COTAN: Co-expression Table Analysis for scRNA-seq data

S. G. Galfrè^{*1}, F. Morandin^{*2}, M. Pietrosanto¹, F. Cremisi^{†3,4}, and
M. Helmer-Citterich^{†1}

¹Department of Biology, University of Rome Tor Vergata

²Department of Mathematical, Physical and Computer Sciences, University of
Parma

³Scuola Normale Superiore di Pisa

⁴Institute of Biophysics, Research National Council of Pisa

May 11, 2020

Abstract

Estimating co-expression of cell identity factors in single-cell transcriptomes is crucial to decode new mechanisms of cell state transition. Due to the intrinsic low efficiency of single-cell mRNA profiling, novel computational approaches are required to accurately infer gene co-expression in a cell population. We introduce COTAN, a statistical and computational method to analyze the co-expression of gene pairs at single cell level, providing the foundation for single-cell gene interactome analysis.

1 Introduction

Single-cell transcriptomes can describe known cell identity states and uncover new ones. This is basically achieved by methods of analysis of correlation of RNA counts among cells and subsequent clustering of cells with consistent global gene expression^{1,2}. The typical pipeline requires to normalize and log-transform raw read counts, yielding expression levels, and perform multivariate analysis^{3,4} on the latter. Unfortunately, the intrinsic low efficiency of scRNA-seq⁵⁻⁷ precludes the detection of many expressed genes in single cells, in particular in droplet-based experiments. This has a critical effect on expression levels, causing the appearance of dropouts artefacts^{8,9}, and often limiting the scope of the research to few tools based

---

^{*}Both authors contributed equally to this work.

[†]Both authors contributed equally to this work.

on zero-inflation and imputation^{10–12}. However, the introduction of Unique Molecular Identifiers¹³ (UMI) greatly reduces amplification noise, and the resulting UMI counts typically fit simple probabilistic models, thus allowing approaches not based on normalization⁸. To exploit this feature with a focus on genes joint distribution, we introduce CO-expression Tables ANalysis (COTAN), which uses UMI count matrices without normalization and does not depend on zero-inflation.

COTAN is based on three main elements: a robust estimation of the UMI detection efficiency (UDE) of each cell, a flexible model for the probability of zero UMI counts, and a generalized contingency table framework for zero/non-zero UMI counts for couples of genes. In fact, COTAN estimates the co-expression of gene pairs by comparing the number of cells that have zero UMI counts for both genes, with the expected value under independence hypothesis. This approach is based on the fact that technical zeros are always independently distributed, while biological zeros may be correlated for genes associated to cell differentiation, providing a way to recover information on the joint distribution.

To estimate the model parameters, COTAN uses raw UMI counts, but then, for computing co-expressions these are coded as zero/non-zero. This choice is a key feature of COTAN and has the twofold aim to increase the sensitivity of the method for genes with low expression level and to better handle genes whose counts deviate from the chosen parametrical distributions.

COTAN can effectively assess co-expression of gene pairs, yielding for each pair a numerical index and an approximate p -value; it can investigate whether single genes are differentially expressed in the population, by scoring them with a global index of differentiation; finally, it provides ways to plot and cluster genes according to their co-expression pattern with other genes, effectively helping the study of gene interactions.

Though the paper is mostly self-contained, with a significant section of methods, the mathematical models are only drafted here: proofs and finer details can be found in the twin paper¹⁴.

2 Results

2.1 Main features

COTAN analysis is performed on post-quality-control UMI counts. It starts with the estimation of the model parameters: UDE for cells (denoted by ν_c in the Methods Section), and mean (λ_g) and dispersion (a_g) for the expression of genes. While UDEs and means are estimated by moment methods, for dispersions we fit the number of cells showing zero UMI counts with its expected value under the negative binomial model. This is done to improve the method robustness, as afterwards we only use dispersions to compute

the probability of zero UMI counts. These probabilities allow to devise an independence test for the expression of two genes (gene-pair analysis, GPA). The test is based on generalized 2×2 contingency tables (also indicated as co-expression tables) which collect the joint occurrence of zero UMI counts for two genes.

COTAN provides an approximate p -value for the GPA test, and a signed co-expression index (COEX), which measures the direction and significance of the deviation from the independence hypothesis. Positive values denote joint expression of two genes, negative values denote disjoint expression and values near 0 are expected when one or both are constitutive genes, or when the statistical power is too low (Figure 1A, 2B).

When GPA is performed genome-wide, it is possible to score genes according to how often their expression shows dependence on the expression of other genes, identifying cell-identity genes. To this purpose, we define the global differentiation index (GDI) of a gene g as a normalization of the average of the 5% lowest p -values obtained by GPA tests of g against all other genes.

We tested our method on neural development datasets. Indeed, brain embryonic structures display high cell diversity, with dividing multipotent progenitor cells, newborn neurons differentiating with many distinct identities and glial cells co-existing in a mixed cell population. This feature makes them particularly suited for scRNA-seq studies aiming to depict cell state identity and relationships between gene expression. We assayed COTAN on a scRNA-seq dataset of embryonic hippocampus¹⁶, focusing on a collection of selected constitutive genes (CGs¹⁷), neural progenitor genes (NPGs^{18,19}) and pan-neuronal genes (PNGs^{18,19}). COTAN's GPA effectively discriminated between CGs, showing COEX near zero against all genes, and NPGs or PNGs, having positive or negative COEX when tested against one another (Figure 1A). Notably, each NPG positively correlated with other NPGs and negatively with PNGs, and vice-versa, indicating COTAN capability to correctly infer joint or disjoint expression of two genes at single cell level.

