Identifying extrinsic versus intrinsic drivers 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[1,9], Andreas Leha[3,10], Arsham Ghahramani[1,4], Oliver J. Culley[1], Annie Kathuria[1], Chloe Hurling[1], Alice Vickers[1], Erika Wiseman[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, UK

[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, UK

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

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

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

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

Abstract

Large cohorts of human induced pluripotent stem cells (iPSCs) from healthy donors are a potentially powerful tool for investigating the relationship between genetic variants and cellular phenotypes. Here we integrate high content imaging, gene expression and DNA sequence datasets from over 100 human iPSC lines to explore the genetic basis of inter-individual variability in cell behaviour. By applying a dimensionality reduction approach, Probabilistic Estimation of Expression Residuals (PEER), we extracted factors that captured the effects of intrinsic (genetic) and extrinsic (environmental) conditions. We identified genes that correlated in expression with intrinsic and extrinsic PEER factors and mapped outlier cell behaviour to expression of genes containing rare deleterious SNVs. Our study thus establishes a strategy for determining the genetic basis of inter-individual variability in cell behaviour.

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. These offer a unique opportunity 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 (HipSci)[5,7]. In addition to creating a comprehensive reference map of common regulatory variants affecting the transcriptome of hiPSC, we performed quantitative assays of cell morphology and demonstrated a donor contribution in the range of 8-23% to the observed variation. In the present study we set out to identify causative genetic variants.

Previous attempts using 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 proliferation 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 hiPSC lines from healthy volunteers, including multiple lines from the same donor. In addition, HipSci 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 hypothesised that subjecting cells to different environmental stimuli increases 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 support cell spreading to differing extents, and assayed the behaviour of single cells and cells in contact with their neighbours. We took a ‘cell observatory’ approach, using high-throughput, high content imaging to gather data from millions of cells 24 hours after seeding. We then used a multidimensional reduction method, Probabilistic Estimation of Expression Residuals (PEER)[16] to reveal underlying structure in the dataset, and correlated cell behaviour with expression of a subset of genes and 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 exploring 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 derived from different clones from the same donor. Seven lines came from 7 individuals with Bardet-Biedl Syndrome and 3 were non-HipSci control lines. Out of the total, 102 of the lines were derived from skin fibroblasts, 2 from hair follicles and 6 from peripheral blood monocytes. All lines were subjected to 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 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 described previously[17]. Cells were disaggregated and resuspended in the presence of 10 µM Rho-associated protein kinase (ROCK) inhibitor to minimise cell clumping. In order to vary the extrinsic 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), with Fn1 representing a suboptimal concentration for cell attachment and spreading. After 24h of culture in the presence of ROCK inhibitor, cells were labelled with EdU for 30 min (to detect proliferative cells), fixed, and stained with DAPI (to visualise nuclei) and CellMask (to visualise cytoplasm) (Figure 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 line, previously reported as A1ATD-iPSC patient 1[18], was included to control for biological variation between experiments.

We extracted a total of 11 measurements, 10 per cell (*i.e.* object) plus the number of cells per well for approximately 2 million cells (Figure 1B-C). Cell features included the derived nucleus/cell area and clump size, a context feature allowing us to distinguish between cells that had attached as individuals and cells in groups (Figure 1B-C). 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 (Figure 1B). The phenotypic features were processed as described previously[17], *i.e.* well-based 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 pixel intensity raw values per cell were used to extract a well-based measure of the fraction of EdU positive cells[17]. This resulted in a final list of 52 features including clumps (Supplementary Table 2).

The scale and complexity of the cell phenotype dataset is illustrated in Figure 1D, in which the mean value of cell area is represented for all cell lines, for three fibronectin concentrations and three biological replicates. This highlights the variance we observed between replicate experiments. It also shows a consistent effect of fibronectin concentration, with cells exhibiting greater cell area on the highest concentration (see also Figure 1C). It reveals some variability for cell lines derived from the same donor, denoted by a common 4 letter code (Figure 1D). Similar results were obtained for other raw phenotypic features (Supplementary Figure 2).

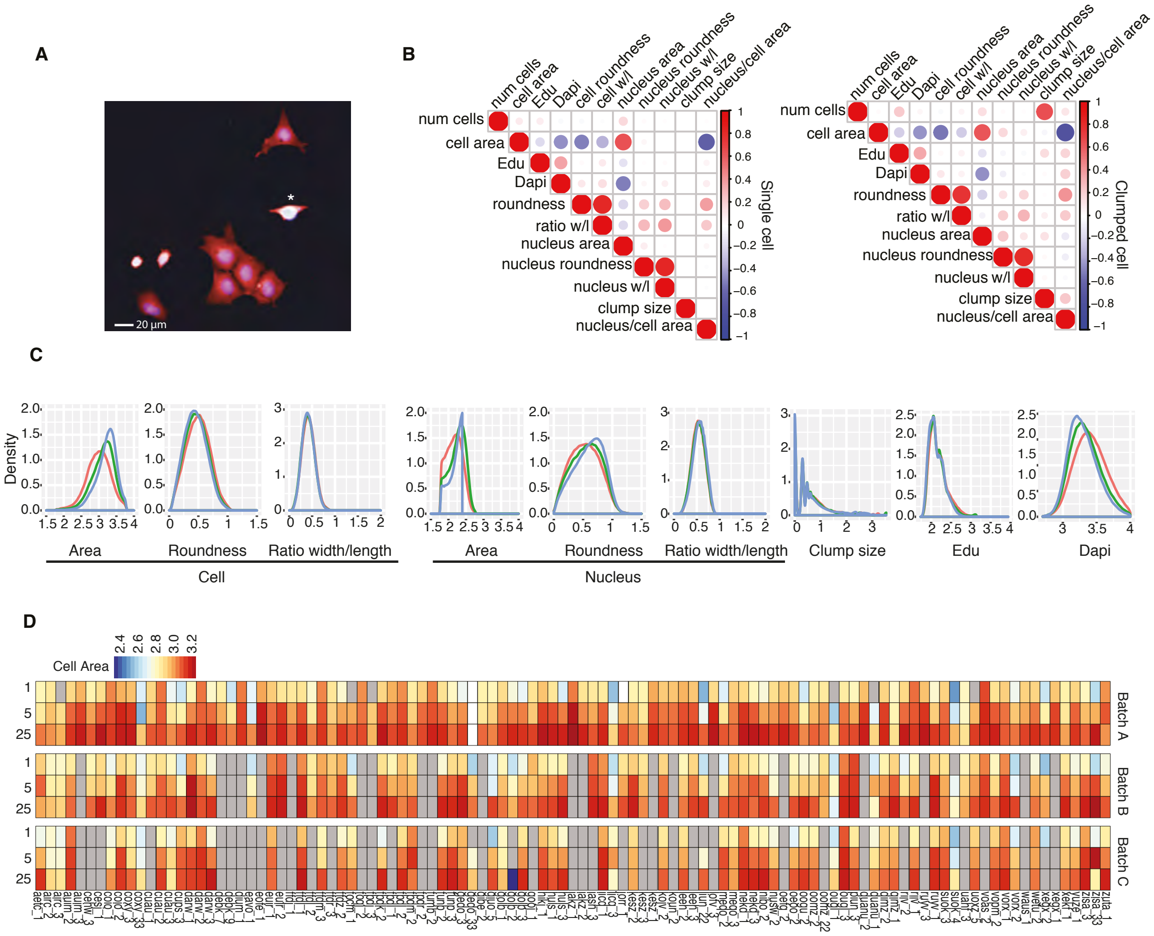


Figure. 1. Description of phenotypic dataset. (A) Microscopic image showing cells 24h after plating. Red: cell mask (cytoplasm); white: EdU incorporation (DNA synthesis, one EdU+ cell marked with asterisk); blue: DAPI (nuclei). Scale bar: 20 µm. **(B)** Correlation of different phenotypic measurements in single (left) and clumped

(right) cells. **(C)** Distribution of main phenotypic features of all cell lines on three fibronectin concentrations (Fn1, red; Fn5, green; Fn25, blue). Y axis: density. **(D)** Heatmap of mean cell area measurements for each cell line on three fibronectin concentrations in three independent experiments. Grey boxes correspond to data unavailable or not satisfying the minimum number of cells per well.

