Yujie Zhang

Mobile : +1-619-458-8969 Email : yujiezhang@hsph.harvard.edu
linkedin.com/in/yujie-zhang-5a0633152 https://yuz682.github.io/

Education

- **Harvard University** Boston, MA
  M.S. in Computational Biology & Quantitative Genetics, Biostats Dept. *Sep. 2021 – May. 2023*
- **University of California, San Diego** San Diego, CA
  B.S. in Bioengineering:bioinformatics, Minor in Mathematics(GPA 4.0), overall GPA 3.80 *Sep. 2017 – May. 2021*

Publications

- Preparing glycomics data for robust statistical analysis with GlyCompareCT *Yujie Zhang, Krishnan Sridevi, Bokan Bao, Wan-Tien Chiang, James T. Sorrentino, Song-Min Schinn, Benjamin P. Kellman, Nathan E Lewis. bioRxiv 2022.05.31.494178; doi: https://doi.org/10.1101/2022.05.31.494178*

- A consensus-based and readable extension of Linear Code for Reaction Rules (LiCoRR) *Kellman, B. P.; Zhang, Y.; Logomasini, E.; Meinhardt, E.; Godinez-Macias, K. P.; Chiang, A. W. T.; Sorrentino, J. T.; Liang, C.; ... & Lewis, N. E. Beilstein J. Org. Chem. 2020, 16, 26452662. doi:10.3762/bjoc.16.215*

- Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis *Bao, B., Kellman, B. P., Chiang, A. W., Zhang, Y., Sorrentino, J. T., York, A. K., ... & Lewis, N. E. (2021). Nature Communications, 12(1), 1-14. doi:10.1038/s41467-021-25183-5*

- Bacterial modification of the host glycosaminoglycan heparan sulfate modulates SARS-CoV-2 infectivity *Martino, C., Kellman, B. P., Sandoval, D. R., Clausen, T. M., Marotz, C. A., Song, S. J., ... & Armingol, E. (2020). bioRxiv. doi: https://doi.org/10.1101/2020.08.17.238444 Preprint*

Experience

- **Hailiang Huang Lab (Stanley Centor at Broad Institute of MIT and Harvard)** Boston, MA
  Graduate Researcher *June 2022 - Present*
  - **Comparison between Statistical Finemapping and Colocalization**: Statistical Finemapping is used in GWAS or eQTL data to infer causal SNP variants; Colocalization is used to determine whether two or more traits share causal variants at a particular locus. The purposes of two methods are different but have similarities. I'm comparing the sensitivity and specificity of both methods under different scenerios by data simulation.
- **Heng Li Lab (Data Science Dept. at Dana Farber Cancer Institute)** Boston, MA
  Graduate Researcher *November 2021 - Present*
  - **High order repeat (HOR) analysis**: Analyzing HOR in human and non-human samples using Satellite Repeat Finder (SRF) Heng Li developed.
  - **Post assembly polishing**: One of the common assembly errors is false variant calling. I'm developing an efficient pipeline to polish the false calling variant using reads mapping information.
  - **Adjust Hifiasm consistency with Oxford Nanopore Technology (ONT) reads**: Hifiasm is a fast haplotype-resolved de novo assembler for PacBio HiFi reads. It does not perform as ideally on ONT reads as on HiFi reads. I'm working on interpreting the inconsistency and improving its performance on ONT reads.
- **Machine Learning for Healthcare at MIT** Boston, MA
  Project leader *January 2022 - July 2022*
  - **Survival analysis of cardiovascular disease risk among Systemic Lupus Erythematosus patients**: Existing cardiovascular disease (CVD) prediction tool perform poorly in Systemix Lupus Erythematosus (SLE) patients. We used semi-parametric, parametric, and deep learning survival models to build a CVD risk prediction tool for SLE patients.
- **Nathan Lewis Lab (Systems Biology And Cell Engineering Lab at UCSD)** La Jolla, CA
  Undergraduate Researcher *June 2019 - May 2022*
  - **Glyco Analysis Command Line Tool development**: GlyCompare is a program our lab developped to analyze glycans through decomposing them to a minimal set of intermediate substructures. I designed and impemented a command line tool called GlyCompareCT as well as developping additional features for GlyCompare. The tool is open-sourced on Github https://github.com/LewisLabUCSD/GlyCompareCT
  - **Linear Code Reaction Rules (LiCoRR) paper**: Linear Code is the most concise and parsable nomenclature for big data analytics of glycans. However, the use of Linear Code by the current field has been inconsistent from each other and from its original setting. In this paper, we are summarizing some accommodations we have seen, together with the original Linear Code implementation rules, to recommend a more consistent version of Linear Code in representing glycosynthesis. We name it Linear Code Reaction Rules (LiCoRR). The paper is published.

