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 - Apr. 2025*

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

- Advisor: Prof. George C. Tseng
- Research Interest: Statistical Computing, Meta-Analysis, High-Dimensional Clustering, Bayesian Analysis
- Revelant Coursework: Longitudinal Data Analysis, Omics Data Analysis, Survival Analysis, Bayesian Data 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 the 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 & Jan. 2024 – May. 2024*

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

Pfizer, Inc *Cambridge, MA*

Mentors: Simon (Xingpeng) Li and Chong Duan *May. 2022 – Aug. 2022*

- 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
- Predicted drug response using baseline transcriptomic data after data integration, comprehensive comparison of AI models such as convolutional neural network, variational auto-encoder, etc., toward precision medicine

Eli Lilly and Company *Indianapolis, IN*

Mentors: Haoyan Hu, Lars Lau Raket, Suktae Choi, Hong Wang *May. 2024 – Aug. 2024*

- Learned the biological mechanism for the side effect of the amyloid-targeting therapies in Alzheimer's disease
- Developed an efficient predictive model to predict the adverse event given the proteomics data at the baseline
- Analyzed the longitudinal high-dimension data to identify the biomarkers with delicate feature selection

SELECTED TALKS

- (August 2024; contributed) "IFDlong: an isoform fusion detector on long-read RNA-seq data", JSM
- (August 2023; invited) "Accurate and Ultra-Efficient p-Value Calculation for Higher Criticism Tests", JSM
- (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,contributed poster) "Accurate and Ultra-Efficient p-Value Calculation for Higher Criticism Tests", ENAR

SELECTED AWARDS AND CERTIFICATES

- ASA Pittsburgh Chapter Student of the Year, 2024
- ASA Sections on Computing and Graphics Student Paper Awards, 2023
- Excellent Graduate of East China Normal University, 2017
- American Society of Actuaries: Financial Mathematics Course, 2017
- First-class & Second-class National Scholarship: for top 3% and 10% students respectively, 2014 & 2015

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.*
2. Yujia Li*, Peng Liu*, **Wenjia Wang***, Wei Zong, Yusi Fang, Zhao Ren, Lu Tang, George C. Tseng (2024). "Outcome-guided Disease Subtyping by Generative Model and Weighted Joint Likelihood in Transcriptomic Applications." *The Annals of Applied Statistics.* (*co-first author)
3. **Wenjia Wang**, Yuzhen Li, Sungjin Ko, Ning Feng, Manling Zhang, Jia-Jun Liu, Songyang Zheng, Baoguo Ren, Yan P. Yu, Jian-Hua Luo, George C. Tseng, and Silvia Liu (2024+). "IFDlong: an isoform and fusion detector for accurate annotation and quantification of long-read RNA-seq data." *Nucleic Acids Research, under review.*

Application

1. **Wenjia Wang**, Weihong Gu, Ron Schweitzer, Omry Koren, Soliman Khatib, George Tseng, Liza Konnikova (2024+). "In utero human intestine contains maternally derived bacterial metabolites." *Microbiome, under review*.
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 (2024). "Proteomes characterization of liver-kidney comorbidity after microbial sepsis". *The FASEB Journal*.
3. Lauren Smith, Eduardo Gonzalez Santiago,Chino Eke, **Wenjia Wang**, Dhana Llivichuzhca-Loja, Tessa Kehoe, Kerri St Denis, Madison Strine, Sarah Taylor, George Tseng, Liza Konnikova (2024). "Maternal and donor human milk support robust epithelial growth and differentiation in a fetal intestinal organoid model". *Gastro Hep Advances*.
4. Silvia Liu, Caroline Obert, Yan-Ping Yu, Junhua Zhao, Baoguo Ren, Kelly Brease, Benjamin Krajacich, **Wenjia Wang**, Kyle Metcalfe, Mat Smith, Tuval Ben-Yehezkel, Jianhua Luo (2024). "Utility Analyses of AVITI Sequencing Chemistry". *BMC Genomics*.
5. 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*.

SELECTED RESEARCH PROJECT

Unsupervised Concensus Clustering by Integrating Multi-Type High-Dimensional Data *May. 2024 – Apr. 2025*
Department of Biostatistics, School of Public health, University of Pittsburgh

- Integrated multiple types of high-dimensional data and identified their multi-facet consensus clustering patterns by using Bayesian latent consensus hierarchical framework
- Selected efficient priors and regularization approaches, and inferred Bayesian estimates by MCMC algorithm
- Conducted simulations to evaluate the multi-facet clusterings, and applied to real microbiome and metabolomic data

Multi-Outcome-Guided Bayesian Concensus Clustering on High Dimensional Omics Data *Dec. 2023 – Dec. 2024*
Department of Biostatistics, School of Public health, University of Pittsburgh

- Developed a Bayesian consensus hierarchical framework embedded with feature selection to identify the latent clustering in the high-dimension omics data, semi-supervised by multivariate outcomes.
- Derived the implementation Markov chain Monte Carlo (MCMC) algorithm to efficiently infer Bayesian estimates
- Conducted extensive simulations to compare our model with existing clustering methods and applied to real data

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 on *The Annals of Applied Statistics*

Combining p-Values: Historical Development, Recent Advances and Future Opportunities *Oct. 2022 – Sep. 2024*
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
- Manuscript ready. Submit to *Annual Review of Statistics and Its Application*

IFDlong: an isoform fusion detector on long-read RNA-seq data *May. 2021 – Sep. 2024*
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
- Submit to *Nucleic Acids Research*

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 on *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 SOFTWARE

`HCp` (GitHub)

- An R package that includes functions of different methods for p-value computation of Higher Criticism test

`ogClust` (GitHub)

- An R package that implements two outcome-guided clustering methods for disease subtyping: the generative model ($ogClust_{GM}$) and the weighted joint likelihood model ($ogClust_{WJL}$).

`IFDlong` (GitHub)

- A unified software performing long-read RNA-seq alignment, isoform/fusion annotation, and isoform/fusion quantification

TECHNICAL SKILLS

- *Programming Language*: Proficient in R, bash, Python, LaTex, SQL, SAS, MATLAB and STATA
- *Statistical Methodologies*: Familiar with various statistical models, Bayesian inference, statistical computation (optimization and sampling computing), 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