

# Shizhao Yang

+1 206 688 9210 | [syang71@uw.edu](mailto:syang71@uw.edu) | [GitHub](#) | [LinkedIn](#)

## EDUCATION

**University of Washington, Seattle, WA**

-expected Mar 2025

*M.S. in Biostatistics (Modeling and Methods pathway)*

**New York University, New York, NY**

May 2023

*B.S. in Data Science, Genomics concentration (Minor: Mathematics)*

## RESEARCH EXPERIENCE

### Scientific Research

***Single-cell Lineage Embedding Contrastive Learning***

Seattle, WA, Nov 2023-present

*Researcher*

*Supervisor: Kevin Lin, University of Washington*

- Designed and implemented a contrastive learning algorithm to learn the high-dimensional embeddings of single-cell data, facilitating the identification of lineage-specific gene expression patterns.
- Develop an evaluation metric for comparing the different embeddings of single-cell data generated by the model.

***Python-based RNA-seq Analysis Algorithm using Negative Binomial GLM***

New York, USA, June-Dec 2022

*Research Assistant*

*Supervisor: Manpreet Katari, New York University*

- Build a python-based Stats Model based on GLM with negative binomial distribution for gene's differential expression analysis.
- Use backtracking line search in dispersion estimation and IRLS (iterative reweighted least squares) in coefficient estimation and apply the Wald Test to the estimated log fold changes.
- Exploit Python's built-in package to enable multiprocessing to shrink the overall runtime from 3 hours to about 3 minutes.
- Test the algorithm on several different datasets and compare the results with actual and estimated values generated by Deseq2.

### SRPMS (Student Research Program in Molecular Science)

***Investigation of Horizontal Transfer in Metagenomics***

Shanghai, China, June 2021-Jan 2022

*Research Assistant*

*Supervisor: Gang Fang, NYU Shanghai, NYU*

- Implemented RNA-seq Analysis on Next-Generation Sequencing data of human gut microbiome, including de novo genome assembly, mapping, gene-calling, and annotation using HPC (high-performance computer) and related pipelines.
- Adopted the Louvain Method to generate pseudo ortholog communities by comparing the similarity between each gene.
- Calculated TPM and self-defined index (PI) related to gene persistence and identified the correlation between the TPM and PI using Python and R.
- Extracted differential expressed genes related to HTG by adopting statistical methods, including ANOVA

### Course Project

**Intro to Math Modeling**

New York, USA

***Refined SIR Model with Vaccination and its Application in 2022 NYC Influenza A Activity Prediction***

September 2022-December 2022

- Implemented a series of ODE methods (including fixed point and its stability, phase plane, and Herd Immunity) to analyze the modified SIR model with the consideration of Vaccination analytically.
- Simulated the NYC influenza data of the past six years using the SIRV model and estimated the transmission rate, removal rate, and reproduction number ( $R_0$ ) using the Quasi-Newton method on Python.
- Compared the estimated coefficients of the past seasons and predicted the infection peak in NYC this year under different vaccination rates.

### Transcriptomics

New York, USA

***RNA-seq Analysis of Brain gene expression in Female and Male *D. melanogaster* post-eclosion***

# Shizhao Yang

+1 206 688 9210| [syang71@uw.edu](mailto:syang71@uw.edu) | [GitHub](#) | [LinkedIn](#)

*March 2022-May 2022*

- Implemented RNA-seq alignment against the reference genome using HPC and hisat2 package.
- Analyzed the differential gene expression with different methods, including PCA(Principal Component Analysis), Clustering Analysis, GO-term Analysis, and GSEA(Gene Set Enrichment Analysis) using R with limma and deseq2 packages.
- Compared the result of differentially expressed genes with those from published papers using similar datasets and presented the analysis of differences and similarities in the report.

## **Machine Learning**

Shanghai, China

### ***Drug sensitivity prediction using machine learning models***

*September 2021-December 2021*

- Found appropriate dataset from CCLE (Cancer Cell Line Encyclopedia) database and implemented data munging method.
- Implemented SVM(Support Vector Machine), Linear Regression Random Forest, Stacking, and ANN(Artificial neural networks) to predict the sensitivity of drugs given the drug information and patients' gene expression information.
- Integrated meaningful outcomes from stimulations and demonstrated them with Python.

## **LANGUAGES & PROFESSIONAL SKILLS**

---

- Chinese (native), English (fluent)
- Programming Languages: Python, R, Linux (Shell), SQL, MongoDB, HTML, MATLAB, Javascript
- Analytics Skills: Statistical modeling, Numerical analysis, Machine Learning
- Bioinformatics Applications: SPAdes, Bowtie2, Hisat2, Samtools, diamond, Blast, Limma, Deseq2