nature methods

Article

https://doi.org/10.1038/s41592-024-02555-5

Multiplexing cortical brain organoids for the longitudinal dissection of developmental traits at single-cell resolution

Received: 18 August 2023

Accepted: 31 October 2024

Published online: 09 December 2024

Check for updates

Nicolò Caporale[1,2,6,7], Davide Castaldi[1,2,6], Marco Tullio Rigoli[1,2,6], Cristina Cheroni[2], Alessia Valenti[1,2], Sarah Stucchi[1,2], Manuel Lessi[1,2], Davide Bulgheresi[2], Sebastiano Trattaro[2], Martina Pezzali[1,2], Alessandro Vitriolo[2], Alejandro Lopez-Tobon[2], Matteo Bonfanti[2], Dario Ricca[2], Katharina T. Schmid[3,4], Matthias Heinig[3,4], Fabian J. Theis[3,4], Carlo Emanuele Villa[2,7] & Giuseppe Testa[1,2,5,7]

Dissecting human neurobiology at high resolution and with mechanistic precision requires a major leap in scalability, given the need for experimental designs that include multiple individuals and, prospectively, population cohorts. To lay the foundation for this, we have developed and benchmarked complementary strategies to multiplex brain organoids by pooling cells from different pluripotent stem cell (PSC) lines either during organoid generation (mosaic models) or before single-cell RNA sequencing (scRNA-seq) library preparation (downstream multiplexing). We have also developed a new computational method, SCanSNP, and a consensus call to deconvolve cell identities, overcoming current criticalities in doublets and low-quality cell identification. We validated both multiplexing methods for charting neurodevelopmental trajectories at high resolution, thus linking specific individuals' trajectories to genetic variation. Finally, we modeled their scalability across different multiplexing combinations and showed that mosaic organoids represent an enabling method for high-throughput settings. Together, this multiplexing suite of experimental and computational methods provides a highly scalable resource for brain disease and neurodiversity modeling.

The polygenic underpinnings of human neurodiversity, in its physiological and pathological unfolding alike, have been eloquently referred to as terra incognita, calling for new maps to trace that unfolding in the authenticity of human genetic backgrounds and thereby render it mechanistically actionable. Developmental stochasticity and environmental triggers add to such complexity, and the increasingly broader range of exposome that is becoming measurable promises to make gene–environment interactions finally tractable at meaningful scales[1–4].

Toward these overarching goals, brain organoid and single-cell multiomic technologies have afforded major strides in the mechanistic dissection of human neurodevelopment, enabling transformative insights from the study of genetic and environmental causes of neuropsychiatric disorders, a community-wide effort to which we and several others have been contributing[2,5–14]. Importantly, our recent benchmark of cortical brain organoids (CBOs) compared to the human fetal cortex confirmed the preservation in CBOs of transcriptional programs pinpointed as relevant for disease modeling[15].

A full list of affiliations appears at the end of the paper. e-mail: giuseppe.testa@fht.org

Despite these advances, the characterization of brain organoids at single-cell resolution from entire cohorts and, in perspective, at population scale remains however an unmet challenge, although an obviously required one if we are to capture how individual genomes and developmental trajectories shape variability in vulnerability and resilience across the spectrum of neurodiversity[16–18]. Scaling up human brain organoid modeling and molecular profiling by single-cell omics would allow us to understand how the molecular causes of neurodevelopmental disorders trigger deviations from physiological trajectories[19], in line with the expanding set of population-level single-cell studies[20,21]. This is however still an experimental and analytical challenge due to high cost and workload and to the inherent batch-to-batch variability of the complex experimental designs required.

To overcome some of these problems, progress has been made in single-cell multiplexing strategies, including methods based on sample barcoding[22–26] and methods leveraging the detection of natural genetic variants[27–30]. These approaches have indeed proven to be instrumental for population genetics and disease-modeling studies, including through co-culture of cell lines derived from multiple donors in a single dish[31–39].

However, multiplexing has not yet been systematically applied to organoids, a challenge that is particularly relevant for the brain given the long-term longitudinal unfolding of highly heterogeneous combinations of cell types, and the field still lacks studies able to establish the experimental and computational viability of applying multiplexing strategies to complex three-dimensional experimental systems.

We thus implemented and benchmarked complementary strategies to multiplex human brain organoidogenesis in vitro, pooling PSCs coming from different individuals either during organoid generation, termed the mosaic model according to standardized guidelines for brain organoid nomenclature we recently contributed to[40], or before scRNA-seq library preparation. To improve genetic-based cell identification when dealing with brain organoid single-cell transcriptomes, we developed an in silico deconvolution method (SCanSNP), which we benchmarked against existing deconvolution tools, producing a consensus pipeline for robust genotype identification across datasets of different quality. Finally, we evaluated the two multiplexing paradigms through a deep reconstruction of neurodevelopmental trajectories, provided proof of principle of their suitability for linking genetic variation to neurodevelopmental trajectory phenotypes and modeled the scalability of the system, analyzing different multiplexing combinations with an increasing number of lines. This provides the community with an enabling resource for scaling up brain organoid modeling to the challenges of human neurodiversity.

Results

Single-cell analysis of multiplexed CBOs

To test the feasibility of multiplexed brain organoid modeling over extended developmental time courses, we developed an experimental design comparing two approaches with distinctive features in terms of scaling, standardization potential and experimental challenges: (1) pooling PSC lines from multiple donors to generate mosaic CBOs (mCBOs) and (2) generating CBOs individually and pooling them only before single-cell droplet encapsulation (hereafter referred to as downstream multiplexing) (Fig. 1a and the Methods). For the first strategy, we pooled PSC lines (5,000 cells per line) before organoid generation and longitudinally profiled the resulting mCBOs through scRNA-seq at 50, 100 and 300 d of differentiation, following the same protocol previously used and benchmarked in our laboratory[5,11,15]. For downstream multiplexing, CBOs were generated individually from the PSC lines and profiled at the same time points, pooling equal amounts of cells per line after organoid dissociation for single-cell library preparation. For both approaches, individual cell identity was demultiplexed through genetic variation (Fig. 1b).

Experimental assessment of mCBOs

We first characterized mCBOs by immunofluorescence for canonical markers of neurodevelopment previously defined for organoids differentiated with the same protocol by us and others[15,40,41] and confirmed their expected expression patterns (Fig. 2a,b and Extended Data Fig. 1a–c). Next, to estimate the stability of the presence of the different PSC lines within mCBOs throughout development, we used a two-pronged approach. First, we longitudinally assessed the proliferation rate of organoids derived from individual lines in two differentiation replicates through flow cytometry-based cell cycle analysis. This showed similar proliferative trends across lines, with the expected decrease in proliferation along differentiation and no substantial variations in the proliferative rate of the different lines over time points (Fig. 2c,d and Supplementary Data 1). Next, we probed the extent to which individual lines, despite homogeneous proliferation rates in isolation, could still yield skewed distributions when grown as mosaics. In line with in vivo evidence of high asymmetric clonality during human brain development[42], the quantification of genotypes from single-cell transcriptomics confirmed a variable degree of balance between individual lines during mCBO differentiation (Extended Data Fig. 2).

Benchmarking of demultiplexing algorithms

The single-cell transcriptomics datasets generated from both mosaic and downstream multiplexed organoids at early, mid and late stages of differentiation (Fig. 1b) were demultiplexed using three different state-of-the-art algorithms available for genetic demultiplexing (demuxlet[27], souporcell[28], Vireo[29]). We observed a high number of predicted doublets with an unexpected read count distribution skewed toward low values when compared to singlets (Extended Data Fig. 3a,b). This was associated with a variable degree of unassigned cells and identity call agreement among tools (Extended Data Fig. 3c,d). We thus set out to develop a new method, SCanSNP (Fig. 3a), to overcome the observed limitations by (1) dividing the classification challenge into two steps, one for identity assignment and one for doublet detection and (2) measuring the genetic purity of each droplet to identify low-quality droplets and separate them from authentic doublets. Finally, to consolidate deconvolution accuracy, we set a consensus call framework considering the strengths and weaknesses of each algorithm, including non-genetic-based tools for doublets and low-quality cell detection[43,44], merging their outcomes into one combined output (Fig. 3b).

Next, we performed demultiplexing again using all the above methods. We found that all three thus far available tools tended to overestimate doublets, if compared to the theoretically expected rate, SCanSNP or the consensus call (Fig. 3c). Notably, doublets identified by SCanSNP and the consensus call were the only ones for which the average log count distribution of unique molecular identifiers was, as expected, higher than that for singlets (Fig. 3e). This suggests a bias of existing algorithms in doublet detection that includes low-quality droplets that SCanSNP is instead able to identify (Extended Data Fig. 3c). Accordingly, the evaluation of agreement rate in the assignment of individual identity across algorithms highlighted a high overall agreement for singlet calls, with cases of lower agreement coinciding with datasets that had higher genotype imbalance (Extended Data Fig. 3d). On the other hand, we observed that doublet detection agreement was consistently lower (Extended Data Fig. 3d). To additionally assess the demultiplexing performance of the different algorithms, we benchmarked them with ground truth priors using (1) five in silico multiplexed datasets with varying degree of balance among genotypes and (2) two ad hoc experiments in which the different genotypes were barcode tagged before pooling. As shown in Fig. 3d, in the simulated datasets, all demultiplexing algorithms performed better in balanced cases, and SCanSNP was the most accurate in both balanced and imbalanced settings. The superior performance (comparison against cellranger multi; Methods) delivered by SCanSNP was confirmed in the barcode-tagged datasets (Extended Data Fig. 4).

https://doi.org/10.1038/s41592-024-02555-5

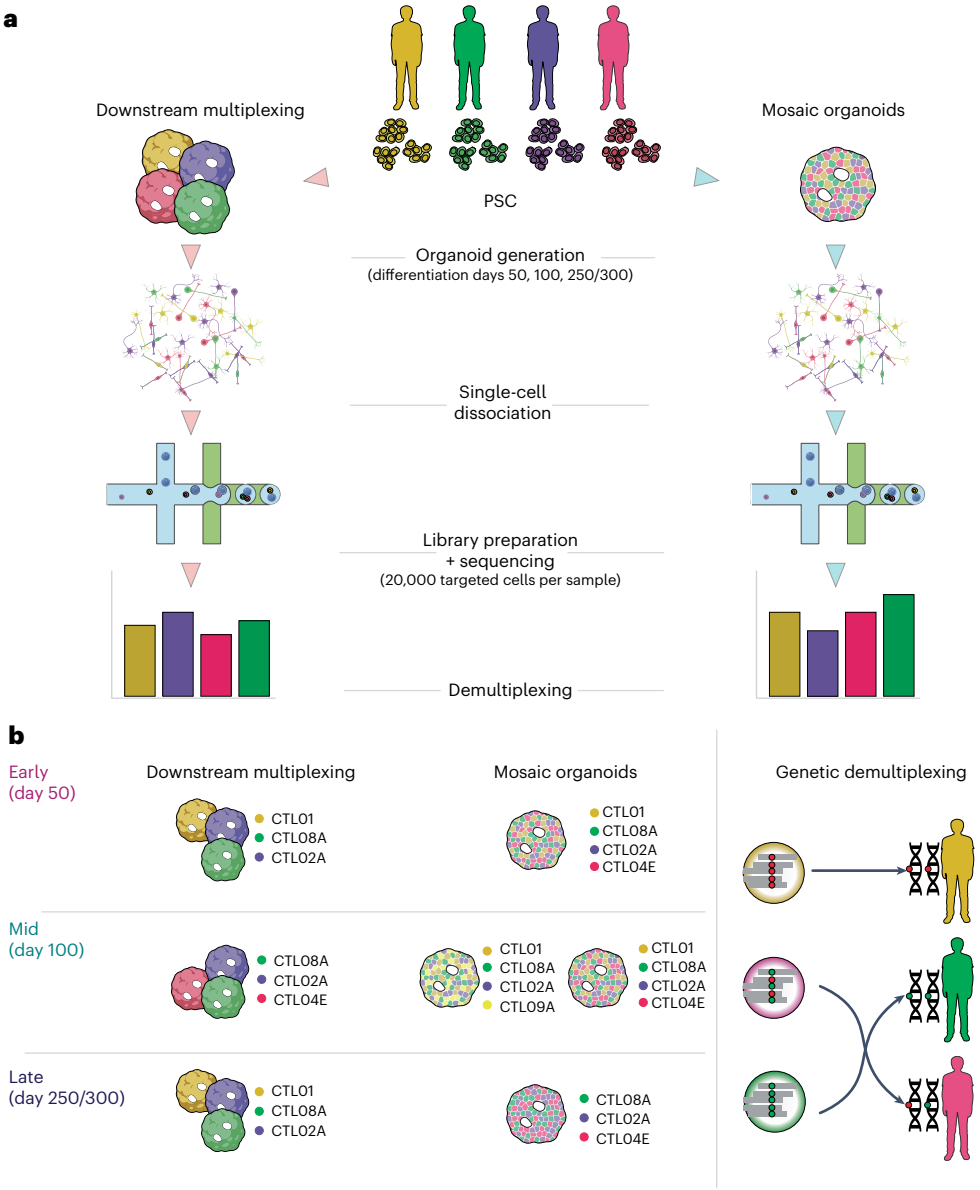


Fig. 1 | Schematic representation of multiplexing paradigms and experimental design. a, Representation of the two explored multiplexing paradigms. Left, downstream multiplexed organoids grown from individual lines and pooled in equal amounts after dissociation at the single-cell level. Right, mosaic organoids generated by pooling equal amounts of multiple PSC lines during organoid seeding. After single-cell dissociation, both paradigms undergo library preparation, sequencing and demultiplexing. PSCs, human pluripotent stem cells. **b**, Representation of experimental design and the demultiplexing approach. The CBO differentiation time points, the number of replicates for each time point and their division between the two multiplexing paradigms and the PSC lines (genotypes) used for each experiment are shown. Figure created with BioRender.com.

Analysis of neurodevelopmental cell types

After deconvolving the identity of each droplet in our dataset by the consensus call, we discarded doublets and low-quality cells and proceeded with cells coming from the PSC lines across multiplexing modalities (Extended Data Fig. 2 and Supplementary Table 1). We thus analyzed the single-cell transcriptomic dataset entailing four PSC lines, their time points and two multiplexing modalities (Fig. 1b). We pre-processed each sample individually, integrated them and carried out multi-tier filtering. A force-directed graph was used to visualize the integrated dataset (Fig. 4a).

We systematically annotated the identity of each cell population by analyzing the expression distribution of relevant markers, defining the gene signatures driving each cluster and projecting the cells onto a reference single-cell human fetal brain dataset (Fig. 4a, Extended Data Fig. 5a–c and Supplementary Data 2)[45]. Consistent with our previous benchmarking[15] and in line with evidence from recent studies characterizing brain organoids at single-cell resolution[46–48], CBOs recapitulated most of the relevant neurodevelopmental cell populations of human corticogenesis. We observed a well-defined group of cells expressing the proliferative markers *CDC20*, *MKI67* and *TOP2A* that were labeled as proliferating progenitor cells. Those cells are in continuity with a bigger cluster of cells expressing *PAX6*, *VIM*, *SOX2* and *NES* and annotated as radial glia. This cluster undergoes a bifurcation, extending either toward cells expressing outer radial glial and/or astrocyte markers *HOPX*, *AQP4* and *S100B* or the neuronal markers *STMN2*, *DCX* and *GAP43*, which we annotated accordingly. We then further divided the neuronal branch into intermediate progenitors expressing *EOMES*, two clusters of excitatory neurons expressing *NEUROD2*, *TBR1*, *SATB2*, *SLC17A6*, *SLC17A7*, *SLA* and *GRIA2*, two interneurons clusters expressing *DLX2*, *DLX5*, *DLX6-AS1*, *GAD1* and *GAD2*, one cluster of migrating

https://doi.org/10.1038/s41592-024-02555-5

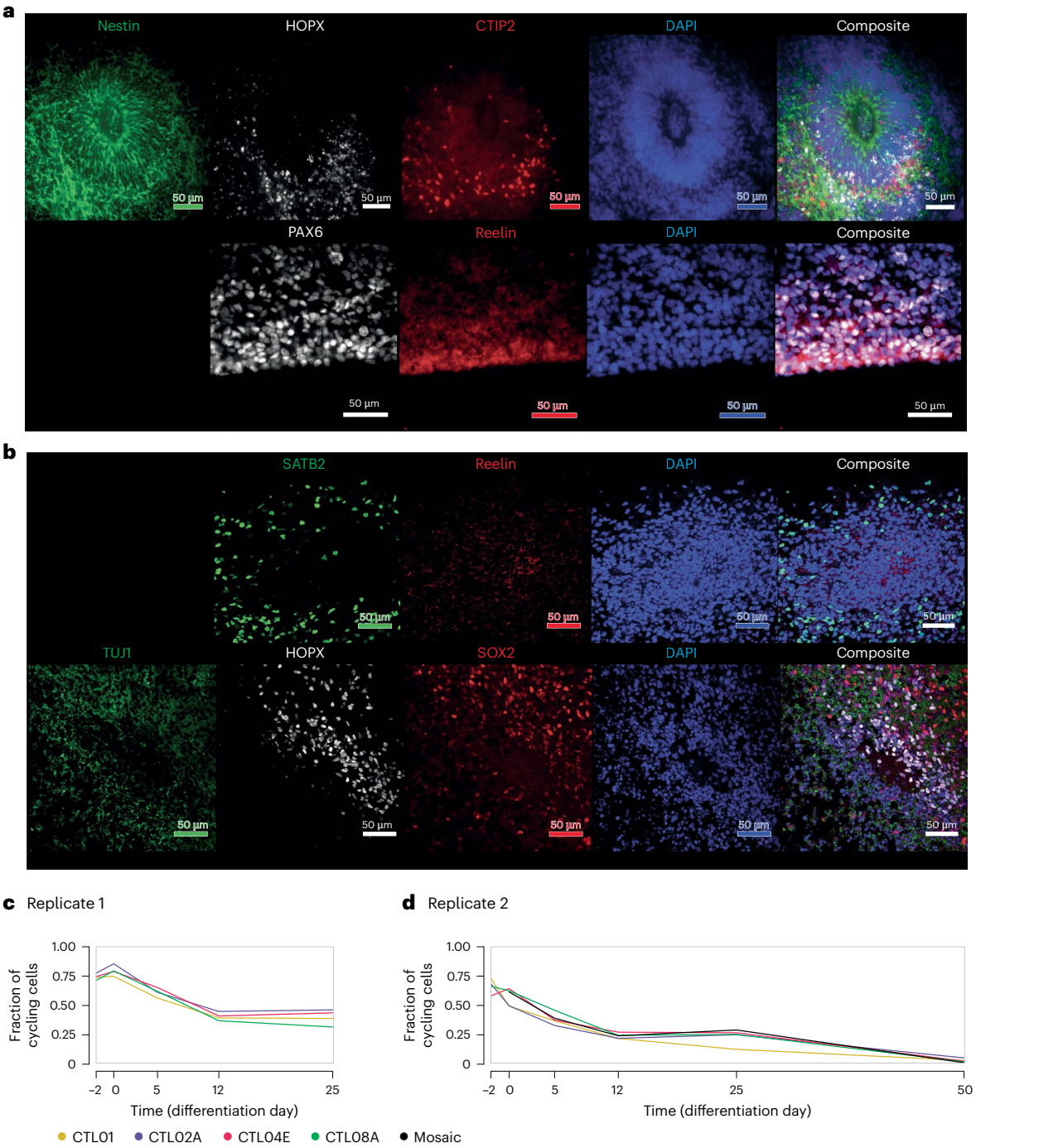


Fig. 2 | Experimental assessment of mCBOs. a,**b**, Immunofluorescence-based benchmarking of neurodevelopmental markers in mCBOs. At differentiation day 50 (**a**), mCBOs show consistent presence of ventricular-like structures positive for the neural stem cell marker nestin and are surrounded by a layer of cells expressing the outer radial glial marker HOPX. The presence of newborn deep-layer, Coup-TFI interacting protein 2 (CTIP2)$^+$ (BCL11B) cortical neurons can be appreciated next to HOPX$^+$ cells. On the outer surface, mCBOs reveal the presence of reelin$^+$ cells. At differentiation day 100 (**b**), mCBOs display more mature ventricular structures characterized by the presence of HOPX$^+$ outer radial glial cells, SATB homeobox 2 (SATB2)$^+$ intermediate layer neurons and SRY-box transcription factor 2 (SOX2)$^+$ neural precursors. DAPI, 4,6-diamidino-2-phenylindole; PAX6, paired box 6. **c**,**d**, Fraction of cycling cells in pure lines and mosaic organoids detected by flow cytometry (Methods) performed on two differentiation replicates at relevant time points, from day −2 (organoid seeding) up to days 25 (**c**) and 50 (**d**).

neurons expressing *FOXP1*, *CA8* and *LHX1* and one cluster expressing *RELN*, *PAX6* and *CBLN1*, annotated as Cajal–Retzius-like cells (Fig. 4a,b and Extended Data Fig. 5b).

