Identifying the genetic basis of variation in cell behaviour in human iPS cell lines from healthy donors

Alessandra Vigilante[1,2,4,7], Anna Laddach[5], Nathalie Moens[1,8], Ruta Meleckyte[1,10], Andreas Leha[3,9], Arsham Ghahramani[1,4], Oliver J. Culley[1], Annie Kathuria[1], Chloe Hurling[1], Alice Vickers[1], Mukul Tewary[6], Peter Zandstra[6], HipSci Consortium, Richard Durbin[3], Franca Fraternali[5], Oliver Stegle[2], Ewan Birney[2], Nicholas M. Luscombe[4,7], Davide Danovi[1*] and Fiona M. Watt[1*]

[1] Centre for Stem Cells and Regenerative Medicine, King's College London, Floor 28, Tower Wing, Guy's Hospital, Great Maze Pond, London, SE1 9RT, UK

[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

[3] The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK.

[4] The Francis Crick Institute, 1 Midland Road, London NW1 1AT

[5] Randall Division, King's College London, New Hunts House, Great Maze Pond, London SE1 9RT

[6] The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Room 1116, Toronto, Ontario, Canada M5S 3E1

[7] UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT

[8]Present address: GlaxoSmithKlein, Gunnels Wood Road, Stevenage, Herts, SG1 2NY

[9]Present address: University Medical Center Göttingen, Georg-August-Universität, Department of Medical Statistics, Humboldtallee 32, 37073 Göttingen, Germany

[10]Present address: Sobell Department, University College London Institute of Neurology, Queen Square House, Queen Square, London WC1N 3GB.

[*] Corresponding authors: davide.danovi@kcl.ac.uk; fiona.watt@kcl.ac.uk

Abstract

Large cohorts of human iPSCs from healthy donors are potentially a powerful tool for investigating the relationship between genetic variants and cellular phenotypes. Here we integrate high content imaging, gene expression and DNA sequence datasets for over 100 human iPSC lines to identify the genetic basis of inter-individual variability in cell behaviour. By applying a dimensionality reduction approach, Probabilistic Estimation of Expression Residuals (PEER), we identified genes that correlated in expression with intrinsic (genetic) and extrinsic (ECM) factors. However, variation in mRNA levels could not account for outlier cell behaviour. Instead, we identified rare, deleterious SNVs in the coding sequence of genes involved in ECM adhesion that occurred in cell lines that were outliers for one or more phenotypes such as cell spreading. These also correlated with altered germ layer differentiation on micropatterned surfaces. Our study thus establishes a strategy for integrating genetic and cell biological measurements for high-throughput analysis.

Introduction

Now that the applications of human induced pluripotent stem cells (hiPSC) for disease modelling and drug discovery are well established, attention is turning to the creation of large cohorts of hiPSC from healthy donors to examine common genetic variants and their effects on gene expression and cellular phenotypes[1-5]. Genome-wide association studies (GWAS) and quantitative trait locus (QTL) studies can be used to correlate single nucleotide polymorphisms (SNPs) and other genetic variants with quantitative phenotypes[6]. As part of this effort, we recently described the generation and characterisation of over 700 open access hiPSC lines derived from 301 healthy donors as part of the Human Induced Pluripotent Stem Cells Initiative[5,7]. In addition to creating a comprehensive reference map of common regulatory variants that affect the transcriptome of hiPSC, we performed quantitative assays of cell morphology and could demonstrate a donor contribution of approximately 8-23% to the observed variation. In the present study we set out to identify the causative genetic variants.

Previous attempts to use lymphoblastoid cell lines to link genetics to in vitro phenotypes have had limited success[8,9]. In that context, confounding effects included EBV viral transformation, the small number of lines analysed, variable cell culture

conditions and line-to-line variation in growth rate. These non-genetic factors decrease the power to detect true relationships between DNA variation and cellular traits[8]. In contrast, we have access to a large number of normal h iPSC lines, including multiple clonal lines from the same donor. In addition, Hipsci cell lines present a substantially lower number of genetic aberrations than reported for previous collections[5,10,11]. Cells are examined at low passage number, and cell properties are evaluated at single cell resolution during a short time frame, using high throughput quantitative readouts of cell behaviour.

Stem cell behaviour reflects both the intrinsic state of the cell[12,13] and the extrinsic signals it receives from its local microenvironment, or niche[14,15]. We reasoned that subjecting cells to different environmental stimuli would increase the likelihood of uncovering links between genotype and cell behaviour. For that reason we seeded cells on different concentrations of the extracellular matrix (ECM) protein fibronectin that supported cell spreading to differing extents, and assayed the behaviour of single cells and cells that were in contact with their neighbours. We took a 'cell observatory' approach, using high-throughput, high content imaging to gather data from many millions of cells 24h after seeding. We used Probabilistic Estimation of Expression Residuals (PEER)[16] to capture variance due to extrinsic and genetic components, and correlated outlier cell behaviour with the presence of rare deleterious SNVs. The strategy we have developed bridges the gap between genetic and transcript variation on the one hand and cell phenotype on the other, and should be of widespread utility in determining the genetic basis of inter-individual variability in cell behaviour.

Results

Generation and characterisation of the lines

We analysed 110 cell lines from the HipSci resource[2] (Supplementary Table 1). Of these, 99 lines were reprogrammed by Sendai virus and 11 using episomal vectors. 100 lines came from 65 healthy research volunteers; thus several lines were different clones from the same donor. 7 lines came from 7 individuals with Bardet-Biedl Syndrome and 3 were non-HipSci control lines. 102 of the lines were derived from skin fibroblasts, 2 from hair and 6 from peripheral blood monocytes. All lines passed the quality controls specified within the Hipsci production pipeline, including high PluriTest (Stem Cell Assays) scores and the ability to differentiate along the three embryonic

germ layers. All the cell lines were reprogrammed on feeders and all but 6 lines were cultured on feeders prior to performing phenotypic analysis (Supplementary Table 1). Cells were examined between passages 15 and 45.

Cell behaviour assays

To quantitate cell behaviour at single cell resolution we used the high-content imaging platform that we have described previously[17]. Cells were disaggregated and resuspended in the presence of 10 μM Rho-associated protein kinase (ROCK) inhibitor (Y-27632; Enzo Life Sciences) to minimise cell clumping. In order to provide varied conditions for cell adhesion and spreading, cells were seeded on 96 well plates coated with 3 different concentrations of fibronectin – 1, 5 and 25 μg/ml (Fn1, Fn5, Fn25). After 24h of culture in the presence of ROCK inhibitor, cells were labeled with EdU for 30 min, fixed, and stained with CellMask (to visualise cytoplasm) and DAPI (to visualise nuclei; Fig. 1A). Under these conditions, over 95% of cells were in the pluripotent state, as evaluated by Oct4 labelling.

Three replicate wells were seeded per cell line and cell lines were analysed in up to three independent experiments. Wells containing technical triplicates of each fibronectin concentration were randomised per column (*e.g.* 1-5-25; 5-25-1) to obviate edge and position effects. Technical replicates of the same cell line were randomised in rows and one hiPSC line, previously reported as A1ATD-iPSC patient 1[18], was included to control for biological variation between experiments.

For each cell line we determined a range of phenotypes, including the number of cells that attached, cell and nucleus morphology, intercellular adhesion and proliferation (EdU incorporation) (Fig. 1B, C). As shown previously[17], there was a clear effect of fibronectin concentration on phenotypic features. For example, both cell and nuclear area increased with increasing fibronectin concentration; additionally nuclear roundness increased, whereas cell roundness decreased with increasing fibronectin concentration (Fig. 1C).

