Anqi (Angie) Zhu, PhD

zhua0513@gmail.com | (919) 599-7989

SUMMARY

- Dedicated statistician with 8 years of experience in developing and applying statistical methods for medical research
- Experienced in collaborating with non-statisticians to strategize and execute experiment design and statistical analysis
- Published 5 methodological papers (first-author) and 7 collaborative papers in peer-reviewed journals
- Effective communicator, collaborator and team-player

EDUCATION

University of North Carolina, Gillings School of Global Public Health — Chapel Hill, NC
- **Ph.D.** in Biostatistics — August 2019
- **M.S.** in Biostatistics — May 2015

Capital University of Economics and Business — Beijing, China
- **B.S.** in Statistics — June 2013

TECHNICAL SKILLS

- R for analysis and software development, SAS for data management, processing and analysis and SQL for data query
- SAS Certified Base Programmer and SAS Certified Advanced Programmer
- Extensive experience with working under the UNIX environment and bash script writing

WORK EXPERIENCE

23andMe, Inc. — South San Francisco, CA
Scientist I, Computational Biology — Sept 2019 - Present
- Apply statistical methods and build computation tools to analyze the genetic database and integrate external genomics datasets to identify novel therapeutic targets
- Provide statistical support to ongoing therapeutic research and development programs

UNC Lineberger Comprehensive Cancer Center — Chapel Hill, NC
Statistical Research Assistant, Biostatistics Core — May 2017 – Aug 2019
- 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

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
Intern, Cancer Immunology June 2018 - August 2018

- 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 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 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

Rank-based Approaches to Cox Model for Interval Censored Data Sept 2014 - May 2015

- Proposed rank-based approaches applying to Cox proportional hazards model for interval censored data that can be implemented easily with SAS
- Proved empirically that using mid-rank and non-uniformly weighted average rank provide better estimates of regression coefficient than other single value imputation methods through simulation studies

SELECT PUBLICATIONS

Method paper

1. **Zhu A**, Zeng D, Zhang P and Li L, (2018). 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.
2. **Zhu A**, Ibrahim JG and Love MI, (2018). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*. 35(12), 2084-2092.
3. **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.
4. **Zhu A**, Zeng D, Shen L, Ning X, Li L, & 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.
5. **Zhu A**, Matoba N, Wilson E, Tapia AL, Li Y, Ibrahim JG, Stein JL and Love MI, (2020). MRLocus: identifying causal genes mediating a trait through Bayesian estimation of allelic heterogeneity. *Biorxiv*

Selected collaboration paper

6. Ehlers M, Bjurlin M, Gore J, Pruthi R, Narang G, Tan R, Nielsen M, **Zhu A**, Deal A, Smith A (2020). A national cross-sectional survey of financial toxicity among bladder cancer patients. In *Urologic Oncology: Seminars and Original Investigations*. In press. Elsevier.
7. Malpica Castillo LE, Palmer S, **Zhu A**, Deal AM, Chen SL, Moll S. 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

For a full list of publications, please visit my google scholar page

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. **Paper Presentation** "Nonparametric expression analysis using inferential replicate counts," JSM2019, Denver CO, July 2019
2. **Paper Presentation** "Application of *t* priors to sequence count data: removing the noise and preserving large differences," ENAR2018, Atlanta GA, March 2018
3. **Poster Presentation** "An empirical Bayes approach for differential expression analysis of RNA-Seq data," BioC2017, Boston MA, June 2017
4. **Workshop Lecture** "Git and R notebook for reproducible research," Cold Spring Harbor NY, June 2017

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