

Discovering Novel Ciliary Genes Through Machine Learning

Torriglia, Emilia¹; Irigoien, Florencia^{2,3}; Romanelli-Cedrez, Laura⁴; Pazos Obregón, Flavio¹

1- Analytical Biochemistry & 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 complete set of genes required for its structure and function remains incompletely characterized. Although cilia have a characteristic protein composition, no consensus sequences or domains enable their systematic identification. We hypothesized that ciliary genes share expression-based and other biological features, independent of DNA sequence similarity, that can be used for their identification. *Caenorhabditis elegans* provides a powerful model for studying ciliary biology due to its well-characterized nervous system, ease of experimental manipulation, and abundant publicly available transcriptomic data.

We constructed a transcriptomic dataset from publicly available single-cell RNA sequencing data (Packer *et al.* 2019; Taylor *et al.* 2021). Agglomerative clustering was performed independently on embryo and adult datasets, and clusters enriched in ciliary genes were identified. Genes co-clustering with known ciliary genes were selected as candidates for further analysis.

Functional Landscape Arrays (FLAs) provide a numerical representation of a gene's genomic context (Pazos Obregón *et al.* 2022). Weighted FLAs (wFLAs) were constructed using genomic windows of different sizes around each gene and quantifying the presence of functional groups within these regions. These features were used to train and evaluate different classifiers, and the best-performing model was applied genome-wide to generate a second set of candidate genes. Finally, candidates from both approaches were integrated, yielding a preliminary set of 56 genes potentially associated with ciliary function.

Methodology

Unsupervised - Transcriptomics

Compilation of annotated ciliary gene sets with **experimental evidence in *C. elegans* (positive genes)** from Gene Ontology

Negative genes chosen randomly from a set of genes excluding all genes with any cilia-related annotation, at least 100 kb away from selected positives, and keeping a 50:50 ratio of positives and negatives within a given chromosome.



Supervised - wFLAs

Embryo and Adult **sc-RNAseq** data (Packer *et al.*, 2019, Taylor *et al.* 2021)

Gene filtering: expression in ≥ 2 cell types and top 60% variance

Z-score normalization

Agglomerative Clustering (12040 genes)

Enrichment analysis in positive genes

Identification and characterization of enriched clusters and **co-clustering genes**

Preliminary candidate list obtained from co-clustering genes classified as **positive** by RandomForest model trained on wFLAs

wFLA calculation with $w = 10, 20, 50$, and 100, and GO terms with depth ≥ 4 and at least 20 annotated genes.

Genes were divided into **training and test sets** at an **80:20** ratio per chromosome, with at least **15 kb** separating genes from the two sets.