We then compared the different time points of our longitudinal dataset, observing, as expected, differences in cell type abundance. While radial glial progenitors, proliferating progenitors and maturing neurons were evenly distributed across time points, we observed that migrating neurons, early excitatory neurons and Cajal–Retzius-like cells were more abundant at the early and middle time points, whereas outer radial glia and/or astrocytes, late excitatory neurons and interneurons were more abundant at the late time point (Extended Data Fig. 5d). Instead, cell type proportions between mosaic organoids

and downstream multiplexed samples were comparable (Extended Data Fig. 5d). We thus tested for statistical significance in differential abundance, leveraging Milo[45], confirming significant differences across time points (Fig. 4c,d). Conversely, the same analysis done comparing mosaic and downstream multiplexed samples showed no significant difference between the two multiplexing paradigms.

To further investigate the molecular impact of mosaic culture, we compared the transcriptomes of the same genotypes between mosaic and downstream multiplexing and, as a reference for technical variability, we compared the same genotypes when grown in two different mosaic batches (Methods).

The analysis highlighted a low number of differentially expressed genes (DEGs) for both comparisons (69 DEGs between multiplexing modalities and four DEGs between two mosaic batches, false discovery rate (FDR) < 0.05, log (fold change (FC)) > 1.5) (Supplementary Fig. 1a and Supplementary Data 3).

Finally, to further probe the transcriptional regulation of specific cell types in the CBOs vis-à-vis the in vivo counterparts, we took advantage of our multigenotype dataset to detect loci with allelic-specific expression (ASE)[46], as the likelihood of observing heterozygous loci is higher. As expected, the number of detected loci with ASE correlated with the total amount of reads (Extended Data Fig. 5e). Moreover, exploring the correlation of the proportion of reads in each allele across cell types, we observed that, also in CBOs, the patterns of ASE are cell type specific, in line with external evidence from the human brain[47] (Extended Data Fig. 5f).

Multiplexed CBOs capture key neurodevelopmental trajectories

One of the most powerful innovations enabled by single-cell analysis of experimental models featuring extensive cellular heterogeneity, such as CBOs, is the possibility to analyze developmental trajectories through pseudotime analysis[49,50]. To deeply explore this aspect in our dataset, we leveraged partition-based graph abstraction (PAGA; Extended Data Fig. 6a)[51] and isolated all biologically relevant lineages recapitulated in CBOs (progenitors → excitatory neurons, progenitors → interneurons, progenitors → astrocytes, progenitors → Cajal–Retzius-like cells, progenitors → migrating neurons). We could thus analyze through diffusion pseudotime (dpt)[52] how the single cells from the different time points and multiplexing modalities were distributed along those neurodevelopmental trajectories (Fig. 5a,b). Moreover, we harnessed the power of trajectory-based differential expression based on generalized additive models (tradeSeq[53]) and identified the key genes driving each lineage of differentiation (Extended Data Fig. 6b and Supplementary Data 4). Of note, for the excitatory lineage, trajectory analysis allowed us to identify two temporally divergent paths of neurogenesis that reconciled toward the end into the excitatory neuron clusters, where *EOMES* was one of the key drivers of the difference (Fig. 5a). This is in line with the recent empirical demonstration of the ability of CBOs to recapitulate in a physiologically relevant way the two different patterns (direct and indirect) of human cortical neurogenesis[54–57].

Linking genetic variation to neurodevelopmental phenotypes

Genome-wide association studies have uncovered hundreds of thousands of genetic variants associated with human traits[48]. Mechanistically linking such variants to phenotypes of interest by molecular quantitative trait locus mapping has been instrumental to understand functional effects of genetic variation[58]. In particular, single-cell expression quantitative trait locus (eQTL) modeling allows identification of the impact of single-nucleotide polymorphisms (SNPs) on cell type-specific molecular mechanisms, as we pioneered in ref. 59, pushing forward previous work by the GTEx Consortium to characterize variation in gene expression across individuals and diverse tissues of the human body[60]. Moreover, because single-cell analysis can quantify longitudinal trajectories, it is possible to assess the effects of SNPs on traits that vary dynamically along a continuous axis[20,49]. The power of these approaches however has yet to be exploited for developmental variation as a trait per se that is now amenable to longitudinal quantitation in the form of single-cell resolved trajectories. We thus set out to leverage our in vitro cohort and probe as proof of principle how genetic variants could be linked to the neurodevelopmental traits quantified in CBOs. First, we empirically tested in our dataset the statistical power to perform single-cell eQTL analysis through our previously developed tool, scPower[50], and identified the number of lines needed to perform single-cell de novo eQTL discovery, given the CBO-specific variability we had determined in terms of gene expression across neurodevelopmental cell types (Extended Data Fig. 6c). Because, as expected with our dataset, we do not have the statistical power for de novo eQTL discovery, we followed a supervised approach to investigate the genetic basis of the differences we observed in neurodevelopmental trajectories between individuals during CBO differentiation.

In particular, we focused on the trajectory of migrating neurons because it displayed the greater and most reproducible genotype diversity in the distribution of cells along pseudotime (computed by dpt; Fig. 5c and Extended Data Fig. 6d). Indeed, along this trajectory, upon downsampling cells per genotype to ensure balance within each time point (50 random sampling iterations), we found differences between two subgroups (of two individuals each) (Fig. 5c). This was also confirmed by PCA performed on the downsampled dataset, where we observed the same genotypes' separation along PC1 (Extended Data Fig. 7), which represents an alternative proxy of the differentiation trajectory of migrating neurons (Fig. 5b). We thus explored in more depth the genetic variants that distinguish the two subgroups. Of the total of 24,987 coding variants passing the quality filters, 1,035 presented a different allelic configuration between the two groups, and we found that 11 of these genetic variants (Fig. 5d and Supplementary Data 5) have been previously identified as single-cell eQTLs in population-scale scRNA-seq profiling of induced PSC (iPSC)-derived neurons (Table 7 from Jerber et al.[35], total reported eVariants from the study, 10,972). Two of these variants are *cis*-acting eQTLs for *SRCIN1*, one of the trajectory-specific highly variable genes (HVGs) we had identified (Fig. 5e) that encodes a protein previously shown to be a regulator of disease-relevant

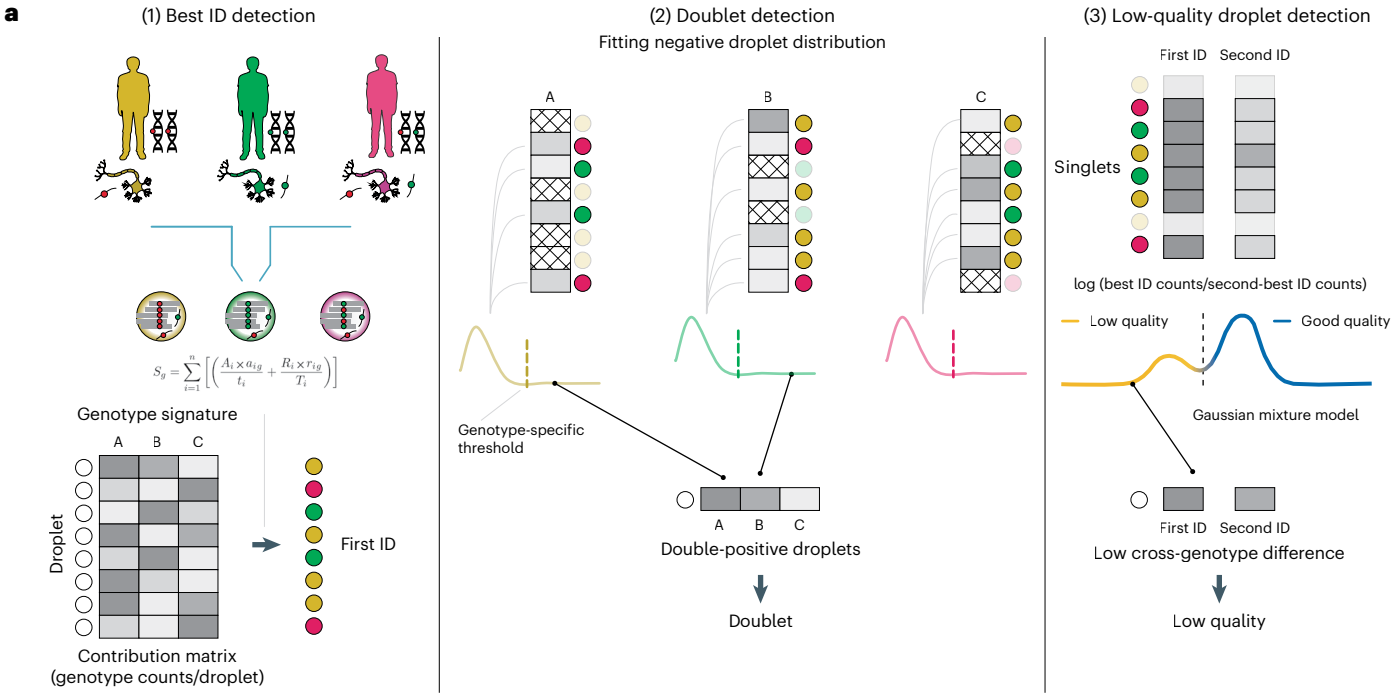
Fig. 3 | SCanSNP and consensus call overview and benchmark. a, Main steps of SCanSNP: (1) best ID assignment, (2) doublet classification and (3) low-quality droplet detection. Colors represent different individuals and droplet identity after best ID assignment; grayscale intensity in matrices represents the number of reads; cross-patterned cells represent droplets not included in the computations; colored curves represent fitted distributions. **b**, Schematic representation of the consensus call. Top, an example of consensus score with weights for each algorithm. Weights are the partial score attributed from different algorithms toward a specific genotype. Bottom, final aggregation rationale. DBL, doublets; LQ, low quality; SoC, souporcell. **c**, Percentage of doublets detected from each algorithm across datasets. *x* axis, datasets ordered by the number of retrieved cells; *y* axis, doublet rate. Lines are colored by algorithm. D, day. **d**, Benchmark of demultiplexing performance against ground truth in an in silico multiplexed dataset. *x* axis reports different algorithms; *y* axis reports the deconvolution accuracy as the rate of barcodes with predicted IDs that match ground truth versus total detected barcodes per dataset. Balanced datasets are colored light green; imbalanced datasets are colored orange. Plot is bound between 0.86 and 0.98 to magnify the differences among and within tools and jittered on the *y* axis for readability. **e**, Distribution of predicted singlets and doublets from the different algorithms for each of the seven datasets. Algorithms are shown on the *x* axis, and mean log counts are shown on the *y* axis. The shape of each symbol corresponds to a dataset. Markers are colored by singlet or doublet predicted labels.

https://doi.org/10.1038/s41592-024-02555-5

alterations in cell migration and hence potentially representing the genetic underpinning of the neurodevelopmental difference in the migrating neuron trajectory observed here between the CBOs from different individuals.

Scalability of multiplexed CBOs

Having shown that multiplexed CBOs enable the parallel assessment of neurodevelopmental trajectories from different lines, we next sought to determine the scalability of the system, given the variable balance



b

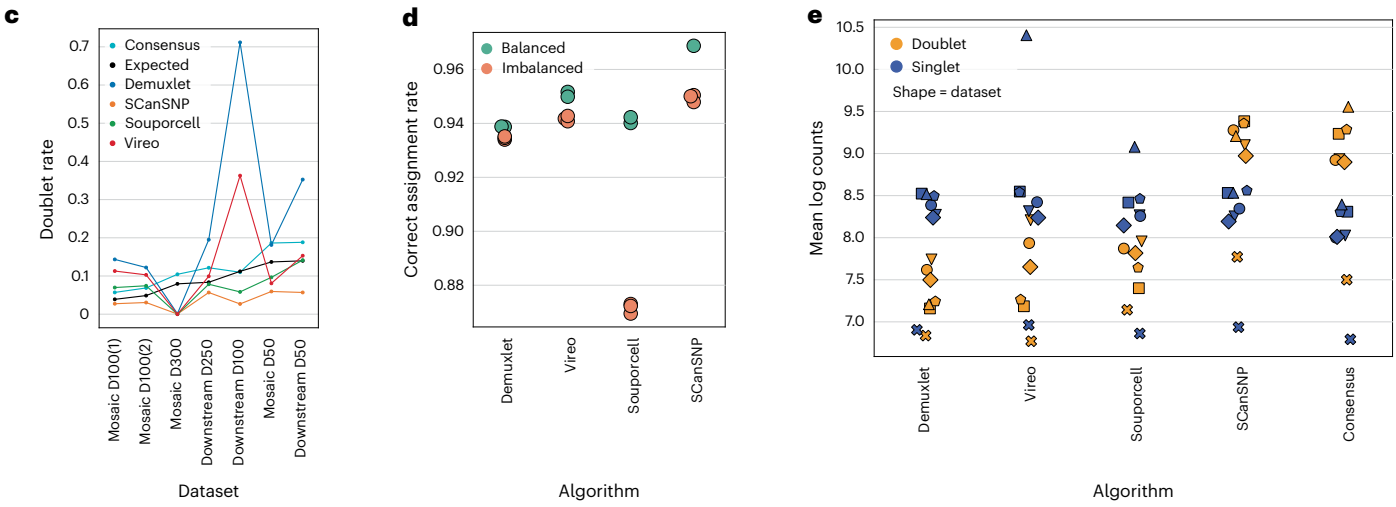
Weights:
Demuxlet = 1.5
SCanSNP = 1.5
Vireo = 1
Souporcell = 1

	Demuxlet	SCanSNP	Vireo	SoC	Consensus score
Droplet 1*j*	1.5	1.5	0	1	4
Droplet n*j*	0	0	1	1	2

Droplet score for *j*th genotype (pass if consensus score ≥ 3)

1	2	3	4
Low quality if Failed consensus and SCanSNP == LQ or Failed consensus and dropkick == LQ	Genetic doublet if SCanSNP == DBL and at least 1 more algorithm == DBL or All algorithms but SCanSNP == DBL	Transcriptomic doublet if ScDblFinder == DBL and The droplet was not already marked as LQ/genetic DBL.	Leftover low quality: Failed consensus and not belonging to previous categories

Consensus call



https://doi.org/10.1038/s41592-024-02555-5

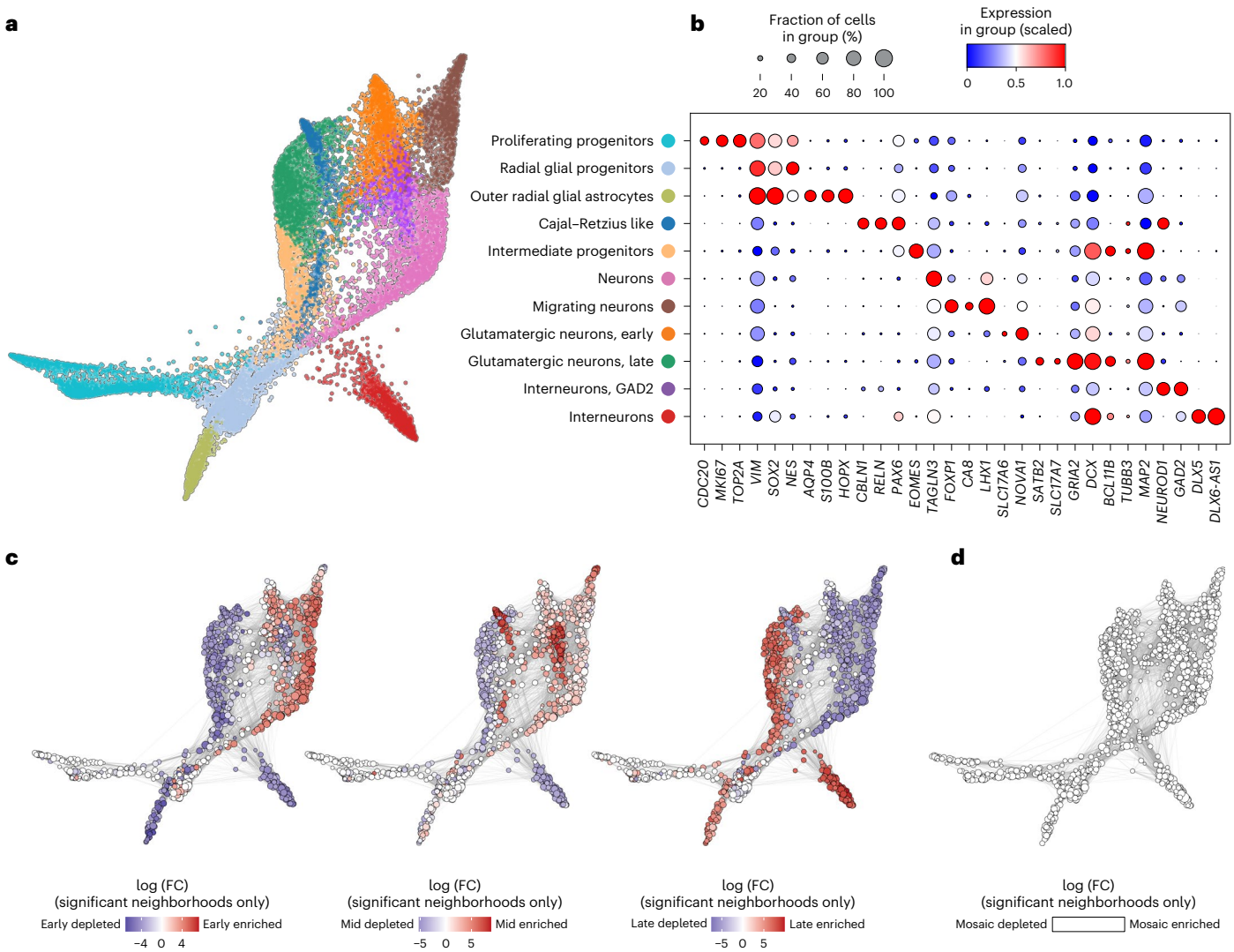


Fig. 4 | Sample multiplexing allows identification of stage-specific neurodevelopmental cell populations. a, Cell embeddings in a force-directed graph after preprocessing and filtering. Each dot is a cell, colored by annotated cell type. **b**, Dot plot of gene expression for some of the relevant markers used in the annotation. Size of dots is proportional to the number of cells expressing a marker, and color encodes the mean expression in the group (scaled log-normal counts). **c**, Differential abundance graph of CBO data across time points. The basis for the visualization is the force-directed graph. Dots represent groups of similar cells; color code indicates enrichment for the indicated categories (only shown for significantly enriched neighborhoods, spatial FDR < 0.1). **d**, Differential abundance graph of CBO data across multiplexing modalities. The basis for the visualization is the force-directed graph. Dots represent groups of similar cells; color code indicates enrichment for the indicated categories (only shown for significantly enriched neighborhoods, spatial FDR < 0.1).

among lines during mosaic organoid differentiation (Extended Data Fig. 2) and recent evidence showing a high degree of clonal variability in brain organoid development[51,52] in line with the physiological clonality of human brain development[42]. We thus increased the number of PSC lines and multiplexing combinations as well as replicates across different organoids of the same combinations to quantitatively measure the variance in balance between different PSC lines in mosaic organoids and derive empirically grounded guidelines for scaling up disease modeling with this approach.

To that end, we undertook an in-depth profiling of the growth rate of 12 different PSC lines by systematic and longitudinal imaging over different passages during pluripotency maintenance and at multiple time points during CBO differentiation. By performing semi-automated image preprocessing and segmentation[53], we modeled the growth dynamics of each PSC line (both in pluripotency and upon CBO differentiation) and of seven different mosaic combinations (Extended Data Figs. 8 and 9).

In parallel, each mosaic combination was profiled by Census-seq[36,37] at five different time points in triplicate to measure the PSC line balance over the mCBO differentiation course. This showed that, starting from a balanced mix of lines when generating mCBOs, the line balance was already altered at early stages of differentiation, resulting, for some of the lines, in very low numbers of cells at later time points (Fig. 6a and Supplementary Data 6). We thus asked whether these imbalance patterns could be predicted by individual lines' growth rates. As shown in Fig. 6a and Supplementary Data 6, for the contribution of each PSC line in mosaics, we found both that it was reproducible across replicates of the same mix and that it reflected line-specific behaviors across different mixes (Fig. 6a), yet such regularity was not explained either by PSC or CBO growth rates (Supplementary Data 6 and Extended Data Fig. 10a). This excludes that a rapid assessment of PSC line growth rates, either in pluripotency or upon differentiation, could be harnessed to preselect optimal mixes (that is, those most likely to preserve balance) for a high-scalability setting.

