A Machine Learning Pipeline for Gene Regulatory Network Inference: A Case Study on Protamines

Puhaleva, Bozhidara (School: Sofia High School of Mathematics)

Protamine 1 (PRM1) and protamine 2 (PRM2) are essential for packaging sperm DNA and maintaining genome stability during spermatogenesis. Changes in their expression levels are associated with male infertility. This study employs a combination of computational approaches to investigate protamine regulation in male sperm. An end-to-end bioinformatic pipeline integrating transcriptomic preprocessing and gene regulatory network inference via an optimized scGeneRAI identified candidate transcription factors regulating PRM1 and PRM2. Among the predicted regulators, HMGB4 and TBPL1 have previously been associated with chromatin regulation and germ cell development, supporting the biological relevance of the predictions. In addition, CREM, a well established regulator of spermatogenesis, was also identified, indicating prioritization of high confidence candidates. For further validation, results were compared with those from other existing methods such as GENIE3, GRNBoost2, and CLR, identifying common transcription factors across approaches that may serve as robust molecular targets for future fertility assessment. These findings may be useful for assessing gamete quality. The pipeline's targeted design enables efficient inference and allows application to diverse scRNA-seq datasets across tissues and cell types.

Awards Won: