From Genes to Clusters: COTAN v2 for Improved scRNA-seq Analysis

Silvia Giulia Galfrè[1], Marco Fantozzi[2], Irene Testa[1], Matteo Tolloso[1], Andrea Alberti[1], Alina Sîrbu[1], Francesco Morandin[2], Corrado Priami[1]

[1] Department of Computer Science - University of Pisa, Italy
[2] Department of Mathematics, Physics and Informatics - University of Parma, Italy

Università di Pisa

1. Abstract

- COTAN (CO-expression Tables ANalysis) presents an innovative approach to analyzing scRNA-seq data through gene co-expression analysis, bypassing the need for imputation, normalization, or feature selection.
- COTAN v2 introduces new features and enhancements: (1) cell clustering and cluster marker detection; (2) improved speed; and (3) user-friendly plotting tools.
- The tool has been rigorously evaluated across various scRNA-seq workflow stages and benchmarked against four state-of-the-art libraries, demonstrating its effectiveness.

2. Methods

- **Gene Correlation**
  Estimated by calculating the deviation between the observed and expected gene co-expression in cells (both represented by 2×2 contingency tables) under the assumption of gene independence. For large datasets, these correlations can be translated into p-values.
- **Differential Gene Expression Analysis**
  The enrichment of gene i in the cell subset A is estimated by computing the deviation between the observed and expected occurrences of gene i expression inside and outside A. This produces an enrichment matrix with genes as rows and cell subsets as columns.
- **Global Differentiation Index** (GDI)
  The GDI of gene i is defined as $\log(-\log(\bar{p}_i))$, where $\bar{p}_i$ represents the average of the lowest 5% of p-values evaluating the independence between gene i and all other genes.
- **Transcriptomic Uniformity** (TU)
  If less than 1% of genes in a cell set have $\texttt{GDI} > 1.4$, the set is considered transcriptionally uniform, indicating a lack of distinct sub-populations.
- **Two-Phase Clustering**
  - *Split phase*: 1) Apply Seurat clustering tool. 2) Keep TU clusters; combine others and reapply Seurat until uniformity is achieved.
  - *Merge phase*: 1) Assess cluster similarity by calculating the cosine distance between the columns of the enrichment matrix. 2) Merge similar clusters if doing so meets the TU criterion.

3. Data

Dataset	Tissue	Method	#Cells
E13.5 Yuzwa	mouse forebrain	Drop-seq	1112
E17.5 Yuzwa	mouse forebrain	Drop-seq	874
E14.5 Loo	mouse forebrain	Drop-seq	10864
E13.5 La Manno	mouse forebrain	10x	4981
E15.0 La Manno	mouse forebrain	10x	8562
E17.5 La Manno	mouse forebrain	10x	2467
CD14	human blood	10x	2434

