Contents

1. Dataset, sample collection and data access

This paper describes data and analyses from a total of 711 iPSC lines and 301 fibroblast lines from 301 unrelated healthy donors. Lines selected for inclusion in this publication were derived from fibroblasts and reprogrammed with Sendai virus. Based on the analysis of Tier 1 assay data, one to two lines from each donor were selected for banking as part of the publicly available HipSci resource. Description of the study sample can be found in **Supplementary Table 1**.

Breakdown of iPSC lines by phenotypic assay.

	hDF	hiPSC Tier 1 lines (donors)	hiPSC Tier 2 lines (donors)
Selected for banking	301	-	417 (273)
Genotyping array /expression array	301	711 (301)	-
Immunochemistry imaging of pluripotency / differentiation markers	-	393 (169)	-
RNA-seq	-	-	239 (166)
Methylation array	-	-	27 (14)
Proteomics	-	-	16 (9)
Cellular morphology	-	-	24 (12)

Samples were collected from consented research volunteer recruited from the NIHR Cambridge BioResource through (http://www.cambridgebioresource.org.uk). Initially, 250 normal samples were collected under ethics for iPSC derivation (REC Ref: 09/H0304/77, V2 04/01/2013), which require managed data access for all genetically identifying data, including genotypes, sequence and microarray data (hereon 'managed access samples'). In parallel we obtained new ethics approval for a revised consent (REC Ref: 09/H0304/77, V3 15/03/2013), under which all data, except from the Y chromosome from males, can be made openly available (Y chromosome data can be used to de-identify men by surname matching), and all samples since October 2013 have been collected with this revised consent (hereon 'open access samples'). The majority of samples were European. In the following we denote lines from these two consents schemes managed access and open access lines respectively.

The assay data used in this publication are listed in the Biostudies archive (https://www.ebi.ac.uk/biostudies/studies) with accession identifier S-BSMS5, which is

specific for this publication, and which represents a subset of HipSci's complete dataset. The biostudies submission includes archive accession identifiers for obtaining the data.

All data can be accessed via the HipSci data portal (http://www.hipsci.org), which references to EMBL-EBI archives that are used to store the HipSci data, according to the assay type and data type, including the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena), European Genome-phenome Archive (EGA; www.ebi.ac.uk/ega/home), European Variation Archive (EVA; http://www.ebi.ac.uk/eva), ArrayExpress (http://www.ebi.ac.uk/arrayexpress) and Proteomics Identifications Database (PRIDE; http://www.ebi.ac.uk/pride). Data types from specialized assays for which none of the existing archives are appropriate are available from the HipSci FTP site (ftp://ftp.hipsci.ebi.ac.uk/vol1/ftp). All assay data generated by the HipSci project are publicly available, including for cell lines ultimately not selected for banking. Genetic and sequencing data of cell lines with managed access are securely archived in the EGA, and access to the data is permitted to bona fide researchers who must accept a data access agreement. The HipSci website (http://www.hipsci.org) has full details of all publicly available data and instructions for researchers to register for access to data in EGA. Intermediate result files for this study, such as processed gene expression levels, can be found at: ftp://ftp.hipsci.ebi.ac.uk/vol1/ftp/data.

Managed access data from all assays are accessible via EGA under the study EGAS00001001465. The open access genotyping array data and RNA-seq data are available from ENA under the studies PRJEB11752 and PRJEB7388, respectively. The open access gene expression array data are available in the ArrayExpress database under accession number E-MTAB-4057. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifiers PXD003903 and PXD005506.

For most up-to-date information of available lines and data see http://www.hipsci.org. HipSci has generated iPSC lines using alternative reprogramming methods as well as from disease samples, which were not considered in the present publication. As of April 2017, 710 iPSC lines have passed QC and are selected for banking as part of a publicly available resource. The standard operating procedures for the HipSci iPSC generation pipeline and differentiation experiments can be found at http://www.hipsci.org/cells/sop and they are also accessible via the FTP site.

2. Reference datasets and software used for the analysis

Software links

LIMIX (https://github.com/PMBio/limix).

BCFtools (https://github.com/samtools/bcftools).

Leafcutter (https://github.com/davidaknowles/leafcutter).

