ARTICLES
https://doi.org/10.1038/s41588-021-00801-6
nature genetics

Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation

Julie Jerber[1,2,10], Daniel D. Seaton[3,10], Anna S. E. Cuomo[3,10], Natsuhiko Kumasaka[2], James Haldane[2], Juliette Steer[2], Minal Patel[2], Daniel Pearce[2], Malin Andersson[2], Marc Jan Bonder[3], Ed Mountjoy[1], Maya Ghoussaini[1], Madeline A. Lancaster[4], HipSci Consortium*, John C. Marioni[2,3,5] ✉, Florian T. Merkle[6,7] ✉, Daniel J. Gaffney[2,11] ✉ and Oliver Stegle[2,3,8,9,11] ✉

Studying the function of common genetic variants in primary human tissues and during development is challenging. To address this, we use an efficient multiplexing strategy to differentiate 215 human induced pluripotent stem cell (iPSC) lines toward a midbrain neural fate, including dopaminergic neurons, and use single-cell RNA sequencing (scRNA-seq) to profile over 1 million cells across three differentiation time points. The proportion of neurons produced by each cell line is highly reproducible and is predictable by robust molecular markers expressed in pluripotent cells. Expression quantitative trait loci (eQTL) were characterized at different stages of neuronal development and in response to rotenone-induced oxidative stress. Of these, 1,284 eQTL colocalize with known neurological trait risk loci, and 46% are not found in the Genotype–Tissue Expression (GTEx) catalog. Our study illustrates how coupling scRNA-seq with long-term iPSC differentiation enables mechanistic studies of human trait-associated genetic variants in otherwise inaccessible cell states.

Human iPSCs are a promising model for assessing the cellular consequences of human genetic variation across different lineages, developmental states and cell types. In particular, human iPSCs facilitate the study of developmental states and stimulation conditions that would be challenging or even impossible to obtain in vivo. The creation of cell banks containing hundreds of iPSC lines[1] provides an opportunity to perform population-scale studies in vitro[2–5]. However, differentiating iPSCs is expensive and labor intensive, with experiments being difficult to compare due to substantial batch variation. Thus, studies of more than a handful of lines remain a substantial challenge. Furthermore, most iPSC differentiation protocols produce heterogeneous cell populations with the target cell type representing only a subset[6–8]. This variability in differentiation outcomes hinders efforts to dissect genetic contributions to cellular phenotypes.

scRNA-seq enables multiplexed experimental designs, in which cells from multiple donors are pooled together[2,9,10]. Multiplexing improves throughput and allows experimental variability between differentiation batches to be rigorously controlled, enabling discrimination of pool effects from differences between lines. However, multiplexed experimental designs have largely been applied to short differentiation protocols and have not captured developmental progression toward a mature cell fate.

Here, we apply a multiplexing strategy to profile the differentiation and maturation of 215 iPSC lines derived from the Human Induced Pluripotent Stem Cell Initiative (HipSci) toward a midbrain neural fate, including dopaminergic neurons (DAs). DAs are involved in motor function and other cognitive processes and play key roles in neurological disorders, including Parkinson's disease[11,12]. Using an established protocol[13], we collected cells at three maturation stages (progenitor-like, young neurons and more mature neurons), covering 52 days of differentiation. We additionally exposed cells to a chemical stressor on day 51 to explore how genetic variation shapes stress response. Using this system, we create an eQTL map at multiple stages of human neuronal differentiation and identify over 500 new trait–eQTL colocalizations. Using estimates of cell population composition based on scRNA-seq data, we identify a strong cell-intrinsic differentiation bias and identify molecular signatures that can predict which iPSC lines fail to efficiently produce neuronal cells.

Results

High-throughput differentiation of midbrain DAs. We selected 215 iPSC lines from the HipSci project[1], each derived from a single healthy donor, for differentiation toward a midbrain cell fate, including DAs[13]. Differentiation experiments were multiplexed in pools containing between seven and 24 lines per experiment, with 35 lines being contained in multiple pools (Supplementary Fig. 1 and Supplementary Table 1). Immunochemistry confirmed that cells differentiated either in pools or individually both expressed protein markers associated with patterning of DAs (LIM homeobox transcription factor 1 α (LMX1A), forkhead box (FOX)A2 and tyrosine hydroxylase (TH)) (Supplementary Fig. 2). To capture transcriptional changes during neurogenesis and neuronal maturation, we

[1]Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, UK. [2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [4]MRC Laboratory of Molecular Biology, Cambridge Biomedical Campus, Cambridge, UK. [5]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [6]Metabolic Research Laboratories and Medical Research Council Metabolic Diseases Unit, Wellcome Trust–Medical Research Council Institute of Metabolic Science, University of Cambridge, Cambridge, UK. [7]Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [8]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [9]Division of Computational Genomics and Systems Genetics, German Cancer Research Center, Heidelberg, Germany. [10]These authors contributed equally: Julie Jerber, Daniel D. Seaton, Anna S. E. Cuomo. [11]These authors jointly supervised this work: Daniel J. Gaffney, Oliver Stegle. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: marioni@ebi.ac.uk; fm436@medschl.cam.ac.uk; dg13@sanger.ac.uk; o.stegle@dkfz-heidelberg.de

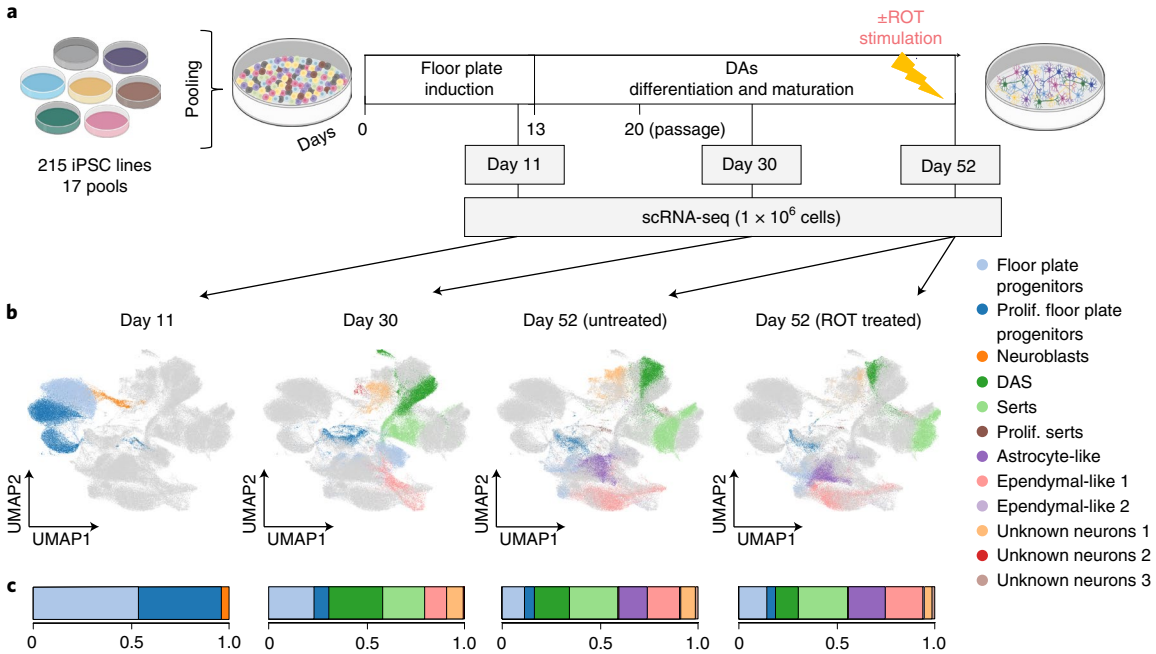


Fig. 1 | Experimental design and cell type heterogeneity in pooled differentiations of iPSCs to a midbrain cell fate. a, Experimental workflow for scRNA-seq analysis of iPSC-derived DAs. Time points at which cells were collected for scRNA-seq profiling (day 11, day 30, day 52) are indicated. On day 51, half of the cells were stimulated with rotenone (ROT) for 24 h to induce oxidative stress. **b**, UMAP plot of all 1,027,401 cells assayed, colored by annotated cell type identity (Methods). Cells that were not collected for a given condition (time point, stimulus) are displayed in light grey. Prolif., proliferating. **c**, Bar plot showing the fraction of cells assigned to each cell type for each condition.

performed scRNA-seq on cells captured at day 11 (midbrain floor plate progenitors), day 30 (young post-mitotic midbrain neurons) and day 52 (more mature midbrain neurons). To mimic an oxidative stress condition, we also profiled day 52 neurons 24 h after exposure to a sublethal dose of rotenone (0.1 µM, 24 h), a chemical stressor that preferentially leads to DA death in models of Parkinson's disease (Fig. 1a)[14].

After quality control (Methods), we obtained a total of 1,027,401 cells across 17 pools[15] and four conditions (Fig. 1a and Supplementary Table 1). The line of origin for each cell in a given pool was inferred from scRNA-seq read information using genotype data from the HipSci consortium (using demuxlet[16]). Adjustment for experimental batch effects using Harmony[17] followed by Louvain clustering[15] identified 26 clusters (six, seven and 13 clusters, respectively, at days 11, 30 and 52, Extended Data Fig. 1a). These clusters were assigned putative cell type labels based on the expression profiles of literature-curated marker genes (Fig. 1b, Extended Data Fig. 1 and Supplementary Methods).

We identified a total of 12 distinct cell types, including six dominant cell types that contained at least 10% of cells in one of the four conditions (Fig. 1 and Extended Data Fig. 1). These included two cell type populations at day 11: proliferating and non-proliferating midbrain floor plate progenitors (both expressing *LMX1A* and *FOXA2* and expressing *MKI67* and *TOP2A* when proliferating[18]). At days 30 and 52, four additional dominant cell types were identified, two of which appeared to be neuronal and two of which appeared to be non-neuronal (characterized by expression and lack of the pan-neuronal markers *SNAP25* and *SYT1*, respectively). The first neuronal population was annotated as midbrain DAs using a comprehensive panel of 75 literature-derived DA marker genes, including *TH*, *NR4A2*, *PBX1* and *TMCC3*[18–21] (Extended Data Fig. 1d). Moreover, we performed transcriptome-wide alignment of this cell population to existing single-cell atlases of human iPSC-derived DAs[18] and human fetal[18] or adult human midbrain samples[22], further supporting their DA identity with mapping rates to reference DA populations of 85–99% (Supplementary Methods). We annotated the second neuronal population as serotonergic-like neurons (serts), because these cells were enriched for *TPH2* and *GATA2* markers, which are also observed in serotonergic neurons in vivo[23]. The two non-neuronal cell types consisted of ependymal-like cells detected at days 30 and 52 (ependymal 1 (ref. [24])) and astrocyte-like cells detected at day 52 (astrocyte-like[25,26]). We also identified a neuroblast population specific to day 11 (4% of cells) expressing pro-neuronal genes (*NEUROD1*, *NEUROG2*, *NHLH1* (refs. [27,28])) and an additional neuronal population (expressing *SNAP25* and *SYT1*) that expressed some midbrain markers but could not be assigned a specific identity (unknown neurons 1, present at day 30 and day 52 at around 7%, Extended Data Fig. 1c–e). Finally, we identified four rare cell types (<2% of cells sampled at any time point), including a second ependymal-like population (ependymal 2), a cell population of proliferating progenitors and serts (proliferating serts) and two additional neuronal populations, which could not be annotated unambiguously (unknown neurons 2 and 3, Fig. 1b and Extended Data Fig. 1c–e).

UMAP projection of cells collected across all time points, stimuli and lines revealed broad co-clustering of cell types but with noticeable differences between time points and stimuli (Fig. 1b,c and Supplementary Fig. 1). For example, the proportion of DAs upon rotenone stimulation was significantly reduced (30% reduction upon stimulation, Fisher's exact test, $P = 2.2 \times 10^{-16}$), consistent with previous observations that DAs are most affected by apoptosis due to oxidative stress[29–31]. In line with this observation, a variance component analysis of gene expression identified treatment as the second most important driver of expression variation after cell type (Supplementary Fig. 1).

Collectively, our population-scale scRNA-seq analysis revealed a diverse repertoire of cell types, enabling the study of both cell line differentiation propensity and the identification of genetic effects on gene expression with cell type resolution.

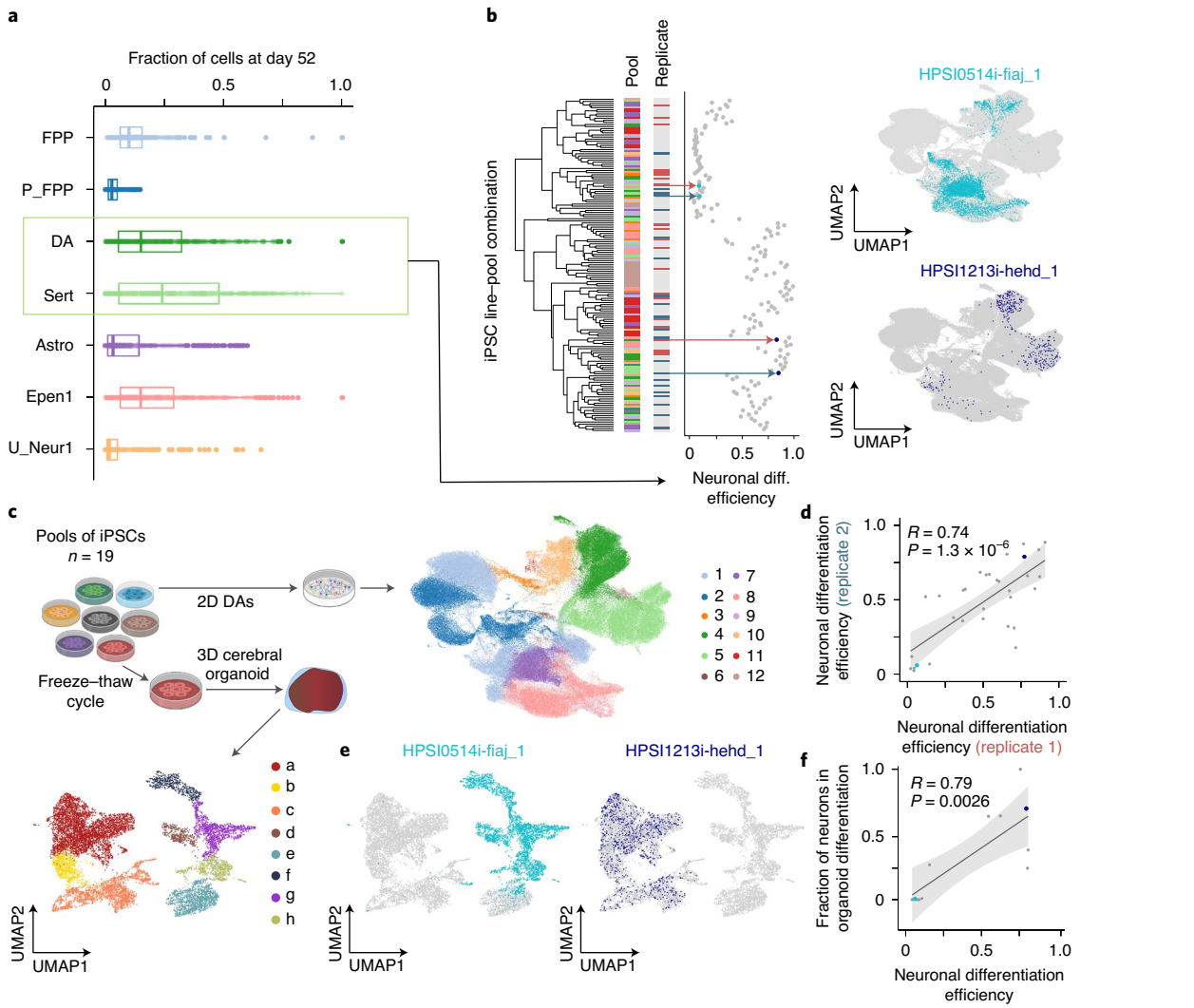


Fig. 2 | Reproducible variation in differentiation trajectories. a, Box plots display the proportions of each cell type at day 52 across cell lines. The sum of the proportions of DA and sert cells at day 52 is defined as neuronal differentiation efficiency. In the box plots, the middle line is the median, and the lower and upper edges of the box denote the first and third quartiles. Astro, astrocyte-like; epen1, ependymal-like 1; FPP, floor plate progenitors; P_FPP, proliferating floor plate progenitors; U_Neur1, unknown neuron 1. **b**, Left, hierarchical clustering of (cell line, pool) combinations by neuronal differentiation (diff.) efficiency. Displayed are data from ten pools with at least ten scRNA-seq profiles across all time points (138 lines). The first bar indicates a line's pool of origin; the second bar indicates the replicate status. Right, uniform manifold approximation and projections (UMAPs), highlighting the distributions of cells on day 52 for two selected cell lines with low (HPSI0514i-fiaj_1, sea green) and high (HPSI1213i-hehd_1, dark blue) neuronal differentiation efficiencies. **c**, Workflow for scRNA-seq analysis of iPSC-derived cerebral organoids. UMAPs summarize the resulting cell populations (1, floor plate progenitors; 2, proliferating floor plate progenitors; 3, neuroblasts; 4, DAs; 5, serts; 6, proliferating serts; 7, astrocyte-like; 8, ependymal-like 1; 9, ependymal-like 2; 10, unknown neurons 1; 11, unknown neurons 2; 12, unknown neurons 3; a, neurons; b, intermediate progenitors; c, radial glial progenitors; d, satellite cells; e, mesenchymal cells; f, myotube; g, paired box (PAX)7$^+$ cells; h, Wnt$^+$ cells; 2D, two-dimensional; 3D, three-dimensional). **d**, Scatterplot of neuronal differentiation efficiency between replicate pools ($n = 32$ cell lines differentiated in two different pools). The two cell lines from **b** are highlighted. **e**, UMAPs of the two cell lines selected in **b**, making non-brain and brain cell types in the organoid study. **f**, Scatterplot of midbrain dopaminergic neuronal differentiation (x axis) versus neuronal differentiation efficiency measured in organoid differentiation (y axis) for a common subset of 12 iPSC lines. The two cell lines from **b** are highlighted. In **d**,**f**, locally estimated scatterplot smoothing (LOESS) curves and 95% confidence intervals are shown alongside Pearson's R and the P value from a two-sided t-test.

