Cluster similarity spectrum integration of single-cell genomics data

Zhisong He[1], Agnieska Brazovskaja[2], Sebastian Ebert[2], J. Gray Camp[3,4,$], Barbara Treutlein[1,2,$]

Affiliations:
[1] Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
[2] Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[3] Institute of Molecular and Clinical Ophthalmology, Basel, Switzerland
[4] University of Basel, Basel, Switzerland.

[$] Corresponding authors

Correspondence: J. Gray Camp (grayson.camp@iob.ch), Barbara Treutlein (barbara.treutlein@bsse.ethz.ch)

Abstract

Technologies to sequence the transcriptome, genome or epigenome from thousands of single cells in an experiment provide extraordinary resolution into the molecular states present within a complex biological system at any given moment. However, it is a major challenge to integrate single-cell sequencing data across experiments, conditions, batches, timepoints and other technical considerations. New computational methods are required that can simultaneously preserve biological signals, while also integrating samples. Here, we propose an unsupervised reference-free data representation, Cluster Similarity Spectrum (CSS), where each cell is represented by its similarities to clusters independently identified across samples. We show that CSS can be used to assess cell heterogeneity and enable differentiation trajectory reconstruction from cerebral organoid single-cell transcriptome data, and to integrate data across individuals and experimental conditions. We compare CSS to other integration algorithms and show that CSS performs comparably well. We also show that CSS allows projection of single-cell genomic data of different modalities to the CSS-represented reference atlas for visualization and cell type identity prediction. We think CSS provides a straightforward and powerful approach to understand and integrate challenging single-cell multi-omic data.

Introduction

Recent advances in molecular, engineering, and sequencing technologies have enabled the high-throughput measurement of transcriptomes and other genomic features in thousands of cells in an experiment. Single-cell RNA sequencing (scRNA-seq) greatly enhances our capacity to resolve the heterogeneity of cell types and cell states in biological samples, as well as to understand how systems change during dynamic processes such as development. However, current scRNA-seq technology only provides a molecular snapshot on limited measured samples at one time. Joint analysis on many samples across multiple

experiments and different conditions is often required. In such a scenario, the biological variation of interest is usually confounded by other factors, including sample sources and experimental batches. This is particularly challenging for developing systems, where the data contains distinct, mature cell types as well as intermediate cell states at different points along various differentiation trajectories. Several computational integration methods, including but not limited to Reference Similarity Spectrum (RSS)[1,2], Harmony[3], Seurat[4], LIGER[5], Scanorama[6] and MNN[7], have been developed to address some of these issues. Among them, RSS represents each cell by its transcriptome similarities to a series of reference samples. Harmony uses iterative clustering-correction procedure based on soft clustering to correct for sample differences. Seurat introduces an anchor strategy, with anchors between samples defined by canonical correlation analysis, to correct batch effect. LIGER adapts integrative non-negative matrix factorization to identifying shared and dataset-specific factors for joint analysis. MNN identifies mutual nearest neighbors between two data sets and derives cell-specific batch-correction vectors for integration. Scanorama generalizes mutual nearest-neighbors matching in two data sets to identify similar elements in multiple data sets in order to support integration of more than two data sets. Benchmarking on these integration methods have revealed that each method presents different performance in various scenarios, and there is no single magic bullet capable of always dissecting out meaningful variations of interest[8].

Here, we propose an unsupervised scRNA-seq data representation namely Cluster Similarity Spectrum or CSS. Instead of using external references as in RSS, CSS considers every cell cluster in each sample for integration as an intrinsic reference, and represents each cell by its transcriptome similarities to clusters across samples. The underlying hypothesis of both RSS and CSS is that the undesired confounding factors, e.g. read coverage of different cells, introduce random perturbation to the observed transcriptomic measures which are not correlated with cell type or cell state identities. Cells with the same identity therefore share similar spectrums, after standardizing across similarities to different references, i.e. similarity spectrum to normalize the global differences across cells introduced by the random perturbation. Additionally, as CSS considers all clusters in different samples as references, standardization is done separately across clusters of different samples, so that global differences across samples can be largely eliminated. We apply CSS to various scenarios focusing on data generated from cerebral organoids derived from human induced pluripotent stem cells (iPSCs). We use CSS to integrate data from different iPSC lines, human individuals, batches, modalities, and conditions. We show that technical variation caused by experimental conditions or protocols can be largely reduced with the CSS representation, and CSS has similar or even better performance compared to other integration methods including Harmony, Seurat v3 and LIGER, which were highlighted in the previous benchmarking[8]. We also show that CSS also allows projection of new data, either scRNA-seq or scATAC-seq, to the CSS-represented scRNA-seq reference atlas for visualization and cell type identity prediction. The CSS codes are available at https://github.com/quadbiolab/simspec.

Results

CSS integrates scRNA-seq data from different organoids, batches, human individuals

To calculate the CSS representation, clustering is first performed on transcriptome data from cells in each sample separately, and average expression profiles are calculated for each cluster (Fig. 1, Supplementary Fig. 1). Transcriptome similarity, represented as the Spearman correlation between gene expression profiles, is then calculated between each cell and each cell cluster. For each cell, the calculated similarities are normalized across clusters of each sample by z-transform and concatenated, resulting in its CSS representation. We applied CSS and other integration approaches to a complex cerebral organoid scRNA-seq dataset[1], where the data was affected by technical variation due to organoid, batch, and iPSC line. Altogether the dataset contained scRNA-seq data from twenty-two-month-old human cerebral organoids, each with a different cell type composition, from seven different ESC/iPSC lines in four batches of in total eleven experiments (Fig. 2a). The UMAP embedding without any integration method reveals variation among cells due to cell type (e.g. brain region), batch, and individual. Based on marker gene expression there were clear signatures of cell heterogeneity that were observed across samples, such as differentiation trajectories from neural progenitor cells (NPC) to neurons, as well as specification of neurons from different brain regions including cortical excitatory neurons (ENs) and ganglionic eminence (GE)-derived inhibitory neurons (INs). In addition, different organoids, especially organoids from different technical batches, also separated cells into distinct clusters (Fig. 2b). Using RSS to fetal BrainSpan RNA-seq data largely integrated data from different experimental batches, allowing interpretable cell type annotations (Supplementary Fig. 2, Fig. 2c). In this case, there are neuronal differentiation trajectories from multiple brain regions including cortex, GE, and non-telencephalic regions. This RSS-based cell type annotation, as provided in the original study[1], was considered as the reference annotation for a comparison of CSS and other integration methods.

Next we applied CSS, Harmony, Seurat v3, and LIGER to this cerebral organoid dataset to compare integration approaches. The UMAP embeddings of all four integration methods show that each method largely improves cell mixing from different organoids and batches (Fig. 2d). For each cortical EN and MGE/CGE IN, we calculated the proportions of neighboring cells being of the same cell type but from different batches. Here, neighboring cells were defined as cells with the smallest Euclidean distances to the cell in the integrated space. All the four integration methods substantially increased such proportions, indicating that they all performed well in cell mixing (Fig. 2e). However, Seurat v3 and LIGER suffered from a severe over-correction problem, where cells from different stages of differentiation or brain regions were mixed together (Fig. 2d-e). On the other hand, cell type heterogeneity remained when the data was integrated with CSS and Harmony, giving results comparable if not better than RSS (Fig. 2d-e). Relative to Harmony, CSS resulted in an UMAP embedding with larger distinction of brain region-specific differentiation trajectories (Fig. 2d), especially the three major cell fates (dorsal telencephalic, ventral telencephalic, and non-telencephalic cells), and less likelihood mixing of different cell types (Fig. 2e).