Identification of outlier cell lines

We next identified outlier cell lines, defined as lines that deviated significantly from modal phenotypic values. To do this we performed a Kolmogorov–Smirnov test of the distributions of each raw phenotypic feature for each individual cell line compared to all cell lines (Figure 2) and defined outliers as cell lines with a statistic (D) value above the 95th percentile. Out of the 110 lines analysed, 36 lines from 30 donors exhibited outlier behaviour for one or more phenotypic feature (Figure 2).

In support of a genetic contribution to outlier cell behaviour, in several cases two independent lines from the same donor exhibited the same outlier behaviour. For example iakz_1 and iakz_2 were outliers for cell roundness, while airc_2 and airc_3 were outliers for DAPI nuclear staining intensity. In addition, where two phenotypes were positively or negatively correlated (*e.g.* cell area and cell roundness) some cell lines were outliers in both categories (*e.g.* iakz_1).

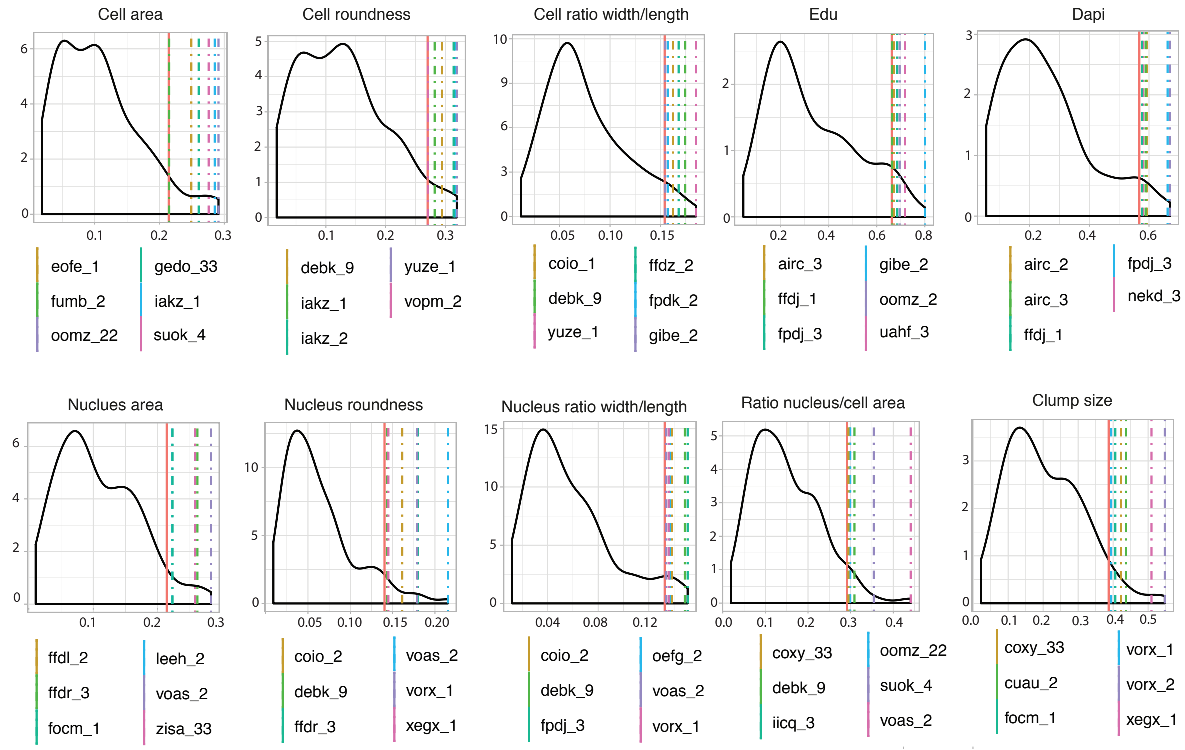


Figure 2. Identification of outlier cell lines for individual phenotypes.

The distribution of the Kolmogorov-Smirnov statistic (D) obtained by performing the Kolmogorov–Smirnov test of the distributions of each raw phenotypic feature for each individual cell line compared to all cell lines. 95th percentile threshold is shown as a red line together with values of individual outlier lines (colour coded).

Contribution of intrinsic and extrinsic factors to variation in cell phenotypes

In order to explore how extrinsic (i.e. different fibronectin concentrations), intrinsic (i.e. cell line or donor specific) and technical or biological components (covariates) contributed to the observed variation in cell phenotypes, we applied a dimensionality reduction approach called Probabilistic Estimation of Expression Residuals (PEER)[16]. PEER was originally implemented for gene expression data and to our knowledge has not been applied previously to multidimensional reduction of phenotypic data. It is a collection of Bayesian approaches that takes as input measurements (i.e. expression value) and covariates and then extract factors that explain hidden portions of the variability. This is fundamental because unobserved, hidden factors, such as cell culture conditions can have an influence on large numbers of cells and many studies have demonstrated the importance of accounting for hidden factors to achieve a stronger statistical discrimination signal[54 55 56].

In our analysis, PEER takes the 52 phenotypic measurements and covariates (*i.e.* fibronectin concentration, experimental replicates, individual donors) as inputs and outputs hidden factors that explain portions of the variance and can be treated as new synthetic phenotypes. Using the 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 (Supplementary Figure 3). For a single factor (PEER factor 1) the effect of fibronectin concentration was apparent and significant between each condition (Figure 3A and Supplementary Figure 4). This factor, which we named the 'extrinsic factor', accounted for ~30% of the total variance (Figure 3A, C, E and Supplementary Figure 3). This contrasted with Principal Component Analysis for which the effects of Fibronectin concentrations was apparent on both PC1 (27.7%) and PC2 (15.3%). Figure 3E shows that the mean of the total number of attached cells, cell and nucleus area and DAPI staining intensity have the largest effect onto the extrinsic factor.

To evaluate whether any PEER factor(s) captured intrinsic variance, we compared their values in cases where more than one cell line from the same donor was available. PEER factor 9 showed no effect of fibronectin concentration (Figure 3B) but the highest genetic concordance (Figure 2D). This factor (which we named the 'intrinsic factor') accounted for 5% of the total variance (Supplementary Figure 3). Phenotypic features describing EdU labelling and other nuclear properties, both in single and clumped cells, loaded onto the intrinsic factor (Figure 3F).

