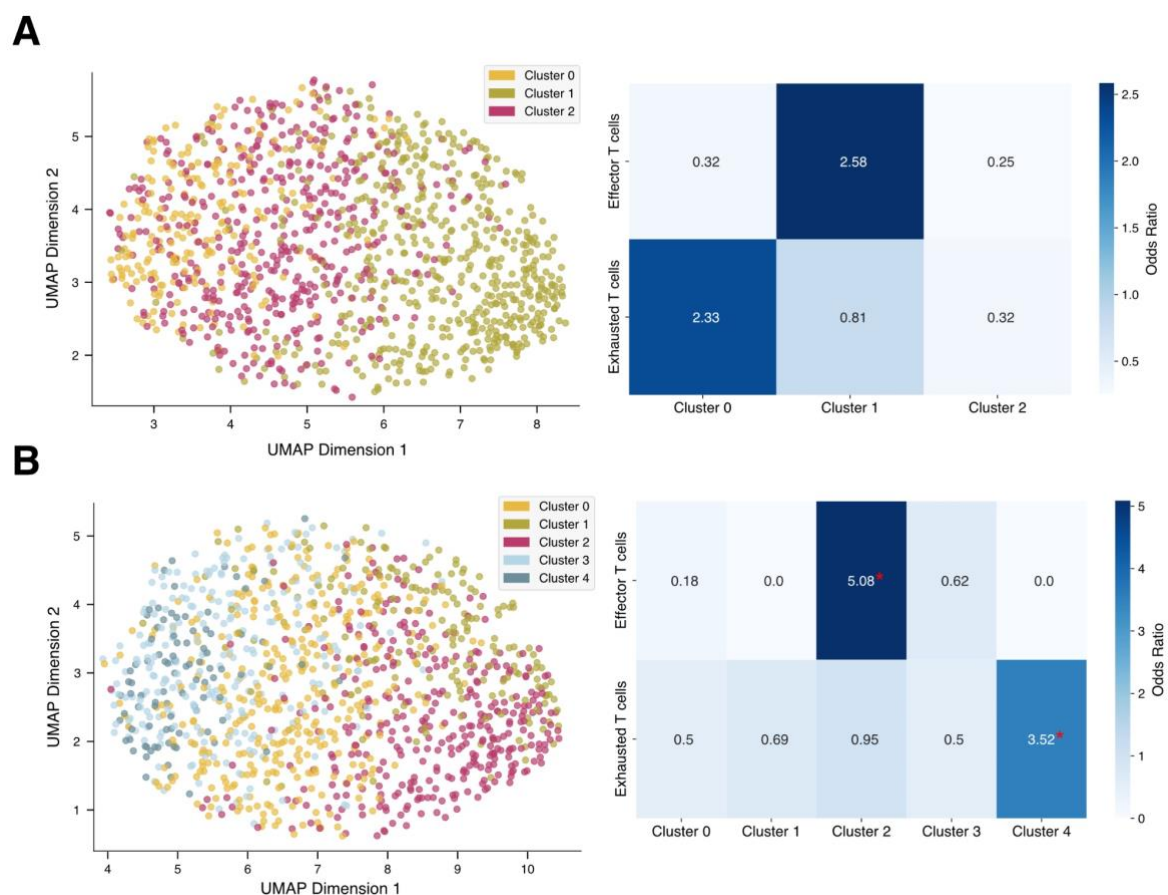


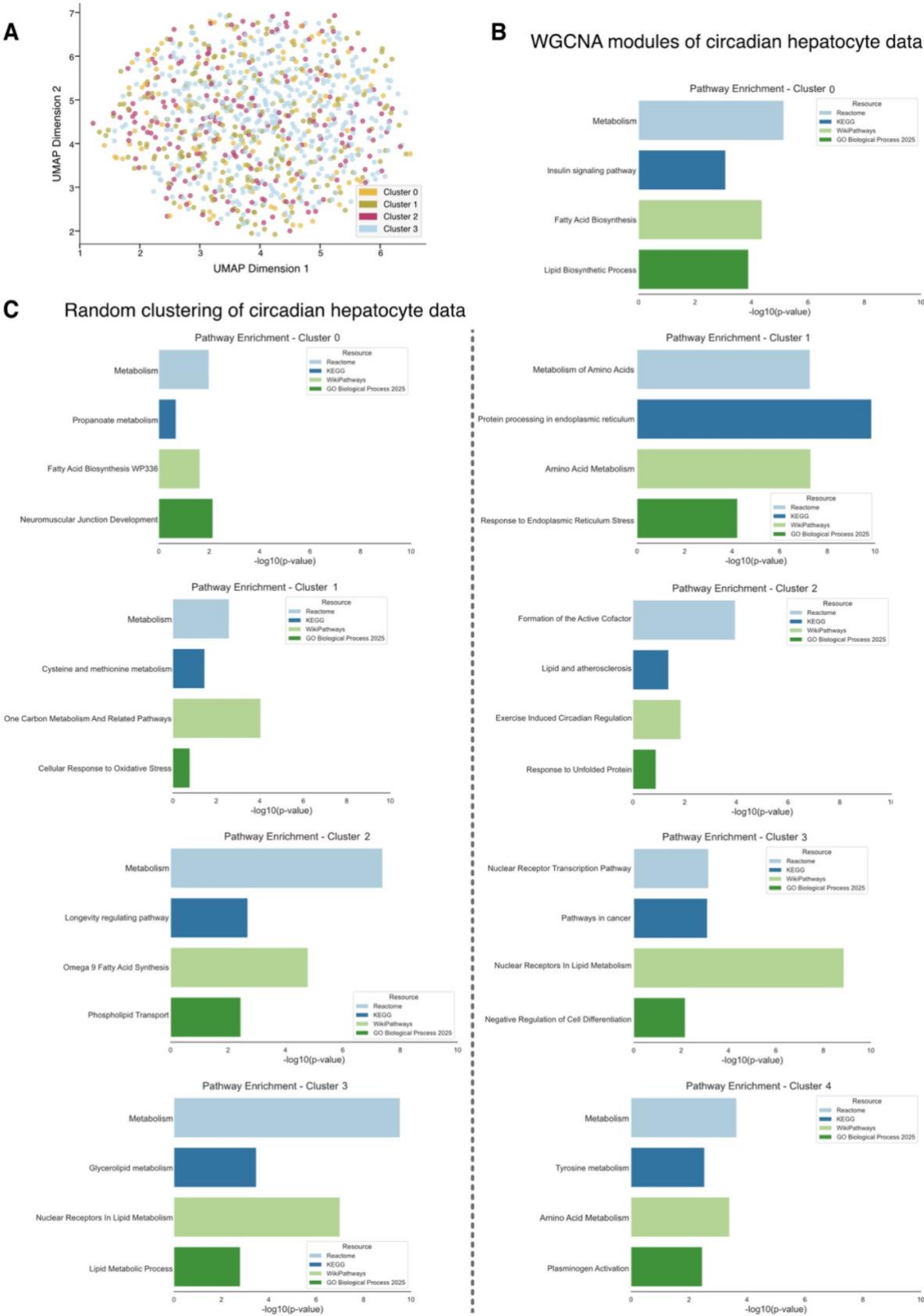
## **InCURA: Integrative gene clustering based on transcription factor binding sites**

Lorna Rinck<sup>1,2</sup>, Ricardo O. Ramirez Flores<sup>3</sup>, Julio Saez-Rodriguez<sup>2,3</sup>, Mahak Singhal<sup>1,4</sup>

**Supplementary Data**  
**(Figure S1-S2 and Table 1)**



**Supplementary Figure S1. Evaluation of different k values in T cell exhaustion case study.** (A) InCURA clustering with  $k = 3$ . Left: UMAP visualization of the resulting gene clusters. Right: Corresponding odds ratios for enrichment of effector T cell and exhausted T cell signatures across clusters. (B) InCURA clustering with  $k = 5$ . Left: UMAP visualization of resulting gene clusters. Right: Corresponding odds ratios for enrichment of effector T cell or exhausted T cell signatures across clusters. A red asterisk indicates statistically significant enrichment ( $p\text{-value} \leq 0.05$ ).



**Supplementary Figure S2. Random clustering and hdWGCNA cannot capture distinction between circadian and metabolic regulatory programs.