We compared COEX to Pearson correlation computing gene expression levels by Seurat¹⁵. COEX proved more accurate in discriminating between CGs, NPGs and PNGs, indicating COTAN as better suited in analyzing the co-expression of couples of genes at single cell level (compare Figure 1A with B). Notably, some CGs (e.g. *Calm1*, *Actb*, *Gapdh*) showed specific, high co-expression or disjoint expression with NPGs or PNGs when analyzed with Pearson correlation (Figure 1B), making them false positive markers of differentiated cell identity. We then compared GDI to the highly variable feature analysis of Seurat¹⁵. GDI efficiently discriminates between CGs, which lay below the median (with two exceptions, *Taf1* and *Zfr*), and NPGs and PNGs, located above the third quartile (Figure 1C). Instead, highly variable features analysis of Seurat (Figure 1D) was much less precise in discriminating between CGs and cell identity genes (compare Figure 1D

Figure 1. GPA and GDI are able to discriminate constitutive genes (CGs) from neural progenitor genes (NPGs) and pan-neuronal genes (PNGs). **A.** COTAN GPA of selected genes. Cell color encodes COEX index: blue indicates genes showing joint expression, red indicates genes showing disjoint expression. White indicates independence, meaning that one or both genes are constitutive, or that the statistical power is too low to detect co-expression (See Methods). Since the co-expression of a gene with itself is irrelevant, the diagonal is made artificially white. **B.** Pearson correlation matrix of the same selected genes as in (A), using Seurat¹⁵ normalized expression levels (obtained following the website vignettes – *Guided Clustering Tutorial*). **C, D.** Comparison between COTAN global differentiation index (GDI, C) and Seurat *highly variable features* (D) analysis. Red labels indicate NPGs, orange labels PNGs, green labels CGs. Dotted blue lines correspond to the median (lowest) and the third quartile (highest). All plots refer to E16.5 mouse hippocampal cells¹⁶ and genes are selected to be characteristic of NPGs, PNGs and constitutive genes with both high and low typical expression.

to 1C).

Correlation analysis approaches are commonly used to identify cell clus-

ters with consistent global gene expression. Conversely, we used COTAN to determine clusters of genes characterized by their co-expression patterns with selected genes in single cell. We chose to investigate mouse embryonic cortex²⁰ because the molecular identity of a number of layer-specific precursor cells and neurons is known¹⁸. We firstly chose from literature¹⁸ robust primary markers for layer I (*Reln*), layers II/III (*Satb2*), layer IV (*Sox5*) and layer V/VI (*Bcl11b*), see Figure 2A. Then, for each marker we selected the most positively associated genes (COEX > 0 and GPA *p*-value < 0.0001 for *Satb2*, *Reln* and *Sox5* or < 0.001 for *Bcl11b*, for a total of 248 genes). For all these genes, we plotted an ordered symmetric heatmap of GPA COEX values with the genes grouped by the marker by which they were selected (Figure 2B). COTAN showed to be well suited to evaluate the co-expression of gene pairs genome-wide. The comparison between groups highlighted impressive consistency of co-expression inside each group and robust disjoint expression between different groups, with the only exception of *Sox5* and *Bcl11b* groups, which resulted as co-expressed. We believe that the *Reln*, *Satb2* and *Sox5/Bcl11b* groups represent genuine gene signatures of distinct cortical cell identity and that similar signatures can be found by unbiased approaches.

Eventually, we assayed the ability of our approach to cluster groups of co-expressed genes regarding cell layer identity (Figure 2C-E). In analogy to other methods²² the analysis was guided by few key genes. We chose ten primary layer identity markers (Greig et al 2013¹⁸): *Reln* and *Lhx5* for layer I (dark red), *Satb2* and *Cux1* for layer II/III (pink), *Rorb* and *Sox5* for layer IV (pale blue), *Fezf2* and *Bcl11b* for layer V/VI (dark blue), *Hes1* and *Vim* for neural progenitors (cyan). We then selected for each primary marker the 25 genes with the lowest GPA *p*-values. In this way we collected a total of 215 secondary markers (35 genes were shared by two primary markers). These were used as variables/features for the dimensionality reduction and cluster analysis. A normalization of COEX served as input data for the analysis (for details see Section 2.3 and 2.4). Figures 2C-E show PCA, t-SNE and MDS, respectively, of the 10% genes with highest localized GDI ($n = 1234$). The dots are colored according to a hierarchical clustering on the same input data (Section 2.5). Notably, each cluster shows univocal correspondence with all the primary markers of one of the major cortical cell identities at the developmental stage of analysis, proving COTAN ability to gather genes with similar nature regarding cell identity.

Eventually, this analysis highlighted a larger number of novel markers of cortical layer identity compared to the ones identified by conventional analysis.

Figure 2. Gene clustering in scRNA-seq analysis. **A.** The six layers of differentiated neurons and progenitor cells of late embryonic cortex are depicted in different colors together with known markers of cell identity^{18,21}. **B.** GPA heatmap of the 248 genes showing strong joint expression with the genes indicated in labels: *Reln*, *Satb2*, *Sox5* and *Bcl11b* respectively markers of layers I, II/III, IV and V/VI. The heatmap shows the reciprocal relationship between all genes pairs; significant joint expression is indicated with blue (positive COEX values) while significant disjoint expression is indicated in red (negative COEX values). **C, D** and **E.** Different dimensionality reduction plots (Principal Component Analysis, t-distributed Stochastic Neighbour Embedding and Multidimensional Scaling, respectively) of 1234 genes (10% highest localized GDI). The primary markers used to set the variables are labeled. Dot colors denote the gene cluster corresponding to a layer or progenitor gene as specified in Section 2.5. Labels corresponding to other known layer identity markers (*Tbr1*, *Mef2c*, *Nes* and *Sox2*) that were not used as primary markers are also shown as additional landmarks. All plots refer to E17.5 mouse cortex cells²⁰.

2.2 Negative and synthetic dataset analysis

COTAN is a tool to study cell differentiation in scRNA-seq datasets. As such, it is important to assess its behaviour under the null hypothesis that there is no differentiation at all, i.e. on datasets that happen to be completely homogeneous. We refer to genes as constitutive if they are assumed to be transcribed in all cells of a population: their UMI counts may still be zero in some cells, but that would be because of random sampling. Therefore for homogeneous datasets all genes should be constitutive, while for

Figure 3. Empirical cumulative distribution functions of negative datasets GPA p -value. Points are a random sample from all $\binom{n}{2}$ gene pairs. The plots are close to the identity function, so this statistics approximates a rigorous p -value. The fraction of expected false positives at significance level α is about $F(\alpha)$. Since deviations from identity are negative, false positive are α or less. **A.** ERCC mRNA dataset formed by 92 different mRNAs and 1015 droplets⁷; **B.** CD14+ cell dataset formed by 11405 genes (before cleaning) and 2370 cells⁷; **C, D.** Homogeneous synthetic datasets with 800 or 4000 cells.

normal multiple cell-type populations, only some of them are. For embryonic cortex datasets in this paper a number of constitutive genes selected from literature¹⁷ were included in our analysis.

