Xinyi LIN

linxy29@connect.hku.hk

Education

The University of Hong Kong | Hong Kong SAR, China
Doctor of Philosophy in Bioinformatics **09.2020-09.2024**
Supervisor: Dr. Joshua Wing Kai, Ho; **Co-supervisor:** Dr. Yuanhua, Huang

- PhD thesis was rated as "excellent (top 10%)"

Tsinghua University | Beijing, China
Visiting student **07.2022-06.2023**
Supervisor: Professor Xuegong, Zhang

Columbia University | NY, US
Master of Science in Biostatistics **09.2018-06.2020**

- GPA: 3.95
- Main Courses: Probability, Data Science, Biostatistical Methods, Epidemiology, Statistical Inference

Sun Yat-sen University | Guangzhou, China
Bachelor of Science in Biological Science **09.2014-06.2018**

- GPA: 3.9/4.0
- Main Courses: Cell Biology, Biochemistry, Genetics, Ecology, Microbiology, Biostatistics, Advanced Mathematics

Statistics (Double Major)

- Main Course: Mathematical analysis, Geometry and algebra, Advanced programming language, Data structure, Database principles and applications

Publication

1. **Lin, X.**, Chau, C., Ma, K. *et al.* DCATS: differential composition analysis for flexible single-cell experimental designs. ***Genome Biol*** 24, 151 (2023). https://doi.org/10.1186/s13059-023-02980-3
2. Meng, Q., Wei, L., Ma, K., Shi, M., **Lin, X.**, Ho, J. W., ... & Zhang, X. (2024). scDecouple: decoupling cellular response from infected proportion bias in scCRISPR-seq. Briefings in Bioinformatics, 25(2), bbae011.
3. Warin, J., Vedrenne, N., Tam, V., Zhu, M., Yin, D., **Lin, X.**, ... & Camus, A. (2024). In vitro and in vivo models define a molecular signature reference for human embryonic notochordal cells. iScience, 27(2).
4. Xue, Y., Su, Z., **Lin, X.**, Ho, M. K., & Yu, K. H. (2024). Single-cell lineage tracing with endogenous markers. Biophysical Reviews, 16(1):125-139.
5. Xue, Y., Chao, Y., **Lin, X**., Huang, Y., Ho, J. W., & Sugimura, R. (2023). Single-cell mitochondrial variant enrichment resolved clonal tracking and spatial architecture in human embryonic hematopoiesis. *bioRxiv*, 2023-09. (Under review in ***Nature Communication***.)
6. Miso, Z., **Lin, X**., Li, J., Ho, J. W., Meng, Q., Zhang, X. (2023). scRecover: Discriminating true and false zeros in single-cell RNA-seq data for imputation. (Accepted in *Statistics in Medicine*.)
7. **X. Lin**, G. Deng, Y. Li, J. Ge, J.W.K. Ho, Y. Liu, GeneRAG: Enhancing Large Language Models with Gene-Related Task by Retrieval-Augmented Generation, bioRxiv (2024). https://www.biorxiv.org/content/10.1101/2024.06.24.600176v1

Conference Presentation

- A long talk named ‘Differential composition analysis in single-cell RNA-seq data analysis’ in **2021 Bioconductor Asia**
- A flash talk named ‘Differential composition analysis in complex single-cell RNA-seq designs’ in NSFC/RGC Workshop on Single-Cell Data Science
- Poster presentation about ‘DCATS: Differential composition analysis in complex single-cell RNA-seq designs’ in **ISMB/ECCB 2023**
- Poster presentation about ‘scRecover: Discriminating true and false zeros in single-cell RNA-seq data for imputation’ in BIIP 2023

Selected Honors

- Bau Tsu Zung Bau Kwan Yeu Hing Research and Clinical Fellowship
- The **Gold Prize** in the national university student innovation competition
- The Third Prize for Entrepreneurship Proposal in the 9th **Hong Kong University Student Innovation and Entrepreneurship Competition**
- Silver presentation award in NSFC/RGC Workshop on Single-Cell Data Science
- Third Class Scholarship in 2014-2015 and 2016-2017
- Second Class Scholarship and Discipline Competition Award in 2015-2016
- First Prize in Freshman Debate Competition (University Level)
- “Best Planner” award and First Prize in School Planning Contest (College Level)