We extracted a total of 11 measurements per cell, distinguishing between cells that had attached as individuals and cells in clumps (Fig. 1B). The phenotypic features were processed as described previously[17], *i.e.* measurements were normalised in

value (log10 or square transformation) and aggregated across the cells in each well by taking the average and standard deviation. For EdU incorporation, median intensity raw values were grouped and characterised on a well-based measure as the fraction of EdU positive cells. These resulted in a final list of 52 phenotypes (Extended Data Table 1). Some features were positively correlated with one another, such as cell area and nuclear area, while in other cases, such as cell area and cell roundness, there was an inverse correlation (Fig. 1B). In total we obtained phenotypic measurements of approximately 2 Million objects/cells.

The complexity of the dataset is represented in Fig. 1D, where the mean value of cell area is represented for all cell lines, for the three fibronectin concentrations in three experiments. This illustrates the variance between replicate experiments. It also shows a clear effect of fibronectin concentration on average cell area, with cells exhibiting greater cell area on the highest concentration (see also Fig. 1C). Furthermore, it is possible to appreciate some intra-donor variability for cell lines derived from the same donor (denoted by a common 4 letter code; *e.g.* airc). Similar results were obtained for other raw phenotypic features (Extended Data Fig. 1).

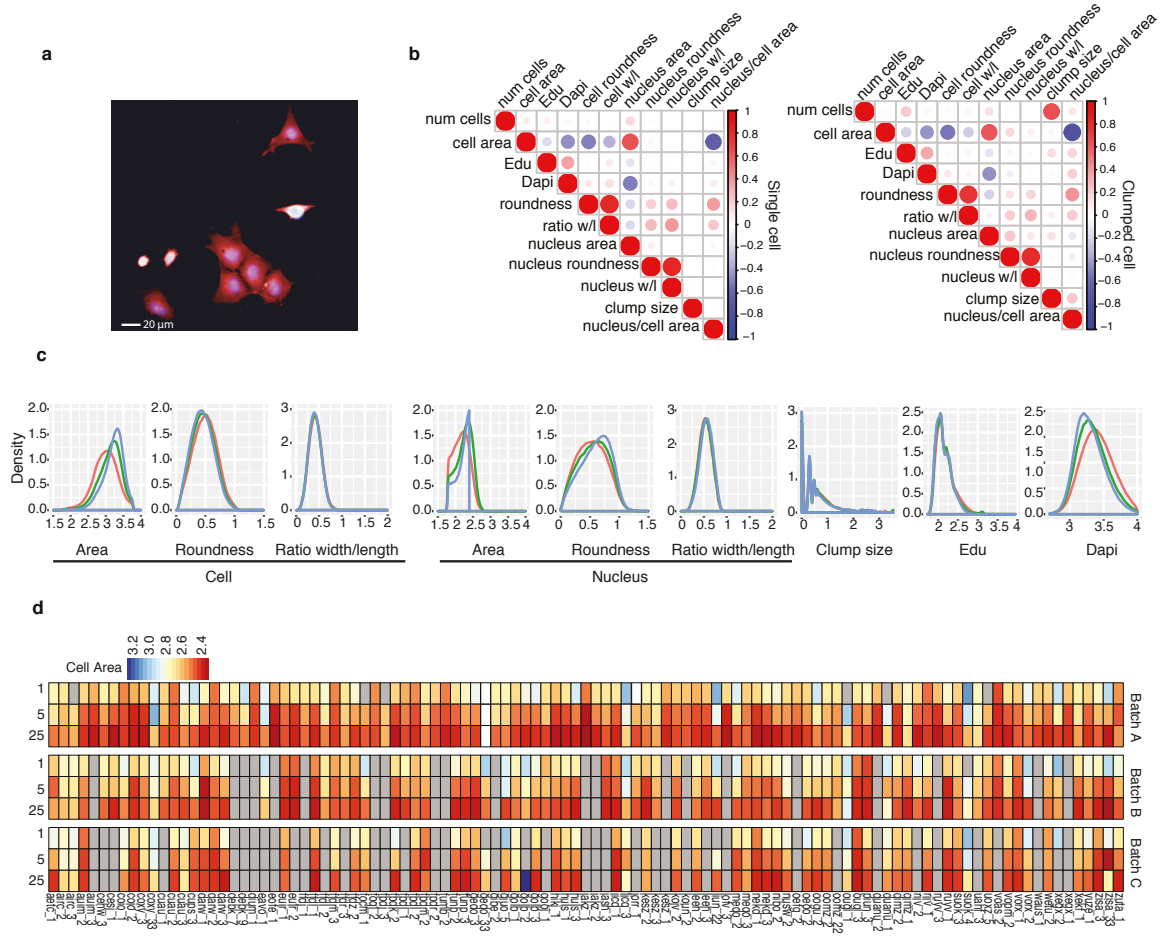

Figure. 1. Description of phenotypic dataset from 110 iPS lines from 85 individuals. a. Image from high content screen showing cells 24h after plating. Red: cell mask (cytoplasm); blue: DAPI (nuclei); white: EdU incorporation (DNA synthesis). EdU+ cells are marked with asterisks. Scale bar: 20 µm**. b.** Correlation of different phenotypic measurements in single (left) and clumped cells (right). **c.** Distribution of the main phenotypic features for all cell lines on three fibronectin concentrations (Fn1, red; Fn5, green; Fn25, blue). Y axis: cell number. **d** Heatmap of mean cell area measurements for each cell line in three conditions in three biological replicates (batches). Grey boxes correspond to data unavailable or not satisfying the minimum number of cells per well threshold.

Contribution of intrinsic and extrinsic factors to variation in cell features

In order to distinguish how extrinsic (*i.e.* different fibronectin concentrations), intrinsic (*i.e.* cell line or donor specific) and technical or biological components affected variation in cell features, we applied a dimensionality reduction approach. We used

Probabilistic Estimation of Expression Residuals (PEER)[16] to identify and extract factors that explain hidden portions of the phenotype variability. In particular, we set out to quantify the fraction of phenotypic variation attributable to donor effects as opposed to different extracellular conditions. PEER was originally implemented for gene expression data and to our knowledge this study is the first using PEER for multidimensional reduction of phenotypic data.

In our analysis, PEER takes an input from the phenotypic measurements and covariates (such as fibronectin concentration, batch, donor) and outputs a suitable number of hidden factors that explain much of the variance and can be treated as new synthetic phenotypes. Using automatic relevance detection (ARD) parameters to understand which dimensions are needed to model the variation in the data[16], the optimal number of PEER factors was set to 9 (Extended Data Fig. 2). Of these, the effect of fibronectin concentrations was apparent only in one (Peer factor 1, named the 'extrinsic factor'), which therefore captured the variance due to extrinsic contributors, accounting for 30% of the total variance (Fig. 2A-C and Extended Data Fig. 3).

To discover whether one or more PEER factors could capture the variation due to the genotype of the cell lines, we compared the value of each factor in different clonal cell lines derived from the same donor (2-3 lines per donor; Extended Data Fig. 3). Factor 9 showed the maximum genetic concordance, capturing the variance attributable to intrinsic components (here named the 'intrinsic factor') and accounting for 5% of the variance (Fig. 2D-F; Extended Data Fig. 3).

We were able to explore and quantitate the contribution of single raw phenotypic features to the synthetic PEER factors. Nucleus and cell area, number of cells and DAPI intensity were the highest to load onto the extrinsic factor (Fig. 2C). Phenotypic features describing other nuclear properties, both in single and clumped cells, and EdU intensity loaded onto the intrinsic factor (Fig. 2F). Using Principal Component Analysis (PCA) we observed the contribution of the diverse fibronectin concentrations to the variance attributable to the first two principal components (45%). In agreement with what we previously observed for one control iPS cell line[17], cells plated on the same concentration of fibronectin albeit now from several different donors, clustered together (Fig. Extended Data Fig. 4).