CSS integrates time course data

We next wanted to apply CSS to a time course scRNA-seq dataset of cerebral organoid development from pluripotency (Fig. 3a). Current scRNA-seq technology provides a snapshot rather than a longitudinal measurement of sample heterogeneity. To study dynamic processes, measuring multiple different samples, each of which represents a different time point, is often mandatory. In such a scenario, temporal variation is unavoidably confounded by differences among individual samples and experimental batches, and it becomes difficult to disentangle the biological and technical portions of variation from each other. Intuitively, meaningful variation is more likely to change transcriptome similarity of a cell to different cell types or cell states compared to technical variation. Under this assumption, CSS has the potential to minimize the influence of technical variation in the data.

To assess whether CSS can integrate data that segregates due to random variation, we applied CSS to the scRNA-seq time course data set from iPSC through four months of human cerebral organoid development[1] (Fig. 3a). This dataset includes seventeen samples from two PSC lines, covering seven different developmental stages. In comparison, Harmony, Seurat v3 and LIGER were also applied to the same data set. The UMAP embedding shows that when no integration was applied, cells of samples at different time points, particularly PSC, embryoid body (EB), neuroectoderm (NEcto) and neuroepithelium (NEpith) stages, separate from each other, forming distinct cell groups (Fig. 3b). Differences between the two PSC lines are also substantial.

Strikingly, we find that CSS largely integrated cells of the early time points including from PSC to NEpith stages, and cell of different lines, without disturbing later time points where different neuronal types have emerged (Fig. 3c). In comparison, Seurat v3 and LIGER again encountered severe over-correction problems and largely mixed PSCs with NPCs in cerebral organoids (Fig. 3d-f). Harmony preserved both temporal and cell type heterogeneity in the data set, although the corresponding UMAP embedding lacked qualitative clarity (Fig. 3d).

To quantitatively compare different integration results, we focused on the PSC/EB time points, which are known to be distinct from cell types in the later time points[9,10]. First of all, all the four integration methods intermixed cells from different PSC lines. However, substantial proportions of non-PSC/EB cells became neighboring cells of the PSC/EB population when Seurat v3 and LIGER were used (Fig. 3g), implying over-correction problem. We also calculated average distances between each PSC/EB cell and cells at different time points (Fig. 3g). The distances to PSC increased along the developmental time course when three of the four tested integration methods with LIGER being the only exception. In general, CSS and Harmony substantially outperformed Seurat v3 and LIGER. Compared to Harmony, CSS showed weaker power in mixing cells from different PSC lines, but better performance in discriminating cells at the later time points from the PSC/EB ones (Fig. 3g).

Interestingly, although CSS in general enhanced the differences between PSC and other cell types, likely by reducing intra-group diversity while retaining inter-group divergence, a continuous linkage was observed from PSC/EB to NEcto cells (Fig. 3c) in the UMAP

176 embedding. Cells between the two cell groups may therefore represent the transition state of
177 neural induction. RNA velocity analysis with scVelo also supported the transition potential
178 from PSC to NEcto cells (Fig. 3h). We screened for cells at PSC/EB or NEcto time points
179 with at least 15 neighbors being both PSC/EB and NEcto cells (Fig. 3h). This resulted in 295
180 cells in total. Comparing those transition cells to both PSC/EB cells and NEcto cells resulted
181 in differentially expressed genes such as CYP26A1 and LITAF (Fig. 3h). Among them,
182 CYP26A1 encodes for a retinoic acid-metabolizing enzyme and has been shown to be
183 essential for body patterning and brain development[11-13].
184

185 *CSS integrates data across technical conditions*

186 To determine if CSS can integrate data across technical conditions, we generated scRNA-
187 seq data from fresh and methanol-fixed cerebral organoid single-cell suspensions. Methanol
188 fixation was applied to half of the dissociated cells prior to the scRNA-seq experiment.
189 ScRNA-seq experiments were done on the fixed and unfixed cells separately. The data
190 suggested moderate difference between the two experimental conditions (Supplementary
191 Fig. 3a-b). RSS based on fetal BrainSpan RNA-seq data integrated cells of the two
192 conditions, and cells were then clustered and annotated based on their RSS representation
193 (Supplementary Fig. 3c).
194
195 We next applied the four data integration methods (CSS, Harmony, Seurat v3 and LIGER)
196 and estimated performances of the methods by comparing each cell with its neighbors. The
197 results suggest that all the four methods significantly increased the proportion of neighbors
198 being of the same cell type but different experimental conditions (Supplementary Fig. 3d).
199 False-integration, quantified by the proportion of neighboring cells of different cell types or
200 branch identities, remained low (Supplementary Fig. 3d). This result indicates that in this
201 common scenario, CSS shows comparable performance as the other integration methods.
202

203 *CSS integrates scRNA-seq data generated by different technologies*

204 We sought to access how CSS may perform when scRNA-seq data of different samples are
205 generated by different technologies. We compared scRNA-seq data generated by 10x
206 Genomics (one cerebral organoid, 4512 cells), Fluidigm C1 and Smart-Seq2 (C1/SS2)
207 technology (685 cells)[1], and inDrop technology (5847 cells, generated in this study). Analysis
208 on the three data sets separately indicates their cell composition, allowing annotation of cells
209 into nine different groups, including NPC and neurons of dorsal forebrain, ventral forebrain
210 and non-telencephalic regions, choroid plexus, mesenchymal-like cells and epithelial-like
211 cells (Fig. 4a-b). Joint analysis with no integration showed that cells primarily separated by
212 data sets by different technologies.
213
214 Next we integrated the three data sets using RSS (to fetal BrainSpan RNA-seq data), CSS,
215 Harmony, Seurat v3 or LIGER. All the methods, except RSS, largely improve cell mixing
216 from different data sets (Fig. 4c-d), although over-integration appeared when LIGER was
217 applied. Focusing on the C1/SS2 data set which only represents 6.2% cells in the joint data
218 set, different methods show different pros and cons. Harmony, Seurat v3 and LIGER
219 performed comparably well in terms of mixing cells from different data sets and better than

CSS. On the other hand, CSS performed the best balancing cell mixing improvement and retaining cell type heterogeneity (Fig. 4d).

CSS allows query data projection to the reference scRNA-seq atlas

As most of the integration methods require data-specific transformation of the expression matrices, when new data is introduced, the integration procedure needs to be rerun. Since CSS representation relies on the normalized expression levels without additional data transformation, it allows projection of new query data to the CSS-represented reference atlas. We tested the feasibility with the two-month-old cerebral organoid scRNA-seq data. In brief, we first extracted cells of two organoids from the Sc102a1 iPSC line as the query data. Next, we built a CSS-integrated reference atlas using the remaining scRNA-seq data (Fig. 5a). The corresponding CSS representations of the query cells in the Sc102a1 organoids were then calculated. The projected UMAP embedding of the query cells, as well as their projected cell type labels, were obtained with the intrinsic UMAP projection mechanism and a k-nearest-neighbor (kNN, k=50, defined in the CSS-space) classifier, respectively. The projected UMAP embedding matches with the fact that the two Sc102a1 cerebral organoids mostly consist of cortical cells (Fig. 5b). More importantly, using the kNN classier to predict cell type of query cells based on the reference cell annotation resulted in 89.4% of the cells predicted to be the exact same cell type as they were annotated. Among the rest, another 8.9% query cells were predicted to be a different cell type as it was annotated, but at the same cell fate branch, i.e. one of cortical cells, GE cells, and non-telencephalic cells (Fig. 5b). This result suggests that the CSS-based data projection procedure is technically straightforward and reliable.