Table 1: *Datasets details.*

4. Results

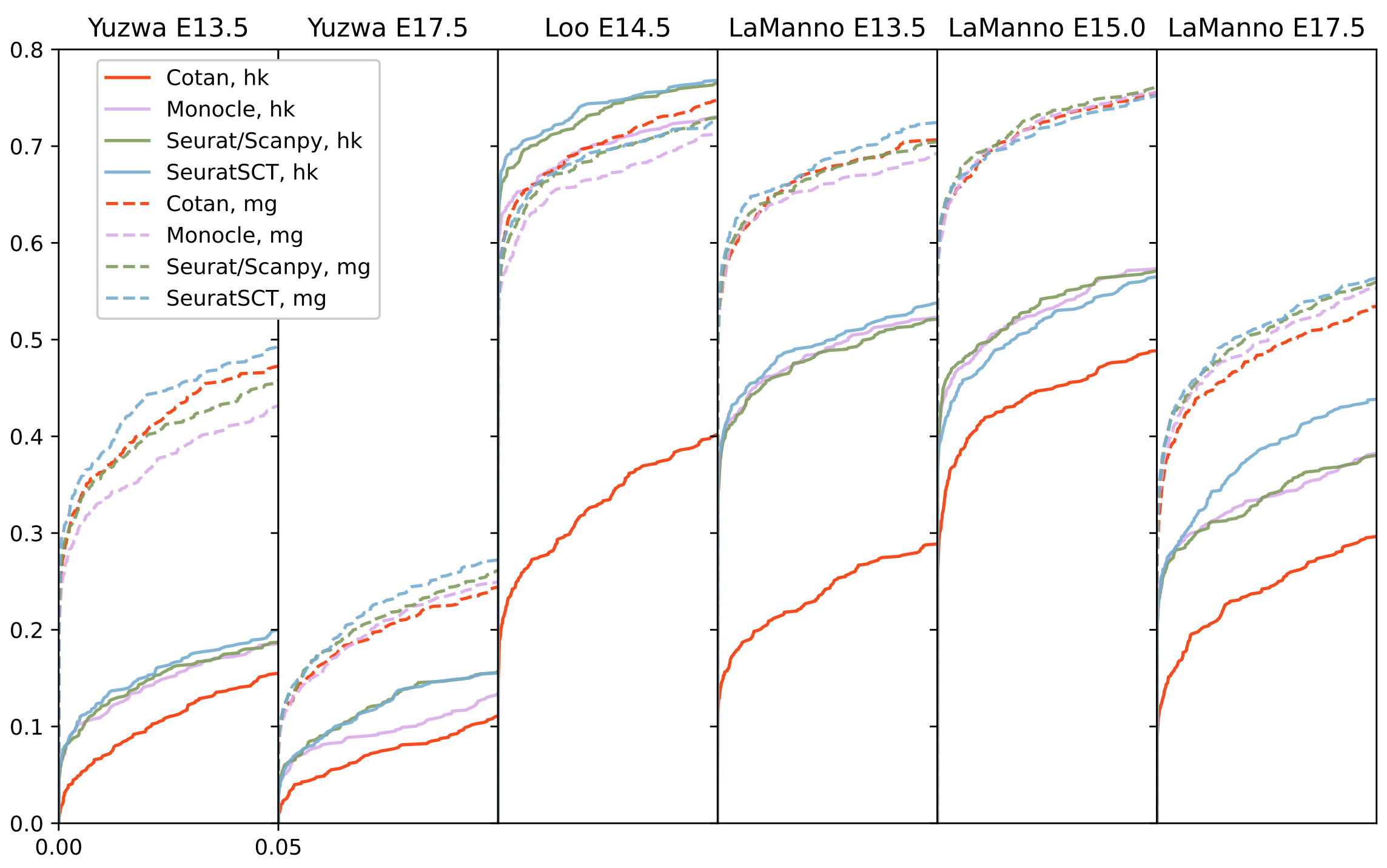

Figure 1: *Empirical distribution functions (eCDF) of p-values for correlations between pairs of neural marker genes (MG, dashed lines) and housekeeping genes (HK, solid lines) across six mouse forebrain datasets.*

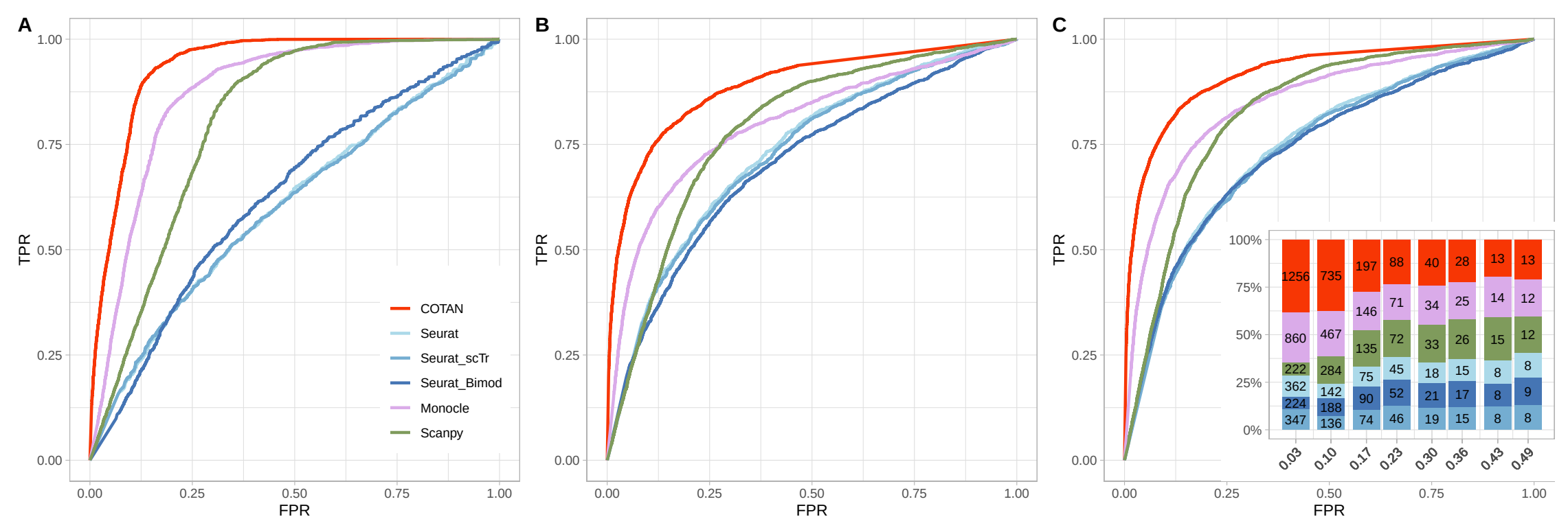

Figure 2: *ROC curves for detecting gene enrichment across subsets of two, three, and five cell types in the La Manno datasets (panels A, B, and C, respectively). The bottom-right bar chart illustrates the number of true positive genes at $FPR = 0.05$, with genes grouped based on their average read counts.*

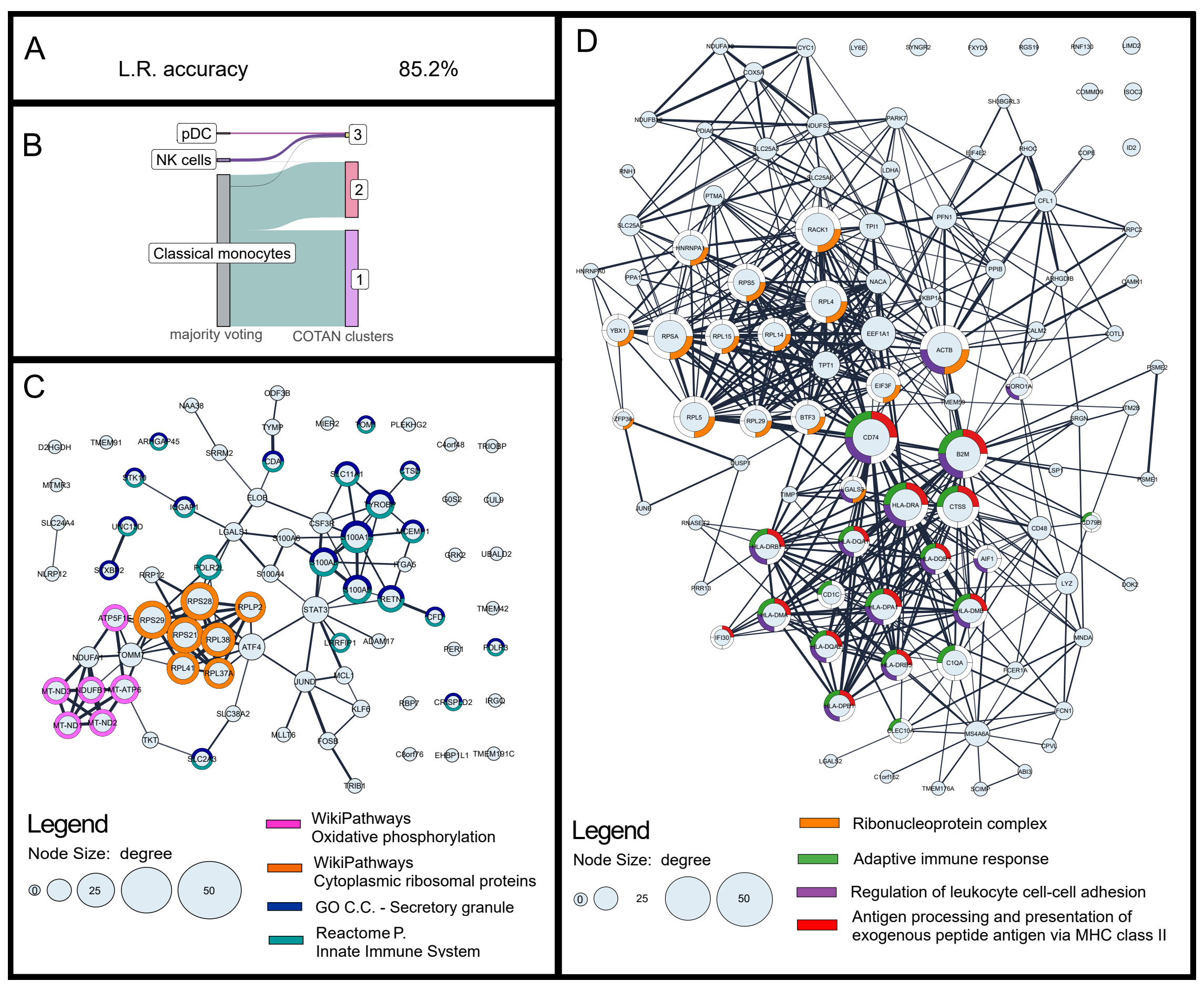
A
L.R. accuracy
85.2%
B
pDC
NK cells
Classical monocytes
majority voting
COTAN clusters
C
D
Legend
Node Size: degree
25
50
WikiPathways Oxidative phosphorylation
WikiPathways Cytoplasmic ribosomal proteins
GO C.C. - Secretory granule
Reactome P. Innate Immune System
Ribonucleoprotein complex
Adaptive immune response
Regulation of leukocyte cell-cell adhesion
Antigen processing and presentation of exogenous peptide antigen via MHC class II