We thus probed an alternative solution to the scalability problem, integrating the longitudinal Census-seq with mCBO imaging data to model mCBO clonal dynamics and thereby determine the upper

https://doi.org/10.1038/s41592-024-02555-5

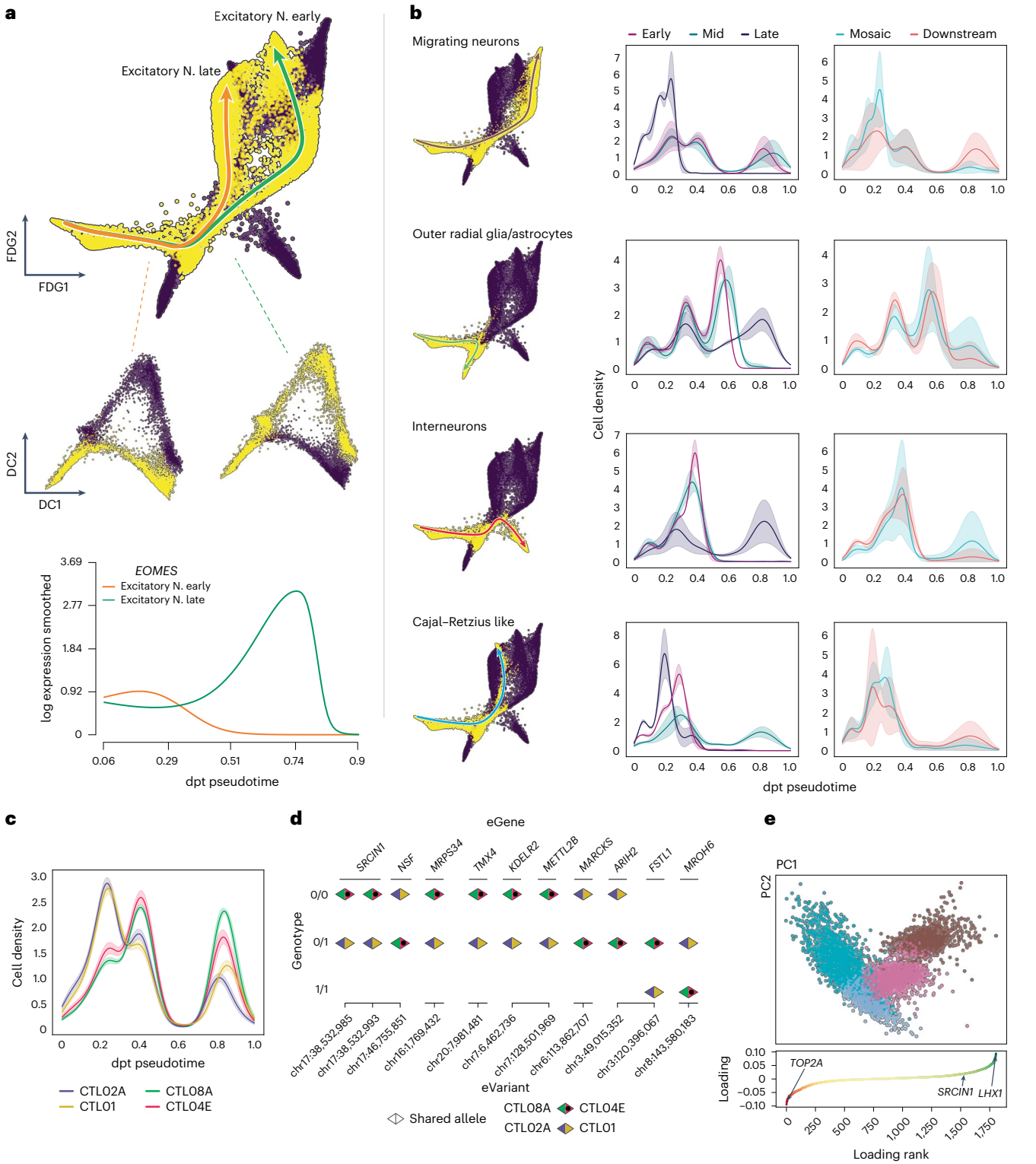


Fig. 5 | Multiplexed CBOs recapitulate key neurodevelopmental trajectories. **a**, Highlight of late and early excitatory neurons (N.) branches. Top, embeddings in a force-directed graph (FDG1 and FDG2 dimensions). Middle, a magnification of isolated branches on the diffusion map (DC1 and DC2 components). Bottom, smoothed expression of *EOMES* for each of the two trajectories along pseudotime. **b**, Highlight of main developmental lineages in a force-directed graph. Next to each force-directed graph, density plots display the number of cells along pseudotime by (1) differentiation time points and (2) multiplexing paradigm. Faded color shows 1 s.d. among dataset replicates; solid line displays the mean value among dataset replicates. **c**, Density plot of cells along pseudotime for the isolated migrating neurons lineage. Cells are divided and colored by genotype. Faded color shows 1 s.d. across random subsampling iterations; solid lines display the mean value among dataset replicates. **d**, Schematics of SNPs classified as eQTLs from Jerber et al.[35] and their allelic configuration (reference (Ref)/Ref, Ref/alternate (Alt), Alt/Alt) for the pairs of genotypes exhibiting difference in pseudotime. chr, chromosome. **e**, Principal-component analysis (PCA) (Harmony integrated) of the isolated migrating neurons branch. Bottom, migrating neurons-specific HVGs. Genes are ordered by their principal component (PC)1 loading, with *TOP2A* (cycling progenitors marker), *LHX1* (migrating neurons marker) and *SRCIN1* (only eGene from **d** also present in HVGs) highlighted.

feasibility limit of the mosaic models as a function of their sensitivity in exposing single-cell endophenotypes. Specifically, we derived the 'mosaic growth rate' of each PSC line by combining mosaic organoid growth rate as measured by imaging and the percentage of specific PSC lines as measured by Census-seq (Fig. 6b and Extended Data Fig. 10b). We then computed the empirical probability distribution of the derived number of cells coming from each PSC line in mCBOs and used it as the basis for a Monte Carlo simulation with two initial parameters: (1) the starting number of PSC lines used to generate mCBOs and (2) the minimum number of cells per line required for single-cell analysis of the main cell types. We found that the number of effectively recovered lines (that is, lines detected at levels that enable single-cell analysis of the main neurodevelopmental cell types) increased with the number of lines that were mixed (despite the reduction in relative representation) (Fig. 6c and Extended Data Fig. 10c).

Given similar clonal dynamics across mCBOs generated by different numbers of PSC lines as well as by different combinations of PSC lines, the scalability of the system is currently limited only by the number of lines that can be accurately counted and multiplexed when generating mCBOs. We could thus compute the impact of the different multiplexing approaches (Fig. 6d and Extended Data Fig. 10d) on the experimental timelines needed to perform large-scale disease-modeling studies. mCBOs emerged from this analysis as an enabling method due to the radical acceleration afforded by parallel processing of different rounds of differentiation experiments, for example, compressing the profiling of 1,000 lines down from 10 to 3 years and hence marking the difference, in the current funding and project management ecosystem of most academic institutions, between a largely unrealistic setting and a routinely feasible design.

Discussion

In this work, we developed multiplexing strategies to tackle the challenges of brain organoid modeling for the systematic study of human neurodiversity at scale. Pooling strategies already proved transforming for iPSC-based disease modeling[34–38]; however, their application to the inherently complex setting of brain organoids, which generate a highly heterogeneous composition of cell types over very extended times, has just started to be undertaken[51,52]. Here, we first benchmarked and improved the demultiplexing steps by leveraging and aggregating the edges of different publicly available state-of-the-art algorithms and developing SCanSNP, a new deconvolution method that more accurately identifies doublets and low-quality cells and flexibly accommodates cases of imbalanced genotypes.

Next, we deeply characterized the neurodevelopmental cell types and trajectories recapitulated by CBOs across time points and multiplexing modalities. Of note, the unicity of our longitudinal design allowed us to capture a very interesting and relevant neurodevelopmental process represented by the transient emergence of Cajal–Retzius-like cells, a population that was also recapitulated in a different brain organoid model but so far not reported in CBOs[54]. In our dataset, Cajal–Retzius-like neurons are enriched at middle stages and then depleted at advanced stages in both downstream CBOs and mCBOs, consistent with in vivo evidence that they are fated to elimination as neuronal networks mature[55]. Future studies will thus enable benchmarking the ability of CBOs to recapitulate also the specific molecular and cellular features of human Cajal–Retzius neurons in vivo.

This design also allowed us to identify the developmental divergence of direct and indirect neurogenesis, highlighting *EOMES* as the key driver guiding the different paths of early-versus-late glutamatergic neuron differentiation in organoids. This confirmed the ability of CBOs to recapitulate the physiological processes of a unique aspect of human neocortical development[56,57], in line with recent work that investigated these biological processes in early fetal tissue and cerebral organoids through imaging[61].

Finally, we observed in our late-stage CBOs also the differentiation of GABAergic neurons[62], a finding that, given the dorsal forebrain patterning, is in line with recent evidence that human dorsal cortical progenitors are also capable of producing GABAergic neurons with the transcriptional characteristics of cortical interneurons[63,64].

Through the analysis of developmental trajectories, we also showed that brain organoid multiplexing can be used to identify how candidate genetic variants are linked to specific neurodevelopmental phenotypes, captured here as proof of principle in the transcriptional regulation of migrating neurons. This paves the way to molecular quantitative trait locus discovery once multiplexing approaches will be applied to CBOs differentiated from large-scale cohorts of PSC lines. Indeed, we showed that the number of lines needed for de novo discovery of eQTLs linked to neurodevelopmental trajectories in brain organoids (which we propose to term eDTL, for expressed developmental trait loci) should be in the range of hundreds of individuals.

Both molecular and imaging analyses showed that mCBOs recapitulate the expected cytoarchitectural organization, including ventricular-like structures and the gradual emergence of relevant neurodevelopmental markers of standard CBOs. As expected, they showed a variable degree of balance among individual PSC lines.

Differently from another multiplexing approach that was recently proposed, which requires however dissociation and re-aggregation of the organoids to keep the balance among individual lines[52], here we aimed to keep the mosaic model unperturbed and thus allow it to recapitulate the physiological phenomenon of clonal asymmetry[51,65,66], proceeding to an in-depth investigation of these dynamics across different numbers of PSC lines, multiplexing combinations and replicates.

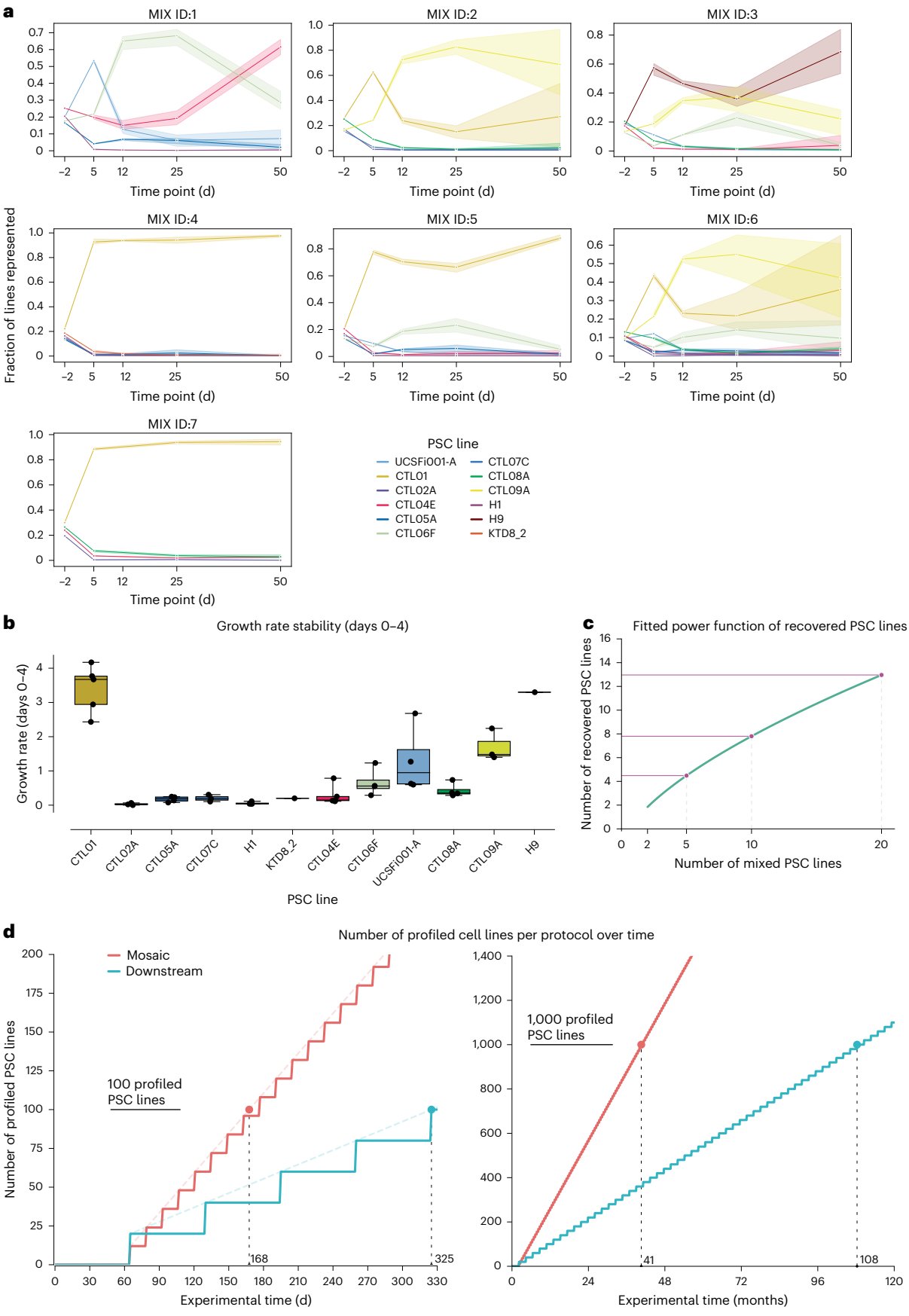
Interestingly, our longitudinal imaging profiling, characterizing with quantitative parameters each PSC line upon passaging in pluripotency and through CBO differentiation, did not explain the contribution of each line in the mosaic models. However, we cannot exclude that a growth rate characterization with even additional time points and/or replicates or at even higher resolution could lead to identification of a relationship between cell cycle and cell fate, as has been described in a different in vitro model[67].

In terms of clonal dynamics and scalability, while we cannot generalize our results for every PSC line, our results showed that:

1. As expected from recent studies[51,52], the line balance in mCBOs is altered already at early stages of differentiation. This result, coupled with the lack of correlation with PSC and CBO growth rates, suggests that interindividual differences across PSC lines in response to the patterning factors used in CBO differentiation

Fig. 6 | Scalability of mCBOs. a, Longitudinal representation of PSC lines in seven mosaic organoids of different composition (MIX IDs, Source Data Fig. 6), as quantified by Census-seq. Each dot represents the average value of representation for a single line at that time point across three different replicates, except day −2, when only one replicate was available. Shading around the line represents the 95% confidence interval around the mean, as depicted by the solid line connecting the dots. **b**, Growth rate of different cell lines for different mosaic mixtures in the interval of day 0 to day 4. Different cell lines are divided along the *x* axis; each dot in a box represents the growth rate of the same cell line in a different experimental mixture. Displayed by each box is the median (horizontal solid line within the box), the interquartile range (upper and lower bound of the boxes) and minimum and maximum values (extension of the whiskers) among mixtures per line. **c**, Fitted power function of cell line recovery (Monte Carlo simulation, Extended Data Fig. 10c). The plot shows the number of mixed lines (*x* axis) and the number of recovered cell lines (*y* axis). **d**, Projection of the number of profilable lines over time across multiplexing approaches. The plot shows the number of profiled cell lines (*y* axis) and the experimental time in days (*x* axis). Vertical dashed lines represent the experimental time to reach 100 and 1,000 profiled cell lines in the left plot and the right plot, respectively. For the left plot, the approximation strict line is displayed for each protocol.

https://doi.org/10.1038/s41592-024-02555-5



protocols (as also shown in ref. 68) may be a key determinant for the clonal asymmetry. However, with our data, we cannot exclude that the imbalance already originates during the first 2 d of the protocol, during which embryoid bodies are generated, a phenomenon that was also shown to be relevant in ref. 51 where the authors showed that clones were lost during the formation of embryoid bodies using a cerebral organoid protocol.

2. Mosaic models showed reproducibility in the relative abundance among different individual lines across replicates of the same multiplexing combinations and across different combinations. This points to the genetic background as a key determinant of clonal dynamics. Naturally, this does not per se rule out that also the epigenetic state of each line at the time of mCBO generation may exert a substantial impact, in line with results from a large collection of PSC lines[69]. Indeed, the lines that were overrepresented in the experiments in Fig. 6 were not overrepresented in previous single-cell transcriptomic experiments over different years and thus from different passages of the same iPSC lines.
3. The comparison between mosaic and downstream multiplexing modalities showed that non-cell-autonomous effects are not evident in the mosaic model, as expected, considering that we only used wild-type PSC lines to generate the single-cell dataset and confirming independent works that found the same result in two-dimensional[38] and three-dimensional[52] cultures. However, considering the few genes that emerged from the differential expression analysis between genotypes in the two multiplexing modalities, it is possible that, by increasing the number of cells and replicates and profiling mosaic models through spatial omics, the molecular impact of cell-to-cell interactions between different lines could come into relief even within the range of genetically normotypical lines. Also, it will be relevant to investigate deeper non-cell-autonomous effects using PSC lines carrying disease-relevant genetic mutations, such as the ones that already showed evidence that non-cell-autonomous mechanisms are relevant for the pathogenesis[70,71].
4. The variance in mCBO clonal dynamics across PSC lines was not dependent on the specific mosaic combinations (where different lines were used) nor on the number of lines at generation, indicating high scalability of the system. Additional experiments will clarify the impact of increasing to hundreds of lines in single mCBOs on these clonal dynamics. However, considering the current limitations in laboratory settings for accurate cell counting, which is needed to generate balanced mCBOs (excluding flow cytometry setups for the logistics of generating mosaic models with many lines), we do not suggest going lower than about 1,000 cells per line during mCBO generation. Thus, our guidelines suggest using a maximum of about 20 lines in mosaic models (because 20,000 cells are needed to generate a mCBO with our current protocol).
5. Monte Carlo simulations incorporating the empirical growth rates from our experiments allowed us to compute the probability of recovering, from each PSC line, a sufficient number of cells for single-cell omics analysis of neurodevelopmental cell types, given the number of lines used at the generation. This in turn enables a precise design of disease-modeling experiments with mosaic models. With our protocol, the number of PSC lines that can be properly analyzed in terms of single-cell transcriptomic characterization of all major neurodevelopmental cell types (considering proliferating progenitors, radial glial progenitors, neurons, migrating neurons, excitatory neurons) from a mosaic experiment is ~12 lines (empirical mean ≈ 12.89; 95% confidence interval, 12.85 to 12.93) if starting from 20 lines and sequencing ~100,000 cells. This means that the ideal settings to apply mCBO designs as a transformative tool are cohort-level screenings for trait-relevant in vitro endophenotypes, drug screening and gene–environment interaction studies[72]. Indeed, if it is not crucial to recover all starting PSC lines, mCBOs allow studying the impact of genetic makeup, environmental chemicals[11] or drugs on the gene expression of specific neurodevelopmental cell types for very large cohorts of PSC lines, even without automated setups, as the only bottleneck is the maintenance of the lines and the generation of the organoids, thereby massively increasing the feasibility of large-scale experiments. For example, by leveraging large cohorts of banked PSC lines (with already thousands of lines available in standardized repositories (Table 2 from ref. 73)) and generating batches of mCBOs with 20 lines, the experimental timeline for profiling at day 50 would be, respectively, halved for 100 lines (about 6 months instead of more than 1 year needed for a downstream multiplexing approach), more than three times lower for 1,000 lines (about 3 years instead of about 10 years with downstream) and substantially lower also than the recently introduced alternative approach based on post-re-aggregation multiplexing[52]. This, in addition to the substantial reduction in the workload (see specific calculations in the Methods), represents a transformative leap for the transcriptomic annotation of polygenic risk scores along neurodevelopment and for precision neurotoxicology and pharmacology.

To conclude, our findings indicate to opt for the downstream multiplexing strategy when the biological question at hand requires strict balance among individual lines, with the mosaic model ideally suited instead, and in fact transformative, for unbiased large-scale studies, in the same way as developmental neuroscientists can leverage pooled CRISPR perturbation strategies for screening the impact of genetic variants as a complementary approach to the validation of single mutations[74–77].

Online content

Any methods, additional references, Nature Portfolio 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/s41592-024-02555-5.

References

1. Silbereis, J. C., Pochareddy, S., Zhu, Y., Li, M. & Sestan, N. The cellular and molecular landscapes of the developing human central nervous system. *Neuron* **89**, 248–268 (2016).
2. Cheroni, C., Caporale, N. & Testa, G. Autism spectrum disorder at the crossroad between genes and environment: contributions, convergences, and interactions in ASD developmental pathophysiology. *Mol. Autism* **11**, 69 (2020).
3. Tărlungeanu, D. C. & Novarino, G. Genomics in neurodevelopmental disorders: an avenue to personalized medicine. *Exp. Mol. Med.* **50**, 1–7 (2018).
4. Hyman, S. E. The daunting polygenicity of mental illness: making a new map. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, 20170031 (2018).
5. López-Tobón, A. et al. Human cortical organoids expose a differential function of GSK3 on cortical neurogenesis. *Stem Cell Reports* **13**, 847–861 (2019).
6. López-Tobón, A. et al. *GTF2I* dosage regulates neuronal differentiation and social behavior in 7q11.23 neurodevelopmental disorders. *Sci. Adv.* **9**, eadh2726 (2023).
7. Mihailovich, M. et al. 7q11.23 CNV alters protein synthesis and REST-mediated neuronal intrinsic excitability. Preprint at *bioRxiv* https://doi.org/10.1101/2022.10.10.511483 (2022).
8. Marangon, D. et al. Novel in vitro experimental approaches to study myelination and remyelination in the central nervous system. *Front. Cell. Neurosci.* **15**, 748849 (2021).