Intrinsic variation in neuronal differentiation efficiency between iPSC lines. iPSC differentiation protocols often generate highly variable results across cell lines, but the reason for this remains obscure, hindering efforts to select cell lines for specific applications[8,32]. We observed substantial variation in the proportions of cell types produced by different iPSC lines at each time point (Fig. 2a, Extended Data Fig. 2a and Supplementary Table 2). Principal component analysis of cell type fractions per line and pool identified the proportion of DAs and serts on day 52 as the largest axis of variation (principal component 1, 47% variance, Extended Data Fig. 2b–d). As DA and sert cells are derived from similar progenitor populations in vivo[33], we considered the combined proportion of these cell types on day 52 as a measure of 'neuronal differentiation efficiency' for each iPSC line (Fig. 2b). Using 32 lines that were represented in two different pools, we confirmed the reproducibility of this measure of neuronal differentiation efficiency (Pearson $R = 0.75$; $P = 2 \times 10^{-6}$; Fig. 2d), and we assessed its robustness when excluding rotenone-treated cells

(Supplementary Fig. 3). Neuronal differentiation efficiency was not associated with the number of lines per pool ($R^2 = 0.04$, $P = 0.46$, *t*-test) but was positively correlated with neuronal maturation[34] (Supplementary Fig. 3 and Extended Data Fig. 1c). Finally, we considered data from six lines that were both differentiated individually and in one pool, finding that pooling neither affected neuronal differentiation efficiency (Extended Data Fig. 3b) nor led to obvious transcriptional differences (Supplementary Fig. 4).

We next investigated whether these observations were generalizable to other neuronal differentiation approaches by differentiating a pool of 18 lines (pool 4) into cerebral organoids for 113 d[35], followed by profiling using scRNA-seq (11,445 cells, Fig. 2c and Methods). We found that the proportion of neurons (Supplementary Fig. 5) produced by each line in the cerebral organoids was also correlated with neuronal differentiation efficiency as estimated from the dopaminergic differentiation (Fig. 2e,f, $R = 0.79$; $P = 3 \times 10^{-3}$; $n = 12$, *t*-test).

The reproducibility of differentiation outcomes in different settings and protocols suggests that variation in iPSC neuronal differentiation efficiencies arises primarily due to cell-intrinsic factors. Furthermore, the consistency of differentiation efficiency suggests that these properties extend to neuronal differentiation more generally.

An iPSC gene expression signature predicts neuronal differentiation efficiency. Motivated by the reproducibility of differentiation outcomes across multiple independent pools, we tested for associations between neuronal differentiation efficiency and experimental and biological factors (Supplementary Table 3). We found no or weak associations with passage number ($P = 0.77$; *F*-test), sex ($P = 0.008$, *t*-test), chromosome X activation status ($P = 0.01$, *F*-test, Supplementary Methods) or PluriTest scores[36] ($P = 0.01$, *F*-test, Supplementary Methods) of the corresponding lines. Using variance component analysis, we assessed the relevance of additional factors, in particular, enabling the comparison of line versus pool effects. This identified line effects as the dominant driver of variability in neuronal differentiation efficiency (Extended Data Fig. 4a,b), although our design does not enable discrimination of line and donor effects.

Next, we assessed whether neuronal differentiation efficiency was associated with gene expression in undifferentiated iPSCs. Using independent bulk RNA-seq data for 184 iPSC lines included in this study[1,37], we identified associations between neuronal differentiation efficiency and gene expression level for 2,045 genes (983 positive and 1,062 negative associations; *F*-test, false discovery rate (FDR) < 5%; Fig. 3b,c, Supplementary Table 4 and Methods). When defining poor differentiation as a binary outcome (neuronal differentiation efficiency <0.2), iPSC gene expression could be used to train a classifier of poor differentiation (logistic regression; 100% precision at 35% recall, assessed using leave-one-out cross-validation; Methods), which we validated in alternative training and testing regimes, using independent hold out lines (Extended Data Fig. 4 and Methods). Using this model, we obtained predicted differentiation scores for 812 HipSci lines with bulk RNA-seq data (Supplementary Table 5), finding that a substantial fraction of HipSci lines (13%) were likely to result in poor differentiation outcomes. We also tested whether the same experimental and biological factors previously associated with neuronal differentiation efficiency were replicated in this larger sample, finding consistent results (Supplementary Table 3 and Methods). Finally, to assess the possibility of a genetic or other donor-specific component of neuronal differentiation efficiency, we assessed the consistency of predicted differentiation outcomes for lines from the same donor, observing poor concordance (Extended Data Fig. 4b). Moreover, we found no association between germline variants and predicted neuronal differentiation efficiency when performing a genome-wide association study (GWAS) (all $P > 5 \times 10^{-8}$, $n = 540$, minor allele frequency (MAF) < 0.05; Methods), although the available sample size was insufficient to rule out weaker effects.

As iPSC cultures are heterogeneous, we hypothesized that the predictive gene signatures might originate from varying proportions of iPSC subpopulations. To test this, we reanalyzed scRNA-seq data from 112 iPSC lines that were assayed previously[2], 45 of which were also included in this study (Methods and Fig. 3a). We identified five clusters, with all but one (cluster 4) expressing high levels of core pluripotency markers (*NANOG*, *SOX2*, *POU5F1*, Extended Data Fig. 5c and Methods). Cells from cluster 2 overexpressed genes associated with inefficient neuronal differentiation (for example, *UTF1*) and downregulated genes associated with efficient neuronal differentiation (for example, *TAC3*; Fig. 3d,e and Extended Data Fig. 5). None of the remaining clusters displayed such an enrichment (Extended Data Fig. 5b and Supplementary Table 6). To more directly assess the relevance of cluster 2 for neuronal differentiation efficiency, we tested for and confirmed an association between the fraction of cells in cluster 2 and neuronal differentiation efficiency for each cell line (Pearson $R = -0.76$, $P = 2.05 \times 10^{-9}$; Fig. 3f and Extended Data Fig. 5d). Using the known relationship between iPSC bulk RNA-seq values and the proportion of cluster 2 cells, we predicted this proportion for 182 cell lines included in our differentiation experiments, confirming the negative correlation with neuronal differentiation efficiency (Pearson $R = -0.49$; $P = 3 \times 10^{-12}$, Extended Data Fig. 5e and Methods).

Finally, we analyzed an additional scRNA-seq dataset from iPSCs derived from lymphoblastoid cell lines[38]. Using our single-cell analysis workflow, we identified a cluster of cells with an expression profile concordant with that of cluster 2 (Methods and Supplementary Fig. 6). In sum, these results suggest that a subpopulation of iPSCs with poor neuronal differentiation capability is consistently detected across different human iPSC banks and that this bias can be robustly predicted using expression markers at the iPSC stage.

eQTL discovery and comparison with in vivo eQTL maps. We next focused on understanding how individual-to-individual genetic variation influenced gene expression across differentiation and in response to stimulation. Specifically, we mapped cis-eQTL separately for each of the 14 distinct cell populations that correspond to the profiled cell type–condition contexts of the dominant cell types. eQTL were mapped using aggregate expression levels for each donor, considering common gene-proximal variants (MAF > 0.05, plus or minus 250 kb around genes; Methods). Variability in neuronal differentiation efficiency between lines resulted in substantial differences in the number of cells from each donor, generating variation in the total number of cells assayed for each context (Extended Data Fig. 6a) and in turn affecting accuracy of the estimates of aggregated expression. To account for this, we adapted commonly used eQTL mapping strategies[2] based on linear mixed models by incorporating an additional variance component into the model (Methods). This increased the number of eQTL discoveries, resulting in 4,828 genes with at least one eQTL in at least one of the contexts (hereafter 'eGene', FDR < 5%, Fig. 4a, Extended Data Fig. 6b and Supplementary Table 7), with expected enrichment of eQTL variants in the vicinity of gene promoters (Extended Data Fig. 7b).

The largest number of eQTL were detected in progenitor cell populations, likely reflecting increased detection power due to the larger number of cells per line assayed (Extended Data Fig. 6a,b). Notably, the cumulative number of genes with an eQTL in each cell type increased when considering contexts further progressed along the differentiation axis, as well as upon stimulation (Fig. 4a). For example, eQTL mapping in matured DAs (day 52) identified an additional set of 441 eGenes compared to those in these cells at day 30. One such time-point-specific eGene was *HSPB1*, which

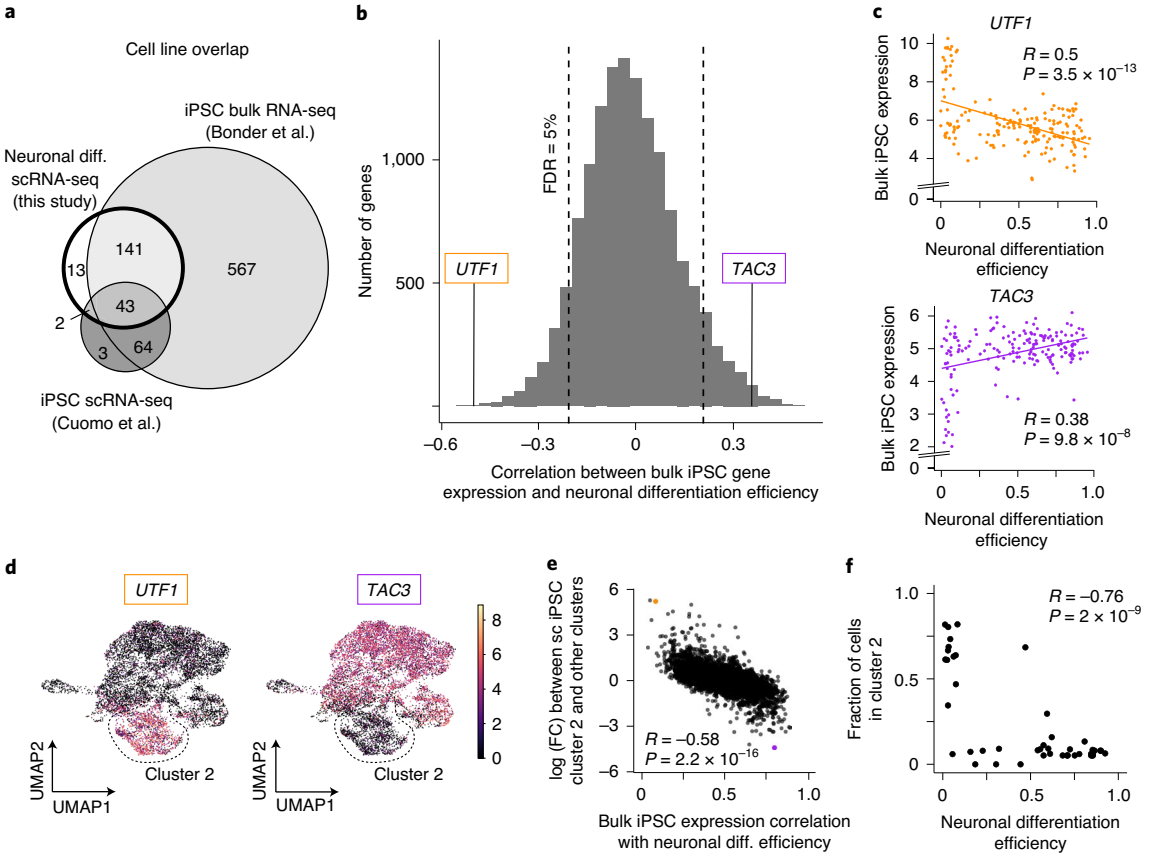


Fig. 3 | A gene expression signature in iPSCs is associated with neuronal differentiation efficiency. a, Venn diagram indicating the overlap of cell lines included in this study and two recent iPSC studies, a bulk RNA-seq study[37] and an scRNA-seq study[2]. **b**, Histogram of Pearson correlation coefficients between variation in gene expression of individual genes (from bulk RNA-seq data[37]) and neuronal differentiation efficiency. Two representative genes with positive (*TAC3*) and negative (*UTF1*) associations are highlighted. **c**, Scatterplot of neuronal differentiation efficiency (*x* axis) versus bulk iPSC gene expression (*y* axis) for *UTF1* and *TAC3*. **d**, UMAPs of scRNA-seq profiles in iPSCs from 112 donors[2]. Colors denote the expression levels of the two selected genes from **b**,**c** (*UTF1* and *TAC3*). Cluster 2 is shown by the dashed lines. **e**, Comparison of marker gene association results with expression markers of cluster 2. For each gene, a scatterplot is shown of the Pearson correlation coefficient of association between iPSC gene expression and neuronal differentiation efficiency (*x* axis, based on bulk gene expression) versus its log fold change between cluster 2 and all other clusters (*y* axis, based on scRNA-seq data). The genes *UTF1* and *TAC3* are highlighted. sc, single-cell. **f**, Scatterplot of neuronal differentiation efficiency (*x* axis) and the proportion of cells assigned to cluster 2 (*y* axis) across 45 cell lines that were included in both sets of experiments. When multiple measurements were available for a given cell line, average values are shown. In **c**,**e**,**f**, the Pearson's *R* and the *P* value from a two-sided *t*-test are indicated.

encodes a heat shock protein that plays a key role in neuronal differentiation[39] and for which SNP rs6465098[T>C] was an eQTL only in cells at day 52 (Fig. 4b). Changes in *HSPB1* expression were observed in neurons after ischemia[40] and were associated with toxic protein accumulation in Alzheimer's disease[41,42].

Similarly, we detected 248 additional eGenes in DA and sert neurons following rotenone treatment. For example, rs12597281[A>G] is an eQTL for *ACSF3* in rotenone-stimulated serotonergic neurons at day 52 but not in unstimulated cells (Fig. 4b). *ACSF3* encodes an acyl-CoA synthetase localized in the mitochondria, and inherited mutations are associated with a metabolic disorder, combined malonic and methylmalonic aciduria, in which patients exhibit a wide range of neurological symptoms, including memory loss, psychiatric problems and/or cognitive decline[43]. We also compared eQTL across the 14 contexts at the level of individual variants (using MASHR[44]; Methods), finding distinct clustering of contexts consistent with the underlying lineages as well as context-specific effects (Extended Data Fig. 6c,d).