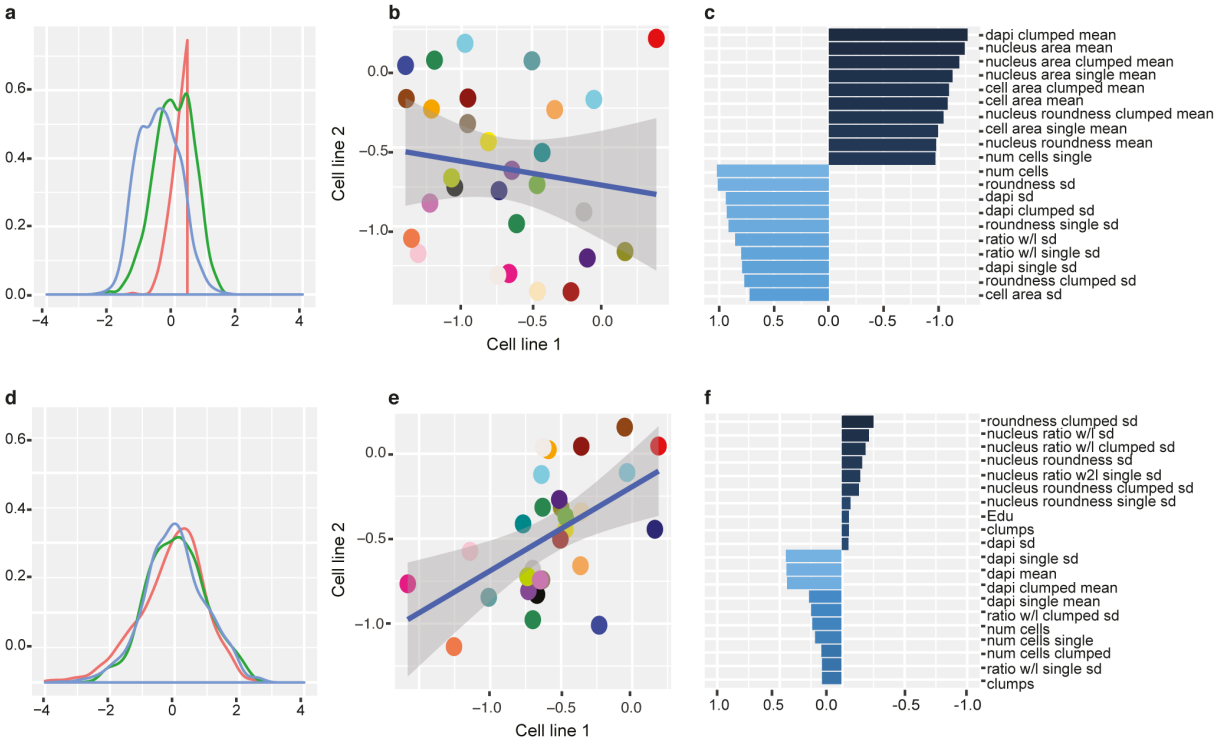

Figure 2. Synthetic phenotypic features capture extrinsic and intrinsic contributions to variance. a,b,c,d,e,f Plots showing the distribution of values for the 'extrinsic' (a; Factor 1) and 'intrinsic' (b; Factor 9) synthetic phenotypic features obtained with PEER dimensionality reduction. The effect of extrinsic conditions (Fn1, red; Fn5, green; Fn25, blue; a, d) and the genetic concordance between two clones of cells from the same donors (b, e) are shown. **c, f.** Loading of phenotypic raw features on the extrinsic (c) and intrinsic (d) factors.

Identification of genes correlating with extrinsic and intrinsic variation in cell phenotype

We next set out to identify how gene expression related with the synthetic PEER factors. We performed a correlation analysis between the factors and gene expression array data generated from cell pellets as part of the HipSci resource[7] and performed Gene Ontology analysis of the genes identified[19] (Fig. 3; Supplementary Table 3). Genes correlating only with the extrinsic factor (PEER factor 1) were enriched in signaling pathways associated with cell surface receptors, proliferation and negative regulation of differentiation (Fig. 3A). Conversely, genes correlating with the intrinsic factor (PEER factor 9) were enriched in a wider range of terms, including intracellular receptor signalling and transport, and positive regulation of transcription (Fig. 3A).

The expression of a total of 4573 genes correlated with either the extrinsic or intrinsic factors or both, in at least one fibronectin concentration. From this list, we applied two different sets of filters. We discarded genes that were not associated with any Ensembl identifiers and removed genes for which multiple probes showed opposite correlation values. The resulting dataset consisted of 3880 genes (Supplementary Table 2).

Based on the phenotypes measured in our study, we further filtered the genes according to the functions of their protein products (GO and PANTHER protein class terms)[20]. Specifically, we selected from this list genes we reasoned were likely to be involved in the distinct cell behaviours observed, belonging to the following four categories: extracellular matrix (ECM) proteins, cell adhesion molecules, signaling molecules and cell junction proteins (Fig. 3B). The resulting data set consisted of 332 genes (Supplementary Table 2) and contained many cadherins and integrins, growth factors and structural proteins that associate with integrin cytoplasmic domains (Fig. 3C). We therefore performed a correlation analysis between their expression and key measured phenotypes depicting cell morphology, adhesion and proliferation. Out of the 332 genes, 160 showed a statistically significant correlation with at least one phenotypic feature in at least one fibronectin concentration (Supplementary Table 2). Expression of 54 of the genes correlated significantly with cell area, clumping, number and/or proliferation. Examples of gene expression variation among cell lines for genes correlating with one, two, three and four phenotypic features are shown in Fig. 3D.

We noted that most genes showed distinct correlations with the intrinsic and extrinsic PEER factors (Fig. 3C). In addition, opposite correlations were found for a given gene and one or more phenotypes (Fig. 3D, Supplementary Table 2). For example, CDH9 and other protocadherins, which mediate intercellular adhesion, were positively correlated with clumping and negatively correlated with proliferation. Conversely, expression of the integrin extracellular matrix receptor ITGAX showed significant and opposite correlations with cell area and proliferation. The growth factor GHRH and the disintegrin ADAMTSL4 that controls the structure and function of extracellular matrix[21], were both positively correlated with clumping and inversely correlated with the number of attached cells. Two genes correlated with four raw phenotypic features but in opposite directions: DSC2, a cadherin-type component of desmosomal junctions[22], showed positive correlations with cell number and clumping and negative correlations

with cell area and proliferation; conversely CD99, a cell surface glycoprotein implicated in cell proliferation, showed the opposite correlation[23].

Altogether these results indicate a robust correlation between RNA expression and the phenotypic features in a large panel of iPSC lines, with specific RNAs associated with intrinsic or extrinsic factors. We next set to examine whether outlier cell behaviour could be predicted on the basis of gene expression level.