9. Drakulic, D. et al. Copy number variants (CNVs): a powerful tool for iPSC-based modelling of ASD. *Mol. Autism* **11**, 42 (2020).
10. Villa, C. E. et al. *CHD8* haploinsufficiency links autism to transient alterations in excitatory and inhibitory trajectories. *Cell Rep.* **39**, 110615 (2022).
11. Caporale, N. et al. From cohorts to molecules: adverse impacts of endocrine disrupting mixtures. *Science* **375**, eabe8244 (2022).
12. Corsini, N. S. & Knoblich, J. A. Human organoids: new strategies and methods for analyzing human development and disease. *Cell* **185**, 2756–2769 (2022).
13. Kelley, K. W. & Pașca, S. P. Human brain organogenesis: toward a cellular understanding of development and disease. *Cell* **185**, 42–61 (2022).
14. Eichmüller, O. L. & Knoblich, J. A. Human cerebral organoids — a new tool for clinical neurology research. *Nat. Rev. Neurol.* **18**, 661–680 (2022).
15. Cheroni, C. et al. Benchmarking brain organoid recapitulation of fetal corticogenesis. *Transl. Psychiatry* **12**, 520 (2022).
16. Koi, P. Genetics on the neurodiversity spectrum: genetic, phenotypic and endophenotypic continua in autism and ADHD. *Stud. Hist. Philos. Sci.* **89**, 52–62 (2021).
17. Baron-Cohen, S. Editorial Perspective: Neurodiversity — a revolutionary concept for autism and psychiatry. *J. Child Psychol. Psychiatry* **58**, 744–747 (2017).
18. Pretzsch, C. M. et al. Neurobiological correlates of change in adaptive behavior in autism. *Am. J. Psychiatry* **179**, 336–349 (2022).
19. Rajewsky, N. et al. LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* **587**, 377–386 (2020).
20. Cuomo, A. S. E., Nathan, A., Raychaudhuri, S., MacArthur, D. G. & Powell, J. E. Single-cell genomics meets human genetics. *Nat. Rev. Genet.* **24**, 535–549 (2023).
21. De Donno, C. et al. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat. Methods* **20**, 1683–1692 (2023).
22. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
23. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
24. Gehring, J., Hwee Park, J., Chen, S., Thomson, M. & Pachter, L. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. *Nat. Biotechnol.* **38**, 35–38 (2020).
25. Hwang, B. et al. SCITO-seq: single-cell combinatorial indexed cytometry sequencing. *Nat. Methods* **18**, 903–911 (2021).
26. Datlinger, P. et al. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* **18**, 635–642 (2021).
27. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
28. Heaton, H. et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).
29. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **20**, 273 (2019).
30. Weber, L. M. et al. Genetic demultiplexing of pooled single-cell RNA-sequencing samples in cancer facilitates effective experimental design. *Gigascience* **10**, giab062 (2021).
31. Yazar, S. et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
32. Perez, R. K. et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
33. Srivatsan, S. R. et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**, 45–51 (2020).
34. 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).
35. Jerber, J. et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* **53**, 304–312 (2021).
36. 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).
37. Wells, M. F. et al. Natural variation in gene expression and viral susceptibility revealed by neural progenitor cell villages. *Cell Stem Cell* **30**, 312–332 (2023).
38. Neavin, D. R. et al. A village in a dish model system for population-scale hiPSC studies. *Nat. Commun.* **14**, 3240 (2023).
39. Farbehi, N. et al. Integrating population genetics, stem cell biology and cellular genomics to study complex human diseases. *Nat. Genet.* **56**, 758–766 (2024).
40. Pașca, S. P. et al. A nomenclature consensus for nervous system organoids and assembloids. *Nature* **609**, 907–910 (2022).
41. Birey, F. et al. Assembly of functionally integrated human forebrain spheroids. *Nature* **545**, 54–59 (2017).
42. Bizzotto, S. et al. Landmarks of human embryonic development inscribed in somatic mutations. *Science* **371**, 1249–1253 (2021).
43. Germain, P.-L., Lun, A., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Res.* **10**, 979 (2021).
44. Heiser, C. N., Wang, V. M., Chen, B., Hughey, J. J. & Lau, K. S. Automated quality control and cell identification of droplet-based single-cell data using dropkick. *Genome Res.* **31**, 1742–1752 (2021).
45. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using *k*-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
46. Cleary, S. & Seoighe, C. Perspectives on allele-specific expression. *Annu. Rev. Biomed. Data Sci.* **4**, 101–122 (2021).
47. Zhao, D., Lin, M., Pedrosa, E., Lachman, H. M. & Zheng, D. Characteristics of allelic gene expression in human brain cells from single-cell RNA-seq data analysis. *BMC Genomics* **18**, 860 (2017).
48. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
49. Cuomo, A. S. E. et al. CellRegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. *Mol. Syst. Biol.* **18**, e10663 (2022).
50. Schmid, K. T. et al. scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nat. Commun.* **12**, 6625 (2021).
51. Lindenhofer, D. et al. Cerebral organoids display dynamic clonal growth and tunable tissue replenishment. *Nat. Cell Biol.* **26**, 710–718 (2024).
52. Antón-Bolaños, N. et al. Brain chimeroids reveal individual susceptibility to neurotoxic triggers. *Nature* **631**, 142–149 (2024).
53. Uozumi, T. et al. Generation of a growth curve for iPS cells in a feeder-free culture by non-invasive image analysis. *Nikon Instruments* www.microscope.healthcare.nikon.com/resources/application-notes/generation-of-a-growth-curve-for-i-ps-cells-in-a-feeder-free-culture-by-non-invasive-image-analysis (2011).
54. Uzquiano, A. et al. Proper acquisition of cell class identity in organoids allows definition of fate specification programs of the human cerebral cortex. *Cell* **185**, 3770–3788 (2022).

55. Elorriaga, V., Pierani, A. & Causeret, F. Cajal–Retzius cells: recent advances in identity and function. *Curr. Opin. Neurobiol.* **79**, 102686 (2023).
56. Pebworth, M.-P., Ross, J., Andrews, M., Bhaduri, A. & Kriegstein, A. R. Human intermediate progenitor diversity during cortical development. *Proc. Natl. Acad. Sci. USA* **118**, e2019415118 (2021).
57. Braun, E. et al. Comprehensive cell atlas of the first-trimester developing human brain. *Science* **382**, eadf1226 (2023).
58. Aguet, F. et al. Molecular quantitative trait loci. *Nat. Rev. Methods Primers* **3**, 4 (2023).
59. van der Wijst, M. et al. The single-cell eQTLGen consortium. *eLife* **9**, e52155 (2020).
60. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
61. Coquand, L. et al. A cell fate decision map reveals abundant direct neurogenesis bypassing intermediate progenitors in the human developing neocortex. *Nat. Cell Biol.* **26**, 698–709 (2024).
62. Sloan, S. A., Andersen, J., Paşca, A. M., Birey, F. & Paşca, S. P. Generation and assembly of human brain region-specific three-dimensional cultures. *Nat. Protoc.* **13**, 2062–2085 (2018).
63. Delgado, R. N. et al. Individual human cortical progenitors can produce excitatory and inhibitory neurons. *Nature* **601**, 397–403 (2022).
64. Andrews, M. G. et al. LIF signaling regulates outer radial glial to interneuron fate during human cortical development. *Cell Stem Cell* **30**, 1382–1391 (2023).
65. Breuss, M. W. et al. Somatic mosaicism reveals clonal distributions of neocortical development. *Nature* **604**, 689–696 (2022).
66. Huang, A. Y. et al. Parallel RNA and DNA analysis after deep sequencing (PRDD-seq) reveals cell type-specific lineage patterns in human brain. *Proc. Natl Acad. Sci. USA* **117**, 13886–13895 (2020).
67. Pauklin, S. & Vallier, L. The cell-cycle state of stem cells determines cell fate propensity. *Cell* **155**, 135–147 (2013).
68. Bertucci, T. et al. Improved protocol for reproducible human cortical organoids reveals early alterations in metabolism with *MAPT* mutations. Preprint at *bioRxiv* https://doi.org/10.1101/2023.07.11.548571 (2023).
69. Kim, S.-K. et al. Individual variation in the emergence of anterior-to-posterior neural fates from human pluripotent stem cells. *Stem Cell Reports* **19**, 1336–1350 (2024).
70. Baek, S. T. et al. An AKT3–FOXG1–reelin network underlies defective migration in human focal malformations of cortical development. *Nat. Med.* **21**, 1445–1454 (2015).
71. Nakagawa, N. et al. Memo1-mediated tiling of radial glial cells facilitates cerebral cortical development. *Neuron* **103**, 836–852 (2019).
72. Seah, C. et al. Modeling gene×environment interactions in PTSD using human neurons reveals diagnosis-specific glucocorticoid-induced gene expression. *Nat. Neurosci.* **25**, 1434–1445 (2022).
73. Huang, C.-Y. et al. Human iPSC banking: barriers and opportunities. *J. Biomed. Sci.* **26**, 87 (2019).
74. Fleck, J. S. et al. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* **621**, 365–372 (2022).
75. Esk, C. et al. A human tissue screen identifies a regulator of ER secretion as a brain-size determinant. *Science* **370**, 935–941 (2020).
76. Meng, X. et al. Assembloid CRISPR screens reveal impact of disease genes in human neurodevelopment. *Nature* **622**, 359–366 (2023).
77. Li, C. et al. Single-cell brain organoid screening identifies developmental defects in autism. *Nature* **621**, 373–380 (2023).

[1]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy. [2]Human Technopole, Milan, Italy. [3]Institute of Computational Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany. [4]Department of Mathematics, Technical University Munich, Munich, Germany. [5]Department of Experimental Oncology, European Institute of Oncology IRCCS, Milan, Italy. [6]These authors contributed equally: Nicolò Caporale, Davide Castaldi, Marco Tullio Rigoli. [7]These authors jointly supervised this work: Nicolò Caporale, Carlo Emanuele Villa, Giuseppe Testa. ✉e-mail: giuseppe.testa@fht.org

Methods

Culture of pluripotent stem cells

PSC lines were cultured under feeder-free conditions on Matrigel-coated plates at 37 °C with 5% CO_2 and 3% O_2. To coat culture dishes, Matrigel solution was prepared by diluting Matrigel (Corning, 354277) 1:40 in ice-cold DMEM/F12 medium (Gibco, 11330057) and stored at 4 °C until use. Before plating cells, 6-cm dishes were coated with 1 ml Matrigel solution and incubated for 30 min at 37 °C. PSCs were maintained in TeSR/E8 medium (Stemcell Technologies, 05990) supplemented with 100 U ml^{-1} penicillin and 100 µg ml^{-1} streptomycin (Thermo Fisher, 15140122) with daily medium changes and passaged 1:8 to 1:10 when confluency reached around 70%. To detach cells, plates were rinsed with 2–3 ml PBS (Gibco, 10010023) and treated with 0.5 ml ReLeSR reagent (Stemcell Technologies, 05872) for 5 min at 37 °C. When single-cell dissociation was needed, Accutase (Sigma-Aldrich, A6964) was used instead of ReLeSR, and 5 µM ROCK inhibitor Y-27632 (Tocris, 1254) was added to the medium to enhance cell survival in the first 24 h. All participants signed an informed consent form, and the use of PSCs was approved by the ethical committee of the University of Milan. All iPSC lines were reprogrammed by at least 15 passages. All PSCs have been routinely verified to be *Mycoplasma* free by routine PCR testing, and their identity was confirmed by short tandem repeat profiling. Details about the PSC lines can be found in Supplementary Table 1.

Multiplexing strategies

In the experimental design, we adopted two distinct multiplexing strategies, namely, mosaic and downstream, that differed in the moment at which the different cell lines were mixed.

In the former multiplexing approach, CBOs were generated by mixing equal amounts of PSCs derived from each cell line to obtain mosaic brain organoids. Briefly, PSC lines were dissociated in parallel at the single-cell level, and cells were counted separately and then mixed in equal proportions to obtain a mosaic cellular suspension. After diluting the cell suspension to the desired concentration of 2×10^5 cells per ml, organoids were generated as explained in the following chapter entitled 'Cortical brain organoids'.

In the downstream approach, CBOs were independently generated from each individual PSC line and grown separately. When reaching the desired analytical time point, organoids from each cell line were dissociated in parallel, and cells were counted and mixed in equal proportions to obtain a pooled cell suspension.

For the comparison of neurodevelopmental cell types and trajectories between mosaic and downstream multiplexed CBOs, the design included four iPSC lines (CTL08A, CTL01, CTL02A, CTL04E) across replicates in both multiplexing modalities. For Census-seq-based assessment of mosaic organoid scalability, mCBOs were generated with different combinations of PSC lines as shown in Fig. 6a.

Cortical brain organoids

Pure line-derived and mosaic brain organoids were generated using an adaptation of the previously described protocol[78], which allows one to obtain dorsal telencephalon cortical organoids, introducing orbital shaking on day 12 of differentiation as previously published by us in refs. 5,11,15. PSCs were grown on Matrigel-coated plates to a confluency of approximately 60–70%, dissociated with Accutase (Sigma, A6964) and resuspended in TeSR/E8 medium supplemented with 5 µM ROCK inhibitor Y-27632 (Tocris, 1254) to reach a final concentration of 2×10^5 cells per ml. The cell suspension (100 µl per well) was seeded in ultra-low-attachment, U-bottom 96-well plates (System Biosciences, MS9096UZ) and then centrifuged for 3 min at 150 rcf to promote formation of embryoid bodies. Plates were incubated at 37 °C with 5% CO_2 and 3% O_2 for 2 d, and then the first medium change was performed, substituting TeSR/E8 with neural induction medium containing 80% DMEM/F12 medium (Gibco, 11330057), 20% knockout serum (Gibco, 10828028), non-essential amino acids (1:100, Sigma, M7145), 0.1 mM cell culture-grade 2-mercaptoethanol solution (Gibco, 31350010), GlutaMAX (1:100, Gibco, 35050061), penicillin at 100 U ml^{-1} and streptomycin at 100 µg ml^{-1} (Thermo Fisher, 15140122), 7 µM dorsomorphin (Sigma, P5499) and 10 µM TGF-β inhibitor SB431542 (MedChemExpress, HY-10431). Since that moment, defined as day 1, cultures were grown in normal oxygen conditions (21% O_2). Medium changes were performed daily for the subsequent 4 d, and, on the fifth day, neural induction medium was substituted with complete neurobasal medium, composed of neurobasal medium (Gibco, 12348017), B-27 supplement without vitamin A (1:50, Gibco, 12587001), GlutaMAX (1:100, Gibco, 35050061), penicillin at 100 U ml^{-1} and streptomycin at 100 µg ml^{-1} (Thermo Fisher, 15140122) and 0.1 mM cell culture-grade 2-mercaptoethanol solution (Gibco, 31350010) supplemented with 20 ng ml^{-1} FGF2 (PeproTech, 100-18B) and 20 ng ml^{-1} EGF (PeproTech, AF-100-15). On day 12, organoids were transferred by pipetting with cut-end pipette tips from 96-well to 9-cm ultra-low-attachment dishes (System Biosciences, MS-90900Z) and placed on a standard orbital shaker (VWR Standard Orbital Shaker, Model 1000). From day 12 onward, medium changes were performed every other day. On day 23, FGF and EGF were replaced with 20 ng ml^{-1} BDNF (PeproTech, 450-02) and 20 ng ml^{-1} neurotrophin 3 (PeproTech, 450-03) to promote differentiation of neural progenitors. From day 42 onward, complete neurobasal medium without BDNF and NT3 was used, performing medium changes every other day.

Cell cycle analysis

Pure line-derived organoids were subjected to cell cycle analysis at multiple time points after their generation along with cell suspensions employed in their generation. Organoids were collected at differentiation days 0, 5, 12, 25, 50 and 75, dissociated at 37 °C for 5 min with trypsin-EDTA (Euroclone, ECB3042) (differentiation days 0, 5 and 12) or for 30 min with papain (Stemcell Technologies, 07466) (differentiation days 25, 50 and 75); papain was more efficient for the dissociation of mature organoids.

As for the first analytical replicate (Fig. 2c), cells were fixed with cold ethanol and stored at +4 °C. Upon completing the longitudinal cohort, cell suspensions were rinsed with PBS, stained overnight with 3 µM propidium iodide solution (Thermo Fisher, P1304MP) in the presence of 25 µg ml^{-1} RNase I (Thermo Fisher, EN0601) and analyzed with a FACSCelesta instrument (BD Biosciences) to measure DNA content. Analyses were performed with FlowJo version 10 software (BD Biosciences).

In the second analytical replicate (Fig. 2d), to avoid the deterioration of early time point samples, 1 million cells per sample were resuspended in PBS supplemented with 0.1% BSA (Sigma, A9418), stained for 20 min at 37 °C with 2 µg ml^{-1} Hoecst 33342 (Sigma-Aldrich, B2261) in the presence of 5 µM verapamil (Sigma, V4629) and analyzed with a CytoFLEX instrument (Beckman Coulter Life Sciences) to measure DNA content.

Analyses were performed with FlowJo (replicate 1, BD Biosciences) or FCS Express 7 software (replicate 2, De Novo Software). Raw counts are shown in Supplementary Data 1, and the gating strategy is shown in Supplementary Fig. 2.

Histological analysis

Organoids were collected on differentiation days 50 and 100, washed with PBS and fixed overnight at 4 °C in 4% paraformaldehyde–PBS solution (Santa Cruz, sc-281692). After rinsing with PBS twice, samples were embedded in 2% low melting agarose, placed in 70% (vol/vol) ethanol and immediately given to the Tissue Processing Unit for paraffin embedding, sectioning and routine hematoxylin–eosin staining.

Deparaffinization and rehydration were achieved by consecutive passages of 5 min each in the following solutions: 2× histolemon (Carlo Erba, 454912), 100% ethanol, 95% ethanol, 80% ethanol and 2× ddH_2O. Sections were then incubated for 45 min at 95 °C with 10 mM

sodium citrate buffer (VWR Chemicals, 27833) with 0.05% Tween-20 (Sigma, P1379) for simultaneous antigen retrieval and permeabilization and then equilibrated at room temperature for at least 2 h. After 30 min of blocking with 5% normal donkey serum (Jackson ImmunoResearch, 017-000-121) in PBS, incubation with primary antibodies in PBS with 5% normal donkey serum was performed. The following primary antibodies were used: anti-PAX6 (rabbit, 1:200, BioLegend), anti-TUJ1 (mouse, 1:1,000, BioLegend), anti-nestin (mouse, 1:500, Millipore), anti-reelin (mouse, 1:400, Millipore), anti-CTIP2 (rat, 1:400, Abcam), anti-SATB2 (mouse, 1:400, Abcam), anti-HOPX (rabbit, 1:500, Sigma), anti-SOX2 (goat, 1:1,000, R&D Systems), anti-MAP2 (guinea pig, 1:200, Synaptic Systems), anti-NeuN (mouse, 1:200, Abcam), anti-Ki-67 (rabbit, 1:250, Abcam). The day after, secondary antibodies conjugated with Alexa Fluor 488, 594 or 647 (donkey, 1:400, Thermo Fisher) were diluted in PBS and applied to the sections for 1 h, followed by a 5-min incubation with 1 µg ml^{-1} DAPI solution. After each incubation, three 5-min washing steps with TBS buffer were performed. After a final rinse with deionized water, slides were dried and mounted using Mowiol mounting solution. Images at ×20, ×40 and ×63 magnification were acquired using a DM6 B MultiFluo microscope (Leica) equipped with an Andor Zyla VSC-04470 sCMOS camera (Fig. 2) or with an ECLIPSE Ti2 Crest microscope (Nikon) coupled with a Photometrics Prime 95B camera at ×20 magnification (Extended Data Fig. 1) and then processed with Fiji software. For Extended Data Fig. 1, maximum intensity projection was performed, choosing the 'maximum intensity' option, and a gamma correction of 0.7 was applied to the AF488 channel.