Research Experience

GeneRAG: Enhancing Large Language Models with Gene-Related Task by Retrieval-Augmented Generation

Project leader 02.2024-Now

- Large Language Models (LLMs) like GPT-4 have revolutionized natural language processing and are used in gene analysis, but their gene knowledge is incomplete. Fine-tuning LLMs with external data is costly and resource-intensive. Retrieval-Augmented Generation (RAG) integrates relevant external information dynamically.
- We develped GeneRAG, a framework that enhances LLMs' gene-related capabilities using RAG and the Maximal Marginal Relevance (MMR) algorithm.
- Evaluations with datasets from the National Center for Biotechnology Information (NCBI) show that GeneRAG outperforms GPT-4o with a 39% improvement in answering gene questions, and more than 40% performance increase in downstream tasks.

Applying mitochondrial mutation in deciphering cell differentiation and development

Project leader 02.2023-Now

- The mutation rate of mitochondrial genome is 10 – 100 fold higher than nuclear genome. It can serve as a kind of endogenous signal for cell lineage tracing.
- The technique of mitochondrial enriched sequencing (MAESTER) was adopted in understanding early embryonic hematopoiesis from hPSC.
- The pipeline of analyzing mitochondria mutation from other sequencing data was adopted and optimized to analyze mitochondria mutation. Multiple waves of erythropoiesis in time-series scRNA-seq data were identified.
- This technique and analysis pipeline were also adopted in understanding the development of Gaint Cell

Tumor of Bone (GCTB) and the mechanism of reoccurrence.

Differential composition analysis in complex single-cell RNA-seq designs
Project leader, Published in **Genome Biology** 10.2020-07.2023

- It remains challenging to effectively detect differential compositions of cell types when comparing samples coming from different conditions or along with continuous covariates, partly due to the small number of replicates and high uncertainty of cell clustering.
- A new statistical model, DCATS, was developed for differential composition analysis in a framework of beta-binomial regression. The use of beta-binomial regression helps to regress out the effects of confounding factors. DCATS also leverages a confusion matrix to correct the clustering bias and allows pre-estimated parameters across all cell types to account for its uncertainty.

Practical Experience

Key member, HKU DeepTech100 **12.2022-12.2023**

- Single cell sequencing is widely used in understanding various biological processes including cell differentiation and disease development. A wealth of cell atlas data has been published in recent years. We propose to fully utilize these extensive datasets, combined with rapidly developing AI techniques, to build digital twins of organs and disease models. These digital twins can accelerate drug development efficiently.
- Our team has been selected for the HKU DeepTech100, a co-ideation program aimed at transforming deep tech research ideas into businesses by providing more than 50, 000 HKD
- We won the Third Prize for Entrepreneurship Proposal in the 9th **Hong Kong University Student Innovation and Entrepreneurship Competition** and **Gold Prize** in the national university student innovation competition**.** We currently hold a patent.

Key Member, The Interdisciplinary Contest in Modeling **01.2017**

- Established a model to assess whether a city meets smart growth with Analytic Hierarchy Process (AHP)
- Chose two cities (Sydney and Yantian in Shenzhen City) as experiment objects, and predicted their future development tendency; conducted sensitivity tests to judge its stability and validity
- Composed a thesis “A Model of Sustainable Smart Growth to Evaluate and Plan smart growth of a City”, and won Honorable Mention

Vice President, Student Union, School of Life Sciences, SYSU **05.2015-05.2016**

- Was responsible for the basic operation and management of Academic Department and PR Department
- Communicate with other vice presidents or presidents, schools, or university departments
- Organized many school activities as the people in charge, including “SUSY Cutting-edge Lecture”, “Lab Tour”, “Biology Festival”, “Biology Experiment and Skills Contest”, etc.
- Worked with presidents from other student union in other universities and organized different activities

Skills

Experiment skills: basic cell biology experiment, microbiology experiment, biochemistry experiment, animal surgery
Computer skills: R, Python, Linux, SAS, MATLAB, C++, Oracle
English skills: CET-4 559; CET-6 556; TOEFL 102; GRE 320+3.5
Language skills: English, Mandarin, Cantonese, and Teochew