Guojie Zhong

PhD candidate, Department of Systems Biology, Columbia University
1130 St. Nicholas Avenue, New York, NY, USA | Phone: (+01)7187305993 | Email: gz2294@cumc.columbia.edu
Website: https://zhongguojie1998.github.io/GuojieZhong/

RESEARCH INTERESTS

I am interested in developing machine learning approaches to understand human disease mechanisms and fundamental biology. My past works expanded several areas including single cell spatial transcriptomics, statistical genetics and missense variant effect predictions. My current research interests include:

1. Disease mechanism discovery utilizing high-throughput screening techniques and machine learning.
2. Generative models for designing proteins, DNAs and RNAs

PUBLICATIONS

Missense variant effect prediction

- **Zhong, G.**, Zhao, Y., Zhuang, D., Chung, W.K. and Shen, Y. PreMode predicts mode-of-action of missense variants by deep graph representation learning of protein sequence and structural context. *bioRxiv*, 2024.02.20.581321 (2024).
- Zhao, Y., **Zhong, G.**, Hagen, J., Pan, H., Chung, W.K., and Shen, Y. A probabilistic graphical model for estimating selection coefficient of nonsynonymous variants from human population sequence data. *medRxiv*, 2023.12.11.23299809 (2023).

Statistical genetics in rare genetic diseases

- **Zhong, G.**, Choi, Y.A. and Shen, Y. VBASS enables integration of single cell gene expression data in Bayesian association analysis of rare variants. *Commun Biol* **6**, 774 (2023).
- **Zhong, G.**, and Shen, Y. Statistical models of the genetic etiology of congenital heart disease. *Curr Opin Genet Dev* **76**, 101967 (2022).
- **Zhong, G.**, Ahimaz, P., Edwards, N.A., Hagen, J.J., Faure, C., Lu, Q., Kingma, P., Middlesworth, W., Khlevner, J., El Fiky, M., et al. Identification and validation of candidate risk genes in endocytic vesicular trafficking associated with esophageal atresia and tracheoesophageal fistulas. *HGG Adv* **3**, 100107 (2022).

Single cell and spatial transcriptomics

- Su, J., Reynier, J.B., Fu, X., **Zhong, G.**, Jiang, J., Escalante, R.S., Wang, Y., Aparicio, L., Izar, B., Knowles, D.A., and Rabadan, R. Smoother: a unified and modular framework for incorporating structural dependency in spatial omics data. Genome Biol **24**, 291(2023).
- Ren, X.#, **Zhong, G.**#, Zhang, Q., Zhang, L., Sun, Y., and Zhang, Z. Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly. *Cell Res* **30**, 763-778 (2020). (#, Contributed Equally)
- Zhang, Q., He, Y., Luo, N., Patel, S.J., Han, Y., Gao, R., Modak, M., Carotta, S., Haslinger, C., Kind, D., Peet G.W., **Zhong, G.**, Lu, S., Zhu, W., Mao, Y., Xiao, M., et al. Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. *Cell* **179**, 829-845 e20 (2019).

CONFERENCES

- **Zhong, G.**, and Shen, Y. (2023). Predicting mode of action of missense variants by graph representation of protein structural context. *American Society of Human Genetics 2023 Annual Meeting, Washington, D.C.*
- **Zhong, G.**, and Shen, Y. (2022). Representation of missense variants for predicting modes of action. *Machine Learning in Structural Biology, Workshop at the 36th Conference on Neural Information Processing Systems.*
- **Zhong, G.**, Choi, Y.A., and Shen, Y. (2022). Integration of gene expression data in Bayesian association analysis of rare variants. *American Society of Human Genetics 2022 Annual Meeting,* Los Angelas, CA.
- **Zhong, G.**, Wang, J., He, S. and Fu, X. (2021). Towards better understanding of developmental disorders from integration of spatial single-cell transcriptomics and epigenomics. *The 2021 ICML Workshop on Computational Biology.*

