# WENJIA WANG

1800 5th Ave, Pittsburgh, PA 15219 | +1 (530) 5649145 | wew89@pitt.edu | https://wenjiaking.github.io/

## EDUCATION BACKGROUND

**University of Pittsburgh** *Aug. 2020 - present*

*Ph.D. in Biostatistics* *GPA: 3.98/4*

- Advisor: Prof. George C. Tseng
- Research Interest: Statistical Computing, Meta-Analysis of Homogeneous or Heterogeneous Signals, High-dimensional Clustering, Multi-omics Analysis, Statistical Genomics and Genetics
- Revelant Coursework: Longitudinal Data Analysis, Omics Data Analysis, Survival Analysis, Machine Learning (Carnegie-Mellon University), Advanced Deep Learning (Carnegie-Mellon University), Epidemiology, Molecular Basis of Human Inherited Deseases, etc.

**University of California, Davis** *Sep. 2018 – Dec. 2019*

*M.S. in Statistics* *GPA: 3.91/4*

- Statistics Coursework: Computational Statistics, Categorical Data Analysis, Stochastic Processes, etc.

**East China Normal University** *Sep. 2013 – Jul. 2017*

*B.S. in Mathematics and Applied Math* *Major GPA: 3.85/4*

- Advanced Math Coursework*:* Abstract Algebra, Real Analysis, Functional Analysis, Numerical Analysis, Topology, etc.

## PROFESSIONAL EXPERIENCE

**Graduate Student Researcher** *Aug. 2020 - present*

*Department of Biostatistics, School of Public health, University of Pittsburgh* *Part Time: 20 Hours per Week*

- Provided statistical consulting for multiple investigators from Pittsburgh Liver Research Center, a partnership of University of Pittsburgh & UPMC, and Liza Konnikova Lab, School of Medicine, Yale University
- Analyzed high-throughput multi-omics data, including RNA-seq (single-cell/bulk) and metabolomics data from raw data preprocessing using command-line tools to downstream analysis such as differentially expressed (DE) analysis, pathway enrichment analysis, trajectory analysis, cell-to-cell interaction analysis, and transcriptomic meta-analysis
- Published and co-authored collaboration papers in cancer, human mucosal immunity and other disease research

**Graduate Teaching Fellow** *Jan. 2023 – May. 2023*

*Department of Biostatistics, School of Public health, University of Pittsburgh*

- Primary instructor of BIOST 2094: Advanced R Computing. Responsible for delivering lectures and lab sessions, preparing assignments and exams, assessing students' final projects, and grading

**Biostatistics Intern** *May. 2022 – Aug. 2022*

*Pfizer, Inc* *Cambridge, MA*

*Mentors: Simon (Xingpeng) Li and Chong Duan* *Full Time: 40 Hours per Week*

- Developed a shiny app to conveniently compare gene signature improvement for the diseases of Atopic Dermatitis and Psoriasis by Pfizer's therapies to competitor drugs
- Developed the Precision Medicine Statistics Shiny Hub to host and well organize Shiny Apps
- Constructed an effective AI model to predict drug response using baseline transcriptomic data after delicate data integration, compresensive comparison of logistic regression, random forest, convolutional neural network and variational auto-encoder, towards to patient stratification and precision medicine

## SELECTED TALKS

- (August 2023; invited) "Accurate and Ultra-Efficient p-Value Calculation for Higher Criticism Tests" , Joint Statistical Meetings (JSM) 2023, Toronto, Ontario, Canada
- (April 2023; invited) "Overview of multi-omics data analysis and horizontal data integration", ASA-SSGG Short Course Series: Selective Introduction to Multi-Omics Analysis, April 2023

## SELECTED HONORS AND AWARDS

- ASA Sections on Computing and Graphics Student Paper Awards, 2023
- Excellent Graduate of East China Normal University, 2017
- First-class & Second-class National Scholarship: for top 3% and 10% students respectively, 2014 & 2015
- University-level First Prize of Daxia Fund Projecrt, 2015

## SELECTED RESEARCH PROJECT

**Outcome-Guided Disease Subtyping Using High-Dimensional Omics Data** *Feb. 2023 – Nov. 2023*
*Department of Biostatistics, School of Public health, University of Pittsburgh*

- Developed a unified latent generative model with feature selection to perform outcome-guided clustering with omics data
- Proposed a weighted joint likelihood model to adaptively emphasize omics pattern or outcome association in clustering
- Extended the models from continuous to survival outcome by incorporating an accelerated failure time model
- Derived the implementation algorithm of our methods based on EM and coordinate descent algorithms
- Conducted extensive simulations to compare our models with existing clustering methods and applied to the lung disease and breast cancer transcriptomic data respectively to identify the disease subtypes
- Published in *The Annals of Applied Statistics*

**Combining p-Values: Historical Development, Recent Advances and Future Opportunities** *Oct. 2022 – Dec. 2023*
*Department of Biostatistics, School of Public health, University of Pittsburgh*

- Comprehensively reviewed the methods of combining p-values in traditional meta-analysis, independent rare and weak singal detection, and dependent signal detection
- Evaluated the theoretical properties, power in finite-sample practice, and computing strategies of all methods
- Discussed variations and weighting schemes of the p-value combining methods for power improvement

**IFDlong: an isoform fusion detector on long-read RNA-seq data** *May. 2021 – Dec. 2023*
*Department of Biostatistics, School of Public health, University of Pittsburgh*