We further sought to determine whether this projection procedure can be applied using scATAC-seq data as query data. We used scATAC-seq data (Fluidigm C1) of micro-dissected cortex-like structure in human cerebral organoids[1] (Fig. 5c). For each cell, its chromatin accessibility pattern was summarized into gene activity profile, defined as the enrichment of accessible regions in promoter and gene body of different genes. CSS representations of those cells towards the scRNA-seq reference atlas were then calculated based on their gene activity profiles. The projected UMAP embedding and cell type labels were then compared to the annotation based on the scATAC-seq data (Fig. 5c). We found that most of the cells (77.4%) in the scATAC-seq data were projected to the dorsal telencephalic branch which matches with the experimental design. Using the cell type annotation by clustering of the scATAC-seq data as the benchmark, most of the cells in the scATAC-seq data that annotated as NPC (93.5%) projected to NPC in the reference atlas, while the majority of the cells annotated as neurons projected to IP or neurons in the reference (67.8%). These results suggest that the two annotation strategies cross validated each other. CSS-based representation and projection of scATAC-seq to the scRNA-seq reference atlas is therefore possible and can be helpful for cell type annotation and interpretation of the scATAC-seq data.

Discussion

Here we presented CSS, a simple but powerful data transformation strategy which can be used to integrate multiple data sets. It represents cells by their transcriptomic similarity to cell clusters in different samples, with similarities to clusters within the same sample normalized. This representation can largely eliminate the influence of technical random variation across different samples. We applied CSS to reanalyze relatively complex scRNA-seq data of developing human cerebral organoids. In our assessment, CSS successfully integrated scRNA-seq data of different samples of different experimental conditions, cell lines, experimental batches and technologies, while retaining cell type heterogeneity. Compared with commonly used Harmony, Seurat and LIGER, CSS performed as well as or outperformed other integration methods. These results suggest that CSS offers a great option for scRNA-seq data integration.

In principle, CSS representation corrects for random variations, which do not influence relative similarities of a cell to different cell types, across samples. In other words, any variation which changes the patterns of transcriptomic similarity, including samples at different conditions, are likely remained. This behavior helps, in certain scenarios, e.g. when integrating time course samples from different temporal stages, and create an integrated embedding allowing visualization of any variation on cells which affects their similarities to cell types in different samples.

On the other hand, the same feature could become a limitation of CSS. For instance, CSS may fail to group corresponding cell types in samples of different conditions or species together, as there are likely differences among cells in those samples which affect their similarity spectrums. In that case, other integration methods which maximize correlations among different samples or conditions, e.g. Seurat and LIGER, may be the preferred solution. In addition, as CSS introduces a standardization across clusters of each sample, it requires at least moderate heterogeneity of the integrated samples so that the standardized spectrum is meaningful. When such a condition is not met, other methods, such as Harmony that integrates data based on iterative soft clustering, may provide better results. Additionally, some technical variations may introduce differences on the measured transcriptome that affects similarity patterns, and therefore won’t be corrected by CSS representation.

One useful feature of CSS representation is its ease of applicability to new data, which makes projections of new data to the CSS-represented reference data feasible. In this study, by splitting the human cerebral organoid scRNA-seq data into reference and query data and applying CSS for data representation, we showed that the CSS projection of query data is reliable and accurate, in terms of both the projected UMAP embedding or the transferred cell type labels. What’s more, our work on projecting human cerebral organoid scATAC-seq data to the corresponding scRNA-seq reference further suggests that CSS could help linking single-cell multi-omic data and assist with annotation across modalities. Altogether, these features highlight the utility of CSS as a simple yet powerful approach for single-cell sequencing data integration from complex samples.

Methods

Principles of RSS and CSS

In general, the calculation of RSS and CSS includes the following steps (Supplementary Fig. 1). For RSS calculation, highly variable genes are first identified in the reference data. Reference samples are represented by their average expression profiles of the identified genes. Next, Pearson correlation coefficients are calculated, between each cell and every reference sample. Finally, resulted correlations are normalized across reference samples for each cell.

For CSS calculation, cells from different samples are first separated and clustered. The average transcriptome profile of each cluster in each sample is calculated. For each cell, its similarity to each of the clusters in all samples, defined as the Spearman correlation coefficient between the two transcriptome profiles, is then calculated. Afterwards, z-normalization is applied separately to the cell's similarities to clusters of each sample. For each cell, its standardized similarity spectrums to different samples are then concatenated as the final CSS representation.

The underlying hypothesis of both RSS and CSS is that the undesired confounding factors, e.g. read coverage of different cells, introduce random perturbation to the observed transcriptomic measures which are not correlated with cell type or cell state identities. Cells with the same identity therefore share similar spectrums, after standardizing across similarities to different references, i.e. similarity spectrum to normalize the global differences across cells introduced by the random perturbation. Additionally, as CSS considers all clusters in different samples as references, standardization is done separately across clusters of different samples, so that global differences across samples can be largely eliminated.

Single-cell RNA-seq data generation of cerebral organoids in different experimental conditions

We acquired the human induced pluripotent stem cell (hiPSC) lines Kucg2 and Sojd3 from the HipSci resource, cultured and differentiated them to cerebral organoids following the same protocol as reported[1]. One organoid from each line at 116-day-old was dissociated as described previously[1]. Cell suspension of each organoid was split into two aliquots. Cells in one aliquot per line were pooled, together with cells of another human cerebral organoid of line Wibj2, also from HipSci resource, and one chimpanzee cerebral organoid, and diluted for an appropriate concentration to obtain approximately 10000 cells in one lane of a 10x microfluidic chip device. Methanol fixation was applied to cells in the remaining aliquots, following the same procedure as described previously[14]. Cells were kept at 4°C all the time. Briefly, between 1 and 2×10^6 cells from a filtered single-cell suspension of cerebral organoids, were pelleted at 300 x g for 5 min. The cell pellet was resuspended in 200 µl HBSS without calcium and magnesium (Gibco). To avoid clumping, 800 ul of methanol (pre-chilled to -20 °C) was added to the cells dropwise while gently vortexing the cell suspension

351 at lowest speed. The methanol-fixed cells were kept on ice for 15 min and then stored at -
352 80 °C for seven days. Afterwards, the fixed cells were rehydrated following the procedure
353 described[14]. In brief, cells were moved from -80 °C to 4 °C and kept on ice during the
354 procedure. Cells were then pelleted at 3000 × g, resuspended in DPBS with 2% (w/v) BSA
355 and 1 U/µl RiboLock RNase Inhibitor (Thermo Fisher). Centrifugation and resuspending
356 were repeated once more and cells were then passed through a 30 µm pre-separation filter
357 (Miltenyi Biotec), counted and diluted to aim for 10000 cells in one lane of the microfluidic
358 device (Chromium™ Single Cell 3' Solution v2, 10x Genomics). Libraries were sequenced
359 on Illumina's Hiseq2500 platform in paired-end mode (26+8 bp, 100 bp).
360

361 Single-cell RNA-seq data generation using inDrop

362 Cerebral organoids were generated from an edited version of the embryonic stem cell line H9,
363 following the protocol previously described[15]. Using CRISPR editing, double-stranded breaks
364 were introduced in the first exon of the PLCB1 gene. One allele contained a 4bp deletion, the
365 other allele a 19 bp insertion. Both mutations led to frame shifts and premature stops. At day-
366 150 since culturing, two cerebral organoids were dissociated into single-cells using a papain-
367 based neural tissue dissociation kit (Miltenyi) as described in Miltenyi's Neural Tissue
368 Dissociation protocol. To increase the accessible surface for the dissociation enzyme
369 organoids were cut into pieces and washed up to 3 times in 1x HBSS without Ca^{2+} and Mg^{2+}
370 (HBSS w/o, Sigma). In order to get sufficient disaggregation samples were gently triturated
371 using wide-bore pipette tips and p1,000 and p200 pipettes. Next, a single-cell suspension was
372 generated by removing clumps of cells filtering suspension through 30 and 20 µm diameter
373 strainers. The single-cell suspension was washed up to 3 times in HBSS w/o. Samples were
374 centrifuged at 300*g* for 5 min and resuspended in HBSS w/o. Finally, cell viability was
375 determined with Trypan blue solution (0.4%) using an automatic cell counter (Countess,
376 Invitrogen). For the single-cell RNA-seq experiment cells were diluted to 90,000 cells/ml in
377 15% Optiprep and 0.02% BSA in PBS.
378
379 Single-cell transcriptome barcoding was performed using inDrops[16] and following the protocol
380 by Zilionis et al, 2017[17]. Shortly, generated single-cell suspension was co-encapsulated with
381 RT-lysis mix and barcoded hydrogel beads (Droplet Genomics) into 3-4 nl droplets. After
382 cDNA synthesis was performed, droplets were broken and cDNA libraries were prepared by
383 second strand synthesis, linear amplification by in vitro transcription, amplified RNA
384 fragmentation, reverse transcription and PCR. Prepared cDNA libraries were sequenced
385 paired-end (100 bp, 50 bp) on an Illumina HiSeq2500 platform on 2 lanes.
386

387 Data retrieve, processing and integration

388 ScRNA-seq data sets of human two-month-old cerebral organoids and cerebral organoid
389 developmental time course from PSC, as well as the metadata and cell type annotation,
390 were retrieved from ArrayExpress (accession E-MTAB-7552)[1] The scATAC-seq data was
391 retrieved from ArrayExpress (accession E-MTAB-8089)[1]. Quality control was done with the
392 same procedure as described in the original publication.
393
394 For the newly generated scRNA-seq data of fixed/unfixed experimental conditions, Cell
395 Ranger, the set of analysis pipelines suggested by 10x Genomics, was used to demultiplex

raw base call files of libraries by 10x Genomics to FASTQ files and align reads to the human genome and transcriptome (hg38, provided by 10x Genomics) with the default alignment parameters. Demultiplexing of human and chimpanzee cells in the unfixed sample was done based on genomic loci with diverged bases between human and chimpanzee, following the same procedure as described[1]. Only human cells were used in the later analysis. Demultiplexing of the three human lines in the unfixed sample was done using demuxlet[18], based on the genotyping information of lines downloaded from HipSci websites. Cells with the best singlet prediction being Wibj2 with likelihood no less than 10 higher than the second best singlet likelihood were discarded.

For the newly generated inDrops scRNA-seq data, preprocessing was done following the Drop-seq tools procedure. In brief, quality control to the FASTQ files was done by removing reads with multiple low-quality bases at cell or molecular barcodes, and polyA sequences (with at least six consecutive As) trimmed using Drop-seq tools (v1.12). Remaining reads were mapped to the hg19 human genome using STAR with default parameters. Count matrices were made using Drop-seq tools. Cells with less than 10000 reads were dropped from the analysis.

To integrate the human two-month-old cerebral organoid scRNA-seq data, RSS to the human fetal BrainSpan RNA-seq reference data was calculated as described in the original publication[1]. For CSS calculation, principal component analysis (PCA) was firstly applied to the data considering the top 5000 highly variable genes. CSS was then calculated, with different organoids as different samples for louvain clustering (with resolution of 0.6) implemented in Seurat, which took the top-20 calculated principal components (PCs) as the input. Harmony was applied with the default parameters and the top-50 PCs, calculated in the same way as above, to integrate different organoids. Seurat v3 was applied following its standard workflow of integration, using 5000 features for anchoring and top-30 PCs in the weighting procedure. LIGER was applied following basic commands tutorial, with variance threshold being 0.3, inner dimension of factorization being 20, convergence threshold being 5E-5, three restarts of integrative non-negative matrix factorization and clustering resolution of 0.4. The same parameters were also applied to the integration of the developmental time course scRNA-seq data of human cerebral organoids from PSC, and the integration of scRNA-seq data of human cerebral organoids in different experimental conditions (fixed/fresh). The only exception is the variance threshold in LIGER, which was set as 0.1 for the developmental time course data, and 0.01 for the two experimental conditions data. Such difference was made to keep the number of variable features used in LIGER integration similar.

To integrate the three scRNA-seq data sets generated by different technologies, 5000 highly variable genes were determined for each data set separately. Genes defined as highly variable in at least two data sets after excluding genes reported as cell-cycle-related genes were used for integration, accounted for 2984 genes, were used for integration. The variance thresholds in LIGER were 0.3, 1 and 0.8 for the 10x, inDrop and C1/SS2 data sets, respectively. Other parameters are the same as described above.

441 Quantitative metrics of integration performance

442 To quantify the performances of different integration methods on the human two-month-old
443 cerebral organoid scRNA-seq data, we calculated k-nearest neighbors (kNN, k=50) for each
444 cell in different integration space (RSS, CSS, Harmony-integrated top-50 PCs, Seurat-
445 integrated top-20 PCs, and LIGER-based quantile aligned factor loadings). Based on the cell
446 type annotation retrieved from the original publication, which is based on RSS, we counted
447 the proportions of neighbors, for cells annotated as cortical excitatory neurons (EN) or
448 MGE/CGE inhibitory neurons (IN), which are annotated as the same cell type of the same
449 organoid, the same cell type but of a different organoid in the same experimental batch, the
450 same cell type but of a organoid in a different experimental batch, and a different cell type. A
451 good integration method should increase the proportion of neighbors annotated as the same
452 cell type but from a different organoid, while keeping the proportion of neighbors annotated
453 as a different cell type low. We focused on the two mentioned cell types, as they are the two
454 cell types which are the most distinct from the others while detected in most of the organoids
455 in the data.
456
457 A similar metric was calculated for the developmental time course scRNA-seq data set of
458 human cerebral organoids from PSC, but focusing on PSC (including PSC and EB time
459 points), as it is the most distinct cell type from the others. The proportions of nearest
460 neighbors of each PSC which was PSC of the same line, PSC of the other line, or cells from
461 any other sample were calculated. A good integration method should show a high proportion
462 of PSC neighbors being PSC of the other line and low proportion of cells from other
463 samples. In addition, average distances between each PSC and cells at PSC/EB,
464 neuroectoderm, neuroepithelium, and cerebral organoids at age of two-to-four months were
465 calculated on each integration space. The resulting average distances were normalized to
466 the mean of average distances between different cells at PSC stage to allow comparison of
467 different integration methods.
468
469 Similarly, we calculated kNN for cells in different integration space to access the
470 performances of different integration methods on different experimental conditions and
471 technologies. Neighbors of each cell are grouped into four groups: cells annotated as the
472 same cell type, which represent 1) the same experimental condition or technology, or 2)
473 different experimental condition or technologies, 3) cells annotated as a different cell type,
474 but of the same major cell fate, and 4) cells annotated as a different cell type of a different
475 major cell fate. Here, major cell fates include cortical excitatory cell fate (cortical
476 NPC/IP/EN), GE inhibitory cell fate (GE NPC, CGE/MGE IN, LGE IN), non-telencephalic cell
477 fate (non-telencephalic NPC/EN/IN), choroid plexus, mesenchymal-like cells, and epithelial-
478 like cells.
479

480 Characterization of transition cell states between PSC and neuroectoderm

481 To identify transition cells between PSC and neuroectoderm cells, for cells from PSC/EB
482 and neuroectoderm samples, we firstly identified their kNNs (k = 50) on the CSS space.
483 Cells with at least 15 PSC/EB neighbors and at least 15 neuroectoderm neighbors are
484 defined as cells at the transition state. Differential expression analysis was done using the R
485 package ‘presto’, to compare transition cells with control PSC/EB cells and control
486 neuroectoderm cells separately. Control PSC/EB cells were defined as cells of PSC/EB

samples with at least 40 of their 50 nearest neighbors being PSC/EB cells. Control neuroectoderm cells were defined in a similar way. Marker genes of the transition state were defined as genes with BH-corrected Wilcoxon rank sum test P<0.01, fold change > 1.5 and AUC > 0.6 in both comparison of transition-vs.-PSC/EB and transition-vs.-neuroectoderm.

Calculation of CSS representation towards the scRNA-seq reference for query data

To calculate CSS representation of scRNA-seq data towards a reference, Spearman correlation coefficients were calculated between transcriptomic profile of each query cell and average transcriptomic profiles of clusters in the reference samples. For each query cell, correlations to clusters of the same reference sample were standardized. The standardized similarities to clusters in different reference samples were then concatenated for the final representation.

To calculate CSS representation of scATAC-seq data towards a scRNA-seq reference, we firstly summarized peak accessibilities to gene activity scores for each cell. The detected peaks were annotated using the R package ChIPseeker[19], against the gene annotation of UCSC (hg19). For each cell, the proportion of detected genic peaks, defined as peaks annotated to be at the promoter, exonic or intronic region of genes, was calculated (denoted as p_i for cell i). For each gene with at least ten genic peaks detected in the data set, its proportion of detected genic peak at each cell was also calculated (denoted as $p_{i,j}$ for gene j in cell i). The gene activity score of gene j in cell i was then defined as the odds ratio $p_{i,j}$ / p_i. CSS representation was then calculated for each cell against the reference sample clusters, by calculating, standardizing and concatenating the Spearman correlation coefficients between the gene activity profile of cells in the scATAC-seq data and the average transcriptomic profiles of the reference sample clusters.

Acknowledgement

We thank E. H. Gustafson, S. Wolfinger and J. A. Knoblich of IMBA, Vienna for providing the cerebral organoids for the inDrops experiment. We thank L. Mazutis, J. Nainys, D. Kučiauskas, K. Simutis from Vilnius University for assisting with inDrops platform setup and for providing barcoded hydrogels (Droplet Genomics), J. Kageyama and M. Dannemann for the computational support, and S. Jansen and S. Kanton for experimental support. This project has been made possible in part by the Chan Zuckerberg Initiative DAF (grant CZF2017-173814), an advised fund of Silicon Valley Community Foundation, European Research Council (Anthropoid-803441, J.G.C.; Organomics-758877, B.T.), Swiss National Science Foundation (Project Grant-310030_184795, J.G.C).

Author contributions

ZH implemented the method and performed the analysis. AB performed scRNA-seq experiments using inDrop. SE performed fixation on organoids and scRNA-seq experiment using 10x. ZH, JGC, BT designed the study and wrote the manuscript.

Author information

Conflict of interest: The authors declare no conflict of interest. Correspondence should be addressed to grayson.camp@iob.ch, barbara.treutlein@bsse.ethz.ch

Accession code

The single-cell RNA-seq data using inDrop and the single-cell RNA-seq data of the fixation experiment were deposited on ArrayExpress with the accession number XXXXX.

Figures

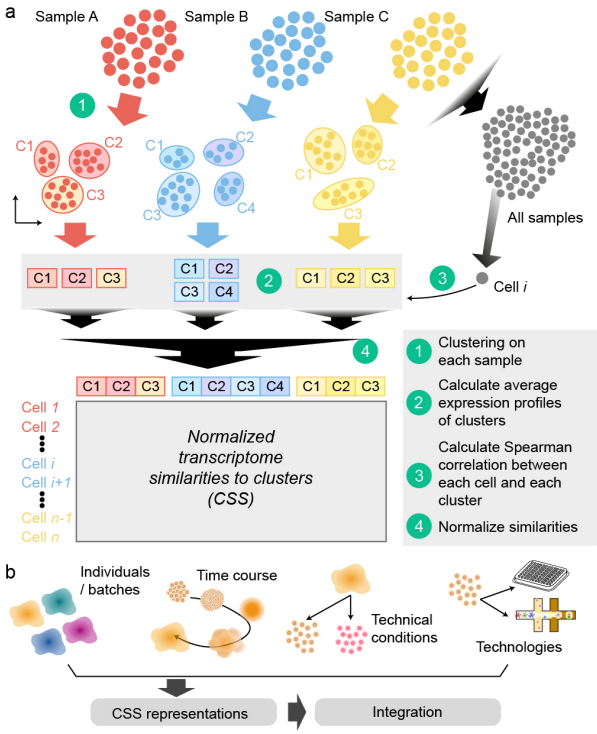


Figure 1: **Schematic of Cluster Similarity Spectrum data representation.** (a) First, clustering is performed on samples separately. Second, the average expression profile of each cluster is calculated. Third, the Spearman correlation is calculated between each cellular transcriptome and the averaged transcriptome of each cluster as their transcriptome similarity. Fourth, for each cell, resulting similarities are then normalized across clusters of each sample, and normalized similarities to different samples are concatenated for the final similarity spectrum of the cell. The CSS vector representation of each cell is used for downstream clustering, embedding, projection, and other analyses. (b) We applied CSS to integrate human cerebral organoid data from different individuals, batches, experimental and technical conditions, and technologies.

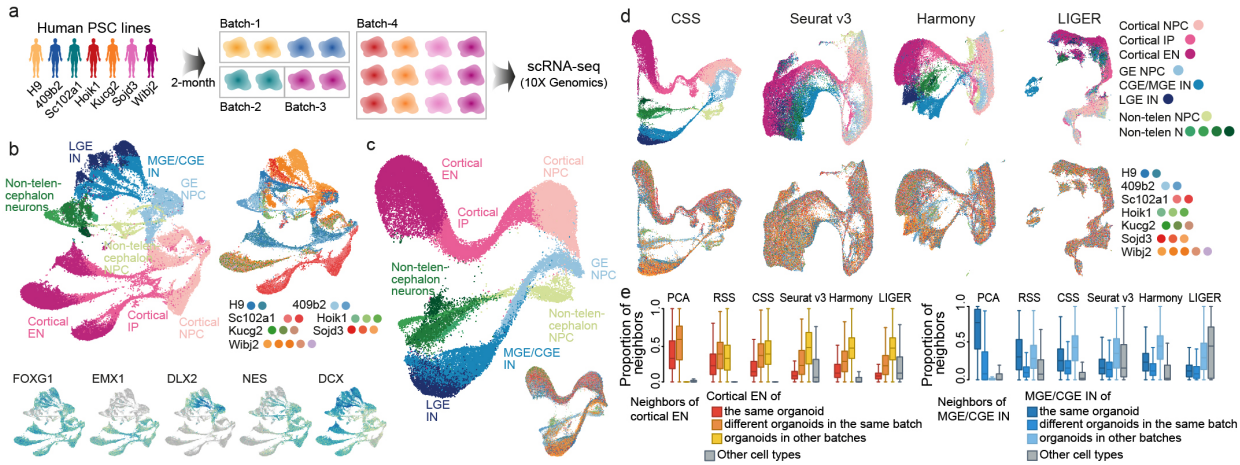


Figure 2: Comparison of different integration methods to resolve cell type heterogeneity in cerebral organoid scRNA-seq data. (a) Schematic of the experimental design. Twenty two-month-old cerebral organoids were generated from seven different human PSC lines in four batches, with their scRNA-seq data generated using 10x Genomics technology. (b) UMAP embedding of cells with no integration, with cells colored by RSS-based cell type annotation (top left) and organoid of origin (top right). Expression patterns of example cell type markers are shown. (c) RSS to fetal BrainSpan RNA-seq data integrates cells of different organoids, and is used as the benchmark of the comparison. The RSS-based UMAP embedding is colored by cell type annotation and organoid, respectively. (d) UMAP embeddings based on CSS and three other integration methods, colored by RSS-based cell type annotation (top row) and organoid (bottom row). (e) Proportion of neighbors of cortical excitatory neurons (EN, left) and MGE/CGE inhibitory neurons (IN, right) in each annotation category. Cell neighbors are defined as cells with the shortest Euclidean distances from the cell in PCA (no integration) or the five different integrated representation spaces.

