Identifying Gene Expression Programs of Cell-type Identity and Cellular Activity with Single-Cell RNA-Seq

Dylan Kotliar[1,2,3]*†, Adrian Veres[1,3,4]*, M. Aurel Nagy[3,5], Shervin Tabrizi[2], Eran Hodis[3,6], Douglas A. Melton[4,7], Pardis C. Sabeti[1,2,7]

[1]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Harvard Stem Cell Institute, Harvard University, Cambridge, MA, USA. [5]Department of Neurobiology, Harvard Medical School, Boston, MA, USA. [6]Biophysics Program, Harvard University, Cambridge, Massachusetts, USA. [7]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA.

* These authors contributed equally to this work.
† Correspondence should be addressed to: D.K. (dylan_kotliar@hms.harvard.edu).

Abstract

Identifying gene expression programs underlying cell-type identity as well as cellular activities (e.g. life-cycle processes, responses to environmental cues) is crucial for understanding the organization of cells and tissues. Although single-cell RNA-Seq (scRNA-Seq) can quantify transcripts in individual cells, each cell's expression profile may be a mixture of both types of programs, making them difficult to disentangle. Here, we develop an adapted non-negative matrix factorization approach, consensus NMF (cNMF), as a solution to this problem. We rigorously benchmark it against existing and novel scRNA-Seq methods, and cNMF performs best, increasing the accuracy of cell-type identification while simultaneously inferring interpretable cellular activity programs in scRNA-Seq data. Applied to published brain organoid and visual cortex scRNA-Seq datasets, cNMF refines the hierarchy of cell-types and identifies both expected (e.g. cell-cycle and hypoxia) and intriguing novel activity programs. We make cNMF and related tools available to the community and illustrate how this approach can provide key insights into gene expression variation within and between cell-types.

Main Text

Genes act in concert to maintain a cell's identity as a specific cell-type, to respond to external signals, and to carry out complex cellular activities such as replication and metabolism. Coordinating the necessary genes for these functions is frequently achieved through transcriptional co-regulation, whereby genes are induced together as a gene expression program (GEP) in response to the appropriate internal or external signal[1,2]. Transcriptome-wide expression profiling technologies such as RNA-Seq have made it possible to conduct systematic and unbiased discovery of GEPs which, in turn, have shed light on the mechanisms underlying many cellular processes[3].

In traditional RNA-Seq, measurements are limited to an average expression profile of potentially dozens of cell-types in a tissue. Any observed changes in gene expression could reflect induction of a program in some specific cell-type(s), an average of many different changes in multiple cell-types, or changes in overall cell-type composition. Single cell RNA-Seq (scRNA-Seq) solves this problem by measuring the expression of many individual cells simultaneously. This makes it possible to identify the cell-types in the sample as well as any additional sources of variation in their gene expression profiles. Exploiting this new technology, large-scale projects such as the Tabula Muris and the Human Cell Atlas are seeking to identify and characterize all the cell types in complex organisms in states of both health and disease[4,5].

Even with the ability to quantify expression in individual cells, it is still challenging to accurately discern GEPs. scRNA-Seq data is noisy and high-dimensional, requiring computational approaches to uncover the underlying patterns. In addition, technical artifacts such as doublets (where two or more distinct cells are mistakenly collapsed into one) can confound analysis. Key methodological advances in dimensionality reduction, clustering, lineage trajectory tracing, and differential expression analysis have helped overcome some of these challenges[6–9].

Here, we focus on a key challenge of inferring GEPs from scRNA-Seq data, the fact that individual cells may express multiple GEPs and we only observe the resulting mixed profile. A cell's gene expression is shaped by many factors including its cell-type, its state in time-dependent processes such as the cell-cycle, and its response to varied environmental stimuli[10]. We group these into two broad classes of expression programs in scRNA-Seq data: (1) GEPs that correspond to the identity of a specific cell-type such as hepatocytes or melanocytes (identity programs) and (2) GEPs that are expressed in any cell that is carrying out a specific activity such as cell division or immune cell activation (activity programs). While identity programs are, by definition, expressed in all

cells of a specific cell-type, activity programs may vary dynamically in cells of one or multiple types. Cells undergoing the activity would therefore express the activity program in addition to their identity program.

Accurately detecting activity programs in scRNA-Seq data is important for three reasons. First, activity programs can be of primary biological importance. For example, programs underlying immune activation, hypoxia, or programmed cell death might be studied with scRNA-Seq either through experimental manipulations or through observation of naturally occurring variation in a population of cells. Second, we might be interested in understanding how the prevalence of activity programs varies across cell-types -- for example, which cell-types are undergoing the highest rates of cell division. Activity programs would thus serve as an additional layer of information in addition to cell-type. Finally, activity programs may confound characterization of the cell-types in which they occur. For example, cell-cycle genes may be spuriously included in the identity programs of proliferative cell-types. Currently, cell cycle is treated as an artifact of scRNA-seq data and is occasionally removed computationally prior to clustering analysis[11,12]. However, this is just a specific instance of the broader problem of confounding of identity and activity programs.

Existing single-cell analysis methods are of limited utility for identifying activity GEPs. In an experimental setting where a stimulus is applied to some cells and not to others, activity programs underlying the response to the stimulus can be identified with differential expression analysis. While this is the best approach for conclusively determining the etiology of an activity program, it cannot be used outside of this experimental setting and thus cannot identify naturally occuring activity GEPs. In a non-manipulated setting, clustering or dimensionality reduction approaches must be used instead. However, simple clustering is limited because it assigns each cell to a single class and therefore cannot simultaneously capture both the cells' type and activity states. While dimensionality reduction methods such as Principal Component Analysis (PCA)[13] don't have this limitation, the dimensions they infer may not necessarily align with biologically meaningful gene expression programs and are frequently ignored in practice.

