Discovering Novel Ciliary Genes Through Machine Learning

Torriglia, Emilia [1] ; Irigoin, Florencia [2, 3] ; Romanelli, Laura [4] ; Pazos Obregón, Flavio [1]

1- Analytical Biochemistry and Proteomics Unit, IIBCE - Institut Pasteur Montevideo, **2-** Academic Unit of Histology and Embryology, Faculty of Medicine, UDELAR, **3-** Human Molecular Genetics Laboratory, Institut Pasteur Montevideo, **4-** Worm Biology Laboratory, Institut Pasteur Montevideo

Background

The primary cilium is an evolutionarily conserved organelle present in many eukaryotic cells, where it plays a central role in cellular signaling and environmental sensing (Hilgendorf *et al.* 2024). Despite its importance, the full set of genes required for its structure and function remains incomplete. Although the cilium has a characteristic protein composition, no consensus sequences or domains have been identified that allow their identification. We hypothesized that genes associated with ciliary function share expression-based features and other biological characteristics, independent of DNA sequence similarity, that can be used for their identification. Caenorhabditis elegans provides a powerful system to study ciliary biology due to its well-characterized nervous system, the ease of genetic and experimental manipulation, and the availability of abundant high-quality publicly accessible transcriptomic data, enabling the identification of conserved mechanisms across species.

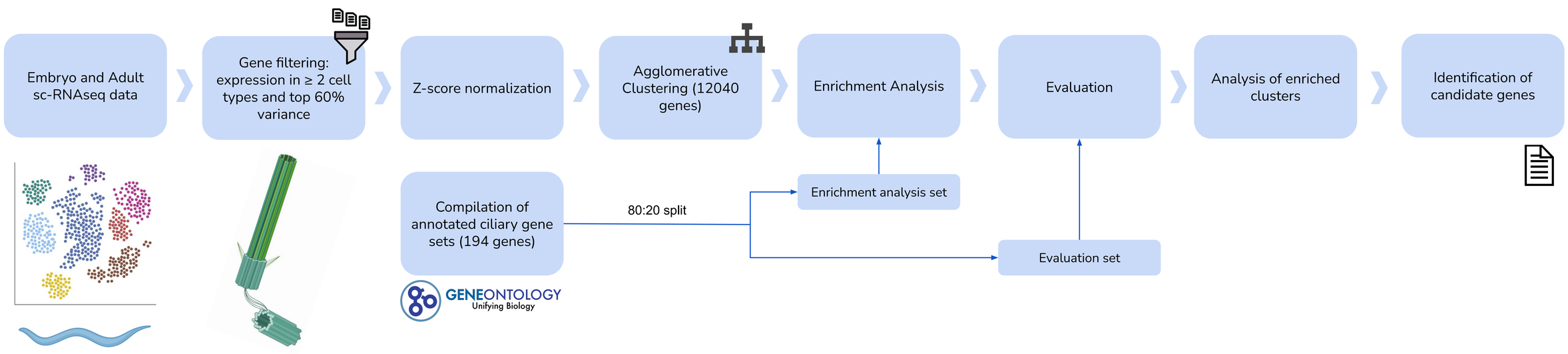
Methodology

To identify candidate ciliary genes, we constructed a transcriptomic dataset from publicly available single-cell RNA sequencing data (Packer *et al.* 2019; Taylor *et al.* 2021). These datasets include 69 neuronal cell types across C. elegans developmental stages, comprising both ciliated and non-ciliated populations, and were integrated prior to analysis. We constructed a combined dataset of neuronal gene expression, merging embryo (31 neuronal types: 21 ciliated, 10 non-ciliated) and adult (38 neuronal types: 18 ciliated, 20 non-ciliated) data.

Genes were filtered based on expression across different cell types and variability, retaining those expressed in at least two neuronal cell types and within the top 60% of variance, yielding 12,040 genes and including 194 annotated as ciliary genes in Gene Ontology. The data was then separated into embryo and adult subsets and Z-score normalized per gene.

Agglomerative clustering (using Ward linkage) was performed independently on embryo and adult datasets, and for ciliated and non-ciliated neurons. In each case, we explored various distance thresholds and selected the one that resulted in the most enriched clusters in ciliary genes. For this analysis, annotated ciliary genes were randomly split in two sets, and we only used 80% of the annotated ciliary genes to identify the enriched clusters.

After removing ciliary genes used to calculate enrichment, genes in enriched clusters were considered candidate genes. Within each developmental stage, the list of candidate genes was further refined by selecting those genes found in both ciliated and non-ciliated neuron enriched clusters. Evaluation of this procedure was performed by testing enrichment in the held-out 20% ciliary genes, and the final set of candidate genes was obtained by intersecting embryo and adult candidates.