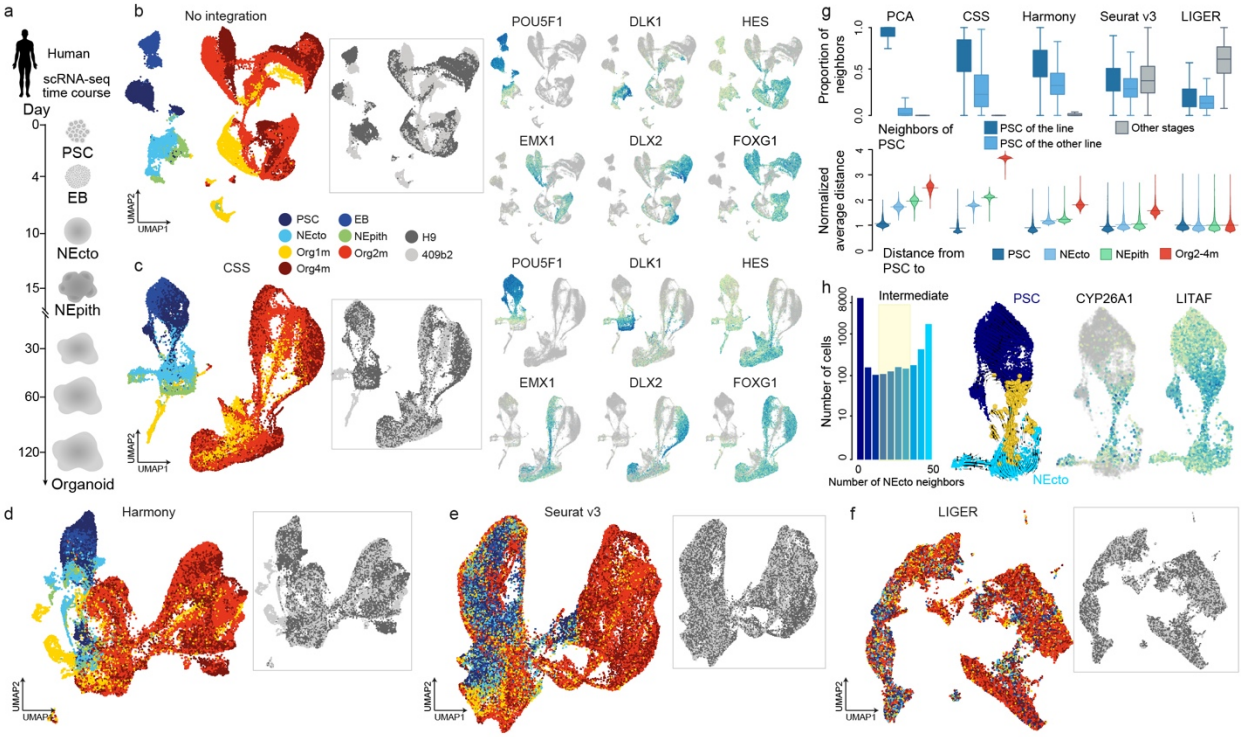


Figure 3: CSS integrates time-course scRNA-seq data of cerebral organoid development from stem cells. (a) Schematic of the experimental design. PSC, pluripotent stem cell; EB, Embryoid Body; NEcto, Neuroectoderm; NEpith, Neuroepithelium. (b) Joint analysis of the time course cerebral organoid development scRNA-seq data without integration. Dots, each representing one cell, are colored by sample time points (left) and PSC lines (right), respectively. Expression patterns of example markers are shown. Org, Organoid; m, month. (c) Joint analysis with CSS representation. The cells in the UMAP embedding are colored by sample time points and PSC lines, respectively. Expression patterns of the same markers are shown. (d-f) UMAP embeddings based on integrations with (d) Harmony, (e) Seurat v3 and (f) LIGER, with cells colored by sample time points (left) and PSC lines (right). (g) Proportion of PSC neighbors in different annotation categories (top), and average Euclidean distances from PSC to cells at later time points in PCA and different integrated representation spaces (bottom). PSC is the union of cells in PSC and EB time points. Neighbors are defined for each cell, as other cells with the shortest Euclidean distance in PCA (no integration) or other integrated representation spaces. (h) Cells at the intermediate state between PSC and neuroectoderm (NEcto) are defined as those with comparable proportion of PSC and NEcto neighbors in the CSS space (left). RNA velocity analysis (middle) indicates potential of cell state transition from PSC (dark blue) to NEcto (light blue) via the defined intermediate cells (yellow). Expression patterns of two marker genes of the intermediate cells, CYP26A1 and LITAF, are shown (right).

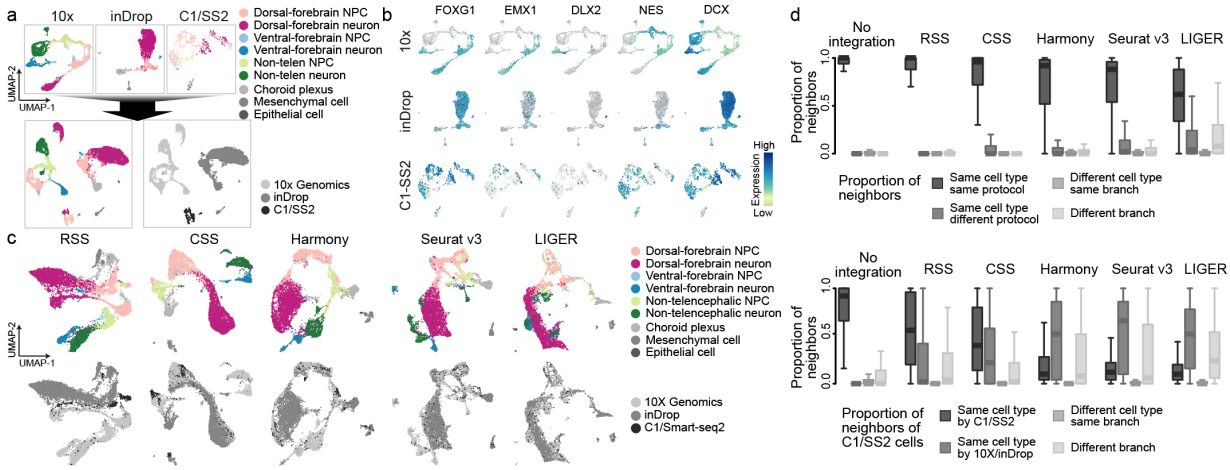


Figure 4: CSS integrates scRNA-seq data generated by different technologies. (a) Joint analysis of cerebral organoid scRNA-seq data generated using different technologies without integration. Cell type annotation is performed in each data set separately (top). Joint UMAP embedding is generated without integration, and colored by cell types (bottom left) and technologies (bottom right), respectively. (b) Expression patterns of example cell type markers on the joint UMAP embedding without integration. (c) UMAP embeddings based on different integration methods, colored by cell type annotation (top) and technology (bottom). (d) Proportion of neighbors in different annotation categories. Neighbors are defined as cells with the shortest Euclidean distances in PCA (no integration) or integrated representation spaces. Upper panel shows distributions of all cells, lower panel shows distribution of cells measured by C1/SS2 technology only.