- **Identifying viruses in CHO cells in silico**: Chinese hamster ovary (CHO) cells are widely used cell lines to manufacture protein therapeutics in biopharmaceutical industries. Therefore carrying exogenous viruses is a huge risk for CHO cells. We designed and implemented a computational pipeline to automatically detect and quantify the viruses in CHO cells. One of its many functions is to avoid virus infected cells passing on to biopharmaceutical manufacturing.
- **Predict glycan motifs using machine learning**: This project is to predict glycan substructure presence at glycosylation sites given the protein surface. The goal is to apply the program to HIV and COVID-19 data. Through machine learning, we want to figure out the optimal sphere radius for predicting glycans and the best information source for predicting glycans.

- **NanoTools BioScience** — La Jolla, CA
  Internship — *March 2019 - March 2021*
  - **Image Analysis**: Independently Designed and implemented a program based on imageJ using macro language to detect the cardiomyocytes (cardiac muscle cell) contraction change under drug screening. The project decomposed to detection and tracking. The detection is realized by the relative grayscale difference. Then the tracking is realized by Frourier transform based single cell tracking algorithm. GitHub: https://github.com/yuz682/CardioDT

Courses

- **Courses taken**
  - **Genomics & Biology**: Genome Analysis, Genetic Epidemiology, Evolutionary and Quantitative Genomics, Analysis of Genetic Association Study, Molecular Sequencing Analysis, Molecular Biology, Biotech Thermodynamics, Biomolecular Engineering, Applied Genomic Technologies
  - **Math & Statistics**: Linear Algebra, Vector Calculus, Differential Equations, Applied Regression Analysis, Statistical Reasoning, Probability and Statistics, Discrete Mathematics, Decision Analysis Methods
  - **Computer Science**: Advanced Data Structure, Machine Learning, Deep learning, Machine Learning for Healthcare, Image Processing, Algorithm Design and Analysis
- **Course projects**
  - **Dynamics of disease transmission and human behavior**: We trained LSTM and BiLSTM models on Google search activity data to predict 5-day COVID-19 confirmed cases and achieved good performance. We also demonstrated the generalizability of the model by comparing the model performances with and without explicit COVID-19 terms.

Professional Activities

- **American Society of Human Genetics (ASHG) Annual Meeting 2022** — Los Angeles, CA
  Attendee — *Oct 2022*
- **X Academy (The largest interdisciplinary innovation summer program in China)** — Shanghai, China
  Computational Biology Academic Leader — *Aug 2021*
- **Undergraduate Bioinformatics Club(UBIC) at UCSD** — UCSD, CA
  Academic Relations Chair — *May 2018 - May 2019*

Award

- **Triton Research and Experiential Learning Scholars (TRELS)** — UCSD, CA
  Glycan Database project — *Sep. 2019*
- **Triton Research and Experiential Learning Scholars (TRELS)** — UCSD, CA
  LiCoRR paper — *Jan. 2020*

Skillset

- **Languages**: **Python**, SQL, SPARQL, MATLAB, TensorFlow, macro, Java, R, C++, **Linux**, Latex, HTML, Flask
- **Software**: ImageJ, Bash, IGV, GCP