Last Updated on 03/25/2023

Zhongyi (James) Guo

(+1) 607-262-3415 ● guozy@stanford.edu ● Palo Alto, CA 94304 ●
GitHub ● LinkedIn ● Twitter ● Personal Website

EDUCATIONS

Stanford University, Palo Alto, CA (Expected) Jun. 2025
M.S. Epidemiology & Clinical Research
GPA: 3.91/4.00

Cornell University, Ithaca, NY May 2023
B.S. Biological Sciences (Computational Biology) and **Biometry & Statistics** (Statistical Genetics)
GPA: 3.57/4.30, Honors: CUM LAUDE, Dean's List

PUBLICATION

Presenter/First Author, Causal effect of type II diabetes on prostate cancer in the East Asian population: A two-sample Mendelian randomization study, AACR Special Conference: Aging and Cancer, 2022 *(Published)*

SKILLS

Programming: R, SAS (Based certified), Python, Java, Swift, SQL, HTML, CSS, UNIX/Linux, LaTeX
Developer Tools: Git, GitHub, Terminal, Jupyter Notebook, RStudio, Eclipse, Overleaf, & Xcode
Core Skills: GWAS, Multi-omics Analysis, GEO Data Mining, scRNA-seq, TCGA, Causal Inference, Matching, Survival Analysis, Statistical Modeling, Machine Learning, Data Science, Data Structure, Shell Scripting, iOS Development, Website Design, Microsoft Office

SELECTED RESEARCH EXPERIENCES

Research Assistant at Graff Lab, San Francisco, CA Oct. 2023 – Present
- Utilized ChemRICH to study ethnic disparities between black and white men in a prostate cancer pilot study, interpreting chemical similarity enrichment analysis results using R
- Developed a web-scraping tool to retrieve Compound IDs and SMILES from PubChem using R
- Replicated and validated t-test results (mean values and p-values) along with Storey's q-value using the False Discovery Rate (FDR) approach based on Metabolon's report on metabolomics data in R

Research Assistant at Lan Lab, Beijing, China Apr. 2021 – Jul. 2021
- Designed a High Throughput Screening (HTS) to study the interaction between high-affinity TCR and pMHC (major histocompatibility complex) in human and used k-means clustering to create prototypes
- Found the optimal set of mutation locations in yeast to inhibit phosphoprotein functions by performing CRISPR/Cas9 every week and then analyzed the returned sequences of purified plasmids extracted

SELECTED PROJECT EXPERIENCES

hurdatPro May 2023
Github Repository
- Developed an R package to analyze Atlantic tropical cyclone activities through collaboration
- Cleaned data, designed functions for storm plotting (track, position, size), identified and predicted U.S. landfalls, computed storm accumulated cyclone energy, and implemented unit tests using testthat

Causal Effect of Type II Diabetes on Prostate Cancer in East Asian Population May 2022 – Dec. 2022
GitHub Repository
- Performed two-sample Mendelian randomization with the inverse variance weighted method while using MR Egger and weighted median methods as sensitivity analysis on GWAS summary data
- Identified proxy SNPs in linkage disequilibrium ($r^2 > 0.8$) and obtained OR = 0.76, 95% CI = [0.76, 0.89], P-value = $2.26 * 10^{-6}$ and similar results in sensitivity analysis.
- Concluded that Type II diabetes has a negative causal effect on prostate cancer using genetic evidence

GWAS Study: Analysis of Lymphoblastoid Cell Lines (LCL) mRNA Levels Apr. 2022 – May 2022
GitHub Repository
- Analyzed genotype & phenotype data and tested whether population and sex as two covariates could influence the GWAS result and cause differences in LCL mRNA level expressions
- Identified significant SNPs from Manhattan & QQ plots and causal polymorphisms
- Concluded that population and sex as two covariates do not impact the GWAS result significantly

GWAS Study: Analysis of Citrulline Levels and Chronic Kidney Disease May 2022
GitHub Repository
- Performed GWAS analysis on citrulline levels and chronic kidney disease data using two PCs obtained from PCA as covariates on genotype data, and Bonferroni correction to reduce Type I error
- Identified 2 significant SNPs from Manhattan plot with 2 covariates included and interpreted the influence of linkage disequilibrium on the result

Weather Data Analysis in Ithaca, NY from 2021.01 to 2022.04 Mar. 2022 – Apr. 2022
GitHub Repository
- Built a Logistic Regression model and a K-Nearest Neighbors (KNN) model to forecast snow in Ithaca, NY, based on daily temperature range using train-test split after data cleaning & EDA
- Reached the model accuracy at 0.809 for the Logistic Regression and 0.786 for the KNN with k = 10 and plotted confusion matrices for two models' tuning & validating and error analysis

α-helix or not? Dec. 2021
GitHub Repository
- Performed feature engineering on the training set by averaging each feature of 5 neighboring amino acids and removed redundant features measured by correlation coefficient
- Trained binary classifiers (Logistic Regression, Decision Tree Regressor, and Random Forest models) using Sklearn to predict α-helix or not using features derived from primary structures of proteins
- Tuned the maximum number of iterations using random search method to optimize the Logistic Regression model, and conducted cross-validation and examined model accuracy (AUROC = 0.625).

Salaries in Big Techs Sep. 2021 – Dec. 2021
GitHub Repository
- Built a multiple linear regression model to predict total yearly salaries based on employee features, including years of experience, gender, race, education, etc., for tech companies in US and overseas
- Established 3 equations for users to optimize their incomes by predicting total yearly salary in the U.S

TEACHING EXPERIENCES

Beta Tester & Teaching Assistant, Introduction to Data Science Jan. 2023 – May 2023
Grader, Probability Models and Inference Aug. 2022 – Dec. 2022
Teaching Assistant, Laboratory in Genetics and Genomics Jan. 2021 – May 2021
Teaching Assistant (Summer), JNC Study Abroad Platform Jul. 2022 – Aug. 2022

INDUSTRY EXPERIENCES

Tencent Ltd., Data Analyst Project Intern (Remote) Jul. 2021 – Sep. 2021
- Extracted e-commerce sales statistics using Python web scraping and using SQL on internal databases
- Developed predictive machine learning models for forecasting sales trends using Sklearn, analyzing customer shopping patterns across product categories, and built multiple linear regression models.
- Refined marketing team's strategies by presenting my elegant and informative visualization, along with model building, in my data analysis report created using Python in Jupyter Notebook.

EXTRACURRICULAR ACTIVITIES

Community HealthEd, Education Branch – Scientific Review Editor Mar. 2022 – May 2023
- Focused primarily on maternal health, prenatal health, neurological & psychiatric health materials
- Visited each scientific paper/website cited in articles to validate the accuracy of the cited information
- Removed technical jargon from each article while retaining the meaning to make articles written clearly and concisely in plain language accessible to the general public as newsletters
- Cooperated efficiently with the authors, copy editors, and community outreach coordinators