Anqi (Angie) Zhu, PhD

(919) 599-7989 ▪ zhua0513@gmail.com ▪ linkedin.com/in/aqzhu91/▪ azhu513.github.io/

SUMMARY

- Dedicated statistician with **11+** years of experience in developing and applying statistical methods for medical research
- Experienced in collaborating with non-statisticians to strategize and successfully execute experiment design and statistical analysis
- Published **17** (**6** first/co-first author) papers in peer-reviewed journals
- Renowned as an effective communicator, adept collaborator and team-player

WORK EXPERIENCE

Genentech, Inc., South San Francisco, CA — Sept 2021 – Present

Principal Scientist, Bioinformatics (July 2024 – Present)
Senior Scientist, Bioinformatics (Sept 2021 – June 2024)

- Lead the development of the statistical analysis workflow of bulk and single-cell amplicon-seq data for validating a CRISPR base editing technique-based platform for accelerated drug resistant variant discovery. The co-first author manuscript is published on Cell Reports.
- Co-lead a cross functional team to build an end-to-end solution for Optical Pooled Screen (OPS) technology development with discovery scientists, tech-dev scientists, pathologists, computational scientists and software engineers, covering from raw image generation, quality check of images, raw image processing to statistical analysis and result visualization.
- Implemented a python workflow for fluorescence cell image processing, in-situ sequencing barcode calling, immunofluorescence and cell paint phenotype features extraction and statistical analysis for colorectal cancer OPS screens, and contributed it to the internal unified OPS software.
- Selected to present the above-mentioned workflow at gRED Group Leader Offsite in 2023.
- Hired and managed two full-time contractors and two summer interns

Rotation, Early Clinical Development (Jan 2024 - June 2024)

- Hands-on work to support on biostatistics and clinical science functions for Vixarelimab phase II study for Idiopathic pulmonary fibrosis (IPF)
- Investigate intercurrent event (ICE) handling strategies with different missing value imputation methods to support the statistical analysis plan (SAP) for phase II Vixarelimab IPF data analysis
- Practiced medical data review and protocol deviation queries

23andMe, Inc., South San Francisco, CA — Sept 2019 - Sept 2021

Scientist I, Computational Biology

- Apply statistical methods and build computation tools to analyze the 23andMe genetic database and integrate external genomics datasets to identify novel therapeutic targets
- Conducted Loss-of-Function variants analysis and build the interactive html report for target evaluation
- Polygenic Risk Score (PRS) modeling to impute biomarker values for 23andMe genetic database using UK-Biobank data
- Deep learning algorithms for variant-to-gene mapping of GWAS hits integrating eQTL, epigenetic annotations, functional genomics data and annotations
- Provide statistical support to ongoing therapeutic research and development programs including trial design for the phase I study of 23ME-006

UNC Lineberger Comprehensive Cancer Center, Chapel Hill, NC May 2017 – Aug 2019

Graduate Research Assistant, Biostatistics Core

- Collaborated with oncologists to strategize and/or conduct experiment design, survey sampling, data management and statistical analysis on multiple clinical and genomics projects
- Provided biostatistics consultation to cancer center members every week

Clinical Projects

- Processed data from multiple sources including survey data, electronic case report forms (eCRFs), electronic medical records (EMR), patient registries, patient reported outcome, clinician reports, experiment data and wearable devices data; familiar with database software like REDCap, OnCore and Qualtrics
- Conducted survival analysis, generalized linear models, mixed effect models and survey sampling methods
- Reviewed study protocols, wrote analysis plans and prepared manuscripts (11 published papers)

Genomics Projects

- Applied Hidden Markov Model (HMM) on ATAC-seq and CHIP-seq data to examine the chromatin status across different cell lines and age groups
- Applied multivariate analysis and linear mixed effects models on the proteomics data of kinases expression over time to explore the drug targets
- Performed Gene Ontology (GO) and network analysis using DAVID software
- Evaluated a quantification pipeline of label-free mass spectrometry-based proteomics built on OpenMS