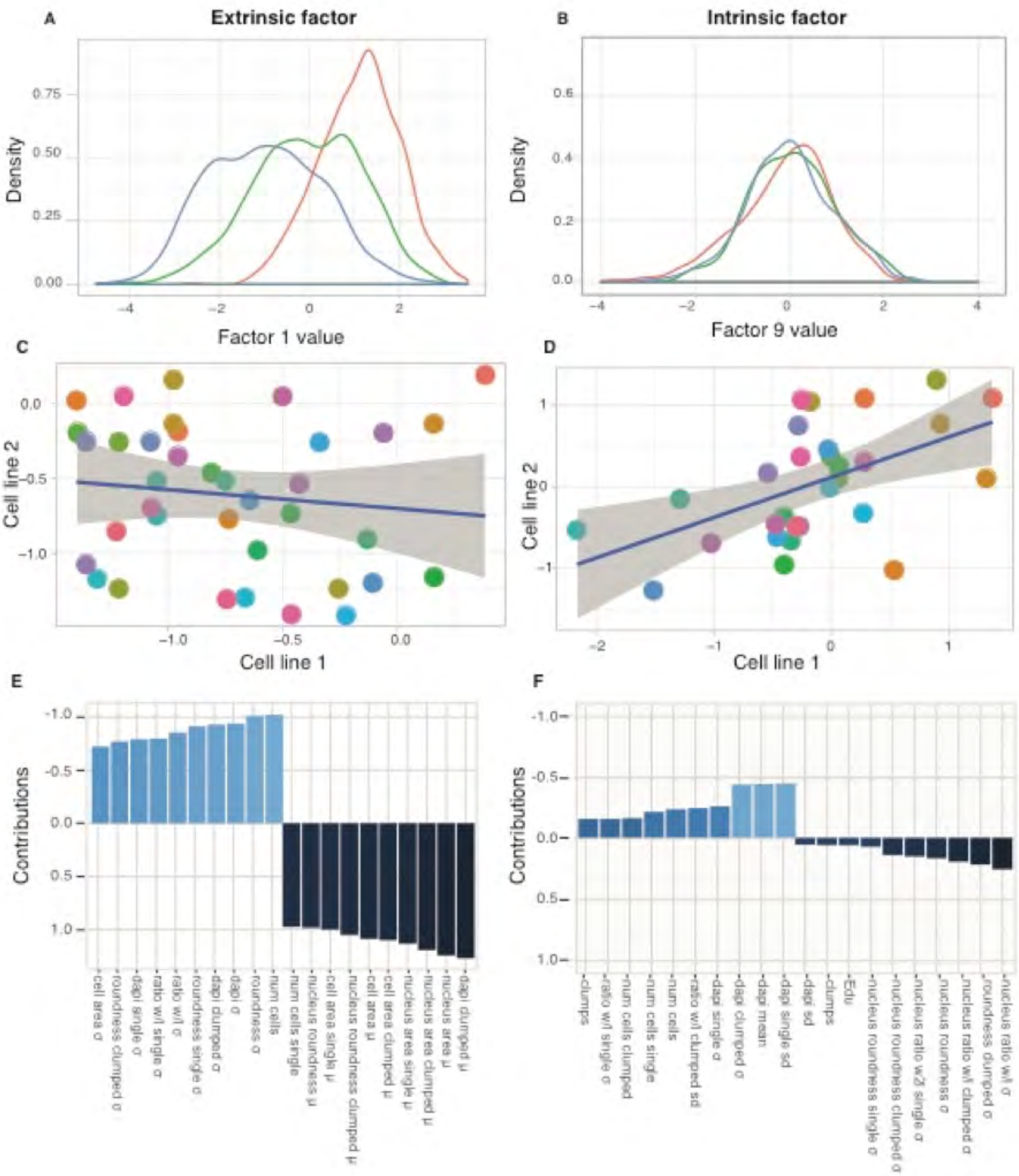


Figure 3. Synthetic phenotypic features capture extrinsic and intrinsic contributions to variance. (A-B-C-D) Plots showing the distribution of values for PEER factor 1 'extrinsic' **(A)** and PEER factor 9 'intrinsic' **(B)** The effect of conditions (Fn1, red; Fn5, green; Fn25, blue;) is visible. The genetic concordance between two clones of cells from the same donors is shown for both factors **(C,D)** and the loading of phenotypic features is detailed **(D,E).**

]

Identification of genes correlating with extrinsic and intrinsic variation

To identify genes whose expression could contribute to phenotypic variance we performed a correlation analysis between the intrinsic and extrinsic factors and gene expression array data independently generated from cell pellets as part of the HipSci resource[7]. Expression of 4573 genes correlated with either the extrinsic or intrinsic factor or both, in at least one fibronectin concentration. From this list, we filtered out genes that were not associated with any Ensembl identifiers. We also removed genes for which multiple probes showed opposite correlation values. The resulting dataset consisted of 3879 genes (Supplementary Table 2), 1321 correlating with the extrinsic factor, 1977 correlating with Intrinsic Factor and 581 with both (Figure 4A).

GO analysis was performed on genes correlating with either or both PEER factors at a threshold value of ± 0.2 of the correlation coefficient (Figure 4B). GO terms associated with the extrinsic factor included cell adhesion and receptor serine/threonine kinase signalling. Terms associated with the intrinsic factor included cell proliferation, response to stress and integrin-mediated cell adhesion. Only two GO terms were associated with both intrinsic and extrinsic factors: membrane organisation and cell-matrix adhesion.

Based on the phenotypes measured in our study, we further filtered the genes according to the functions of their protein products. We selected 175 genes we reasoned were likely to be involved in the distinct cell behaviours observed, belonging to six GO categories describing cell-cell adhesion, regulation of cell cycle and regulation of cell proliferation, cell matrix adhesion, membrane organization and transmembrane receptor signalling pathways. The expression of 98 out of the 175 genes showed a statistically significant correlation with the original phenotypic measurements depicting cell morphology, adhesion and proliferation.

Examples of gene expression variation among cell lines for genes correlating with one, two, three and four phenotypic features are shown in Figure 4C. We noted that most genes showed distinct correlations with the intrinsic and extrinsic PEER factors (Fig. 4C). In addition, opposite correlations were found for a given gene and one or more phenotypes. For example, ITGAL, which mediates intercellular adhesion, was positively correlated with clumping and negatively correlated with proliferation.

38 out of 175 genes showed outlier expression in one or more cell lines (5th and 95th percentiles) (Figure 4D). Almost all of them (32 out of 38) were outliers in outlier cell lines (Figure 4E and Supplementary Table 5). The only outlier gene exclusively associated with the extrinsic PEER factor was SRC, Proto-oncogene tyrosine-protein kinase. In cases in which two cell lines from the same donor were outliers for the same raw phenotypic features (Figure 2), this could not be explained by the overexpression or lack of expression of the same set of genes.

In conclusion, we could identify a set of genes known to regulate cell phenotypic features and observed a correlation between mRNA expression in cell pellets and cell phenotypic features on three concentrations of Fibronectin. Nonetheless, the expression of specific mRNAs was not correlated with outlier cell behaviour.