We tested COTAN on three different types of homogeneous datasets: 1) a technical negative dataset formed by External RNA Controls Consortium mRNA (ERCC⁷), 2) a biological negative dataset formed by FACS sorted CD14+ cells⁷ and 3) two synthetic datasets with a single cell type (see Section 3.5).

We analyzed GPA and GDI on these datasets. Figure 3 shows that GPA's p -value is a good approximation for the true p -value of the independence test. If the test is formally performed as usual by comparing the p -values with some given significance level, the false positives can be controlled efficiently.

COTAN's GDI, applied on negative datasets (Figure 4), shows no unwanted dependence from total gene expression, and only few genes exceed the empirical threshold value 1.5 for differential expression, giving less than

Figure 5. GPA COEX heatmap for the 53 genes with GDI greater than 1.5 in the CD14+ cell negative biological dataset. Most genes associated by positive co-expression are coherent with the existence of a dendritic cell subpopulation, and therefore their high GDI does not represent a false positive. Genes marked with * are those found by Enrichr (see text).

in the CD14+ cell dataset, in agreement with our findings.

2.3 Gene co-expression space plots

Because COTAN was specifically developed to analyze joint or disjoint expression of two genes, it is possible to analyze which genes present similar behaviour in the dataset. The basic idea is to select a set V of informative genes (as variables/features) and then use COTAN to represent all genes, genome wide, according to their pattern of co-expression with respect to those in V . This multivariate representation allows then to cluster genes by their co-expression pattern and, after a dimensionality reduction, to plot them on the plane.

We chose a comprehensive set V of layer-associated genes to be used as variables (Figure 6). To build V we selected a set of ten known primary markers of cortical layer identity (*Reln* and *Lhx5* for layer 1, *Satb2* and *Cux1* for layer 2/3, *Rorb* and *Sox5* for layer 4, *Bcl11b* and *Foxf2* for layer 5/6 and *Vim* and *Hes1* for progenitor cells), together with the top 25 genes most correlated with each of them, for a total of 215 secondary markers (this number is lower than 250 because some of these genes were shared).

Figure 6. Schematic of the procedure to build the sets of genes V and U used for dimensionality reduction analyzes (see text).

For all genes in the dataset, we computed the *localized GDI* relative to the genes in V (see Section 2.4 below) and used it to filter only the 10% mostly differentially expressed genes, with the highest values, in order to get a meaningful graphical representation and better input data for the subsequent cluster analysis (Figure 6). These selected genes formed the set U of units/points to represent.

Then, as described in Section 3.6, we filled the matrix $U \times V$, applied the normalization $\tanh(\sqrt{m} \text{COEX})$, yielding the gene co-expression space, and finally performed three different dimensionality reduction analyzes on the resulting index: PCA (Figure 2C), t-SNE (Figure 2D) and MDS (Figure 2E). t-SNE plot was performed using 20 initial dimensions, 3000 maximum iterations, perplexity 30, eta = 200 and theta = 0.4. Genes were colored according to the cluster analysis results (see Section 2.5).

2.4 Localized GDI

Global differentiation index (GDI) as defined in Section 3.4 is a method to score genes according to their propensity to show statistically significant joint or disjoint expression against *all* other genes. Although this is not a rigorous test of differential expression, we remark that values around 1.3 are typical for average-expression, constitutive genes, while values above 1.5 are evidence of differential expression. To make this tool more specific, it is possible to restrict the set of genes against which GDI is computed to a selected subset V , with the recommendation to include a consistent fraction of cell-identity genes (for instance markers of layer identity). In this case the index is referred to as GDI localized to V .

Localized GDI may uncover information that is hidden or confounded in the genome-wide GDI. We stress that the differentiation threshold 1.5 is referred to the latter, and that the former should be used only to score genes and not to infer whether they are in fact constitutive for the cell population or not (Figure 7).

2.5 Gene clustering

We performed a cluster analysis on the 10% genes selected in Section 2.3, using their co-expression space representation (see Section 3.6) as an input for a hierarchical clustering with the Ward's method. The resulting tree presents a natural cutting distance at seven clusters (possible alternatives being at 2, 3 or 5 clusters – see Figure 8).

In five of the seven clusters we found the five pairs of primary markers, each one in its own. From them we assigned the identity of the five clusters and in particular: cluster 1, containing *Vim* and *Hes1*, was identified as genes related to progenitors identity, cluster 2, containing *Sox5* and *Rorb*, was identified as genes related to layer IV identity, cluster 5, containing *Cux1* and *Satb2*, was identified as genes related to layer II/III identity, cluster 6, containing *Reln* and *Lhx5*, was identified as genes related to layer I identity and, finally, cluster 7, containing *Bcl11b* and *Fezf2*, was identified as genes related to layer V/VI identity.

To assess the ability of COTAN gene clustering to detect new markers we compared our clusters with the literature²⁵. Results are summarized in

Figure 7. Comparison between genome-wide and localized GDI for all genes of E17.5 mouse cortex²⁰. Labels and respective colors as in Figure 2, in particular only primary markers selected for layer identity together with landmark layer identity genes are labelled. Colored dots are those secondary markers (i.e. the set V of Section 2.3) that have positive COEX against a primary marker, with which it shares the color. Dotted red line is the identity function. Localized GDI has typically higher values than global GDI, and its right tail is more spread. Notably, many informative genes of layer identity have intermediate global GDI, below the differentiation threshold 1.5. It would be difficult to detect them without localized GDI. The vertical line corresponds to the 90-percentile of the localized GDI and its right side identifies the 10% genes used for cluster analysis.

Tables 1 and 2.

We can observe from the tables that all markers used or identified by Loo et al.²⁵ are also detected by COTAN clustering and the agreement in the layer assignment is quite good. There are only 4 genes out of 26 that have been assigned to different layers. *Plxna4*, which is the only one with known cortical expression pattern in Allen Brain Atlas database, is recognized by Loo et al. as marker for layer V/VI at stage E14.5, and by our analysis as marker for layer II/III, at stage E17.5. Accordingly, ISH Allen Brain Atlas in Figure 9 shows that this protein is actually localized principally in layer V/VI at early stages, but it co-localizes with layer II/III at later stages.