Haplotype reference imputation panel

Available from the Sanger Imputation Service:

https://imputation.sanger.ac.uk/.

Reference assembly for microarray probe remapping

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz

Proteomics quantification references

Original reference downloads from:

ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000005640_9606.fasta.gz

ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000005640_9606_additional.fasta.gz

Final reference database:

http://www.ebi.ac.uk/pride/archive/projects/PXD003903/UP000005640_9606.sp.20150427.fastahttp://www.ebi.ac.uk/pride/archive/projects/PXD003903/UP000005640_9606.sp.20150427.fasta

RNA-seq quantification assembly and annotation

GRCh37 primary assembly with the human decoy sequence 37d5:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz

Exon-intron junctions derived from Gencode v19:

ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz

Chromatin annotations

Roadmap segmentations:

http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/

Encode annotations:

https://www.encodeproject.org/data/annotations

3. Statistics of cell line and assay generation

Samples for the project were collected over a period of 42 months between February 2013 and July 2015 during which we received a total of 796 skin punch biopsies from healthy, unrelated research volunteers, the vast majority of which were of Northern European ancestry recruited through the NIHR Cambridge BioResource (http://www.cambridgebioresource.org.uk). We successfully cultured fibroblast outgrowths from skin explants of 701 individuals (88.1%) of which, by the time of the current study, 600 have been taken forward to the reprogramming stage. Using a Sendai viral vector system [1] on a feeder layer of mouse embryonic fibroblasts we successfully produced pluripotent colonies from 427 (71.2%) donors 34 days post transduction on average. Of the 427 successfully reprogrammed samples, 301 were sufficiently advanced in our experimental pipeline to be included in the current study.

We established multiple independent lines from most donors (82% of donors had >1 line, 50% had three or more lines) resulting in a total of 711 iPSC lines that were subjected to an initial set of genetic and phenotypic assays (hereafter 'Tier 1' assays) (**Fig. 1a)**. Tier 1 assays included array based genotyping and gene expression profiling of the iPSCs and their fibroblast progenitors. For 307 lines we quantified protein expression of *NANOG*, *POU5F1* (*OCT4*), and *SOX2* using immunohistochemistry followed by quantitative image analysis using the Cellomics (Thermo Fisher Scientific) high content imaging system. We also differentiated 372 lines into neuroectoderm (dEC), mesoderm (dME), and endoderm (dEN), using a defined culture system [2], and measured the expression of three lineage-specific differentiation markers (**Fig. 1a**, **Extended Data Fig. 1**).

The Tier 1 assay data were used to select 1-2 high quality lines (**Methods**) for each donor for further phenotyping and cell line banking, minimising the number of genetic abnormalities and maximizing pluripotency. For this study, 239 lines (hereafter 'selected lines') from 166 donors were selected based on Tier 1 assay data, and profiled using RNA-seq, with lines from 27 donors subjected to DNA methylation profiling, 9 donors to quantitative proteomics and 12 to cell morphological imaging using an Operetta (Perkin Elmer) high content imaging system (hereafter 'Tier 2' assays) (**Supplementary Table 1**).

4. Estimating statistical significance of CNA recurrence

The significance of recurrently observed trisomies was estimated from the complementary cumulative distribution function of the binomial distribution as follows. Let m = 23 be the number of tested chromosomes, n = 24 be the number of trisomies observed across all samples, and k the maximum number of trisomies observed for a given chromosome (trisomies in multiple lines of the same donor are counted only once). We calculate *P*-value as:

$$P(x \geq k) = 1 - P(x \leq k-1) = 1 - (n-k+1)\frac{n}{k-1}\int_0^{1-1/m} t^{n-k}(1-k)^{k-1}dt$$

