

Búsqueda mediante aprendizaje automático de nuevos genes ciliares

Torriglia, Emilia ¹ ; Irigoín, Florencia ^{2,3} ; Romanelli, Laura ⁴ ; Pazos Obregón, Flavio ¹

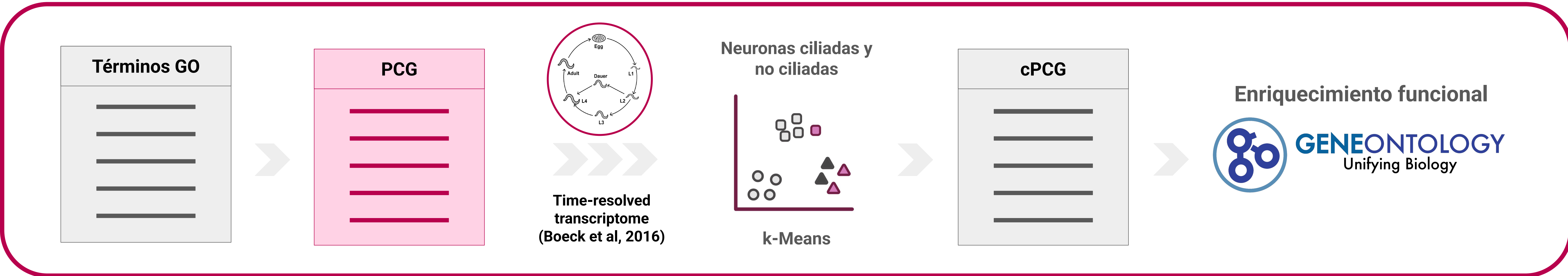
1- Unidad de Bioquímica y Proteómica Analíticas, IIBCE - Institut Pasteur Montevideo, 2- U.A. de Histología y Embriología, Facultad de Medicina, UDELAR, 3- Laboratorio de Genética Molecular Humana, Institut Pasteur Montevideo, 4- Laboratorio de Biología de Gusanos, Institut Pasteur Montevideo

