Identification of critical cell-types using genetic modules: A case study of neurodevelopmental disorders

Julie Chow\textsuperscript{1,}*, Marketa Tomkova\textsuperscript{1,}*, Ashleigh Thomas\textsuperscript{2}, Elior Rahmani\textsuperscript{3}, Sagiv Shifman\textsuperscript{4}, and Fereydoun Hormozdiari\textsuperscript{1,5,*}

\textsuperscript{1}Genome Center, UC Davis, Davis, CA, USA
\textsuperscript{2}University of British Columbia, Vancouver, BC, Canada
\textsuperscript{3}Electrical Engineering and Computer Sciences, UC Berkeley, CA, USA
\textsuperscript{4}Dept. of Genetics, Hebrew University, Jerusalem, Israel
\textsuperscript{5}Department of Biochemistry and Molecular Medicine, UC Davis, Sacramento, CA, USA

*Joint first authors

\textsuperscript{*Corresponding authors: fhormozd@ucdavis.edu}

July 4, 2023

Abstract

Identifying the critical cell-types impacted by various diseases is crucial for understanding disease mechanisms and developing targeted therapeutics. Recent advances in disease genetic module discovery and single-cell technologies provide a unique opportunity to study critical cell-types based on functional pathways and modules. Disease genetic modules are defined as sets of genes with correlated expression that are part of the same biological pathways and are disrupted in the disease. Critical cell-types for a biological function are defined as clusters of similar cells most “active” or “involved” in that biological function. In this paper, we provide a formal problem definition for the critical cell discovery problem using the recently introduced local correlation concept, and show that the proposed problem is intractable in theory. We propose a novel method, MoToCC (Module To Critical Cell-types), to find sets of similar cells with local correlated gene expression activity for input modules. We evaluated MoToCC on four neurodevelopmental disorder modules using single-cell expression data from the developing human cortex. Finally, we demonstrate that the objective value returned by MoToCC for the tested modules is an acceptable approximation to the optimal solution. Overall, our work provides a valuable tool for studying critical cell-types and their role in disease mechanisms, which could lead to the development of more effective targeted therapeutics. The MoToCC package is available at https://github.com/jchow32/MoToCC

keywords: critical cell-types, genetic module, single-cell RNA-sequencing, Linear Programming

1 Introduction

The advent of single-cell and single-nucleus RNA-sequencing (scRNA-seq and snRNA-seq) technologies has revolutionized the study of cellular decomposition in tissues by providing researchers with unbiased tools. These technologies enable the discovery of novel cellular subtypes and the genes that are differentially expressed in each of these subtypes. The identification of these subtypes has the potential to reveal the underlying etiology of complex diseases and common traits [1] [2]. Recent single-cell technologies have allowed the exploration of molecular mechanisms in a wide range of biological systems, including tumor cells and their associated microenvironments, as well as previously uncharacterized neuronal subtypes [3] [4]. Single-cell analysis enables the dissection of cellular heterogeneity and the identification of specific molecular
targets for drug intervention, populations of cells with coordinated expression relevant to common biological functions, and the origins of disease pathogenesis, in contrast to bulk RNA-seq analysis [5, 6, 7]. Thus, single-cell and single-nucleus RNA sequencing have opened up new avenues for understanding the molecular basis of disease and for developing targeted therapies.

Genetic modules are groups of genes with similar biological functions that are distinct from other modules. These modules usually consist of co-expressed genes that are highly connected in protein-protein interaction networks. A genetic module that its disruption causes a specific disease is denoted as disease genetic module. Disease genetic modules often exhibit a significant enrichment of genetic variants or epigenetic modifications in affected cases compared to controls. Several methods for genetic module discovery have been developed, including those using gene expression, gene co-expression networks, protein-protein interaction networks, signaling pathways, and GO-terms as biological signals [8, 9, 10, 11]. The maturity of these methods and the ability to test modules in vitro and in vivo have led to the discovery of a growing list of disease genetic modules. To better understand the functional impact of a disease genetic module, it is crucial to identify the specific cell types in which the disease module is most active, and to understand how its disruption contributes to the disruption of critical cell types.

