ZUOLIN CHENG

• zuolincheng166@gmail.com • (+1)571-385-8606 • Cupertino, CA 95014
• https://zuolincheng.github.io/

SUMMARY

Highly skilled Ph.D. in Bioinformatics with 8+ years of experience in designing/implementing multi-omics data analysis pipelines & developing novel machine-learning/statistical models. Proficient in multiple programming languages, machine-learning/data-analysis libraries, and bioinformatics tools. Proven track record of successful model development & omics data analysis that has enabled significant biological discoveries in cross-functional, collaborative research.

CORE COMPETENCIES

- **AI / Machine Learning:**
  - Model Development: Large Language Models (LLMs**)**, Neural Networks, Supervised Learning (Random Forests, SVM), Unsupervised Learning, Statistical Modeling
  - AI Applications: AI for Biological Insight (Driving Transcriptional Factor Identification, etc.)
- **Bioinformatics & Omics Analysis:**
  - Data Types: Omics Data Analysis (scRNA-seq, Spatial Transcriptomics, scATAC-seq, ChIP-seq, CUT&RUN, proteomics, metabolomics)
  - Techniques: Proficient in Omics Technologies, End-to-End Omics Data Analysis Pipeline Development, Gene Set Enrichment Analysis, Regulatory Network Analysis, Cell Type Annotation
- **Programming & Tools:**
  - Languages: Python, R, C++, MATLAB, Bash Scripting.
  - Tools & Environment: Git / GitHub, Linux Environment

WORK EXPERIENCE

Virginia Tech Cupertino, CA, USA
Postdoctoral Research Associate Feb. 2023 – Present
Lead Next Generation Sequencing (NGS) data analysis in collaborative biological research studies, effectively facilitating scientific discoveries. Develop state-of-the-art NGS data analysis models and methods, providing user friendly R packages to benefit a larger audience. Mentor & train junior graduate students in research practices.

- ***Project 1: GO-BERT, Cell Type Annotation for scRNA-seq Data Using Large Language Model (LLM)***
  - *Data*: scRNA-seq data, Gene Ontology (GO); *Language*: Python, PyTorch
  - Innovatively **integrated domain-specific priors** (gene ontology) into neural network architecture during supervised fine tuning, serving as regularizations to guide model toward more plausible predictions
  - Significantly enhanced cell type annotation performance, increasing F1 score to 0.76 (previously 0.69)
  - **Reduced 31% parameters** in fine tuning without loss of performance compared to the state-of-the-art
  - Propose a **domain-insight-based pretraining strategy** for large-scale single-cell data, overcoming limitations of existing models (e.g. scBERT, scGPT, Geneformer) in capturing awareness of global context

- ***Project 2: A Machine Learning Model Unifying Various NGS Analysis Tasks***
  - *Data*: various types of NGS data; *Language:* C++, R, MATLAB
  - Unified various NGS analysis tasks (e.g. single cell RNA-seq data normalization, cancer gene detection, GO term activity inference, etc.) by pinpointing the shared key mathematical problem behind them
  - Developed **a novel model of joint inference of multiple hidden factors** by integrating sophisticated statistical/machine-learning techniques, enhancing confounding factor control in downstream analyses
  - Proposed a highly efficient customized optimizer leveraging model structure & code optimization (C++), achieving **scalability on large-scale real-world data** (reduced running time: days to minutes)
  - Developed user-friendly R packages on different OS (Windows, Linux, iOS), and distributed via **GitHub**

Virginia Tech, Computational Bioinformatics and Imaging Lab Arlington, VA, USA
Graduate Research Assistant Aug. 2015 – Dec. 2022

Developed innovative AI/ML and statistical tools for NGS data analysis. Designed and implemented NGS analysis pipelines in cross-functional projects, collaborating with biologists and clinicians from leading institutions such as Stanford University, Cincinnati Children's Hospital Medical Center, and the Salk Institute.

- ***Project 1: RNA-seq Analysis Pipeline to Uncover Microglial Heterogeneity During Development***
  - *Data*: bulk and single cell RNA-seq data; *Language*: R
  - Designed and implemented **end-to-end analysis pipeline**, including sequencing alignment, quality control (QC), RNA-seq normalization, clustering, data visualization, DEG analysis, GSEA, etc.
  - **Customized** some modules in the pipeline, achieving best solutions to **project-specific requests**
  - Enabled biological discoveries, collaborating with Stanford University: **revealed a novel microglia subset**

- ***Project 2: Machine Learning Model for Driving Transcription Factor (TF) Prediction***
  - *Data*: scRNA-seq, ChIP-seq Database, DNase-seq, Motif Database (JASPAR); *Language*: R
  - **Integrated** ChIP-seq, motif databases, and epigenetic data to predict driving TF behind biological process
  - **Discovered a long-ignored key factor** and integrated it into ML model, boosting the best record of disease driving TF detection based on large TF-gene binding database (**38.8% to 49.0%** on benchmark)
  - Validated predictions (Zfp36l1 TF) via **wet-lab collaboration**, published in *Cell Stem Cell*

- ***Project 3: Full-length RNA-seq Data Normalization***
  - *Data*: scRNA-seq; *Language*: R, C, C++
  - Identified a key confounding factor (effective cDNA-length during PRC) ignored by previous analyses in full-length RNA-seq raw counts, by **combining empirical data mining** (high-throughput intermedia data of sequencing alignment) and **theoretical reasoning** (**deep understanding of sequencing technologies**)
  - Proposed a ML algorithm to learn and address the factor during RNA-seq data normalization, enhancing the accuracy of downstream differential expressed gene (DEG) detection (AUC: 0.886 to 0.938)

EDUCATION

Virginia Tech Arlington, VA, USA
PhD, Electrical Engineering Aug. 2015 – Dec. 2022
- *Research areas:* Bioinformatics; Machine Learning

Peking University Beijing, China
M.S., Electronic Science and Technology
- GPA: 3.87/4.0; **Top Ranked**: 1/30

PUBLICATIONS

TOTAL PUBLICATIONS: 13; TOTAL CITATIONS: 1600 +

https://scholar.google.com/citations?user=6pHncW4AAAAJ&hl=en&oi=ao

- Li Q*, **Cheng Z* (co-first author)**, Zhou L, Darmanis S, … Tony Wyss-Coray, Ben A Barres. Developmental heterogeneity of microglia and brain myeloid cells revealed by deep single-cell RNA sequencing. ***Neuron***. 2019. (**Best of Neuron 2018-2019**).
- Barclay K, Nora A, **Cheng Z**, Kim M, Zhou L, Yang J, Rustenhoven J, et al. An inducible genetic tool to track and manipulate specific microglial states reveals their plasticity and roles in remyelination. ***Immunity***. 2024.
- **Cheng Z**, Wei S, Wang Y, Wang Y, Lu R, Wang Y, Yu G. An Efficient and Principled Model to Jointly Learn the Agnostic and Multifactorial Effect in Large-Scale Biological Data. *bioRxiv* (2024): 2024-04.
- **Cheng Z,** Wei S, Yu G. A Single-Cell-Resolution Quantitative Metric of Similarity to a Target Cell Type for scRNA-seq Data. *In 2022 IEEE International Conference on Bioinformatics and Biomedicine (**BIBM**)* 2022.