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 and cellular processes is a crucial step toward understanding the organization of cells and tissues. Although single-cell RNA-Seq (scRNA-Seq) can quantify transcripts in individual cells, each cell's expression may derive both from programs determining cell-type and from programs facilitating dynamic cellular activities such as cell-division or apoptosis, which cannot be easily disentangled with current methods. Here, we introduce clustered non-negative matrix factorization (cNMF) as a solution to this problem. We show with simulations that it deconvolutes scRNA-Seq profiles into interpretable programs corresponding to both cell-types and cellular activities. Applied to published brain organoid and visual cortex datasets, cNMF refines the hierarchy of cell-types and identifies both expected (e.g. cell-cycle and hypoxia) and intriguing novel activity programs. In summary, we show that cNMF can increase the accuracy of cell-type identification while simultaneously inferring interpretable cellular activity programs in scRNA-Seq data, thus providing useful insight into how cells vary dynamically within 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 differentiation. 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. The development of scRNA-Seq avoids this problem by measuring the expression of thousands of individual cells simultaneously. This permits determination of the cell-types in the sample as well as changes to any of 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 identify 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 to overcome these challenges[6,7].

Here, we focus on one specific challenge of inferring GEPs from scRNA-Seq data, the fact that individual cells may express multiple gene expression programs which must be computationally disentangled. To see this, it is helpful to think about two broad classes of programs in scRNA-Seq data: (1) GEPs that determine identity as a specific cell-type such as a hepatocyte or melanocyte (identity programs) and (2) GEPs expressed in any cell that is carrying out a specific activity such as cell division or apoptosis (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 more types. Cells

undergoing the activity would therefore express the activity program in addition to their identity program.

Accounting for activity programs in scRNA-Seq data can be 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 may be interested in understanding how activity programs vary 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 sometimes removed computationally prior to clustering analysis[8,9]. Broadly, confounding of identity and activity programs is a more general case of this problem.

Here, we use non-negative matrix factorization (NMF) to directly model the fact that cells may express multiple programs. NMF is an algorithm for deconvoluting a mixed signal into its constituent sources[10]. In the context of scRNA-Seq, NMF represents each cell's expression profile as a weighted sum of multiple GEPs. These programs are shared across all cells in the dataset and are represented as a vector of expression values for each gene. NMF simultaneously learns the programs as well as their usage in each cell (I.e. what proportion of the cell's transcripts are attributed to each program). NMF assigns cells a high usage for the identity GEP corresponding to their cell-type as well as for the activity GEPs corresponding to any processes they are executing. This is in contrast to standard clustering where cells are effectively assigned one gene expression program that is shared by all other cells in the same cluster. An additional benefit of NMF for scRNA-Seq data is that it would likely assign doublet artifacts to have high usage of the two identity GEPs of the merged cells rather than arbitrarily assigning it to one of the two classes as in clustering.

An important caveat of NMF is that it is mathematically under-constrained and therefore, repeated runs can yield widely varying solutions. To address this, we adapt an approach from cancer mutational signatures discovery--running NMF multiple times on the dataset and clustering the results to obtain a final representation of the components (Fig. 1a, Online Methods)[11]. This approach, which we term clustered NMF (cNMF), is necessary to provide a robust final estimate of the GEPs. While mixed membership models such as NMF have been applied to bulk and single-cell RNA-Seq, these reports

did not directly address stochasticity of the solutions in this way nor did they distinguish between activity and identity programs[12–14].

We demonstrate the utility of cNMF for inferring GEPs using simulated scRNA-Seq data. We simulated 15,000 cells, 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) (online methods). We then examined its accuracy for identifying the cell-identity and activity programs, as well as doublets. First we selected the number of NMF components by comparing the trade-off between predictive accuracy and solution stability as described in Alexandrov et al[11] (Online methods). We found that this dataset was well-explained by 14 GEPs, as expected (Supplementary Fig. 1). cNMF yielded robust clusters of components that recapitulated the simulated GEPs (Fig. 1c-d). By contrast, low dimensional embeddings of the data with PCA or ICA yielded components that were mixtures of the true programs and were therefore less biologically interpretable than those identified by cNMF (Supplementary Fig. 2).