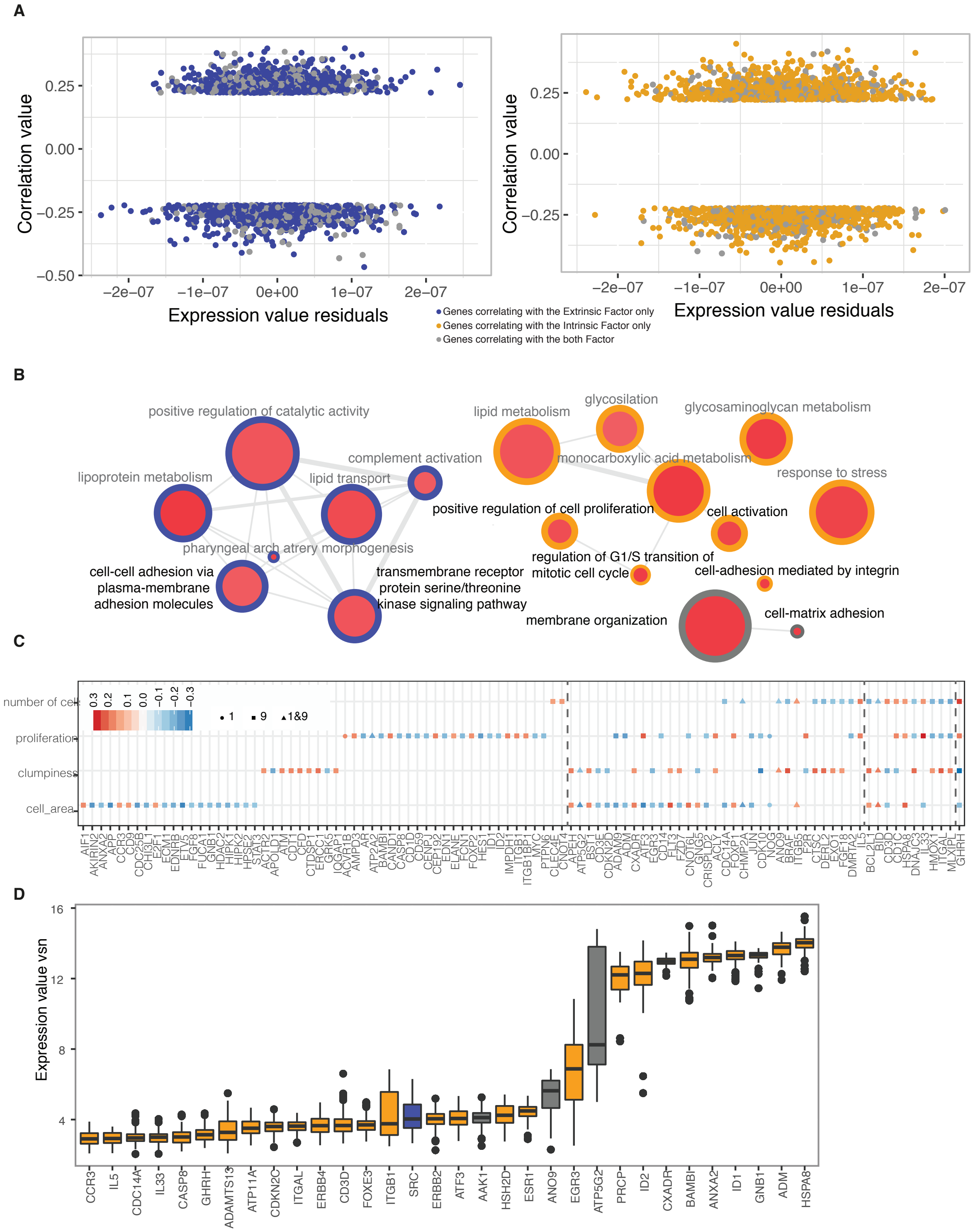

A
Correlation value
0.25
0.00
−0.25
−0.50
−2e−07
−1e−07
0e+00
1e−07
2e−07
Expression value residuals
● Genes correlating with the Extrinsic Factor only
● Genes correlating with the Intrinsic Factor only
● Genes correlating with the both Factor
B
positive regulation of catalytic activity
lipid metabolism
glycosilation
glycosaminoglycan metabolism
complement activation
monocarboxylic acid metabolism
lipoprotein metabolism
lipid transport
response to stress
positive regulation of cell proliferation
cell activation
pharyngeal arch atrery morphogenesis
regulation of G1/S transition of mitotic cell cycle
cell-cell adhesion via plasma-membrane adhesion molecules
transmembrane receptor protein serine/threonine kinase signaling pathway
cell-adhesion mediated by integrin
membrane organization
cell-matrix adhesion
C
number of cells
proliferation
clumpiness
cell_area
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
• 1
▪ 9
▴ 1&9
AIF1 AKIRIN2 ANXA2 APP CCR3 CD9 CDC25B CHI3L1 E2F1 ECM1 EDNRB ETV5 FGF8 FUCA1 GNB1 HDAC2 HIPK1 HIPK2 HPSE2 STAT3 ACTR2 APOLD1 ATM CDT1 CFD CTDSP1 ERCC1 GRK5 IQGAP1 ACVR1B AMPD3 AR ATP2A2 BAMBI CAND1 CASP8 CD1D CD59 CENPJ CEP192 EDN1 ELANE FCN1 FOXP2 HES1 ID1 ID2 IMPDH1 ITGB1 ITGB1BP1 MYC PTPN6 CLEC4E CNOT4 APEH ATP5G2 BST1 CD3E CDKN2D ADAM9 ADM CXADR ATF3 EGR3 CD14 FLT3 FZD7 CNOT6L GNG5 CRISPLD2 ACLY CDC14A FOXP1 CHMP2A JUN CDK10 NOG ANO9 BRAF ITGB5 F2R CTSC DERL2 EXO1 FGF18 DMRTA2 IL5 BCL2L1 BID CD3D CD1C HSPA8 DNAJC3 IL33 HMOX1 ITGAL MLXIPL GHRH
D
Expression value vsn
16
12
8
4
CCR3 IL5 CDC14A IL33 CASP8 GHRH ADAMTS13 ATP11A CDKN2C ITGAL ERBB4 CD3D FOXE3 ITGB1 SRC ERBB2 ATF3 AAK1 HSH2D ESR1 ANO9 EGR3 ATP5G2 PRCP ID2 CXADR BAMBI ANXA2 ID1 GNB1 ADM HSPA8

Figure 4. Using the 'extrinsic' and 'intrinsic' factors to identify genes associated with specific outlier phenotypes. (A) A total of 3879 genes correlated with either the 'extrinsic' PEER factor (a, blue spots), the 'intrinsic' factor (b, yellow spots) or both (a, b, grey spots). (**B)** GO analysis of genes correlating with either or both PEER factors. Circle colours as in (a). Node size (red fill) indicates the p-value associated to each term. 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. Terms with a black description have been used to select the list of 175 genes. (**C)** In a total of 98 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). **(D)** Boxplots showing the expression values (vsn) of 32 out of the 38 genes with outlier gene expression in one or more outlier cell lines. Color code as in (a).

Identification of SNVs in cell adhesion genes that correlate with outlier cell phenotypes

To complement the analysis of gene expression variation, we explored whether the presence of Single Nucleotide Variants (SNVs) in gene exons affecting protein function could account for outlier cell behaviour. We observed that donors for which one line was not selected further in the HipSci project for genomic analysis tended to show less genetic concordance (see Figure 2D).

We searched for SNVs in the 3879 genes identified with the extrinsic and intrinsic PEER factors (Supplementary Table 6). Of the 9831 SNVs identified, 4339 were classified as rare, based first on the 1000 Genomes Project [24] and ExAC[25], and second on the frequency in our cell lines (present in fewer than 5 out of 110 lines, Figure 5A). We further filtered the SNVs using the computational model DUET[26] to predict SNVs that would be deleterious and the computational model Condel[27] to predict a final list of 95 rare, deleterious and destabilising SNVs that would impair protein structure. The genes that we identified (Supplementary Table 4) encoded proteins that were associated with cell adhesion, including integrins, cytoskeleton and ECM proteins.

33 out of 95 rare, deleterious and destabilising SNVs identified occurred in cell lines that were outliers for one or more phenotypes (Figure 5B and Figure supplement 5). An interesting example is the presence of deleterious and destabilising SNVs in the integrin ITGA6 (position 27) in the cell line yuze_1, detected as an outlier for cell and nucleus ratio width/length and for EdU and DAPI intensity. These cells showed reduced spreading (enhanced cell roundness) particularly on the lowest concentration of Fibronectin (Figure 5C 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 in the HipSci resource 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, which like the ITGA6 SNV in yuze_1, maps to the ligand binding domain. When plated on Fibronectin, the ffdc_11 line exhibited reduced attachment and spreading on the lowest Fibronectin concentration (Figure 5C), indicative of the predicted outlier phenotype.