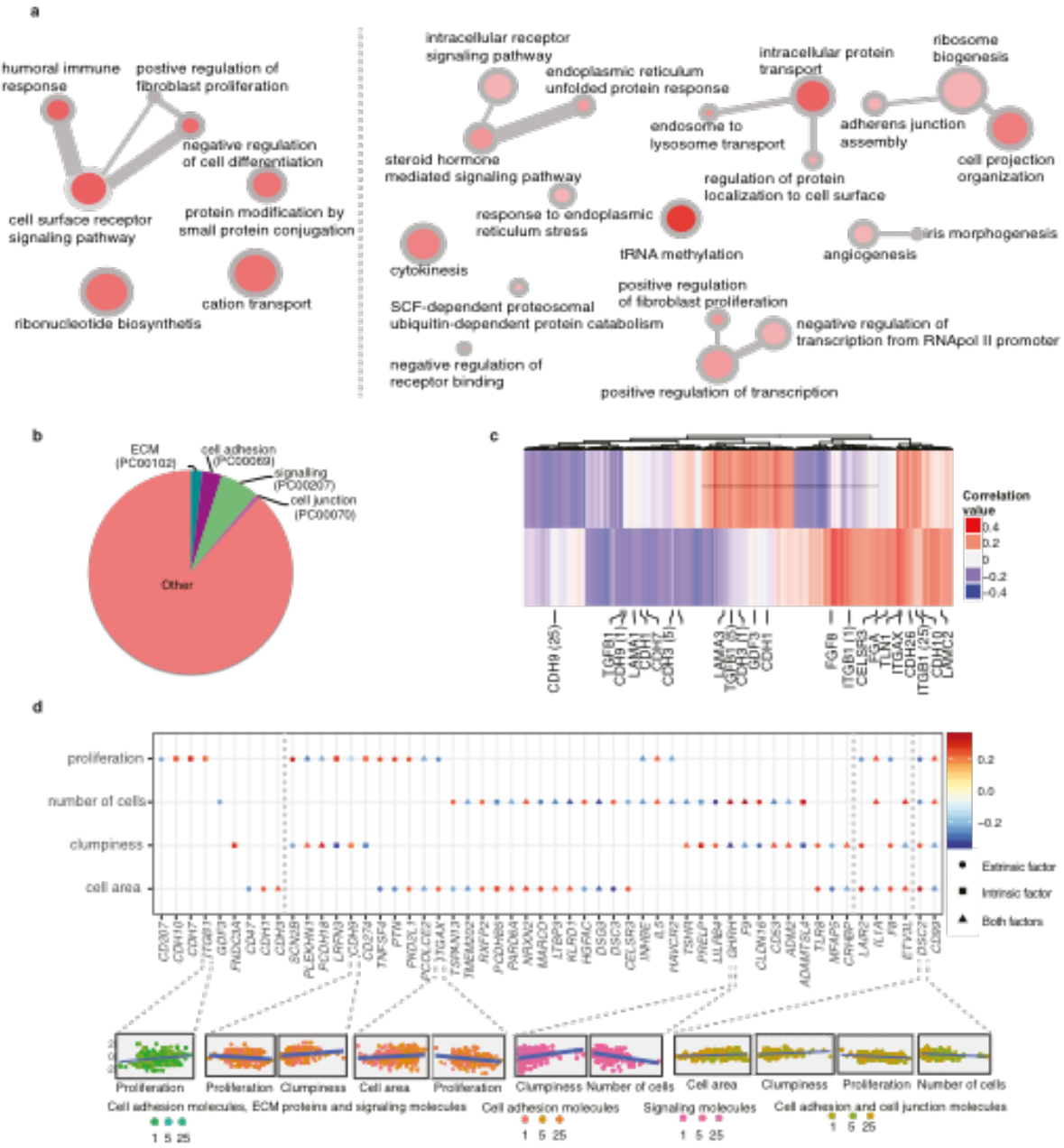

Figure 3. Using the 'extrinsic' and 'intrinsic' factors to identify genes associated with specific phenotypic features. a. GO analysis of genes correlating with 'extrinsic'

(PEER Factor 1; left) or 'intrinsic' (PEER Factor 9; right) factors. Node colour indicates the p-value (threshold: p-value < 0.0001): the greater the intensity, the lower the p-value. Node size indicates the frequency of the GO term in the GOA database. Each gene was mapped to the most specific terms applicable in each ontology. Highly similar GO terms are linked by edges, edge width depicting the degree of similarity. **b.** Genes were filtered using Panther protein classes and a total of 332 genes belonging to the 4 groups shown were found. **c.** Out of the 332 genes, the expression of 160 genes shown in the heatmap correlates significantly with either or both PEER factors in at least one fibronectin condition. Names of some of the genes are shown. **d.** In a total of 54 genes, gene expression correlated significantly with cell area, tendency to form clumps ('clumpiness'), number of cells and/or proliferation. The colours of the points correspond to the correlation values, while the shapes indicate correlation of a specific gene to the extrinsic and/or intrinsic factors. Grey dotted vertical lines separate genes correlating with one, two or four phenotypes (left to right). The entire list of genes correlating with each phenotype is shown in Extended Data Table 2.

Identification of SNVs in cell adhesion genes that account for outlier cell phenotypes

While the bulk gene expression data yielded correlations with the behaviour of cell populations, cells that deviated significantly from modal phenotypic values did not show a corresponding variation in gene expression levels (Extended Data Fig. 5). We therefore explored whether the presence of SNVs in gene exons affecting protein function could account for outlier cell behaviour. We searched for SNVs in the 332 genes whose expression correlated with the intrinsic and extrinsic PEER factors (Supplementary Table 2). Of the 516 SNVs identified, 298 were classified as rare based, first, on the 1000 Genomes Project [24] and ExAC[25], and, secondly, on the frequency in our cell lines (present in fewer than 5 out of 110 lines) (Fig. 4A). We further filtered the SNVs according to whether they were predicted to be deleterious by DUET[26] and to destabilise protein structure according to Condel[27]. The genes that we identified (Supplementary Table 4) encoded proteins that were associated with cell adhesion, including integrins, cytoskeleton and ECM proteins.

The vast majority of the SNVs (18 out of 28) identified in this way occurred in cell lines that were outliers for one or more phenotypes (Fig. 4B and Extended Data Fig. 5). An

interesting example is the presence of deleterious and destabilising SNVs in the integrins ITGA6 (position 27) and ITGB6 (position 417) in the cell line yuze_1, which was detected as an outlier for cell and nucleus ratio width/length and for EdU and DAPI intensity. These cells show reduced spreading (enhanced cell roundness) on the lowest concentration of Fibronectin, and increased cell width to length ratio on the higher Fibronectin concentrations (Fig. 4C and Supplementary Table 4).

Having acquired cell behaviour data for 110 of the HipSci lines, we next explored the genome sequences of the full collection of over 700 lines to test whether the presence of rare, deleterious and destabilising SNVs could correctly predict outlier cell behaviour. We identified a SNV in the integrin ITGB1 in the cell line ffdc_11 and a SNV in the FHL2 gene in pamv_1 (also present in another independent clone from the same donor, pamv_3). The ffdc_11 line showed a distinct morphology indicating an outlier phenotype characterised by reduced cell spreading on Fn1. In contrast pamv_1 and pamv_3 (Fig. 4D) were indistinguishable from controls (Fig. 4E). Although ffdc_11 and the two pamv clones exhibited different spreading on fibronectin, this did not correlate with ITGB1 expression (Fig. 4F), consistent with the observation that gene expression levels did not predict outlier cell behaviour.

Since the phenotypic measurements were carried out on cells in the pluripotent state, we next investigated whether cells harboring potentially deleterious SNVs also exhibited outlier phenotypes when induced to differentiate. Recent work has shown that human pluripotent stem cells cultured on 1000 μm diameter circular micropatterned substrates in the presence of BMP4 self-organise into the three germ layers, recapitulating their *in vivo* spatial arrangement Warmflash et al.[28] and Tewary et al.[29] have developed the technology to enable screening of cells in 96 well plates. We therefore evaluated whether the ffdc_11 and pamv_1 lines differed in their capacity to differentiate into the three germ layers (Fig. 4F-G). We seeded each cell line on micropatterned substrates arrayed in 96-well plates, added BMP4 for 48 hours and then fixed and labelled the cultures with antibodies to Sox2 (ectoderm), EOMES (mesoderm) and Sox17 (endoderm marker). As previously reported, we observed spatial segregation of each cell type (Fig. 4F-I). Warmflash et al[28] observed that Sox2+ cells were most centrally located, surrounded by a zone of EOMES+ cells, and – at the periphery – Sox17+ cells. Compared to a control cell line with no SNV detected, in the

case of ffdc_11 cells, Sox2+ cells centrally, surrounded by EOMES+ cells (**with increased levels of expression)**, and a peripheral concentration of SOX17+ cells (Fig. 4F, G). However, in pamv_1 cells the concentration of SOX2+ cells shifted towards the periphery. Thus, both ffdc11 and pamv1 cells presented with a SNV and with altered differentiation patterns.