Furthermore, cNMF could infer cell-identity programs with greater accuracy than clustering for cell-types that expressed the activity program. We illustrate this using ordinary least squares regression to associate genes with programs defined by cNMF or by Louvain Clustering[15,16] (Online Methods). Louvain clustering accurately separated cells by cell-type but couldn't account for the mixed pattern of cells expressing the activity program (Supplementary Fig. 3a). As a result, it incorrectly associated activity program genes with the identity programs of the cell-types that execute the activity, leading to an elevated false discovery rate (Fig. 1e). However, cNMF and clustering recovered cell-identity programs with nearly identical accuracy for cell-types that did not express the activity program (Supplementary Fig. 3b). While we only compared cNMF with Louvain clustering, similar behavior would be expected for any form of hard clustering. Thus, our analysis suggests that cNMF can improve the accuracy of cell-type identification in scRNA-Seq datasets that contain significant signal from activity programs.

cNMF could also accurately infer which cells expressed the activity program (Fig. 1f). With an expression usage threshold of 10%, 1152 out of 1206 (95.5%) cells expressing the program were accurately classified while 10250 out of 10366 (99.1%) cells not-expressing the program were correctly classified. Moreover, we observed a high Pearson correlation between the inferred and simulated usages in cells that expressed the program (R=0.86). Thus, cNMF could infer which cells express an activity program, as well as what proportion of their transcripts derive from that program.