Previously proposed ideas for predicting the critical cell-types of certain diseases have involved linking the expression of modules to a specific defined tissue or cell-types [12, 13]. The simple approach of selecting the critical cell-type based on the highest expression of the input module assumes that the set of cell-types is static and predefined. In reality, the molecular profiles of individual cells are highly heterogeneous, thereby complicating the absolute assignment of a cell to any given cell-type or subtype.

In this paper, we propose a formal framework for the discovery of critical cell-types associated with a module. We have developed software named Module To Critical Cell-types (MoToCC), a novel linear programming approach that returns a subset of cells that selectively express module genes for given gene expression data. Users may vary the desired maximum number of cells to return as a solution, permitting the user to visualize relevant, distinct groups of cells that have similar expression levels at different scales of resolution. MoToCC source code and associated scripts are freely available at [https://github.com/jchow32/MoToCC].

2 Problem Formulation

In this section, we will first provide an informal intuition about the critical cell discovery problem, next provide a formal problem definition, and finally give an NP-completeness proof of the proposed problems. We define a critical cell-type for a given genetic module as a set of cells that are the main target and impacted by the biological function of the input genetic module or pathway. Such approaches also provide a tool for finding the associations between human diseases, cell-types, and cellular processes by integrating various biological signals [14]. The level of expression of the genes in an input module can be considered as one factor in selecting the critical cell-types for an input module, however, we propose that the most important factor is the whole functional activity of the module that determines the critical cells. This is similar to utilizing gene co-expression (correlation between the expression of two genes) that has proven very effective in finding genetic modules.

Our formulation is motivated by the recently introduced concept of “local correlation” [12]. The local correlation between two genes is defined by considering their expression over all input cells, denoted by set $C$, available in the single-cell data [12]. Formally, they define local correlation statistics between two genes $p$ and $q$ as $h_{p,q} = \sum_{i,j \in C} w_{i,j} \times (p_i q_j + p_j q_i)$, where $w_{i,j}$ is the similarity score between cells $i$ and $j$, and $p_i, q_i, p_j, q_j$ are the normalized and scaled expressions of the genes $p, q$ in respective cells $i$ and $j$. Note that the scaled and normalized gene expression guarantees that for every gene $p$, the $E[p_i] = 0$ and $Var[p_i] = 1$ over all cells in $C$. Here, we propose to extend the definition of pairwise local correlation to a local correlation of a genetic module $M$ and a set of cells $S \subset C$ as

$$H_{M,S} = \sum_{p \in M} \sum_{q \in M, q \neq p} \sum_{i \in S} \sum_{j \in S} w_{i,j} (p_i q_j + p_j q_i)$$

We are now ready to give formal problem definitions:
• First, we formally define the discovery of the most locally correlated cells (MLCC) problem. Assume a genetic module \( M \), normalized single-cell gene expression, and a user-defined parameter \( k \) are provided. The objective of the MLCC problem is to select a subset cells \( S \subseteq C \) (where \( |S| \leq k \)) to maximize \( H_{M,S} \).

• Next, we formally define the discovery of the critical cell-type problem as finding a subset cells \( S \subseteq C \) (where \( |S| \leq k \)) to maximize \( H_{M,S} \) which are also “similar” based on available single-cell expression data. For the similarity constraint we use the condition that selected cells have to be strongly connected in the K-NN similarity graph. Note that the K-NN similarity graph is built using expression of all the genes \([15]\).