Cortical brain organoid processing for single-cell transcriptomic analysis

Organoids were collected on differentiation days 50, 100, 250 and 300 (±3 d). Three to five organoids per condition were dissociated by incubation with a solution of 0.5 mg ml^{-1} Collagenase/Dispase (Sigma) with 0.22 mg ml^{-1} EDTA (Euroclone) and 10 µl DNase I at 1,000 U ml^{-1} (Zymo Research) for 30–45 min according to organoid size. Digested suspensions were filtered through 70-µm-pore Flowmi Cell Strainers (Sigma, BAH136800040), resuspended in PBS and counted using a TC20 Automated Cell Counter (Bio-Rad). For the day 100 downstream sample, the single-cell suspension was further centrifuged at 500 rcf for 3 min and resuspended in 200 µl cold PBS–0.04% BSA. Pre-chilled (800 µl) 100% methanol was added dropwise for a final concentration of 80%. Cells were fixed for 30 min and stored at −80 °C for 6 months. For recovery of a single-cell suspension, fixed cells were thawed at 4 °C (all steps at 4 °C) and centrifuged at 1,000 rcf for 5 min, the supernatant was completely removed, and pre-chilled SSC cocktail (3× SSC, 0.04% BSA, 1% SUPERase-In, 40 mM DTT) was added. Cells were counted and resuspended at a concentration 1,000 cells per µl. Droplet-based single-cell partitioning and scRNA-seq libraries were generated using the Chromium Single Cell 3′ Reagent v2 Kit (10x Genomics) following the manufacturer's instructions[79]. Briefly, a small volume (6–8 µl) of single-cell suspension at a density of 1,000 cells per µl was mixed with RT–PCR master mix and immediately loaded together with Single Cell 3′ gel beads and partitioning oil into a Single Cell 3′ Chip. The gel beads were coated with unique primers bearing 10x cell barcodes, unique molecular identifiers and poly(dT) sequences. The chip was then loaded onto a Chromium instrument (10x Genomics) for single-cell GEM generation and barcoding. RNA transcripts from single cells were reverse transcribed within droplets to generate barcoded full-length cDNA. After emulsion disruption, cDNA molecules from each sample were pooled and pre-amplified. Finally, amplified cDNA was fragmented, and adaptor and sample indices were incorporated into finished libraries that were compatible with Illumina sequencing. The final libraries were quantified by real-time quantitative PCR and calibrated with an in-house control sequencing library. The size profiles of the pre-amplified cDNA and sequencing libraries were examined on the Agilent Bioanalyzer 2100 using a High Sensitivity DNA chip (Agilent). Two indexed libraries were pooled equimolarly and sequenced on the Illumina NovaSeq 6000 platform using the v2 Kit (Illumina) with a customized paired-end, dual-indexing (26/8/0/98-bp) format according to the recommendation of 10x Genomics. Using proper cluster density, a coverage of around 250 million reads per sample (2,000–5,000 cells) was obtained, corresponding to at least 50,000 reads per cell.

Generation of CellPlex libraries

Two independent downstream multiplexed datasets were generated for external validation of the SCanSNP pipeline, labeling each genotype with a specific barcode from the 3′ CellPlex Kit (10x Genomics, PN-1000261). Briefly, organoids were collected on day 50 and dissociated with an HBSS-based solution containing 30 U ml^{-1} papain (Stemcell, 07466) and 125 U ml^{-1} DNase I (Zymo Research) for 30–45 min according to organoid size. After filtering through 70-µm-pore Flowmi Cell Strainers, cells were counted, and 1 million live cells per sample were taken and labeled with a unique barcode from the 3′ CellPlex Kit Set A according to the manufacturer's instructions. Once the washing steps were completed, 1×10^5 live cells per sample were pooled together, and the concentration was adjusted to be within $1.2–1.6 \times 10^6$ cells per ml for droplet-based single-cell partitioning. Libraries were generated using the Chromium Next GEM Single Cell 3′ v3.1 Kit (10x Genomics) similarly as described previously. Target coverage for gene expression was set at 50,000 reads per cell, whereas, for CellPlex multiplexing oligonucleotides, a sequencing depth of 5,000 reads per cell was chosen as recommended.

Generation of bulk RNA-seq-based variant-calling file (VCF)

Reference genotypes for the deconvolution were obtained from bulk RNA-seq data of pure line-derived organoids.

Total RNA was isolated from pure line organoids with the RNeasy Micro Kit (Qiagen) according to the manufacturer's instructions. RNA quantification and integrity was assessed by electrophoretic analysis with the Agilent 2100 Bioanalyzer. The TruSeq Stranded Total RNA LT Sample Prep Kit (Illumina) was used to run the library for each sample using 500 ng total RNA as the starting material. Sequencing was performed with the Illumina NovaSeq 6000 platform, sequencing on average 35 million 50-bp paired-end reads per sample.

First, raw reads were aligned to the GRCh38 version 93 Ensembl reference genome using STAR[80] in two-pass mode to learn splicing junctions from data, subsequently, read group tags were made uniform, and optical and PCR-duplicated reads were marked using GATK MarkDuplicates. The resulting BAM file (in silico multiplexed BAM creation is described in the Supplementary Methods) was recalibrated using BaseRecalibrator and ApplyBQSR with the sorted 00-All.vcf.gz (b151_GRCh38p7) file from dbSNP. Next, we performed variant calling using GATK HaplotypeCaller[81] with the option 'dont-use-soft-clipped-bases' –standard-min-confidence-threshold-for-calling set to 30 and –min-base-quality-score equal to 20. We applied thresholds for MIN_DP, DP and quality (GQ) of 10, 10 and 20, respectively. Finally, we used CombineGVCFs and GenotypeGVCFs exclusively to merge together GVCFs of genotypes mixed in the corresponding single-cell experiment, ignoring the aggregated INFO field.

Alignment and genotype demultiplexing

scRNA-seq data were aligned using Cell Ranger 3.0.0 count and the matched reference provided from 10x. For subsequent demultiplexing and downstream analyses, only droplets passing the Cell Ranger filter were considered. For demultiplexing, we applied demuxlet, souporcell, Vireo and SCanSNP. The final identities used are the result of the consensus call (consensus call setup is described in the Supplementary Methods). With the exception of souporcell, all the tools were provided with bulk RNA-seq-derived VCFs and were embedded in a collective

singularity image. To optimize the deconvolution process for our samples, we used demuxlet with default parameters and the setting '–doublet-prior' according to the number of retained droplets after Cell Ranger filtering and V3 kit specs doublet expectation. CellSNP, the Vireo companion tool for single-cell variant calling, was launched 'with -p 10–minMAF 0.1–minCOUNT 20', and subsequently Vireo was launched on the combined VCF file cleaned from loci containing missing calls. Souporcell was launched using the available singularity image specifying only the number of genotypes in the mixture (*k*). hiPSC lines included in this analysis are listed in Supplementary Table 1.

Single-cell data preparation

All downstream analyses were performed within the SCANPY single-cell analysis framework[82].

Basic filtering was done right after importing count matrices from Cell Ranger. We inspected the number of genes, mitochondrial gene counts and ribosomal gene count distributions and adopted dataset-specific thresholds to remove droplets with likely technical issues. Next, we removed droplets called for low quality or doublets according to the consensus call. After merging the seven datasets, cell counts were normalized and log transformed using the 'sc.pp.normalize_total' and 'sc.pp.log1p' SCANPY functions. Finally, we regressed out the effect of total counts and the percentage of mitochondrial transcripts with the 'sc.pp.regress_out' and 'sc.pp.scale' functions. All functions were run with default parameters.

Cell filtering and annotation

We relied on multiple tiers of annotation, filtering and dimensionality reduction. First, we removed proteoglycan (*WLS*, *ANXA2*, *TPBG*, *RSPO3*, *DCN*, *BGN*, *MYH3*)-expressing cells, considered non-relevant for our downstream analyses. We next iteratively partitioned cells and assessed the top marker via Leiden and rank_gene_groups, respectively, to remove cells highly expressing adhesion (*WLS*, *TPBG*) and stress (*BNIP3*, *PGK1*, *MT-CO2*) markers while manually annotating cell types using literature markers. Finally, we used SCANPY's score_genes function providing ER stress and hypoxia signatures to remove clusters of cells scoring >0 and >0.3, respectively. Finally, we repartitioned the remaining cells with a Leiden resolution of 0.6 and manually transferred the annotation to new clusters, merging, if needed, different partitions into coherent cell types when key markers were overlapping.

Highly variable gene selection

For HGV detection, we took advantage of having at least two datasets per time point and detected HVGs by time point with the 'sc.pp.highly_variable_genes' SCANPY function, providing each dataset as a separate batch and min_mean = 0.0125, max_mean = 5 and min_disp= 0.5 as parameters. For each time point, we kept only genes found as HVGs in at least two datasets. After major filtering and single-trajectory isolation, HVGs were recomputed on the cell subset in the same fashion. Moreover, we included some relevant neurodevelopmental genes from the literature regardless of whether they were detected as HVGs (Supplementary Data 2).

Dimensionality reduction and dataset integration

After filtering, PCA was computed on defined HVGs, and the seven datasets were integrated via SCANPY's Harmony implementation[83] with a maximum of 20 iterations; the cells' neighborhood graph was then computed on top 15 PCs, specifying 100 as the neighborhood size (*n*). Upon single-trajectory isolation, PCA was recomputed on each cell subset and branch-specific HVGs, Harmony was run with a maximum of 20 iterations, lambda = 2 and theta = 1, and cell neighborhood graphs were computed with ten PCs and 50 *n*; in the case of the excitatory neuron lineage, we used nine PCs and 60 *n*, given the greater differences between early and late branches than those within lineage-continuous cell states of other cases.

Trajectory isolation

To analyze the cell state transitions with trajectory-wise magnification, we isolated the most relevant neurodevelopmental trajectories. For this purpose, we first partitioned cells with the Leiden algorithm with double resolution (1.2) with respect to the previous partitioning and used it as the basis for PAGA[84], obtaining a PAGA graph. The PAGA graph was refined by removing edges with weight <0.05. After complementing the edges to 1 to obtain the equivalent distance graph, we computed the shortest path from root *r* to each endpoint *e* using NetworkX[85] with the Bellman–Ford method, where *r* is the partition of cells with highest counts for the '*TOP2A*' gene and *e* are partitions with the highest rate of the cell types considered one of the endpoints (astrocytes, Cajal–Retzius-like neurons, early excitatory neurons, late excitatory neurons, interneurons and migrating neurons).

Differential abundance analysis

We adopted Milo[45] for differential abundance analysis. We tested for differential abundance between the mosaic organoid dataset and non-mosaic datasets via direct comparison following the Milo standard workflow. The differential abundance test was between each time point and the other two simultaneously; therefore, we provided as model contrast $CT_n = T_n - (T_x + T_y)/2$ for each of the *n* assessed time points, where *x* and *y* are the other two time points, *T* is the timepoint and CT_n is the model contrast for timepoint *n*. Plots were generated displaying enriched/depleted cells' neighbors for each time point with spatial FDR < 0.1.

Differential expression analysis

To compare the molecular impact of mosaic co-culture, we performed direct differential expression analysis between the same genotypes either grown individually (downstream multiplexing) or in mosaics. Given the impact of methanol fixation on the day 100 downstream dataset, (Supplementary Fig. 1b), we relied on day 50 datasets (Supplementary Fig. 1c). To provide a reference of the expected batch-to-batch variability, we also compared the same genotypes when grown in two different mosaic batches (Supplementary Fig. 1d). For both comparisons, we kept the two most abundant genotypes (Supplementary Fig. 1c,d, left) and the three most abundant cell types (that is, proliferating progenitors, radial glial progenitors and neurons; Supplementary Fig. 1c,d, right) present in both experiments. For each genotype and cell type combination, single-cell counts were aggregated to obtain two pseudoreplicates. Gene filtering and normalization were carried out within edgeR[86] using the functions filterByExpr() and calcNormFactors(), respectively; the latter was performed to account for different numbers of cells aggregated into the various pseudoreplicate combinations. Finally, count fitting (glmQLFit edgeR function) and the DE test (glmQLFTest edgeR function) were repeated individually for each cell type comparing the condition (multiplexing paradigm or replicate) while providing the genotype as the blocking variable.

Developmental trajectory analysis

We aimed to assess the distribution of cells along pseudotime by different covariates for migrating neuron, astrocyte, Cajal–Retzius-like neuron and interneuron trajectories after their isolation. We wished to assess whether (1) time point differences mirror the asynchronous development of specific cell types in our in vitro system, (2) the multiplexing paradigm impact on the developmental timing of such populations and (3) whether we had the resolution to capture developmental differences among control genotypes. For each trajectory individually, we computed diffusion map[87] and dpt[88]. Next, for the different time points and the multiplexing paradigm, we computed the kernel density of each dataset (sklearn.neighbors.KernelDensity, kernel = 'gaussian', bandwidth corresponding to 5% of the whole pseudotime window) and plotted mean and ±1 s.d. among dataset densities. For genotype comparison, we kept the most relevant time points for each trajectory

(early and mid for migrating neurons and Cajal–Retzius-like neurons; early, mid and late for the other trajectories) and genotypes for which at least 50 cells were retrieved at each retained time point. Finally, the genotypes were balanced via random sampling to have the same amount of cells across time points. Mean and standard deviations were computed on kernel densities of 50 sampling iterations to assess the stability.

For the isolated migrating neuron trajectory, we confirmed the dpt-based genotype grouping (Fig. 5c) by assessing their behavior along PC1. We started by confirming that PC1 was mainly driven by differentiation, and thus could be used as an alternative of pseudotime measure. To do so, we quantified the PC1 variance (adjusted R^2) explained by the annotated Leiden covariate using the ordinary least-square implementation of the statsmodels Python library (Extended Data Fig. 7a, left). Similarly, we then assessed the PC1 variance explained by the genotype (Extended Data Fig. 7a, right) and genotype distributions along PC1 (Extended Data Fig. 7b).

Subsequently, we used tradeSeq[89] (fitGAM function was run with nknots = 8 after trimming cells in the first and 99th dpt percentiles to increase stability at lowly sampled extremes) and detected pseudotime-driven transcriptional difference within each lineage for all but excitatory neuron trajectories, whereas, for excitatory neurons, we used tradeSeq to find key transcriptional differences between early–mid and late neuronal trajectories.

Within lineage differences, to extract key driver genes along pseudotime, we isolated HVGs between lineage extremes (tradeSeq startVsEndTest(), pVal ≤ 0.001 and log (FC) ≥ 2).

For early-versus-late excitatory neuron differences, to detect key differences between the two lineages, we considered both the greatest divergently expressed genes at the terminal states (tradeseq diffEndTest(), pVal ≤ 0.001 and log (FC) ≥ 2) and the transiently varying genes, defined as simultaneously low ranked from the tradeSeq:diffEndTest function and high ranked from the tradeSeq:patternTest(pVal ≤ 0.001) function.

Allele-specific expression analysis

We leveraged our multigenotype design to carry out ASE analysis for each cell type annotated. We first produced the read pileup at genomic (biallelic-only) loci displaying variability within our cohort using SCanSNP pileup mode, which provided an anndata[82] with nCells × nLoci dimensionality and two layers: 'Refreads' and 'Altreads', with the count of reads presenting or not presenting the variant, respectively. To perform the analysis, we summed the reads mapping to variant sites (at least one heterozygous genotype) of the same cell type. If multiple genotypes were heterozygous at the given location, reads were included in the sum regardless of the original genotype. As it is not granted that, among different individuals, if present, the dominant allele is the same, we first checked that, for loci with detected ASE, the dominant allele was coherent before merging coverages from different individuals. We observed minimal discrepancy, that is, at a given locus, the dominant allele was vastly the same (always Ref or Alt) across genotypes (information available in the Github repository of the paper, Notebook 04_ASE/13.1_SanityCheck). For each cluster, we kept only loci with at least 20 reads, computed the β value (Alt reads/(Alt reads + Ref reads)) and performed binomial test and FDR correction ($q < 0.05$) provided by the SciPy[90] and statsmodels implementations, respectively. To calculate the correlation among cell types, we used β values of loci with detected ASE in at least one cell type and covered with at least 20 reads in all cell types. Correlation was computed in pandas with 'spearman' metrics.

SCanSNP

In our benchmark, the presence of low-quality droplets and doublets was observed to be an open challenge also for well-established methods when assigning genotypes (IDs) to droplets in the genetic demultiplexing. With those challenges in mind, we developed SCanSNP (available at https://github.com/GiuseppeTestaLab/SCanSNP) by dividing the demultiplexing and filtering in 3(+1) steps:

1. Best ID detection per droplet: here, as for other approaches, we leveraged the accessibility of bulk RNA-seq data to generate a function that maximizes the score difference of each ID to the sequenced droplets:

$$S_g = \sum_{i=1}^{n}\left(\left(\frac{A_i \times a_{ig}}{t_i} + \left(\frac{R_i \times r_{ig}}{T_i}\right)\right)\right),$$

where S_g is the score for ID g in each droplet, i are the loci for which allelic information is accessible from bulk RNA-seq, A and R are, respectively, the number of reads supporting alternative and reference alleles, a and r are the number of alternative and reference alleles in g and t and T are total alternative and reference alleles in the cohort at locus i.
2. Second-best ID determination: given *an* m-by-g contribution matrix, where m are the droplets and g are the multiplexed IDs containing the number of reads supporting genotype-specific alleles. We used this matrix to iteratively train a multinomial logistic regression model to predict which is the most likely ID after the first one, assuming ambient contamination consistent across droplets. We split the contribution matrix into groups of droplets sharing the best ID according to step 1; for each group, we trained the model on counts and labels from other groups to predict the second-best ID of barcodes in the current group.
3. Doublet detection: to allow doublet detection to be specific and flexible while accommodating genetic contributions ranging from balanced doublets to the presence of a cell and debris in the same drop, we implemented a method similar to the one adopted in ref. 22. Starting from the previous m-by-g contribution matrix, for every genotype g, we define as negative droplets the ones that do not contain that genotype as the best ID according to the first step and fit a negative binomial distribution via the fitdistrplus[91] R function on counts supporting private g alleles. We therefore used the 99% quantile of the fitted distribution as the positivity threshold. Droplets positive for more than one ID are considered multiplets.
4. We finally took advantage of the mixed-genotype design to structure an added layer of a low-quality droplet detection to be used during consensus call aggregation.

We applied a Gaussian mixture model expectation-maximization algorithm (implemented through the R mixtools[92] package) to separate droplets with 'low' and 'high' signal-to-noise ratios by computing log (FC) between the first- and second-best predicted IDs. We started by preparing a new contribution matrix similar to the one in passage 2 but considering only non-ambiguous loci between each possible pair of best and second-best IDs in the dataset. Additionally, before log (FC) calculation, we add pseudocounts, which mimics average ambient RNA contamination coming from each ID, calculated as the average rate of reads deriving from the other genotype's unambiguous reads when they are not labeled as first ID or second ID across all droplets (according to the contribution matrix); similar to the approach proposed in the hashedDrops function from the package MarioniLab/DropletUtils[93,94], this step ensures that log (FC) is always defined for all droplets. Given the nature of the model, the resulting classification assumes the presence of two distinct populations that can be separated based on the proportion of the two IDs, and, given that it is computed after doublet detection, it will likely detect those droplets that embed enough ambient RNA to pass the Cell Ranger emptyDrops filter, while it should not be used if any sort of prior filtering of low-quality droplets has already been done. Benchmarking of SCanSNP and genetic demultiplexing in barcode-tagged samples are described in the Supplementary Methods.

Power estimation for single-cell eQTLs

We estimated the eQTL power using our R package scPower (version 1.0.2)[50] for sample sizes between 25 and 200 and for number of cells per sample between 250 and 1,500, keeping the read depth as in our experiment. We fitted the required expression priors per cell type using the complete scRNA-seq dataset, combining the different multiplexing strategies and time points, and took effects from a previously published single-cell eQTL study in IPS cells[35], excluding eQTLs from ROT-treated cells. Genes were defined as expressed with at least three counts in at least 9.5% of the samples.

Modeling mCBO clonal dynamics

We integrated the longitudinal Census-seq data with mCBO imaging to model mCBO clonal dynamics (modeling of PSC and CBO growth with imaging; Census-seq and Census-seq ranking of iPSC lines in mosaic organoids are described in the Supplementary Methods). First, we used already published imaging of CBOs stained for nuclear markers to estimate the density of cells inside CBOs (0.000767495 per μm^3). Next, for each replicate of each mosaic combination that we longitudinally profiled with imaging, we multiplied the measured area of the mCBOs (converting this to the equivalent sphere volume simply through the sphere volume formula) by the density to estimate the number of cells present in each mCBO at each time point. We computed the average of this value for each mosaic combination across the replicates for each time point. We then multiplied this value by the average contribution of each iPSC line in each mosaic combination, as given by the corresponding Census-seq data (correlation across multimodal ranking is described in the Supplementary Methods). This represents the estimated number of cells for each PSC line in each mCBO at the different time points. The ratio between sequential time points was computed to estimate the line-specific growth rates, and their stability across mosaic combinations (Fig. 6b and Extended Data Fig. 10b) as defined by $n_{it} = r_{it} \cdot n_t$ and $g_i = (n_{it+1})/n_{it}$ with n_t as the total number of cells in the organoid at time t, r_{it} is the ratio of cells measured by Census-seq, n_{it} is the number of cell per genotype at time t, g_i is the growth rate of the genotype i. The empirical distribution is defined by expanding the distribution of $g_i \forall$ genotypes i from our measured data. We used the computed line-specific growth rates to estimate an empirical distribution of possible growth rates. We then generated the growth rates of the lines inside the mCBO randomly (10,000 cycles of sampling) from the probability distribution calculated over the domain. The process was repeated for sample sizes between two and 20 theoretically multiplexed lines.

Because we had a constant number of starting cells (20,000) when generating mCBOs, we multiplied this by the Monte Carlo simulated growth rates, thus modeling the expected number of cells per line for all the possible conditions (from two to 20 starting lines).

To translate this into useful guidelines for experimental disease-modeling pipelines, we considered that at least 100 cells per neurodevelopmental cell type should be recovered for each PSC line grown in mCBOs and that 100,000 cells can be sequenced. On the basis of these parameters, Supplementary Fig. 10c shows the number of PSC lines (y axis) retrieved with the highest empirical probability given the number of PSC lines used to generate mCBOs (x axis). We finally employed the curve_fit function from the SciPy python library to estimate the coefficients a and b of the power function $N = aI^b + c$ linking the number of mixed cell lines I and the average number of recovered lines N (Fig. 6c). The scalability of mosaic experiments in terms of experimental timeline is described in the Supplementary Methods.