GEP deconvolution is better framed as a signal separation problem rather than a dimensionality reduction because we aim for the inferred dimensions to be biologically meaningful programs that can teach us about the cell-types and cellular activities in the data. In a signal separation problem, the observed samples each represent a different weighted mixture of a smaller number of component signals that we seek to infer[14]. When signals are assumed to combine linearly, signal separation amounts to a matrix

factorization; the data matrix is approximated as a product of two lower rank matrices, one representing the latent signals (in our case, gene expression programs) and another specifying how they should be combined for each cell. We refer to the second matrix as a 'usage' matrix as it specifies how much each GEP is 'used' by each cell in the dataset. Commonly used matrix factorization algorithms to consider for this purpose include PCA, Independent Component Analysis (ICA)[15,6], Latent Dirichlet Allocation (LDA)[16], and Non-Negative Matrix Factorization (NMF)[17]. A priori, it is unclear which matrix factorization, would be most appropriate for identifying clear and biologically meaningful gene expression programs.

Here, we simulate scRNA-Seq data to develop and then systematically benchmark these matrix factorization methods for deconvoluting GEPs in our data. We first implement a meta-analysis approach to reduce the stochastic variability in the solutions for LDA, NMF, and ICA (Fig. 1a). We find that with this adaptation, this adaptation of NMF, which we call consensus NMF (cNMF) is able to deconvolute activity and identity GEPs from scRNA-Seq data and significantly outperforms PCA and clustering. We then use cNMF to analyze several previously published scRNA-Seq datasets where it clarifies the cell-types and identifies both expected and novel activity programs, providing insight into how cells vary within and between cell-types.

Results

We set out to identify and benchmark approaches to deconvolute GEPs from simulated single-cell RNA-Seq data. We simulated 15,000 cells made up of 13 cell-types, one cellular activity program that is active in a subset of cells of four cell-types, and a 6% doublet rate (Fig. 1b). To rigorously evaluate performance, we conducted 20 replicates of this simulation each at three different 'signal to noise' ratios (Online methods).

We first examined the performance of ICA, LDA, and NMF, and adapted them for GEP deconvolution. We found that running these methods multiple times with different random initializations yielded variable results; when we ran each algorithm 200 times on a dataset and clustered the resulting 200x14=2800 components with KMeans (K=14), many of the components were poorly correlated with the median of their cluster (Supplementary Fig. 1). While ICA gave the most consistent results, consistency decreased in all methods for lower signal to noise ratios. However, when we filtered outlier replicate components (identified by low similarity to their nearest neighbors), and took the median of the remaining GEPs in a cluster, the resulting estimates correlated well with the true simulated programs (Fig. 1c) (Online methods). We refer to this approach as consensus matrix factorization and to its application to these specific

algorithms as cLDA, cNMF and cICA respectively. We note that this approach is conceptually analogous to consensus clustering[18] and is very similar to procedures commonly used to infer mutational signatures from cancer sequencing data[19].

All consensus matrix factorization methods were well able to deconvolute the identity programs for the 2 highest signal to noise ratios. However, cNMF performed better than the others at identifying the activity program, successfully deconvoluting it in 20 out of 20 simulations at the highest signal to noise level and 17 out of 20 simulations at the next highest level, compared to 19 and 12 out of 20 for cICA and 17 and 5 out of 20 for cLDA (Fig. 1d, Supplementary Fig. 2). When a method, successfully deconvoluted the GEPs, it also correctly inferred the cells that used each program. We obtained a final usage estimate by fixing the component matrix to the cluster median values and running one last iteration of the matrix factorization. These inferred usages recapitulated the true simulated usage of the GEPs in each cell (Fig. 1e, Supplementary Fig. 3a). We thus arrived at an aggregation approach to increase the robustness of LDA, NMF, and ICA for GEP deconvolution and demonstrated that it works well on simulated scRNA-Seq data. W

Next, we benchmarked the consensus matrix factorization methods against PCA and simple clustering using the 20 simulation replicates at the highest signal to noise level. For clustering, we considered both the commonly used Louvain algorithm[20,21] as well as an 'optimal' clustering derived by using our ground-truth simulated values to assign each cell to a single cluster corresponding to its maximum contributing GEP, with cells with >40% usage of the activity program assigned to an activity cluster (Supplementary Fig. 3b). This optimal clustering allows us to put an upper bound on how well any discrete clustering could perform on this dataset in which many cells have profiles derived from two GEPs. While the other matrix factorizations were able to identify all 14 simulated programs, PCA could only identify 13 and would systematically leave out the 14th. The 13 identified programs were combinations of the true programs rather than complete deconvolutions (Supplementary Fig. 3c). Because, simple clustering couldn't assign mixed cluster membership to cells expressing an identity and activity program, it could not fully separate out the activity GEPs from the identity GEPs of cell-types with the activity (Supplementary Fig. 3b,c).

We next quantitatively benchmarked the accuracy of the methods by comparing how well they could infer the genes that constitute each simulated expression program. We used multiple ordinary least squares regression to associate the expression of each gene with the inferred usage of each GEP across cells. We then determined for every association strength threshold, what proportion of the genes associated with a GEP

were truly in that program and what proportion were false positives (Online Methods). We combined all genes from the 20 simulation replicates at the highest signal to noise level to obtain final accuracy estimates. cNMF had the highest accuracy at inferring the activity program, with a sensitivity of 62% at an FDR cutoff of 5% (Fig. 1f). Optimal clustering and cICA were the next most accurate methods with 55%, and 45% sensitivity at the 5% FDR cutoff respectively. While cICA and cLDA outperformed the optimal clustering at activity GEP detection in the majority of runs, their overall performance was significantly decreased by the few replicates where they failed to deconvolute it (Supplementary 4a). Of note, cNMF also performed best at inferring the identity program of cell-types that express the activity program (Supplementary Fig. 4b). Louvain clustering performed poorly at inferring both these identity programs and the activity program as it incorrectly associated activity program genes with the identity program, and vice versa, leading to elevated false positive rates. Except for PCA, which had significantly lower accuracy, all other methods performed comparably well for identifying genes associated with the identity programs of cell-types that don't express the activity program (Supplementary Fig. 4b).