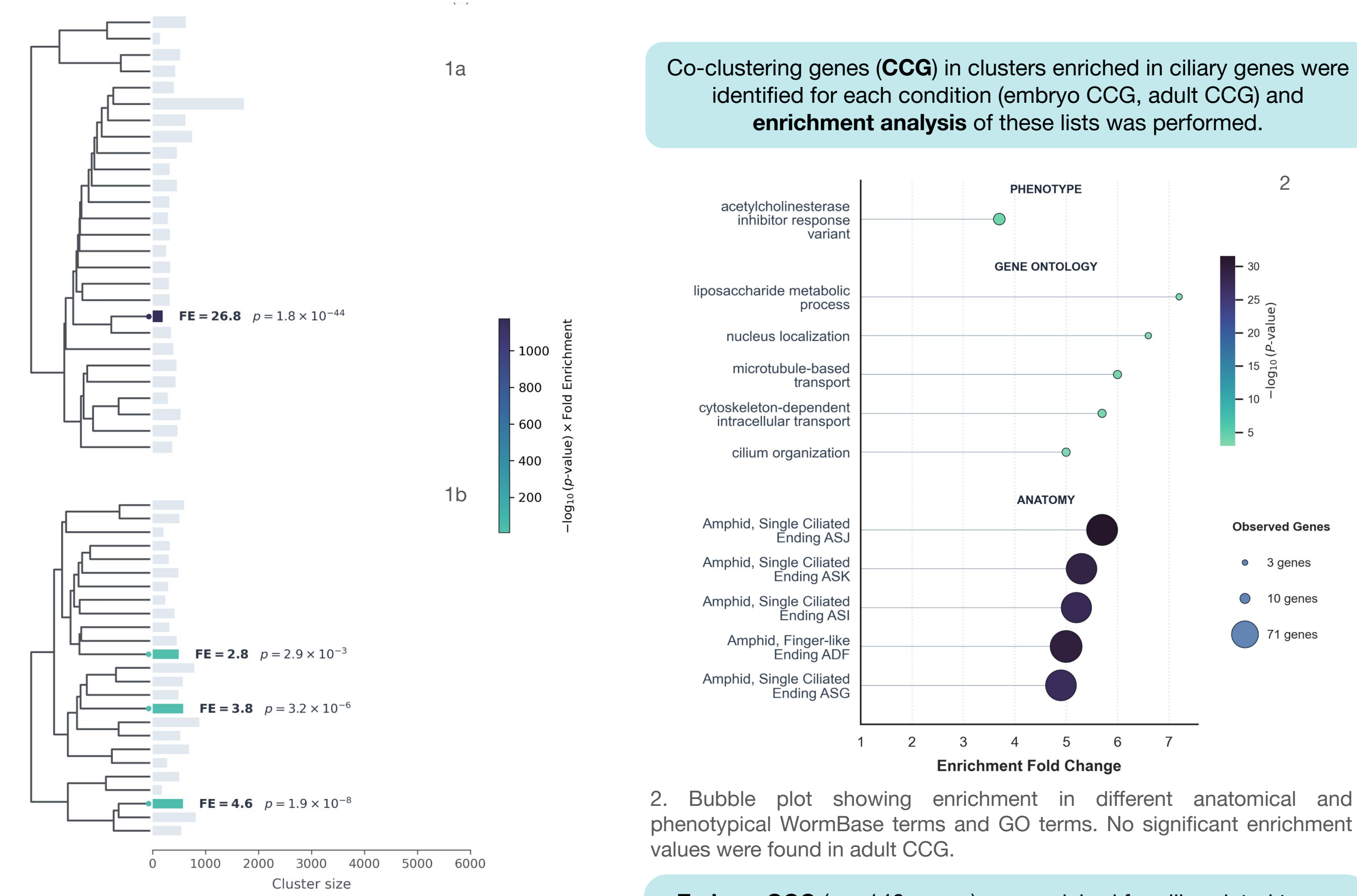
RandomForest training, hyperparameter and decision threshold optimization

Model evaluation

Classifier applied to whole genome to obtain probability of belonging to the **positive class**