- Developed a bioinformatic tool for isoform and fusion Detection on long-read RNA sequencing data called IFDlong
- With the input of long RNA sequences in fastq file, the pipeline embeds aligment step and can annotate the long reads with known gens and isoforms, detect novel isoforms, quantify isoform expression by a novel estimation maximization algorithm, as well as discover novel fusions and quantify fusion transcripts at isoform level.
- Conducted comprehensive comparison with the existing tools for long-read RNA sequencing data analysis by extensive simulations, artificially data and real cancer data

**Accurate and Ultra-Efficient p-Value Calculation for Higher Criticism Tests** *Jan. 2021 – Dec. 2022*
*Department of Biostatistics, School of Public health, University of Pittsburgh*

- Proposed the cross-entropy based importance sampling method with appropriate importance density to efficiently compute the null distribution for Higher Criticism (HC) test statistic and benchmark the other computing methods.
- Modified the existing analytic computing to improve the numerical stability and flexibility
- Achieved fast vectorizing computation for enormous number of HC tests by constructing the quantile-probability curves
- Won ASA Sections on Computing and Graphics Student Paper Awards, 2023
- Published in *Journal of Computational and Graphical Statistics*

**Neurons Extraction from in-vivo Calcium Imaging Data by Statistical Learning** *Aug. 2019 – Feb. 2020*
*Department of Statistics, University of California, Davis*

- Explored the structures of various imaging data, theoretically analyze the efficiency and drawbacks of unsupervised basis learning approaches in extracting neurons from different calcium imaging data
- Conducted simulations to remove neuropil/out-of-focus contamination and extract subcellular compartments

## SELECTED PUBLICATIONS

### Statistical Methodology

*1.* **Wenjia Wang**, Yusi Fang, Chung Chang, George Tseng (2023+). "Accurate and ultra-efficient p-value calculation for higher criticism tests." *Journal of Computational and Graphical Statistics, to appear.*

2. Yujia Li*, Peng Liu*, **Wenjia Wang***, Wei Zong, Yusi Fang, Zhao Ren, Lu Tang, George C. Tseng (2023+). "Outcome-guided Disease Subtyping by Generative Model and Weighted Joint Likelihood in Transcriptomic Applications." *The Annals of Applied Statistics, minor revision.* (*co-first author)

### Application

1. Liu Silvia, Yu Yan-Ping, Ren Bao-Guo, Ben-Yehezkel Tuval, Obert Caroline, Smith Mat, **Wang Wenjia**, Ostrowska Alina, Soto-Gutierrez Alejandro, Luo Jian-Hua (2023). "Long-read single-cell sequencing reveals expressions of hypermutation clusters of isoforms in human liver cancer cells." *eLife, to appear*.
2. Yuan Gui, Yanbao Yu, **Wenjia Wang**, Yuanyuan Wang, Hanyue Lu1, Sarah Mozdzierz, Eskander, Kirollos, Yi-Han Lin, Hanwen Li, Xiaojun Tian, Silvia Liu, Dong Zhou (2023+). "Proteomes characterization of liver-kidney comorbidity after microbial sepsis." *Molecular Metabolism, submitted*.
3. Silas A. Buck, Sophie A. Rubin, Tenzin Kunkhyen, Christoph D. Treiber, Xiangning Xue, Lief E. Fenno, Samuel J. Mabry, Varun R. Sundar, Zilu Yang, Divia Shah, Kyle D. Ketchesin, Darius D. Becker-Krail, Iaroslavna Vasylieva, Megan C. Smith, Florian J. Weisel, **Wenjia Wang**, M. Quincy Erickson-Oberg, Emma I. O'Leary, Eshan Aravind, Charu Ramakrishnan, Yoon Seok Kim, Yanying Wu, Matthias Quick, Jonathan A. Coleman, William A. MacDonald, Rania Elbakri, Briana R. De Miranda, Michael J. Palladino, Brian D. McCabe, Kenneth N. Fish, Marianne L. Seney, Stephen Rayport, Susana Mingote, Karl Deisseroth, Thomas S. Hnasko, Rajeshwar Awatramani, Alan M. Watson, Scott Waddell, Claire E. J. Cheetham, Ryan W. Logan, Zachary Freyberg (2023+) "Sexually dimorphic mechanisms of VGLUT-mediated protection from dopaminergic neurodegeneration" *cell, submitted.*

## SELECTED SOFTWARE

HCp

- An R package that includes functions of different methods for p-value computation of Higher Criticism test
- Tutorials available at https://github.com/wenjiaking/HCp/tree/master

ogClust

- An R package that implements two outcome-guided clustering methods for disease subtyping: the generative model ($ogClust_{GM}$) and the weighted joint likelihood model ($ogClust_{WJL}$).
- Tutorials available at https://github.com/wenjiaking/ogClust

## TECHNICAL SKILLS

- *Programming Language*: Proficient in R, bash, Python, LaTex, SQL, SAS, MATLAB and STATA
- *Statistical Methodologies*: Familiar with various statistical models, AI models including CNN, RNN, Hidden Markov Models, Bayesian Networks, Reinforcement Learning, Recommender Systems, and computer vision such as image processing, feature extraction, segmentation and classification
- *Statistical Computing Skills*: Hands-on experience in analyzing omics and clinical data with R packages limma, DEseq2, monocle3, Seurat, etc., Python packages scanpy, gseapy, phate, etc., and CellPhoneDB
- *Communication and Writing Skills*: Strong communication and writing skills from long-term collaborations with non-statistical researchers of various backgrounds