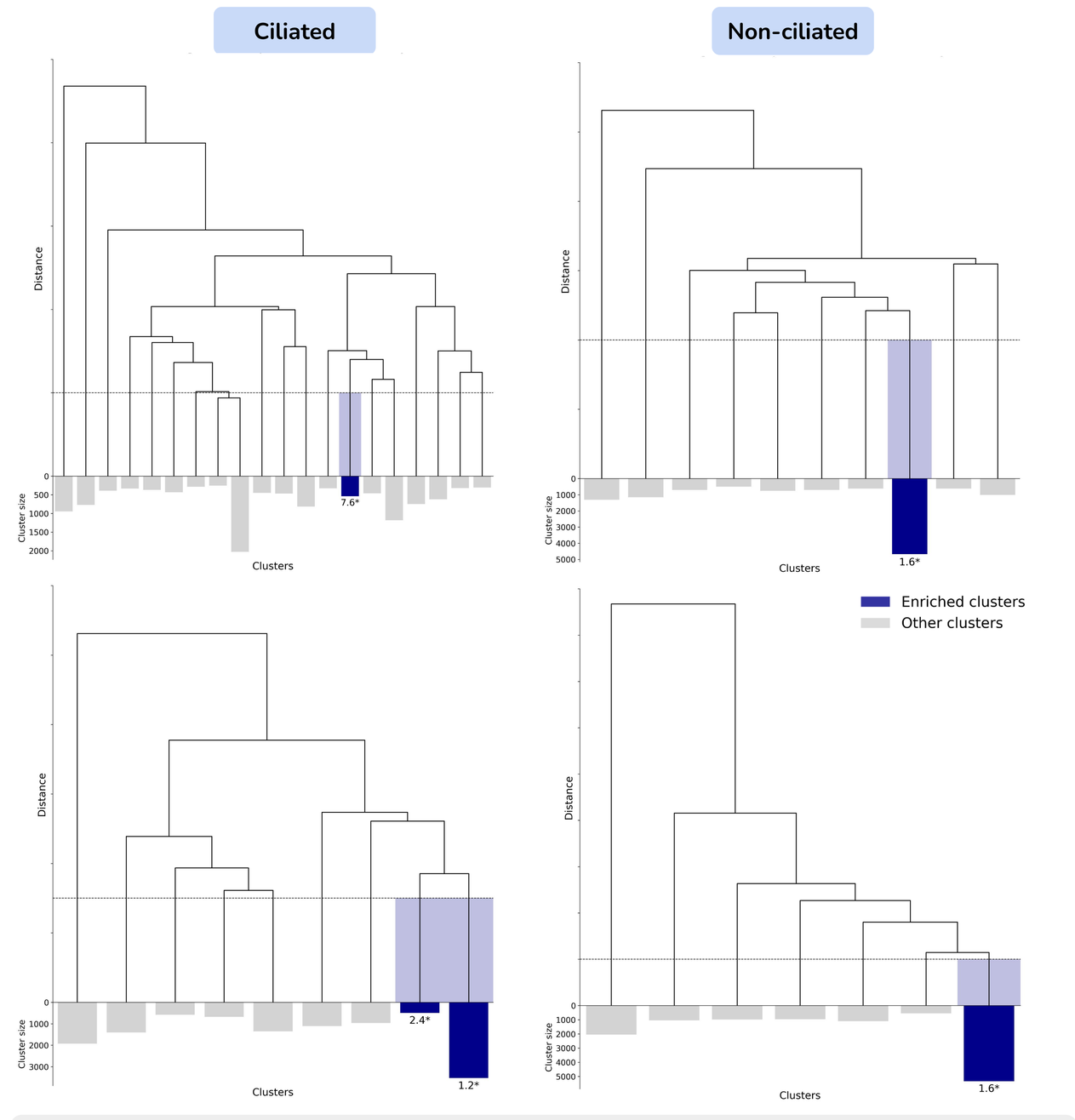
Results

A total of **5 clusters** significantly enriched in ciliary genes were identified across all conditions. The set of genes resulting from the intersection of these clusters was significantly enriched in the held-out 20% of ciliary genes. The whole procedure yielded a candidate set of **55 genes** that share expression patterns across both ciliated and non-ciliated neurons in embryos and adults, and prove to be interesting prospects for further analysis.

Conclusions and perspectives

This study identified **55 genes** that can be considered as candidates to be associated with ciliary function in *C. elegans*. By intersecting enriched clusters across developmental stages and cell types, we obtained a set of genes that co-clusters with known ciliary genes.

Our next steps include refining this list of candidates by integrating unsupervised analysis of other sequence-independent variables, such as functional landscape arrays and patterns of structural motif localization, evaluating alternative ciliary gene sources and other clustering algorithms. Furthermore, we will combine this approach with the results of supervised learning models, aiming to refine our list of candidates before performing experimental validation.



Dendrograms showing enriched clusters (blue) for all scenarios. Dotted line corresponds to chosen threshold. Fold enrichment is shown below the highlighted bars (p<0.05)

References

Hilgendorf K, Myers B, Reiter J. (2024). Emerging mechanistic understanding of cilia function in cellular signalling. Nature reviews. Molecular cell biology.

Packer J. *et al.* (2019). A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution science.

Taylor S. (2021). Molecular Topography of an Entire Nervous System. Cell.