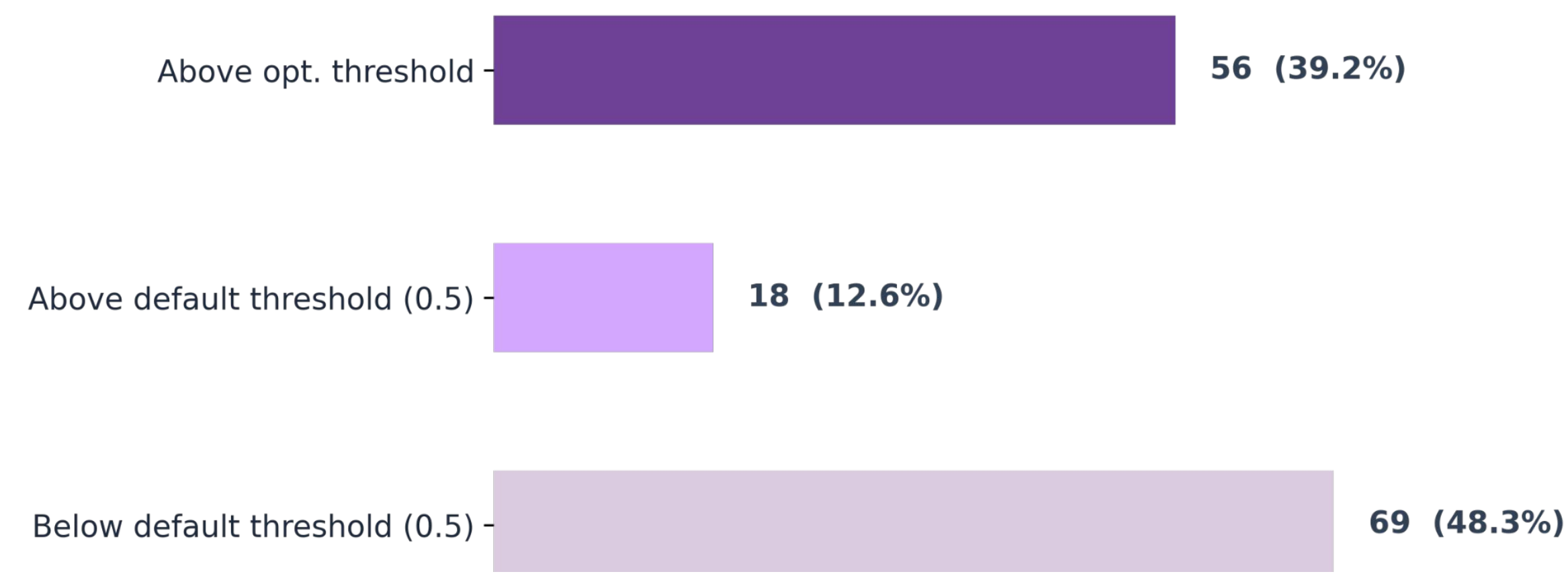
Selection of genes with probability scores over optimized threshold