Genentech, Inc., South San Francisco, CA June 2018 - August 2018

Intern, Cancer Immunology

- Analyzed RNA-Seq expression data sets of two single-cell experiments to discover potential targets for combinatorial immune-therapy and to understand the biological functions of a transcription factor
- Integrated an analysis pipeline of single-cell RNA-Seq data sets from data quality control, normalization, controlling nuisance covariates, clustering to gene and gene sets differential expression analysis
- Collaborated with immunologists on analysis plan and results delivery with biological evidences, resulting in a presentation to a team of immunologists

RESEARCH EXPERIENCE

Statistical Methods for Sequencing Count Data and Integrative Functional Genomics (Doctoral Dissertation) Oct 2016 – Aug 2019

- Proposed an empirical Bayes method with adaptive priors to the coefficients in generalized linear models (GLMs) and various approximation techniques to provide the approximate posterior distribution
- Applied the proposed Bayesian method to the RNA-Seq count data for shrinkage estimates of the log-2 fold changes (LFC) that is more accurate and provides better inference for biological decisions
- Proposed nonparametric rank tests with Gibbs or bootstrap samples to test differential expression of RNA-seq with quantification uncertainty using inferential replicate counts with accommodation to multiple experiment designs, and proposed using permutation to build the null distribution

- Applied the proposed nonparametric test procedure to bulk and single cell RNA-seq data and showed improved control of false discovery rate in particular with transcripts with high inferential uncertainty
- Proposed Bayesian hierarchical models to determine the colocalization of GWAS and eQTL signals and further estimate the causal effect of genes to complex traits

Causal Effects of Drugs / Drug-Drug Interactions on Adverse Event using EMR Data (Advisor: Dr. Donglin Zeng) Sept 2014 - May 2017

- Worked on matched case-control design samples with five million patient records from Electronic Medical Record (EMR) database to determine the drug/drug-drug interactions (DDI) causal effect on the adverse event (ADE) myopathy
- Proposed a conditional causal log-Odds Ratio (OR) definition to characterize individual causal effects, and a control-based propensity score method with spline estimation to estimate this conditional causal log-OR with observational data for better confounders adjustment that provided consistent estimates
- Proposed a nonparametric supreme testing procedure of weighted combination of multiple single effects with logistic regression models to optimize the power in detecting DDI induced adverse events that improved power and controlled type I error in simulation study
- Tackled computational challenges of large-scale health record databases by dimension reduction techniques and meta-analysis methods
- Implemented proposed methods to the EMR data in R resulted in discovery of 70 single drug and some novel DDIs

SELECT PUBLICATIONS (For a full list of publications, please visit my google scholar page)