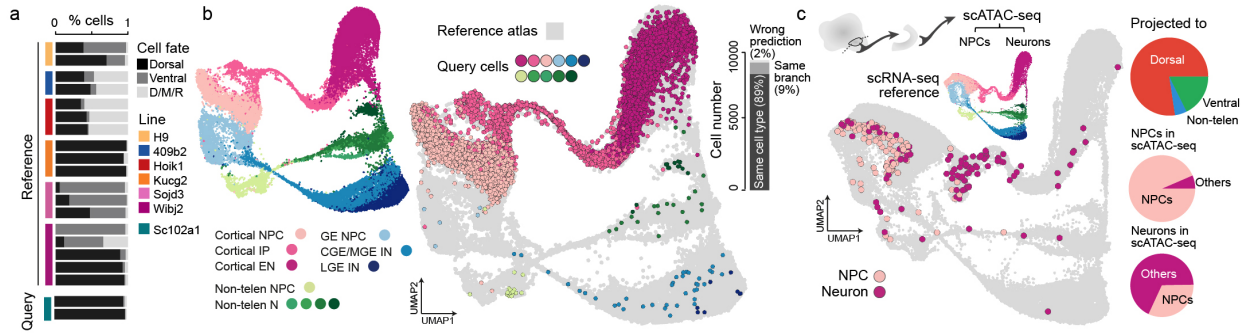
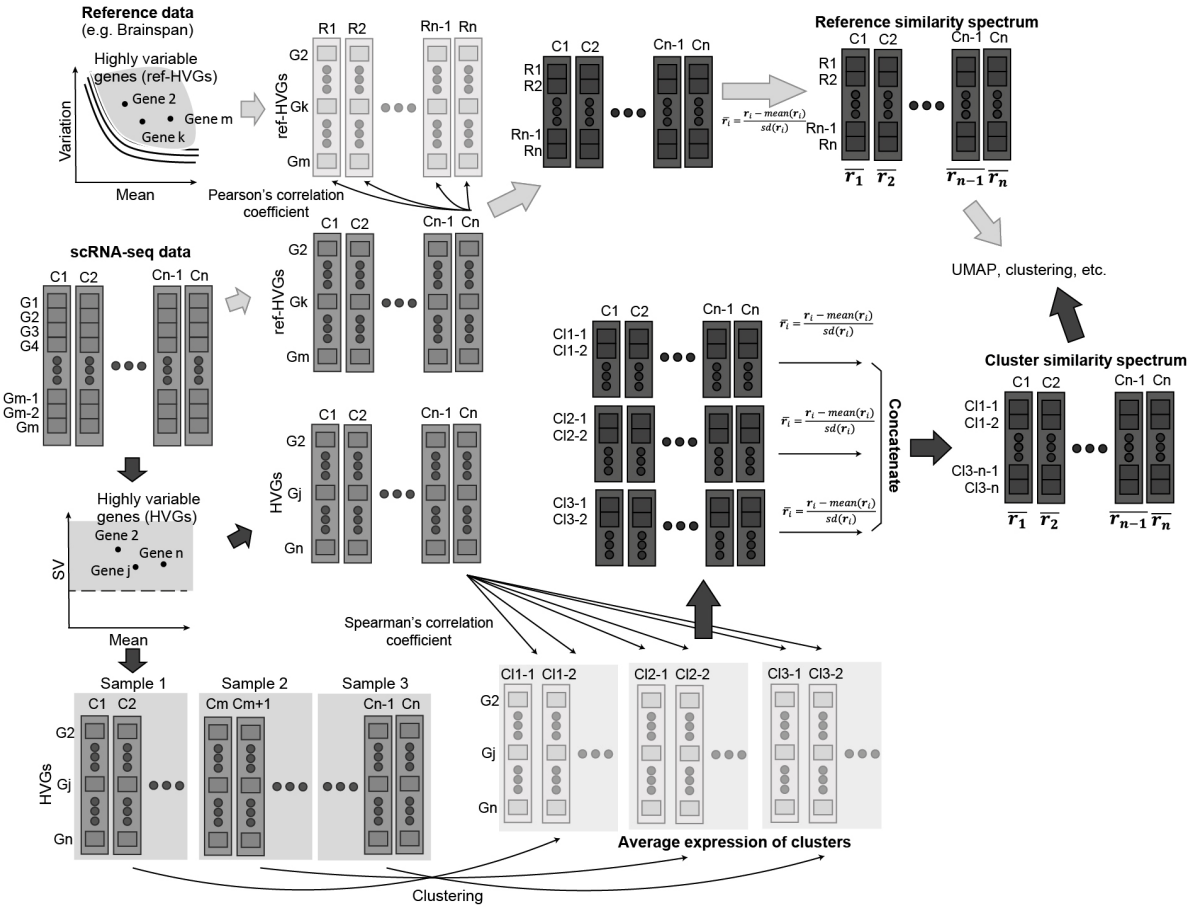
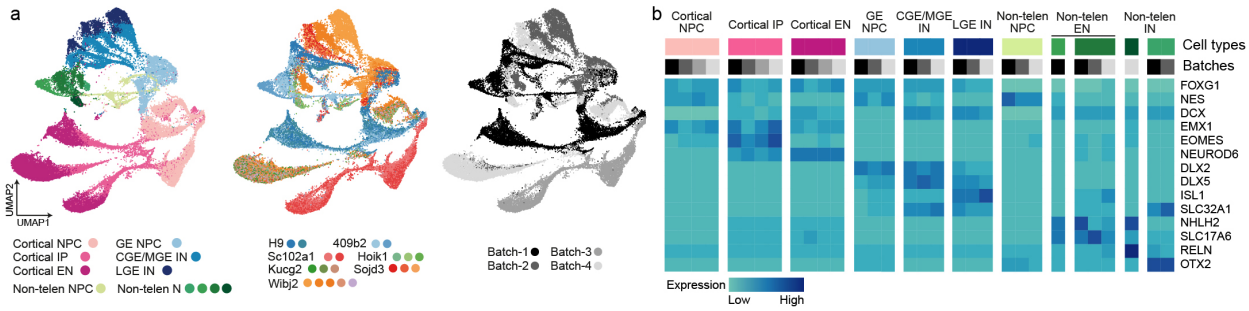