In summary, we found that cNMF was the most accurate at inferring the genes in both activity and identity gene-expression programs from scRNA-Seq data. While cICA had somewhat lower accuracy, we note that independent replicates of ICA were more consistent meaning that it was less dependent on the consensus matrix factorization approach than NMF. Thus, ICA may be useful in situations where running multiple replicates isn't practical. However, we also note that NMF and other non-negative factorizations yielded like LDA yield more interpretable solutions than ICA. With a non-negative factorization, the GEP matrix can be easily converted to interpretable expression units such as the transcripts per million (TPM), whereas this isn't possible for ICA (Online Methods). Moreover, when normalized to sum to 1, each row of the cNMF usage matrix represents the proportion of gene-expression that can be attributed to each GEP in a cell. Because of the negative values, this kind of interpretation isn't possible for the usage matrix in ICA either. Based on the greater accuracy and the increased interpretability of its solutions, we proceeded with cNMF to analyze the real scRNA-seq datasets.

Beyond identifying the activity program itself, we found that cNMF could accurately infer which cells expressed the activity program and what proportion of their expression was derived from the activity program (Fig. 1g). With an expression usage threshold of 10%, 91% of cells expressing the program across all simulations were accurately classified while 94% of cells that did not express the program were correctly identified. Moreover, we observed a high Pearson correlation between the inferred and simulated usages in cells that expressed the program (R=0.735). Thus, cNMF could infer which cells

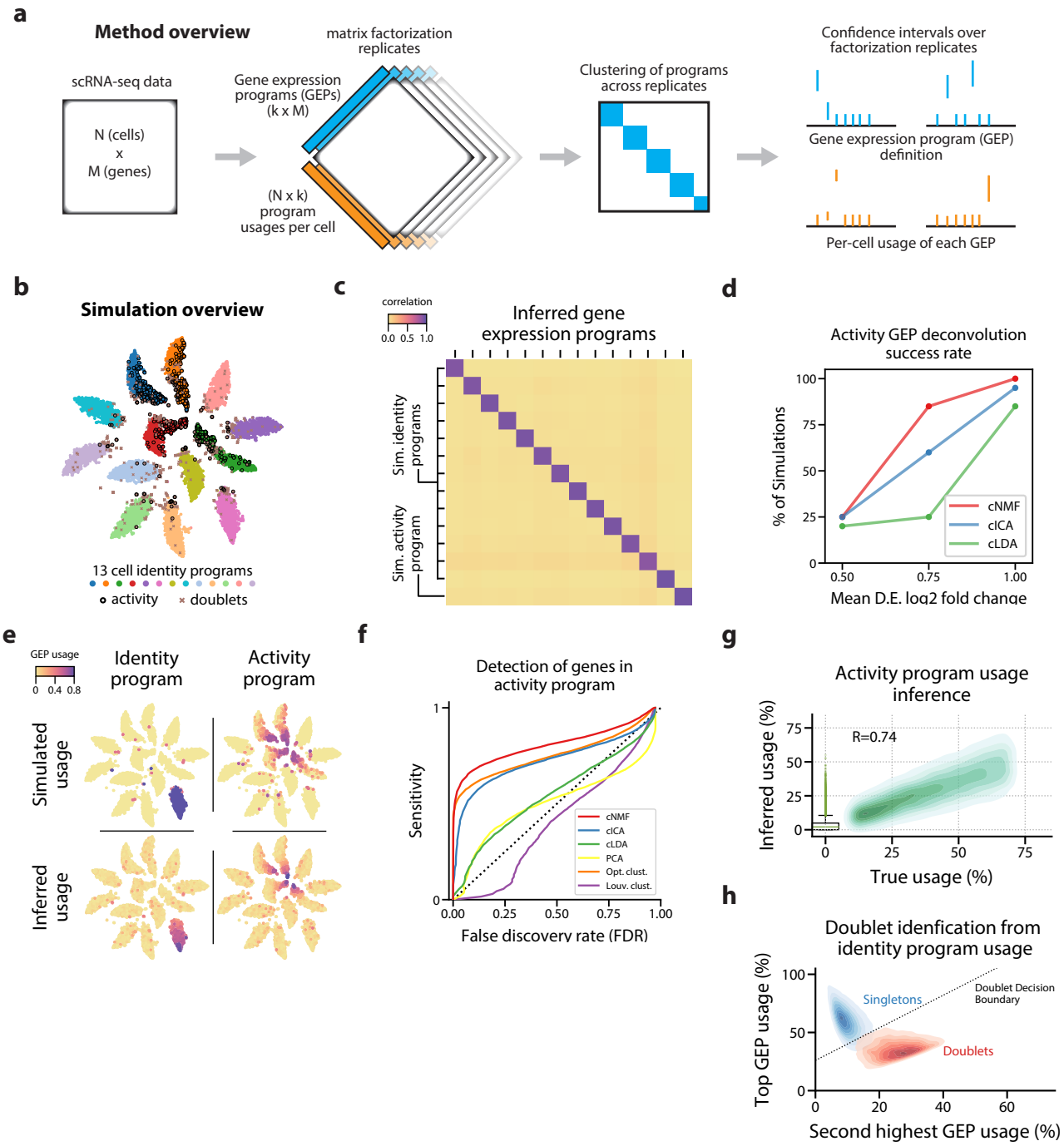


Figure 1: cNMF infers identity and activity expression programs in simulated data. **(a)** Schematic of the consensus matrix factorization pipeline. **(b)** t-distributed stochastic neigbour embedding (tSNE) plot of an example simulation showing different cell-types with marker colors, doublets as gray Xs, and cells expressing the activity gene expression program (GEP) with a black edge. **(c)** Pearson correlation between the true GEPs and the GEPs inferred by cNMF, both in units of gene variance normalized transcripts per million for the simulation in (b). **(d)** Percentage of 20 simulation replicates where an inferred GEP had Pearson correlation greater than 0.75 with the true activity program for each signal to noise ratio (parameterized by the mean log2 fold-change for a differentially expressed gene). **(e)** Same tSNE plot as (b) but colored by the simulated true usage or the consensus non-negative matrix factorization (cNMF) inferred usage of an example identity program (left) or the activity program (right). **(f)** Receiver Operator Characteristic curve (except with false discovery rate rather than false positive rate) showing accuracy of prediction of genes associated with programs by each method for all simulations with differential expression log2 fold-change of 1.0. **(g)** Scatter plot comparing the simulated activity GEP usage and the usage inferred by cNMF. For cells with a simulated usage of 0, the inferred usage is shown as a box and whisker plot with the box corresponding to interquartile range and the whiskers corresponding to 5th and 95th percentiles. **(h)** Kernel density plot of the top and the second highest identity GEP usage inferred by cNMF for all cells that don't express the activity program. Markers are colored by whether the cell is a doublet or a singlet.

express the activity program, as well as what proportion of their transcripts derive from that program.