The significance of sub-chromosomal events was estimated using a permutation test, which

was performed separately for insertions and deletions. We split the genome into 200 Kb regions. Then in each permutation we generate as many intervals as there were CNAs observed in real data, and place them randomly onto the genome. The size of these intervals matches the length of the CNAs. Each of the regions R_i is now associated with two numbers: C_i, the number of real CNAs overlapping the region, and S_i, the number of simulated intervals overlapping the region. Define the indicator function $I(S_i \geq C_i)$ to be 1 if the number of intervals randomly placed in the region R_i is bigger or equal to the number of real CNAs observed in the region and 0 otherwise. After running $M = 5\text{x}10^8$ tests, we calculated the *P*-value as:

$$P(R_i) = \left[1 + \sum_{1}^{M} I(S_i \geq C_i)\right] / M$$

5. Modelling of the Cellomics phenotype and normalization

In order to differentiate non-responding cells from the cells that express the marker of interest, we fit a mixture model of two gamma distributions ($G(\alpha_0, \beta_0)$ and $G(\alpha_1, \beta_1)$) to the intensity profile in each well:

$$p(x|\alpha_0, \beta_0, \alpha_1, \beta_1) = \pi_0 G(x|\alpha_0, \beta_0) + (1 - \pi_0) G(x|\alpha_1, \beta_1)$$

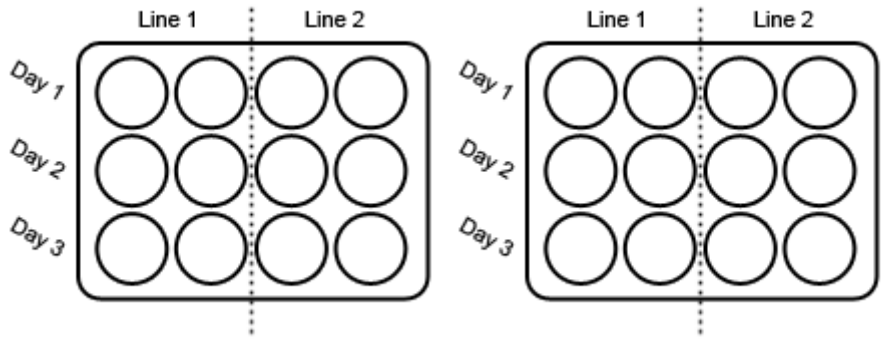
Briefly, $G(\ \ )$ denotes the gamma distribution and $(1 - \pi_0)$ corresponds to the proportion of responding cells. The parameters of each mixture component (α_0, β_0 and α_1, β_1) were estimated using maximum likelihood. To improve the fitting process, we used wells with background signal (secondary staining only). The mixture component that corresponds to responding cells (parameters α_1, β_1) was fit to data from wells with primary antibody staining only. In contrast, the mixture component to model background signal from non-responding cells was fit to model cells in the primary wells as well as to model background (secondary only staining wells) (**Extended Data Fig. 1**). The final model fit was used to estimate both the proportion of responding cells as well as the overall intensity (expression) of the responding cells.

The proportion of responding cells corresponds to the mixture coefficient $(1-\pi_0)$. The intensity estimates (expression level) were derived from weighting this proportion by the mean intensity of the corresponding mixture component. Wells with fewer than 5,000 cells were discarded from the analysis. Additionally, wells for which the background signal on the matching plate exceeded a threshold value that corresponds to 10% responding cells (estimated using the reference line) were removed.

6. Functional characterisation of copy number duplications

6.1. Growth curve assay

A week prior to seeding, iPSCs were thawed and plated on to six well plates [Corning, 3506], coated with Vitronectin [StemCell Technologies, UK, 07180]. The cells were plated using ROCK inhibitor, Y-27632 [Sigma-Aldrich, UK, Y0503] at a concentration of 1:1000 in Essential 8 complete medium consisting of basal medium DMEM/F-12(HAM) 1:1 (Life technologies, UK, A15169-01) supplemented with E8 supplement (50X) (Life technologies, UK, A15169-01) and 1% Pen/Strep (Life Technologies, UK, 15140122). After approximately three to four days the cells were passaged using PBS-EDTA [Invitrogen, UK, 15575-038] back onto a 6 well plate at a 1:3 to 1:6 split ratio. Once confluent, the cells were enzymatically dissociated with Accutase [Gibco, A11105-01] and incubated at 37C / 5% CO2 for 10 minutes. The cells were harvested and counted using a Nucleocounter NC200 and then seeded onto six 12 well plates [Corning, 3513], again coated with Vitronectin, at a density of 30,000 cells/cm2. To account for plate to plate variation, four replicates were seeded across two plates.



Experimental design allowing four replicates to be split across two plates in a three-day collection window.