Notably, COTAN identified a much higher number of layer markers compared to the conventional methods applied by Loo et al. (see supplement-

Figure 8. Hierarchical clustering of genes. The dotted line denotes the height of the tree cut forming seven clusters. Branches and leaves colors indicate cluster identity: cluster 1, in cyan, indicates progenitor identity, cluster 2, in light blue, indicates layer IV identity, cluster 5, in pink, denotes layer II/III identity, cluster 6, in dark red, indicates layer I identity and cluster 7, in dark blue, indicates layer V/VI identity. The two clusters in gray (3 and 4) do not contain primary markers and are likely inconsistent with projection neuron identity (see text). The gene names reported are the ones identified as secondary markers (see Section 2.3); among these, the ones initially linked to a primary marker not coherent with the final cluster are boxed with the color related to their initial marker.

tary file **STable1.csv**). Among all possible new layer markers detected by COTAN, we analysed the ones presenting nucleic acid binding gene ontology (GO:0003676). Complete tables are attached as supplementary files (**STable1.csv** and **STable2.csv**). Figure 10 shows the E18.5 ISH collection of the genes available from Allen Brain Atlas website. Most of the genes

Figure 9. In Situ Hybridizations data from Allen Brain Atlas website²⁶ for mouse cortex at E15.5 and E18.5. It is evident that *Plxna4* protein at early stage is principally localized in layers V/VI, while at later stages it is more present in superficial layers, in accordance with both our finding and the work from Loo et al.²⁵

Figure 10. E18.5 ISH from Allen Brain Atlas website²⁶ of markers with nucleic acid binding gene ontology identified by COTAN. In red, markers detected also by conventional analysis are indicated.

show ISH pattern consistent with layer identity as identified by COTAN, with few exceptions.

Among the seven natural clusters detected by hierarchical clustering, two clusters do not contain any primary markers. To assess their identity we performed a gene enrichment analysis on the Enrichr web site^{23,24}. We found that cluster 3 is enriched in septofimbrial nucleus genes, and cluster 4 is enriched in accumbens nucleus genes (in the Allen Brain Atlas up-regulated genes dataset – data tables attached as supplementary files).

		Markers from Loo et al. ²⁵			Prog
		Layer I	Layer II/III	Layer V/VI	
Markers detected by COTAN	Layer I	4			
	Layer II/III		2	1	
	Layer V/VI			9	
	Prog				7
	Other cluster			2	1

Table 1. Number of layer markers found by Loo et al.²⁵ with their respective layer according to the original paper (columns) and according to COTAN (rows). Bold text denotes consistent identification by the two methods. *Plxn4* in purple, see text.

		Markers from Loo et al. ²⁵			Prog
		Layer I	Layer II/III	Layer V/VI	
Markers detected by COTAN	Layer I				
	Layer II/III		1	1	
	Layer V/VI			5	
	Prog				4
	Other cluster			2	1

Table 2. Same as Table 1, but excluding all genes belonging to the set V of secondary markers, as these were selected by co-expression with the primary markers and hence their assignment to the correct clusters might be favored by the method.

3 Methods

3.1 Mathematical framework

We refer for all the details to the dedicated paper¹⁴. We resume here the main aspects of the mathematical framework.

3.1.1 Model and parameters estimation

For each gene g and cell c , let $R_{g,c}$ denote the UMI count and $\Lambda_g^{(c)}$ the corresponding biological expression¹. Conditional on the value of the latter, the read count has $\text{Poisson}(\nu_c \Lambda_g^{(c)})$ distribution, where ν_c denotes the UMI detection efficiency (UDE) of cell c , which is not supposed to depend on the gene (the pipeline in Section 3.7 includes a step to check this assumption). Then the expected UMI count is given by $\mu_{g,c} := E(R_{g,c}) = \nu_c E(\Lambda_g^{(c)}) =: \nu_c \lambda_g$. We introduce two different frameworks for estimating these parameters: *linear* and *square-root*.

¹The biological expression is here considered random, with unknown distribution of mean λ_g , and measured not in real molecules, but in expected reads, taking the extraction efficiency into account.

With the symbol $*$ used to denote averaging along all the values of the corresponding index, and under the convenient assumption that $\nu_* = 1$, then $R_{g,*}$, $R_{*,*}$ and $R_{*,c}$ are unbiased estimators of λ_g , λ_* and $\nu_c \lambda_*$, and $\frac{R_{*,c}}{R_{*,*}}$, is an estimator of ν_c . These are called the *linear estimators* and may often be used with good results, though sometimes they may not be very robust, when there is a large biological variability of the highly expressed genes. (In particular, the estimator of ν_c may end up being correlated with the expression of these genes.)

A possible alternative is using the variance-stabilizing transformation $X_{g,c} = \sqrt{R_{g,c}}$ which yields the *square-root estimators*, defined as follows:

$$\begin{aligned}\tilde{\lambda}_g &:= \psi(X_{g,*}) + \frac{1}{2}\psi''(X_{g,*}) \cdot \left[\frac{m-1}{m} S_{g,*}^2 - \psi(X_{g,*}) + X_{g,*}^2 \right], & g \in G \\ \tilde{\nu}_c &:= \frac{\tilde{\nu}_c}{\tilde{\nu}_*} := \frac{\tilde{\nu}_c}{\frac{1}{m} \sum_{u \in C} \tilde{\nu}_u}, & c \in C \\ \tilde{\mu}_{g,c} &:= \tilde{\lambda}_g \tilde{\nu}_c,\end{aligned}$$

where ψ is the inverse of the function $x \mapsto \sum_{k \geq 1} \sqrt{k} \frac{x^k}{k!} e^{-x}$ on $[0, \infty)$, where $S_{g,*}^2$ denotes the sample variance of $X_{g,c}$ for fixed g , and where the “tilde” estimator of ν is the analogous for cells of the “check” estimator for λ , so:

$$\tilde{\nu}_c := \psi(X_{*,c}) + \frac{1}{2}\psi''(X_{*,c}) \cdot \left[\frac{n-1}{n} S_{*,c}^2 - \psi(X_{*,c}) + X_{*,c}^2 \right], \quad c \in C,$$

We remark that $\psi(x) \approx x^2$ and in particular the above formulas are just a second order correction to the naive attempt at commuting of the square root and the average.

Accuracy and precision of these estimators were tested on synthetic datasets with heterogeneous cell types, for which the true values of ν and λ were known (Figure 11 and Section 3.5).