To test how our eGene discovery relates to previous studies, we compared the number of eGenes identified in this study with bulk eQTL maps from in vivo tissues from the GTEx consortium[45] (Methods). Although we observed fewer eQTL in cell populations of individual contexts than in GTEx tissues of similar sample size, the aggregate numbers of eGenes identified across contexts were similar to those from eQTL maps from primary tissue of the same sample size (Fig. 4c and Extended Data Fig. 8c).

A key question about eQTL maps from in vitro iPSC-based models is how closely they resemble eQTL maps from primary tissues that differ in cell composition. To explore this, we tested the extent to which regulatory variants were shared between eQTL maps in three resources: (1) the current study, (2) GTEx brain tissues ($n = 13$ tissues) and (3) bulk RNA-seq profiles of HipSci iPSC lines[2,37], as measured by genome-wide consistency of eQTL effect sizes (using MASHR[44]; Methods). We observed that, as iPSCs were differentiated to increasingly mature neuronal cell types, the extent of eQTL sharing tended to increase (Fig. 4d), although this trend could in part be explained by increasing fractions of GTEx brain eGenes that were expressed in different condition–cell type contexts (Extended Data Fig. 8a). Interestingly, while globally iPSC-derived eQTL maps mimic in vivo GTEx brain eQTL maps, we also identified 2,366 eQTL that could not be detected in GTEx brain tissues (*Q* value >0.05 in any of 13

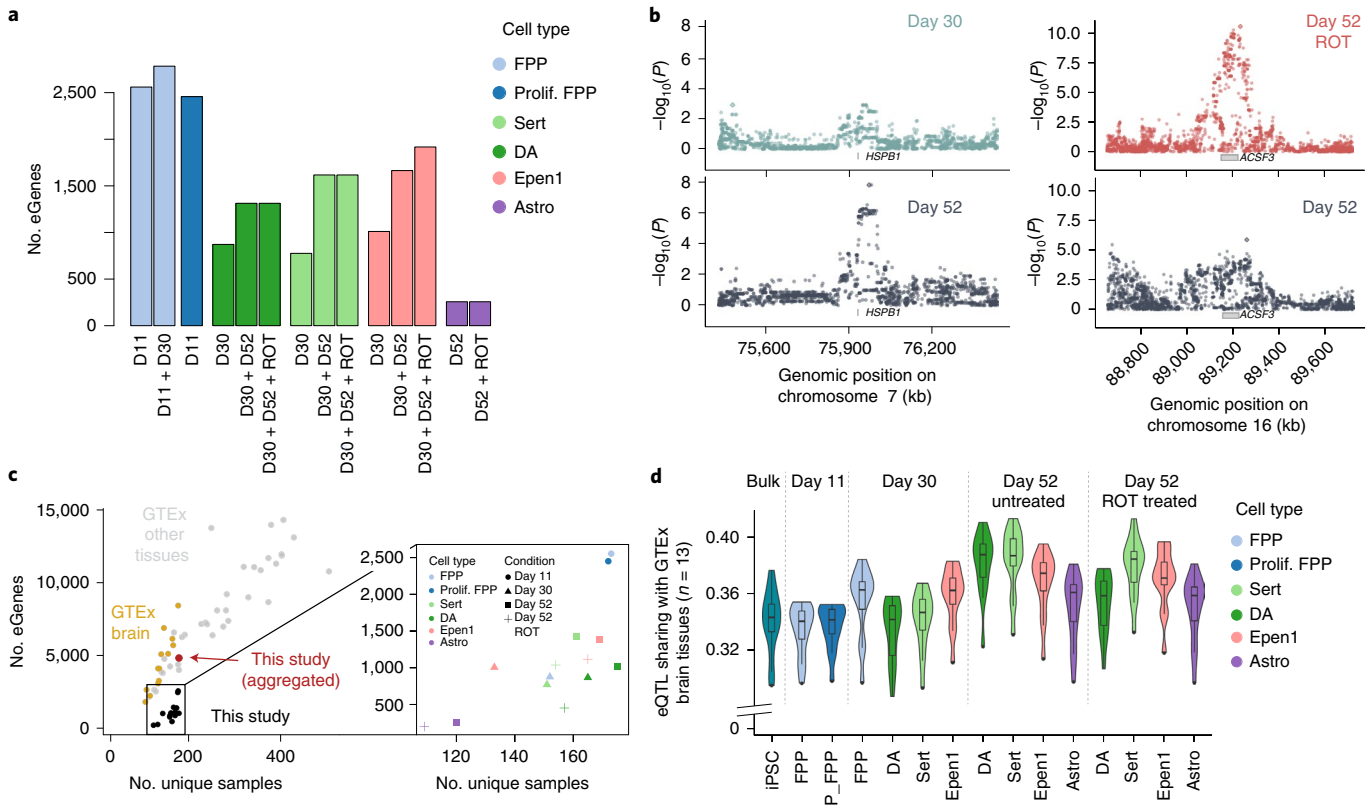


Fig. 4 | Mapping of cis-eQTL in 14 distinct cell contexts (cell types–conditions) for six dominant cell types identified across midbrain differentiation. **a**, Cumulative number of eGenes for each cell type and condition (ROT, rotenone stimulation). D11, day 11; D30, day 30; D52, day 52. **b**, Left, day 52-specific eQTL for *HSPB1* in DAs (rs6465098; FDR < 5%, Methods). Manhattan plots for DAs at day 30 (top) and unstimulated DAs at day 52 (bottom) are shown. Right, a rotenone stimulus-specific eQTL for *ACSF3* in serts (rs12597281, right). Shown are Manhattan plots for rotenone-stimulated serts at day 52 (top) and unstimulated serts at day 52 (bottom). **c**, Comparison of the number of eGenes for individual eQTL maps (FDR < 5%; *y* axis) as a function of the effective sample size (number of unique donors, *x* axis) across studies and cell types. Left, results from overlapping eQTL results in this study with in vivo eQTL maps from GTEx data, divided into brain tissues (yellow) and non-brain tissues (grey). The result from this study when aggregating across all 14 eQTL maps is shown in red. Right, a magnified view of the results from our study colored by cell type and shaped by condition. **d**, Sharing of eQTL signals discovered across 14 cell contexts in this study, as well as using bulk RNA-seq in iPSCs[2,37] with in vivo brain eQTL maps (from the GTEx catalog). Violin plots show the extent of eQTL sharing (Methods) with each of 13 GTEx brain eQTL maps. In the box plots, the middle line is the median, and the lower and upper edges of the box denote the first and third quartiles, while the violin plots show the distribution.

tissues), demonstrating the ability of our approach to discover new regulatory relationships.

Colocalization of eQTL with disease risk variants. The identified cell-type-specific eQTL maps across different differentiation contexts provide an opportunity to understand human disease traits and their genetic risk factors as identified by GWASs. To test for such colocalization events, we applied coloc[46] (Methods) to the summary statistics from 25 neurological traits, eQTL discovered in our study, as well as eQTL obtained from GTEx data (Methods and Supplementary Tables 8 and 9).

We identified 1,284 eQTL in our study with evidence of colocalization with at least one disease trait (Fig. 5a,b). Of these, 597 were found only in our dataset, corresponding to an additional >10% of colocalization events of GWAS variants compared to those of eQTL across all GTEx tissues (5,028 across 48 tissues, Fig. 5b). Notably, the majority of these genes (98%) were expressed in GTEx tissues but either had no eQTL (65%) or had an eQTL that did not give rise to a significant colocalization (34%, posterior probability (PP)3 < 0.5 for all GTEx brain tissues). Furthermore, when considering the relevance of different cell type contexts for explaining these specific colocalization events, we observed that 401 (67%) of the colocalizations in our data were associated with eQTL detected in later differentiation stages (day 52) or upon stimulation (day 52 with rotenone, Supplementary Fig. 7). Finally, we considered a colocalization analysis when using aggregate pseudobulk results across all cell types in our data at day 52 (untreated cells), which yielded a markedly lower number of colocalizations, suggesting that the cell type specificity of our approach is a key factor in explaining the additional colocalizations (Extended Data Fig. 8d).

One notable colocalization event was an eQTL for *SFXN5*, a mitochondrial amino acid transporter[47], that was specific to the rotenone-stimulated serotonergic neurons at day 52 and colocalized with a schizophrenia hit (PP4 = 0.78, Fig. 5c and Supplementary Fig. 7). Exposure to rotenone is known to induce oxidative stress by inhibiting the mitochondrial respiratory chain complex[48,49], suggesting that the specific genetic signal observed for the mitochondrial gene *SFXN5* in serotonergic neurons might modulate environmental stress response.

Another example that colocalized with a schizophrenia GWAS variant was an eQTL for *FGFR1*, detected both in proliferating and non-proliferating floor plate progenitors at day 11 (PP4 = 0.93 and 0.88 respectively, Fig. 5d). Previous studies showed that nuclear *FGFR1* plays a key role in regulating neural stem cell proliferation and central nervous system development, in part by binding to the promoters of genes that control the transition from proliferation to

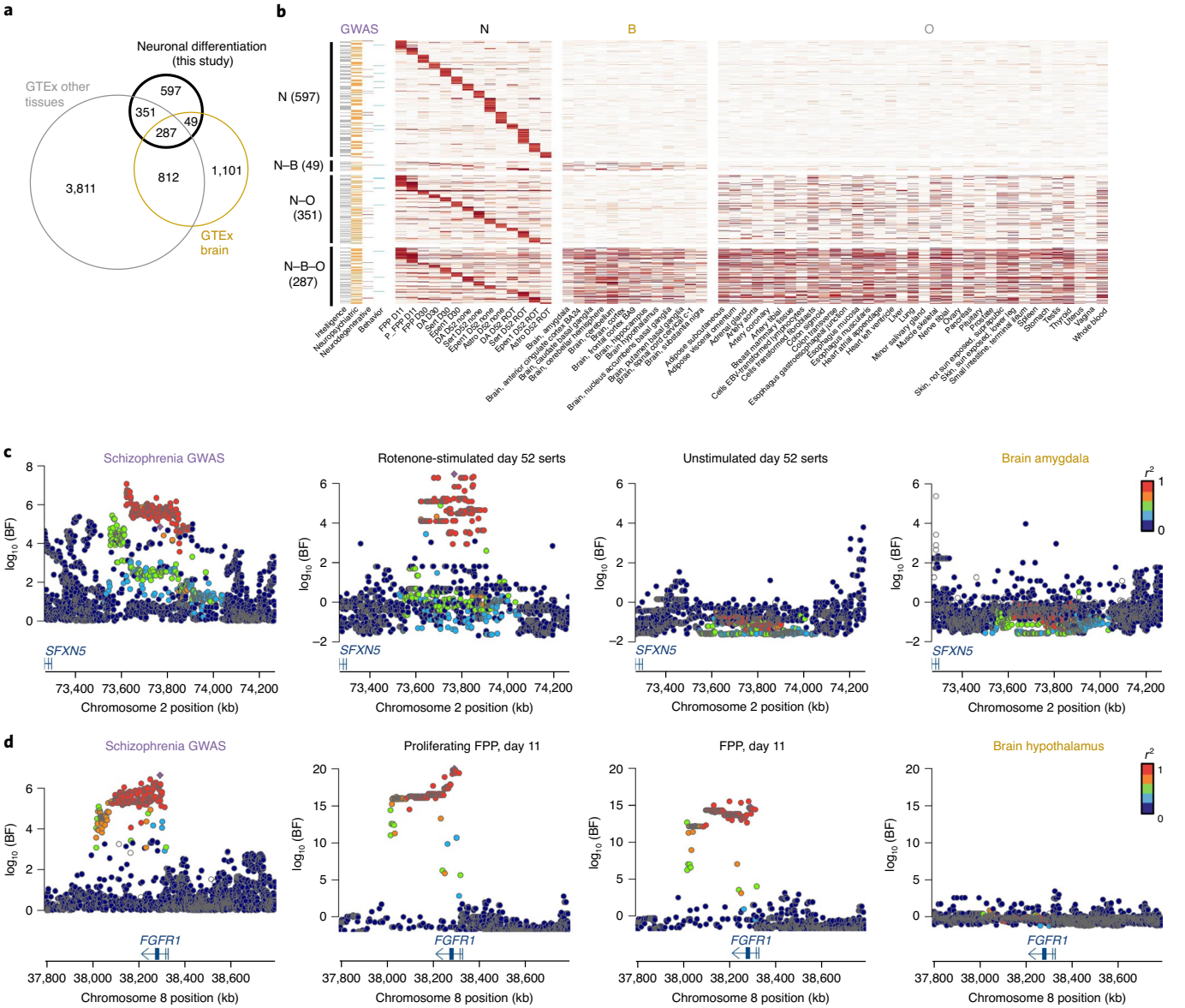


Fig. 5 | Colocalization analysis of eQTL with 25 neuro-related GWAS traits. a, Venn diagram showing the number and the overlap of colocalization events discovered using eQTL maps from this study, GTEx brain and other GTEx tissues. **b**, Heatmap showing the posterior probability of colocalization (PP4 from coloc; Methods) for eQTL that colocalized with one or more GWAS traits. N, neuronal differentiation (this study); B, GTEx brain; O, other GTEx tissues. **c**, Locus zoom plots around the *SFXN5* gene. The schizophrenia GWAS association (left) is colocalized with the eQTL in rotenone-stimulated serts at day 52 (second panel from the left). No colocalization signal was detected in unstimulated serotonergic neurons at day 52 (third panel from the left) or in any other brain GTEx tissues as illustrated here with GTEx brain amygdala (rightmost panel). The lead variant is indicated with a purple diamond, and other points are colored according to the linkage disequilibrium index (r^2 value) with the lead variant. BF, Bayes factor. **d**, A midbrain progenitor-specific eQTL for *FGFR1* associated with schizophrenia. We identified a colocalization event with this eQTL in both proliferating (second panel from the left) and non-proliferating floor plate progenitors (third panel from the left) at day 11. No colocalization was found in any other cell type profiled in this study (not shown) nor in any brain GTEx tissues (shown with GTEx brain hypothalamus, rightmost panel).

cell differentiation[50]. Additionally, it was shown that altered *FGFR1* signaling is linked to the progression of the cortical malformation observed in schizophrenia[51].

These examples suggest that a combination of genetic and environmental factors during early development might contribute to schizophrenia pathology and illustrate how these data represent a valuable resource for understanding the molecular basis of complex neurological disease.

Discussion

Characterizing the function of human trait-associated genetic variation requires large-scale studies performed in disease-relevant cell types and states. Here, we demonstrate how human iPSCs can be efficiently profiled at scale throughout a long-term differentiation to a midbrain cell fate. We uncover a highly reproducible, cell-intrinsic neuronal differentiation bias and show how this bias can be predicted from gene expression profiling of the pluripotent cell state. This sets the stage for optimized design of future large-scale iPSC experiments, in which cell lines can be rationally selected a priori without laborious testing of differentiation capacity.

Despite a modest sample size, our study identified a large number of new disease–eQTL colocalizations compared with those in GTEx tissues of equivalent sample size. For example, the numbers of new disease–eQTL colocalizations added by GTEx liver and cerebellar

hemisphere ($n = 208$ and 215, respectively) are 80 and 107, respectively, compared to 597 colocalizations in this study. This does not necessarily indicate that the eQTL we discovered here are disproportionately more likely to be disease relevant, as it is challenging to account for differences in the number of comparisons; individual GTEx tissues constitute a single eQTL map, whereas our data comprise multiple maps. As a result, there is an implicit multiple-testing burden that is not accounted for in existing methods for colocalization analysis (for example, 20,201 tests in GTEx liver versus 153,350 tests across all our maps). A biological explanation for additional colocalization events is that our experiment profiled expression states that are difficult to capture using post-mortem tissue, including time points during neuronal differentiation and following rotenone exposure. Additionally, we detected many eQTL that were specific to individual cell types, enabled by the single-cell resolution of our study. These signals, while present, are challenging to detect in bulk tissue because the relevant cell types are often rare. The relevance of cell-type-specific colocalization events is also supported by our colocalization analysis using pseudobulk profiles at day 52, which identified a considerably lower number of colocalizations (Extended Data Fig. 8d). In sum, these results suggest that many 'missing' but disease-relevant eQTL likely remain to be discovered using single-cell sequencing of both primary tissue and in vitro cell models.