**Theorem 1.** The MLCC problem is NP-complete.

**Proof.** We provide a reduction from the standard subset sum problem to the MLCC problem. Given a set of numbers, \( X \), the decision version of the subset problem asks if there exists a nonempty subset of these numbers where their summation is equal to zero. We construct an instance of finding the most locally correlated cells problem using module \( M \) and set of cells \( C \), i.e., finding the set \( S \subseteq C \) that maximizes the \( H_{M,S} \) given the module \( M \) and cells \( C \), using numbers in set \( X \) as follows. We first assume module \( M \) consists of only two genes \( p \) and \( q \) (i.e., \( M = \{p, q\} \)), and for any number \( x \in X \) we consider a cell in set \( C \) with gene expression \( x \) and \(-x\) assigned to genes \( p \) and \( q \) respectively (i.e., \( p_x = x \), \( q_x = -x \)). Note that total number of cells in the instance of the MLCC problem is equal to cardinality of \( X \) (i.e., \( |C| = |X| \)). We also assume the value \( w_{i,j} = 1 \) for every pair of cells \( i \) and \( j \). Thus, solving this instance of the MLCC problem is finding a set of cells \( S \subseteq C \) that maximize \( H_{M,S} = \sum_{i \in S} \sum_{j \in S} (p_i q_j + p_j q_i) = \sum_{i \in S} p_i \sum_{j \in S} q_j + \sum_{i \in S} q_i \sum_{j \in S} p_j \). Now, based on our instance construction assumption, we have \( \sum_{i \in S} p_i = -\sum_{i \in S} q_j \). Thus \( H_{M,S} = -2(\sum_{i \in S} p_i)^2 \leq 0 \). Finally, we conclude that a solution to this instance of the MLCC problem would also solve the decision version of the subset sum problem give set of numbers \( X \).

**Cells with up-regulation activity of module \( M \)**: Note that, in theory the above formulation the solution can produce a set of cells that the input module \( M \) is down-regulated in them as optimal solution. However, in practice we are mostly interested in critical cells that the input module \( M \) is up-regulated in them. Thus, we limit the above formulation by imposing an additional constraint such that module \( M \) is up-regulation in the selected cells.

We next propose a linear programming formulation for solving the MLCC problem, and then use this formulation with heuristic rounding to solve the critical cell-types problem.

### 3 Methods

We propose a two step framework, MoToCC, to discover the most critical cells given the module \( M \). The inputs to MoToCC consist of a genetic module \( M \), normalized single-cell gene expression, and a user-defined maximum number of critical cells \( k \). The similarity matrices, Shared Nearest-Neighbor (SNN) and K nearest-neighbor (K-NN) graphs, are built using the normalized single-cell gene expression \([16]\).

We first provide a linear programming (LP) formulation for solving the MLCC problem with a simple rounding solution of real values returned for the LP. Next, we will utilize the real value solution from the linear programming proposed for the MLCC problem to find the critical cell-types by imposing a cell-cell similarity constraint during the rounding step. The cell-cell similarity constraint in the critical-cell problem is that selected cells induced on the K-NN graph have to be a strongly connected component.

#### 3.1 Linear programming formulation for MLCC problem

For solving the MLCC problem, we represent each cell as a node in a graph and for every pair of nodes \( i \) and \( j \), we define a weight \( w_{i,j} \) indicating the cell-to-cell similarity as retrieved from the SNN graph.
Given the genes in the input module \( M \), we pre-calculate for every pair of cells \( i \) and \( j \) the score \( z_{i,j} = \sum_{p \in M} \sum_{q \in M, p \neq q} w_{i,j} (p_i q_j + q_i p_j) \) where \( p_i \), \( p_j \), \( q_i \), and \( q_j \) represent the normalized expression of genes \( p \) and \( q \) (\( \in M \)) in cells \( i \) and \( j \) respectively. The objective of the MLCC problem is to select a maximum of \( k \) cells such that the summation of \( z_{i,j} \) for every pair of cells selected is maximized. We note that the MLCC problem is NP-complete (Theorem 1) and thus we provide a heuristic solution for solving it.