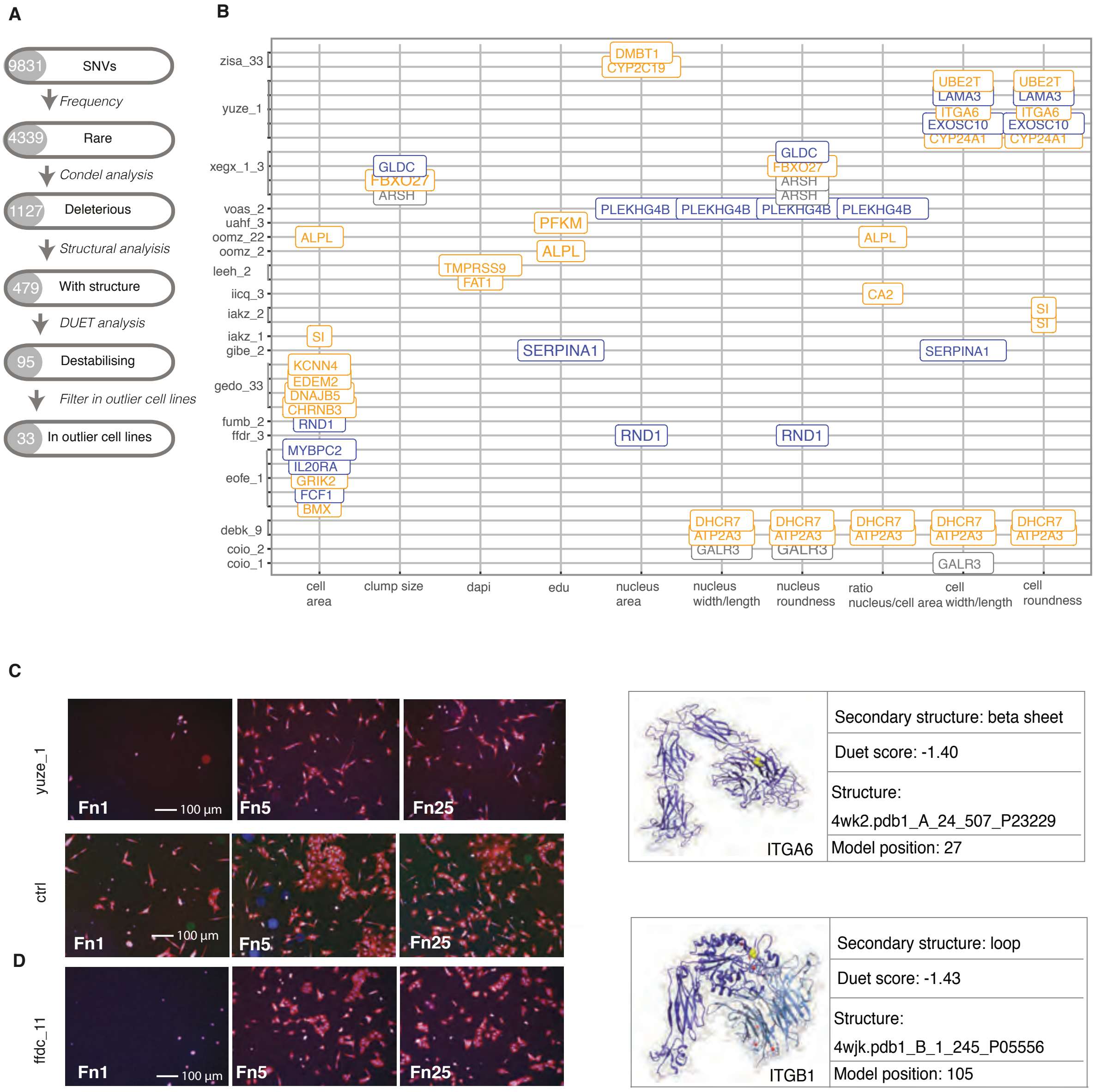


Figure 5. (A) Analysis pipeline for selection of genes. (**B)** Genes with at least one rare, deleterious and destabilizing SNV in at least one cell line (y-axis) found to be an outlier for one or more phenotype (x-axis). See Figure 2 for outlier KS analysis. **(C)** Panels show representative images of outlier cell line yuze1, with SNP in ITGA6 control cell line Bob and line ffdc11 harbouring a SNP in ITGB1, right. 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 a cell lines not analysed in the original screen (ffdc_11) harbouring a potentially disruptive

SNVs in ITGB1 gene. Panels show representative images of cells plated on different fibronectin concentrations (Fn1, Fn5, Fn25).

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 replacement 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, differentiation state as these cells do not self-renew in vitro 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 an 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.

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 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 genetic variants. 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 their ability 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 associating specific RNAs with intrinsic or extrinsic factors and discovering SNVs that account for outlier cell behaviour. This represents a major advance in attempts to map normal genetic variation to phenotypic variation.

Materials and Methods

Cell line derivation and culture All HipSci 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^{TM}, 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; 4μl of 100mg/ml RNase solution (Qiagen) was added and DNA was purified through the provided spin-column and eluted in 150μl. 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-clear-black 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. The raw Image data are available in the Image Data Resource <placeholder idr0037?> following on from the previous dataset http://dx.doi.org/10.17867/10000107.

Harmony v4.1 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].

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 https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-4057/. PEER analysis was performed taking as input the vsn expression values with the following parameters: K = 36; covariates = cell line and batch; maximum iterations = 10,000. The residual gene expression matrix was used to perform a correlation analysis with both intrinsic/extrinsic factor and raw phenotypes using cor() function in R (method Spearman's). A p-value threshold < 0.05 was applied to select the statistical significant correlations and the cut-off of the correlation values was set to ± 0.2.

Gene Ontology analysis

Gene Ontology analysis was performed using the Gorilla web-service (http://cbl-gorilla.cs.technion.ac.il/) and the output has been visualised with ReviGO (http://revigo.irb.hr/). Three analyses were performed separately for the genes correlating with the extrinsic factor, the intrinsic factor and both factors.

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].

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. We thank members of the Watt lab and Jeremy Green for scientific discussions and Mia Gervasio, Sabrina Munir, Ayaulim Nurgozhina, Fatima Chowdhury, Darrick Hansen, Zuming Tang, Christopher Sibley-Allen and Fran Molina 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).

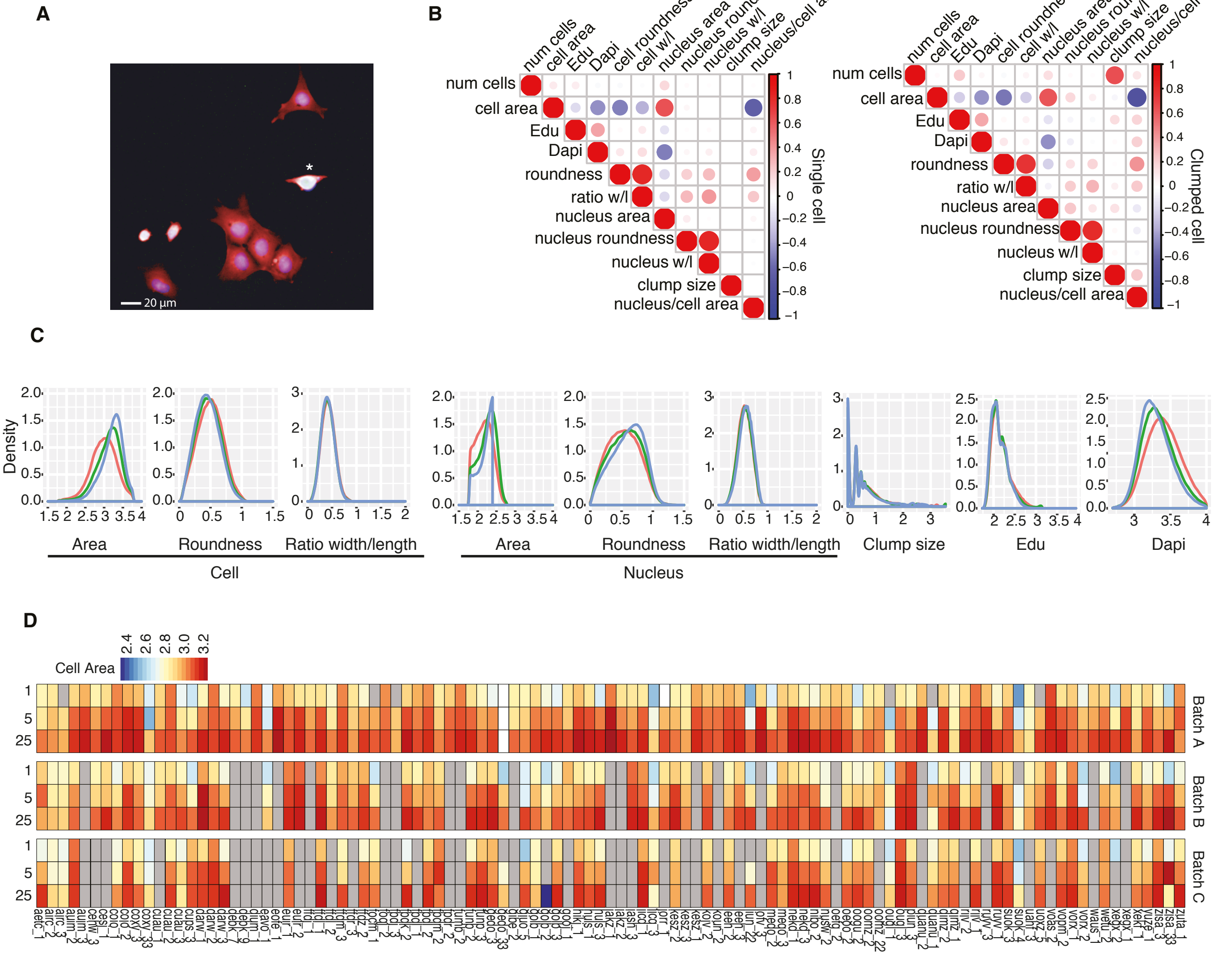
52 Webb, B. a. S., A. . Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* **54**, 5.6.1-5.6.37 (2016).