Statistics and reproducibility

Statistical analyses were carried out using tests appropriate for each assessed modality using SCANPY, tradeSeq and Milo for single-cell transcriptomics analyses.

The threshold for statistical significance was spatial FDR < 0.1 for differential abundance (Milo) and $P < 0.05$ for other statistical tests. All details on sample size, number of replicates, statistical tests and significance are provided in the relevant figure legends. CBOs were differentiated in multiple independent batches, and the number of replicates was chosen on the basis of previous published studies on brain organoids.

The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assignments.

Graphics and figures

Final figure panels were assembled using Adobe Illustrator version 27.0.1. For the organoids, cells and human shapes in Fig. 1, templates were downloaded from BioRender and subsequently modified.

Reporting summary

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

Data availability

The scRNA-seq data generated in this study are accessible via ArrayExpress (accession E-MTAB-14574). WGS and low-pass WGS sequencing data have been deposited at the European Genome–Phenome Archive with the study identifier EGAD50000000978. Additional resources include the reference genome Ensembl GRCh38 version 93, dbSNP version b151 GRCh38p7 (00-All.vcf) and single-cell eQTLs from Jerber et al.[35] (Table 7). Source data are provided with this paper.

Code availability

Full code used for the analyses can be retrieved at https://github.com/GiuseppeTestaLab/organoidMultiplexing_release. The latest release of SCanSNP and the docker image link are available at https://github.com/GiuseppeTestaLab/SCanSNP.

References

78. Yoon, S.-J. et al. Reliability of human cortical organoid generation. *Nat. Methods* **16**, 75–78 (2019).
79. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
80. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
81. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at *bioRxiv* https://doi.org/10.1101/201178 (2017).
82. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
83. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
84. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
85. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. In *Proc. Python in Science Conference (SciPy)* (eds Varoquaux, G. et al.) https://doi.org/10.25080/TCWV9851 (SciPy Proceedings, 2008).
86. Chen, Y., Lun, A. T. L. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res.* **5**, 1438 (2016).
87. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
88. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
89. Van den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).

90. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
91. Delignette-Muller, M. L. & Dutang, C. fitdistrplus: an R package for fitting distributions. *J. Stat. Softw.* **64**, 1–34 (2015).
92. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. S. mixtools: an R package for analyzing mixture models. *J. Stat. Softw.* **32**, 1–29 (2010).
93. Lun, A. T. L. et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
94. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.* **9**, 2667 (2018).

Acknowledgements

We thank present and former members of G.T.'s laboratory for collaborating to various extents on this study with both technical help and conceptual discussion. We also thank the Flow Cytometry and Imaging facilities of Human Technopole and the European Institute of Oncology (IEO) and colleagues in the Genomics Unit of the IEO and the Center for Genomic Science of the Istituto Italiano di Tecnologia for helpful discussion at the onset of the project. We thank the European School of Molecular Medicine (SEMM) at which D.C., M.T.R., A. Valenti, S.S., M.P., S.T. and M.L. are/were enrolled as students for their PhD degree program in Systems Medicine. Some components of the schematics were adapted from BioRender. The project, carried out in G.T.'s laboratory at the IEO and at Human Technopole, has been funded by European Union Horizon 2020 research and innovation program grants EDC-MixRisk (634880), ENDpoiNTs (825759), NEUROCOV (101057775), RE-MEND (101057604), R2D2-MH (101057385) and PNRR (Centro Nazionale CN3: RNA, 'National Center for Gene Therapy and Drugs based on RNA Technology').

Author contributions

N.C., D.C. and M.T.R. contributed equally, are listed in alphabetical order and have the right to list their name first in their CV. D.C., M.T.R., A. Valenti, S.S., M.L., S.T. and M.P. are/were PhD students at the European School of Molecular Medicine (SEMM). N.C., C.E.V. and G.T. conceived the project and, with M.T.R. and D.C., implemented the experimental and analytical design; M.T.R. has driven the experimental activities with the help of S.S., M.L., D.B. and S.T.; D.C. has driven the computational work with the help of C.C., A. Valenti, M.B. and K.T.S.; M.P., A. Vitriolo, A.L.T., D.R., M.H. and F.J.T. contributed to the study design and critical discussions and interpretation of the results; N.C., C.E.V. and G.T. supervised wet and computational activities; N.C., D.C., M.T.R., C.E.V. and G.T. wrote the paper with input from all other authors. All authors read and approved the final paper.

Competing interests

F.J.T. consults for Immunai, Singularity Bio, CytoReason and Omniscope and has ownership interest in Dermagnostix and Cellarity. The other authors declare no competing interests.

Additional information

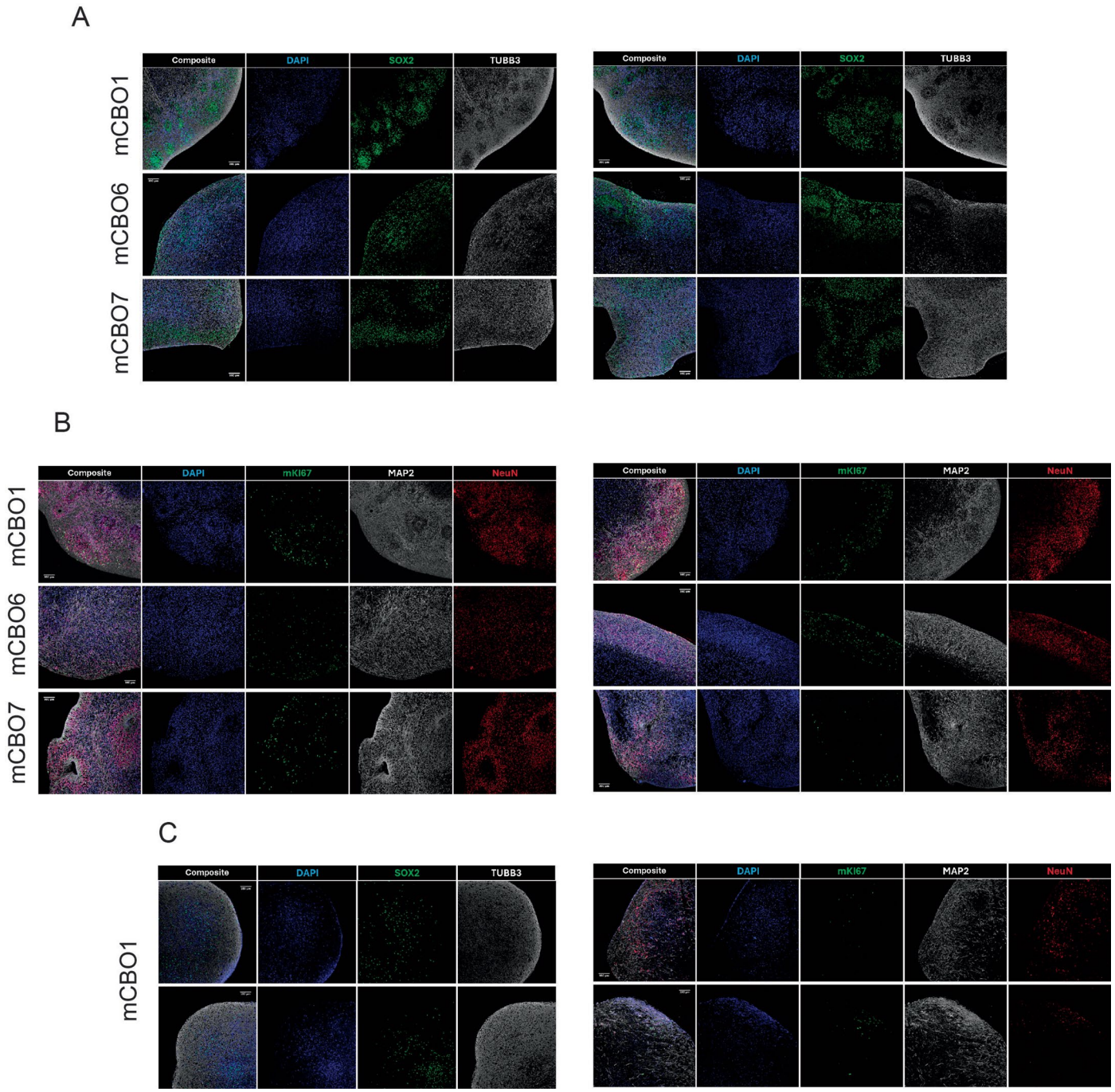
Extended data is available for this paper at https://doi.org/10.1038/s41592-024-02555-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02555-5.

Correspondence and requests for materials should be addressed to Giuseppe Testa.

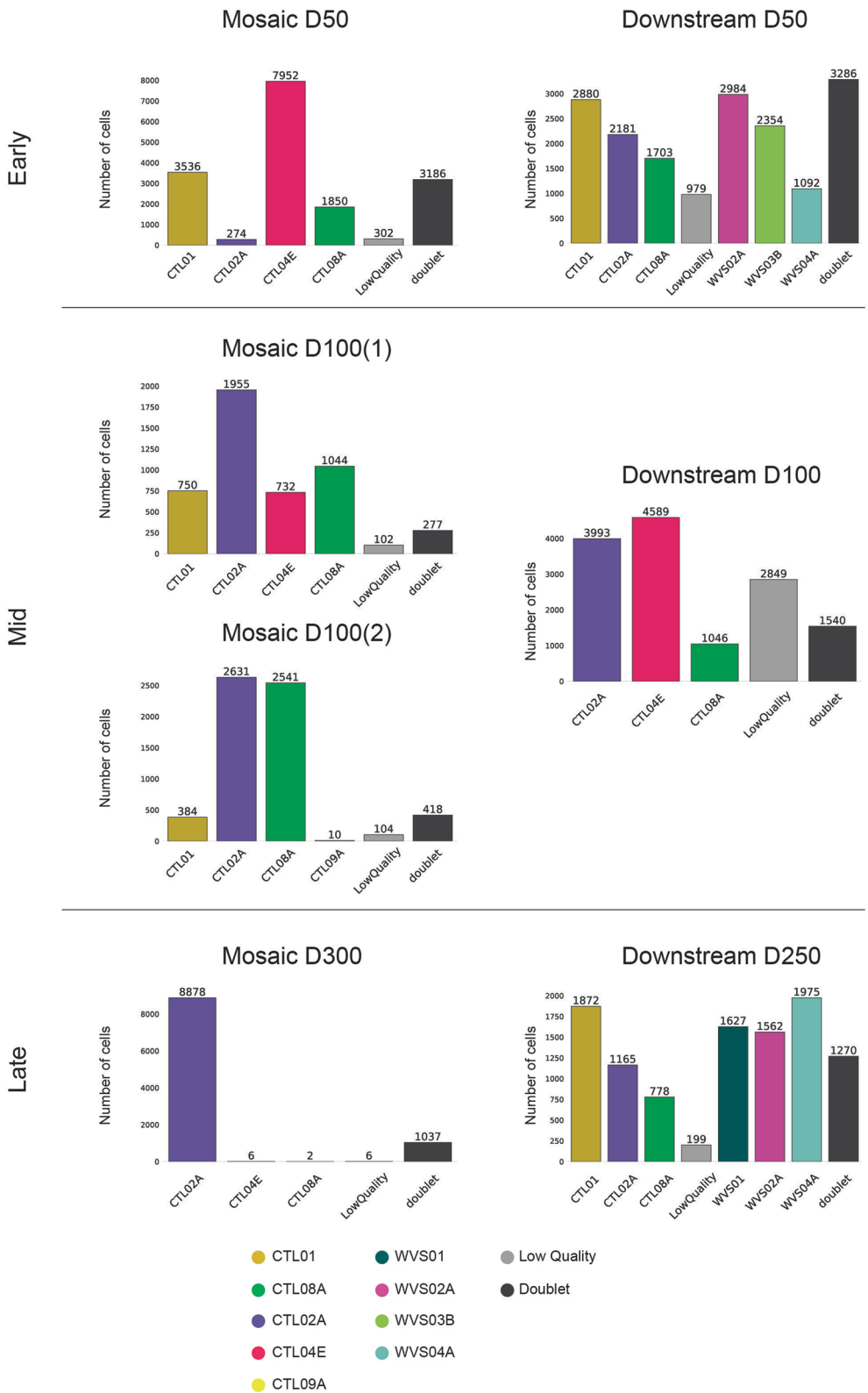
Peer review information *Nature Methods* thanks Carlo Colantuoni and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

https://doi.org/10.1038/s41592-024-02555-5



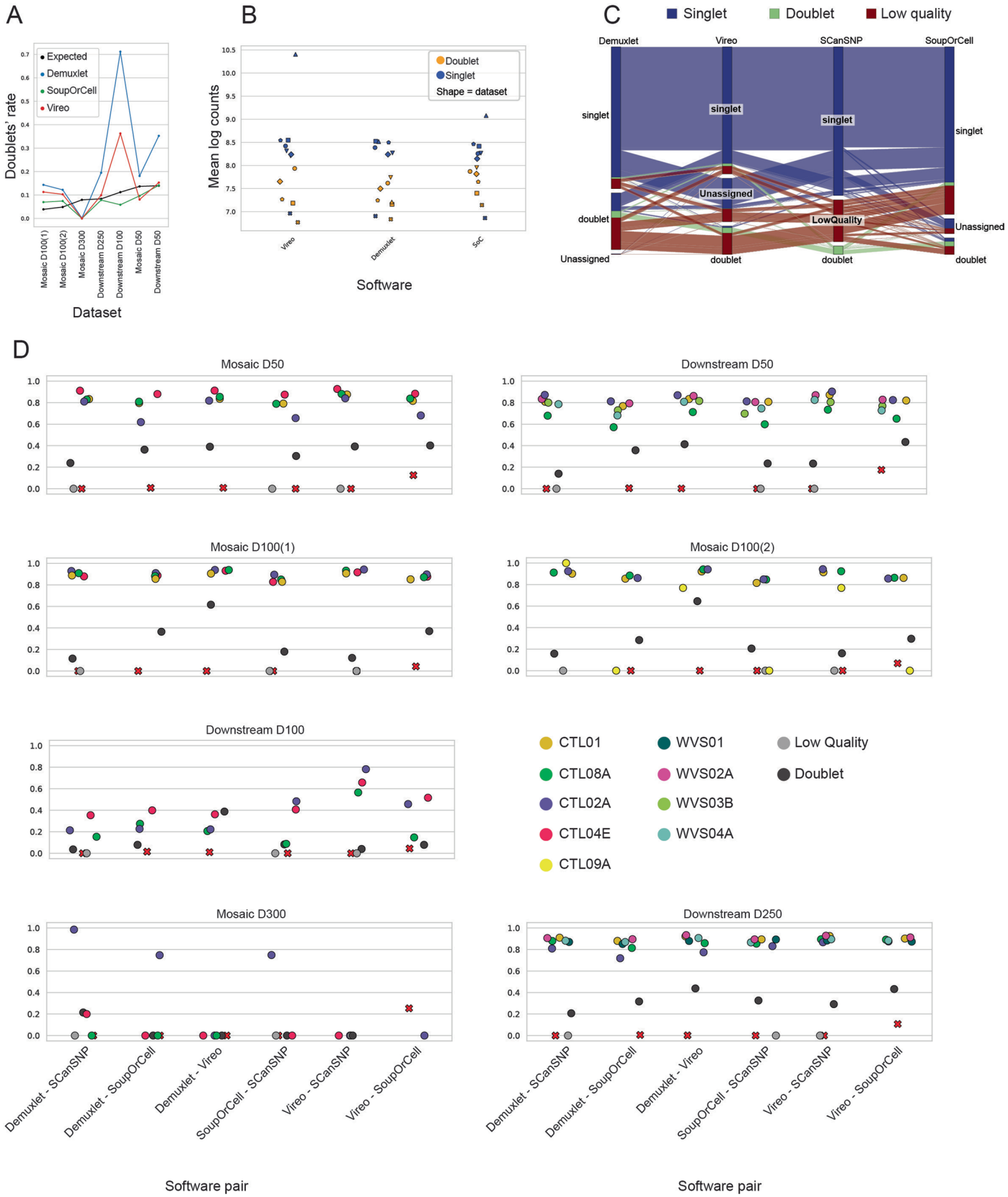
Extended Data Fig. 1 | mCBO immunofluorescence characterization. Immunofluorescence-based benchmarking of different mosaic CBOs combinations. At differentiation day 50 (**a**, **b**), mosaic CBOs mixes 1, 6, and 7 show consistent expression of the neuronal lineage-specific tubulin TUBB3 as well as the presence of ventricular-like structures positive for the neural stem cell marker SOX2 (**a**). Similarly to the *in vivo* counterpart, these structures display high rates of proliferation as shown by the focal enrichment in mKI67 positive cells (**b**). Outside ventricular-like structures, the presence of neurons can be appreciated by the broad presence of NeuN positive nuclei as well as by the uniform presence of MAP2 positive cellular processes (**b**). At differentiation day 135 (**c**), mosaic CBOs mix1 display more mature ventricular-like structures characterised by reduced luminal area and a reduced and scattered expression of both SOX2 and mKI67 positive cells, whereas both NeuN positive nuclei and the sharpness of TUBB3 and MAP2 signal appears increased with longer cellular processes being clearly detected by anti-MAP2 staining.

https://doi.org/10.1038/s41592-024-02555-5



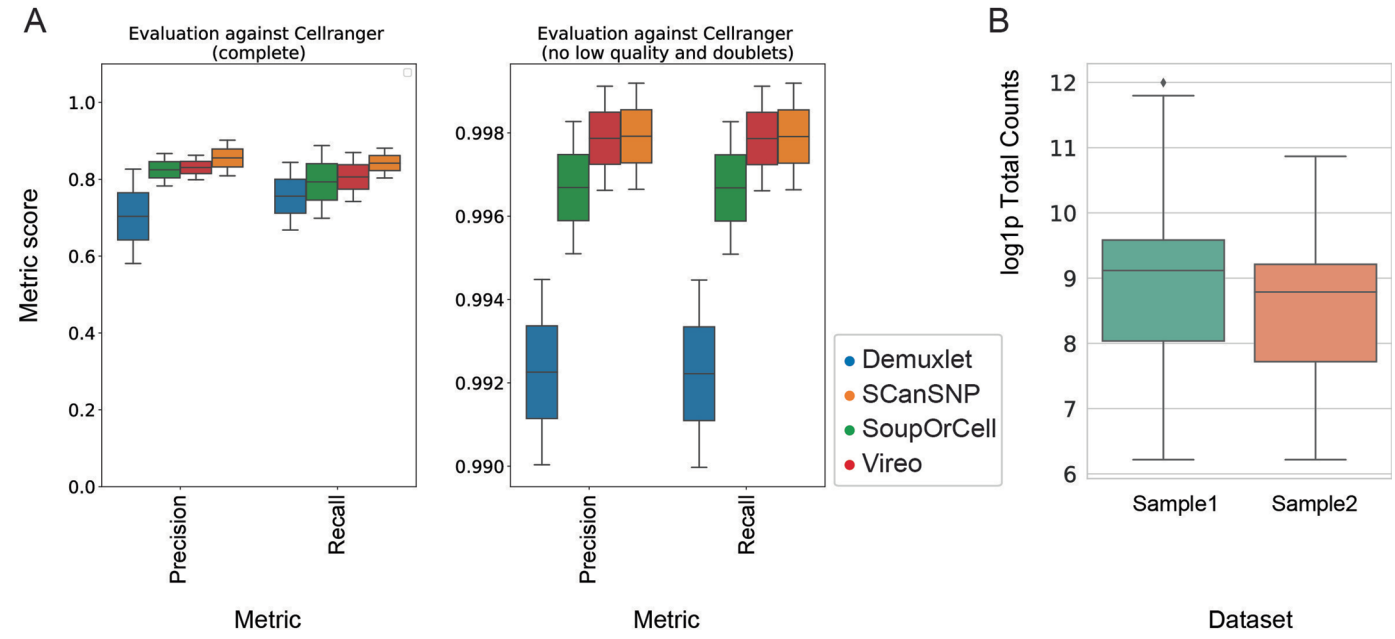
Extended Data Fig. 2 | Dataset composition by genotype. Barplot representing number of cells by genotype according to the consensus call prior to filtering. WVS01H, WVS02A, WVS03B, WVS04A, CTL09A were not included in downstream analysis since there were no replicates across multiplexing modalities.