For a period of approximately five to seven days, the cells were collected on a 24-hour schedule. Each day the cells were dissociated with Accutase for 10 minutes at 37C / 5% CO2, washed with 1ml of complete E8, collected and centrifuged at 120 RCF for three minutes. After centrifugation, the supernatant was aspirated and the pellet resuspended in 200µl-1ml of E8, dependent upon pellet size. A sample was then extracted and counted using the Nucleocounter system. This process was repeated until the cells started to reach a stationary phase, or the plates became unusable due to excessive cell detachment.

6.2. Tunel Assay

One well of confluent iPSCs six well was rinsed with HBSS and 1mL/well TrypLE (1 ml) was added. The plate was then transferred to the incubator for 8 minutes (37°C; 5% CO2; 20% O2), until cells could be removed by gentle pipetting. The collected cells were transferred into E8 (1 ml) and centrifuged at 400 RPM for four minutes. Once the supernatant was removed, E8+ Rock inhibitor (10µM) (1 ml) was added to re-suspend the pellet and the cell were counted. 3000 cells were plated onto one well of 96 well plate that was coated with vitronectin. The cells were allowed to adhere to the plate for two hours. There were three conditions to the assay, treated, untreated and control. The treated condition had E8 + 0.1 µg/ml mitomycin, untreated condition had only E8 and control condition E8 + 10 µg/ml mitomycin. The first two conditions had three technical replicates.

The cells were placed in the incubator for twenty-four hours and then fixed using 8% PFA (para-formaldehyde) and rinsed three times with PBS. The cells were stained for DAPI (4',6-diamidino-2-phenylindole) and tunel (Click-iT® TUNEL Alexa Fluor® 488 Imaging Assay). The images were acquired using Operetta High-Content Imaging System at 10X. All

experiments were replicated at least three times.

6.3. EdU Assay

One well of a confluent iPSCs six- well plate was rinsed with HBSS and 1mL/well TrypLE (1 ml) was added. The plate was then transferred to the incubator for 8 minutes (37°C; 5% CO2; 20% O2), until cells be removed by gentle pipetting. The collected cells were transferred into E8 (1 ml) and centrifuged at 400 RPM for four minutes. Once the supernatant was removed, E8+ Rock inhibitor (10μM) (1 ml) was added to re-suspend the pellet and the cell were counted. 3000 cells were plated onto one well of 96 well plate that was coated with vitronectin. The cells were allowed to adhere to the plate for two hours. There were three conditions to the assay, treated, untreated and control. The treated condition had E8 + 0.1 μg/ml mitomycin, untreated condition had only E8 and control condition E8 + 10 μg/ml mitomycin. The first two conditions had three technical replicates. The cells were placed in the incubator for twenty-four hours; Edu was added to the cells for half hour before fixation using 8% PFA. The cells were stained for DAPI and Edu (Click-iT® EdU Alexa Fluor® 488). The images were acquired using Operetta High-Content Imaging System at 10X. All experiments were replicated at least three times using the iPS cell lines mentioned in the figure.

7. Definition of proxy variants for eQTL replication

To define tissue-specific eQTLs, replication of the eQTL effect was tested on the level of individual eQTL variants between all pairs of tissues (iPSC and 44 somatic tissues from GTEx). As detailed in **Methods**, if the exact same lead variant could not be tested in a tissue, a proxy variant was selected as follows:

For each eGene in a discovery tissue:
Is the original lead variant available to measure in the query tissue?

1. If YES, does it replicate?
    If YES -> end.
    If NO, does the lead variant have any high-LD proxies?
        If YES, are they available to measure in the query tissue?
            If YES, do any of them replicate?

If YES -> end.
If NO -> end.
If NO -> end.
If NO -> end.
2. If NO, does the lead variant have any high-LD proxies?
If YES, are they available to measure in the query tissue?
If YES, do any of them replicate?
If YES -> end.
If NO -> end.
If NO, does best available variant* replicate?
If YES -> end.
If NO -> end.
If NO, does best available variant* replicate?
If YES -> end.
If NO -> end.