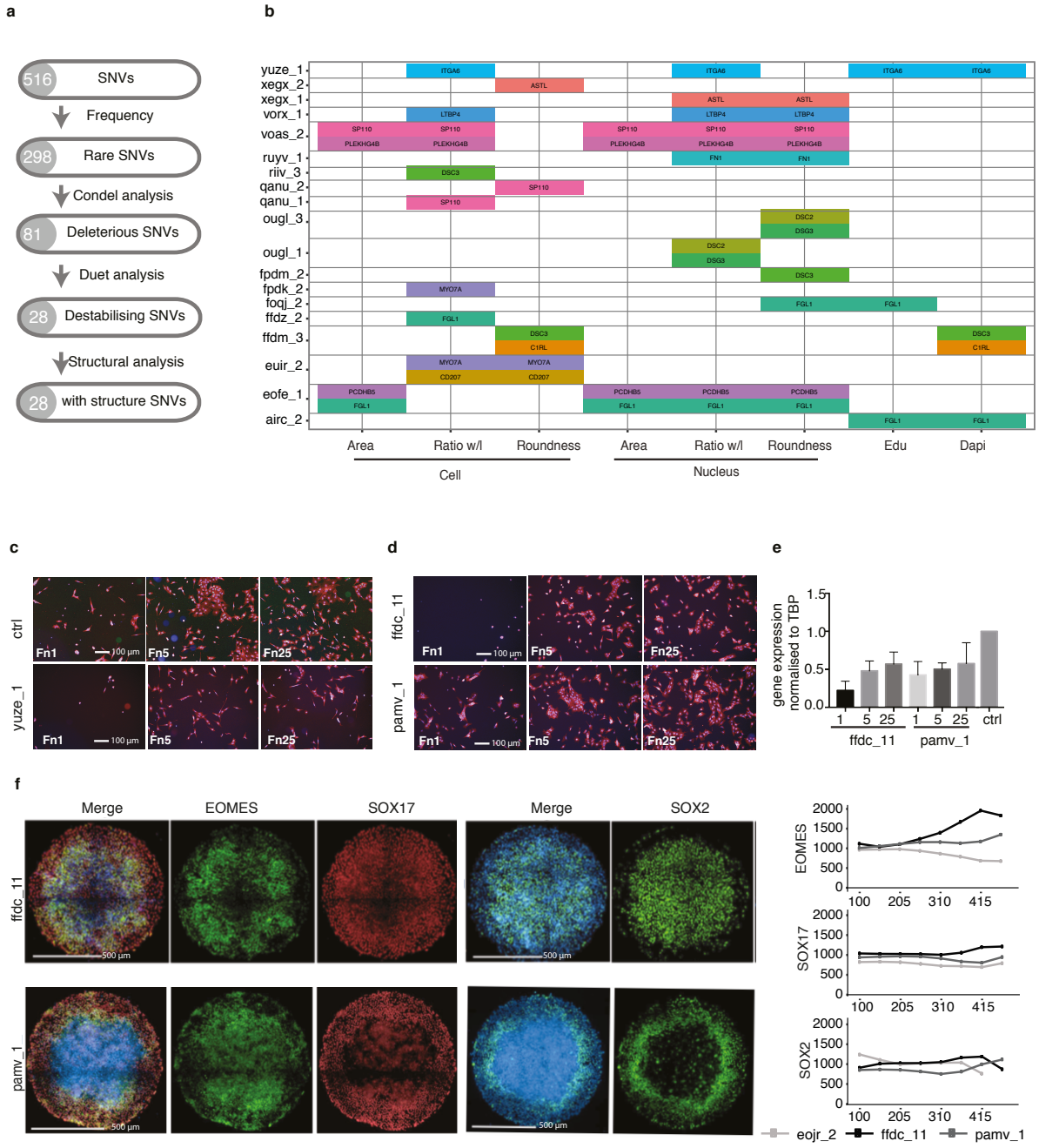

Figure 4. Identification of rare polymorphic gene variants that associate with outlier phenotypic features. a. Analysis pipeline for selection of genes. **b.** Genes with at least one rare, deleterious and destabilizing SNV in at least one cell line (y-axis) found to be an outlier for one or more phenotype (x-axis). See Extended Data Fig. 5

for outlier KS analysis. **c.** ITGA6 protein structure depicting position of disruptive SNV in extended beta-strand in yuze_1. Panels show representative images of yuze_1 cells (bottom) and cells from control line BOB (top) plated on different fibronectin concentrations (Fn1, Fn5, Fn25). **d.** Evaluation of potentially disruptive SNVs in ITGB1 and FHL2 genes in three cell lines that were not analysed in the original screen (ffdc_1 and pamv_1 respectively - pamv_3 not shown). Panels show representative images of cells plated on different fibronectin concentrations (Fn1, Fn5, Fn25). **e.** Q-PCR of ITGB1 expression normalized to TBP does not show any significant differences between ffdc_1 and cells lacking the ITGB1 SNV. **f.** Cells were plated on micropatterned islands and induced to differentiate into the three germ layers, then labelled with antibodies to EOMES (mesoderm, green), SOX17 (endoderm, red) and SOX2 (ectoderm, green), plus DAPI (DNA, blue). The mean intensity of SOX2, EOMES and SOX17 labelling as a function of distance from the island centre is shown. Analysis was performed using a two-way ANOVA test. Data represent n=3 independent experiments, each performed in duplicate, with a minimum of 8 colonies analysed per cell line per experiment. Scale bars: 100 μm (C, D), 500 μm (F).

Discussion

Genetic mapping provides an unbiased approach to discovering genes that influence disease traits and responses to environmental stimuli such as drug exposure[30]. The attractions of developing human in vitro models that reflect in vivo genetics and physiology for mechanistic studies are obvious, and include quantitation via high content image analysis and the avoidance of animal experiments. The concept that human disease-causing mutations result in alterations in cell behaviour that can be detected in culture is well established, as in the case of keratin mutations affecting the properties of cultured epidermal cells[31]. In addition, human lymphoblastoid cell lines have long been used to model genotype-phenotype relationships in healthy individuals, although limitations include the confounding effects of biological noise and in vitro artefacts such as variation in passage number and growth rate[8,9].

There has been renewed interest in applying human iPSC for pharmacogenomics, disease modelling and uncovering genetic modifiers of complex disease traits[32,33]. For example, studies with iPS cell derived neurons[34] support the 'watershed model'[35],

whereby many different combinations of malfunctioning genes disrupt a few essential pathways to result in the disease. For these reasons we decided to extend the iPSC approach in an attempt to identify genetic modifiers of cell behaviour in healthy individuals. We have recently reported that in an analysis of over 700 well-characterised human iPSC lines there is a 8-23% genetic contribution to variation in cell behaviour[5]. Our ability to detect this contribution depended on the use of simple, short-term, quantitative assays of cell behaviour; the application of multiple environmental stimuli (different concentrations of fibronectin; single cells versus cell clumps); and homogeneous starting cell populations for the assays. The concept that genetic background contributes to variability of human iPSC is supported by a number of earlier studies[13,36,37].