We also found that cNMF could detect doublets and infer the two cell-types that contributed to each doublet (Fig. 1h). We examined this by first excluding cells that expressed the activity program and then classifying a cell as a doublet if it's top identity GEP usage was >26% higher than the usage of it's second highest identity GEP. With this cutoff, 91% of doublets and 92% of singletons across all simulations were correctly classified into their respective classes. 91% of correctly identified doublets had the correct two programs as their top two GEPs. This illustrates how cNMF can in principle be used to determine the cell-types that combine to form doublets. However, we note that cells with high usage of two or more GEPs may also represent a developmental transition state or cells that otherwise cannot be neatly resolved between two related cell-types. Thus, the presence of differentiation or many cell-types with similar gene expression programs in a dataset may complicate the use of this approach for doublet detection.

Having demonstrated its good performance on simulated data, we used cNMF to re-analyze a published scRNA-Seq dataset of 52,600 single cells isolated from human brain organoids[22]. The initial report of this data confirmed that organoids contain excitatory cell-types homologous to those in the cerebral cortex and retina as well as unexpected cells of mesodermal lineage, but further resolution can be gained on the precise cell-types and how they differentiate over time. As organoids contain many proliferating cell-types, we sought to use this data to detect activity programs -- in this case, cell cycles programs -- in real data.

cNMF identified 31 distinct programs in this data (Supplementary Fig. 5). Most cells had high usage of just a single GEP (Fig. 2a). However, when cells used multiple GEPs, those programs typically had correlated expression profiles, suggesting that they correspond to identity programs of closely related cell-types. (Supplementary Fig. 6). By contrast, 3 GEPs were co-expressed with many distinct and uncorrelated programs suggesting that they represent activity programs that occur across diverse cell-types. Consistent with this, the 28 suspected identity programs were well separated by the cell-type clusters reported in Quadrato et. al. while the three suspected activity programs were expressed by cells across multiple clusters (Supplementary Fig. 7).

Our 28 identity programs further refined the 10 primary cell-type clusters originally reported for this dataset. For example, we sub-classified the mesodermal cluster into three populations expressing genes characteristic of (1) immature skeletal muscle (e.g.

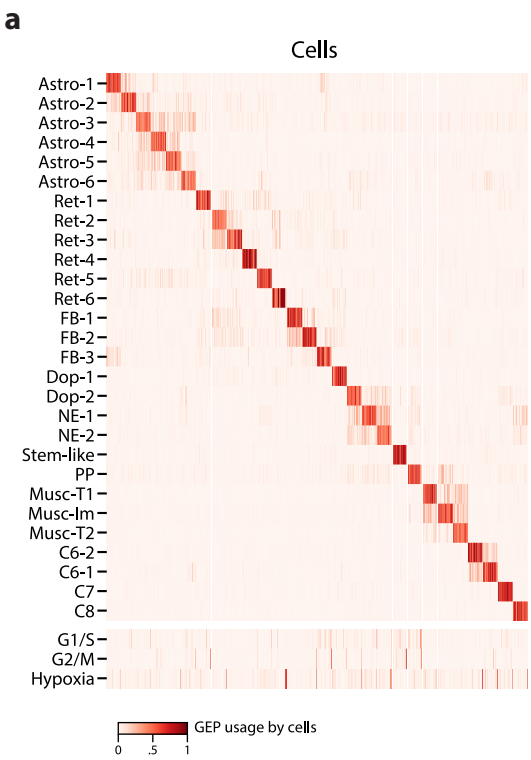
MYOG, *TNNT2*, *NES*), fast-twitch muscle (e.g. *TNNT3*, *TNNC2*, *MYOZ1*), and slow-twitch muscle (e.g. *TNNT1*, *TNNC1*, *TPM3*) (Fig. 2c). This unexpected finding suggests that 2 distinct populations of skeletal muscle cells -- excitatory cell-types with many similarities to neurons -- are differentiating in these brain organoids.

Of the three activity programs, we found that two were strikingly enriched for cell cycle Gene Ontology (GO) sets, suggesting that they correspond to separate phases of the cell cycle (Fig. 2b). One showed stronger enrichment for genesets involved in DNA replication (e.g. DNA Replication $P=6x10^{-55}$ compared to $P=2x10^{-14}$) while the other showed stronger enrichment for genesets involved in mitosis (e.g. Mitotic Nuclear Division, $P=5x10^{-61}$ compared to $P=2x10^{-46}$). These enrichments and inspection of the genes most associated with these programs implied that one represents a G1/S checkpoint program and the other represents a G2/M checkpoint program (Fig. 2d).

The third activity program is characterized by high levels of well-known hypoxia related genes (e.g. *VEGFA*, *PGK1*, *CA9*, *P4HA1*, *HILPDA*) suggesting it represents a hypoxia program (Fig. 2d). This is consistent with the lack of vasculature in organoids which makes hypoxia an important growth constraint[23]. This GEP was significantly enriched for genesets related to protein localization to the endoplasmic reticulum and nonsense mediated decay ($P=3x10^{-37}$, $P=4x10^{-31}$) (Fig. 2b), consistent with literature showing that hypoxia post-transcriptionally increases expression of genes that are translated in the ER[24] and modulates nonsense mediated decay activity[25].

Having identified proliferation and hypoxia activity programs, we sought to quantify their relative rates across cell-types in the data. We found that 2644 cells (5.0%) expressed the G1/S program and 1437 cells (2.7%) expressed the G2/M program (with usage>=10%). Classifying cells into cell-types according to their most used identity program, we found that many distinct populations were replicating. For example, we noticed a rare population, included with the forebrain cluster in the original report, that we label as “stem-like” based on high expression of pluripotency markers (e.g. *LIN28A*, *L1TD1*, *MIR302B*, *DNMT3B*) (Supplementary table 1). These cells showed the highest rates of proliferation with over 38% of them expressing a cell-cycle program in addition to the “stem-like” identity GEP (Fig. 2e).