3.1.2 Probability of zero UMI count

Whichever the estimator of $\mu_{g,c} = \lambda_g \nu_c$, its value is the starting point to approximate the probability that $R_{g,c} = 0$. In fact by the Poisson conditional distribution of the UMI counts, we get

$$P(R_{g,c} = 0) = E[e^{-\nu_c \Lambda_g^{(c)}}] =: e^{-\eta_g(\mu_{g,c})} \quad (1)$$

where η_g denotes the log-mgf function of $\Lambda_g^{(c)}/\lambda_g$. We don't fix a particular model for the distribution of either $\Lambda_g^{(c)}/\lambda_g$ or $R_{g,c}$, but we make the

Figure 11. Violin plots of linear and square-root estimator error. The six synthetic datasets differ for the number of cell types and the total number of cells. Homogeneous datasets are for reference. In the legend 10X and DSQ (Drop-seq) refer to the method of the real dataset used as a source. H and M indicate if the dataset is homogeneous or with multiple cell types; 800 and 4000 refer to the number of cells simulated, and *sqrt* and *linear* indicate the type of estimation used for ν and λ .

assumption that at least the probability mass in zero for $R_{g,c}$ can be approximated by a negative binomial model with mean $\mu_{g,c}$ and some dispersion a_g depending only on the gene. Then equation (1) becomes

$$P(R_{g,c} = 0) \approx \left(\frac{a_g^{-1}}{\mu_{g,c} + a_g^{-1}} \right)^{a_g^{-1}}.$$

The value of a_g is estimated by imposing the marginal condition that the expected number of cells with zero UMI count for the gene, $\sum_c e^{-\eta_g(\mu_{g,c})}$, equals their total number: with $\#$ denoting cardinality,

$$\sum_c \left(\frac{a_g^{-1}}{\mu_{g,c} + a_g^{-1}} \right)^{a_g^{-1}} = \#\{c, R_{g,c} = 0\}.$$

In the case of the genes that would require a negative dispersion, because $\sum_c e^{-\mu_{g,c}} > \#\{c, R_{g,c} = 0\}$, we choose instead to impose a zero dispersion model (Poisson distribution) with *increased mean* $(1 + b_g)\lambda_g$, yielding $P(R_{g,c} = 0) \approx e^{-(1+b_g)\mu_{g,c}}$. This choice is consistent with the intended universality of the approach, because no distribution of $\Lambda_g^{(c)}$ would give a negative dispersion and because λ_g is anyway an estimated quantity and therefore noisy. In the typical dataset about 30% of all genes fall in this case, with values of b_g no larger than 0.02 in average and with no single value higher than 0.15. In the implementation, a_g and b_g are encoded in a single parameter (Section 3.7.3).

3.2 Co-expression tables

COTAN at its core is a tool for measuring co-expression of two genes. In this section we take advantage of the estimators introduced above, and show how to create and test the co-expression tables on which COTAN’s *gene pair analysis* (GPA) is based.

Gene co-expression is a powerful tool to study cell differentiation²⁷. When the population of cells is completely homogeneous, each gene is assumed to be either expressed in all cells or not expressed in any. In this situation the UMI counts of the expressed genes should be independent random variables across the cells. On the other hand, if the population is non-homogeneous and there are different cell types, each one expressing its own genes, then two genes could have non-zero UMI counts in the same cells, more (or less) often than should be expected if the population was homogeneous.

COTAN is built on the biological assumption that cell differentiation will typically *shun to zero* the expression of several genes, and that UMI count fold change is not as much informative as the switch of a gene from present to absent (or vice versa). In fact fold change is also difficult to assess for genes with very low expression, which are the most common in scRNA-seq. Significantly, in these datasets several constitutive genes have very few UMI counts (Figure 12). However, being constitutive, their low counts cannot be considered indication for absent or negligible gene expression. Consequently, for low-expression genes the main information is whether their UMI count is positive or zero.

From a mathematical point of view, moreover, the very low typical expression of most genes is also the main cause of dropout artefacts and more in general of the difficulties in modelling gene expression levels^{8,9}. In fact log-normalization, even with pseudocounts, essentially assumes to act on larger numbers, so that they may be thought of as continuous variables, but scRNA-seq datasets exhibit instead discrete counts, typically with 95% counts of 2 or less and 98% of 4 or less.

Based on the previous biological assumption, and on the mathematical need to deal with the discrete UMI count variable, we choose to collapse the values $R_{g,c}$ in the two categories $\{R_{g,c} = 0\}$ and $\{R_{g,c} \geq 1\}$, neglecting to consider values larger than 1 as different. COTAN GPA main test compares the number of cells with zero read count in couples of genes (jointly versus marginally), in a way similar to 2×2 contingency tables, but generalized to experimental units with different UDE.

For example consider the constitutive genes *Cox7c* and *Tars* in the following case:

Figure 12. Violin plots of UMI counts + 1, with y -axis in logarithmic scale. On the left column (**A**, **B**, **C**) plots of a Drop-seq mouse cortex dataset²⁰ at four development times (E11.5, E13.5, E15.5 and E17.5). On the right column (**D**, **E**, **F**) plots of a 10X dataset¹⁶ showing the development of mouse dentate gyrus (E16.5, P0 and P5). **A**, **D**. Constitutive genes. **B**, **E**. Selected neural markers. **C**, **F**. Cell identity transcription factors. Notably, the UMI counts of some constitutive genes, such as *Taf1b*, *Taf2*, *Tbp* and *Gapdh*, are quite low. Neural markers and identity genes with similar UMI count levels, such as *Sox1*, *Notch1*, *Tle1* or *Satb2*, are expected to be expressed with biological relevance.

<i>Cox7c</i> vs <i>Tars</i>	$Tars \geq 1$	$Tars = 0$	tot
$Cox7c \geq 1$	262	1006	1268
$Cox7c = 0$	249	1376	1625
tot	511	2382	2893

This contingency table *cannot* be tested with the standard statistics, which are built either on the hypergeometric distribution (Fisher's test) or on deviation from expected values given by the proportion of the marginals (chi-squared test). In fact, these methods are based on the assumption that all experimental units have the same probabilities of ending in the four counts. In scRNA-seq data, the different UDE ν_c of different cells makes this hypothesis false, and the standard test would give *spurious correlations*, as can be seen by computing the expected values in the standard way:

<i>Cox7c</i> vs <i>Tars</i>	$Tars \geq 1$	$Tars = 0$	tot
$Cox7c \geq 1$	224.0	1044.0	1268
$Cox7c = 0$	287.0	1338.0	1625
tot	511	2382	2893

Cox7c and *Tars* are constitutive genes, and hence they should not show a significant co-expression, but with classical statistics the resulting p -value is very low (0.019% with chi-squared test and 0.022% with Fisher-Irwin

test). In fact the cells with lower UDE have higher probability to end in the $\{R_{g_1} = 0, R_{g_2} = 0\}$ box and so they raise the frequency of that outcome above the expected 1338, yielding in this case a total of 1376 cells. With the classical contingency table interpretation, these would be explained by a positive co-expression of the two genes, but this is incorrect: the two genes are constitutive, and should be really expressed in all cells in the sample, it is random sampling that causes some cells to have zero UMI count. Random sampling of the two genes should also be independent, *conditionally* on the cell UDE, but if UDE is not taken into account, then there is correlation of the two genes *through UDE variability*.