** (A) UMAP visualization of gene clusters identified by random, based on DEGs derived from the analysis of single-cell RNA-seq data from adult mouse hepatocytes of REV-ERB $\alpha/\beta$  double-knockout vs. control liver tissues. (B) Enriched pathways in hdWGCNA modules as indicated by the  $-\log_{10}(\text{p-values})$  based on four different resources: Reactome (light blue), KEGG (dark blue), WikiPathways (light green) and GO Biological Processes (dark green). From top to bottom: Module/cluster 0 - 4. (C) Enriched pathways in random clusters as indicated by the  $-\log_{10}(\text{p-values})$  based on four different resources: Reactome (light blue), KEGG (dark blue), WikiPathways (light green) and GO Biological Processes (dark green). From top to bottom: Cluster 0 - 3.

**Supplementary Table 1. External data used in case studies**

Case Study	Data Type		Species	Accession	Authors	Year	Link
T cell exhaustion	Bulk	RNA-seq	mouse	GSE132987	Khan <i>et al.</i>	2019	-
SLE B cells	Bulk-cell RNA-seq		human	GSE110999	Wang <i>et al.</i>	2018	-
Circadian double knockout	Single-cell RNA-seq		mouse	GSE143528	Guan <i>et al.</i>	2021	-
Mouse gastrulation	DEGs only		mouse	GSE169210	Mittenzweig <i>et al.</i>	2021	<a href="https://apps.tanaylab.com/MCV/embflow/">https://apps.tanaylab.com/MCV/embflow/</a>
MCF7 breast cancer cells	Genes linked to DARs		human	GSE174152	Wang <i>et al.</i>	2021	-