Resumen

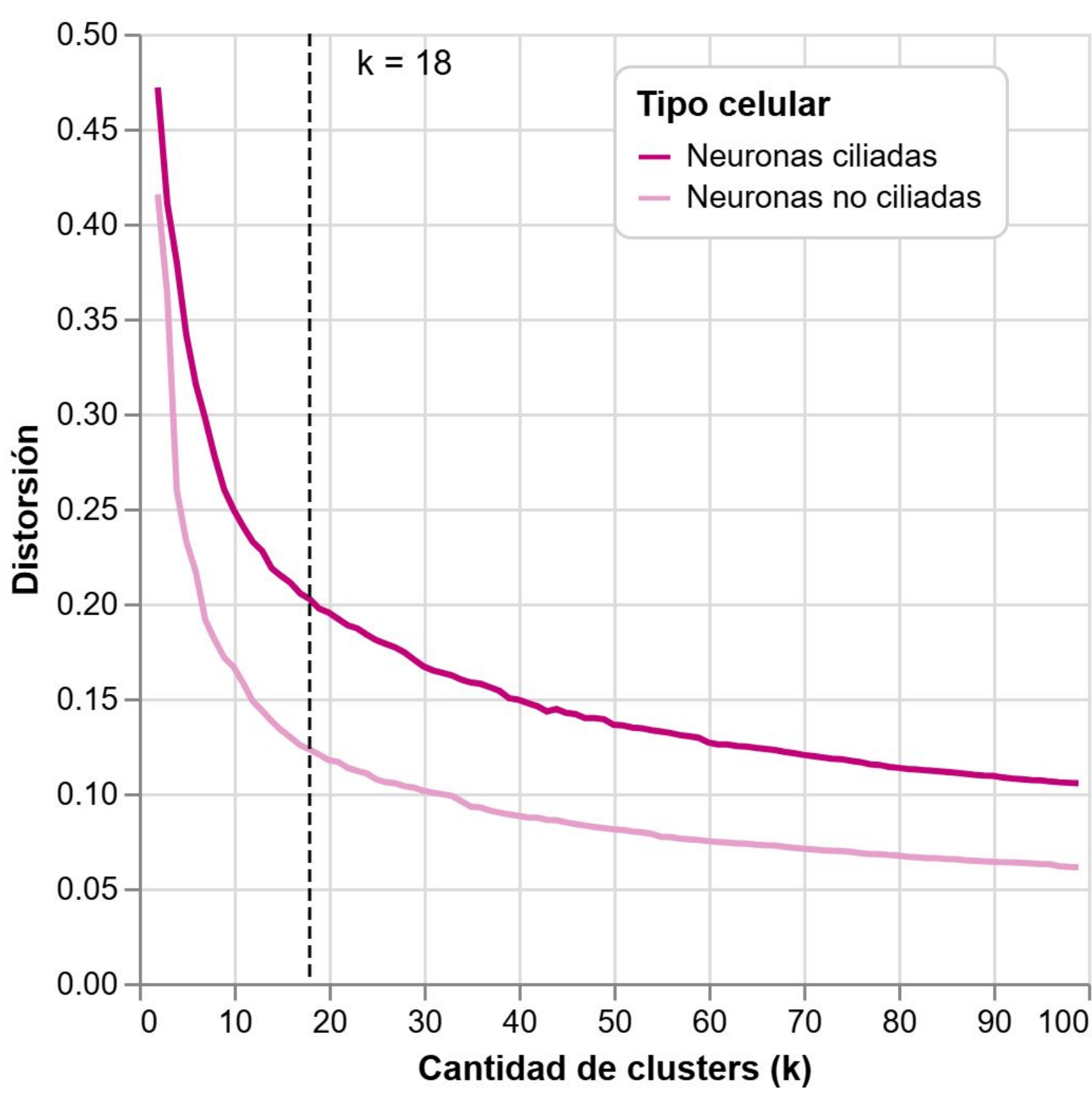
La cilia primaria es un organelo conservado a nivel evolutivo, presente en la mayoría de las células de vertebrados y en algunas células neuronales de invertebrados. Su función principal es la señalización y comunicación de la célula con el entorno, aunque aún no se conoce por completo el conjunto de proteínas que conforman y mantienen el funcionamiento de la cilia. *Caenorhabditis elegans* es un modelo ampliamente utilizado en el estudio de este organelo debido a que su fisiología y anatomía permiten simplificar procedimientos y trasladar conocimientos adquiridos a otras especies, como *Homo sapiens* o *Mus musculus*. Nuestra hipótesis es que los genes ciliares de *C. elegans* comparten características independientes de la homología de secuencia que permiten su identificación.

