schuldt, B.M., DeBoever, C., Arias, A.D., Garcia, M., Nelson, B.C.*, et al.* (2017). iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. Stem cell reports *8*, 1086-1100.

Panopoulos, A.D., Yanes, O., Ruiz, S., Kida, Y.S., Diep, D., Tautenhahn, R., Herrerias, A., Batchelder, E.M., Plongthongkum, N., Lutz, M.*, et al.* (2012). The metabolome of induced pluripotent stem cells reveals metabolic changes occurring in somatic cell reprogramming. Cell Res *22*, 168-177.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S.*, et al.* (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

Ruiz, S., Brennand, K., Panopoulos, A.D., Herrerias, A., Gage, F.H., and Izpisua-Belmonte, J.C. (2010). High-efficient generation of induced pluripotent stem cells from human astrocytes. PLoS One *5*, e15526.

Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811-1817.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics *31*, 2032-2034.

Thurman, R.E., Day, N., Noble, W.S., and Stamatoyannopoulos, J.A. (2007). Identification of higher-order functional domains in the human ENCODE regions. Genome research *17*, 917-927.

Tischler, G., and Leonard, S. (2014). biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol Med *2014*, 13.

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nature methods *12*, 1061-1063.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J.*, et al.* (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics *43*, 11 10 11-33.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546-1558.

Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome biology *11*, R14.