A second implication of our study is that, despite growth competition between cell lines, multiplexing experiments retain sufficient cells per donor to perform robust genetic analysis, even following extended periods in culture (Extended Data Fig. 6a and Supplementary Table 2). Nevertheless, although cell lines were pooled at similar numbers, we observed extensive variation throughout our experiment in the numbers of cells produced by different lines (Fig. 2a). Future technical improvements, such as better differentiation methods, more precise matching of growth rates of cell lines within pools or line selection based on predicted differentiation capacity using markers in the iPSC state, may further increase the utility of multiplexed iPSC differentiation.

The 'quality' of human iPSCs was previously carefully examined using both genetic and functional genomic data[36,52–55]. Despite these efforts, differentiation bias among cell lines was widely appreciated but poorly understood. The underlying mechanisms were hypothesized to involve epigenetic factors, environmental factors, such as culture conditions, changes acquired by cells over time in culture or cell type of origin. Our work systematically surveys differentiation biases at the scale of an entire cell bank. The results cannot arise due to differences in the cell type of origin[56] because all HipSci lines were derived from skin. We observed weak relationships between neuronal differentiation efficiency and other biological factors, including X chromosome inactivation status, which was described as relevant for other lineages[2]. However, our results clearly demonstrate that variability in differentiation outcomes is due to cell-intrinsic factors that are maintained over multiple freeze–thaw cycles. We found that this was unlikely to be the result of donor-specific effects, as there was poor correspondence in predicted differentiation outcomes between lines derived from the same individual (Extended Data Fig. 4). Additionally, we did not detect significant effects in a genome-wide association analysis with predicted differentiation outcomes. Given these results, we suggest the two most likely candidates for future investigation are somatic genetic changes or persistent epigenetic changes that arise early in cellular reprogramming or under suboptimal culture conditions.

Our analysis identified a negative association between neuronal differentiation efficiency at day 52 and the proportion of cells stemming from a specific subpopulation (cluster 2) of pluripotent cells that express the transcription factor *UTF1* and other genes at elevated levels. Counterintuitively, the abundance of cluster 2 cells was positively correlated with the proportion of neuroblast cells on day 11. One possible explanation is that cell lines that commit earlier to a neuronal fate disproportionately lose neurons upon passaging at day 20. We speculate that culture methods that reduce iPSC heterogeneity may reduce the fraction of iPSC lines that resist efficient neuronal differentiation. We note that our findings do not explain all of the variance in neuronal differentiation capacity, and future studies will be required to better understand the biological basis of the differentiation bias observed here.

Based on molecular markers that predict differentiation bias, we estimate that 13% of iPSC lines in the HipSci resource produce very few neuronal cell types under the conditions tested. Importantly, these predictions generalize to previously untested lines. While the production of neuronal cells was intrinsically limited in these cell lines, the fact that this effect was associated with particular cell lines but not with particular donors suggests that cell banks that contain multiple lines per donor can be most effectively used for applications involving neural differentiation by the rational selection of cell clones. This a priori selection is enabled by gene expression profiling data from the pluripotent state that is easily obtainable and often already available.

In summary, our study demonstrates how iPSC differentiation combined with scRNA-seq unlocks population-level studies in complex, dynamic and biologically realistic cellular models. We anticipate that future uses of this model system will focus on experimental settings that are challenging or impossible with primary cells. These could include high-resolution sampling along extended differentiation times to more complex differentiation trajectories or systems, such as organoids, or involve large panels of disease-relevant stimuli and drug exposures.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00801-6.

Received: 5 May 2020; Accepted: 25 January 2021;
Published online: 4 March 2021

References

1. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
2. Cuomo, A. S. E. et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 810 (2020).
3. Strober, B. J. et al. Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).
4. Schwartzentruber, J. et al. Molecular and functional variation in iPSC-derived sensory neurons. *Nat. Genet.* **50**, 54–61 (2018).
5. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
6. D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. In vitro differentiation of human iPSC-derived retinal pigment epithelium cells (iPSC-RPE). *Bio-protocol* **9**, e3469 (2019).
7. Banovich, N. E. et al. Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res.* **28**, 122–131 (2018).
8. Volpato, V. et al. Reproducibility of molecular phenotypes after long-term differentiation to human iPSC-derived neurons: a multi-site omics study. *Stem Cell Rep.* **11**, 897–911 (2018).
9. Nguyen, Q. H. et al. Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. *Genome Res.* **28**, 1053–1066 (2018).
10. Mitchell, J. M. et al. Mapping genetic effects on cellular phenotypes with 'cell villages'. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.29.174383 (2020).
11. Osborn, T. & Hallett, P. J. Seq-ing markers of midbrain dopamine neurons. *Cell Stem Cell* **20**, 11–12 (2017).
12. Stoddard-Bennett, T. & Pera, R. R. Stem cell therapy for Parkinson's disease: safety and modeling. *Neural Regen. Res.* **15**, 36–40 (2020).

13. Kriks, S. et al. Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson's disease. *Nature* **480**, 547–551 (2011).
14. Xiong, N. et al. Mitochondrial complex I inhibitor rotenone-induced toxicity and its potential mechanisms in Parkinson's disease models. *Crit. Rev. Toxicol.* **42**, 613–632 (2012).
15. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
16. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
17. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
18. La Manno, G. et al. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167**, 566–580 (2016).
19. Park, C.-H. et al. Acquisition of in vitro and in vivo functionality of Nurr1-induced dopamine neurons. *FASEB J.* **20**, 2553–2555 (2006).
20. Ramonet, D. et al. PARK9-associated ATP13A2 localizes to intracellular acidic vesicles and regulates cation homeostasis and neuronal integrity. *Hum. Mol. Genet.* **21**, 1725–1743 (2012).
21. Arenas, E., Denham, M. & Villaescusa, J. C. How to make a midbrain dopaminergic neuron. *Development* **142**, 1918–1936 (2015).
22. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
23. Cummings, K. J. & Hodges, M. R. The serotonergic system and the control of breathing during development. *Respir. Physiol. Neurobiol.* **270**, 103255 (2019).
24. Campbell, J. N. et al. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.* **20**, 484–496 (2017).
25. Sloan, S. A. et al. Human astrocyte maturation captured in 3D cerebral cortical spheroids derived from pluripotent stem cells. *Neuron* **95**, 779–790 (2017).
26. Zhang, Y. et al. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* **89**, 37–53 (2016).
27. Bertrand, N., Castro, D. S. & Guillemot, F. Proneural genes and the specification of neural cell types. *Nat. Rev. Neurosci.* **3**, 517–530 (2002).
28. Lacomme, M., Liaubet, L., Pituello, F. & Bel-Vialar, S. NEUROG2 drives cell cycle exit of neuronal precursors by specifically repressing a subset of cyclins acting at the G_1 and S phases of the cell cycle. *Mol. Cell. Biol.* **32**, 2596–2607 (2012).
29. Sherer, T. B. et al. Mechanism of toxicity in rotenone models of Parkinson's disease. *J. Neurosci.* **23**, 10756–10764 (2003).
30. Knönagel, H. & Karmann, U. Autologous blood transfusions in interventions of the pelvis using the cell saver. *Helv. Chir. Acta* **59**, 485–488 (1992).
31. Cannon, J. R. et al. A highly reproducible rotenone model of Parkinson's disease. *Neurobiol. Dis.* **34**, 279–290 (2009).
32. D'Antonio-Chronowska, A. et al. Association of human iPSC gene signatures and X chromosome dosage with two distinct cardiac differentiation trajectories. *Stem Cell Rep.* **13**, 924–938 (2019).
33. Ye, W., Shimamura, K., Rubenstein, J. L., Hynes, M. A. & Rosenthal, A. FGF and Shh signals control dopaminergic and serotonergic cell fate in the anterior neural plate. *Cell* **93**, 755–766 (1998).
34. He, Z. & Yu, Q. Identification and characterization of functional modules reflecting transcriptome transition during human neuron maturation. *BMC Genomics* **19**, 262 (2018).
35. Lancaster, M. A. et al. Guided self-organization and cortical plate formation in human brain organoids. *Nat. Biotechnol.* **35**, 659–666 (2017).
36. Müller, F.-J. et al. A bioinformatic assay for pluripotency in human cells. *Nat. Methods* **8**, 315–317 (2011).
37. Bonder, M. J. et al. Identification of rare and common regulatory variants using population-scale transcriptomics of pluripotent cells. *Nat. Genet.* https://doi.org/10.1038/s41588-021-00800-7 (2021).
38. Sarkar, A. K. et al. Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet.* **15**, e1008045 (2019).
39. Miller, D. J. & Fort, P. E. Heat shock proteins regulatory role in neurodevelopment. *Front. Neurosci.* **12**, 821 (2018).
40. Bartelt-Kirbach, B. et al. HspB5/αB-crystallin increases dendritic complexity and protects the dendritic arbor during heat shock in cultured rat hippocampal neurons. *Cell. Mol. Life Sci.* **73**, 3761–3775 (2016).
41. Shimura, H., Miura-Shimura, Y. & Kosik, K. S. Binding of tau to heat shock protein 27 leads to decreased concentration of hyperphosphorylated tau and enhanced cell survival. *J. Biol. Chem.* **279**, 17957–17962 (2004).
42. Wilhelmus, M. M. M. et al. Small heat shock proteins inhibit amyloid-β protein aggregation and cerebrovascular amyloid-β protein toxicity. *Brain Res.* **1089**, 67–78 (2006).
43. Tucci, S. Brain metabolism and neurological symptoms in combined malonic and methylmalonic aciduria. *Orphanet J. Rare Dis.* **15**, 27 (2020).
44. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
45. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
46. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
47. Kory, N. et al. SFXN1 is a mitochondrial serine transporter required for one-carbon metabolism. *Science* **362**, eaat9528 (2018).
48. Palmer, G., Horgan, D. J., Tisdale, H., Singer, T. P. & Beinert, H. Studies on the respiratory chain-linked reduced nicotinamide adenine dinucleotide dehydrogenase. XIV. Location of the sites of inhibition of rotenone, barbiturates, and piericidin by means of electron paramagnetic resonance spectroscopy. *J. Biol. Chem.* **243**, 844–847 (1968).
49. Betarbet, R. et al. Chronic systemic pesticide exposure reproduces features of Parkinson's disease. *Nat. Neurosci.* **3**, 1301–1306 (2000).
50. Ma, D. K., Ponnusamy, K., Song, M.-R., Ming, G.-L. & Song, H. Molecular genetic analysis of FGFR1 signalling reveals distinct roles of MAPK and PLCγ1 activation for self-renewal of adult neural stem cells. *Mol. Brain* **2**, 16 (2009).
51. Stachowiak, E. K. et al. Cerebral organoids reveal early cortical maldevelopment in schizophrenia-computational anatomy and genomics, role of FGFR1. *Transl. Psychiatry* **7**, 6 (2017).
52. International Stem Cell Initiative. Assessment of established techniques to determine developmental and malignant potential of human pluripotent stem cells. *Nat. Commun.* **9**, 1925 (2018).
53. Tsankov, A. M. et al. A qPCR ScoreCard quantifies the differentiation potential of human pluripotent stem cells. *Nat. Biotechnol.* **33**, 1182–1192 (2015).
54. Bock, C. et al. Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439–452 (2011).
55. Kajiwara, M. et al. Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells. *Proc. Natl Acad. Sci. USA* **109**, 12538–12543 (2012).
56. Hu, S. et al. Effects of cellular origin on differentiation of human induced pluripotent stem cell-derived endothelial cells. *JCI Insight* **1**, e85558 (2016).

HipSci Consortium

Daniel J. Gaffney[2,11] and Oliver Stegle[2,3,8,9,11]

Methods

Human iPSC lines. Human iPSCs were obtained from the HipSci project[1] (http://www.hipsci.org, Supplementary Table 10). Briefly, primary fibroblasts from skin biopsies were collected from consenting research volunteers recruited from the NIHR Cambridge BioResource. iPSC derivation was performed either using the Sendai reprogramming kit or, for two lines (HPSI0213i-nawk_55, HPSI0813i-ffdc_1), with episomal plasmids. Following transfer to feeder-free culture and expansion, each line was submitted to quality control, and the criteria for line selection were (1) level of pluripotency, as determined by the PluriTest assay[36], (2) number of copy number abnormalities and (3) the ability to differentiate into each of the three germ layers (Kilpinen et al.[1]). All lines selected were healthy and of European descent.

Human iPSC culture. Lines were thawed onto tissue culture-treated plates (Corning, 3516) coated with 10 µg ml^{-1} Vitronectin XF (Stemcell Technologies, 07180) using complete Essential 8 (E8) medium (Thermo Fisher, A1517001) and 10 µM ROCK inhibitor (Sigma, Y0503). After thawing, cells were expanded in E8 medium for at least two passages, using 0.5 µM EDTA, pH 8.0 (Thermo Fisher, 15575-020) for cell dissociation. The last passage was always 3–4 d before plating for differentiation. Cell line synchronization was performed by adjusting the splitting ratio of each line, aiming to reach 60–80% of confluence on the pooling day.

Pooling and differentiation of midbrain DAs. iPSC colonies were dissociated into a single-cell suspension using Accutase (Thermo Fisher, A11105-01) and resuspended in E8 medium containing 10 µM ROCK inhibitor. Cells were counted using an automated cell counter (Chemometec NC-200), and a cell suspension containing an equal amount of each iPSC line was prepared in E8 medium containing 10 µM ROCK inhibitor and seeded at 2×10^5 cells per cm^2 on plates coated with 1% Geltrex (Thermo Fisher, A1413202). Each pool of lines contained cells from between seven to 24 donors. After plating (24 h), neuronal differentiation of the pooled lines to a midbrain lineage was performed as described in ref. [13] with the following minor modifications: (1) SHH C25II was replaced with 100 nM SAG (Tocris, 6390) in the neuronal induction phase, and (2) on day 20, the cells were passaged with Accutase containing 20 U ml^{-1} papain (Worthington, LK00031765) and plated at 3.5×10^5 cells per cm^2 on plates coated with 1% Geltrex for final maturation. The step-by-step protocol can be found at https://www.protocols.io/view/generation-of-ipsc-derived-dopaminergic-neurons-bjpgkmjw.

Rotenone stimulation. On day 51 of differentiation, cells were exposed for 24 h to freshly prepared 0.1 µM rotenone (Sigma, R8875, HPLC purity ≥95%) diluted in neuronal maturation medium[13]. The final DMSO concentration was 0.01% in all exposure conditions. Unstimulated control samples (that is, treated with DMSO only) were taken concurrently.

Generation of cerebral organoids. Cerebral organoids were generated according to the enCOR method as previously described[35]. Briefly, one pool of 18 iPSC lines was thawed and expanded for one passage before seeding 18,000 cells onto PLGA microfilaments prepared from Vicryl sutures. The STEMdiff Cerebral Organoid kit (Stemcell Technologies, 08570) was used for organoid culture with timing according to manufacturer's suggestion and Matrigel embedding as previously described[57]. From day 35, the medium was supplemented with 2% dissolved Matrigel basement membrane (Corning, 354234) and processed for scRNA-seq after 113 d of culture.

Generation of single-cell suspensions for sequencing. On collection days, cells were washed once with 1× Dulbecco's (D) PBS (Thermo Fisher, 14190-144) before adding either Accutase (day 11) or Accutase containing 20 U ml^{-1} papain (days 30 and 52). The cells were incubated at 37 °C for up to 20 min (day 11) or up to 35 min (days 30 and 52) before adding DMEM/F12 (Thermo Fisher Scientific, 10565-018) supplemented with 10 µM ROCK inhibitor and 33 µg ml^{-1} DNase I (Worthington, LK003170, only for cells collected on days 30 and 52). The cells were dissociated using a P1000 and collected in a 15-ml tube capped with a 40-µm cell strainer. After centrifugation, the cells were resuspended in 1× DPBS containing 0.04% BSA (Sigma, A0281) and washed three additional times in 1× DPBS containing 0.04% BSA. Single-cell suspensions were counted using an automated cell counter (Chemometec NC-200), and concentrations were adjusted to 5×10^5 cells per ml.