We first introduce the linear programming (LP) relaxation of the MLCC formulation and use a simple rounding approach to get the final solution. For each cell/node \( i \) we define variable \( y_i \) to indicate the selection of that cell/node. For every pair of cells/nodes \( (i, j) \) we also define the variable \( x_{i,j} \) to indicate that both pairs of nodes \( i \) and \( j \) are selected (i.e., \( y_i = 1 \) and \( y_j = 1 \)). We provide a linear programming formulation for the relaxed version of the MLCC problem as follows:

\[
\begin{align*}
\text{Maximize} & \quad \sum_{i \in C} \sum_{j \in C} z_{i,j} x_{i,j} \\
\text{such that} & \quad \sum_{i \in C} y_i \leq k \quad \text{maximum number of cells selected} \\
& \quad \sum_{i \in C} \sum_{p \in M} y_i p \geq 0 \quad \text{limit selected cells to be upregulated for module } M \\
& \quad x_{i,j} \leq y_i \\
& \quad x_{i,j} \leq y_j \\
& \quad x_{i,j} \geq y_i + y_j - 1 \\
& \quad 0 \geq x_{i,j} \geq 1 \\
& \quad 0 \geq y_i \geq 1 \\
\end{align*}
\]

**Rounding of MLCC solution**: The proposed real value solution to the LP needs to be rounded to an integer solution indicating selected cells. As we are constrained by the number of cells we can select to be \( \leq k \), the rounding we propose is to select the top \( k \) cells with the highest calculated \( y_i \) score. Finally, the associated objective function values for cells in the LSCC are returned as the final solution.

### 3.2 Critical cell-types discovery problem

We propose a heuristic for predicting the critical cell-types that builds on the MLCC linear programming formulation. We utilize the same LP as the MLCC problem, however, during rounding we impose the constraint that selected cells need to be strongly connected in the K-NN graph. First, we select the initial solution returned by linear programming of cells with \( y_i > 0 \), which is referred to as potential “candidate cells”. Next, among the candidate cells, the largest strongly connected component (LSCC) in the K-NN graph is rounded to 1 and returned as the final solution. We utilize the largest strongly connected component (LSCC) function implemented in the networkx library (version 1.11) [17]. While, we don’t have an asymptotic approximation factor for this proposed rounding approach, we can experimentally calculate a lower bound on the approximation factor for various input modules using this solution (see section 4.4).

### 3.3 Implementation notes

For solving the linear programming we used the academic version of IBM ILOG CPLEX Optimizer [18]. To maximize the objective function (Section 2.2), edge weights \( (z_{i,j}) \) are calculated between cells that have non-zero similarity \( (w_{i,j}) \) as per the SNN graph. If pruning (–prune) is enabled, only edge weights outside of one standard deviation from the mean edge weight are retained. For a module, edge weights only need to be calculated once. Thus, if the user wishes to run MoToCC using varied \( k \), the quickstart parameter (–quickstart) can be enabled to load edge weights previously calculated by MoToCC using the same module.

### 3.4 Return and refinement of initial solution

MoToCC returns the silhouette score associated with a 2D t-SNE dimensionality reduction and K-means clustering (K=2) of cells in the LSCC. Given multiple silhouette scores for varied \( k \), breakpoints at which
distinct groups of cell-types are selected can be viewed. To visualize the selected cells of the refined solution compared to unselected cells (via 2D t-SNE, 3D t-SNE, and UMAP) and to plot silhouette scores versus varied k, additional scripts and their usage are described at https://github.com/jchow32/MoToCC.

4 Results

4.1 Data

We evaluated the performance of the MoToCC pipeline on scRNA-seq gene expression datasets from the developing human cortex brain [13]. To analyze these datasets, we used four modules previously identified in studies of neurodevelopmental disorder: two NDD modules (M1_NDD and M2_NDD) that were discovered using de novo variants [19], an epilepsy module (M_SCN1A) centered on the seed gene SCN1A [20], and an autism module (ASD_mod13_Group1) enriched in chemical synaptic transmission genes [21]. We obtained and normalized the data using the standard Seurat pipeline.

4.2 Overview of experimental analysis

We compared the critical cells identified by MoToCC to those identified by Seurat and examined their agreement with known biological functions and cell subtypes associated with each module. To evaluate the statistical significance of critical cells selected per module, we compared the objective function value of the selected cells to that of randomly selected modules of the same size. To ensure that the gene expression distribution of the randomly selected modules was similar to that of the input module, we binned all genes into 10 bins based on their average expression across all cells and randomly selected the same number of genes from each bin as in the original input module. Finally, we compare the MoToCC results against a naïve heuristic approach that selects the subset of cell clusters found by Seurat [16] with the highest average gene expression, denoted as $h_S$ - i.e., the average expression of module genes in selected cells S (see section 4.5).