Agglomerative Clustering of sc-RNA seq data

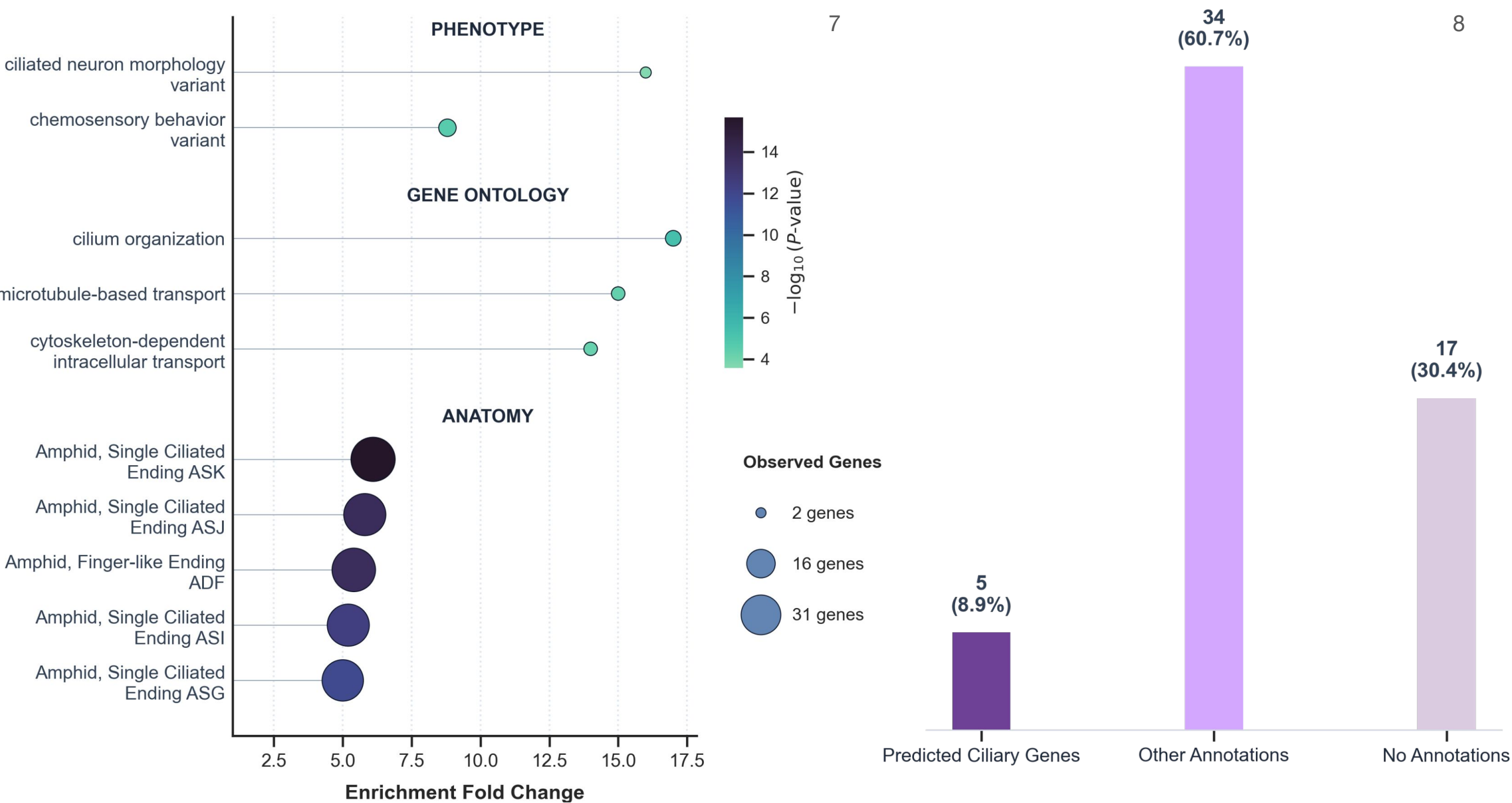


Classification of embryo CCG