We also found that a cell cluster labeled in Quadrato et al., 2017 as “proliferative precursors” based on high expression of cell-cycle genes is composed of multiple cell-types including immature muscle and dopaminergic neurons (Supplementary Fig. 7). The predominant identity GEP of cells in this cluster is most strongly associated with the gene PAX7, a marker of self-renewing muscle stem cells[26] (Supplementary table 1).



b

rank	G1/S Activity Program GO term	FDR(p-val	G2/M Activity Program GO term	FDR(p-val	Hypoxia Activity Program GO term	FDR(p-val
1	Cell Cycle	$3x10^{-84}$	Mitotic Cell Cycle	$5x10^{-77}$	Establish. of Protein Loc. To Endoplasmic Reticulum	$3x10^{-37}$
2	Mitotic Cell Cycle	$1x10^{-80}$	Cell Cycle Process	$4x10^{-68}$	Protein Loc. To Endoplasmic Reticulum	$2x10^{-35}$
3	Cell Cycle Process	$4x10^{-77}$	Cell Cycle	$1x10^{-63}$	Translation Initiation	$6x10^{-32}$
4	Chromosome Organization	$4x10^{-70}$	Mitotic Nuclear Division	$5x10^{-61}$	Nuclear Transcribed mRNA Catabolic Process NMD	$4x10^{-31}$
5	DNA Metabolic Process	$4x10^{-69}$	Organelle Fission	$3x10^{-53}$	rRNA Metabolic Process	$8x10^{-28}$
6	DNA Repair	$6x10^{-55}$	Sister Chromatid Division	$1x10^{-47}$	Ribosome Biogenesis	$2x10^{-27}$

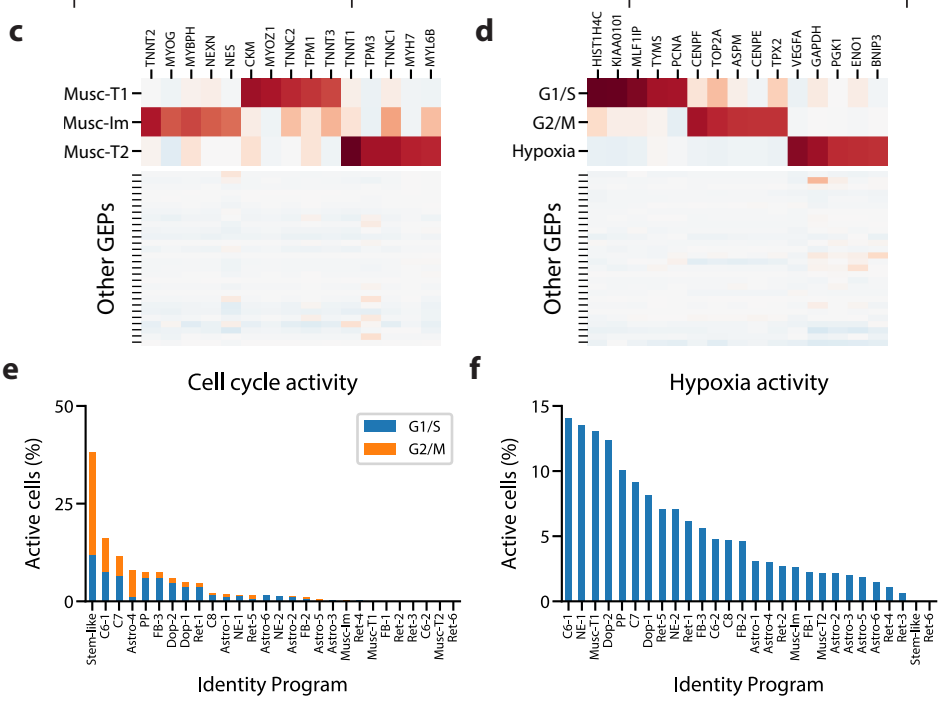


Figure 2: Deconvolution of cell-cycle programs from cell identity in brain organoid data. (a) Heatmap showing percent usage of all GEPs (rows) in all cells (columns). Identity GEPs are shown on top and activity GEPs are shown below. **(b)** Table of P-values for the top six Gene Ontology geneset enrichments for the three activity GEPs. **(c)** Heatmap of Z-scores of top genes associated with three mesodermal programs in those programs (top) and in all other programs (bottom). **(d)** Heatmap of Z-scores of top genes associated with three activity GEPs in those programs (top) and in all other programs (bottom). **(e)** Proportion of cells assigned to each identity GEP that express the G1/S or G2/M program with a percent usage greater than 10%. **(f)** Proportion of cells assigned to each identity GEP that express the hypoxia program with a percent usage greater than 10%.

Indeed, this GEP has high (>10%) usage in 52% of cells that express the immature muscle program, suggesting it may be a precursor of muscle cells. This relationship was not readily identifiable by clustering because the majority of genes associated with the cluster were cell-cycle related. This highlights the ability of cNMF to refine cell-types by disentangling identity and activity programs.

We also saw a wide range of cell-types expressing the hypoxia program, with the highest rates in C6-1, neuroepithelial-1, type 2 muscle, and dopaminergic-2 cell-types. The lowest levels of hypoxia program usage occurred in forebrain, astroglial, retinal, and type 1 muscle cell-types (Fig. 2f). This illustrates how inferring activity programs in scRNA-Seq data using cNMF makes it possible to compare the rates of cellular activities across cell-types.

Next we turned to another published dataset to illustrate how cNMF can be combined with scRNA-Seq of experimentally manipulated cells to uncover more subtle activity programs. We re-analyzed scRNA-Seq data from 15,011 excitatory pyramidal neurons or inhibitory interneurons from the visual cortex of dark-reared mice that were suddenly exposed to 0 hours, 1 hours, or 4 hours of light[27]. This allowed the authors to identify transcriptional changes induced by sustained depolarization, a phenomenon believed to be critical for proper cortical function. Given that the authors identified heterogeneity in stimulus-responsive genes between neuronal subtypes, we wondered if cNMF could identify a shared program and whether it could tease out any broader patterns in what is shared or divergent across neuron subtypes.