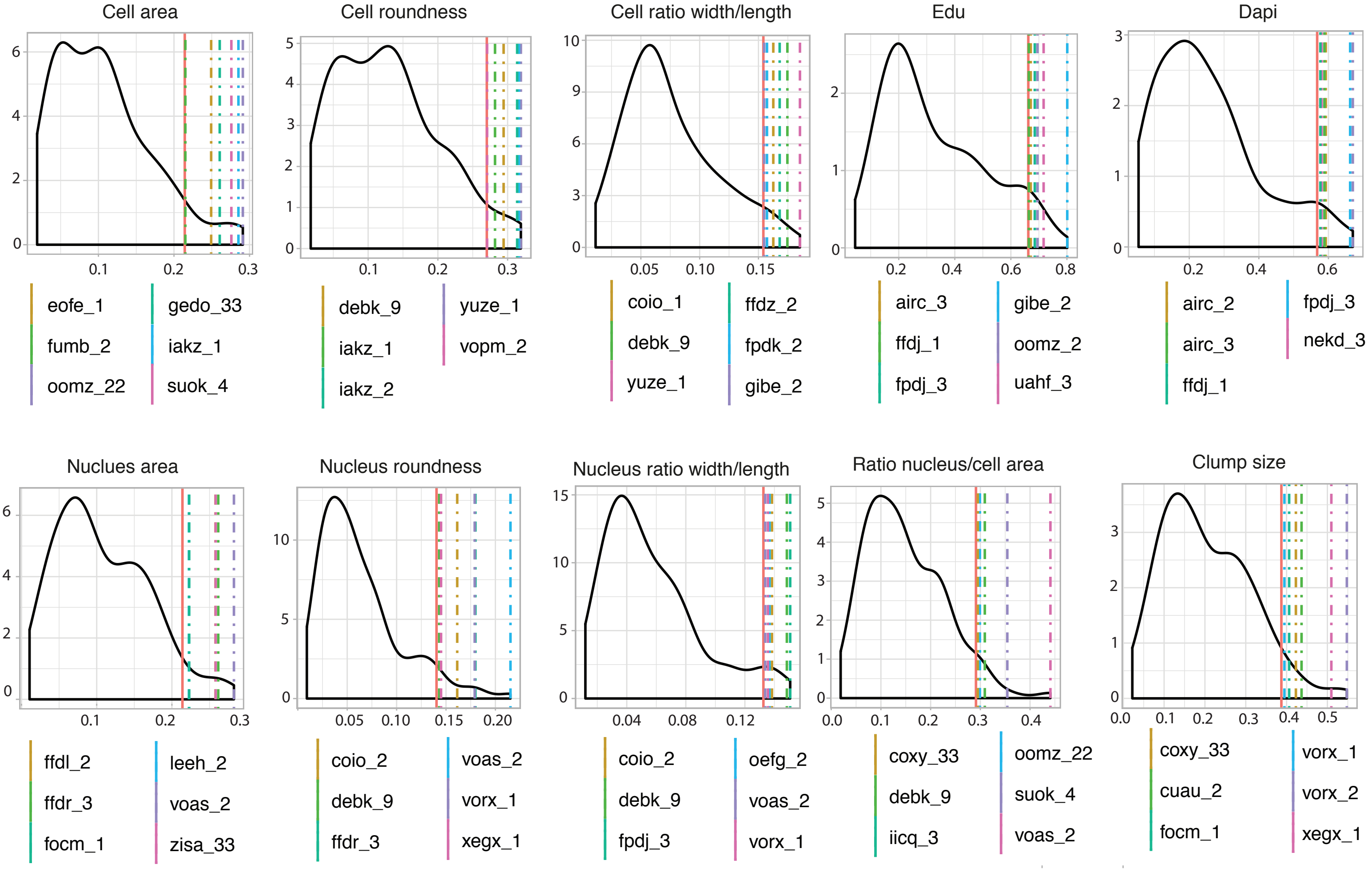
53 Shen, M. Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15**, 2507-2524, doi:10.1110/ps.062416606 (2006).

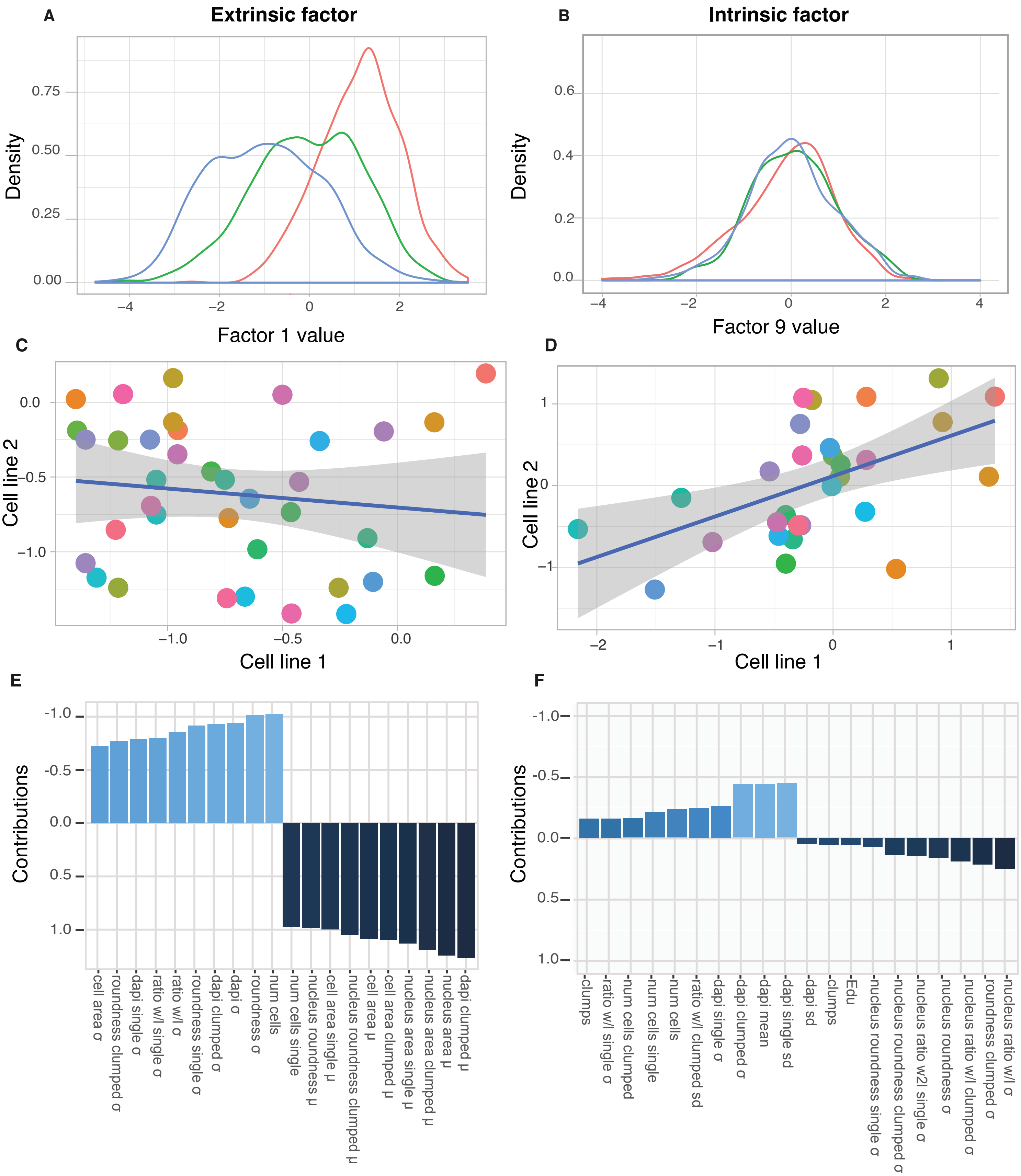
54 Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 9, 1724-35 (2007)