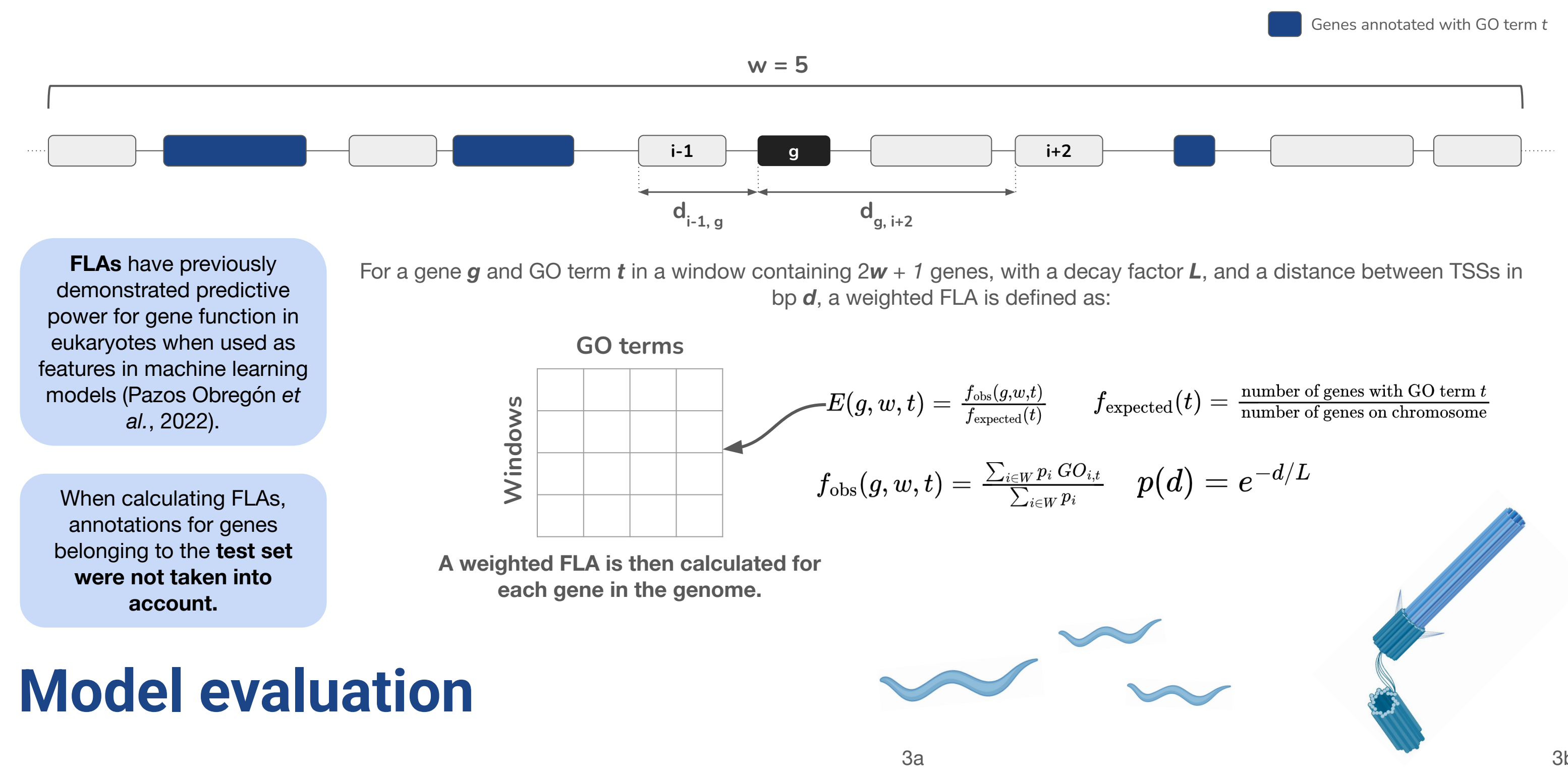
The classifier was applied to the embryo CCG list. The results were as follows:



