Tanaya Datar

Email | Portfolio | LinkedIn | Research Articles | Github

Education

University of Toronto, St. George — Toronto, ON
BSc in Bioinformatics & Computational Biology, Minor in CS & Immunology — *Sept 2022 – Apr 2026*

Machine Learning Specialization — DeepLearning AI, Stanford
Built & deployed machine learning models in Python (NumPy, scikit-learn, TensorFlow) — *Jun 2025 – Oct 2025*

Experience

Computational Biology Research Associate – Pai Lab — Sept 2025 – Present
Ontario Institute for Cancer Research (OICR) — *Toronto, ON*

- Developed and supported end-to-end algorithms and data processing pipelines for single-cell tumor analysis across 7+ cerebellar datasets (750,000+ cells) in R, including quality control, batch integration, dimensionality reduction, and unsupervised clustering to characterize genomic alterations and clinical behavior in a cancer context
- Conducted statistical analysis and differential expression analysis to identify transcriptomic signatures linked to tumor biology and clinical outcomes; supported innovative analyses of experimental datasets to inform precision cancer medicine approaches
- Maintained and curated sequencing datasets and databases from public repositories and collaborators on shared high-throughput Linux servers, applying rigorous quality control and documentation to support reproducibility and data integrity
- Prepared processed data, publication-quality figures, and manuscript sections for display and distribution to external collaborators; collaborated cross-functionally with wet-lab scientists and research teams, and supported informal mentoring of junior lab members

Data Analysis Research Internship – Goeva & Agrawal Labs — May 2025 – Sept 2025
Dept. of Molecular Genetics | Dept. of Ecology and Evolutionary Biology, University of Toronto — *Toronto, ON*

- Built reproducible data processing pipelines converting raw sequencing data (FASTQ, STAR, SAMtools, DESeq2) into analysis-ready formats compatible with statistical analysis and visualization, on Linux/UNIX systems using Bash
- Analyzed bulk RNA-seq data across 246 samples using R and statistical methods, applying algorithms for quality control, alignment, and normalization to support high-throughput analysis pipelines
- Evaluated the performance of bioinformatics algorithms through statistical validation; assessed robustness and sensitivity of analytical approaches applied to large genomic datasets

Bioinformatics Researcher – Provart Lab — Sept 2024 – Apr 2025
Dept. of Cell & Systems Biology, University of Toronto — *Toronto, ON*

- Developed gene regulatory network algorithms by integrating large-scale biological datasets and databases; implemented statistical and computational methods for data integration and analysis using Python and R — GitHub
- Built and maintained a reproducible, version-controlled web tool using HTML and JavaScript for display and distribution of processed transcription factor network data, integrating with AIV2/ePlant databases to assist researchers without technical expertise in hypothesis generation

Projects

ExprCompareR – Interactive Omics Data Reporting Tool | *R, Shiny* — GitHub

- Sole developer of an R package and data display tool integrating RNA-seq and protein-expression datasets across human tissues; enables statistical comparison and visualization of transcript-protein relationships to support functional genomics and precision oncology research

BMP7 Signaling in MB: Bulk RNA-seq & Pathway Analysis | *R, Bioconductor, Cytoscape*
Data Processing | Pathway Analysis

- Recreated and extended the analysis behind Ohata et al. (2025) on BMP7-driven oncogenic signaling in medulloblastoma: curated, cleaned, and normalized the bulk RNA-seq dataset (GSE229150) and performed differential gene expression analysis
- Conducted thresholded and non-thresholded gene set enrichment analysis (GSEA) on DE results and built an enrichment map in Cytoscape to visualize and interpret dysregulated pathways

Pathogenic SNV Prediction Tool | *Python, Bash, Machine Learning – Cancer Genomics / Rare Disease* — Devpost

- Built a machine learning classification algorithm predicting pathogenic vs. benign SNVs using ClinVar labels; processed and annotated 13,000+ SNVs with regulatory features from genomic databases using Bash

Research Output

Co-Authored Manuscript – Dr. Vijay Ramaswamy Lab – Developmental Cell *(In Press)* August 2026
- Analyzed single-cell gene expression across tumor datasets to characterize the developmental origins of ependymomas & medulloblastomas

Co-Authored Manuscript – Dr. Nicholas Provart – Genetics *(In Press)* January 2027
- Methods-focused manuscript documenting shared bioinformatics tools and pipelines for integrating and visualizing plant regulatory interaction networks

Poster Presentation – Medical Biophysics Department, University of Toronto August 2026
- "Developing a single-cell atlas of the mouse cerebellum to uncover developmental origins of Group 3 and Group 4 medulloblastoma"; integrated single-cell sequencing datasets across developmental stages to study tumor lineage relationships and clinical behavior

Technical Skills

Programming & Tools: Python (proficient), R (proficient), Bash/Shell scripting (proficient), Git, Docker, VS Code
Bioinformatics & Genomics: Seurat, Scanpy, FASTQ, STAR, SAMtools, DESeq2, MultiQC, NCBI databases and genomic databases, high-throughput data pipelines
Data Analysis: Bulk and single-cell tumor analysis, differential expression analysis, dimensionality reduction, unsupervised clustering, curating and maintaining large datasets
Statistical & Computational Methods: Supervised/unsupervised machine learning, feature engineering, anomaly detection, statistical hypothesis testing, algorithm design for genomic and clinical outcomes data