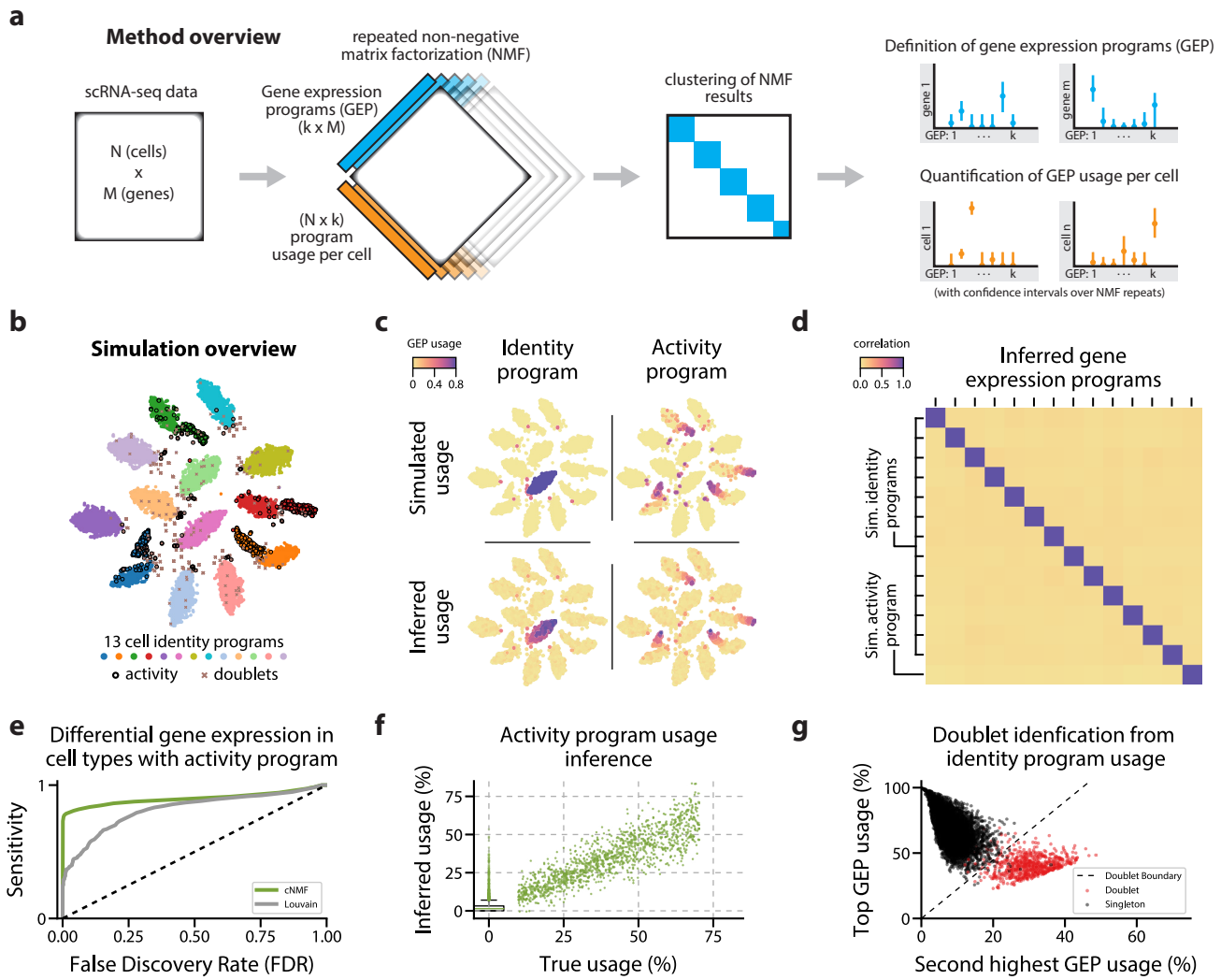


Figure 1: cNMF infers identity and activity expression programs in simulated data. (a) Schematic of the cNMF pipeline. **(b)** tSNE plot of the simulated data showing different cell-types with marker colors, doublets in gray, and cells expressing the activity GEP with a black edge. **(c)** Same tSNE plot as (b) but colored by the simulated true usage or the inferred usage of a sample identity program (left) or the activity program (right). **(d)** Pearson correlation between the true GEPs and the inferred GEPs, both in units of gene variance normalized TPM. **(e)** Receiver Operator Curve showing accuracy of prediction of genes associated with programs by cNMF or Louvain clustering (K=13) for a range of 500 evenly spaced Z-score thresholds. **(f)** 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 whisker corresponding to 5th and 95th percentiles. **(g)** Scatter 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. A decision boundary for classifying doublets is drawn as a dashed line.

Finally, we found that cNMF could accurately detect doublets, and could infer the two cell-types that contributed to each doublet (Fig. 1g). 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 less than three times the usage of it's second highest identity GEP. With this cutoff, 647/658 (98.3%) of doublets were correctly classified and 10226 out of 10336 (98.9%) of singletons were correctly classified. 609 out of 647 (94.3%) identified doublets had the correct two programs as their top two GEP usages. This illustrates how cNMF can be used to determine the cell-types that combine to form doublets.

Having demonstrated its utility on simulated data, we used cNMF to re-analyze a published scRNA-Seq dataset of 52,600 single-cells isolated from human brain organoids[17]. 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. 4). 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. 5). 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. 6).

Our identity programs further refined the cell-types identified in 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 organoids.

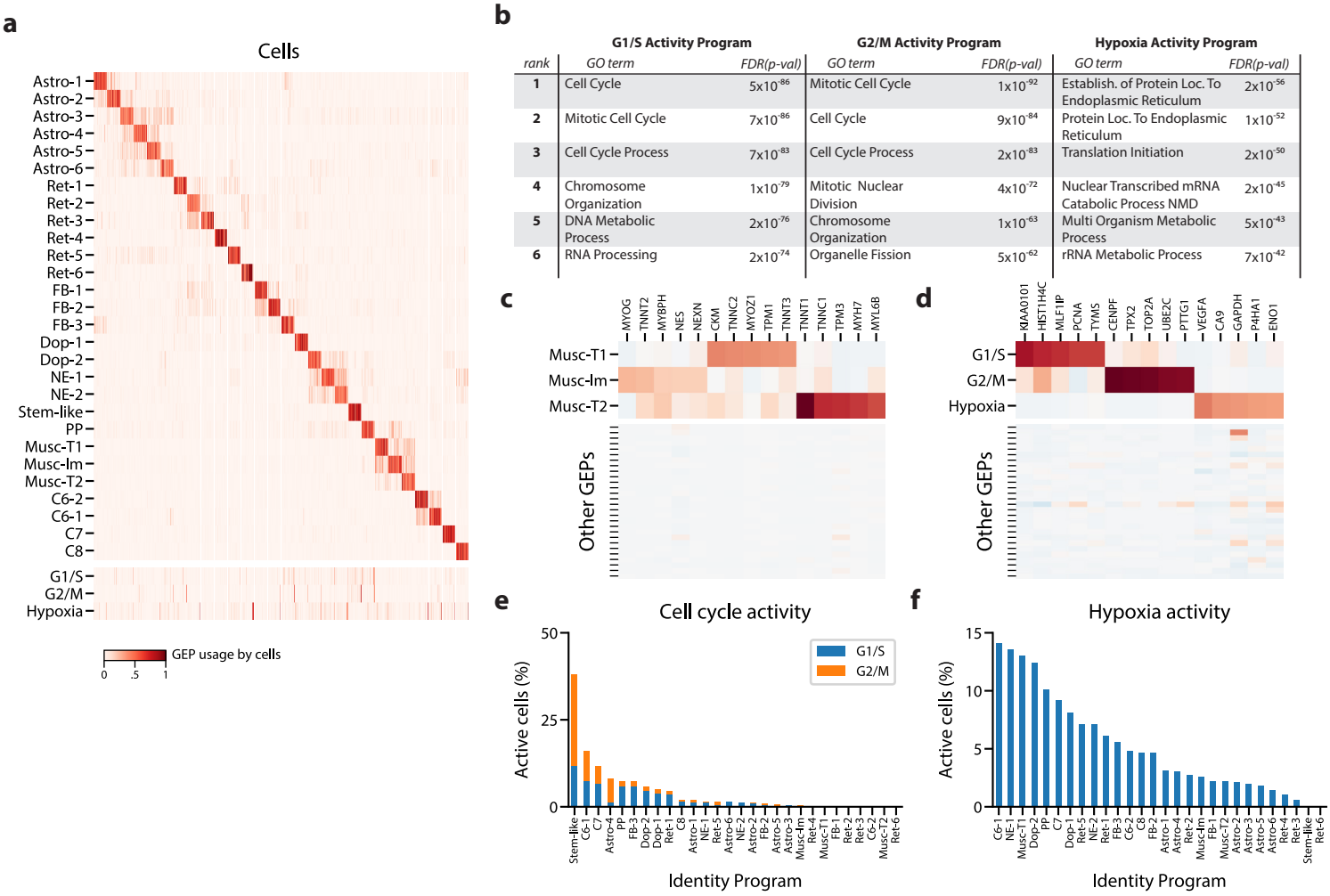
Next, we found that two of the activity programs were strikingly enriched for cell cycle Gene Ontology (GO) sets, suggesting that they correspond to separate phases of the

cell cycle (Fig. 2b). One showed stronger enrichment for genesets involved in DNA replication (e.g. DNA Replication $P=1x10^{-55}$ compared to $P=2x10^{-11}$) while the other showed stronger enrichment for genesets involved in mitosis (e.g. Sister Chromatid Segregation, $P=1x10^{-60}$ compared to $P=3x10^{-41}$). 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*, *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[18]. This GEP was significantly enriched for genesets related to protein localization to the endoplasmic reticulum and nonsense mediated decay ($P=1.6x10^{-56}$, $P=2.3x10^{-45}$) (Fig. 2b), consistent with literature showing that hypoxia post-transcriptionally increases expression of genes that are translated in the ER[19] and modulates nonsense mediated decay activity[20].