Fortunately, with the estimates of $P(R_{g,c} = 0)$ available, we can approximate the expected count of the $\{R_{g_1} = 0, R_{g_2} = 0\}$ box with $\sum_c P(R_{g_1,c} = 0) \cdot P(R_{g_2,c} = 0)$. In the example above, this yields expected values which are much closer to the observed values:

<i>Cox7c</i> vs <i>Tars</i>	<i>Tars</i> ≥ 1	<i>Tars</i> = 0	tot
<i>Cox7c</i> ≥ 1	256.1	1011.9	1268
<i>Cox7c</i> = 0	254.9	1370.1	1625
tot	511	2382	2893

3.3 Test for GPA and co-expression index

The proposed GPA model yields an approximate table of expected cell counts under the hypothesis of independence of two genes g_1 and g_2 , therefore it is possible to build the usual statistics S_{g_1, g_2} for contingency tables, by comparing observed counts $O_{i,j}$ to expected counts $\epsilon_{i,j}$ (dependence on g_1 and g_2 implicit):

$$S := \sum_{i,j=0}^1 \frac{(O_{i,j} - \epsilon_{i,j})^2}{1 \vee \epsilon_{i,j}}$$

The small correction on the denominators is needed because we are forced to use this statistics also when some of the expected values are small, since Fisher's test cannot deal with variable UDE's. The correction is conservative, in the sense that it will not increase false positives.

One can then formally compute a $\chi^2(1)$ p -value as usual as $p_{g_1, g_2} := P(\chi^2(1) > S_{g_1, g_2})$. For example, in the case of the two constitutive genes of the previous section, one gets $p_{Cox7c, Tars} = 56.4\%$, in accordance with the null hypothesis that the two genes are independently found in the cells.

Rigorously, the computed probability is not a true p -value for this test, since one cannot easily deduce its distribution under the null hypothesis, but it has the usual qualitative meaning, and tests on negative synthetic datasets (see Section 3.5) show it to be near to the uniform distribution on the unit interval (Figure 3C-D). Figure 13 shows the distribution of GPA's p -values for several genes in a normal biological sample with different cell types²⁰.

Figure 13. Violin plots of GPA’s p -values of CGs (left) and NGs (right), from²⁰ E15.5 . Each gene was tested for independence against all other genes and the resulting approximate p -values are represented. CG values are typically uniformly distributed, as expected (see Section 3.4). Small values are instead more frequent for NGs, as these are either co-expressed or disjointly expressed with most other non-constitutive genes.

We introduce also the GPA *co-expression index* (COEX) of the two genes, defined as $\pm\sqrt{\frac{S}{m}}$, where m denotes the number of cells in the sample, and with the sign chosen according to whether the deviation of $O_{i,j}$ from $\epsilon_{i,j}$ is in the direction of joint expression (positive) or disjoint expression (negative). Hence we could write $\text{COEX} = \text{sgn}(O_{0,0} - \epsilon_{0,0})\sqrt{\frac{S}{m}}$ or other equivalent formula.

When the null hypothesis of the test is true, COEX distribution will be approximately Gaussian with mean zero and small variance $1/m$, so values near zero are in accordance with the independence hypothesis. On the other hand, when the null hypothesis is false, COEX will converge, for $m \rightarrow \infty$, to some deterministic coefficient in $[-1, 1]$ depending on the joint expression of the two genes in the population, hence values away from 0 may indicate bonds between the genes². Figures 1A and 2B show typical examples of COEX values for constitutive and cell-identity genes: it is apparent that COEX is a powerful tool for studying genes relations and cell differentiation.

3.4 Global differentiation index

When GPA is performed genome-wide, between all pairs of genes of the sample, it becomes possible to score genes based on their propensity to show either joint or disjoint expression against any other gene. Consider in particular a constitutive gene. It should be (biologically) present in all cells and its possible disappearance from the UMI counts of some cells should be only due to random sampling, independently from any other gene. Therefore, for any constitutive gene, the distribution of the GPA p -values against other genes should be approximately uniform on the unit interval (Figure 13, left).

²The test explained above can give statistical significance to such a hypothesis.

On the other hand, for a cell-identity gene, we expect a considerable fraction of p -values to be very small, as the independence hypothesis will be false for all the related cell-identity genes (Figure 13, right).

The global differentiation index (GDI) of a gene is defined as a normalization of the average of the smallest 5% of all the GPA p -values against other genes. The normalization chosen, $f(x) = \ln(-\ln(x))$, helps to make the index distribution more evenly spread and easy to grasp. Analysis of synthetic and negative datasets show that a value above 1.5 can be usually associated with non-constitutive genes, with $< 1\%$ false positives (Figure 4). For a constitutive gene with average expression level, a typical value for the GDI should be around $\ln(-\ln(0.05/2)) \approx 1.31$. Observe that because the GDI comes from p -values, it is expected to depend on the sample size for non-constitutive genes: a larger cell number makes the statistical test more powerful producing GDIs generally higher, for differentially expressed genes.

3.5 Synthetic datasets creation

We generated six synthetic datasets to assess the performance of COTAN and the quality of the estimators.

To make them more realistic, the parameters were obtained from two biological datasets: an E17.5 mouse cortex Drop-seq (DSQ) dataset²⁰ and a P0 hippocampal mouse 10X dataset¹⁶. For both of these, cells were clustered adapting the Louvain-Jaccard Clustering algorithm used in Loo et al.²⁵, finding 6 and 15 cell types respectively. We assumed conditional $\text{Poisson}(\nu_c \Lambda_g^{(c)})$ distributions for UMI counts, and for $\Lambda_g^{(c)}$ we assumed gamma distributions with parameters depending on cell type and on the gene. Then we performed maximum likelihood estimation on each cluster separately (with an iterative optimization procedure using Tensorflow²⁸), to get the UDE ν_c for each cell and the gamma parameters for each cell type and for each gene.