Organoids were washed twice in 1× DPBS before adding EBSS (Worthington, LK003188) dissociation buffer containing 19 U ml^{-1} papain, 50 µg ml^{-1} DNase I and 22.5× Accutase. Organoids were incubated in a shaking block (750 r.p.m.) at 37 °C for 30 min. Every 10 min, the organoids were triturated using a P1000 and BSA-coated pipette tips until large clumps were dissociated. Dissociated organoids were transferred into a new tube capped with a 40-µm cell strainer and pelleted for 4 min at 300*g*. After centrifugation, cells were resuspended in EBSS containing 50 µg ml^{-1} DNase I and 2 mg ml^{-1} ovomucoid (Worthington, LK003150). EBSS (0.5 volume), followed by 0.5 volume of 20 mg ml^{-1} ovomucoid, was added to the top of the cell suspension, and the cells were mixed by flicking the tube. After centrifugation, the cells were resuspended in 1× DPBS containing 0.04% BSA. Single-cell suspensions were counted using an automated cell counter, and concentrations were adjusted to 5×10^5 cells per ml. The step-by-step protocol can be found at https://www.protocols.io/view/dissociation-of-neuronal-culture-to-single-cells-f-bh32j8qe.

Immunohistochemistry. Cells were fixed in 4% paraformaldehyde (Thermo Fisher Scientific, 28908) for 15 min, rinsed three times with 1× PBS (Sigma, D8662) and blocked with 5% normal donkey serum (AbD Serotec, C06SBZ) in PBST (1× PBS with 0.1% Triton X-100, Sigma, 93420) for 2 h at room temperature. Primary antibodies were diluted in PBST containing 1% normal donkey serum and incubated overnight at 4 °C. Cells were washed five times with 1× PBS and incubated with secondary antibodies diluted in 1× PBS for 45 min at room temperature. Cells were washed three more times with 1× PBS, and Hoechst stain (Thermo Fisher Scientific, H3569) was used to visualize cell nuclei. Image acquisition was performed using a Cellomics ArrayScan VTI (Thermo Fisher Scientific).

The following antibodies were used: anti-FOXA2 (Santa Cruz, sc101060, 1:100), anti-LMX1A (Millipore, AB10533, 1:500), anti-TH (Santa Cruz, sc-25269, 1:200), anti-MAP2 (Abcam, 5392, 1:2,000), donkey anti-chicken AF647 (Thermo Fisher Scientific, A21449), donkey anti-mouse AF488 (Thermo Fisher Scientific, A11008), donkey anti-mouse AF555 (Thermo Fisher Scientific, A31570), donkey anti-rabbit AF488 (Thermo Fisher Scientific, A21206) and donkey anti-rabbit AF555 (Thermo Fisher Scientific, A27039).

Chromium 10x Genomics library and sequencing. Single-cell suspensions were processed by the Chromium Controller (10x Genomics) using the Chromium Single Cell 3′ Reagent kit version 2 (PN-120237). On average, 15,000 cells from each 10x reaction were directly loaded into one inlet of the 10x Genomics chip (Supplementary Table 1). All steps were performed according to the manufacturer's specifications. Barcoded libraries were sequenced using HiSeq 4000 (Illumina, one lane per 10x chip position) with 50-bp or 75-bp paired-end reads to an average depth of 40,000–60,000 reads per cell.

Single-cell data preprocessing. Sequencing data generated from the Chromium 10x Genomics libraries (see above) were processed using Cell Ranger software (version 2.1.0) and aligned to the GRCh37/hg19 reference genome. Counts were quantified with the Cell Ranger 'count' command, using the Ensembl 84 reference transcriptome (32,738 genes) with default parameters.

For each of 17 pooled experiments, donors (that is, cell lines) were demultiplexed using demuxlet[16], using genotypes of common (MAF > 1%) exonic variants available from the HipSci bank, and a prior doublet rate of 0.05. Only cells with successful donor identification were retained for further analysis.

Further quality control steps led to the exclusion of seven 10x samples, where a 10x sample is defined as the cells sequenced from one inlet of a 10x chip. In particular, samples for pool 10 on day 11 were excluded because of an issue in the library preparation. Samples for pool 12 on day 52 were excluded on the basis of low cell viability (72.1% in the rotenone-stimulated sample) and outlying gene expression (the first principal component in gene expression separated this sample from others at the same time point). One 10x reaction for pool 1 on day 30 was excluded on the basis of low quality, with <30% of cells successfully mapping to a cell line. Finally, cells from an outlier cell line (HPSI0913i-gedo_33) were excluded. This cell line contributed 91% of cells to pool 14 and had outlying gene expression, suggestive of a large-effect somatic mutation (Supplementary Table 1).

Normalization, dimensionality reduction and clustering. Two sets of analyses were performed: (1) analysis of each time point independently and (2) a combined analysis of a subsample of 20% of cells from all time points (used only for visualization purposes; Fig. 1).

First, independent analysis of time points allowed efficient batch effect correction (as all samples from the same time point contained similar mixtures of cell types), as well as reductions in computational demands. Counts were normalized to the total number of counts per cell using the Scanpy function pp.normalize_per_cell, and only genes with non-zero counts in at least 0.5% of cells were retained. The top 3,000 most variable genes were then selected, after controlling for mean–variance dependence in expression data using the Scanpy function pp.filter_genes_dispersion. The first 50 principal components were calculated. Batch correction was applied on the level of principal components using Harmony[17], with each 10x sample treated as a distinct batch. UMAP and clustering were performed using these transformed principal components. Clustering was performed using Louvain clustering with ten nearest neighbors, identifying 26 clusters across time points (six, seven and 13 clusters, respectively, at days 11, 30 and 52, Extended Data Fig. 1). Analysis steps besides batch correction were carried out using the Scanpy package (version 1.4)[58]. Cell type annotation was performed using a literature-curated set of relevant marker genes (Supplementary Methods)[59–69].

For the combined analysis of all time points (for visualization purposes only), the same steps were applied, except that only a random subsample of 20% of cells was included in the analysis (following filtering for cells with donor assignment), and the definition of batches for the Harmony batch correction step was modified. In particular, for each batch to have a similar mixture of cell types, each pool (rather than each 10x sample) was considered as a distinct batch.

Batch correction and clustering of the organoid dataset. The same steps described above were applied to the cerebral organoid data. This identified eight clusters that were mapped to different cell types (neurons, intermediate progenitor cells, radial glial progenitor cells, satellite cells, mesenchymal cells, myotube and Wnt-positive and PAX7-positive cells) using 24 marker genes (Supplementary Fig. 5).

Batch correction and clustering of the two single-cell iPSC datasets. The same dimensionality reduction, batch correction and clustering steps described above were applied to the two single-cell iPSC datasets analyzed[2,38] (Extended Data Fig. 5 and Supplementary Fig. 6). For the Cuomo et al. dataset[2], normalized (by CPM) and log-scaled data were taken from the original publication, and no further normalization was performed. For the Sarkar et al. dataset[38], count data were normalized and log scaled as described above for our data (normalized to total counts per cell). Note that, in both cases, only cells that passed quality control (as defined in the original publications) from these datasets were included. This analysis identified five and four clusters in the two different datasets, respectively.

Definition of neuronal differentiation efficiency. We computed cell type proportions for each cell line in each pool (that is, all (cell line, pool) combinations) at each time point. Based on these proportions (cell line, pool), combinations were clustered based on Euclidean distance (Fig. 3b and Extended Data Fig. 3). Only (cell line, pool) combinations for which at least ten cells were present at all time points were included in the heatmap in Extended Data Fig. 3a and in the principal component analysis shown in Extended Data Fig. 3b–d.

Neuronal differentiation efficiency was defined as the sum of the proportion of serts and DAs present on day 52. The neuronal differentiation efficiency of iPSC lines was calculated as the average of the efficiencies across all pools in which that cell line was included.

Predictive model of neuronal differentiation efficiency from iPSC gene expression. A logistic regression classifier was trained to predict midbrain neuron differentiation failure (neuronal differentiation efficiency <0.2, from above definition) of iPSC lines from their gene expression at the iPSC stage (using independent bulk RNA-seq data[37]). For cell lines that were differentiated multiple times, average values of neuronal differentiation efficiency across replicate experiments were used. Gene expression data were available from independent bulk experiments. The feature set was all expressed genes (that is, genes with mean $\log_2(\mathrm{TPM}+1)>2$, $n=13{,}475$), and the model was trained using the scikit-learn (version 0.21.3) Python package (sklearn.linear_model.LogisticRegression). L1 regularization was used, with default parameter settings (inverse regulation strength, 1.0). Precision–recall was evaluated using leave-one-out cross-validation. We also considered the same model trained on the first half of the dataset (pools 1–8) and assessed its performance on the second half of the dataset (pools 9–17), which was generated sequentially over the course of the study (Extended Data Fig. 3e). This approach yielded similar results.

When using the trained model for making predictions, we used the predicted score that corresponds to recall of 35% and precision of 100% (Extended Data Fig. 3c,d). When predicted scores for two lines from the same donor were present, we classified donors into concordant good differentiators when both lines had positive differentiation scores (>0.02231; $n=209$) and concordant poor differentiators when they were both predicted to be poor differentiators (<0.02231; $n=13$). Finally, 'discordant' donors were donors for which one line was predicted to be a bad differentiator and the other was not ($n=49$). Bulk RNA-seq expression of *UTF1* and *TAC3* for these lines was consistent with these predictions (Extended Data Fig. 4).

Finally, we considered data from two additional pools (pools 20 and 21) that were not considered in the main analysis. These pools included data from 16 lines in total, 11 of which were not contained in any other pool. We applied the predictor trained on the main dataset to obtain predicted scores of neuronal differentiation capacity for these lines and compared the predicted scores to the observed neuronal differentiation efficiency. For ten of 11 lines, the predicted differentiation score indicated potent differentiation (score >0.02231), and all but one (ten of 11) demonstrated positive neuronal differentiation (neuronal differentiation efficiency >0.2). One line was predicted to be a poor differentiator (score <0.02231) and indeed performed poorly (neuronal differentiation efficiency <0.2). Overall, this corresponds to 100% precision at 50% recall, which is within expected ranges based on the chosen thresholds.

Differential expression and gene set enrichment analysis. Differentiation expression analysis between each cluster and the others in the single-cell iPSC datasets was performed using Scanpy's function 'tl.rank_genes_groups', grouping by each cluster at a time[58]. Multiple-testing correction was performed using the Benjamini–Hochberg approach with an FDR threshold of 5% (Supplementary Table 6). We used gprofiler[70] (https://biit.cs.ut.ee/gprofiler/gost) to perform biological process enrichment analysis using all upregulated genes with a log fold change greater than or equal to one as an input. The top 20 hits are presented in Supplementary Table 11.

Cis-eQTL mapping. For cis-eQTL mapping, we followed the procedure in Cuomo et al.[2] and adopted a strategy similar to approaches commonly applied in conventional bulk eQTL analyses[1]. We considered common variants (MAF > 5%) within a cis region spanning 250 kb up- and downstream of the gene body for cis-QTL analysis. Association tests were performed using a linear mixed model, adapting the approach in ref. [2]. Specifically, we considered an additional random effect term to account for varying numbers of cells per donor. Briefly, for each donor, we introduced a variance term $1\times n^{-1}$, accounting for the varying numbers of cells used to estimate mean expression level for each donor. All models were fitted using LIMIX[71,72], using likelihood ratio tests to assess significance. To adjust for experimental batch effects across samples, we included the first 15 principal components calculated on the expression values in the model as fixed effect covariates. To adjust for multiple testing, we employed an approximate permutation scheme, analogous to the approach proposed in ref. [73]. Briefly, for each gene, we generated 1,000 permutations of the genotypes while retaining the relationship between covariates, random effect terms and expression values. We then adjusted for multiple testing using this empirical null distribution. To control for multiple testing across genes, the Storey's Q value procedure was applied[74]. Genes with significant eQTL were reported at FDR < 5%.

eQTL mapping was performed as described above for 14 contexts (cell type, time point and stimulus status), considering six major cell types (top four cell types per condition with at least 20% cells; Supplementary Table 7). Gene expression for each donor was calculated as the mean of log-transformed counts-per-cell-normalized expression across cells (including cells from different pools, when applicable). A line was considered in the eQTL analysis of given context (cell type, condition) if at least ten cells were captured from the corresponding context. Genes were considered for eQTL analysis if they were expressed (UMI count >0) in at least 1% of cells across all lines (10,993–12,789 genes tested across contexts). Lead associations per eGene and context are reported in Supplementary Table 7. The robustness of this eQTL analysis approach was confirmed by comparison to alternative eQTL mapping strategies (Supplementary Methods and Extended Data Fig. 7).

Sharing of eQTL signal between cell types and with GTEx brain tissues. To quantify the extent of sharing between eQTL maps, MASHR software was used[44]. For all MASHR analyses, lead eQTL SNPs were considered per gene and context. Four random SNPs per gene were selected as a background for the calculation of the data-driven covariance; a canonical covariance matrix was also included as recommended (https://stephenslab.github.io/mashr/articles/eQTL_outline.html.). Next, we extracted the posterior β values using MASHR and estimated pairwise sharing between conditions and/or tissues. Effects were defined as shared if the effect direction and effect size were within a factor of 0.5 of each other, when considering SNP–gene pairs passing a significance threshold (local false sign rate <0.05) in at least one of the two conditions considered.

We applied this workflow to perform the following three comparisons.

- eQTL maps from our cell types and conditions ($n=14$ eQTL maps): 9,641 genes assessed (Extended Data Fig. 6c)
- eQTL maps from all cell types and conditions ($n=14$ eQTL maps), as well as bulk RNA-seq-derived eQTL in iPSCs, as well as all of the GTEx brain tissue data ($n=13$ eQTL maps): 7,975 genes assessed (Fig. 4d and Supplementary Fig. 7b)
- All of our cell types and conditions ($n=14$ eQTL maps), as well as all of the GTEx tissues ($n=49$ eQTL maps): 7,586 genes assessed (Supplementary Fig. 7a)

As an alternative strategy to assess the sharing of eQTL, we assessed the fraction of eQTL from GTEx brain tissues that were recapitulated in our eQTL maps. Briefly, we considered a nominal definition of eQTL replication, based on nominal significance of lead eQTL variants discovered in each of the 13 GTEx brain tissue eQTL maps in each of our 14 eQTL maps (Extended Data Fig. 8a,b). Notably, this comparison allows for distinguishing lack of replication versus lack of assessment of an eGene because of differences in expression. Among eGenes identified at FDR < 5% in each of the GTEx maps, approximately 50% were tested in the eQTL contexts in this study. For the shared fraction of genes assessed, 20–40% eQTL were nominally significant ($P<0.05$) across the 14 maps. Cumulatively, this means that 10–20% of the eQTL from a GTEx brain map could be rediscovered in our single-cell maps.

Colocalization analysis between neuro-related GWAS traits. We collected summary statistics for 25 GWAS traits that were either neurodegenerative and/or neuropsychiatric diseases or related to behavior and intelligence and that had at least five genome-wide significant loci ($P=5.0\times10^{-8}$; Supplementary Table 8). We then defined GWAS subthreshold loci as 1-Mb-wide genomic windows with at least one SNP at $P<10^{-6}$, centering the window around the index variant (variant with minimum P value in the window). If there were multiple subthreshold loci within a 1-Mb window, we merged them and took the index variant with the minimum P value overall. Statistical colocalization analysis between 14 eQTL maps from our study and 48 eQTL maps from GTEx (version 7, https://www.gtexportal.org/home/datasets) and those GWAS loci was performed using the coloc package[46], implemented in R with default hyperparameter settings. We tested any gene for which the transcription start site and eQTL lead variant (SNP with minimum P value for the gene) were both within the 1-Mb window centered at each GWAS

index variant. We tested all SNPs located between the GWAS index variant and the top eQTL (within the window) variant with 500-kb extensions on either side. We matched SNPs between eQTL and GWAS based on chromosomal position and reference and/or alternative alleles. Genes with the posterior probability of colocalization (PP4) greater than 0.5 were defined as GWAS colocalized.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Managed access data from scRNA-seq are accessible in the European Genome–phenome Archive (EGA, https://www.dev.ebi.ac.uk/ega/) under the study number EGAS00001002885 (dataset EGAD00001006157). Open access scRNA-seq data are available in the European Nucleotide Archive (ENA) under the study ERP121676 (https://www.ebi.ac.uk/ena/browser/view/PRJEB38269). Processed single-cell count data and eQTL and colocalization summary statistics are available on Zenodo at https://zenodo.org/record/4333872. The two iPSC single-cell datasets are available on Zenodo (https://zenodo.org/record/3625024) and GEO (GSE118723) for the datasets described in Cuomo et al.[2] and Sarkar et al.[38], respectively. iPSC bulk RNA-seq data from Bonder et al.[37] are available on the EGA (study ID, EGAS00001000593, https://www.ebi.ac.uk/ega/studies/EGAS00001000593) and the ENA (ERP007111, https://www.ebi.ac.uk/ena/browser/view/PRJEB7388). Chip genotypes for HipSci lines are available from the EGA (EGAS00001000866) and the NCBI (PRJEB11750).