1. Dorighi K*, **Zhu A***, Fortin JP, Lo JH, Sudhamsu J, Durinck S, Callow M, Foster S, Haley B, (2024). Accelerated drug resistant variant discovery with an enhanced, scalable mutagenic base editor platform. *Cell Reports*, 43(6), e114314.
2. **Zhu A***, Matoba N*, Wilson E, Tapia AL, Li Y, Ibrahim JG, Stein JL and Love MI, (2021). MRLocus: identifying causal genes mediating a trait through Bayesian estimation of allelic heterogeneity. *PLoS genetics* 17 (4), e1009455.
3. **Zhu A**, Zeng D, Shen L, Ning X, Li L, and Zhang P, (2020). A super-combo-drug test to detect adverse drug events and drug interactions from electronic health records in the era of polypharmacy. *Statistics in Medicine,* 39(10), 1458-1472.
4. **Zhu A**, Ibrahim JG and Love MI, (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*. 35(12), 2084-2092.
5. **Zhu A**, Srivastava A, Ibrahim JG, Patro R and Love MI, (2019). Nonparametric expression analysis using inferential replicate counts, *Nucleic Acids Research*, 47(18), e105.
6. **Zhu A**, Zeng D, Zhang P and Li L, (2019). Estimating causal log-odds ratio using the case-control sample and its application in the pharmaco-epidemiology study. *Statistical methods in medical research*, p.0962280217750175.
7. Szeto A, Bucci T, Deal A, **Zhu A**, Ahmad M, Cass A, Sketch M, Kemper R, Zeidner J, Foster M, Muluneh B, and Crona D, (2022). Response to Tyrosine Kinase Inhibitors in Real-World Patients With Chronic Myeloid Leukemia. *Annals of Pharmacotherapy 56 (7), 753-763.*
8. Ehlers M, Bjurlin M, Gore J, Pruthi R, Narang G, Tan R, Nielsen M, **Zhu A**, Deal A, and Smith A, (2021). A national cross-sectional survey of financial toxicity among bladder cancer patients. *Urologic Oncology: Seminars and Original Investigations 39 (1), 76. e1-76. e7.*
9. Malpica Castillo LE, Palmer S, **Zhu A**, Deal AM, Chen SL, and Moll S, (2020). Incidence and time course of neutropenia in patients treated with rituximab-based therapy for non-malignant immune-mediated hematologic diseases. *Am J Hematol.* 2020;95(5):E117-E120. doi:10.1002/ajh.25751.

R PACKAGES

1. **Zhu A**, Ibrahim JG and Love MI (2017). **apeglm**: Approximate Posterior Estimation for GLM Coefficients. R package with the *Bioconductor* project.
2. **Zhu A,** Srivastava A, Ibrahim J, Patro R, Love M, (2019). **fishpond**: differential transcript and gene expression with inferential replicates. R package with the *Bioconductor* project.

SELECT PRESENTATIONS AND LECTURES

1. **Oral presentation** "Identifying new targets for colorectal cancer using optical pooled screens", gRED Group Leader Offsite, Napa CA, June 2023
2. **Invited Talk** "Understanding gene-to-trait effect with GWAS and eQTL summary statistics with Bayesian hierarchical modeling", ICSA2021, online, September 2021
3. **Paper Presentation** "Nonparametric expression analysis using inferential replicate counts," JSM2019, Denver CO, July 2019
4. **Paper Presentation** "Application of *t* priors to sequence count data: removing the noise and preserving large differences," ENAR2018, Atlanta GA, March 2018

PROFESSIONAL TRAININGS AND SERVICES

- **Deep Learning Specialization Certificate**
  DeepLearning.AI, Coursera, May 2020
- **Meetup Organizer**, Bay area Biotech-pharma Statistical Workshop (BBSW)
  San Francisco Bay Area, June 2020 - Present
- **2019 JSM Diversity Workshop and Mentoring Program**
  Attendee, JSM2019, Denver CO, July 2019
- **Session Chair**
  Contributed Session: Methods for Single Cell Analysis, ENAR 2018, Atlanta GA, 2018

EDUCATION

University of North Carolina at Chapel Hill, Chapel Hill, NC

- **Ph.D.** in Biostatistics Advisor: Dr. Joseph G. Ibrahim and Dr. Michael I. Love August 2019
  Dissertation: *Statistical Methods for Sequencing Count Data and Integrative Functional Genomics*
- **M.S.** in Biostatistics Advisor: Dr. Michael G. Hudgens May 2015
  Thesis: *Rank-based Approaches to Cox Model for Interval Censored Data*

Capital University of Economics and Business, Beijing, China

- **B.S.** in Statistics (China National Scholarship and Honor Graduate) June 2013

TECHNICAL SKILLS

- R for analysis and software development, SAS for data management, processing and analysis and SQL for data query
- Python for fluorescence microscopy cell image processing, quality control, feature extraction, data analysis and visualization
- SAS Certified Base Programmer and SAS Certified Advanced Programmer
- Extensive experience with working under the UNIX environment and bash script writing