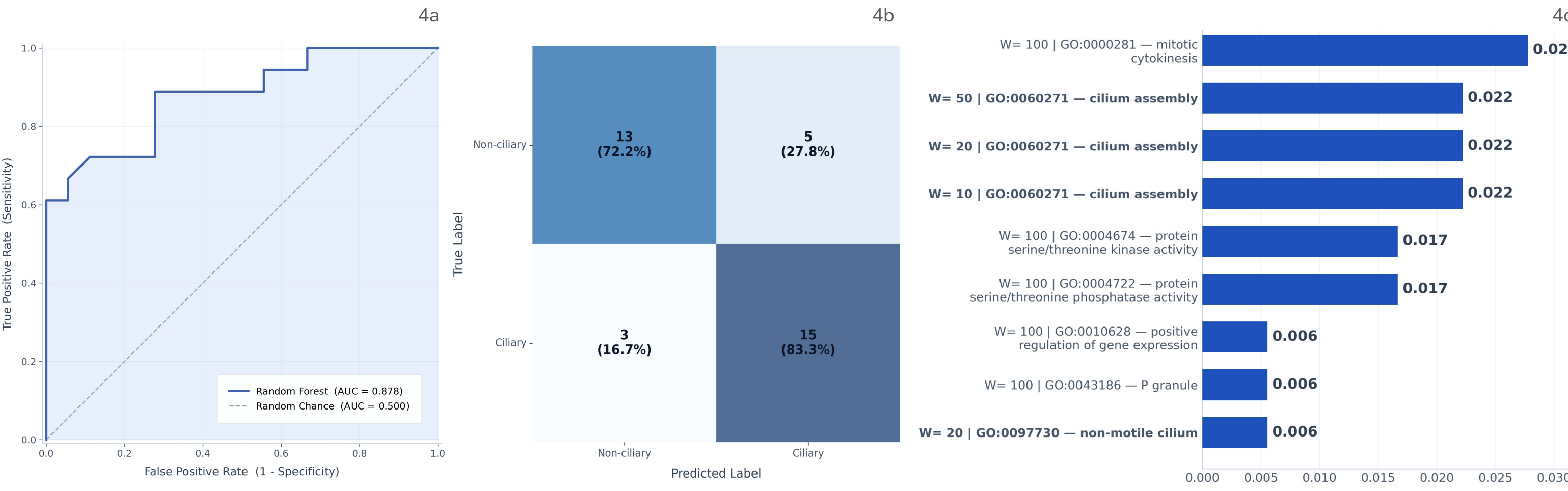
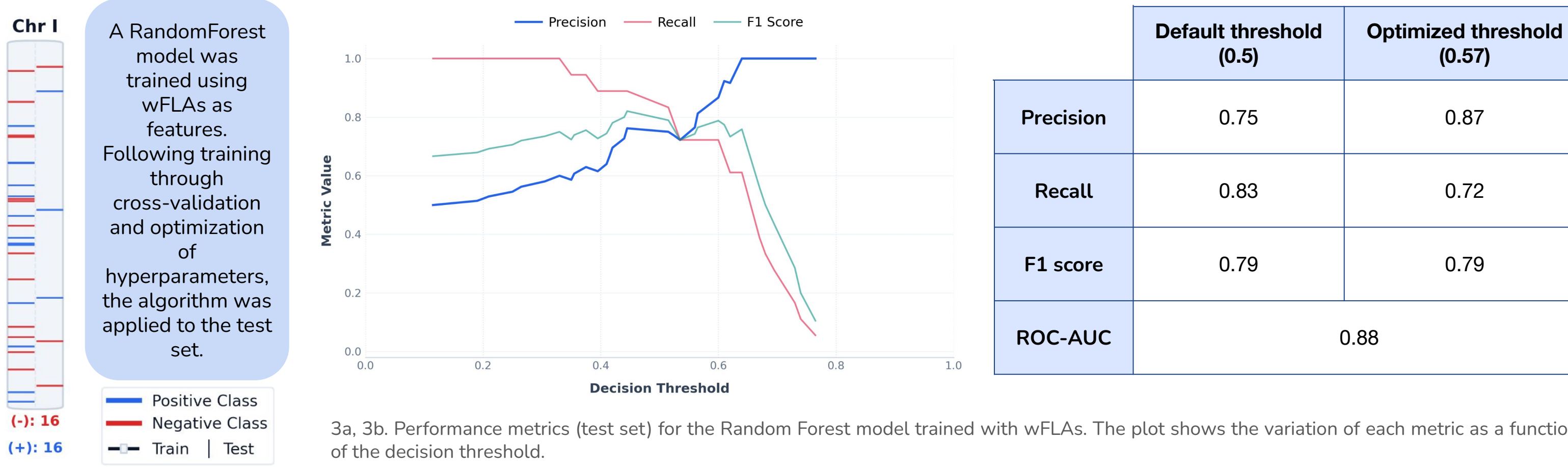
The **56 genes** with probability scores **above the optimal threshold** were selected as a **preliminary candidate list**. Enrichment analysis was performed and previously existing annotations were examined.