Code availability

All scripts are available in the following github repository: https://github.com/single-cell-genetics/singlecell_neuroseq_paper/. The standalone predictor for neuronal differentiation capacity is available at https://github.com/single-cell-genetics/singlecell_neuroseq_paper/tree/master/differentiation_prediction_model/. The eQTL mapping pipeline is available at https://github.com/single-cell-genetics/limix_qtl/.

References

57. Lancaster, M. A. & Knoblich, J. A. Generation of cerebral organoids from human pluripotent stem cells. *Nat. Protoc.* **9**, 2329–2340 (2014).
58. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
59. Ferri, A. L. M. et al. Foxa1 and Foxa2 regulate multiple phases of midbrain dopaminergic neuron development in a dosage-dependent manner. *Development* **134**, 2761–2769 (2007).
60. Andersson, E. et al. Identification of intrinsic determinants of midbrain dopamine neurons. *Cell* **124**, 393–405 (2006).
61. Loo, L. et al. Single-cell transcriptomic analysis of mouse neocortical development. *Nat. Commun.* **10**, 134 (2019).
62. Ren, J. et al. Single-cell transcriptomes and whole-brain projections of serotonin neurons in the mouse dorsal and median raphe nuclei. *eLife* **8**, e49424 (2019).
63. Huang, K. W. et al. Molecular and anatomical organization of the dorsal raphe nucleus. *eLife* **8**, e46464 (2019).
64. Okaty, B. W. et al. A single-cell transcriptomic and anatomic atlas of mouse dorsal raphe neurons. *eLife* **9**, e55523 (2020).
65. Mercurio, S., Serra, L. & Nicolis, S. K. More than just stem cells: functional roles of the transcription factor Sox2 in differentiated glia and neurons. *Int. J. Mol. Sci.* **20**, 4540 (2019).
66. Wu, Y., Liu, Y., Levine, E. M. & Rao, M. S. Hes1 but not Hes5 regulates an astrocyte versus oligodendrocyte fate choice in glial restricted precursors. *Dev. Dyn.* **226**, 675–689 (2003).
67. Wiese, S., Karus, M. & Faissner, A. Astrocytes as a source for extracellular matrix molecules and cytokines. *Front. Pharmacol.* **3**, 120 (2012).
68. Redies, C. Cadherins in the central nervous system. *Prog. Neurobiol.* **61**, 611–648 (2000).
69. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
70. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
71. Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. Preprint at *bioRxiv* https://doi.org/10.1101/003905 (2014).
72. Casale, F. P., Rakitsch, B., Lippert, C. & Stegle, O. Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods* **12**, 755–758 (2015).
73. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
74. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).

Acknowledgements

All data for this study were generated under the Open Targets project OTAR039. J.J. was supported by a postdoctoral fellowship from Open Targets, A.S.E.C. was supported by a PhD fellowship from the EMBL International PhD Programme, and D.D.S. was supported by a postdoctoral fellowship from the EMBL Interdisciplinary Postdoctoral Programme. M.A.L. was funded by the Medical Research Council (MC_UP_1201/9). N.K. and D.J.G. were funded by the Wellcome Trust grant WT206194. F.T.M. is a New York Stem Cell Foundation, Robertson Investigator and is supported by the New York Stem Cell Foundation (NYSCF-R-156), the Wellcome Trust and Royal Society (211221/Z/18/Z) and the Chan Zuckerberg Initiative (191942) and by the NIHR Cambridge BRC. J.C.M. acknowledges core support from EMBL and Cancer Research UK (C9545/A29580). O.S. is supported by core funding from EMBL and the DKFZ, as well as the BMBF, the Volkswagen Foundation and the EU (810296). We thank the MRC Metabolic Diseases Unit Imaging Core Facility for assistance with imaging. We thank the staff at the Cellular Generation and Phenotyping and Sequencing core facilities at the Wellcome Sanger Institute and the imaging core facility of the Wellcome–MRC Institute of Metabolic Science. We thank H. Kilpinen and P. Puigdevall Costa for useful discussions regarding data analysis.

Author contributions

The main analyses and data preparation were performed by J.J., D.D.S. and A.S.E.C. N.K. performed the colocalization analysis. Cell culture experiments were performed by J.J., J.H., J.S., D.P. and M.A., and M.A.L. performed the experimental work on the organoid dataset. J.J. and M.P. oversaw the cell culture experiments. E.M. and M.G. processed GWAS summary statistics for colocalization analysis. J.J., D.D.S., A.S.E.C., J.C.M., F.T.M., O.S. and D.J.G. wrote the manuscript; N.K. assisted in editing the manuscript; J.J., F.T.M., O.S. and D.J.G. conceived and oversaw the study.

Competing interests

D.J.G. and E.M. were employees of Genomics PLC, and D.D.S. was an employee of GSK at the time the manuscript was submitted.

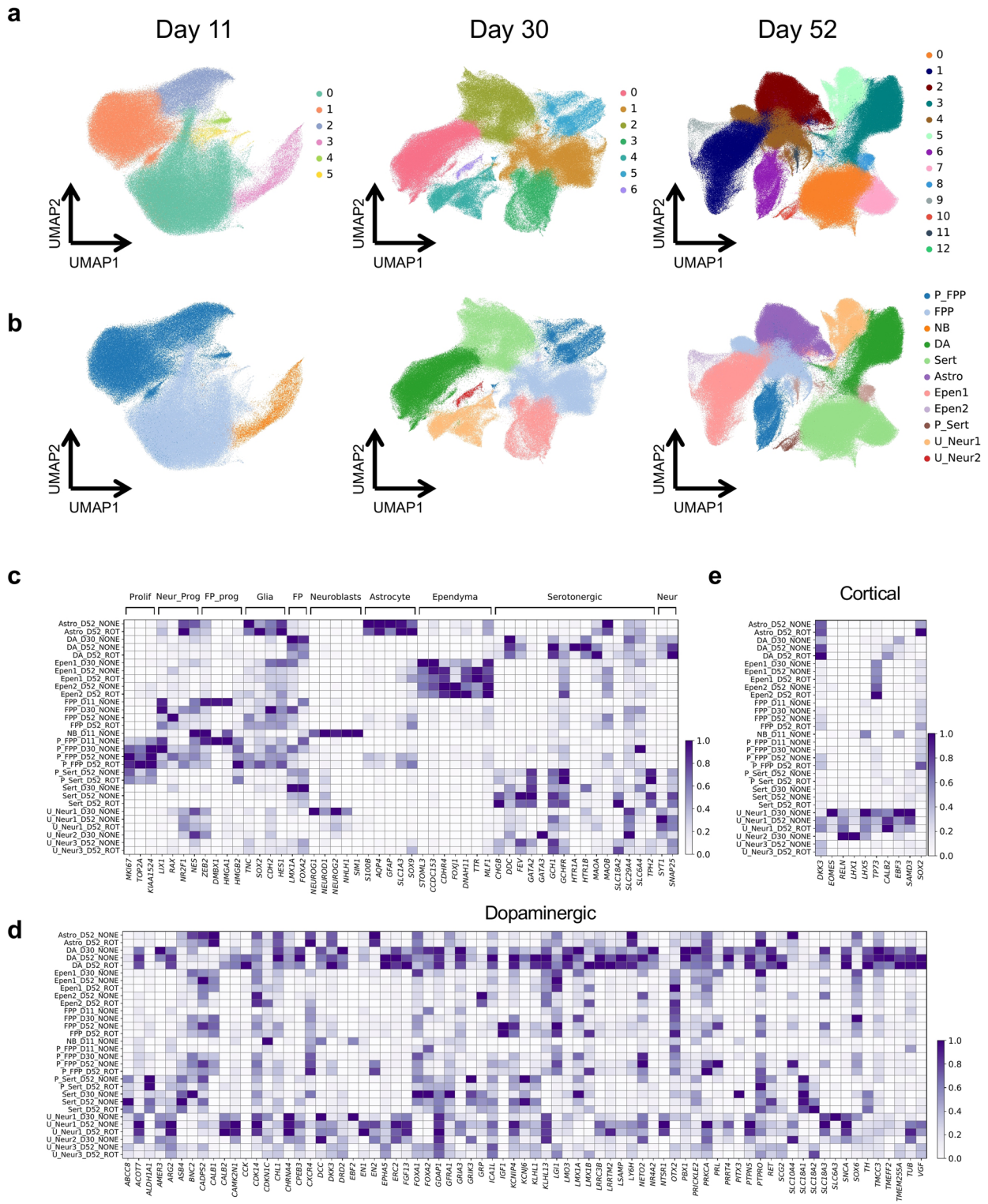
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00801-6.

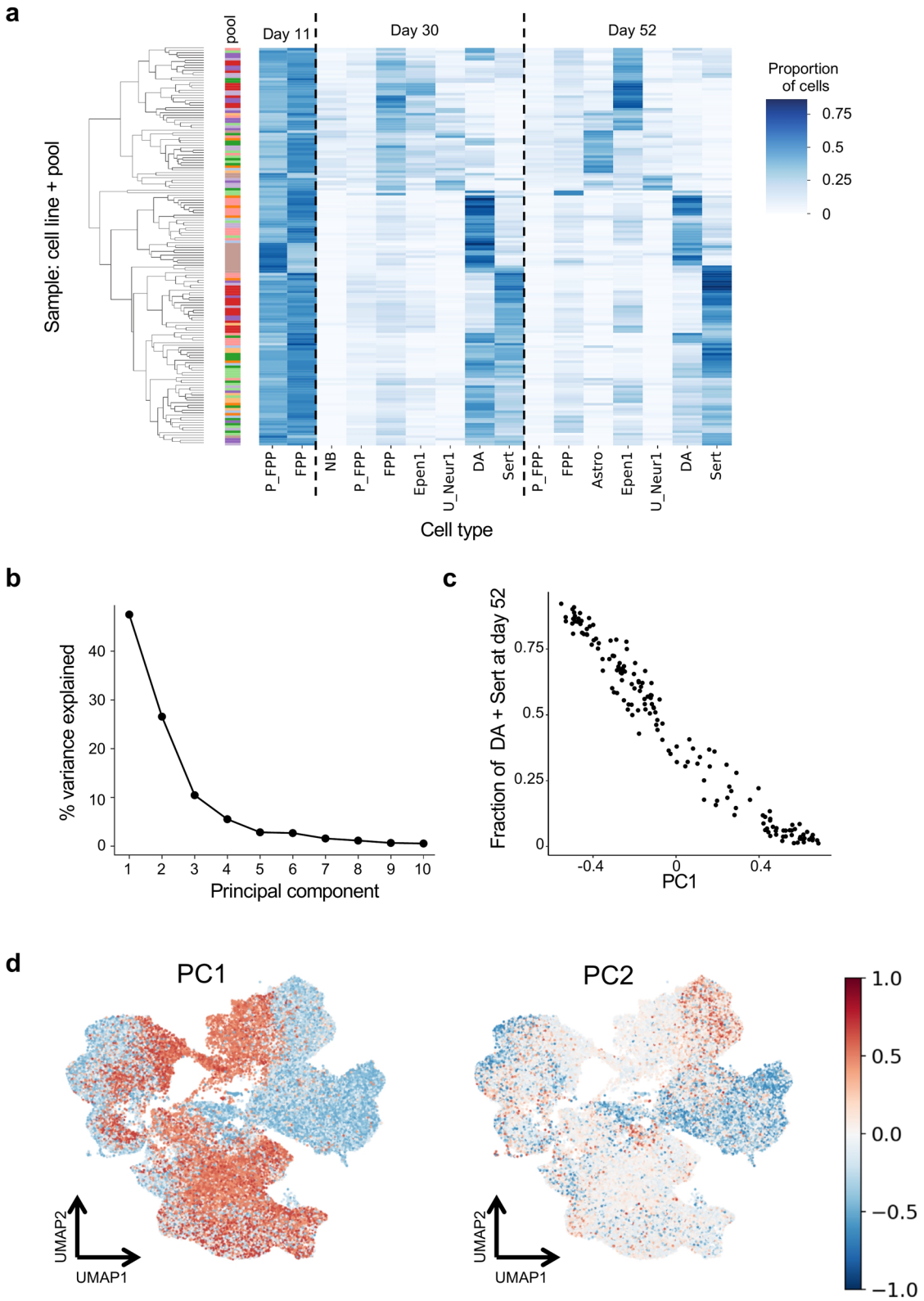
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00801-6.

Correspondence and requests for materials should be addressed to J.C.M., F.T.M., D.J.G. or O.S.

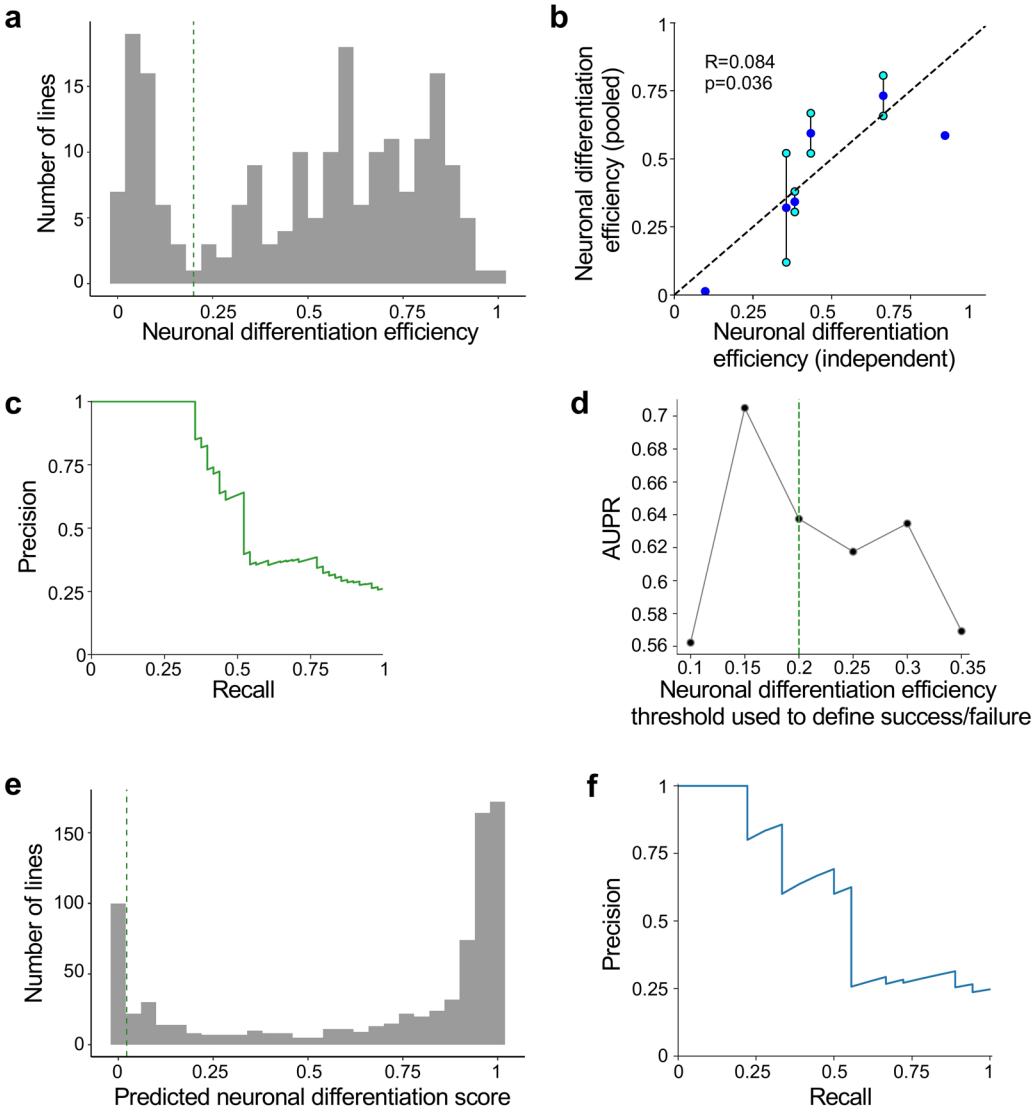
Peer review information *Nature Genetics* thanks Kristen Brennand and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