The main results in support of the critical cells predicted by MoToCC:

- The critical cells predicted by MoToCC for all test modules are enriched in cell types associated with the disrupted module.
- For almost all the $k$ values, MoToCC returns a significantly higher objective score using the real genetic module compared to randomly selected modules of the same size and similar average gene expression.
- MoToCC outperforms the naïve heuristic approach in selecting cell types with higher average gene expression for the input module.

4.3 Neurodevelopmental disorders critical cell-types

We investigated the critical cell-types impacted in NDDs using four different genetic modules previously reported [19,20,21]. For neurodevelopmental disorder modules, we consider two NDD modules found using MAGI [19], the epilepsy module found using MAGI-S with SCN1A as a seed gene, and a recent autism module found [21] respectively denoted as M1_NDD, M2_NDD, M_SCN1A and ASD_mod13_Group1.

The M1_NDD module (80 genes) is significantly enriched in chromatin remodeling and the Wnt pathway, while the M2_NDD module (19 genes) is significantly enriched in chemical synaptic transmission and long-term potentiation. The module M_SCN1A (36 genes) is significantly enriched in non-synonymous de novo mutations from epilepsy cohorts, known epilepsy genes, and pathways such as long-term potentiation, chemical synaptic transmission, and regulation of neurotransmitter activity. The ASD_mod13_Group1 module (53 genes) was discovered via hierarchical clustering of a topological protein interaction network and was enriched in genes including SYNGAP1, SHANK2, and SHANK3 and the synaptic transmission pathway.

The critical cells returned by MoToCC for all the NDD modules had significantly larger initial and final total objective function value at all $k$ compared to modules of the same size consisting of random genes.
Figure 1: A) The largest increases in silhouette scores are highlighted in red. B) For the M1 module, selected cells from largest increase in silhouette score are primarily of the progenitor cell-type (PgS, PgG2M and IP). C) The final objective function of critical cells for the M1 module (red) are compared to values from 20 same-sized modules consisting of random genes (black). D) Two-dimensional t-SNE plots including selected critical cells (red) at \(k\) corresponding to the greatest increases in silhouette score (red title).

selected to have similar average gene expression. This indicates that the associated predicted critical cells with the input module have potential biological significance.

4.3.1 Neurodevelopmental disorder module - M1_NDD

The M1 module exhibits functional enrichment in the regulation of transcription, chromatin remodeling, and Wnt pathway, rather than neurotransmitter secretion. However, CSEA [22] failed to identify any significant enrichment of the M1 module in any specific cell type, emphasizing the need to consider single-cell data and specialized methods to identify critical cells.

Silhouette scores can indicate disparate clustering and shifts in percent composition. In the case of the M1_NDD module, red points highlight the largest increases in silhouette score (see Figure 1). Two-dimensional t-SNE plots were also displayed for several \(k\) cell values (50, 500, 1000, 1500, and 2000), with cells selected by LSCC appearing in red. The silhouette score analysis showed that the \(k\) cells between 750 and 1250 exhibited the sharpest increase, and the selected cells were mainly progenitor cells (PgS, PgG2M, and IP). This finding aligns with recent reports that emphasize the crucial role of the Wnt pathway and primary autism genes in this module, including \(CHD8\), in neural progenitor cells [23] [24].
4.3.2 Neurodevelopmental disorder module - M2_NDD

In this study, we also investigated M2_NDD, an NDD module identified using MAGI. The module M2_NDD is significantly enriched in synaptic transmission and long-term potentiation. Using MoToCC, we identified critical cells for M2_NDD across all k values, with the majority of the selected cells belonging to the excitatory deep layer 1 and 2 neurons (ExDp1, ExDp2) (Figure 2). Specifically, more than 67% of all cells of the ExDp2 cell-type were selected when k = 750, and the proportion of ExDp2 cells tends to increase as k increases for M2_NDD.