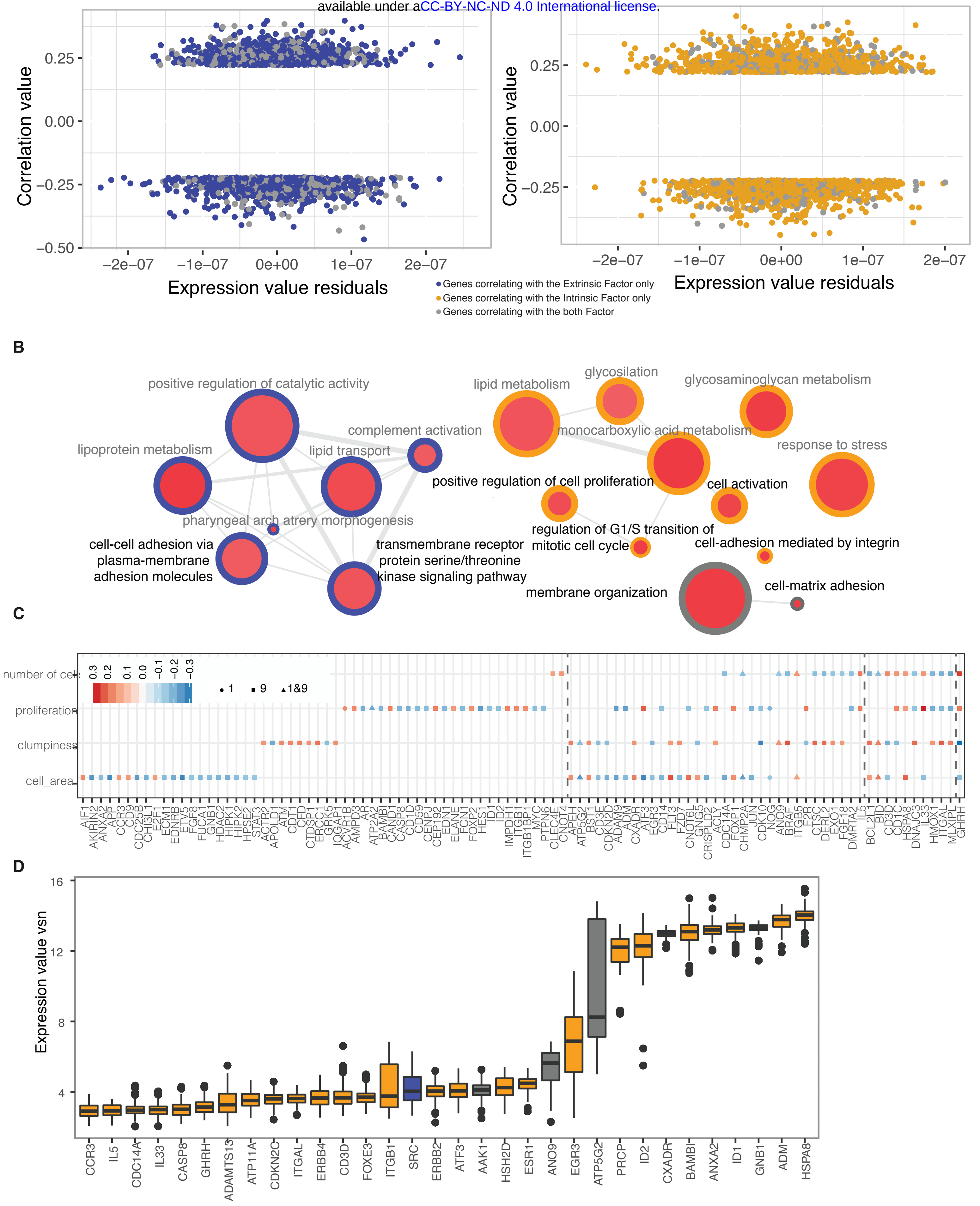
55 Stegle O, Kannan A, Durbin R, Winn J Accounting for non-genetic factors improves the power of eQTL studies. Research in Computational Molecular Biology, 411–422 (2008)

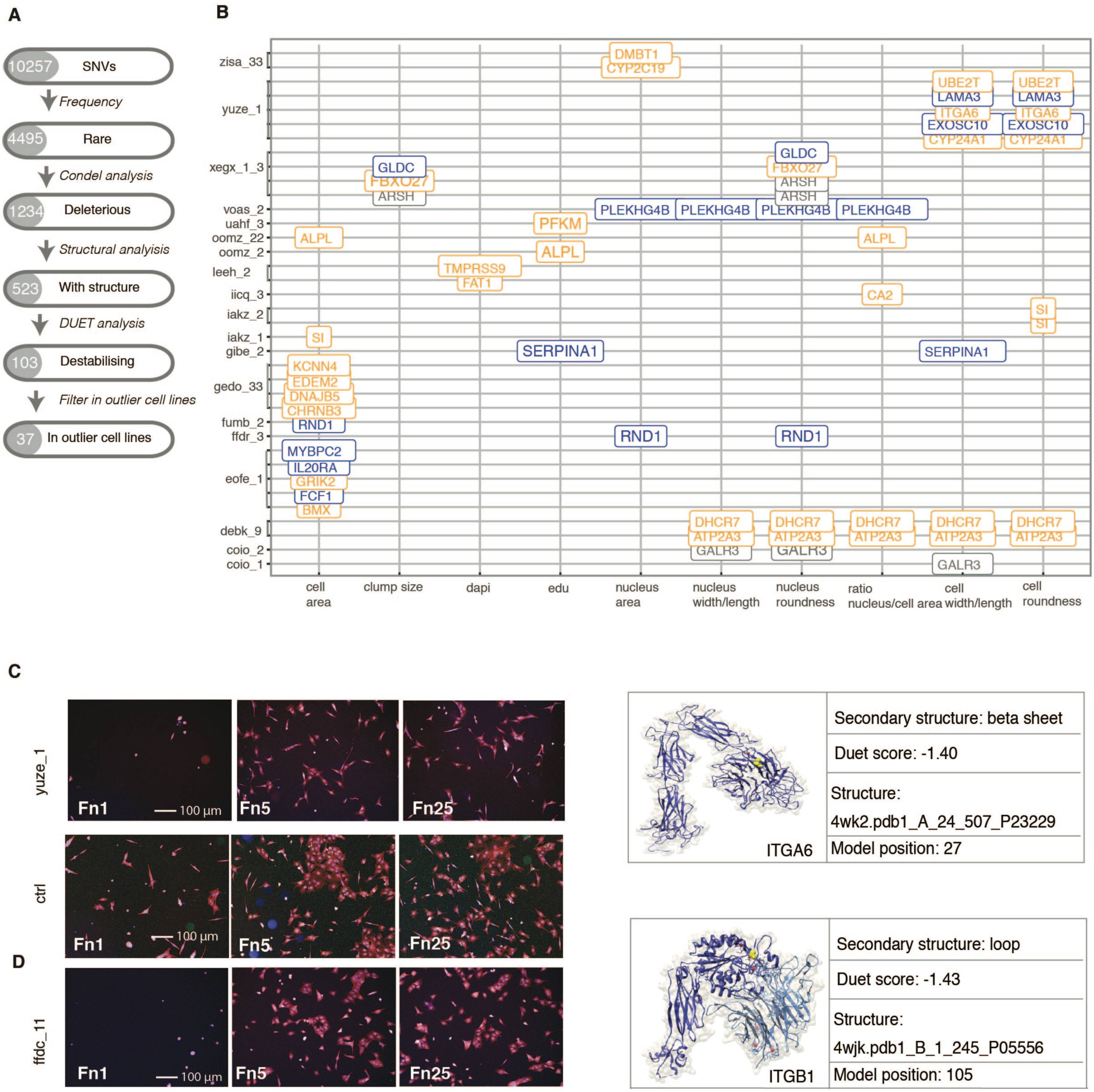
56 Kang HM, Ye C, Eskin E Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics 180, 1909–1925 (2008)