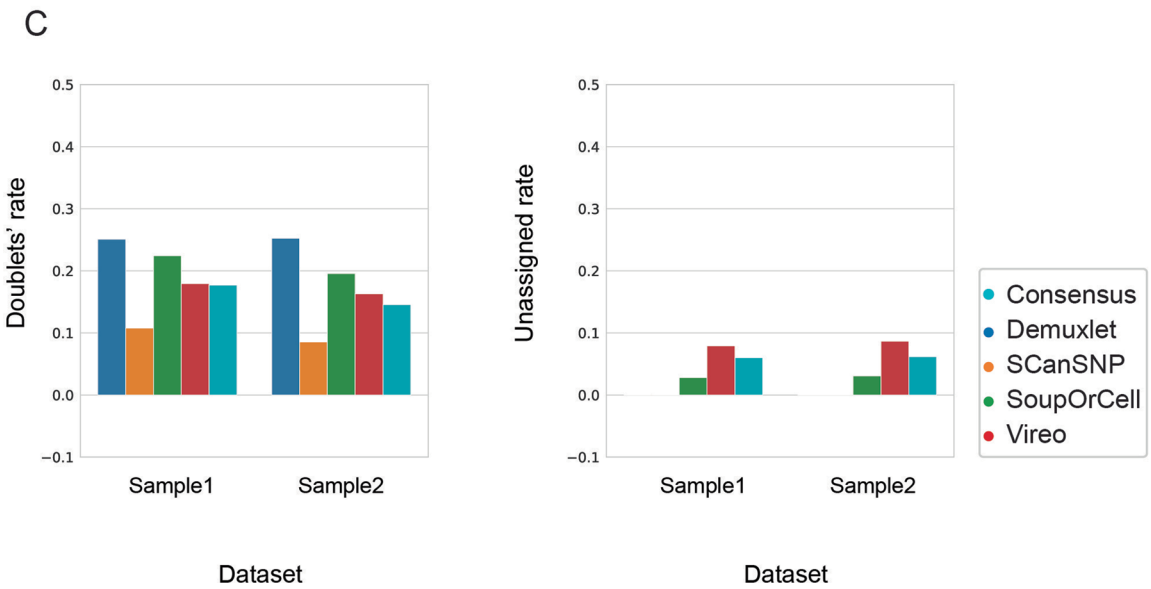
https://doi.org/10.1038/s41592-024-02555-5



Extended Data Fig. 3 | Demultiplexing performance assessment. a) Doublet rate by dataset and algorithm. Lines are coloured by demultiplexing algorithm, datasets (x axis) are ordered by number of retrieved cells. **b**) Average log counts distribution by demultiplexing algorithm. Dots are coloured by predicted singlet or doublet identity by each algorithm; the shape of the dots encodes each of the 7 datasets. **c**) Alluvial plot displaying Singlets, Doublets, Unassigned and Low-quality classes mappings across demultiplexing algorithms. Cells (rows) are coloured according to SCanSNP assignment class. **d**) Pairwise agreement among software, divided by dataset. The agreement is expressed as a Jaccard similarity of each called identity between 2 software; x represent unassigned cells.

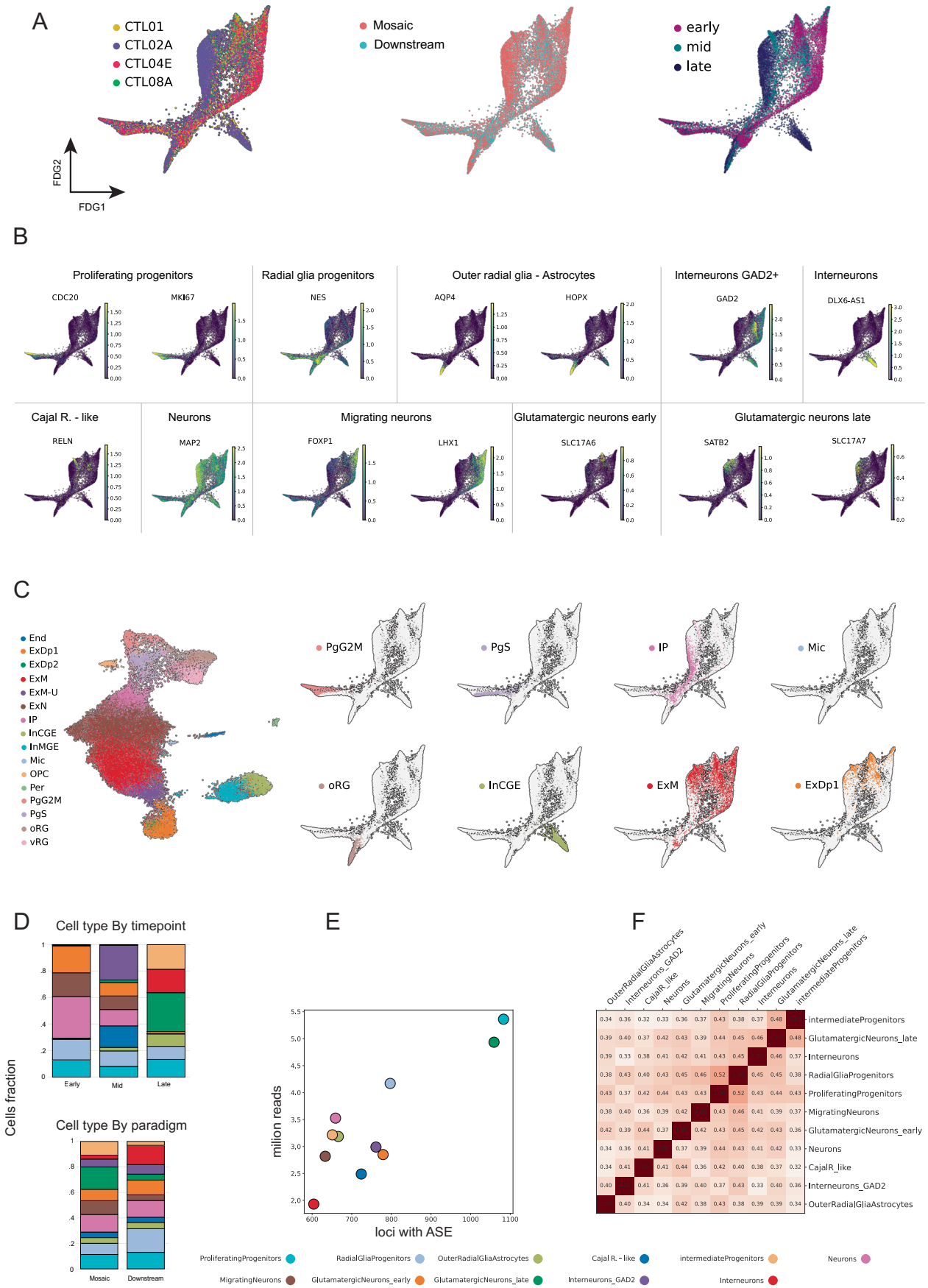
https://doi.org/10.1038/s41592-024-02555-5





Extended Data Fig. 4 | SCanSNP benchmarking. a) Boxplots of precision and recall scores for evaluation of the classification of Demuxlet v2, Souporcell, Vireo and SCanSNP against barcode-based identity demultiplexing performed by cellranger multi. Displayed by each box is the median (horizontal solid line within the box), interquartile range (upper and lower bound of the boxes), min and max values (extension of the whiskers) for two independent datasets. On the right, droplets classified as low quality and doublets by either algorithm were included, on the left they were not taken into account in the comparison. Change in the y-axis scale reflects the higher performance of all algorithms against the ground truth when only singlets and good quality barcodes are considered. **b**) Natural logarithm of the total counts +1 in each sample for two independent datasets. Displayed by each box is the median (horizontal solid line within the box), quartiles' range (upper and lower bound of the boxes), min and max values (extension of the whiskers). Outliers are computed as a function of the inter-quartile range and shown as points outside the minimum and maximum range. **c**) Barplot showing differences in doublets and unassigned droplets rates by algorithm.

https://doi.org/10.1038/s41592-024-02555-5

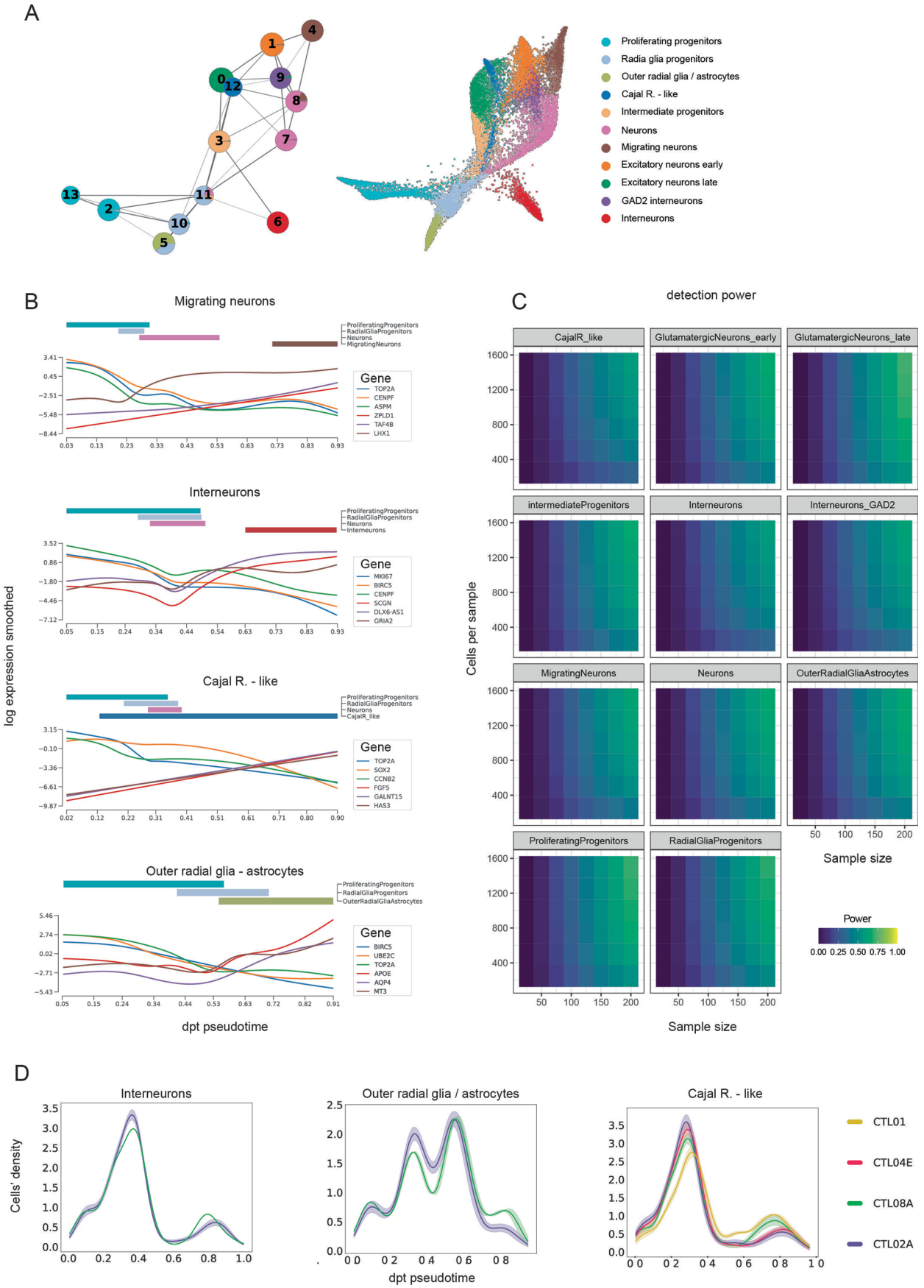


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41592-024-02555-5

Extended Data Fig. 5 | Single cell datasets characterization. a) Embedding of cells from all datasets of force-directed graph. From left to right cells are coloured by genotype, multiplexing paradigm and stage. **b**) Force-directed graph coloured by expression of relevant markers. Plotted markers are divided by the cell type they are most relevant for. **c**) Force-directed graph coloured by transferred label from *Poliudakis et al*. dataset. End: endothelial; ExDp1: excitatory deep-layer1; ExDp2: excitatory deep-layer2; vRG: ventral radial microglia; oRG: outer radial glia; ExN: newborn excitatory; ExM: maturing excitatory; ExM-U: excitatory upper-layer-enriched; IP: intermediate progenitors; inCGE: interneurons caudal ganglionic eminence; inMGE: interneurons medial ganglionic eminence; Mic: microglia; OPC: oligodendrocyte precursors; Per: pericytes; PgG2M: G2M phase proliferating progenitors; PgGS: S phase proliferating progenitors; **d**) Plot of fraction of cells for each cell type, divided by timepoint (upper panel), and by multiplexing paradigm (lower panel). **e**) The scatterplot shows the number of loci with detected allele specific expression on the x axis and the total number of reads expressed in millions on the y axis; each dot represents a cell type. **f**) Spearman correlation on reads bringing alternative alleles / total reads (bValues) among the observed cell types. Correlation is calculated on loci that displayed allelic imbalance (binomial test fdr < 0.05) in at least one cell type and with at least 20 reads in each cell type.

https://doi.org/10.1038/s41592-024-02555-5

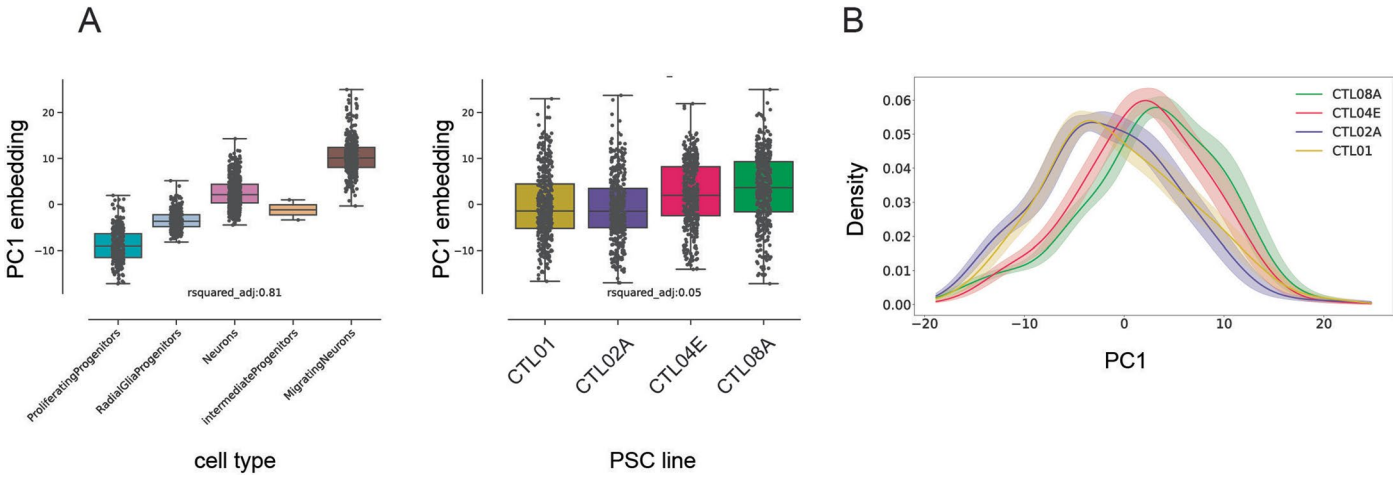


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41592-024-02555-5

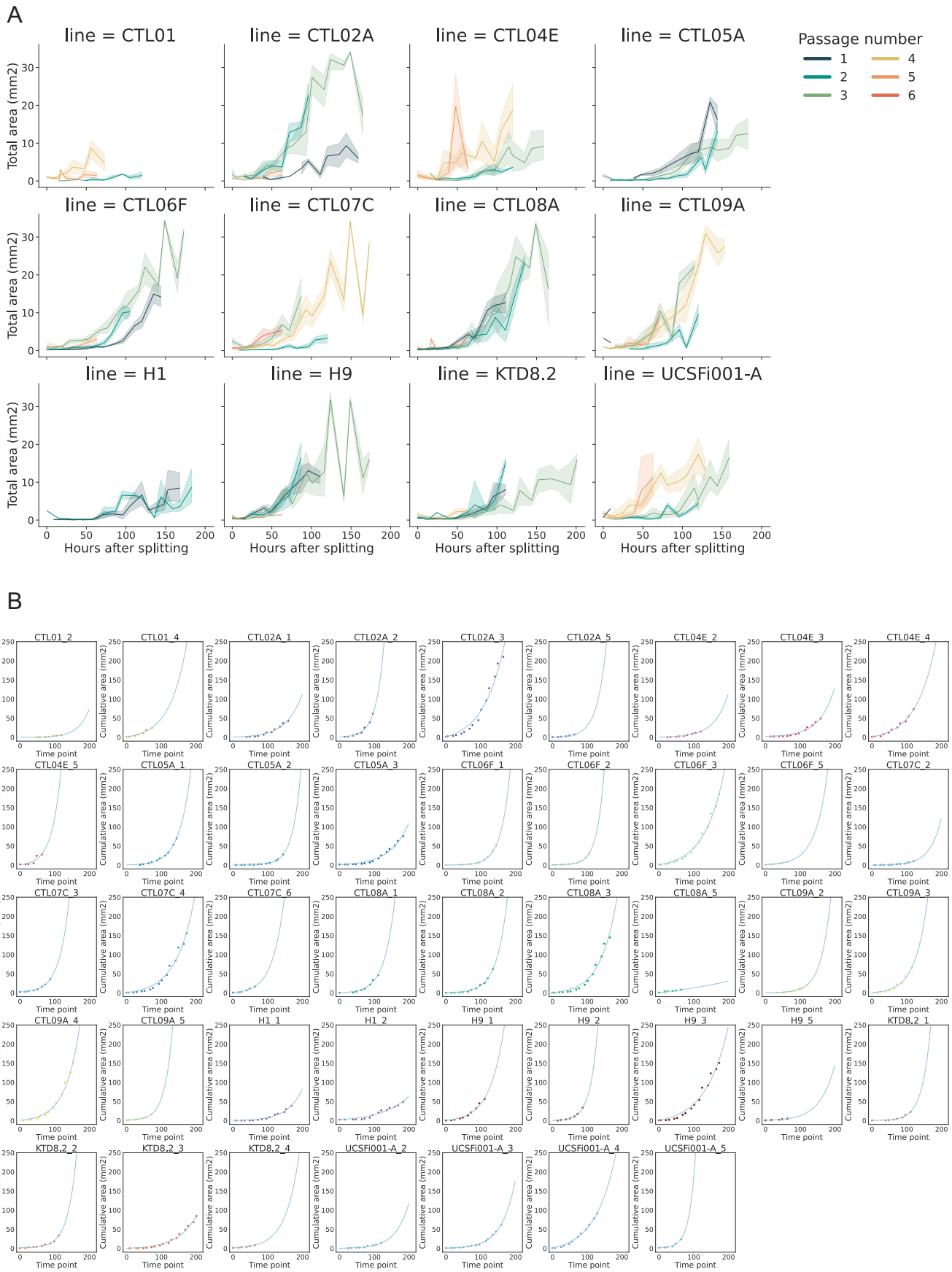
Extended Data Fig. 6 | Developmental trajectory analysis and power analysis. **a**) On the left: Partition-based graph abstraction (PAGA) plot. Each circle represents a Paga cluster, circles are partitioned according to the fraction of cells per annotated cell type (shown as reference on the right side), weighted edges among PAGA clusters encode their transcriptional similarity. **b**) Plot of smoothed gene expression - obtained via tradeseq - along pseudotime (Methods). For each lineage the three most relevant decreasing and increasing genes (sorted by pVal and absolute logFC) are shown. Above each expression panel, bars coloured by cell type indicate the occupancy of each cell type along pseudotime. **c**) SCPOWER: Estimation of single cell eQTL power per specific cell type, depending on the sample size and numbers of cells per sample. **d**) Distribution of genotypes along pseudotime for Interneurons, Outer radial glia / astrocytes and Cajal Retzius-like lineages. Within each differentiation stage cells were balanced to the same amount across genotypes for correct comparison. If too few/no cells were retrieved at any differentiation stage, the whole genotype was removed from the comparison. Faded colour shows 1 standard deviation across random subsampling iterations, solid line display the mean value across random subsampling iterations.

https://doi.org/10.1038/s41592-024-02555-5



Extended Data Fig. 7 | Migrating neurons PC1 analysis. a) Boxplot of cells' embedding on PC1 for migrating neurons trajectory, grouped by cell type and PSC line. Above the x axis residuals values are reported for each covariate. Each dot is the embedding of a cell in PC1, boxplot display median, interquartile range, minimum and maximum values among cells of each group. **b**) Distribution of different genotypes along PC1 after genotypes balancing per timepoint. Solid line represents the mean, and faded colour shows 1 standard deviation value, upon 50 random subsampling iterations.

