Yujie Zhang

Mobile : +1-619-458-8969 Email : zyj1729@gmail.com
linkedin.com/in/yujie-zhang-5a0633152 https://yuz682.github.io/

Education

- **University of California, San Francisco** San Francisco, CA
  Ph.D. in Biological & Medical Informatics *Sep. 2023 – 2028*
- **Harvard University** Boston, MA
  M.S. in Computational Biology & Quantitative Genetics, Biostats Dept., GPA 3.8 *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 3.8 *Sep. 2017 – May. 2021*

Publications

- De novo reconstruction of satellite repeat units from sequence data *Zhang, Y., Chu, J., Cheng, H. and Li, H. (2023). arXiv. doi: https://doi.org/10.48550/arXiv.2304.09729 Preprint*
- Preparing glycomics data for robust statistical analysis with GlyCompareCT *Zhang, Y., Krishnan, S., Bao, B., Chiang, A. W., Sorrentino, J. T., Schinn, S. M., ... & Lewis, N. E. (2023) STAR protocols, 4(2), 102162. doi: https://doi.org/10.1016/j.xpro.2023.102162*
- 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. (2020) 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*

Research Experience

- **Hailiang Huang Lab (Stanley Center at Broad Institute of MIT and Harvard)** Boston, MA
  Graduate Researcher *June 2022 - May 2023*
  - **Comparison between Statistical Finemapping and Colocalization**: Simulated eQTL and GWAS data based on UK BioBank genotypes to compare the sensitivity and specificity of two causal inference methods, statistical finemapping and colocalization, under varying genetic architectures.
- **Heng Li Lab (Data Science Dept. at Dana-Farber Cancer Institute)** Boston, MA
  Graduate Researcher *November 2021 - May 2023*
  - **High order repeat (HOR) analysis**: Analyzed the distribution and abundance of HOR in human and non-human samples. HOR is tandem array of larger repeat units consisting of multiple basic repeat units.
  - **Post assembly polishing**: Developed a computational pipeline to polish the false calling variants on assemled genome contigs using reads mapping information. The pipeline ourperforms the commonly-used polishing tool, Racon, in sensitivity and specificity.
- **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**: Fitted semi-parametric, parametric, and deep-learning-based survival models to identify cardiovascular disease (CVD) risk factors specifically for Systemic Lupus Erythematosus (SLE) patients. Found five consistently significant CVD risk factors for SLE patients.
- **Nathan Lewis Lab (Systems Biology And Cell Engineering Lab at UCSD)** La Jolla, CA
  Undergraduate Researcher *June 2019 - May 2022*
  - **Bioinformatics tool development for glycomic analysis**: Built a command line version of GlyCompare, a Python-based bioinformatics tool for large glycomic data analysis, to simplify its usage. Significantly improved its running memory and implemented parallel computing for runtime gain. Open-sourced on Github https://github.com/LewisLabUCSD/GlyCompareCT

- **Linear Code Reaction Rules (LiCoRR) paper**: Summarized inconsistent usage of a glycan nomenclature called Linear Code among different research groups. Recommended a more consistent version of Linear Code, named Linear Code Reaction Rules (LiCoRR), in representing glycosynthesis. Published at Beilstein Journal of Organic Chemistry.
- **Single cell virus detection and quantification**: Created a computational pipeline to automatically detect and quantify the viruses in Chinese hamster ovary (CHO) cells. One of its many functions is to avoid virus infected cells passing on to biopharmaceutical manufacturing.
- **Glycan motifs prediction using machine learning**: Predicted glycan substructure presence at glycosylation sites given the protein surface structures. Fitted machine learning models including Naive Bayes and Random Forest to assess the optimal sphere radius and the best information source for predicting glycans.

- **NanoTools BioScience** — La Jolla, CA
  Internship — *March 2019 - March 2021*
  - **Microscopic video analysis**: Designed and implemented a program based on ImageJ, an image and video analysis tool, to detect and track the cardiomyocytes (cardiac muscle cells) contraction change under drug screening in microscopic videos. GitHub: https://github.com/yuz682/CardioDT

Coursework

- **Courses taken**
  - **Genomics & Biology**: Statistical Genetics, 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 project**
  - **Dynamics of disease transmission and human behavior**: Trained LSTM and BiLSTM models on Google search activity data to predict 5-day COVID-19 confirmed cases. Achieved good performance. 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 2023**
  Poster presentation selected — *Nov 2023*
- **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*

Awards

- **Cum Laude graduate honor** — UCSD, CA
  Top 14% graduating seniors — *May. 2021*
- **Triton Research and Experiential Learning Scholars (TRELS)** — UCSD, CA
  LiCoRR paper — *Jan. 2020*
- **Triton Research and Experiential Learning Scholars (TRELS)** — UCSD, CA
  Glycan Database project — *Sep. 2019*

Skillset

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