Linhua **Wang**

Computational Biologist | Multi-omics data scientist

+1 412 956 7722
linhuaw@bcm.edu
linkedin.com/in/linhuaw
Personal website
github.com/linhuawang
8181 Fannin St, Houston, TX 77054 - USA

I am passionate about using computational methods to advance biomedical discoveries from genomics and transcriptomics data. My Ph.D. research focused on developing techniques for analyzing spatial transcriptomics and multi-omics data from single cells, which has taught me the importance of collaborating with biologists and clinicians to create practical, usable tools. With a background in both biology and computer science, I have the skills and expertise to apply computational approaches such as Machine Learning, Deep Learning, and Convex Optimization to address biological questions. I have developed several tools and packages that have proven beneficial in a range of areas, including cancer research, genetic disease diagnosis, neurodegenerative diseases, and protein function predictions.

Education

Current — **Baylor College of Medicine**
PhD. Quantitative & Computational Biosciences
Research : Spatial Transcriptomics, single-cell Multi-omics, AI-aided rare-disease diagnosis

2017 — **Carnegie Mellon University**
MSc. Computational Biology
Concentration : Machine Learning, Programming Language, Algorithm

2013 — **Sun Yat-sen University**
BSc. Biotechnology
Concentration : Molecular Biology, Cell Biology, Cancer Biology

Research Experience

Current – December 2019 — **Texas Children's Hospital - Zhandong Liu's Lab | Data Science - Graduate Student**
Department of Pediatrics – Mentor : Dr. Zhandong Liu

- Developed a computational tool, MIST, to detect molecular regions and impute gene expression values for Spatial Transcriptomics (ST) data (Published at **Nature Communications**)
- Designed a strategy, ReSort, to generate an internal reference for ST's cell type deconvolution.
- Simulated ST datasets and demonstrated that ReSort increased the accuracies of state-of-the-art reference-based ST deconvolution methods.
- Identified macrophage polarization in Breast Cancer with Epithelial-Mesenchymal Transitions using ReSort.
- Validated ReSort's discoveries using external The Cancer Genome Atlas (TCGA) datasets and immunochemical staining of cell type markers (Under review at **Genome Biology**)
- Constructed a machine learning pipeline, MARRVEL AI, to aid rare genetic-disease diagnosis by collaborating with domain experts. Resulted in >10% top-5 accuracies than other academic and commercial tools. (Oral presentation by coauthor at **ASHG, 2022**)
- Published a bioinformatics Python package, SEAGAL, for Spatial Enrichment Analysis of Gene Associations using L-index, after understanding the needs by interviewing biologists.
- SEAGAL allows identifying and visualizing spatial co-localization or exclusion of immune cell types at specific spatial niches (Submitted to **Bioinformatics**).

Computational Biology | Spatial Transcriptomics | Single-cell Multi-omics | Genetic Diagnosis | Data Visualization

Sep 2022 – May 2022 — **Ancestry - DNA Sciences - Genomics Data Scientist Intern**
Applied Machine Learning – Mentor : Dr. Milos Pavlovic

- Processed feature vectors with 30 million rows and 3000 features using chunked data processing.
- Reduced the number of features from >2000 to <100 using feature selection techniques.
- Constructed scalable XGBoost-based machine learning pipelines that reduced the number of classifiers (>2000) to a log-scale (<20).
- Delivered accurate geographical community assignments to 30 million customers based on genetic features using XGBoost with improved precision and recall scores.

DNA Science | Pandas | AWS | Scikit-learn | Sparse Matrix Processing

April 2019 – July 2017

Icahn School of Medicine at Mount Sinai - The Pandey Lab | Machine Learning - Full-Time Bioinformatician

Department of Genetics and Genomic Sciences – Supervisor : Dr. Gaurav Pandey

- Built and maintained ensemble models combining 11 base WEKA classifiers, including Naive Bayes, SVM, Logistic Regression, and so on.
- Scaled up the tool leveraging parallel and distributed computing on high-performance computing (HPC) systems.
- Improved predictions' performance (F-score) on 277 gene ontology terms for 63,449 amino acid sequences from 19 clinically relevant bacterial pathogens.
- Packaged and maintained the tool on GitHub and wrote Docker files for version control.

Machine Learning | Version Control | Protein Function Prediction | High-Performance Computing (HPC)

Dec 2016 – May 2016

University of Pittsburgh - Lu Lab | Cancer Biology - MS research

Department of Biomedical Informatics – Mentor : Dr. Xinghua Lu

- Performed cancer marker discovery analysis for Brain Tumors, including Lower Grade Glioma and Glioblastoma samples from The Cancer Genome Atlas (TCGA) using R programming.