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. One rare cell-type which we label “stem-like” due to its high expression of pluripotency markers (e.g. *LIN28A*, *L1TD1*, *MIR302B*, *DNMT3B*) showed the highest rates of proliferation with over 38% of cells expressing a cell-cycle program (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. 6). However, the predominant identify GEP of cells in this cluster is most strongly associated with the gene PAX7, a marker of self-renewing muscle stem cells[21] (Supplementary table 1). 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 mitosis 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



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	5×10^{-86}	Mitotic Cell Cycle	1×10^{-92}	Establish. of Protein Loc. To Endoplasmic Reticulum	2×10^{-56}
2	Mitotic Cell Cycle	7×10^{-86}	Cell Cycle	9×10^{-84}	Protein Loc. To Endoplasmic Reticulum	1×10^{-52}
3	Cell Cycle Process	7×10^{-83}	Cell Cycle Process	2×10^{-83}	Translation Initiation	2×10^{-50}
4	Chromosome Organization	1×10^{-79}	Mitotic Nuclear Division	4×10^{-72}	Nuclear Transcribed mRNA Catabolic Process NMD	2×10^{-45}
5	DNA Metabolic Process	2×10^{-76}	Chromosome Organization	1×10^{-63}	Multi Organism Metabolic Process	5×10^{-43}
6	RNA Processing	2×10^{-74}	Organelle Fission	5×10^{-62}	rRNA Metabolic Process	7×10^{-42}

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 the top genes associated with three mesodermal programs in those programs (top) and in all other programs (bottom). **(d)** Heatmap of Z-scores of the 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%.

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[22]. 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. 7). As with 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. 8).

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 ($P=8.6x10^{-34}$ for the ERP and genes induced at 1H; $P=1.3x10^{-22}$, $P=2.5x10^{-13}$ 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[22]. 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.