The elevated proportion of ExDp1 and ExDp2 cells in the M2_NDD module is consistent with our functional enrichment analyses, which reveal enrichment in synaptic transmission and regulation of neurotransmitter receptor and cation channel activity. Furthermore, this finding complements the increased expression levels of relevant genes such as GABRB2, GRIN2B, and STXBP1 in ExDp2 cells. Consistent with these results, CSEA [22] also highlights the relevance of M2 genes in deep cortical neurons.

Taken together, our results suggest that the critical cells in M2_NDD are primarily excitatory deep layer 1 and 2 neurons, which are enriched in synaptic transmission and associated regulatory mechanisms. These findings provide valuable insights into the molecular and cellular mechanisms underlying neurodevelopmental disorders, and may help to inform the development of new therapeutic strategies.

4.3.3 Epilepsy module - M_SCN1A

The M_SCN1A and M2_NDD modules display similar functional enrichment due to the large proportion (> 30%) of shared genes among modules. Like M2_NDD, critical cells of the M_SCN1A module are primarily labeled as excitatory deep layer neurons. At smaller values of k (k < 250) the M_SCN1A module’s critical cells initially consist of most existing ExDp2 cells (Figure 3). The percent composition of critical cells for higher k values is mostly ExDp1 and ExDp2. This is in agreement to what is reported as importance of excitatory neurons for SNC1A deficit samples [25, 26].

4.3.4 Autism model (ASD Module 13, Group 1)

In a previous study, by Li et al. 2014 [21] found that the ASD_mod13_Group1 module exhibited high expression specificity in the corpus callosum, a brain region mainly populated by oligodendrocyte progenitor cells (OPCs) and astrocytes. Unlike the M1_NDD, M2_NDD, and M_SCN1A modules, ASD_mod13_Group1 shares no common genes except for the seed gene SCN1A in M_SCN1A. When using MoToCC to identify critical cell types at k=200 and 250, the algorithm selected only OPCs, which were followed by outer radial glial (oRG) and ventral radial glial (vRG) cells (see Figure 4). For k values below 150, MoToCC identified endothelial cells as the critical cell type, which showed the highest increase in silhouette score.

4.4 Estimation of approximation factor of critical cell discovery problem

We further investigate the empirical approximation factor of the proposed rounding heuristic for critical cell discovery using the four modules for various values of k. We calculated the ratio of final objective function value of $H_{M,S}$ for selected cells over the objective function value returned by the linear programming as an empirical estimation of the approximation factor (Figure 5). Note that, as the objective value returned by the linear programming (using real value solution) is an upper bound on the true objective function we can conclude the ratio calculated an lower bound on the approximation factor (Figure 5). Based on observed ratio for the four different modules for various k it seems the approximation factor is constant and quite close to real solution (ratio is > 0.5 for most k).

4.5 Comparison of MoToCC results against naïve approach

We provide a comparison of the MoToCC results against a naïve approach, in which the cells labeled as belonging to the cell-type with the largest average normalized expression values ($h_s$) were selected as a solution. While a ground truth in most cases is unknown, we compare the average expression ($h_s$) for
Figure 2: A) The largest increases in silhouette scores at \( k = (750; 1250; 2000) \) are highlighted in red. B) For the M2_NDD module, selected cells are primarily of the excitatory deep layer 1 and 2 (ExDp1, ExDp2) cell-types. C) The final objective function of critical cells for the M2_NDD module (red) are compared to values from 20 same-sized modules consisting of random genes (black). At all \( k \), the corresponding final total objective function value of the M2 module is significantly greater than that of the randomized modules. D) Two-dimensional t-SNE plots including selected critical cells (red) at \( k \) corresponding to the greatest increases in silhouette score (red title). Cell-type abbreviations are provided in Supplementary table S1.
Figure 3: A) The largest increases in silhouette scores at \( k = (500; 750; 1750) \) are highlighted in red. B) For the M\(_{SCN1A}\) module, selected cells are primarily of the excitatory deep layer 1 and 2 (ExDp1, ExDp2) cell-types at lower \( k \), followed by increased selection of the maturing excitatory upper enriched (ExM-U). C) The final objective function of critical cells for the M\(_{SCN1A}\) module (red) are compared to values from 20 same-sized modules consisting of random genes (black). At all \( k \), the corresponding final total objective function value of the M\(_{SCN1A}\) module is significantly greater than that of the randomized modules. D) Two-dimensional t-SNE plots including selected critical cells (red) at \( k \) corresponding to the greatest increases in silhouette score (red title).
Figure 4: A) The largest increases in silhouette scores at $k = (100; 250; 1,000)$ are highlighted in red. B) For the ASD_mod13_Group1 module, selected cells for $k < 150$ are primarily endothelial cells and for $k > 150$ of the oligodendrocyte progenitor cell (OPC) cell-types. C) The final objective function of critical cells for the ASD_mod13_Group1 module (red) are compared to values from 20 same-sized modules consisting of random genes (black). D) Two-dimensional t-SNE plots including selected critical cells (red) at $k$ corresponding to the greatest increases in silhouette score (red title).
Figure 5: Ratio of final and initial objective function values for modules at varied k.