Metodología

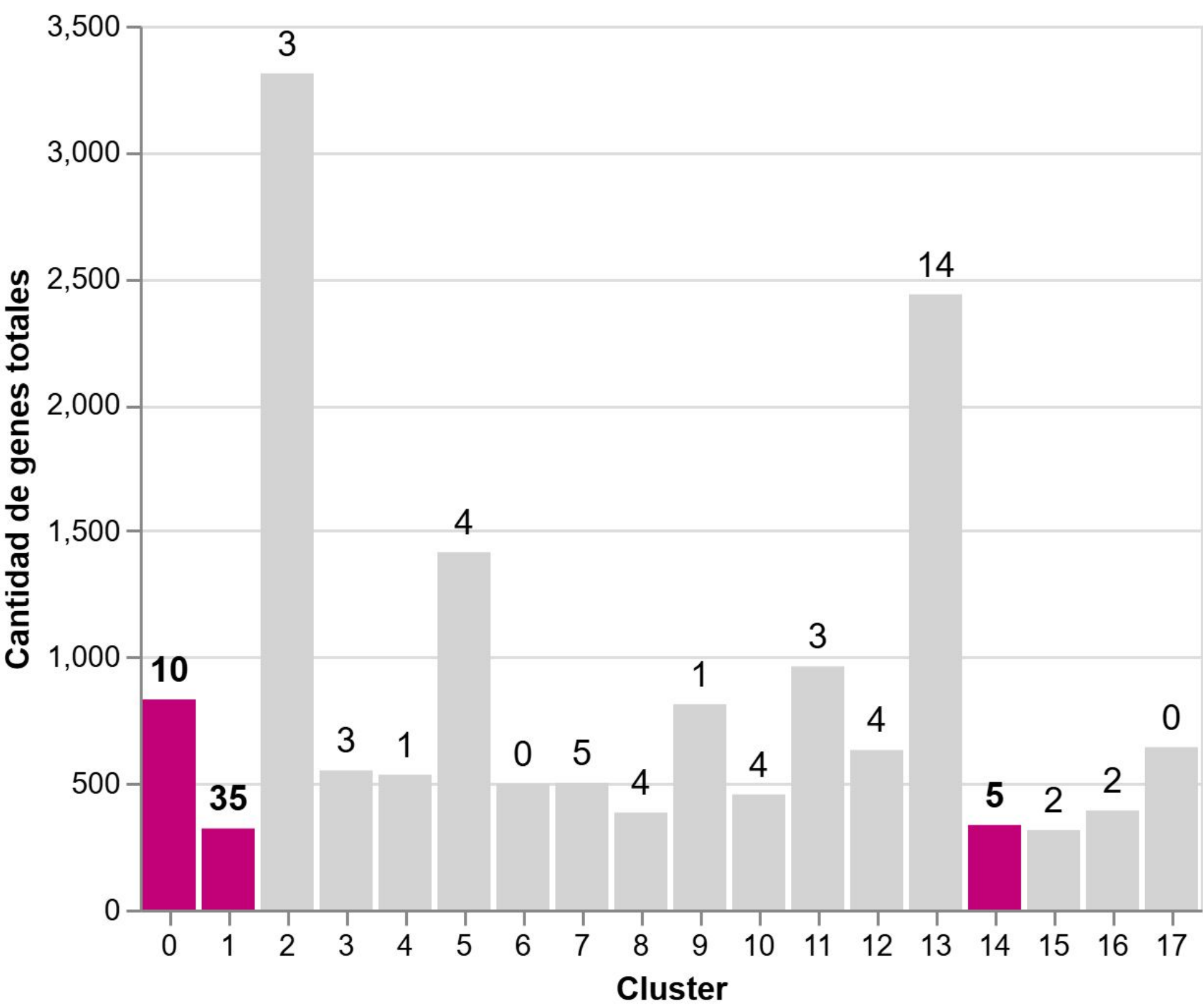
Se recopiló una lista de 141 genes de *C. elegans* con anotaciones vinculadas a cilia primaria (genes de cilia primaria, PCG) en Gene Ontology. Se aplicó k-Means a un set de datos de dominio público de transcriptómica *single cell* de neuronas ciliadas y no ciliadas de embrión de *C. elegans* (Boeck et al., 2016). Estos datos fueron sometidos a un filtrado previo en el cual se descartaron genes con muy bajos niveles de expresión o varianza, y se aplicó análisis de componentes principales a los datos. En cada caso (neuronas ciliadas y no ciliadas), el k óptimo a utilizar se determinó aplicando el método del codo y se calculó el enriquecimiento en PCG de los *clusters* obtenidos. Los genes pertenecientes a los *clusters* con enriquecimiento en PCGs estadísticamente significativo ($p < 0,05$, aplicando test de Fisher) tanto en neuronas ciliadas como en no ciliadas y que no estuviesen incluidos en la lista inicial de PCGs son considerados genes PCG candidatos (cPCG). Utilizando la herramienta PANTHER se determinó el enriquecimiento en términos GO de estas listas de genes.



