Yaqin Si

Raleigh, NC

siyaqin.syq@gmail.com — 9197988635 — Google Scholar — GitHub — LinkedIn

Core Skill Set

- Excellent Programming: Data cleaning, visualization, algorithm implementation, reproducible pipelines, package development.
- Tools: R, Python, SAS, Julia, SQL, TreeAge, Tableau, GitHub, Shell scripts; Linux; HPC platform; AWS
- Statistics & Machine Learning: Advanced statistical modeling, simulation studies, experimental design, deep learning and large language models
- Data types: genomics data (WGS, CNV/SV, SNP) and real-world heterogeneous healthcare data
- Collaboration: Extensive experience working with interdisciplinary teams across academia, public health agencies, and industry
- Communication: Scientific writing, technical documentation, and presentations for both technical and non-technical audiences

Education

- **North Carolina State University** — Raleigh, NC
  Ph.D. Candidate in Bioinformatics – Advisors: Dr. Jung-Ying Tzeng & Dr. Wenbin Lu — Aug 2020 – May 2026
  Comajor: Master in Statistics (GPA 3.9)
  Relevant Coursework: Bioinformatics, Statistical Theory/Methods, Neural Networks, Functional genomics, Computation for Statistical Research, Computational Methods for Molecular Biology
- **Peking University** — Beijing, China
  Master in Epidemiology and Health Statistics, GPA 3.7/4.0 — Aug 2016 – Jul 2018
  Bachelor of Preventive Medicine & Bachelor of Economics, GPA 3.7/4.0 — Aug 2011 – Jul 2016

Experience

- **Research Assistant in Bioinformatics** — Raleigh, NC
  Bioinformatics Research Center, NC State University — Aug 2020 – present

DNA structural variants association analysis (Method Development)

- Developed penalized regression models using line-graph-guided penalties (Weighted Fusion) and Lasso to perform variable selection and effect estimation of phenotype-associated DNA structural variants.
- Designed simulations to benchmark statistical power, estimation accuracy, and robustness against existing methods. **The manuscript in revision at *the Annals of Applied Statistics***

R package development & Scalable Implementation

- Released CNVreg R package on CRAN implementing genome-wide CNV association analysis.
- Achieved 1,600+ downloads by the end of 2025, with full documentation and reproducible vignette.

Joint analysis of DNA structural variants and SNPs

- Built scalable computational pipelines to support joint analysis across CNV/SV and SNPs,
- Integrated haplotype imputation, heritability estimation, rare haplotype clustering, and polygenic risk score construction to improve association power in whole-genome sequencing studies.

- **Data Scientist** — Beijing, China
  Beijing HealthCom Data Technology Company — Jun 2018 – May 2019
  - Performed large data cleaning and feature engineering from large-scale datasets using SQL on Hadoop.
  - Built automated, reproducible pipelines for statistical analysis and reporting.
  - Linked public climate data to build time-series models to forecast disease burden forecasting.
  - Built interactive Tableau dashboards to communicate insights to non-technical stakeholders.

- **Research Assistant Intern** Zhejiang, China
  Center of Disease Control, Yinzhou District Jun 2015 – Mar 2016
  - Conducted epidemiological surveillance and statistical analysis of chronic disease.
  - Integrated electronic health records, insurance claims, and surveillance data for longitudinal analysis of incidence, prevalence, and mortality.
  - Built multivariate regression and survival models to identify risk factors and predict lifetime disease risk.
  - Delivered data-driven recommendations for public health prevention strategies.

PROJECT

- **Bioinformatics Research Center, North Carolina State University** 2020 – present

  Genome Assembly from PacBio Sequencing Reads (Co-Lead)

  Extract potential target reads based on reference genome using DIAMOND blastx and Kraken2;
  Performed genome assembly with Hifiasm for 2 target species;
  Evaluated assembly quality with QUAST and BUSCO, summarizing the genome size, contig N50, and gene completeness and redundancy.

  Terrain Identification from wearable device signals using Neural Networks(Co-Lead)

  Processed the time series data with overlapping window slicing;
  Implemented the Fully-connected Neural Network, CNN, and RNN with LSTM cell using Keras of TensorFlow;
  Tuned the network structure and hyperparameters based on validation loss and F1 score.

  Naturally-inspired algorithm for multiple sequence alignment(Co-Lead)

  Devided long sequences into more manageable pieces with Devide&Conquer strategy using genetic algorithm;
  Implemented Ant Colony Optimization algorithm for multiple DNA sequence aslignment, tuning hyperparameter (such as pheromone and evaporation rate) for optimal alignment paths.

  Hands-on experience on Large Language Models and Generative AI (Co-Lead)

  Built interactive text and image processing chatbots using ChatGPT and LLaMA APIs.
  Developed user-facing web applications with Streamlit and Flask.
  Applied fine-tuning and Retrieval-Augmented Generation (RAG) for diagnostic modeling in plant and human disease applications.

- **School of Public Health, Peking University** 2016 – 2018

  Disease prevention strategy evaluation based on risk prediction models

  Conducted a systematic review and meta-analysis of statin effect for cardiovascular disease prevention (Lead);
  Built Cox Proportional Hazards Model for cardiovascular disease risk prediction (Participate).
  Evaluated public health intervention strategies using Markov Chain Monte Carlo simulations in TreeAge;
  Quantified saved quality-adjusted life year and cost-effectiveness ratios to assist decision-making (Lead)

PUBLICATION (REFER TO GOOGLE SCHOLAR FOR A FULL LIST) AND DELIVERY

- Si Y., Lu W., Tzeng J, et al, CNV-profile regression for CNV association analysis with whole genome sequencing data. **In revision at The Annals of Applied Statistics;**
  Preprint available at bioRxiv (2024) https://www.biorxiv.org/content/10.1101/2024.11.23.624994v1
- Si Y, Holloway S, Tzeng, J. CNVreg: An R package for CNV association analysis
  CRAN https://cran.r-project.org/web/packages/CNVreg/index.html

ADDITIONAL INFORMATION

- GGB symposium (2022) Best Student Presentation Award
- Teaching assistant: Ph.D-level course ST721 Statistical Genetics (2023)