# Chenghuiyun (Cassie) Xu

cathawennn@gmail.com
(404) 247- 4724

## SKILLS

**Software**: R (shiny, phyloseq, lme4), bash (plink, admixture), Python, SAS, STATA, SQL

**Language**: Mandarin (native), English (fluent)

**Research/Admin**: Genetics, Gene Alignment, Genetic Ancestry, RNA-Sequencing, Cancer Research, TCGA, Modeling, Forecasting, Machine Learning (Optimization, Multiclass Classification), Time Series, Advanced Mathematics and Statistics (Multivariate Regression, Generalized Linear Model), Bayesian Statistics, Data Mining, Deep Learning

## RESEARCH & WORK EXPERIENCE

**Biostatistician (intern)** *June 2023 – September 2023*

*Jazz Pharmaceuticals, Palo Alto, CA*

- Replaced third-party tools such as EAST, for clinical trial accrual and time-to-event prediction, by writing my own R application to optimize future investment (capital and resource) allocation and allow custom features and models to be added to better align with company and client requirements.
- Significantly boosted phase III clinical trial accrual and time-to-event prediction accuracy, achieving estimates within a 3-month window when using 50% real data and within a 1-month window with 75% real data when applying statistical models such as AFT exponential and Weibull models, piecewise models, and Bayesian models.
- Delivered a 45-minute presentation on my program during an all-hands meeting to hundreds of people including the CEO, introducing potential trial applications for my tool.

**Data Scientist (intern)** *June 2021 – September 2021, June 2022 – September 2022*

*InterVenn Biosciences Corporation, South San Francisco, CA*

- Improved on the quality control procedure for glycoproteomic data produced by liquid chromatography–mass spectrometry which included
  - building an R Shiny app to visualize Levey-Jennings charts for biomarkers of interest,
  - testing a batch effect removal tool, Combat-Seq, on healthy controls from different batches,
  - and developing revised tools using both Bayesian and Frequentist regression methods to significantly improve the reduction of batch effects.
- Programmed automated report for clients that generates analytical results such as data quality checks, differential expression analysis, and machine learning performance assessment from user-submitted proteomic and glycoproteomic data.
- Improved visualizations and user experience of the report by making them interactive via HTML and JavaScript.
- Implemented feature engineering to extract features from existing glycoprotein data using non-parametric dimensional reduction methods such as t-SNE, and compared the result with that of using PCA.
- Assessed the prediction performance of the extracted features on patient characteristics data using 5-fold cross validation and achieved an AUC of 95%.

**Predoctoral Researcher** *September 2019 – Present*

*University of California – Davis, CA*

- Performed gene alignment (STAR), quality control, subtype classification (PAM50), differential expression analysis (limma-voom and DESeq2) and pathway analysis (GSEA) on 271 whole exome RNA-sequencing data from formalin-fixed, paraffin-embedded (FFPE) breast tumor tissues collected on more than 2000 patients from the Instituto Nacional de Enfermedades Neoplasicas (INEN), in Lima, Peru.
- Conducted quality control, ancestral estimations and PCA analysis on germline genome-wide genotype data with 795,842 variants using Plink and R.
- Wrote shell scripts to obtain data (29.17 TB) from the NIH GDC Data Portal for the TCGA-BRCA study. Applied the PEGEN-BC analytical pipeline on the TCGA-BRCA data to compare their results, and discovered a more aggressive profile of Luminal subtypes in the studied samples from Peru.

**Principal Research Analyst** *May 2018 – August 2019*

*Weill Cornell Medicine, New York, NY*

*Project I:*

- Led a team of 3 to estimate treatment effects of a-VISTA, Cyclophosphamide and Radiation Therapy on breast cancer by analyzing 533-sample 16S rRNA gene sequence data of human and mouse microbiomes using QIIME2 and R.

- Revamped phylotree structure by rewriting the relationship between parent and child nodes, and created an interactive visualization where selected nodes are highlighted with their adjusted p-values.