EDUCATION

Department of Systems Biology, **Columbia University**, New York, NY, USA. *2019.08 – present*
PhD in Systems Biology, Integrated program in cellular, molecular and biomedical sciences
Thesis Advisor: Dr. Yufeng Shen
Thesis Committee: Dr. Wendy K. Chung, Dr. Mohammad AlQurashi, Dr. David Knowles

Department of Integrated Biology, **University of California, Berkeley**, CA, USA. *2017.08 – 2017.12*
Exchange Student in Berkeley Bioscience Study Abroad Program (BBSA)

Yuanpei College, **Peking University**, Beijing, China. *2015.09 – 2019.07*
Bachelor of Science in Integrated Science Program
Thesis Advisor: Dr. Zemin Zhang
Thesis: Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly

RESEARCH EXPERIENCES

PhD Student, Department of Systems Biology, Columbia University
Advisor: Dr. Yufeng Shen

PreMode: Predict mode-of-action of missense variants by graph representation of protein sequence and structural context. *2022.02-present*

- Accurate prediction of the functional impact of missense variants is one of the bottlenecks in discovering genetic causes of diseases and implementing genomic medicine. Current methods focused on generating an binary prediction score to distinguish pathogenic and benign variants while overlooking the fact that variant effect should be multi-dimensional.
- Pathogenic missense variants in the same gene may act through different modes of action. Molecularly, they may disrupt different functions such as loss of folding stability, binding affinity or enzymatic activity. Genetically, they can manifest as gain or loss of function (G/LoF).
- Developed PreMode, a foundation method to learn universal representation of sequence variation from protein context utilizing pre-trained protein language models, protein structures and multiple sequence alignments.
- We applied PreMode to predicting molecular and genetical mode-of-action of several protein and protein families through transfer learning, where it achieved the state-of-the-art performances.

Integration of gene expression data in Bayesian association analysis of rare variants. *2019.08-2022.02*

- The statistical power to identify risk genes by rare *de novo* variants is generally low due to rarity of genotype data. Previous studies have shown that disease risk genes usually have high expression in relevant cell types, although for many diseases the identity of these cell types are largely unknown. Recent efforts in single cell atlas in human and model organisms produced large amount of gene expression data.
- Developed VBASS, that integrate expression data to improve power of rare variants association analysis.
- VBASS models the association of disease risk as a function of expression profiles of relevant tissue or cell types in Bayesian frameworks. VBASS uses both analytical likelihood function and neural network approximations in joint probability calculation, and it learns the importance of cell types jointly from expression and genetics data.
- On simulated data, VBASS showed proper error rate control and better statistical power than state-of-the-art Bayesian methods.
- We applied VBASS to published datasets and identified more candidate risk genes than previous methods with supports from literature or data from independent cohorts.

Undergraduate Researcher, Biomedical Pioneering Innovation Center (BIOPIC), Peking University.
Advisor: Dr. Zemin Zhang, Dr. Xianwen Ren

3D single cell interaction network reconstruction based on ligand-receptor mediated self-assembly. *2018.03-2019.07*

- Developed CSOmap, an unsupervised machine learning algorithm that inference of cellular spatial organization and cellular interaction from scRNA-seq data. CSOmap predict cell relative distances based on ligand-receptor expression levels between cells.
- We applied CSOmap to five published cancer datasets, including liver carcinoma, lung carcinoma, carcinoma of colon and rectum, melanoma, head and neck cancer. CSOmap can successfully recapitulate spatial characteristics of corresponding tumor microenvironment, prioritize molecular determinants of cellular interactions, and generate biological insights consistent with literature via *in silico* interference.

AWARDS

Dean's fellowship, Graduate School of Arts and Science, Columbia University 2019
2014-2015 academic year Outstanding Freshman Scholarship, Peking University 2015
The 28th National Olympiad in Chemistry in Provinces, 1st prize, Zhejiang Province, China 2014

Skills

- Solid background in applications of Machine Learning, Deep Learning, Geometric Deep Learning and Statistical Learning to biological questions.
- Solid programming experience in Python (PyTorch), R, Matlab.
- Solid experiences single cell transcriptomics analysis and whole genome sequencing (WGS) analysis.