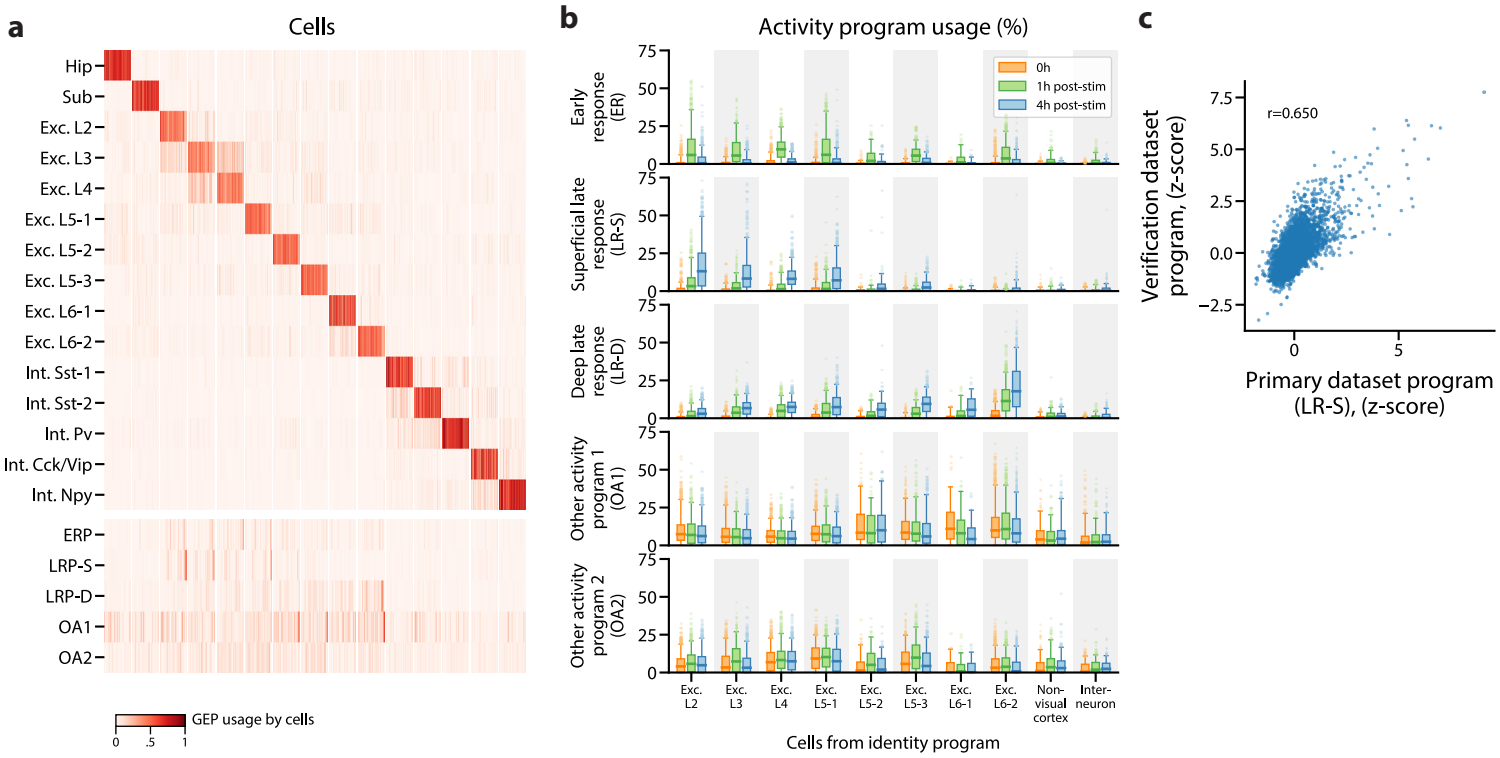


Figure 3: Identification of early and late activity induced transcriptional programs in neurons of the visual cortex. (a) Heatmap showing percent usage of all GEPs (rows) in all cells (columns). Identity GEPs are shown on top and activity GEPs are shown below. **(b)** Box and whisker plot showing the 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.

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[23]. In this dataset, cNMF identified a matching GEP for all visual cortex cell-types found in the original data except for one (Supplementary Fig. 9a). 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=104.0, $P=1.2x10^{-123}$, 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. 9b). 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*[24], *H2-Q4*, *YWHAZ*[25], *CADM1/SYNCAM1*[26], *NCAM1*[27], and *BICD1*[28] (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*[29], *ELAVL3*, *GLG1*[30], *GM20594*, *RTN1*[31]). 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 propose cNMF as a new approach for inferring gene expression programs in scRNA-Seq data. We show that it can increase the accuracy of cell-type identification while simultaneously characterizing activity programs that vary dynamically across cell-types. We illustrate with published data how inferring activity programs provides 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.

Methods

Simulations

Our simulation framework is based on Splatter[32] 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[33] 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.

Clustered 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[11] 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[34]. 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 the one identified as 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. Scialdone, A. *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85,** 54–61 (2015).
9. 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).

10. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791 (1999).
11. 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).
12. Dey, K. K., Hsiao, C. J. & Stephens, M. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet.* **13,** e1006599 (2017).
13. Shao, C. & Höfer, T. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics* **33,** 235–242 (2017).
14. Zhu, X., Ching, T., Pan, X., Weissman, S. M. & Garmire, L. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ* **5,** e2888 (2017).
15. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008,** P10008 (2008).
16. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162,** 184–197 (2015).
17. Quadrato, G. *et al.* Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545,** 48–53 (2017).
18. Kelava, I. & Lancaster, M. A. Dishing out mini-brains: Current progress and future prospects in brain organoid research. *Dev. Biol.* **420,** 199–209 (2016).
19. 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).

20. 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).
21. 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).
22. Hrvatin, S. *et al.* Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21,** 120–129 (2018).
23. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19,** 335–346 (2016).
24. 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).
25. 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).
26. Robbins, E. M. *et al.* SynCAM 1 adhesion dynamically regulates synapse number and impacts plasticity and learning. *Neuron* **68,** 894–906 (2010).
27. 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).

28. 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).
29. 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).
30. 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).
31. 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).
32. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18,** 174 (2017).
33. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19,** 15 (2018).
34. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201 (2015).