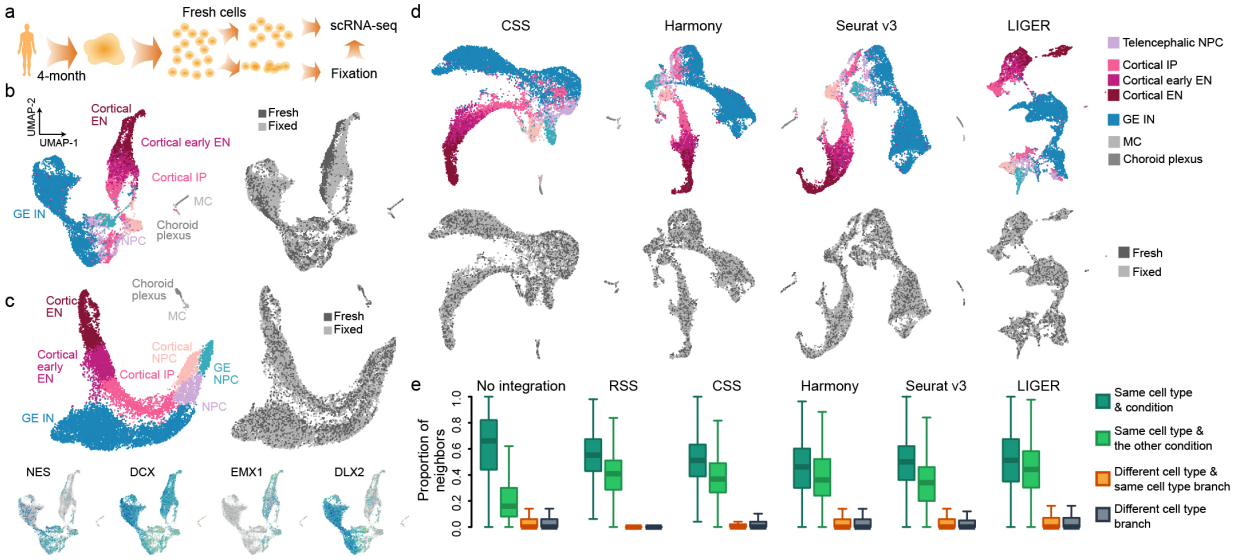
Figure 5: Projection of scRNA-seq and scATAC-seq data to the reference scRNA-seq atlas using CSS. (a) Projection of scRNA-seq data of two Sc102a1 cerebral organoids to the reference atlas built using scRNA-seq data of the other eighteen cerebral organoids. The barplot shows proportion of cells per organoid that are in the dorsal or ventral telencephalon or diencephalon, mesencephalon and rhombencephalon (D/M/R) neuronal branches. (b) UMAP embedding of the reference atlas based on CSS representation. Cells are colored by cell type annotation (left). Cells in the Sc102a1 organoids are projected to the reference atlas, with cells in the projected embedding colored by their cell type annotation (middle). Proportion of cells with predicted cell type consistent with annotation is shown in the barplot (right). (c) Projection of scATAC-seq data of dissected cortex-like areas in cerebral organoids to the scRNA-seq reference atlas including twenty cerebral organoids. Projected loci of cells in the scATAC-seq data set is shown in the reference UMAP embedding, colored by the independent cell type annotation. Pie charts show proportion of cells in the scATAC-seq data set projected to dorsal, ventral or D/M/R neuronal branches (top right), proportion of predicted cell types of NPCs (middle right), and proportion of predicted cell types of neurons (bottom right).



Supplementary Figure 1: Detailed schematic of RSS and CSS representation calculation. To calculate RSS, highly variable genes in the reference data are first identified. Transcriptome similarity between each cell in the scRNA-seq data is calculated to each of reference samples across the identified genes. The resulting correlations of one cell are normalized using z-transform to obtain its RSS representation. To calculate CSS, clustering is applied to cells in each sample separately. Average transcriptome profiles of clusters in samples are calculated. Similarities between transcriptome of each cell and the cluster average transcriptome profiles are calculated. For each cell, similarities to clusters of one sample are normalized by z-transform, with the normalized similarities to different samples concatenated to obtain its CSS representation.



Supplementary Figure 2: Cell type annotation of the cerebral organoid scRNA-seq data. (a) Joint analyzing scRNA-seq data of all the twenty cerebral organoids with no integration. The UMAP embedding is colored by RSS-based cell type annotation (left), organoids (middle), and experimental batches (right), respectively. Cells are primarily grouped by major cell fates of cortical EN, GE IN and non-telencephalon neurons, while differences of different organoids and batches are substantial. (b) Average expression patterns of selected cell type markers in different cell types, with cells from organoids of different experimental batches separated.



Supplementary Figure 3: Integration of scRNA-seq data of cerebral organoids with fixed and unfixed experimental conditions. (a) Schematic of the experimental design. Fresh and methanol-fixed single-cell suspensions of the same four-month-old cerebral organoid were measured using scRNA-seq. (b) UMAP embedding of the scRNA-seq data before integration. Cells are colored by cell types (left, MC for mesenchymal-like cells) and experimental conditions (right). Expression patterns of four example genes are shown. (c) RSS-based integration was used as the benchmark of the analysis. The RSS-based UMAP embeddings are colored by cell type (left) and experimental condition (right). (d) UMAP embeddings based on CSS, Harmony, Seurat v3 and LIGER integration. Cells are colored by cell type (first row), and by experimental condition (second row). (e) Proportion of neighbors of each cell in different categories, with neighbors defined as cells with the shortest Euclidean distances with a cell in PCA (no integration) or different integration spaces.

662

663 ## References

664 1 Kanton, S. *et al.* Organoid single-cell genomic atlas uncovers human-specific
665 features of brain development. *Nature* **574**, 418-422, doi:10.1038/s41586-019-1654-
666 9 (2019).
667 2 Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates
668 cellular heterogeneity in human colorectal tumors. *Nat Genet* **49**, 708-718,
669 doi:10.1038/ng.3818 (2017).
670 3 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with
671 Harmony. *Nat Methods* **16**, 1289-1296, doi:10.1038/s41592-019-0619-0 (2019).
672 4 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902
673 e1821, doi:10.1016/j.cell.2019.05.031 (2019).
674 5 Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts
675 Features of Brain Cell Identity. *Cell* **177**, 1873-1887 e1817,
676 doi:10.1016/j.cell.2019.05.006 (2019).
677 6 Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell
678 transcriptomes using Scanorama. *Nat Biotechnol* **37**, 685-691, doi:10.1038/s41587-
679 019-0113-3 (2019).
680 7 Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-
681 cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat*
682 *Biotechnol* **36**, 421-427, doi:10.1038/nbt.4091 (2018).
683 8 Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell
684 RNA sequencing data. *Genome Biol* **21**, 12, doi:10.1186/s13059-019-1850-9 (2020).
685 9 Kim, D. *et al.* Generation of human induced pluripotent stem cells by direct delivery
686 of reprogramming proteins. *Cell Stem Cell* **4**, 472-476,
687 doi:10.1016/j.stem.2009.05.005 (2009).
688 10 Yu, J. *et al.* Human induced pluripotent stem cells free of vector and transgene
689 sequences. *Science* **324**, 797-801, doi:10.1126/science.1172482 (2009).
690 11 Ribes, V., Fraulob, V., Petkovich, M. & Dolle, P. The oxidizing enzyme CYP26a1
691 tightly regulates the availability of retinoic acid in the gastrulating mouse embryo to
692 ensure proper head development and vasculogenesis. *Dev Dyn* **236**, 644-653,
693 doi:10.1002/dvdy.21057 (2007).
694 12 Uehara, M. *et al.* CYP26A1 and CYP26C1 cooperatively regulate anterior-posterior
695 patterning of the developing brain and the production of migratory cranial neural crest
696 cells in the mouse. *Dev Biol* **302**, 399-411, doi:10.1016/j.ydbio.2006.09.045 (2007).
697 13 White, R. J., Nie, Q., Lander, A. D. & Schilling, T. F. Complex regulation of cyp26a1
698 creates a robust retinoic acid gradient in the zebrafish embryo. *PLoS Biol* **5**, e304,
699 doi:10.1371/journal.pbio.0050304 (2007).
700 14 Alles, J. *et al.* Cell fixation and preservation for droplet-based single-cell
701 transcriptomics. *BMC Biol* **15**, 44, doi:10.1186/s12915-017-0383-5 (2017).
702 15 Lancaster, M. A. *et al.* Cerebral organoids model human brain development and
703 microcephaly. *Nature* **501**, 373-379, doi:10.1038/nature12517 (2013).
704 16 Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to
705 embryonic stem cells. *Cell* **161**, 1187-1201, doi:10.1016/j.cell.2015.04.044 (2015).
706 17 Zilionis, R. *et al.* Single-cell barcoding and sequencing using droplet microfluidics.
707 *Nat Protoc* **12**, 44-73, doi:10.1038/nprot.2016.154 (2017).
708 18 Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural
709 genetic variation. *Nat Biotechnol* **36**, 89-94, doi:10.1038/nbt.4042 (2018).
710 19 Yu, G., Wang, L. G. & He, Q. Y. ChIPseeker: an R/Bioconductor package for ChIP
711 peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382-2383,
712 doi:10.1093/bioinformatics/btv145 (2015).

713