A
B
C
D
20 µm
*
num cells
cell area
Edu
Dapi
cell roundness
cell w/l
nucleus area
nucleus roundness
nucleus w/l
clump size
nucleus/cell area
roundness
ratio w/l
Single cell
Clumped cell
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
Density
Area
Roundness
Ratio width/length
Cell
Nucleus
Clump size
Edu
Dapi
Cell Area
2.4
2.6
2.8
3.0
3.2
1
5
25
Batch A
Batch B
Batch C


Cell area
0
2
4
6
0.1
0.2
0.3
eofe_1
fumb_2
oomz_22
gedo_33
iakz_1
suok_4
Cell roundness
0
1
2
3
4
5
0.1
0.2
0.3
debk_9
iakz_1
iakz_2
yuze_1
vopm_2
Cell ratio width/length
0
3
6
9
10
0.05
0.10
0.15
coio_1
debk_9
yuze_1
ffdz_2
fpdk_2
gibe_2
Edu
0
1
2
0.2
0.4
0.6
0.8
airc_3
ffdj_1
fpdj_3
gibe_2
oomz_2
uahf_3
Dapi
0
1
2
3
0.2
0.4
0.6
airc_2
airc_3
ffdj_1
fpdj_3
nekd_3
Nuclues area
0
2
4
6
0.1
0.2
0.3
ffdl_2
ffdr_3
focm_1
leeh_2
voas_2
zisa_33
Nucleus roundness
0
5
10
0.05
0.10
0.15
0.20
coio_2
debk_9
ffdr_3
voas_2
vorx_1
xegx_1
Nucleus ratio width/length
0
5
10
15
0.04
0.08
0.12
coio_2
debk_9
fpdj_3
oefg_2
voas_2
vorx_1
Ratio nucleus/cell area
0
1
2
3
4
5
0.0
0.1
0.2
0.3
0.4
coxy_33
debk_9
iicq_3
oomz_22
suok_4
voas_2
Clump size
0
1
2
3
0.0
0.1
0.2
0.3
0.4
0.5
coxy_33
cuau_2
focm_1
vorx_1
vorx_2
xegx_1


A
Extrinsic factor
Density
0.75
0.50
0.25
0.00
−4
−2
0
2
Factor 1 value
B
Intrinsic factor
Density
0.6
0.4
0.2
0.0
−4
−2
0
2
4
Factor 9 value
C
Cell line 2
0.0
−0.5
−1.0
−1.0
−0.5
0.0
Cell line 1
D
Cell line 2
1
0
−1
−2
−1
0
1
Cell line 1
E
Contributions
-1.0
-0.5
0.0
0.5
1.0
cell area σ
roundness clumped σ
dapi single σ
ratio w/l single σ
ratio w/l σ
roundness single σ
dapi clumped σ
dapi σ
roundness σ
num cells
num cells single
nucleus roundness μ
cell area single μ
nucleus roundness clumped μ
cell area μ
cell area clumped μ
nucleus area single μ
nucleus area clumped μ
nucleus area μ
dapi clumped μ
F
Contributions
-1.0
-0.5
0.0
0.5
1.0
clumps
ratio w/l single σ
num cells clumped
num cells single
num cells
ratio w/l clumped sd
dapi single σ
dapi clumped σ
dapi mean
dapi single sd
dapi sd
clumps
Edu
nucleus roundness single σ
nucleus roundness clumped σ
nucleus ratio w2l single σ
nucleus roundness σ
nucleus ratio w/l clumped σ
roundness clumped σ
nucleus ratio w/l σ


A
Correlation value
0.25
0.00
−0.25
−0.50
−2e−07
−1e−07
0e+00
1e−07
2e−07
Expression value residuals
Correlation value
0.25
0.00
−0.25
−2e−07
−1e−07
0e+00
1e−07
2e−07
Expression value residuals
● Genes correlating with the Extrinsic Factor only
● Genes correlating with the Intrinsic Factor only
● Genes correlating with the both Factor
B
positive regulation of catalytic activity
complement activation
lipoprotein metabolism
lipid transport
pharyngeal arch atrery morphogenesis
cell-cell adhesion via
plasma-membrane
adhesion molecules
transmembrane receptor
protein serine/threonine
kinase signaling pathway
lipid metabolism
glycosilation
glycosaminoglycan metabolism
monocarboxylic acid metabolism
response to stress
positive regulation of cell proliferation
cell activation
regulation of G1/S transition of
mitotic cell cycle
cell-adhesion mediated by integrin
membrane organization
cell-matrix adhesion
C
number of cell
proliferation
clumpiness
cell_area
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
● 1
■ 9
▲ 1&9
AIF1
AKIRIN2
ANXA2
APP
CCR3
CD9
CDC25B
CHI3L1
E2F1
ECM1
EDNRB
ETV5
FGF8
FUCA1
GNB1
HDAC2
HIPK1
HIPK2
HPSE2
STAT3
ACTR2
APOLD1
ATM
CDT1
CFD
CTDSP1
ERCC1
GRK5
IQGAP1
ACVR1B
AMPD3
AR
ATP2A2
BAMBI
CAND1
CASP8
CD1D
CD59
CENPJ
CEP192
EDN1
ELANE
FCN1
FOXP2
HES1
ID1
ID2
IMPDH1
ITGB1
ITGB1BP1
MYC
PTPN6
CLEC4E
CNOT4
APEH
ATP5G2
BST1
CD3E
CDKN2D
ADAM9
ADM
CXADR
ATF3
EGR3
CD14
FLT3
FZD7
CNOT6L
GNG5
CRISPLD2
ACLY
CDC14A
FOXP1
CHMP2A
JUN
CDK10
NOG
ANO9
BRAF
ITGB5
F2R
CTSC
DERL2
EXO1
FGF18
DMRTA2
IL5
BCL2L1
BID
CD3D
CD1C
HSPA8
DNAJC3
IL33
HMOX1
ITGAL
MLXIPL
GHRH
D
Expression value vsn
16
12
8
4
CCR3
IL5
CDC14A
IL33
CASP8
GHRH
ADAMTS13
ATP11A
CDKN2C
ITGAL
ERBB4
CD3D
FOXE3
ITGB1
SRC
ERBB2
ATF3
AAK1
HSH2D
ESR1
ANO9
EGR3
ATP5G2
PRCP
ID2
CXADR
BAMBI
ANXA2
ID1
GNB1
ADM
HSPA8


A
10257 SNVs
Frequency
4495 Rare
Condel analysis
1234 Deleterious
Structural analyisis
523 With structure
DUET analysis
103 Destabilising
Filter in outlier cell lines
37 In outlier cell lines
B
zisa_33
yuze_1
xegx_1_3
voas_2
uahf_3
oomz_22
oomz_2
leeh_2
iicq_3
iakz_2
iakz_1
gibe_2
gedo_33
fumb_2
ffdr_3
eofe_1
debk_9
coio_2
coio_1
cell area
clump size
dapi
edu
nucleus area
nucleus width/length
nucleus roundness
ratio nucleus/cell area
cell width/length
cell roundness
DMBT1
CYP2C19
UBE2T
LAMA3
ITGA6
EXOSC10
CYP24A1
GLDC
FBXO27
ARSH
PLEKHG4B
PFKM
ALPL
TMPRSS9
FAT1
CA2
SI
SERPINA1
KCNN4
EDEM2
DNAJB5
CHRNB3
RND1
MYBPC2
IL20RA
GRIK2
FCF1
BMX
DHCR7
ATP2A3
GALR3
C
yuze_1
ctrl
Fn1
Fn5
Fn25
100 µm
D
ffdc_11
Secondary structure: beta sheet
Duet score: -1.40
Structure:
4wk2.pdb1_A_24_507_P23229
Model position: 27
ITGA6
Secondary structure: loop
Duet score: -1.43
Structure:
4wjk.pdb1_B_1_245_P05556
Model position: 105
ITGB1