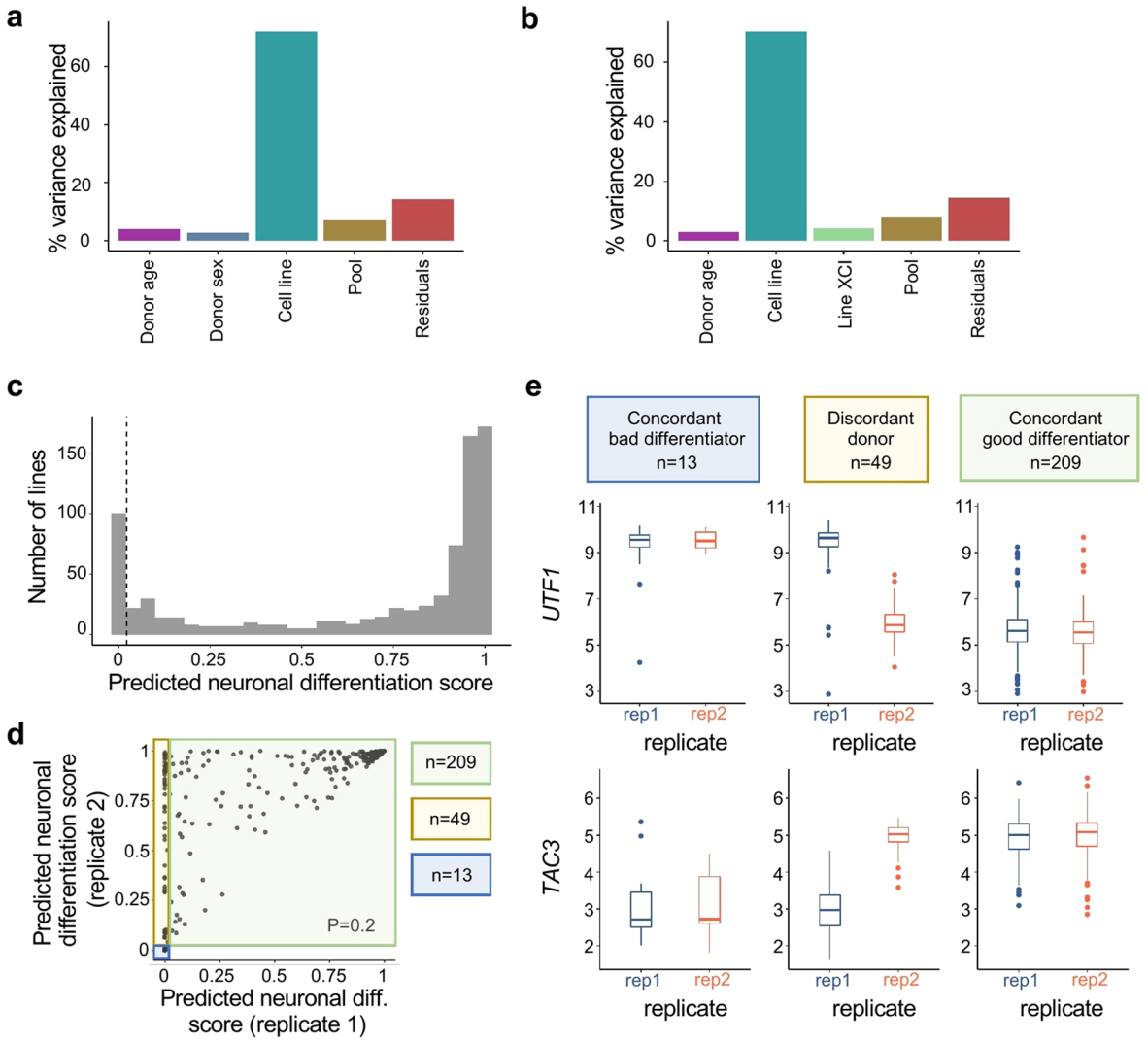
Extended Data Fig. 1 | scRNA-seq clustering and cell type annotation. Cells at each time point pooled across lines were clustered using Louvain clustering, after normalization and batch correction using Harmony (**Methods**). Subsequently, clusters were annotated to cell types using known marker genes; when two clusters showed the same gene set enrichment they were assigned to the same cell type identity (**Methods**). **a,** UMAPs of cells sampled at each time point and coloured by cluster identity. **b,** Same UMAPs as in **a**, with cells this time coloured by cell type annotation. **c,** Heatmap showing average expression profiles of canonical marker genes across the identified cell types (as in **b**, excluded dopaminergic neuronal markers; expression scaled between 0 and 1 for each gene). **d,e,** Heatmaps similar to **c**, showing average expression profile of marker genes across the identified cell types for **d**) dopaminergic neurons using literature-curated markers **e**) cortical hem/Cajal retzius cells using markers (**Methods**). Legend: Astro: Astrocyte-like, DA: Dopaminergic neurons, Epen1/2: Ependymal-like 1/2, FPP: Floor Plate Progenitors, NB: Neuroblasts, P_FPP: Proliferating Floor Plate Progenitors, P_Sert: Proliferating serotonergic-like neurons, Sert: Serotonergic-like neurons, U_Neur: Unknown Neurons.



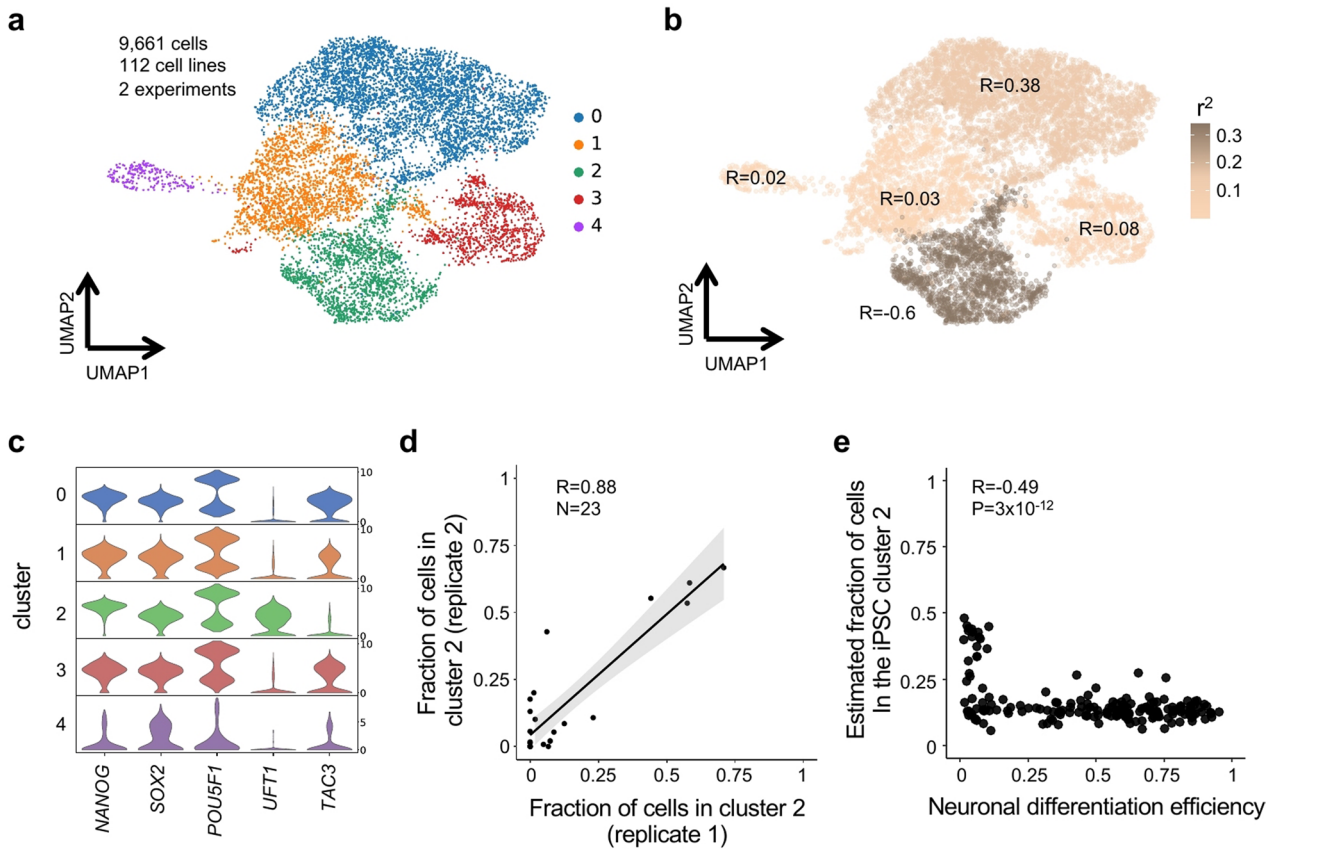
Extended Data Fig. 2 | Cell type proportions across time points and definition of neuronal differentiation efficiency at day 52. a, Heatmap of the cell proportion matrix. The colour code in the first bar indicates the assignment of lines to pools. Rows (that is cell line, pool combinations) were hierarchically clustered according to their Euclidean distance (as in Fig. 2b). Cell proportions were estimated for each cell type and time point for all combinations of cell lines, considering 10 pools with at least 10 cells at all time points (138 lines). **b,** Proportion of variance explained by each principal component calculated from the cell proportions matrix from **a**. **c,** Comparison of the first principal component (PC1) with the sum of fractions of dopaminergic and serotonergic-like neurons present on day 52. **d,** UMAP of the scRNA-seq profiles used for the cell proportions matrix from **a**, with cells coloured by the loading of PC1 (left) and PC2 (right) of the cell proportion matrix. Legend: Astro: Astrocyte-like, DA: Dopaminergic neurons, Epen1: Ependymal-like 1, FPP: Floor Plate Progenitors, NB: Neuroblasts, P_FPP: Proliferating Floor Plate Progenitors, Sert: Serotonergic-like neurons, U_Neur1: Unknown Neurons 1.



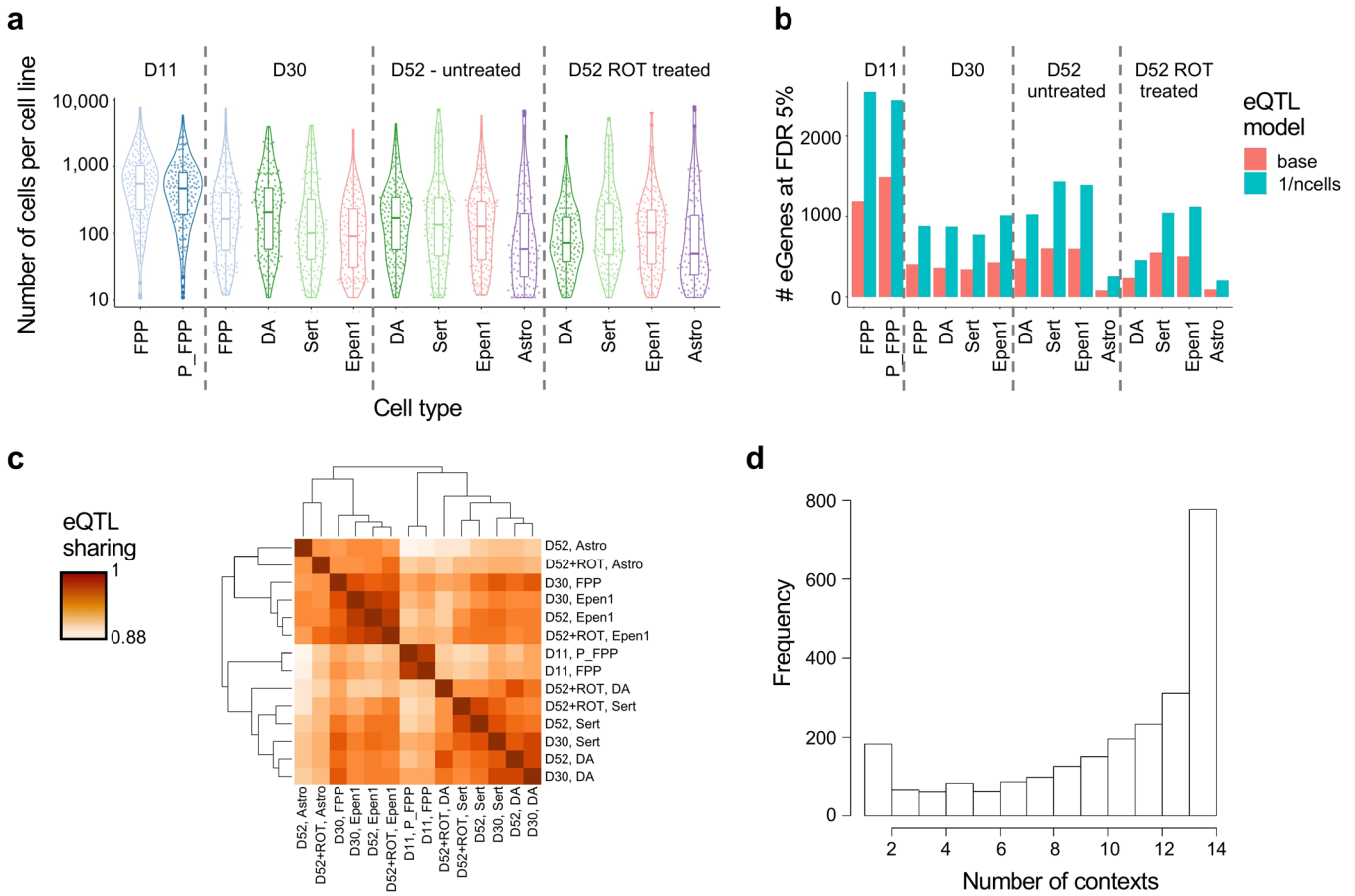
Extended Data Fig. 3 | Prediction of differentiation failure from iPSC gene expression. a, Histogram of neuronal differentiation efficiencies across cell lines. The dashed line denotes the threshold to define differentiation success or failure (that is efficiency=0.2). **b,** Effect of pooling on neuronal differentiation efficiency. Shown is a scatterplot of neuronal differentiation efficiency, estimated from independent single-line differentiations (x-axis) vs differentiation efficiency defined from pooled data from the corresponding lines (y-axis). Bars connecting cyan and blue points indicate differentiation efficiencies for replicates of the same cell line in different pools (cyan points), and the average of those replicates (blue points). For cell lines differentiated in multiple pools, average differentiation efficiencies are shown by blue points. The Pearson R and p-value were computed from the average values (blue points) only. **c,** Precision-recall curve for a logistic regression model trained to predict differentiation failure from iPSC gene expression data (Methods). Shown is precision versus recall, as assessed using leave-one-out cross validation. **d,** Area under the precision-recall curve (AUPR) for models as presented in **c**, when considering alternative threshold values to define differentiation failure. **e,** Histogram of the predicted differentiation based on iPSC gene expression for 812 HipSci cell lines. The threshold used to define potent differentiators corresponds to 35% recall, 100% precision, when using 0.2 as threshold (as in **a, b**). **f,** Cross validation of differentiation outcome prediction. The dataset was split in half (pools 1–8, 9–17) to define independent training and test fractions. All processing steps (merging, clustering, batch correction, etc.) were performed separately for these two fractions, following identical steps and parameter settings to those used for the main analysis. We trained a predictive model using data from pools 1–8, and assessed its performance on pools 9–17. Only cell lines not contained in pools 1–8 were considered for performance assessment.