wFLA (Weighted Functional Landscape Array)

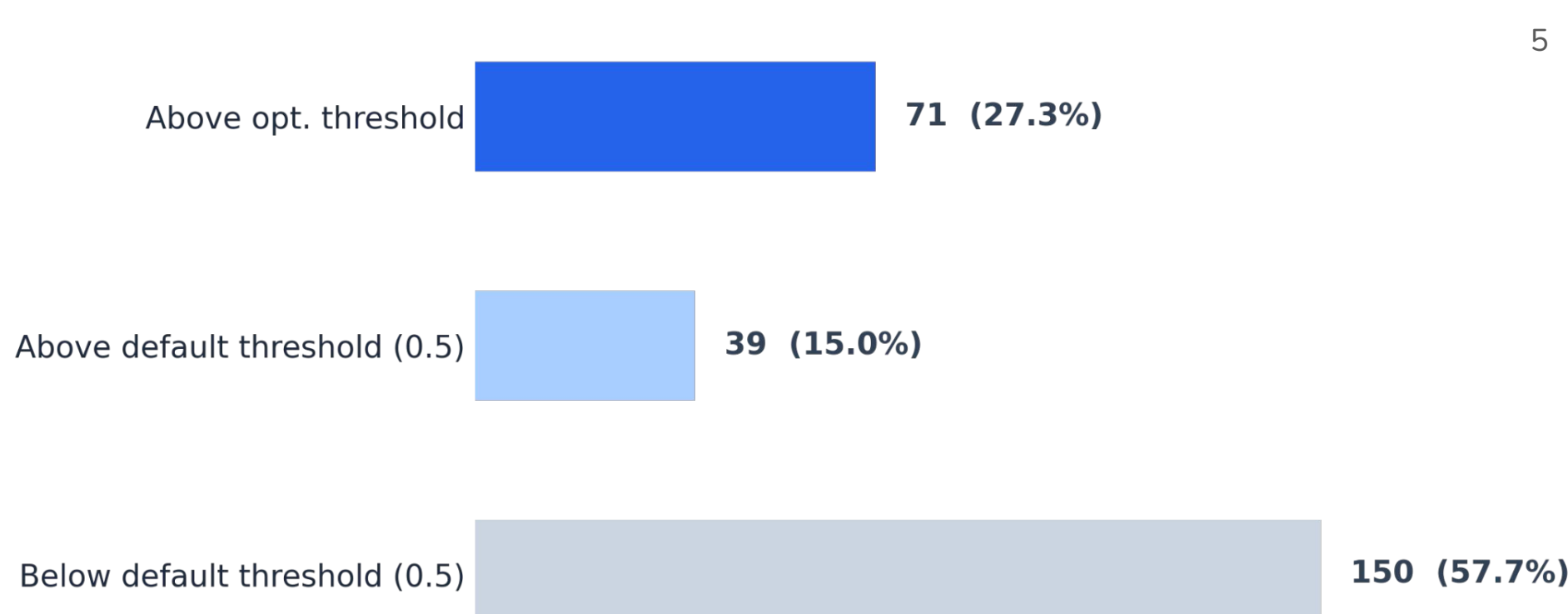


Model evaluation



4a. ROC curve, 4b. confusion matrix for the model and 4c. top 9 values of mean decrease in impurity (permutation importance) for the features in wFLAs.

The model was evaluated a second time by classifying predicted ciliary genes with no experimental annotations. Of 260 genes, the results were as follows:



5. Bar chart showing the distribution of assigned probability scores of belonging to the positive class for ciliary genes with no experimental annotations.

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.
- Pazos Obregón F. *et al.* (2022). Gene function prediction in five model eukaryotes exclusively based on gene relative location through machine learning. Nature Scientific Reports.
- Gaudet P *et al.* (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Briefings in Bioinformatics.
- Simonis N *et al.* (2009) Empirically-controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. Nature Methods.

Conclusions & perspectives

This study identified 56 genes that can be considered candidate genes associated with ciliary function in *C. elegans*. By identifying enriched clusters across developmental stages and cell types, we obtained a set of genes that co-cluster with known ciliary genes. We then trained a classifier using weighted Functional Landscape Arrays (wFLAs), which was used to identify genes of interest among these co-clustering genes.

Our next steps include refining this list of candidates by integrating supervised and unsupervised analyses of other sequence-independent features, such as patterns of structural motif localization and additional transcriptomic datasets. We will also evaluate alternative sources of known ciliary genes, as well as other clustering algorithms and classifiers. Furthermore, we aim to improve the wFLA model by reducing the effect of physical proximity, which results in highly similar wFLAs among neighboring genes regardless of their function. Ultimately, we aim to refine the candidate list using complementary computational and biological criteria before proceeding to experimental validation.