Figure 3: *Analysis of CD14 dataset clustering.* ***A*** *Accuracy of logistic regression using COTAN's clusters as targets.* ***B*** *Sankey plot comparing cluster assignments from CellTypist (left) and COTAN (right).* ***C*** *and* ***D*** *STRING network analysis, performed using Cytoscape, of the top 100 enriched genes for COTAN's clusters 1 and 2. Cluster 1 is enriched in Innate Immune System genes, while cluster 2 is enriched in Adaptive Immune System genes.*

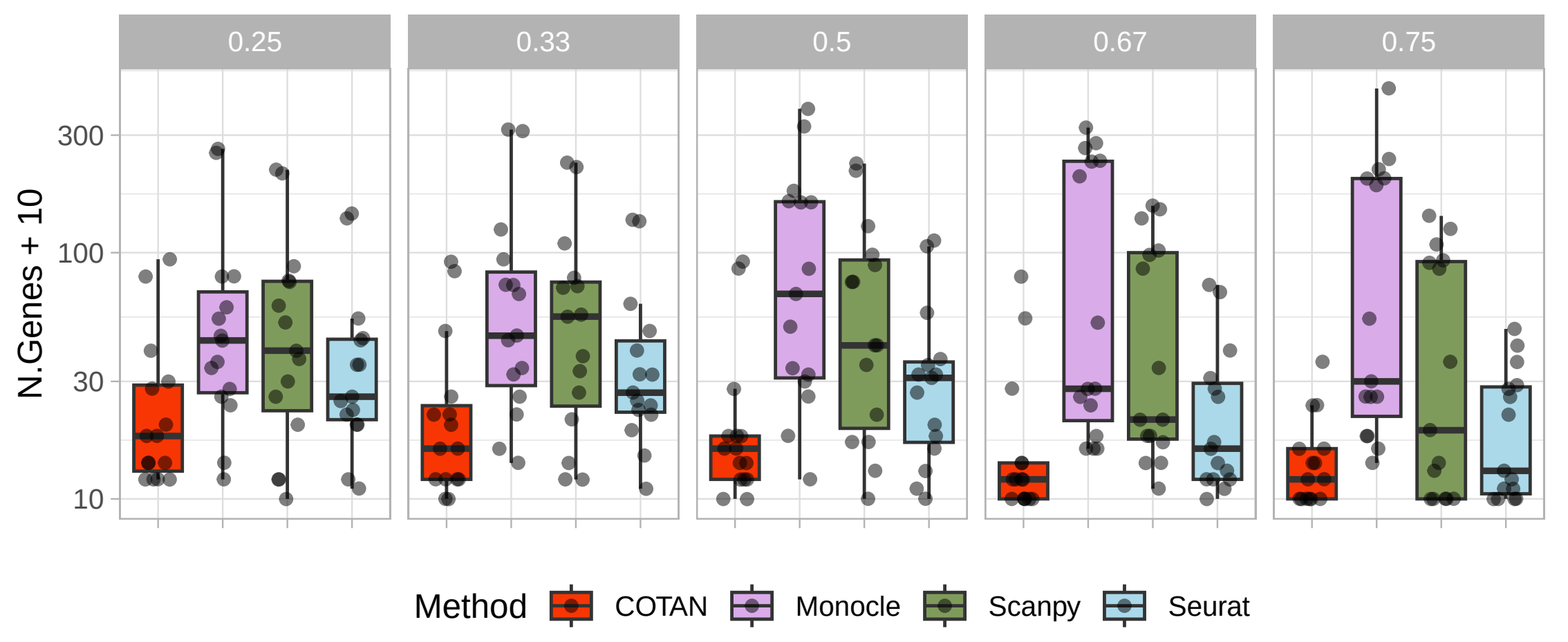

Figure 4: *Number of genes erroneously detected as significantly enriched across different partitions of 15 uniform clusters from the La Manno datasets. The clusters were randomly split into high and low library size cells in five different proportions (indicated at the top of each panel).*