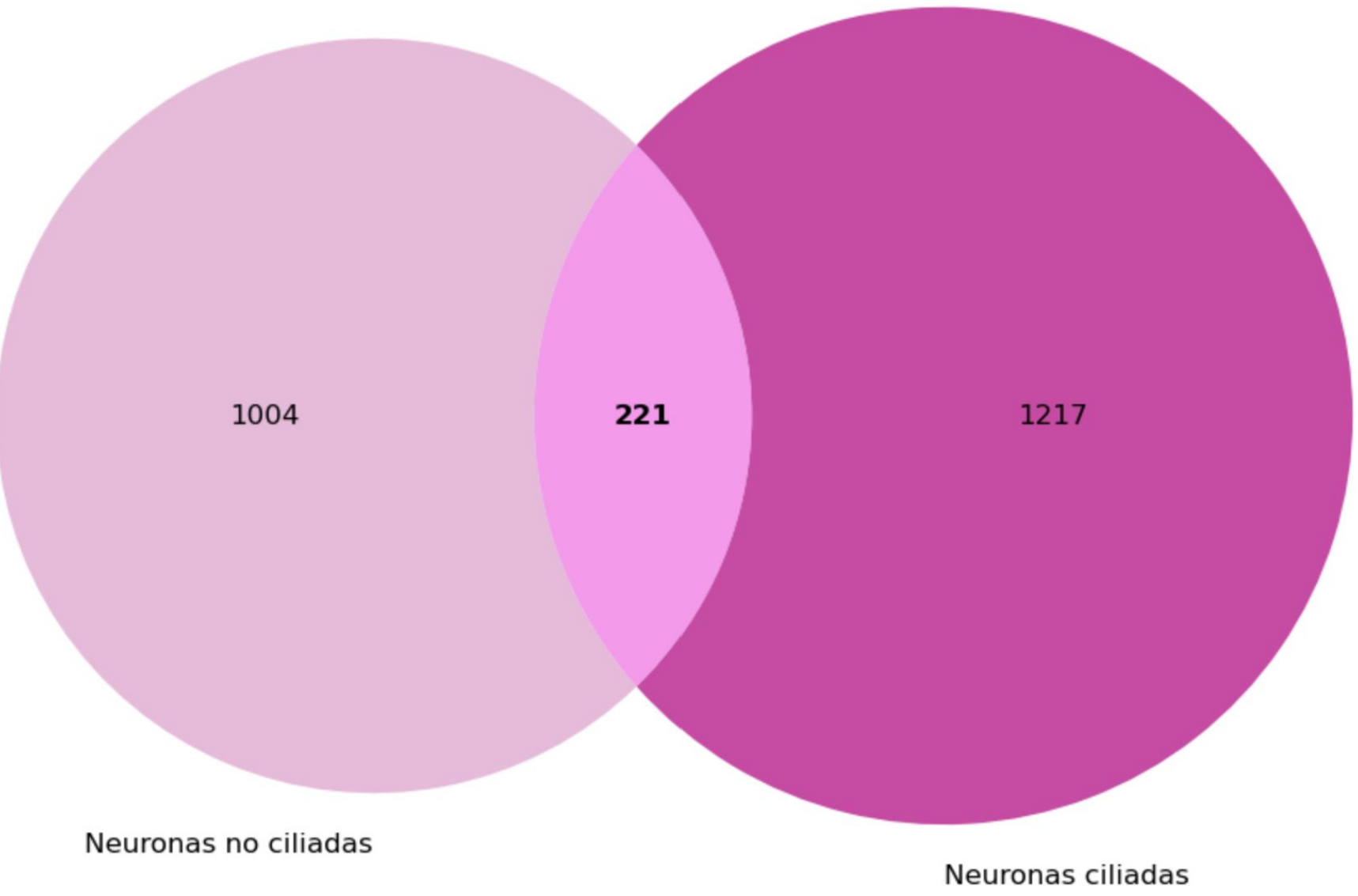
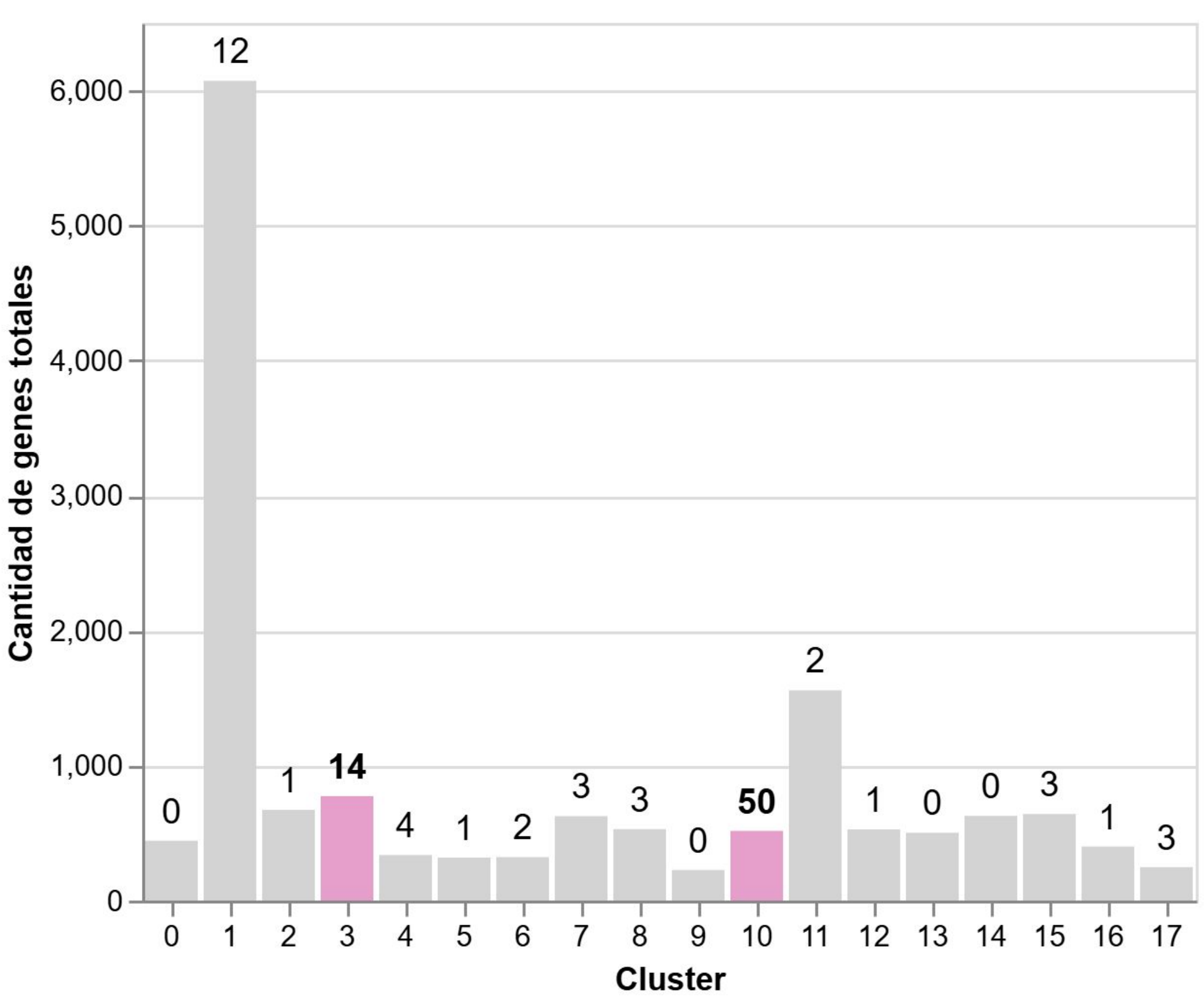
Resultados