We ran cNMF on neurons combined from all three exposure conditions and identified 15 identity and 5 activity programs (Fig. 3a, Supplementary Fig. 8). As we saw in the organoid data, the activity programs were co-expressed with many distinct and uncorrelated GEPs, while the identity programs only overlapped in highly related cell-types. In addition, the identity programs were well separated by the published clusters which we subsequently used to label the GEPs (Supplementary Fig. 9).

Three activity programs were correlated with the stimulus, which indicates that they are induced by depolarization (Fig. 3b). One of these was induced at 1H and thus corresponds to an early response program (ERP). The others were primarily induced at 4H and thus correspond to late response programs (LRPs). These programs overlapped significantly with differentially expressed genes reported in Hrvatin et. al. 2017 ($P=9x10^{-34}$ for the ERP and genes induced at 1H; $P=2x10^{-22}$, $P=4x10^{-14}$ for the LRPs and genes induced at 4Hs).

Intriguingly, one LRP was more induced in superficial cortical layers while the other was more induced in deeper layers. This provides support for a recently proposed model where the ERP is predominantly shared across excitatory neurons while LRPs vary more substantially across neuron subtypes[27]. It also illustrates cNMF's sensitivity: in the initial report, only 64 and 53 genes were identified as differentially expressed in at least one excitatory cell-type at 1H and 4Hs (FC≥2, FDR<.05). Nevertheless, cNMF was able to find this program in the data without knowledge of the experimental design.

cNMF was also able to identify a depolarization-induced program in visual cortex neurons that were not experimentally manipulated to elicit them. We re-analyzed an additional scRNA-Seq dataset of 1,573 neurons from the visual cortex of adult mice that, unlike in the primary dataset, were not reared in darkness or treated with a specific light stimulus[28]. In this dataset, cNMF identified a matching GEP for all visual cortex cell-types found in the original data except for one (Supplementary Fig. 10a). Moreover, it identified a GEP that showed striking concordance with the superficial late response program found in the primary dataset (Fisher Exact Test of genes with Z-score>1.5, OR=114.7, $P=9x10^{-109}$, Pearson Correlation=.65) (Fig. 3c). This program was predominantly expressed in excitatory cells of the more superficial layers of the cortex as would be expected based on the results in the primary dataset. For example, over 50% of the excitatory neurons of cortical layer 2 expressed this activity program (Supplementary Fig. 10b). This demonstrates that cNMF could also find the depolarization-induced activity program in scRNA-Seq of cells that had not been experimentally manipulated.

Finally, cNMF identified 2 intriguing activity programs in the primary visual cortex dataset that were not well correlated with the light stimulus but were expressed broadly across excitatory neurons and inhibitory neurons of multiple cortical layers (Fig. 3b). One activity program (labeled other activity 1 - OA1) is characterized by genes involved in synaptogenesis. In ranked order, the top genes associated with this GEP were *MEF2C*[29], *H2-Q4*, *YWHAZ*[30], *CADM1/SYNCAM1*[31], *NCAM1*[32], and *BICD1*[33] (an example reference is included for genes with a published link to synapse formation). The top genes associated with the remaining activity program (OA2) include several that are involved in cerebral ischemic injury: *MEG3*[34], *GLG1*[35], *RTN1*[36], *ELAVL3*, *CMIP*). These functional interpretations of OA1 and OA2 are speculative but they highlight the ability of cNMF to identify intriguing novel gene expression programs in an unbiased fashion.

In summary, we have shown that matrix factorization approaches can be used to infer identity and activity gene expression programs from scRNA-Seq data. We developed a