After parameters estimation, the synthetic datasets were generated all in the same way (described below) but with different sources (DSQ or 10X), number of cells (800 or 4000) and cell types (1 or 6 for DSQ, or 15 for 10X). The two synthetic datasets with homogeneous (H) cell type (DSQ_H_800 and DSQ_H_4000) are considered negative datasets and are used to assess the false positive rate of GPA and the distribution of approximated p -values and GDI under the hypothesis of independence (see Section 2.2). The four synthetic datasets with multiple (M) cell types (DSQ_M_800, DSQ_M_4000, 10X_M_800 and 10X_M_4000) are considered heterogeneous samples and are used to assess the performances of the parameters estimators (see Section 3.1.1).

For each dataset we generated the UDE of each cell c by sampling from the empirical distribution obtained with COTAN estimators on the biological samples (see Figure 16B-C for typical distributions).

We assigned randomly each cell c to a type $t(c) \in \{1, 2, \dots, K\}$, (where K was 1, 6 or 15, depending on the dataset), with probabilities propor-

tional to the cluster sizes. Then we generated the potential expression $\Lambda_g^{(c)} \sim \text{gamma}(\eta_{g,t(c)}, \theta_{g,t(c)})$ for each gene, with parameters depending on the gene and the cell type, and finally we generated the UMI counts $R_{g,c} \sim \text{Poisson}(\nu_c \Lambda_g^{(c)})$.

The two homogeneous synthetic datasets contain a single type of cells, whose parameters were estimated from the largest cluster in the biological sample. Since in these datasets there is by construction no cell differentiation, but just variable cell extraction efficiency, all tests should result in accepting the null hypothesis, and hence they can be used to estimate the false positive rate and to check the distribution of the GPA’s p -value (see Figure 3C-D) and genome-wide GDI under the null hypothesis (see Figure 4C-D).

3.6 Gene co-expression space

GPA can be used to represent genes in a low-dimensional space and cluster them by co-expression. We already pointed out that two genes are shown to be co-expressed when their COEX is large, but if the aim is to cluster genes according to their differential expression, this information is not rich enough to infer something precise about their distance. For example it would be nice to include the information that both genes have strong negative COEX against a third gene. To be able to build a proper distance, we generalize this idea by comparing the whole co-expression patterns of the two genes against other genes: if these turn out to be similar, then the two genes should be probably considered “close” in the co-expression space.

The procedure starts by fixing a set V of genes to use as variables and a set U of genes to use as units to analyze (Figure 6). The two sets may be equal, but it is not required. Both sets, but in particular V , should contain only genes that are known or suspected to be differentially expressed in the sample, to reduce noise and improve quality.

To measure co-expression we consider the GPA COEX matrix for these genes. Applying the transform $\tanh(\sqrt{m} \text{COEX})$ yields roughly binary values, since for many components $\sqrt{m} \text{COEX}$ will have large absolute value³. There will be components near ± 1 for statistically significant joint and dis-joint expression, and components randomly spread in $(-1, 1)$ for the other couples. In this way the Euclidean distance between points becomes similar to a Manhattan distance for the co-expression pattern and hence these transformed matrix can be used as input for dimensionality reduction techniques and clustering.

The space in which genes g are represented with components $g = (g_v)_{v \in V}$

$$g_v = \tanh(\sqrt{m} \text{COEX}(g, v))$$

---

³This is because its square is the statistics S from GPA test for independence (see Section 3.3).

for all genes $v \in V$ is the *co-expression space* relative to the set V . This space is suitable to perform dimensionality reductions, graphical representations and cluster analysis, as its Euclidean geometry is based on the gene co-expression patterns with genes in V . The choice of V will affect all these analysis and in fact a focus on peculiar subsets V may uncover information that is hidden or confounded in the genome-wide co-expression space (see also Section 2.3 of Results and Section 2.4 on the related concept of localized GDI).

3.7 COTAN pipeline

In the following subsections we describe the detailed COTAN pipeline (Figure 14). We developed and tested our method on drop-based techniques and in particular on 10X datasets and Drop-seq datasets. Analysis requires and starts from a matrix of raw UMI counts after removing possible cell doublets or multiplets and low quality or dying cells (with too high mtRNA percentage).

3.7.1 Data cleaning

The first step is removing genes that are not significantly expressed (default threshold is 0.3% of cells) or unwanted (such as mitochondrial ones) to reduce the matrix memory size and to increase the analysis speed without losing information.

The user chooses either the linear or the square-root estimation framework (see Section 3.1.1), then begins an iterative procedure to filter out outlier cells (such as blood cells in a brain cortex dataset). In each iteration the UDE is estimated for all cells and UMI counts are simply normalized dividing by its value. Cells are then represented on the PCA plane and clustered by Mahalanobis distance (two clusters, A , B , complete linkage clustering). The clustering algorithm will detect outlier cells which will fall into the smallest cluster B (Figure 15A). A subsequent plot displays the most abundant genes expressed in B , to allow the user to check if they are peculiar in any way (Figure 15C). The user may choose to drop the cells in B and do another iteration, or to stop the procedure, when the PCA plot does not show obvious outliers (Figure 15B-D).

After the last iteration there are two final quality checks on the estimated UDE of cells. Firstly the PCA plot colored by UDE should not show a clear separation of cells with high and low UDE (Figure 16A). In fact COTAN builds on the assumption that UDE is not gene-dependent (see Section 3.1.1) and if the PCA plot is polarized by UDE, this may be false. Secondly, the plot of sorted UDE values will detect if the efficiency drops notably for a small fraction of cells. If this is the case, we usually want to exclude cells below the elbow point (Figures 16B-C, notice that UDE values are

Figure 14. Pipeline diagram.

normalized to have mean 1, so there is not absolute threshold for acceptable efficiency). If cells are removed, another estimation iteration is due.

Figure 15. Cleaning plots. **A, B.** PCA plots of the cells clustered by Mahalanobis distance in a larger *A* (blue) and a smaller *B* (red – see text) set. **C, D.** Plots for the most expressed genes in the *B* cell group. **A, C** show a dataset that must be cleaned, with a small well separated cluster *B* and high expression on unusual genes. **B, D** show an example of the final stage, when no more cleaning iterations are required.