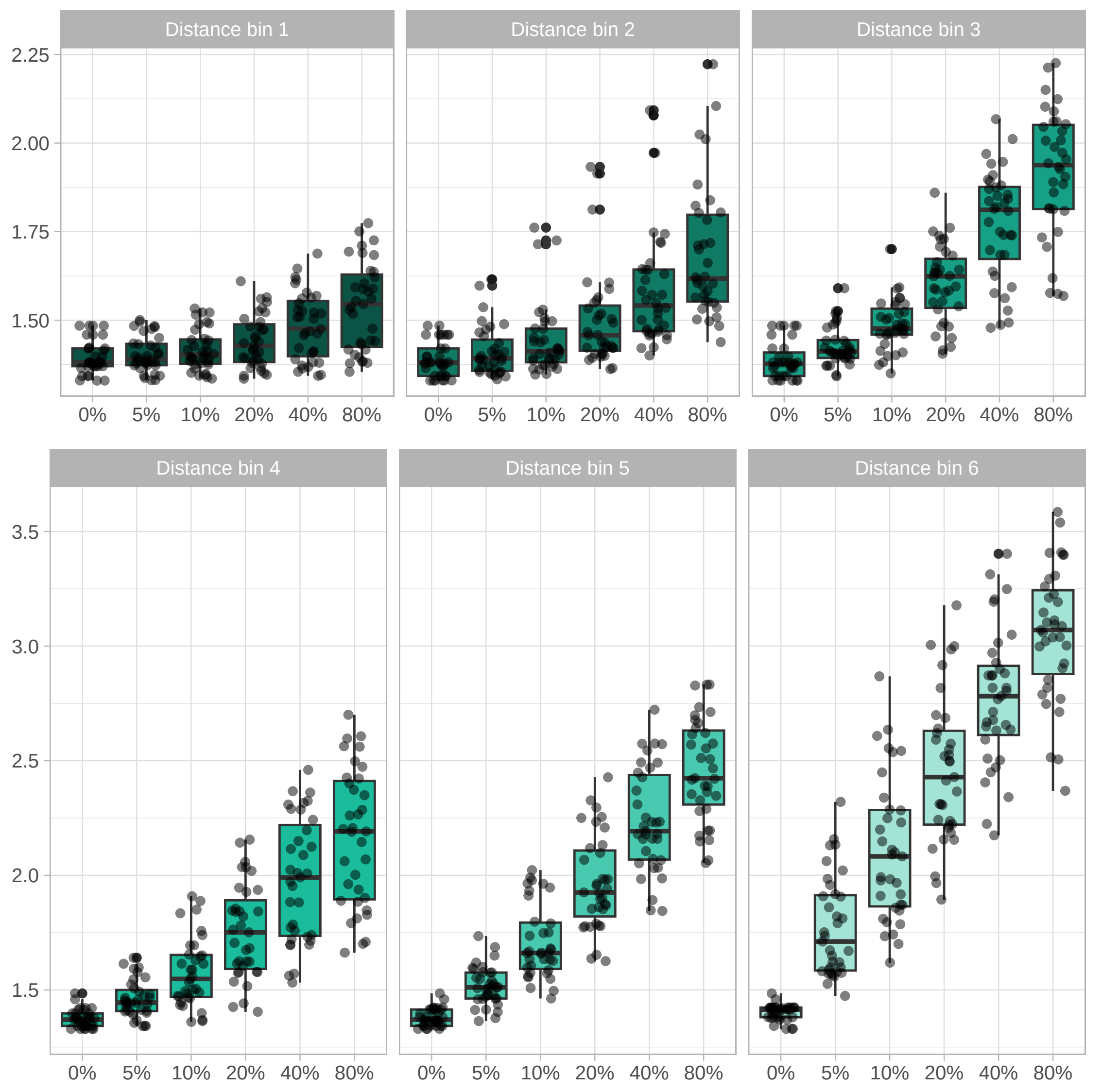

Figure 5: *Variation of the top 1% GDI values as the proportion of cells from a second type increases in two-population subsets of the La Manno datasets. The datasets were grouped into six bins based on the distance between the two populations of cells.*

Dataset	Label	Cluster Size	Base Acc.	Val. Acc.
E15.0 La Manno	Cl525	334, 316, 176	0.36	0.76
E15.0 La Manno	Cl432	403, 183	0.57	0.84
E15.0 La Manno	Cl511	397, 143	0.61	0.95
E13.5 La Manno	Cl432	289, 247	0.50	0.92
E13.5 La Manno	Cl186	246, 177	0.51	0.87
E15.0 La Manno	Cl184	228, 94	0.59	0.80

Table 2: *Evaluation of COTAN's clustering through logistic regression. Analysis involved normalizing and log-transforming raw counts, selecting up to 150 highly variable genes (100 with the highest variance and 50 with the highest GDI), and training a logistic regression model with an 80 : 20 training-validation split. "Base Acc." is the expected accuracy for a random partition with the same proportions; "Val Acc." is the accuracy on the validation set.*

5. Conclusions

- **Unique Approach**: Innovative method for analyzing scRNA-seq data by focusing on gene co-expression and incorporating lowly expressed genes.
- **Sparse Data Handling**: Efficiently handles sparse datasets without requiring imputation, normalization, or feature selection, preventing biases and information loss.
- **Enhanced Usability**: Reduces computational time with Torch-optimized functions and offers new, user-friendly plotting tools.
- **Robustness**: Demonstrates robustness against false positive correlations in gene expression (Figure 1).
- **Superior Performance in Marker Detection**: Outperforms other tools in gene marker detection by enhancing statistical power (Figure 2) and reducing false positives (Figure 4).
- **Biologically Informed Clustering**: Creates clusters with uniform transcriptomic profiles using the GDI score, validated for sensitivity (Figure 5) and effectiveness (Table 2), and demonstrated to yield biologically relevant clusters (Figure 3).
- **Availability**: COTAN is available as an R package on Bioconductor and GitHub.