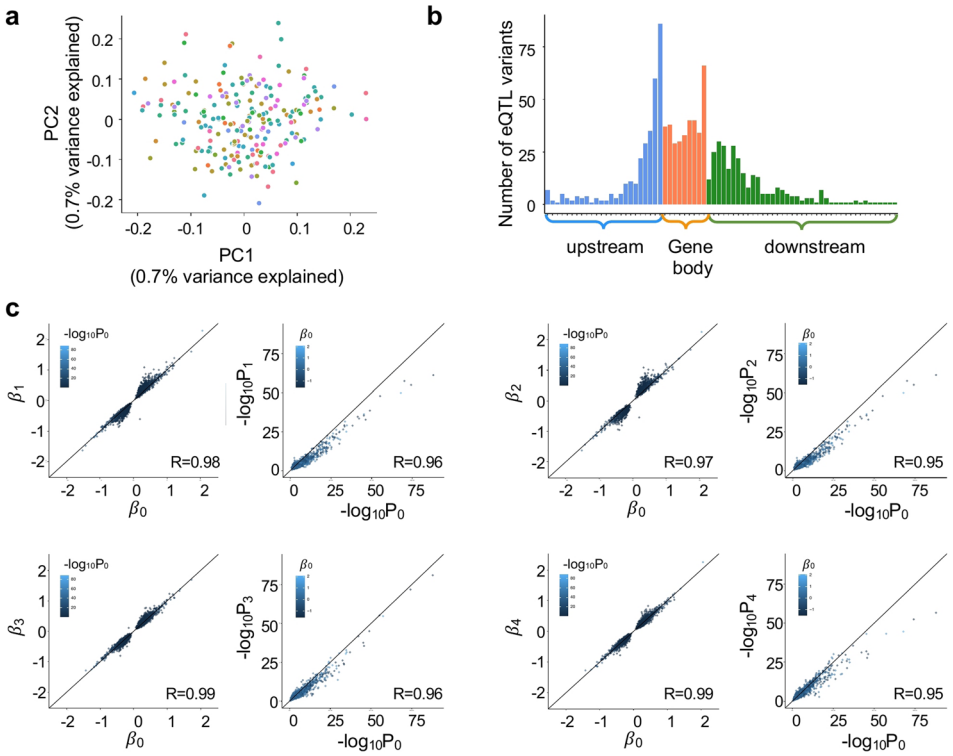
Extended Data Fig. 4 | Predicted neuronal differentiation capacity across replicate iPSC lines derived from the same donor. a,b, Variance component analysis of neuronal differentiation efficiency. **a,** Variance component breakdown of a model that explains neuronal differentiation efficiency as a function of cell line, pool, sex, age and noise (n = 230; fitted using lme4). **b,** In order to assess the effect of XCI status, we fit an analogous variance component model as in **a**, however considering only female lines (n = 115), and explaining neuronal differentiation efficiency as a function of cell line, pool, XCI status, age, and noise. **c,** Histogram of predicted neuronal differentiation efficiency based on iPSC gene expression for 812 HipSci cell lines. The vertical line indicates the prediction threshold that corresponds to 35% recall, 100% precision (c.f. Extended Data Fig. 3). **d,** Scatter plot of predicted neuronal differentiation efficiency for two replicate lines from the same donor. Shown are data from n = 271 donors contained in HipSci with RNA-seq data from two independent reprogramming events. Replicate 1 is chosen as the line with the lower predicted score. Colours indicate three categories of donors, according to the concordance of predicted neuronal differentiation capacity: both lines predicted to fail (blue, n = 13), both lines predicted to be potent differentiators (green, n = 209), discordant predictions, with one potent and one failing differentiator (yellow, n = 49). To assess whether this was significantly different from what we would expect for any two lines taken by chance, we performed a chi square test comparing the expected frequencies for any two given lines (based on the overall results) and the observed frequencies for pairs of lines from the same donor, obtaining a non-significant result (p = 0.1991). **e,** Bulk RNA-seq expression of *UTF1* and *TAC3* for the two replicate lines for the same donor, stratified by the categorisation as in **d**. In the box plots, the middle line is the median and the lower and upper edges of the box denote the first and third quartiles.



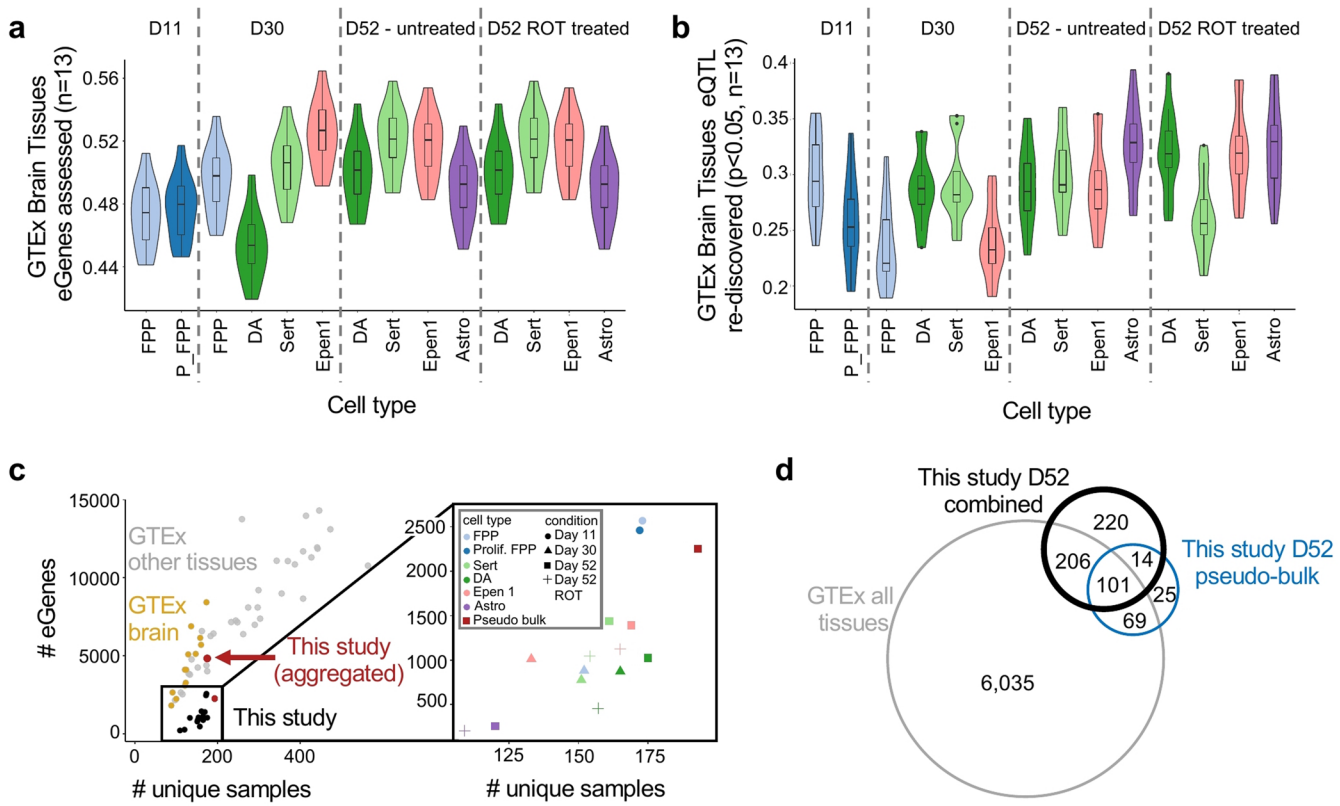
Extended Data Fig. 5 | Analysis of iPSC scRNA-seq data reveals a subpopulation characterised by expression of predictive marker genes associated with lower differentiation efficiency. a, UMAP overview of the dataset. iPSC scRNA-seq data from (Cuomo *et al*. 2020) were analysed following the analogous batch adjustment and clustering steps as applied to the neuronal differentiation data, identifying 5 clusters. **b,** UMAP as in **a**, coloured by the squared correlation coefficient R2 between correlation of bulk expression and differentiation efficiency (R values are indicated, as in Fig. 3b) and log fold change between one cluster and all others. **c,** Violin plots of gene expression for selected pluripotency genes (*NANOG, SOX2, POU5F1*) as well as marker genes that are upregulated and downregulated respectively in cluster 2 (*UTF1, TAC3*, from Fig. 3). **d,** Scatter plot of the proportion of cluster 2 cells between replicate experiments (based on n = 23 lines differentiated in two separate pools in Cuomo *et al*. paper 2020). LOESS curve and 95% confidence interval are included. **e,** Scatter plot between neuronal differentiation efficiency (x-axis) and the proportion of cells assigned to cluster 2 (y-axis) analogous to Fig. 3f, however using computational estimates of the proportions of cluster 2 cells based for a larger set of HipSci lines (using Decon-cell, based on bulk RNA-seq, n = 182; **Methods**).



Extended Data Fig. 6 | eQTL mapping strategies and sharing between eQTL maps. a, Distribution of the number of cells per cell line with scRNA-seq available for eQTL mapping for each context (cell type-condition). Dots correspond to individual cell lines (number of cell line per context ranging between 104 and 173). **b,** Number of genes with at least one eQTL (that is eGenes) for each context (cell type-condition) detected using either a traditional linear model (coral) or using a linear mixed model that accounts for heterostochastic noise due to variation in the number of cells assayed for each line (seagreen; **Methods**). **c,** Sharing of eQTL signal between 14 eQTL maps across all contexts (cell type-conditions), as estimated using MASHR (**Methods**). **d,** Distribution of the number of contexts (cell type-conditions) in which a given eQTL is identified (from 1 to 14, lfsr < 0.05, quantified using MASHR). Legend: Astro: Astrocytes-like; DA: Dopaminergic neurons, Epen1: Ependymal-like1, FPP: Floor Plate Progenitors, P_FPP: Proliferating Floor Plate Progenitors, Sert: Serotonergic-like neurons.



Extended Data Fig. 7 | eQTL mapping robustness across methodologies. a, Scatter plot of the first two principal components of the kinship matrix, revealing no evidence for pronounced population structure or relatedness between lines. **b,** Genomic location of eQTL lead variants relative to normalized gene coordinates, considering 1,024 eQTL identified in DA day 52 untreated cells (using Model 0, see below). **c,** Scatter plot of effect size estimates (left) and negative log p-values (right) for eQTL lead variants (FDR < 5%), comparing the model considered in this study (Model 0, x-axis) versus 4 alternative eQTL models (Model 1–4, y-axis). Inlined is Pearson' R. Shown are results obtained on day 52 untreated DA cells, comparing the following models: Model 0: y = PC1:15 + SNP + 1/n + noise (1,024 eGenes), Model 1: y = pool + sex + SNP + 1/n + noise (1 pool per line selected; (608 eGenes, 574 of which also in Model 0), Model 2: y = pool + sex + SNP + K + noise (320 eGenes, 312 of which shared with Model 0), Model 3: y = PCs + K + noise (471 eGenes, 457 of which shared with Model 0), Model 4: y = pool + SNP + K + 1/n + noise (856 eGenes, 734 of which shared with Model 0).



Extended Data Fig. 8 | eQTL and colocalisation in relation to GTEx. a, Fraction of GTEx brain eGenes that could be assessed in each of the considered contexts (cell type-conditions; Methods). **b,** Fraction of GTEx brain eQTL that were replicated in this study (nominal $p < 0.5$; fraction relative to the set of assessed genes from **a**). **c,** Figure analogous to main text Fig. 4c, additionally including eQTL counts from a pseudobulk eQTL analysis (top red dot on the left, red square on the right; calculated using cells from all day 52 cells untreated pooled). **d,** Figure analogous to main text Fig. 5a, additionally including colocalisation results from a pseudobulk eQTL analysis (using cells from all day 52 cells untreated pooled). In the box plots, the middle line is the median and the lower and upper edges of the box denote the first and third quartiles, while the violin plots show the distribution. Legend: Astro: Astrocytes-like; DA: Dopaminergic neurons, Epen1: Ependymal-like1, FPP: Floor Plate Progenitors, P_FPP: Proliferating Floor Plate Progenitors, Sert: Serotonergic-like neurons.

nature research

Corresponding author(s): Dan Gaffney, Oliver Stegle, Florian Merkle, John Marioni

Last updated by author(s): Jan 21, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	Published software: a combination of NGS analysis tools, CellRanger (version 2.1.0), STAR (version 020201), GATK (version 3.7.0), demuxlet (version 1.0), R (Harmony, MNN) and Python (Scanpy version 1.4, see details in Methods). Unpublished data analysis scripts are available at https://github.com/single-cell-genetics/singlecell_neuroseq_paper/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Managed access data from single-cell RNA sequencing are accessible in the European Genome-phenome Archive (EGA, https://www.dev.ebi.ac.uk/ega/) under the study number EGAS00001002885 (dataset: EGAD00001006157).
Open access single-cell RNA sequencing data are available in the European Nucleotide Archive (ENA) under the study ERP121676 (https://www.ebi.ac.uk/ena/browser/view/PRJEB38269).
Processed single cell count data and eQTL summary statistics are available from Zenodo: https://zenodo.org/record/4333872.

Data availability statement including URLs for both the new data generated here and other publicly available datasets used are reported in the main manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was based on results from many previous human eQTL studies (e.g. PMID: 29022597), where above 100-200 samples provide power to map eQTL, with power increasing with sample size.
Data exclusions	Single-cell RNA-seq data were excluded based on quality control criteria, as described in the Methods
Replication	We consider the main dataset (sc-RNA-seq from differentiating neurons) to be a single dataset, albeit including data from many repeated differentiation assays. In particular, the differentiation assay was repeated across 19 experiments and the measure of differentiation efficiency was robust across replicates of the same cell line in different experimental batches. While some additional differentiation assays failed in the course of generating the dataset, these were not attempts to replicate any findings we report. For 12 lines, the measure of differentiation efficiency was also replicated in another independent scRNA-seq cerebral organoid dataset. The identification of a subpopulation of iPSC cells was replicated in an independent dataset (see Extended Figure 3).
Randomization	iPSC lines were allocated to groups/experimental batches at random.
Blinding	Investigators were blinded during collection, as the origin of each assayed cell was only determined during data analysis (by assessment of RNA-sequencing data).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	The following antibodies were used: FOXA2 (Santa Cruz, sc101060 - 1/100),LMX1A (Millipore, AB10533 - 1/500), TH (Santa Cruz, sc-25269 - 1/200), MAP2 (Abcam, 5392 - 1/2000), Donkey anti-chicken AF647 (Thermo Fisher Scientific, A21449), Donkey antimouse AF488 (Thermo Fisher Scientific, A11008), Donkey anti-mouse AF555 (Thermo Fisher Scientific, A31570), Donkey antirabbit AF488 (Thermo Fisher Scientific, A21206), Donkey anti-rabbit AF555 (Thermo Fisher Scientific, A27039)
Validation	All commercially available antibodies have been validated by the manufacturer with supporting publications found on manufacturer websites. Antibodies were further validated in-house using relevant positive and negative controls.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	All cell lines are iPSCs from the HipSci project (www.hipsci.org)
Authentication	All cell lines were matched to the original donor by genotype, using RNA-sequencing data.
Mycoplasma contamination	Cells were not tested for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	NA

Human research participants

Policy information about studies involving human research participants

Population characteristics	25 to 79 years old phenotypically 'healthy' donors (male and female) with no diagnosed genetic disease.
Recruitment	All samples for the HipSci resource were collected from consented research volunteers recruited from the NIHR Cambridge BioResource (http://www.cambridgebioresource.org.uk).
Ethics oversight	Samples were collected initially under ethics for iPSC derivation (REC Ref: 09/H0304/77, V2 04/01/2013), with later samples collected under a revised consent (REC Ref: 09/H0304/77, V3 15/03/2013).

Note that full information on the approval of the study protocol must also be provided in the manuscript.