https://doi.org/10.1038/s41592-024-02555-5

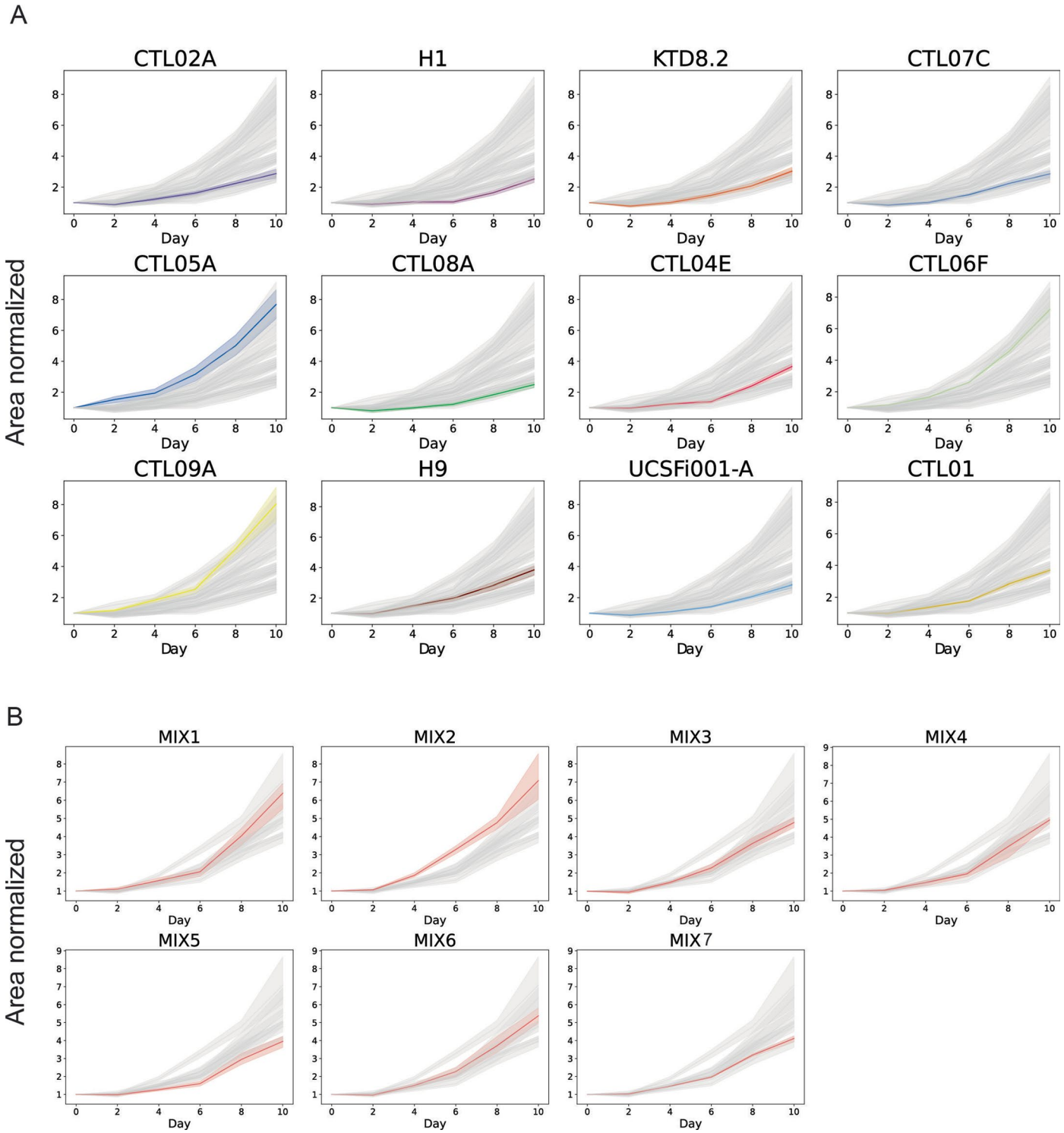


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41592-024-02555-5

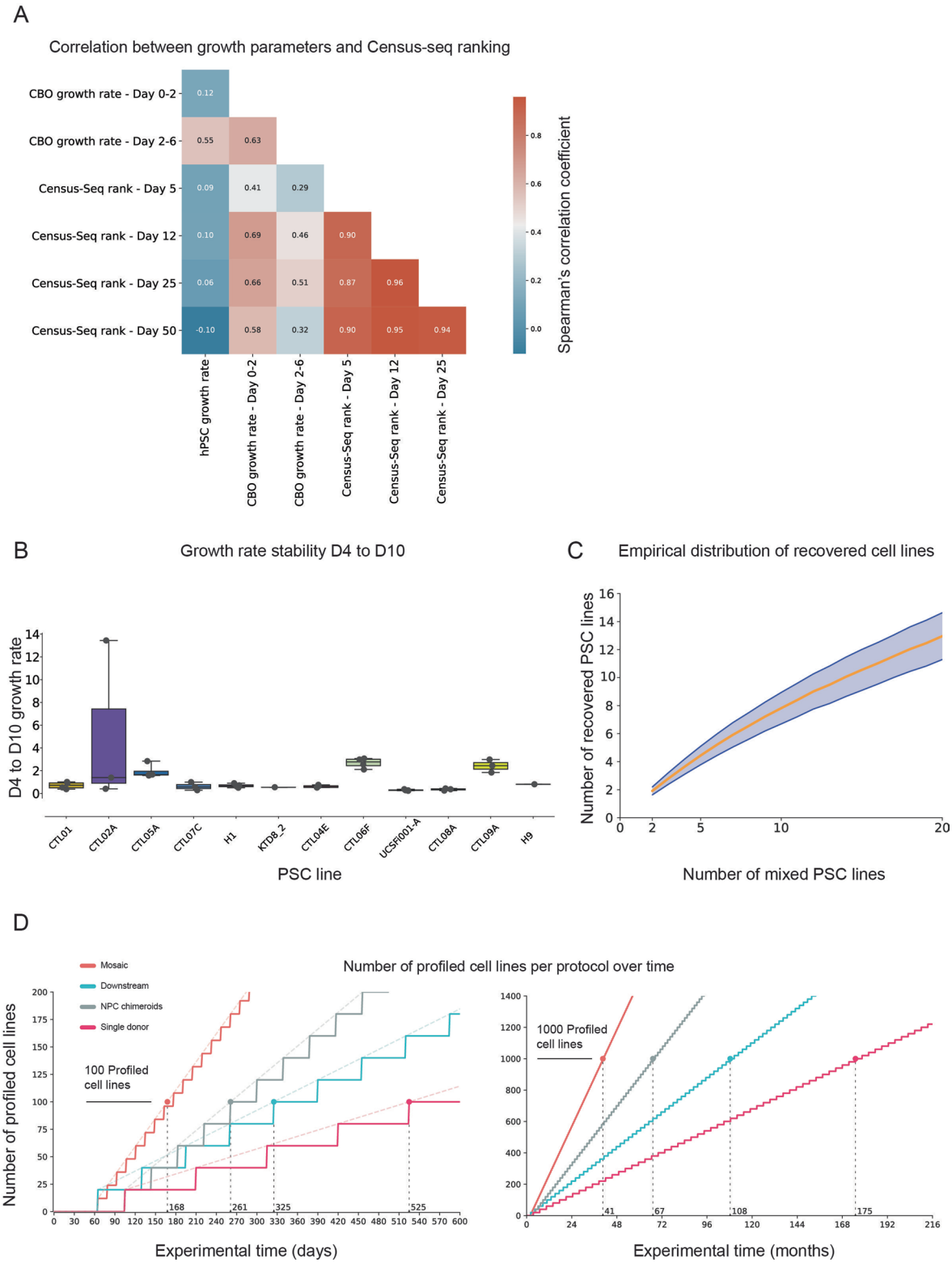
Extended Data Fig. 8 | PSC growth curves. a) Growth curves of each line coloured according to the number of passage (that is split) post-thawing. On the x axis the time in hours after splitting the cells from one plate to a new one (see Methods), the y axis the total area detected in mm^2. The line depicts the mean area at that time point across the field of views, the shade shows the 95% confidence interval. **b**) Cumulative mean area of each PSC line at each different passage fitted as an exponential curve, as depicted by the solid line. The dots represent the empirical values.

https://doi.org/10.1038/s41592-024-02555-5



Extended Data Fig. 9 | CBO growth curves. a, **b**) Growth curves of pure-line CBO (**a**) or mosaic CBO (**b**). The area detected at each time point was normalized on the area of the organoid at day 0. The lines depict the mean area across five independent replicates for all the PSC lines at all time points except CTL09A and MIX4 at day 10, when only 4 replicates were available. The shade around the mean is representative of the 95% confidence interval. The coloured line is representative of the subplot title PSC line while all the others are shown in light grey.

https://doi.org/10.1038/s41592-024-02555-5



Extended Data Fig. 10 | See next page for caption.

https://doi.org/10.1038/s41592-024-02555-5

Extended Data Fig. 10 | PSC growth dynamics in mCBO. a) Heatmap coloured by Spearman's correlation coefficients computed between the rate of cumulative growth (hPSC growth rate), slopes of the linear fitting (CBO growth rate) and Census-Seq weighted ranking normalized at each available time point. The text on the heatmap shows the Spearman's correlation coefficient **b**) Growth rate of different cell lines for different mosaic mixtures (different dots) in the interval day 4 to day 10. Different cell lines are divided along x axis, each dot of a box represent the growth rate of the same cell line in a different experimental mixture. Displayed by each box is the median (horizontal solid line within the box), interquartile range (upper and lower bound of the boxes), minimum and maximum values (extension of the whiskers) among mixtures per line. **c**) Monte Carlo simulation of cell lines recovery. The plot shows the number of theoretically mixed (x axis) and recovered cell lines (y axis). The yellow line indicates the mean number of recovered cells and the faded blue indicates 1 standard deviation upon 100'000 simulations for each value of x. **d**) Estimation of the experimental workload and time required for large-scale experiments (see "Scalability of mosaic experiments in terms of experimental timeline") also for the "chimeroid" approach. In this case NPC-chimeroids are dissociated and reaggregated after 25 days of differentiation, thus the same considerations of the downstream multiplexing design applies until that timepoint. The plot shows the number of profiled cell lines (y axis) and the experimental days (x axis). Vertical dashed lines represent the experimental time to reach 100 and 1000 profiled cell lines in left plot and right plot respectively. For the left plot, the approximation strict line is displayed for each protocol.

nature portfolio

Corresponding author(s): Giuseppe Testa

Last updated by author(s): Oct 2, 2024

Reporting Summary

Nature Portfolio 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 Portfolio 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	Code used for the analysis is available at https://github.com/GiuseppeTestaLab/organoidMultiplexing_release Latest SCanSNP release is available at https://github.com/GiuseppeTestaLab/SCanSNP SC RNA-seq data were aligned using cellranger3.0.0 count and the matched reference provided from 10x, with the exception of the Cellplex experiment data that were aligned with cellranger count and cellranger multi both v6.1.2 and the same matched reference. Bulk RNA-seq was aligned using STAR v2.6.1 and variant call was subsequently performed via GATK version 4.2.1 and 4.0.1 (Base Recalibration only) Nextflow-sarek 3.1.1 and relative dependencies: https://github.com/nf-core/sarek/blob/master/CITATIONS.md Deepvariant 1.4.0 Census-seq 2.5.4 and relative dependencies https://github.com/broadinstitute/Drop-seq/blob/master/doc/Census-seq_Computational_Protcools.pdf Python 3.8.5 Python libraries: scikit-learn 1.3.2, scanpy 1.9.3, numpy 1.20.3, matplotlib 3.6.0, leidenalg 0.8.3, harmonypy 0.0.5, anndata 0.8.0, anndata2ri 1.0.6 R 4.0.5 R libraries: Milo 1.4.0, tradeSeq 1.4.0, scRNAseq 2.4.0, leiden 0.3.6, IRanges 2.24.1, GenomicRanges 1.42.0, GenomicFeatures 1.42.3, edgeR 3.32.1 , DropletUtils 1.10.3, DESeq2 1.30.1, org.Hs.eg.db 3.12.0

All additional Python and R packages used and their versions are specified in the github repository

Additional Datasets/Databases/Resources
Reference genome: GRCh38 v93 Ensembl reference genome
dbSNP: v.b151_GRCh38p7 (00-All.vcf)
Jerber et. al 2021: sc-eQTLs Table 7

Some of the images were adapted from Biorender templates (Fully licensed - Human Technopole). Figures were assembled in Adobe Illustrator v.27.0.1

Singularity 3.8.7 was used to create environments

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Full code used for the analyses can be retrieved at https://github.com/GiuseppeTestaLab/organoidMultiplexing_release. SCanSNP latest release and the docker link are available at https://github.com/GiuseppeTestaLab/SCanSNP.
All omics data are being deposited into approved public repositories upon acceptance

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the 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 determined according to the number of control PSC lines available for which sufficient characterization had been previously carried out, and to meet the requirements of PSC and brain organoids based disease modeling. Germain, P.-L. & Testa, G. Taming Human Genetic Variability: Transcriptomic Meta-Analysis Guides the Experimental Design and Interpretation of iPSC-Based Disease Modeling. Stem Cell Reports 8, 1784–1796 (2017).
Data exclusions	Statistical analyses were implemented according to the best practices of each specific technique. More in detail, for omics approaches, data were subjected to bioinformatics pipelines for the pre-processing steps (e.g. sequencing alignment for genomics and transcriptomics) and a quality control phase so to detect potential technical co-variables or sub-optimal samples. Standardized computational pipelines, including filtering, normalization, batch correction, dimensionality reduction, clustering, differential expression analysis and other downstream analysis were applied according to state of the art algorithms for each specific technique. Cells from non-control genotypes were excluded after genetic demultiplexing, since they were included in the multiplexed 10x library

preparation to lower the costs but their analyses are within the scope of different projects.

Replication: Both mosaic- and pure lines-organoids were differentiated and analyzed in multiple replicates, as specified in figure legends.

Randomization: N/A all the cell lines employed in the study were both multiplexed into mosaic CBOs and profiled through downstream multiplexing in parallel as for the goal of the current study. Setup of mosaic mixtures for prediction of genotypes growth in mosaic models was based on the similarity of the growth curves of the single PSC lines.

Blinding: Experimenters were aware of the experimental conditions (mosaics or downstream multiplexed CBOs) during their differentiation and the following analysis steps

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
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Antibodies

Antibodies used:

Pax6 (Rabbit): Biolegend Cat. No. #901301, Clone Poly19013, Lot No. #B354380, used 1:200;
Tuj1 or TUBB3 (Mouse): Biolegend Cat. No. #801202, Clone TUJ1, Lot No. #B354042 , used 1:1000;
Nestin (Mouse): Millipore Cat. No. #MAB5326, Clone 10C2, Lot No. ND , used 1:500;
Reelin (Mouse): Millipore Cat. No. #MAB5364, Clone G10, Lot No. #3927128 , used 1:400;
CTIP2 (Rat): Abcam Cat. No. #ab18465, Clone 25B6, Lot No. #GR3427932-1, used 1:400;
SATB2 (Mouse): Abcam Cat. No. #ab51502, Clone SATBA4B10, Lot No. #GR3451030-1, used 1:400;
HOPX (Rabbit): Sigma-Aldrich Cat. No. #HPA030180, Polyclonal, Lot No. #000042633, used 1:500;
SOX2 (Goat): R&D Systems Cat. No. #AF2018, Polyclonal, Lot No. #KDY0521031, used 1:1000
MAP2 (Guinea Pig): Synaptic Systems Cat. No. #188004, Polyclonal, Lot. No. ND, used 1:200;
NeuN or Neuronal Nuclei (Mouse): Abcam, Cat. No. #ab104224, Monoclonal, Clone No 1B7, Lot. No. GR3408621-1, used 1:200;
mKI67 (Rabbit): Abcam .Cat. No. #ab15580, Polyclonal, Lot. No. ND, used 1:250

AlexaFluor Plus 488-conjugated donkey anti-mouse: ThermoFisher Cat. No. #A32766 , Lot No. ND, used 1:400
AlexaFluor Plus 555-conjugated donkey anti-rat: ThermoFisher Cat. No. #A48270, Lot No. ND, used 1:400
AlexaFluor Plus 555-conjugated donkey anti-mouse: ThermoFisher Cat. No. #A32773, Lot No. ND, used 1:400
AlexaFluor Plus 555-conjugated donkey anti-goat: ThermoFisher Cat. No. #A32816, Lot No. ND, used 1:400
AlexaFluor Plus 647-conjugated donkey anti-rabbit: ThermoFisher Cat. No. #A32795, Lot No. ND, used 1:400

Validation:

The antibodies used in this study have been selected from published literature and their use has been optimized in house. Human reactivity for all of the antibodies used in the present study has been validated by the manufacturer and the antibodies were validated in house on human cortical brain organoids. Proof of validation from the manufacturer includes:
Pax6 Biolegend Poly19013: Validated on human cell extract in WB (1:2000); validated on mouse FFPE sections through IHC (1:100)
Tuj1 or TUBB3 Biolegend TUJ1: Validated on human FFPE sections through IHC (1:1000)
Nestin Millipore 10C2: Validated in human for ICC, IHC and IF on FFPE sections (1:200)
Reelin Millipore G10: validated in WB on rat brain lysates, was used in IHC on human samples
CTIP2 Abcam 25B6: validated for IHC-P, Flow cytometry, WB and IF in human samples (1:500 in IF)
SATB2 Abcam SATBA4B10: Validated for ICC, IP and WB in human samples
HOPX Sigma HPA030180: Validated by the Human Protein Atlas project on an IHC tissue array of 44 nurmal human tissues.
SOX2 R&D AF2018: Validated for WB, CHIP, ICC, IHC in human samples
MAP2 Synaptic Systems 188004: validated for WB, IP, ICC and IHC on human samples (1:500 in IHC)
NeuN Abcam ab1o4224: Validated in IHC-P on human samples
mKI67 Abcam ab15580:: validated in IHC-P, ICC and IF on human samples

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s):

WTSIi018-B-1 (in house name: CTL08A), hiPSC line, male, derived from healthy donor fibroblasts. Purchased from Wellcome Trust Institute
UMILi026-A (in house name: CTL01), hiPSC line, female, derived from healthy donor fibroblasts. Reprogrammed in house

	UOSi001-A (in house name: CTL02A), hiPSC line, male, derived from healthy donor fibroblasts. Reprogrammed in house UMILi024-A (in house name: CTL04E), hiPSC line, female, derived from healthy donor fibroblasts. Reprogrammed in house UMILi025-A (in house name: CTL09A), hiPSC line, male, derived from healthy donor fibroblasts. Reprogrammed in house UMILi006-A (in house name: CTL05C), hiPSC line, male, derived from healthy donor fibroblasts. Reprogrammed in house UMILi007-A (in house name: CTL06F), hiPSC line, female, derived from healthy donor fibroblasts. Reprogrammed in house UMILi008-A (in house name: CTL07C), hiPSC line, female, derived from healthy donor fibroblasts. Reprogrammed in house UMILi012-A (in house name: WVS02A), hiPSC line, male, derived from Weaver syndrome donor fibroblasts. Reprogrammed in house UMILi013-A (in house name: WVS01H), hiPSC line, female, derived from Weaver syndrome donor fibroblasts. Reprogrammed in house UMILi014-A (in house name: WVS04A), hiPSC line, male, derived from Weaver syndrome donor fibroblasts. Reprogrammed in house UMILi015-A (in house name: WVS03B), hiPSC line, female, derived from Weaver syndrome donor fibroblasts. Reprogrammed in house UCSFi001-A hiPSC line, male, derived from healthy donor PBMCs. Purchased from Coriell Institute for Medical Research WAe001-A (in house name: H1) ESC line, male. No disease diagnosed. Purchased from WiCell WAe009-A (in house name: H9) ESC line, female. No disease diagnosed. Purchased from WiCell KTD8.2 is a WTSIi018-B-1 isogenic jiPSC line carrying EZH2 point mutation of UMIL1012-A. All cell lines except KTD8.2 have been registered at the Human Pluripotent Stem Cell Registry (https://hpscreg.eu/)
Authentication	All the cell lines are routinely checked through STR analysis (GenePrint 10 System, Promega).
Mycoplasma contamination	We confirm that all cell lines are maintained mycoplasma free. Routine screening for mycoplasma contamination is performed monthly on all the cell lines and their derivatives, including organoids.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell line has been used in this study.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	In replicate 1, organoids were dissociated, cells were filtered with 70um cell strainers to remove clumps and fixed in ice-cold Ethanol. The day prior to the analysis, Cells were incubated overnight with 3 uM Propidium Iodide in the presence of 25ug/ml RNAse I and subsequently analyzed by flow cytometer. In replicate 2, organoids were dissociated and cells filtered with 70um cell strainers to remove clumps. Cell suspensions were incubated for 30' at 37 C with Hoechst33342 2ug/ml and 5uM Verapamil to minimize Hoechst efflux through pGP transporter. ToPro3 dye was added to the suspension right before flow cytometer analysis to identify and exclude dead cells.
Instrument	Raplicate 1 has been analyzed with FACS Celesta (BD Biosciences); Replicate 2 has been analyzed with Cytoflex LX flow cytometer (Beckman Coulter)
Software	Replicate 1 has been analyzed with FlowJo software (BD Biosciences) ; Replicate 2 data have been analyzed using FCS Express 7 software (DeNovo software)
Cell population abundance	In replicate 1, all the cells from each sample were profiled. The specifics for each sample can be found in Supplementary table 1. In replicate 2, 15000 events of interest per sample were recorded as shown in Supplementary fig. 8B.
Gating strategy	In replicate 1 we profile ethanol-fixed cells (all events) and therefore we subselected singlets from the entire population according to their Propidium Iodide (PI)- height vs area signal. Using Propidium Iodide signal intensity, we discarded cellular fragments (SubG1 population, fig. S7A) before proceeding with cell cycle analysis using the analytical software-embedded algorithm. In replicate 2, we first selected cells (GATE: cells) from the general population (all events) according to their FSC-A/ SSC-A ratio, then we applied a Live Cells gate according to ToPro3 Live/dead stain intensity, keeping the ToPro3 negative cells. Of those, we selected singlets according to the HOECST area vs height signal. 15000 events were acquired on this latter hierarchical gate.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.