In order to identify the nature of the genetic contribution to variation in cell behaviour we developed new computational approaches to integrate genomic, gene expression and cell biology datasets. The first was to apply a dimensionality reduction approach, PEER, to capture variance due to extrinsic contributors (different fibronectin concentrations) and genetic concordance. This revealed a robust correlation between RNA expression and the phenotypic features in a large panel of iPSC lines, with specific RNAs associated with intrinsic or extrinsic factors. Carcamo-Orive et al. (2017)[3] also found that human iPSC lines retain a donor-specific gene expression pattern. However, in that study cells were not exposed to different environmental stimuli, and the potential regulators of variability identified, including GATA4, GATA6, EOMES, APOA2, LINC00261, FOXQ1, CER1, were not significant in our experiments.

The majority of human iPS cells we screened responded in the same way to all microenvironmental stimuli. This likely reflects canalisation, the process by which normal organs and tissues are produced even on a background of slight genetic abnormalities[38,39]. However, we did identify a number of cell lines that exhibited outlier behaviour that could not be accounted for by variation in gene expression levels, leading us to hypothesise that outlier phenotypes might instead be attributable to rare SNVs. We identified rare SNVs that were predicted to be deleterious and for which protein structural information was available. Most of the SNVs identified by this approach occurred in cell lines that were outliers for one or more phenotypes such as cell spreading. The identification of SNVs in integrin genes is of particular interest,

because integrins are highly polymorphic and some of the previously reported SNVs alter adhesive function of cancer cells[40,41]. These SNVs not only affected cells in the pluripotent state, but also altered the ability of cells to undergo ectodermal differentiation in vitro, providing proof of principle that our approach can uncover SNVs with lineage-specific effects.

In conclusion, our platform has been successful in discovering specific RNAs associated with intrinsic or extrinsic factors and SNVs that account for outlier cell behaviour. This represents a major advance in attempts to map normal genetic variation to phenotypic variation in vitro.

Materials and Methods

Cell line derivation and culture All samples were collected from skin biopsies of consented research volunteers recruited from the NIHR Cambridge BioResource (http://www.cambridgebioresource.org.uk). Human iPSC were generated from fibroblasts by transduction with Sendai vectors expressing hOCT3/4, hSOX2, hKLF4, and hc-MYC (CytoTune™, Life Technologies, Cat. no. A1377801). Cells were cultured on irradiated or Mitomycin C-treated mouse embryonic fibroblasts (MEF-CF1) in advanced DMEM (Life technologies, UK) supplemented with 10% Knockout Serum Replacement (KOSR, Life technologies, UK), 2 mM L-glutamine (Life technologies, UK) 0.007% 2-mercaptoethanol (Sigma-Aldrich, UK), 4 ng/mL recombinant Fibroblast Growth Factor-2, and 1% Pen/Strep (Life technologies, UK). Pluripotency was assessed based on expression profiling[42], detection of pluripotency markers in culture and response to differentiation inducing conditions[43]. Established iPSC lines were passaged every 3-4 days approximately at a 1:3 split ratio. The ID numbers and details for each cell line are listed in Supplementary Table 1.

Mycoplasma testing and STR profiling For mycoplasma testing 1 ml of conditioned medium was heated for 5min at 95°C. A PCR reaction was set up with the following primers: forward (5'GGGAGCAAACAGGATTAGATACCCT3'); reverse (5'TGCACCATCTGTCACTCTGTTAACCTC3'). PCR products were loaded on a 1% w/v agarose gel, run at 110 V for 30 minutes in TAE buffer and observed with Gel Dox XR+ imaging system (Bio-Rad). To confirm cell line identity, DNA extraction was

performed using the DNeasy Blood & Tissue Kit (Qiagen). Confluent cells were dissociated from 6-well plates and lysed in protein K solution; 4□l of 100mg/ml RNase solution (Qiagen) was added and DNA was purified through the provided spin-column and eluted in 150ul. DNA quality was confirmed with nanodrop spectrophotometer (Nanodrop 2000, Thermo scientific) and in 1% agarose gel. DNA samples were sequenced using STR profiling at the Wellcome Trust Sanger Institute.

Fibronectin adhesion assays 96-well micro-clearblack tissue culture plates (Greiner cat. No. 655090) were coated with three concentrations of human plasma fibronectin (Corning) in alternating columns in a randomised fashion[17]. Cells were incubated for 8 min with Accutase (Biolegend) to create a single cell suspension. As the cells began to separate and round up, pre-warmed medium containing 10 μM Rho- associated protein kinase (ROCK) inhibitor (Y-27632; Enzo Life Sciences) was added and cells were removed from culture wells by gentle pipetting to form a single cell suspension. Cells were then collected by centrifugation, aspirated and resuspended in medium containing 10 μM Rock inhibitor. Cells were counted using a Scepter 2.0 automated cell- counting device (Millipore) and seeded onto the fibronectin- coated 96-well plate using Viaflo (INTEGRA Biosciences) electronic pipettes.

Cell line plating was randomised within rows, with three wells per condition for each line to obviate edge and position effects. One control line (A1ATD-iPSC patient 1)[18], kindly provided by Tamir Rashid and Ludovic Vallier, was run as internal control in the majority of plates. For each well, 3,000 cells were plated for 24 hours prior to fixation. Paraformaldehyde 8% (PFA, Sigma–Aldrich) was added to an equal volume of medium for a final concentration of 4%, and left at room temperature for 15 min. Cells were labelled with EdU (Click-iT, Life Technologies) 30 minutes before fixation. Fixed cells were blocked and permeabilised with 0.1% v/v Triton X-100 (Sigma–Aldrich), 1% w/v bovine serum albumin (BSA, Sigma–Aldrich) and 3%v/v donkey serum (Sigma–Aldrich) for 20 min at room temperature and stained with DAPI (1 microM final concentration, Life Technologies) and cell mask (1:1000, Life Technologies). EdU was detected according to manufacturer's instructions, except that the concentration of the azide reagent was reduced by 50%.

Images were acquired using an Operetta (Perkin Elmer) high content device. Border

wells were avoided to reduce edge effects. Harmony software was used to derive measurements for each cell. Measurements included intensity features (DAPI, EdU), morphology features (cell area, cell roundness, cell width to length ratio, nucleus area, nucleus roundness, nucleus width to length ratio) and context features related with cell adhesion properties (number of cells per clump). Processing quantification and normalisation of data was performed as previously described[17].

Differentiation on micropatterned islands

96 well plates containing 1000μm diameter circular adhesive islands were prepared by UV lithography[29]. Briefly, 1000μm patterns were transferred onto Polyethylene Glycol-coated glass coverslips (110mmx74mm) by photo-oxidising select regions of the substrate using Deep UV exposure. The patterned slides were then glued to bottomless 96-well plates to produce microtitre plates with patterned cell culture surfaces. To activate the carboxyl groups on the photo-activated regions of the plates prior to cell seeding, the wells were incubated with N-(3-dimethylaminopropyl)-N'-ethylcarbodiimide hydrochloride (Sigma-Aldrich) and N-hydroxysuccinimide (Sigma-Aldrich) (20 minutes, RT). After washing with ddH_2O, the wells were coated with Geltrex (1:150) overnight at 4°C. Immediately before seeding, the wells were washed five times with PBS to remove any passively adsorbed ECM protein.