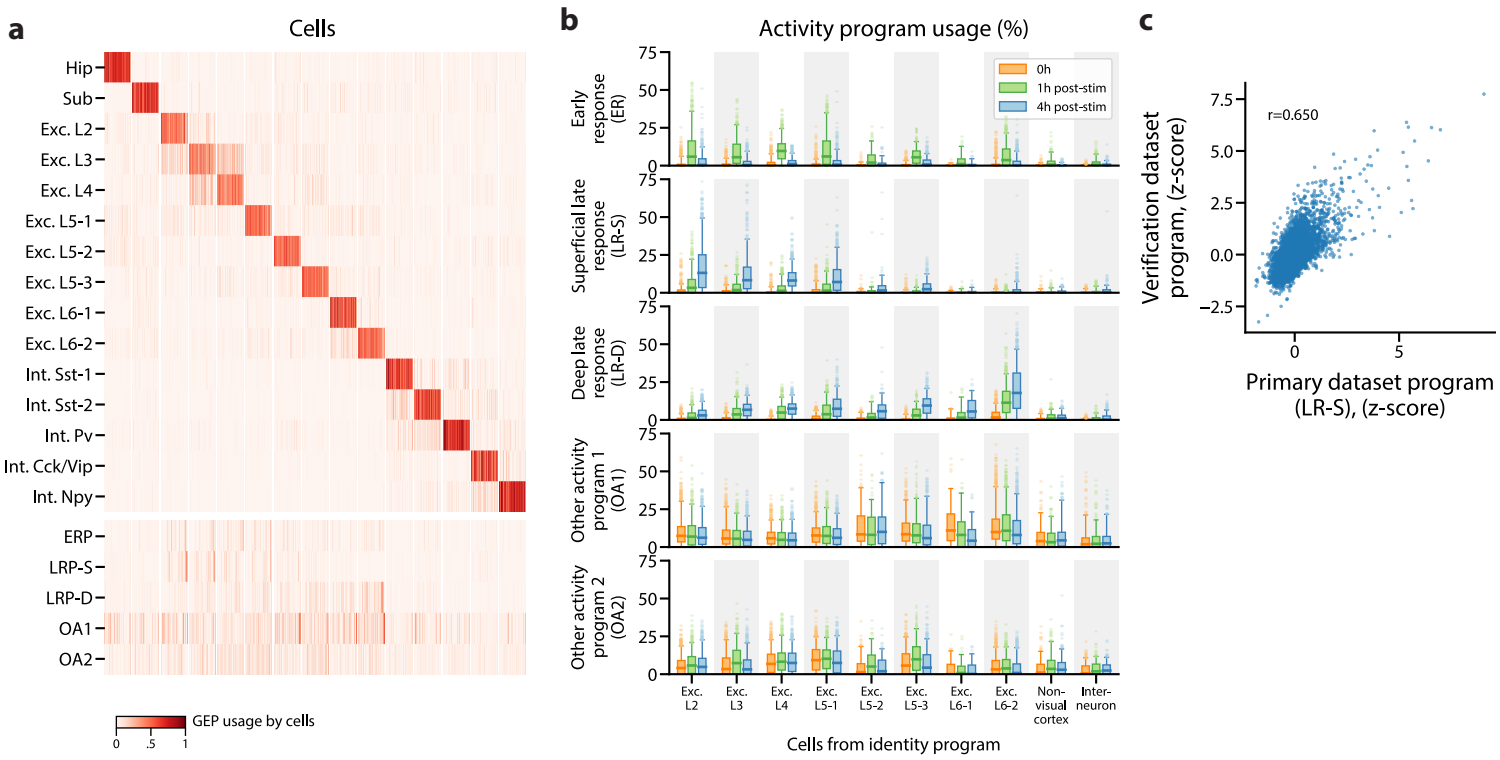


Figure 3: Identification of early and late activity induced transcriptional programs in neurons of the visual cortex. (a) Heatmap showing percent usage of all GEPs (rows) in all cells (columns). Identity GEPs are shown on top and activity GEPs are shown below. **(b)** Box and whisker plot showing the percent usage of activity programs (rows) in cells classified according to their maximum identity GEP (columns) stratified by the stimulus condition of the cells (hue). **(c)** Scatter-plot of Z-scores of genes in the superficial late response GEP from the primary visual cortex dataset against the corresponding program in the Tasic et. al., 2016 dataset for all overlapping genes.

consensus approach that increases the accuracy and robustness of several matrix factorizations, rigorously benchmarked them against each other and against state-of-the-art approaches, and found that our method, cNMF, provides the most accurate results. We demonstrate with simulated data that cNMF can increase the accuracy of cell-type identification while simultaneously characterizing activity programs that vary dynamically across cell-types. We then analyze several published datasets and illustrate how inferring activity programs can provide an additional layer of information on top of cell-types and can shed light on important biological phenomena such as depolarization-induced neuronal adaptation. As scRNA-Seq data further increases in RNA capture efficiency and throughput, it will likely become possible to detect more refined GEPs, further increasing our ability to disentangle activity and identity programs. Here, we have illustrated how this might be leveraged to understand the dynamic activities of cells within tissues.

Acknowledgements

We thank Allon Klein, Samuel Wollock, Aubrey Faust, Yakir Reshef, the CGTA discussion group, and members of the Sabeti Laboratory for useful discussions and feedback on the manuscript. We thank the Arlotta, Greenberg, and Zeng laboratories for generating the primary datasets we analyze in this manuscript. The project described was supported by award Number T32GM007753 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. DK was supported by NIH NIAID R01AI099210.

Author contributions

DK and AV conceived of the project, developed the method, analyzed the data, and wrote the manuscript with input from the other authors.

MAN provided crucial guidance in analyzing the visual cortex data.

ST helped with implementing early versions of the method.

EH, DAM, and PCS provided long-term guidance on the project.

Online methods

Simulations

Our simulation framework is based on Splatter[37] but is re-implemented in Python and adapted to allow simulation of doublets and shared gene-expression programs. Gene-expression programs were simulated as in Splatter. Cells were then randomly assigned an identity program with uniform proportions. 30% of cells of 4 cell-types were randomly selected to express the shared activity program at a usage uniformly distributed between 10% and 70%. Then their mean gene-expression was computed as the weighted sum of their cell-identity program and the activity program. Doublets were constructed by randomly selecting pairs of cells, summing their gene counts, and then randomly down-sampling the counts to the maximum of the two cells. We simulated 25984 genes, 1000 of which were associated with the activity program. The probability of a gene being differentially expressed in a given cell-identity program was set to 2.5%. The differential expression location and scale parameters were 1.5 and 1.0 for cell identity and activity programs. Other splatter parameters were: lib.loc=7.64, libscale=0.78, mean_rate=7.68, mean_shape=0.34, expoutprob=0.00286, expoutloc=6.15, expoutscale=0.49, diffexpdownprob=0, bcv_dispersion=0.448, bcv_dof=22.087. These values were inferred from 8000 randomly sampled cells of the Quadrato et al., 2017 organoid dataset using Splatter.

To compare against cNMF, we used Louvain clustering as implemented in scanpy[38] and determined corresponding GEPs as the cluster centroids. We used 15 principal components to compute distances between cells and used 15 nearest neighbors to define the KNN graph.

Concensus non-negative matrix factorization (cNMF)

We used non-negative matrix factorization implemented in scikit-learn (version 20.0) with the default parameters except for random initialization, tolerance for the stopping condition of 10^{-4}, and a maximum number of iterations of 400.

Each replicate of cNMF was run with a randomly selected seed. The component matrices from each replicate were concatenated into a single matrix where each row was a component from one replicate. Each of these components were normalized to have L2 norm of 1. Then components that had high average euclidean distance from their K nearest neighbors were filtered out. We set K to be .3 times the number of

bootstraps for the organoid and brain datasets and .5 times the number of bootstraps for the simulated dataset. The threshold on average euclidean distance was set by inspecting the histogram and truncating the long tail (Supplementary Fig. 1, 4, 7). Next the bootstrap-components were clustered using KMeans with euclidean distance and K set to the number of components used for the NMF run. Then each cluster was collapsed to a single component by taking the median across each of its dimensions. GEP components were then normalized to sum to 1. And lastly, a final usage matrix was fit by running one last iteration of NMF with the component matrix fixed to the median cluster components.