*Project II:*

- Explored the effect of an EHR prescribing redesign on both opioid prescribing choices and keystrokes across the 2 sites with 22,113 patients received a new short-acting opioid prescription from 821 providers by applying segmented regression to conduct interrupted time series analysis (ITS).
- Discovered that an unobtrusive "nudge" involving changing the default opioid prescription option was associated with an increase in CDC guideline concordance in a setting with low baseline concordance, but not in a different setting where concordance was already high.

**Research Analyst** *January 2017 – May 2018*

*Weill Cornell Medicine, New York, NY*

*Project I:*

- Built logistic regression model in R to evaluate the association between two biomarkers, exosome level and exosome number, and autism status for children from 1 to 15 years old.
- Assessed out-of-sample prediction performance of 70% by performing five-fold cross validation repeated ten times to obtain the average area under receiver operating characteristic curve and its corresponding 95% confidence interval. The optimal cutoff points were identified using Youden's approach.

*Project II:*

- Perform a literature review to analyze the regression models used in previous studies to study the relationship between the components of frailty syndrome: body composition, strength, and physical performance level, and develop new applicable models.
- Perform structural equation model analysis on the data collected in the Rancho Bernardo Study.

## EDUCATION

**University of California, Davis** September 2019 – Present

*Ph.D. in Biostatistics* Davis, CA

*Research Topic:* Pathway analysis of breast cancer RNA-Sequencing data for a comparison of cancer intrinsic subtypes in different populations classified by ancestral components (Project Advisor: Dr. David Rocke & Dr. Laura Fejerman)

*Awards:* Dean's Graduate Summer Fellowship Award, Biostatistics Fellowship Stipend and Assistantship

---

**Cornell University** September 2017 – December 2018

*Master of Science in Biostatistics and Data Science* New York, NY

---

**Agnes Scott College** August 2013 – May 2016

*Bachelor of Arts in Mathematics and Economics* Decatur, GA

*Awards:* Phi Beta Kappa, Summa Cum Laude, Omicron Delta Epsilon, Omicron Delta Kappa

## PUBLICATIONS

- Ancker, Jessica S., J. Travis Gossey, Sarah Nosal, **Chenghuiyun Xu**, Samprit Banerjee, Yuming Wang, Yulia Veras, Hannah Mitchell, and Yuhua Bao. "Effect of an electronic health record "nudge" on opioid prescribing and electronic health record keystrokes in ambulatory care." *Journal of general internal medicine 36, no. 2 (2021): 430-437*.
- RoyChoudhury, Arindam, and **Chenghuiyun Xu**. "A dataset on body composition, strength and performance in older adults." *Data in brief 29 (2020): 105103*.
- RoyChoudhury, Arindam, Thuy-Tien L. Dam, **Chenghuiyun Xu**, Jonathan H. Diah, Deepa Chaganty, Jonathan Solares, and Linda P. Fried. "Feed-forward loop between body composition, strength and performance in older adults." *Mechanisms of Ageing and Development 183 (2019): 111130*.
- Kang, J., K. A. Pilones, C. Daviaud, J. Kraynak, M. E. Rodriguez-Ruiz, S. Demaria, J. E. Park, **C. Xu**, X. K. Zhou, and S. C. Formenti. "VISTA Blockade Immunotherapy in a MULTI-Modal Approach to Triple Negative Breast Cancer (TNBC) in MICE and IMPACT on Microbiome." *International Journal of Radiation Oncology, Biology, Physics 105, no. 1 (2019): S88-S89*.
- Katherine Brooke*, Denisse Saucedo*, **Chenghuiyun Xu***. "Second-Order Linear Recurrence Relations and Periodicity" (2017). *The Onyx Review: The Interdisciplinary Research Journal, Vol. 2, No. 2, pp. 7-12.* *All authors contributed equally to this work and manuscript.