selected cells for the tested modules compared to naive selection of the cell-type associated with the largest $h_s$ value (table 1). Note that, in all the five module while the MoToCC objective was not to maximize $h_s$ value, yet it was able to achieve a higher $h_s$ than the naive approach to select cells that maximize this score (table 1). For the M1

<table>
<thead>
<tr>
<th>Modules</th>
<th>#Cells</th>
<th>$h_s$</th>
<th>cell type</th>
<th>#Cells</th>
<th>$h_s$</th>
<th>cell type</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1_NDD</td>
<td>524</td>
<td>23.87</td>
<td>PgS, PgG2M, IP</td>
<td>48</td>
<td>11.95</td>
<td>Mic</td>
</tr>
<tr>
<td>M2_NDD</td>
<td>615</td>
<td>15.09</td>
<td>ExDp1, ExDp2</td>
<td>166</td>
<td>12.44</td>
<td>ExDp2</td>
</tr>
<tr>
<td>M_SCN1A</td>
<td>423</td>
<td>18.52</td>
<td>ExDp1, ExM-U</td>
<td>166</td>
<td>16.66</td>
<td>ExDp2</td>
</tr>
<tr>
<td>ASD_mod13</td>
<td>74</td>
<td>39.91</td>
<td>End</td>
<td>306</td>
<td>13.31</td>
<td>OPC</td>
</tr>
</tbody>
</table>

Table 1: Comparison of average expression ($h_s$) between MoToCC and naive (Seurat based) critical cell selection for selected $k$ at elevated silhouette scores. Cell-type abbreviations are provided in Supplementary table S1.

5 Discussion

Understanding the critical cells affected by disrupted pathways in each disease is crucial for better understanding the etiology of diseases. To address this challenge, we have developed a novel approach called MoToCC that utilizes disrupted disease modules and single-cell expression data to identify the critical cells impacted in a disease associated with the input module.

MoToCC uses a linear programming approach to identify a set of cells that selectively express genes from a given genetic module. Specifically, MoToCC maximizes an objective function that considers both the correlated co-expression (local correlation) among module genes and cell-cell similarity. This optimization
approach yields an initial solution, which is then refined using the derived K-nearest neighbor graph to select
the largest strongly connected component (LSCC) among cells in the initial solution. Therefore, the selected
critical cells in the LSCC consist of cells that share a high degree of local correlation in module genes and
display similar overall gene expression profiles compared to all other cells in the dataset.

To assess the performance of MoToCC, we analyzed normalized single-cell gene expression data from the
developing human cortex which were obtained from publicly available datasets [13]. We focused on four mod-
ules that are relevant to neurodevelopmental disorders (M1_NDD, M2_NDD, M_SCN1A, and ASD_mod13_Group1).

To identify the critical cells impacted by the disrupted pathways associated with each module, we used
MoToCC to solve the linear programming problem and vary the upper bound $k$ of the number of cells
to return as a solution from 250 to 5,000 cells. By examining changes in silhouette score, we identified
emphasized clusters of cells that were most relevant to the module and the targeted phenotype.