We determined the number of components for cNMF using the approach described in Alexandrov et al, 2013[19] with a few modifications. We ran NMF on normalized data matrices rather than count matrices and therefore did not resample counts but simply repeated NMF with different randomly selected seeds. We still determined the number of components by considering the trade-off between mean Frobenious error of the NMF repeats, and stability of the solutions from the distinct runs. As in Alexandrov et al, 2013, stability was computed as the Silhouette score of the KMeans clustering on the NMF components (prior to filtering outliers). However we used Euclidean distance on L2 normalized components as the metric rather than Cosine distance. Silhouette score was calculated using the Scikit-learn silhouette_score function.

Data preprocessing

For each of the datasets, we filtered out cells with fewer than 1000 unique umis detected. We also filtered out genes that were not detected in at least 1 out of 500 cells. Then we selected the 2000 genes with the most over-dispersion as determined by the v-score[39]. Then, each gene was scaled to unit variance before running cNMF. Note, we did not perform any TPM normalization prior to cNMF. This is because cells with more counts should contribute more information. Variance due to capture efficiency is captured in the Usage matrix rather than the GEP matrix. However, after cNMF, the GEP profiles was determined for all genes (including low variance ones) and the programs were converted to units of TPM. This was accomplished by running the last step of NMF with the full TPM cell x gene matrix as input, and the usage matrix fixed to that identified above.

Finding genes associated with programs

We associated genes with programs with Ordinary Least Squares Regression. We used the Usage matrix (cells x programs), row normalized to sum to 1 for each cell as the

predictor. We fit this to mean and variance normalized TPM profiles for each gene. The regression coefficients for a program therefore correspond to how many standard deviations above average a cell's expression would be expected to be if all of its usage derived from that program.

Testing enrichment of genesets in programs

We used the regression coefficients identified as above as inputs for a ranksum test (with tie correction) comparing the median of genes in each geneset to that of genes not in the geneset.

References

1. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34,** 166–176 (2003).
2. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 14863–14868 (1998).
3. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1,** 417–425 (2015).
4. The Tabula Muris Consortium, Quake, S. R., Wyss-Coray, T. & Darmanis, S. Transcriptomic characterization of 20 organs and tissues from mouse at single cell resolution creates a Tabula Muris. (2017). doi:10.1101/237446
5. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6,** (2017).
6. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).
7. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502

(2015).

8. Amir, E.-A. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31,** 545–552 (2013).
9. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11,** 740–742 (2014).
10. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34,** 1145–1160 (2016).
11. Scialdone, A. *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85,** 54–61 (2015).
12. Chen, M. & Zhou, X. Controlling for Confounding Effects in Single Cell RNA Sequencing Studies Using both Control and Target Genes. *Sci. Rep.* **7,** 13587 (2017).
13. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
14. Cardoso, J. F. Blind signal separation: statistical principles. *Proc. IEEE* **86,** 2009–2025 (1998).
15. Comon, P. Independent component analysis, A new concept? *Signal Processing* **36,** 287–314 (1994).
16. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **3,** 993–1022 (2003).
17. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791 (1999).

18. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Mach. Learn.* **52,** 91–118 (2003).
19. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3,** 246–259 (2013).
20. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008,** P10008 (2008).
21. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162,** 184–197 (2015).
22. Quadrato, G. *et al.* Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545,** 48–53 (2017).
23. Kelava, I. & Lancaster, M. A. Dishing out mini-brains: Current progress and future prospects in brain organoid research. *Dev. Biol.* **420,** 199–209 (2016).
24. Staudacher, J. J. *et al.* Hypoxia-induced gene expression results from selective mRNA partitioning to the endoplasmic reticulum. *Nucleic Acids Res.* **43,** 3219–3236 (2015).
25. Gardner, L. B. Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. *Mol. Cell. Biol.* **28,** 3729–3741 (2008).
26. Pawlikowski, B., Lee, L., Zuo, J. & Kramer, R. H. Analysis of human muscle stem cells reveals a differentiation-resistant progenitor cell population expressing Pax7 capable of self-renewal. *Dev. Dyn.* **238,** 138–149 (2009).

27. Hrvatin, S. *et al.* Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21,** 120–129 (2018).

28. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19,** 335–346 (2016).

29. Barbosa, A. C. *et al.* MEF2C, a transcription factor that facilitates learning and memory by negative regulation of synapse numbers and function. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 9391–9396 (2008).

30. Ramser, E. M. *et al.* The 14-3-3ζ protein binds to the cell adhesion molecule L1, promotes L1 phosphorylation by CKII and influences L1-dependent neurite outgrowth. *PLoS One* **5,** e13462 (2010).

31. Robbins, E. M. *et al.* SynCAM 1 adhesion dynamically regulates synapse number and impacts plasticity and learning. *Neuron* **68,** 894–906 (2010).

32. Hata, K., Maeno-Hikichi, Y., Yumoto, N., Burden, S. J. & Landmesser, L. T. Distinct Roles of Different Presynaptic and Postsynaptic NCAM Isoforms in Early Motoneuron-Myotube Interactions Required for Functional Synapse Formation. *J. Neurosci.* **38,** 498–510 (2018).

33. Aguirre-Chen, C., Bülow, H. E. & Kaprielian, Z. C. elegans bicd-1, homolog of the Drosophila dynein accessory factor Bicaudal D, regulates the branching of PVD sensory neuron dendrites. *Development* **138,** 507–518 (2011).

34. Yan, H., Yuan, J., Gao, L., Rao, J. & Hu, J. Long noncoding RNA MEG3 activation of p53 mediates ischemic neuronal death in stroke. *Neuroscience* **337,** 191–199 (2016).

35. Zhang, R. L. *et al.* E-selectin in focal cerebral ischemia and reperfusion in the rat. *J.*

Cereb. Blood Flow Metab. **16,** 1126–1136 (1996).

36. Gong, L. *et al.* RTN1-C mediates cerebral ischemia/reperfusion injury via ER stress and mitochondria-associated apoptosis pathways. *Cell Death Dis.* **8,** e3080 (2017).
37. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18,** 174 (2017).
38. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19,** 15 (2018).
39. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201 (2015).