8. Comparison of pipeline effects between HipSci and GTEx

There are important differences between the eQTL calling pipeline used in this study and the approach taken by the GTEx project, which could affect the comparisons between the respective sets of eQTLs. Differences between the pipelines include, but are not limited to: i) RNA-seq read alignment with STAR (HipSci) versus TopHat 2 (GTEx), ii) expression quantification on the level of Gencode v19 genes using HTSeq (HipSci) versus custom set of exons based on Gencode v19 using RNA-SeQC (GTEx), iii) eQTL mapping on PEER K30 residuals using LIMIX (HipSci) versus on PEER K15 factors using Matrix-eQTL, and iv) correction for population stratification using a kinship matrix as a random effect (HipSci) versus top three principal components of the genotypes (GTEx).

Since reprocessing the full GTEx dataset with the HipSci pipeline is computationally demanding, we took the following steps to evaluate the effects of pipeline differences:

1. Re-processing of two GTEx tissues using the HipSci read mapping pipeline
2. Re-processing of HipSci data with a ‘GTEx-like’ QTL mapping pipeline

First, we implemented the GTEx V6p quantification and mapping pipeline as closely as possible without re-aligning the RNA-seq reads, based on methods information available on the GTEx web portal (www.gtexportal.org), hereon referred to as the 'GTEx-like' pipeline. To validate that results obtained with this pipeline indeed resemble the original GTEx V6p eQTL results, we re-processed two GTEx tissues (Adrenal Gland and Esophagus Gastroesophageal Junction, hereon referred to as 'AG' and 'EGJ') using the HipSci read alignment pipeline, followed by the GTEx-like expression quantification and eQTL mapping. We then compared the eQTL results obtained from the re-processed tissues to the original V6p results released by GTEx.

To assess how the original eQTLs replicated in the reprocessed tissues, we tested for their nominal significance in the reprocessed tissue ($P < 0.01$) and found that 72.2% and 72.0% of the V6p lead eQTLs (AG and EGJ, respectively) replicated in the corresponding reprocessed data ('GTEx-like' pipeline). As an alternative, we also calculated the global π_1 statistic ($\pi_1 = 1 - \pi_0$; using the *qvalue* package in R) between the original and reprocessed eQTLs, again confirming that the reprocessed and original GTEx eQTL maps are very similar ($\pi_1 = 0.96$ and 0.94, AG and EGJ, respectively). Of note, while the variant-specific replication test considers only lead eQTLs of the discovery tissue, the π_1 analysis considers all significant variants per eGene, which is likely to explain the difference between the two tests. Taken together, this analysis indicates that the eQTLs discovered using the GTEx-like pipeline resemble the original GTEx mapping approach to a sufficient extent, despite underlying differences in RNA-seq technology (such as read length and insert size) and read alignment.

Second, we reprocessed the HipSci iPSC RNA-seq data using the 'GTEx-like' quantification and eQTL mapping pipeline, which yielded 3,747 genes with a significant eQTL at a false discovery rate of 5% (compared to 6,631 when using the HipSci pipeline). This difference in power is most likely due to differences in gene expression quantification. Further, the GTEx-like pipeline adjusts for fewer PEER factors than our pipeline (15 and 30, respectively, determined by sample size in GTEx), which in our iPSC dataset may additionally contribute to power differences. Next, we used this second eQTL map in HipSci (hereon referred to as 'iPSC-alternative') to assess the extent to which iPSC eQTLs replicate in GTEx tissues. We observed the same overall trends than with the original HipSci pipeline, although the proportion of iPSC-specific eQTLs was slightly lower (22% versus 32%, alternative and

original, respectively; **Extended Data Fig. 8a**). This moderate difference is most likely due to a combination of reduced power and more similar pipelines. Finally, we used the alternative iPSC eQTLs to repeat the functional enrichment analysis at transcription factor binding sites. While the overall degree of fold enrichment was higher with this alternative (potentially more conservative) set, the overall profile of enriched factors was very similar to the results obtained with the original HipSci pipeline (**Extended Data Fig. 8d**), confirming that the biological conclusions derived from the iPSC eQTLs are robust across pipelines.

References

1 Fusaki, N., Ban, H., Nishiyama, A., Saeki, K. & Hasegawa, M. Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. *Proc Jpn Acad Ser B Phys Biol Sci* **85**, 348-362 (2009).

2 Vallier, L. *et al.* Signaling pathways controlling pluripotency and early cell fate decisions of human induced pluripotent stem cells. *Stem Cells* **27**, 2655-2666, doi:10.1002/stem.199 (2009).