Método del codo. Gráfico de método del codo para neuronas ciliadas y no ciliadas.



Cantidad de genes por cluster. Los clusters con enriquecimiento significativo ($p < 0,05$) en PCGs se muestran en color magenta (neuronas ciliadas) o rosa (neuronas no ciliadas). La cantidad de PCGs en cada cluster se muestra en la parte superior de cada barra.



Complemento de PCG en clusters enriquecidos. Se listaron los genes no pertenecientes a la lista original de PCG presentes en los clusters enriquecidos (neuronas ciliadas y no ciliadas). Luego se determinó la intersección de ambas listas y se calculó el enriquecimiento funcional de esta intersección (denominada lista de genes en clusters enriquecidos, LGCE).

Término GO	Fold Enrichment
Axonemal microtubule (GO:0005879)	38,62
Intraciliary transport particle B (GO:0120170)	24,03
Ciliary plasm (GO:0097014)	19,12
Intraciliary transport (GO:0042073)	14,08
Cilium assembly (GO:0060271)	11,86

Enriquecimiento funcional. Términos GO con mayor enriquecimiento en la LGCE ($p < 0,05$ en todos los casos).

Ontología	% de genes sin anotar
Función molecular	62
Proceso biológico	45
Componente celular	59

Genes sin anotaciones funcionales. Porcentaje de genes en la LGCE sin anotaciones funcionales en la base de datos PANTHER.

Discusión

Los resultados obtenidos indican que al aplicar clustering sobre datos de transcriptómica los genes ciliares tienden a co-clusterizar. Estos resultados nos animan a continuar explorando la posibilidad de identificar nuevos genes ciliares utilizando aprendizaje automático. Nuestros próximos incluyen la aplicación de otros algoritmos de clustering en combinación con otras variables predictivas, como la ubicación relativa de los genes en el genoma, así como la implementación de modelos de aprendizaje supervisado.