R programming | Cancer Research | Statistical Analysis

Publications

Unraveling Spatial Gene Associations with SEAGAL : a Python Package for Spatial Transcriptomics Data Analysis and Visualization — Under Review

Linhua Wang, Chaozhong Liu, Zhandong Liu

Bioinformatics — Spatial Transcriptomics | Immune Co-localization | Bi-variate Spatial Correlation

scGREAT : Graph-based regulatory element analysis tool for single-cell multi-omics data — Under Review

Chaozhong Liu, Linhua Wang, Zhandong Liu

Bioinformatics — Single-cell | Multi-omics Data Analysis | Cis-regulatory Elements

Accurate cell type deconvolution in spatial transcriptomics using a batch effect-free strategy — Unver Review

Linhua Wang, Ling Wu, Chaozhong Liu, Wanli Wang, Xiang H.-F. Zhang, Zhandong Liu

Genome Biology — Spatial Transcriptomics | Cell Type Deconvolution | Tumor Microenvironment

Region-specific denoising identifies spatial co-expression patterns and intra-tissue heterogeneity in spatially resolved transcriptomics data — 2022

Linhua Wang, Mirjana Maletic-Savatic, Zhandong Liu

Nature Communications — Spatial Transcriptomics | Clustering | Modularity Detection | Imputation | Low-rank approximation

Single-cell multi-omics integration for unpaired data by a siamese network with graph-based contrastive loss — 2023

Chaozhong Liu, Linhua Wang, Zhandong Liu

BMC Bioinformatics — Deep Learning | Single-cell Multi-omics Integration | kNN-graph Imputation

Integrating multimodal data through interpretable heterogeneous ensembles — 2022

Yan Chak Li, Linhua Wang, Jeffrey N Law, T M Murali, Gaurav Pandey

Bioinformatics Advances — Ensemble Learning | Data Integration | COVID-19 Mortality Prediction

Predicting youth diabetes risk using NHANES data and machine learning — 2022

Nita Vangeepuram, Bian Liu, Po-hsiang Chiu, Linhua Wang, Gaurav Pandey

Scientific Reports — Diabetes | Machine Learning

Large-scale protein function prediction using heterogeneous ensembles — 2018

Linhua Wang, Jeffrey Law, Shiv D Kale, TM Murali, Gaurav Pandey

F1000Research — Ensemble Learning | Machine Learning | Protein Function Predictions

Skills

Programming	Python, OOP, R, MATLAB, Bash, GitHub, Docker, LaTex
Data Science	Pandas, ScanPy, AnnData, Scikit-Learn, AWS, TensorFlow, PyTorch
Data Visualization	Seaborn, Matplotlib, ggplot2
Research	Spatial Transcriptomics Analysis, Machine Learning, Computational Biology, Single-cell Multi-omics Analysis, Genetics
High Performance Computing	LSF, Distributed Computing, Parallel Computing
Soft Skills	Communication, Creativity, Critical Thinking

Invited Talks

Large-scale assessment of protein function prediction using heterogeneous ensembles 2018

The 26th Intelligent Systems for Molecular Biology (ISMB)

Linhua Wang, Jeffrey Law, Shiv D Kale, TM Murali, Gaurav Pandey

Position of Responsibility

2022-present **MDPI - Multidisciplinary Digital Publishing Institute**

Invited Reviewer

- Review at least two-paper per month for various journals at MDPI, including Cells, Cancers, Diagnostics, and so on.

2020 **MLCB - Machine Learning in Computational Biology**

Scientific Program Committee

- Invited for review of conference papers.

Achievements & Recognitions

2015-2017 **Academic Achievement Fellowship** by Department of Computational Biology, CMU

2022 **Honorable Mentions** in 2022 **Empower22 HACKATHON**, Ancestry, Lehi, UT.

2021 **Second place** (over 250 participants, subtask) in 2021 **Multimodal Single-Cell Data Integration (NeurIPS Competition 2021)**

2018 **F1000 Outstanding Presentation Prize** (one in six) at **Intelligent Systems for Molecular Biology (ISMB)**, Chicago.

2018 **Travel Fellowship** for **Intelligent Systems for Molecular Biology (ISMB)**, Chicago

Languages

English ●●●●●

Chinese ●●●●●