In general, at all $k$, the total objective function values associated with initial and final solutions tended
to be significantly smaller for randomized modules of the same size as true modules, suggesting that cells
were non-randomly selected according to their correlated gene expression.

The M1_NDD module is widely expressed in the human brain during neurodevelopment, with higher
expression in the fetal brain. Functional analysis shows enrichment in chromatin remodeling, the Wnt
pathway, and the Notch pathway. MoToCC identifies critical cells enriched in progenitor cells PgG2M and
PgS, consistent with recent findings linking these cells to autism genes involved in chromatin remodeling and
the Wnt pathway, such as CHD8.

M2_NDD and M_SCN1A modules contain genes associated with synaptic transmission or neurotransmit-
ter receptors, enriched in terms related to the regulation of neurotransmitter secretion. CSEA analysis shows
selective expression in deep cortical neurons, crucial for functions such as memory formation and perception.
For M2_NDD, ExDp1 and ExDp2 cell types were primarily selected, capturing over 67% of all ExDp2 cells.
Example commands, guidelines, and associated scripts for pre-processing and data visualization are freely
available at [https://github.com/jchow32/MoToCC](https://github.com/jchow32/MoToCC).

6 Acknowledgment

This project was partly funded by the BSF US-Israel Binational Science Foundation grant number 2021092
to FH and SS.

References

over single-cell rna sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis.

[2] Trygve E Bakken, Rebecca D Hodge, Jeremy A Miller, Zizhen Yao, Thuc Nghi Nguyen, Brian Aever-
mann, Eliza Barkan, Darren Bertagnolli, Tamara Casper, Nick Dee, et al. Single-nucleus and single-cell


applications of next-generation sequencing. Cold Spring Harbor Perspectives in Medicine, 9(10):a026898,
2019.


<table>
<thead>
<tr>
<th>Tissue</th>
<th>Abbreviation</th>
<th>Cell-type name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cortex</td>
<td>Mic</td>
<td>Microglia</td>
</tr>
<tr>
<td></td>
<td>Per</td>
<td>Pericyte</td>
</tr>
<tr>
<td></td>
<td>End</td>
<td>Endothelial</td>
</tr>
<tr>
<td></td>
<td>OPC</td>
<td>Oligodendrocyte progenitor cell</td>
</tr>
<tr>
<td></td>
<td>lnCGE</td>
<td>Interneuron caudal ganglionic eminence</td>
</tr>
<tr>
<td></td>
<td>lnMGE</td>
<td>Interneuron medial ganglionic eminence</td>
</tr>
<tr>
<td></td>
<td>ExDp2</td>
<td>Excitatory deep layer 2</td>
</tr>
<tr>
<td></td>
<td>ExDp1</td>
<td>Excitatory deep layer 1</td>
</tr>
<tr>
<td></td>
<td>ExM-U</td>
<td>Maturing excitatory upper enriched</td>
</tr>
<tr>
<td></td>
<td>ExM</td>
<td>Maturing excitatory</td>
</tr>
<tr>
<td></td>
<td>ExN</td>
<td>Migrating excitatory</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Intermediate progenitor</td>
</tr>
<tr>
<td></td>
<td>PgG2M</td>
<td>Cycling progenitors (G2/M phase)</td>
</tr>
<tr>
<td></td>
<td>PgS</td>
<td>Cycling progenitors (S phase)</td>
</tr>
<tr>
<td></td>
<td>oRG</td>
<td>Outer radial glia</td>
</tr>
<tr>
<td></td>
<td>vRG</td>
<td>Ventricular radial glia</td>
</tr>
<tr>
<td>Midbrain</td>
<td>DaNs</td>
<td>Dopaminergic neurons</td>
</tr>
<tr>
<td></td>
<td>OPCs</td>
<td>Oligodendrocyte progenitor cells</td>
</tr>
</tbody>
</table>

Table S1: Cell-type abbreviations.

### 7 Supplementary Material