Noor Sohail

nsohail19.github.io | github.com/nsohail19 | Boston, MA

Summary & Skills

Computational biologist with 5+ years of experience with end-to-end analyses of large, high-dimensional datasets (single-cell, spatial, bulk sequencing). Proven track record of partnering with domain experts to answer biological questions with robust statistical workflows. Passionate about communication, data visualization, and teaching.

Bioinformatics: Single-cell (RNA, ATAC, CITE-seq), spatial (Xenium, Visium, STEREO-seq), bulk RNA-seq.
Programming & Analytics: Python (pandas, NumPy, scanpy), R (tidyverse, Bioconductor, Seurat), C++, bash, SQL.
Infrastructure & Tools: Linux/UNIX, HPC (slurm, LSF), AWS S3, Docker, git, report generation (Jupyter, Quarto).

Education

University of Michigan — May 2020
B.Sc. in Life Science Informatics — Ann Arbor, MI

Professional Experience

Harvard T.H. Chan School of Public Health — Apr. 2023 – Present
Bioinformatician II, Harvard Chan Bioinformatics Core — Boston, MA

- Oversee end-to-end analysis of large-scale transcriptomic datasets across immunology, skin, gastrointestinal, and neuroscience studies.
- Apply clustering, dimensionality reduction (e.g., PCA, UMAP), gene set enrichment, differential expression, neighborhood analysis, and related methods to uncover statistically significant disease mechanisms.
- Designed a workflow that improves gene co-expression estimates in sparse single-cell count matrices, enabling more reliable identification of potential drug targets for industry partners.
- Collaborate with clients to plan projects, set goals, and interpret findings by delivering publication-ready visualizations and written reports within established timelines.
- Create and teach workshops on programming basics and advanced data analysis, equipping hundreds of Harvard researchers with practical computational skills.

Memorial Sloan Kettering Cancer Center — Apr. 2021 – Apr. 2023
Computational Biologist, Single-cell Analysis Innovation Lab — New York, NY

- Served as the primary data analyst for translational projects, including studies of how CRISPR deletions in chromosome 9p alter immune responses.
- Assembled an 8M-cell COVID-19 dataset from published data by aggregating cells/data with similar phenotypes to reduce sparsity and scale of the dataset.
- Created and parallelized custom pipelines starting from raw sequencing files, one to recover weak interferon signals and another to track lineages using mitochondrial mutations.
- Implemented scalable WDL/AWS/Docker workflows and QC frameworks for processing datasets, supporting hundreds of samples across dozens of projects.

Research Experience

Weill Cornell Medicine — Jan. 2020 – Mar. 2020
Summer Research Intern — Ann Arbor, MI

- Merged protein phosphorylation scores with independently generated public datasets (GTEx, HPA, CCLE) to predict tissue-specific protein interaction patterns.

Van Andel Research Institute May 2019 – Aug. 2019; May 2018 – Aug. 2018
Summer Research Intern Grand Rapids, MI

- Maintained and extended MTseeker, an R/Bioconductor package for mitochondrial DNA variant detection.
- Analyzed large public datasets to define the mitochondrial mutation landscape in pediatric cancers, contrasting tumor and normal samples to identify disease-specific variants.

University of Michigan Oct. 2017 – Dec. 2019
Undergraduate Researcher Ann Arbor, MI

- Evaluated drug efficacy by simulating protein–ligand interactions between β-Secretase 1 and drug candidates with free energies of binding, where were computed with molecular dynamics (CHARMM) and statistical mechanics.
- Compared accuracy and computational cost of Multisite λ-Dynamics vs. TI/MBAR alchemical free energy methods, providing guidance on method selection for drug discovery projects.

Publications

[5] Kaplan HS, BL Logeman, Zhang K, Yawitz TA, Santiago C, **Sohail N**, . . . , Dulac C. "Sensory input, sex and function shape hypothalamic cell type development." *Nature*, 2025. (link)

[4] Persad S, Choo ZN, Dien C, **Sohail N**, . . . , Pe'er D. "SEACells infer transcriptional and epigenomic cellular states from single-cell genomics data." *Nature Biotechnology*, 2023. (link)

[3] Barriga FM, Tsanov KM, Yu-Jui H, **Sohail N**, . . . , Lowe SW. "MACHETE identifies interferon-encompassing chromosome 9p21.3 deletions as mediators of immune evasion and metastasis." *Nature Cancer*, 2022. (link)

[2] Vilseck JZ, **Sohail N**, . . . , Brooks CL III. "Overcoming Challenging Substituent Perturbations with Multisite λ-Dynamics: A Case Study Targeting β-Secretase 1." *Journal of Physical Chemistry Letters*, 2019. (link)

[1] Triska P, Kaneva K, Merkurjev D, **Sohail N**, . . . , Gai X. "Landscape of Germline and Somatic Mitochondrial DNA Mutations in Pediatric Malignancies." *Cancer Research*, 2019. (link)

Selected Presentations

Sohail N, Billingsley JM, Geist F, Hoong H, Ho SH. "Co-Expression Workflow for Single-Cell RNA Sequencing Data." Poster presentation at ISMB, 2024.

Sohail N, Moorman A, Pe'er D. "SEACells: Using Single Cell Aggregation to Facilitate Data Integration." Oral presentation at Chan Zuckerberg Initiative Assembling Tissue References Workshop, 2022.

Sohail N, Triche TJ. "MTseeker: Mitochondrial Variant Analysis Tools for Bioconductor." Poster presentation at Bioconductor, 2018.

Projects & Training Materials

Actively Maintained Workshops

- **Introduction to single-cell RNA Sequencing (scRNA-seq)** (link)

  Hands-on course in end-to-end analysis of high-dimensional single-cell count data, including experimental design, QC, clustering, differential expression, and reproducible workflows in R.
- **Introduction to Spatial Transcriptomics** (link)

  Practical introduction to working with large, structured spatial datasets (Visium HD), pairing imaging coordinates with molecular profiles.
- **Pseudobulk and Related Approaches for scRNA-seq** (link)

  Workshop on aggregating high-noise, sparse observations into sample-level features to account for variability and enable statistically robust analysis of differences between conditions and proportions.
- **Introduction to Python** (link)

  Zero-to-hero Python course centered on data analysis workflows, covering core data structures, data manipulation with pandas, real-world data wrangling, and visualization with Matplotlib and Seaborn.

Projects

- **Co-expression Workflow for scRNA-seq** (GitHub)
  R pipeline for computing co-expression scores in zero-inflated, high-dimensional single-cell datasets; includes preprocessing, imputation, and visualization of correlation structure.
- **Template for scRNA-seq Quality Control** (GitHub)
  Jupyter notebook template to standardize scRNA-seq cleanup with quality control metrics, filtering, and visualizations for reproducible analysis.
- **Tidy Tuesday** (GitHub)
  Ongoing collection of weekly visualization projects in R with ggplot2.

Other Activities

University of Michigan Jan. 2020 – Mar. 2020
Lab Prep Assistant Ann Arbor, MI

- Prepared reagents and materials for an advanced biology lab course.

Office of Academic Multicultural Initiatives Jan. 2018 – Apr. 2018
Mathematics Tutor Ann Arbor, MI

- Mentored students one-on-one in Calculus I and II.

Wolverine CuiZine Sep. 2017 – Mar. 2020
Writer Ann Arbor, MI

- Composed food-related articles aligned with each magazine issue's theme.

Model United Nations at the University of Michigan Sep. 2016 – Mar. 2020
Committee Director Ann Arbor, MI

- Created educational materials on historical and global issues to facilitate debate among student committees.