A single cell suspension of hiPSCs ($6x10^5$ cells/ml in 74% DMEM, 20% KOSR, 1% Penicillin/Streptomycin, 0.1mM β-mercaptoethanol, 1% non-essential amino acids, 1% Glutamax and 2% B27 minus retinoic acid, supplemented with 20ng/ml bFGF (all ThermoFisher) and 10μM ROCK inhibitor (ROCKi; Reagents Direct) was seeded at a density of 60,000 cells/well and incubated for 3 hours, after which the medium was replaced with SM without ROCKi. When cells had reached confluence (typically 24 hours after seeding), gastrulation-like differentiation was induced using N2B27 differentiation medium (DM) consisting of 93% DMEM, 1% Penicillin/Streptomycin, 0.1mM β-mercaptoethanol, 1% non- essential amino acids, 1% Glutamax, 2% B27 minus retinoic acid, 1% N2 supplement and supplemented with 50ng/ml BMP4, 10ng/ml bFGF and 100ng/ml NODAL (all ThermoFisher). Cells were incubated with N2B27 DM for 48 hours at 37.5°C.

At the end of the incubation period, cells were fixed in 4% paraformaldehyde for 15 minutes at room temperature and washed three times with PBS. Cells were

permeabilised with TritonX-100/PBS (0.1%) and then blocked using 5% donkey serum (1 hour, room temperature). Primary antibodies were rabbit anti-SOX2 (Cell Signalling; reference 2748), rabbit anti-EOMES (AbCam; ab23345) and goat anti-SOX17 (Santa Cruz; sc-17355). Secondary antibodies were donkey anti-rabbit Alexa Fluor 488 and donkey anti-goat Alexa Fluor 633 (Life Technologies). Primary antibodies were diluted in 1% donkey serum/PBS and applied to cells overnight at 4°C. Following three washes with PBS, cells were incubated with secondary antibodies and DAPI (1:5000) for 1 hour at room temperature.

Images were acquired using the Operetta CLS (PerkinElmer) confocal microscope with a 20x 1.0NA water objective and analysed using Harmony 4.5 software (PerkinElmer). The images shown were tiled together from 4 fields per micropatterned island. For each patterned colony the mean intensity of each channel was calculated relative to the colony centre. A pipeline was built to: identify each patterned colony; select quality controlled patterned colonies (*e.g.* circular in shape, not attached to side of wells); determine the geometrical centre; and divide each colony into 8 concentric rings radiating out from the centre. The average staining intensity for each germ layer marker in each ring was calculated and normalised relative to a control line, eojr_2. Statistical analysis (two-way ANOVA) was performed using Prism software. Results are presented as mean and standard deviation.

Dimensionality reduction approach

We applied a Bayesian factor analysis model called PEER[16] to the phenotype data in each cell line. This approach uses an unsupervised linear model to account for global variance components in the data, and yields a number of factor components that can be used as synthetic phenotype in further analysis. We tested a wide range of parameter settings for the model (the *k* number), controlling the amount of variance explained by it. We ran PEER with the full pre-normalized dataset with the following parameters: K = 9; covariates = cell line, fibronectin and batch; maximum iterations = 10,000.

Gene expression profiling

Gene expression profiles were measured with Illumina HumanHT-12 v4 Expression BeadChips and processed as described in[5]. Probe intensity estimates were normalised separately for the two cell types using the variance-stabilizing transformation implemented in the R/Bioconductor vsn package[44]. After normalisation, the datasets were limited to the final remapped set of probes (nprobes = 25,604). We refer to this version of the gexarray data by vsn log2(iPS cell/somatic). Open access gene expression array data are available in the ArrayExpress database (https://www.ebi.ac.uk/arrayexpress/) under accession number E-MTAB-4057. Correlation analyses were performed between gene expression data and both intrinsic/extrinsic factor and raw phenotypes using cor() function in R (method Spearman's).

Detection of outlier cell lines

In order to detect outlier cell lines, deviating significantly from modal values in at least one phenotypic feature, we applied a Kolmogorov-Smirnov test (ks.test() function in R) to test if cells belonging to the same cell lines behave differently compared to all the other cell lines for all the phenotypic measurements. Distributions of the D scores obtained where used to detect outlier cell lines (Extended Data Fig. 5).

Single Nucleotide Variations (SNVs) analysis

All nsSNVs identified from the "INFO_04_filtered" VCF files from the latest release of the exomeseq data, which have been filtered for higher confidence variants using Impute2, were mapped to protein sequences using ANNOVAR[45]. Those nsSNVs which mapped to genes in our set of genes were selected for further analysis.

Rare nsSNVs were defined as those with a minor allele frequency (MAF) < 0.005 in both the 1000 Genomes Project[24] and ExAC database[25]. Protein domain boundaries were obtained by scanning UNIPROT [46] protein sequences against the PFAM[47] seed libraries using HMMER[48]. UniProt proteins (with mapped nsSNVs) were assigned resolved protein structures/homologs from the PDB biounit database[49] using BLAST [50]. Hits were accepted with a sequence identity > 30% and E-value < 0.001. BLAST searches were carried out using both the entire protein sequences and domain sequences.

For each protein with mapped nsSNVs the structural homolog with the highest identity was chosen as a template for homology modelling. In the case of ties the modelling process was performed using each template. The portion of the template and query sequences relating to a BLAST hit were aligned using T-COFFEE[51]. 10 homology models for each query template alignment were created using the MODELLER software[52]. In each case the model with the lowest zDOPE score[53] was selected for further analysis. Where models were created using several templates the model with the lowest zDOPE out of all created models was selected for further analysis.

The impact of all nsSNVs were assessed using a primarily sequence-based consensus predictor of deleteriousness, Condel[27]. Where structural information was available, the impact of nsSNVs on protein structural stability was also predicted using DUET[26].

RNA extraction and real time qPCR.

Total RNA was isolated from culture cells using the RNeasy kit (Qiagen). Complementary DNA was generated using the QuantiTect Reverse Transcription kit (Qiagen). qPCR analysis of cDNA was performed using qPCR primers (published or designed with Primer3) and Fast SYBR green Master Mix (Life Technologies). RT-qPCR reactions were run on CFX384 Real-Time System (Bio-Rad). 18S rRNA was used as housekeeping gene for normalization.

mRNA isolation and qPCR

RNA was purified with Purelink RNA microkit (Invitrogen) and reverse transcribed with SuperScriptIII (Qiagen). qPCRs were performed with the RT² Profiler PCR Arrays (Growth-Factor-Array: PAMM-041ZE-1; Wnt-Signalling-Array: PAMM-243ZE-1) using RT²-SYBR-Green-qPCR-Mastermix (Qiagen). Analysis was performed by delta-Ct method, using one-way ANOVA with Geisser-Greenhouse correction and Holm-Sidak's multiple comparisons test.

Acknowledgements

We are grateful to the Wellcome Trust and MRC for funding through the Human Induced Pluripotent Stem Cell Initiative (WT098503). We also gratefully acknowledge funding from the Department of Health via the National Institute for Health Research

comprehensive Biomedical Research Centre award to Guy's & St Thomas' National Health Service Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust. AV and NML gratefully acknowledge the support of The Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001110), the UK Medical Research Council (FC001110), and the Wellcome Trust (FC001110). AV and NML were also supported by funding from a Wellcome Trust Investigator Award. Gernot Walko and Priya Viswanathan for critical review of the paper Mia Gervasio, Fatima Chowdhury and Darrick Hansen for technical support.

References:

1 Warren, C. R. *et al.* Induced Pluripotent Stem Cell Differentiation Enables Functional Validation of GWAS Variants in Metabolic Disease. *Cell Stem Cell* **20**, 547-557 e547, doi:10.1016/j.stem.2017.01.010 (2017).

2 Pashos, E. E. *et al.* Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. *Cell Stem Cell* **20**, 558-570 e510, doi:10.1016/j.stem.2017.03.017 (2017).

3 Carcamo-Orive, I. *et al.* 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, doi:10.1016/j.stem.2016.11.005 (2017).

4 DeBoever, C. *et al.* 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, doi:10.1016/j.stem.2017.03.009 (2017).

5 Kilpinen, H. *et al.* Corrigendum: Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 686, doi:10.1038/nature23012 (2017).

6 Panopoulos, A. D. *et al.* 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, doi:10.1016/j.stemcr.2017.03.012 (2017).

7 HipSci Consortium <http://www.hipsci.org/>

8 Choy, E. *et al.* Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet* **4**, e1000287, doi:10.1371/journal.pgen.1000287 (2008).

9 Jack, J., Rotroff, D. & Motsinger-Reif, A. Lymphoblastoid cell lines models of drug response: successes and lessons from this pharmacogenomic model. *Curr Mol Med* **14**, 833-840 (2014).

10 *International Society for Stem Cell Research*, <www.isscr.org> (2011).

11 Laurent, L. C. *et al.* Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell* **8**, 106-118, doi:10.1016/j.stem.2010.12.003 (2011).

12 Choi, J. *et al.* Genomic landscape of cutaneous T cell lymphoma. *Nat Genet* **47**, 1011-1019, doi:10.1038/ng.3356 (2015).
13 Kyttala, A. *et al.* Genetic Variability Overrides the Impact of Parental Cell Type and Determines iPSC Differentiation Potential. *Stem Cell Reports* **6**, 200-212, doi:10.1016/j.stemcr.2015.12.009 (2016).
14 Lane, S. W., Williams, D. A. & Watt, F. M. Modulating the stem cell niche for tissue regeneration. *Nat Biotechnol* **32**, 795-803, doi:10.1038/nbt.2978 (2014).
15 Reimer, A. *et al.* Scalable topographies to support proliferation and Oct4 expression by human induced pluripotent stem cells. *Sci Rep* **6**, 18948, doi:10.1038/srep18948 (2016).
16 Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500-507, doi:10.1038/nprot.2011.457 (2012).
17 Leha, A. *et al.* A high-content platform to characterise human induced pluripotent stem cell lines. *Methods* **96**, 85-96, doi:10.1016/j.ymeth.2015.11.012 (2016).
18 Rashid, S. T. *et al.* Modeling inherited metabolic disorders of the liver using human induced pluripotent stem cells. *J Clin Invest* **120**, 3127-3136, doi:10.1172/JCI43122 (2010).
19 Huang, D. W. *et al.* DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* **35**, W169-175, doi:10.1093/nar/gkm415 (2007).
20 Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* **44**, D336-342, doi:10.1093/nar/gkv1194 (2016).
21 Kelwick, R., Desanlis, I., Wheeler, G. N. & Edwards, D. R. The ADAMTS (A Disintegrin and Metalloproteinase with Thrombospondin motifs) family. *Genome Biol* **16**, 113, doi:10.1186/s13059-015-0676-3 (2015).
22 Runswick, S. K., O'Hare, M. J., Jones, L., Streuli, C. H. & Garrod, D. R. Desmosomal adhesion regulates epithelial morphogenesis and cell positioning. *Nat Cell Biol* **3**, 823-830, doi:10.1038/ncb0901-823 (2001).
23 Kreppel, M. *et al.* Suppression of KCMF1 by constitutive high CD99 expression is involved in the migratory ability of Ewing's sarcoma cells. *Oncogene* **25**, 2795-2800, doi:10.1038/sj.onc.1209300 (2006).
24 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
25 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).
26 Pires, D. E. V., Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research* **42**, W314-W319, doi:10.1093/nar/gku411 (2014).
27 Gonzalez-Perez, A. & Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* **88**, 440-449, doi:10.1016/j.ajhg.2011.03.004 (2011).
28 Warmflash, A., Sorre, B., Etoc, F., Siggia, E. D. & Brivanlou, A. H. A method to recapitulate early embryonic spatial patterning in human embryonic stem cells. *Nat Methods* **11**, 847-854, doi:10.1038/nmeth.3016 (2014).

29 Tewary, M. *et al.* A stepwise model of Reaction-Diffusion and Positional-Information governs self-organized human peri-gastrulation-like patterning. *Development*, doi:10.1242/dev.149658 (2017).
30 McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-369, doi:10.1038/nrg2344 (2008).
31 Tan, C. S., Cho, B. C. & Soo, R. A. Next-generation epidermal growth factor receptor tyrosine kinase inhibitors in epidermal growth factor receptor -mutant non-small cell lung cancer. *Lung Cancer* **93**, 59-68, doi:10.1016/j.lungcan.2016.01.003 (2016).
32 Barral, S. & Kurian, M. A. Utility of Induced Pluripotent Stem Cells for the Study and Treatment of Genetic Diseases: Focus on Childhood Neurological Disorders. *Front Mol Neurosci* **9**, 78, doi:10.3389/fnmol.2016.00078 (2016).
33 Kathuria, A. *et al.* Stem cell-derived neurons from autistic individuals with SHANK3 mutation show morphogenetic abnormalities during early development. *Mol Psychiatry*, doi:10.1038/mp.2017.185 (2017).
34 Brennand, K. J. & Gage, F. H. Concise review: the promise of human induced pluripotent stem cell-based studies of schizophrenia. *Stem Cells* **29**, 1915-1922, doi:10.1002/stem.762 (2011).
35 Cannon, T. D. & Keller, M. C. Endophenotypes in the genetic analyses of mental disorders. *Annu Rev Clin Psychol* **2**, 267-290, doi:10.1146/annurev.clinpsy.2.022305.095232 (2006).
36 Burrows, C. K. *et al.* Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in iPSCs. *PLoS Genet* **12**, e1005793, doi:10.1371/journal.pgen.1005793 (2016).
37 Rouhani, F. *et al.* Genetic background drives transcriptional variation in human induced pluripotent stem cells. *PLoS Genet* **10**, e1004432, doi:10.1371/journal.pgen.1004432 (2014).
38 Rutherford SL, L. S. Hsp90 as a capacitor for morphological evolution. . *Nature* **396**, 336-342 (1998).
39 McLaren, A. Signaling for germ cells. *Genes Dev* **13**, 373-376 (1999).
40 Ferreira, M., Fujiwara, H., Morita, K. & Watt, F. M. An activating beta1 integrin mutation increases the conversion of benign to malignant skin tumors. *Cancer Res* **69**, 1334-1342, doi:10.1158/0008-5472.CAN-08-3051 (2009).
41 Evans, R. D. *et al.* A tumor-associated beta 1 integrin mutation that abrogates epithelial differentiation control. *J Cell Biol* **160**, 589-596, doi:10.1083/jcb.200209016 (2003).
42 Muller, F. J. *et al.* A bioinformatic assay for pluripotency in human cells. *Nat Methods* **8**, 315-317, doi:10.1038/nmeth.1580 (2011).
43 Robinton, D. A. & Daley, G. Q. The promise of induced pluripotent stem cells in research and therapy. *Nature* **481**, 295-305, doi:10.1038/nature10761 (2012).
44 Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96-104 (2002).
45 Wang, K., Li, M., Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Research* **38** (2010).
46 The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

47 Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279-285, doi:10.1093/nar/gkv1344 (2016).

48 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29-37, doi:10.1093/nar/gkr367 (2011).

49 Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**, 980, doi:10.1038/nsb1203-980 (2003).

50 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

51 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217, doi:10.1006/jmbi.2000.4042 (2000).

52 Webb, B. a. S., A. . Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* **54**, 5.6.1-5.6.37 (2016).

53 Shen, M. Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15**, 2507-2524, doi:10.1110/ps.062416606 (2006).