3.7.2 Co-expression tables

This is a genome-wide procedure to compute the tables of observed counts for each pair of genes (see Section 3.2). If n is the number of genes in the sample, four $n \times n$ matrices store the number of cells in each of the four conditions (expressing both genes, only the first one, only the second one or none). Constitutive genes that appear in every single cell cannot be used and are removed in this step (saving a list of them).

Figure 16. UDE plots for data cleaning. **A.** PCA plot of the cells colored by estimated UDE; in this case they show uniform distribution, so COTAN assumptions are reasonable. **B.** Plot of ordered UDE values for all cells. **C.** Magnification of the previous plot showing the elbow point below which cells should be removed.

3.7.3 Expectation of co-expression table values

Before computing the expected values for the co-expression tables, it is necessary to estimate the dispersion parameter a_g of each gene g following the mathematical framework of Section 3.1. The approximated values are determined iteratively by simple bisection. Genes for which the estimated dispersion would be negative are instead modelled with zero dispersion and mean expression increased to $(1 + b_g)\lambda_g$. The positive parameter b_g is en-

coded as $-a_g$ so that one single parameter can account for both cases. The fraction of genes with negative a_g is reported.

Afterwards the tables of expected values are computed like the observed counts, as four $n \times n$ matrices corresponding to the same four conditions.

3.7.4 GPA p -values and COEX matrix

The script computes, for each pair of genes, the GPA test statistics S , the corresponding $\chi^2(1)$ p -value and COEX index (Section 3.3) and saves them in three $n \times n$ matrices. The last two files are the primary output of COTAN analysis.

3.8 Seurat pipeline for Figure 1

Seurat (3.1.0) workflow was performed on E16.5 hippocampal dataset¹⁶ following the Guided Clustering Tutorial²⁹, with modifications. Data import (CreateSeuratObject) was done using min.cells 3 and min.features 200. The selected range for the number of features was between 200 and 4000; the maximum allowed fraction of mitochondrial genes per cell was 7.5%. Normalization was done using the default parameters. The correlation was then calculated on the whole Seurat normalized data matrix and the heatmap was plotted subsetting this (Figure 1B). Figure 1D was plotted by calling the function FindVariableFeatures with selection method VST.

4 Conclusion

In conclusion, we introduced a new method for the analysis of scRNA data and described its application to datasets of embryonic hippocampus and cerebral cortex, which show high and documented cell identity diversity. We found that COTAN is well suited to identify gene pairs which are jointly or disjointly expressed in single cell. Moreover, COTAN GDI is a new parameter to assess differentially expressed genes in a dataset.

Finally, COTAN could cluster genes with consistent co-expression in single cell rather than cells with consistent gene global expression. Notably, COTAN features makes it a novel tool that, unlike previous indirect approaches³⁰, can directly infer single-cell gene interactome from scRNA-seq dataset.

References

- [1] Stuart, T. and Satija, R. Integrative single-cell analysis. *Nature Reviews Genetics* **20**(5) 257–272 (2019).

- [2] Kiselev, V. Y., Andrews, T. S., and Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20**(5) 273–282 (2019).
- [3] Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., and Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nature Communications* **10**(1) 1–11 (2019).
- [4] Luecken, M. D. and Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15**(6) (2019).
- [5] Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**(5) 1202–1214 (2015).
- [6] Zhang, X. et al. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular cell* **73**(1) 130–142.e5 (2019).
- [7] Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* (2017).
- [8] Townes, F. W., Hicks, S. C., Aryee, M. J., and Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology* **20**(1) 295 (2019).
- [9] Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* **38**(2) 147–150 (2020).
- [10] Pierson, E. and Yau, C. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology* **16**(1) 241 (2015).
- [11] Van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**(3) 716–729 (2018).
- [12] Huang, M. et al. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature methods* **15**(7) 539–542 (2018).
- [13] Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* **11**(2) 163–166 (2014).
- [14] Galfrè, S. G. and Morandin, F. A mathematical framework for raw counts of single-cell RNA-seq data analysis. *arXiv:2002.02933* (2020).
- [15] Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**(5) 411–420 (2018).

- [16] Hochgerner, H., Zeisel, A., Lönnerberg, P., and Linnarsson, S. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nature Neuroscience* **21**(2) 290–299 (2018).
- [17] Ramsköld, D., Wang, E. T., Burge, C. B., and Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Computational Biology* **5**(12) (2009).
- [18] Greig, L. C., Woodworth, M. B., Galazo, M. J., Padmanabhan, H., and Macklis, J. D. Molecular logic of neocortical projection neuron specification, development and diversity. *Nature Reviews Neuroscience* **14**(11) 755–769 (2013).
- [19] Bertrand, N., Castro, D. S., and Guillemot, F. Proneural genes and the specification of neural cell types. *Nature Reviews Neuroscience* **3**(7) 517–530 (2002).
- [20] Yuzwa, S. A. et al. Developmental Emergence of Adult Neural Stem Cells as Revealed by Single-Cell Transcriptional Profiling. *Cell Reports* **21**(13) 3970–3986 (2017).
- [21] Molyneaux, B. J., Arlotta, P., Menezes, J. R., and Macklis, J. D. Neuronal subtype specification in the cerebral cortex (2007).
- [22] Nitzan, M., Karaiskos, N., Friedman, N., and Rajewsky, N. Gene expression cartography. *Nature* **576**(7785) 132–137 (2019).
- [23] Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**(W1) W90–W97 (2016).
- [24] Chen, E. Y. et al. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14** (2013).
- [25] Loo, L. et al. Single-cell transcriptomic analysis of mouse neocortical development. *Nature Communications* **10**(1) 134 (2019).
- [26] Allen Brain Atlas data portal. <http://developingmouse.brain-map.org/>. [Online; accessed 31-March-2020] (2020).
- [27] Crow, M. and Gillis, J. Co-expression in Single-Cell Analysis: Saving Grace or Original Sin? *Trends in Genetics* **34**(11) 823–831 (2018).
- [28] Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org (2015).

- [29] Seurat Guided Clustering Tutorial. https://satijalab.org/seurat/v3.1/pbmc3k_tutorial.html. [Online; accessed 20-February-2020] (2020).
- [30] Mohammadi, S., Davila-Velderrain, J., and Kellis, M. Reconstruction of cell-type-specific interactomes at single-cell resolution. *Cell Systems* **9**(6) 559–568 (2019).