Douglas Yao

douglasyao@g.harvard.edu • (408) 510-4019 • douglasyao.github.io

EDUCATION

Harvard University

- Ph.D. in Computational Biology — Jul 2018 – Jun 2023 (expected)

University of California, Los Angeles

- B.S. in Molecular Biology (GPA: 3.82, magna cum laude) — Sep 2014 – Jun 2018
- Additional coursework: Discrete Structures (Math 61, Grade: A), Algorithms and Complexity (CS 180, Grade: A), Probability (Stat 100A, Grade: A+), Mathematical Statistics (Stat 100B, Grade: A), Linear Algebra (Math 33A, Grade: A-), Computational Genetics (CS 124, Grade: A), Advanced Programming (PIC 10C, Grade: A), Machine Learning (CS 226, Grade: A-)

EXPERIENCE

Harvard University — Jul 2018 – Present

Computational Biology Researcher

- PIs: Alexander Gusev and Brian Cleary
- Led research projects in two different fields, statistical genetics and functional genomics. Wrote two first-author publications published/in revision in top journals.
- Project 1: Conceived and developed software for a novel statistical method to estimate disease heritability mediated by gene expression levels. Applied method to real human disease genetics data from UK Biobank and RNA-seq data from the Genotype-Tissue Expression Consortium. Wrote first-author manuscript published in Nature Genetics (#1 genetics journal by impact factor). Paper cited >150 times since publication in 2020.
- Project 2: Developed a novel computational and experimental framework to increase the efficiency of Perturb-seq, a type of experimental assay that combines pooled CRISPR screening with single-cell RNA-seq, by up to 20x using random composite perturbations. Worked closely with experimental scientists to implement and test framework in a macrophage cell line. Wrote first-author manuscript currently in revision at Nature Biotechnology (#1 biotechnology journal by impact factor).

nference, Inc. — Aug 2020 – Mar 2021

Biomedical Data Scientist

- Developed software integrating data from publicly available data sources and electronic health records from the Mayo Clinic to identify novel drug targets and drug combinations.

University of California, Los Angeles — Feb 2016 – Jun 2018

Computational Biology Researcher

- PIs: Eleazar Eskin and Thomas Graeber
- Conceived and led a project applying linear mixed models to identify genes whose expression is associated with genomic instability in cancer. Analyzed RNA-seq data from The Cancer Genome Atlas. Wrote first-author publication in Scientific Reports.

SKILLS

Programming languages: Python, R, Unix/Linux

Statistics/machine learning: Dimensionality reduction (PCA, NMF, UMAP), clustering, penalized regression (scikit-learn), generalized linear models (glmnet), convex optimization (cvxpy), matrix completion (softImpute), Bayesian statistical inference (Stan, Tensorflow Probability), deep learning (Keras, Tensorflow), variance component models

Genetics/genomics: Single-cell/bulk RNA-seq, SNP array data, WGS/WXS, Perturb-seq, differential expression analysis (DESeq2), read alignment (STAR), gene set enrichment analysis (GSEA)

Genomics datasets: UK Biobank, 1000 Genomes Consortium, The Cancer Genome Atlas, Genotype Tissue Expression Consortium

SELECTED PUBLICATIONS

Yao DW, Binan L, Bezney J, Simonton B, Freedman J, Frangieh CJ, Dey K, Geiger-Schuller K, Eraslan B, Gusev A*, Regev A*, Cleary B*. Compressed Perturb-seq: highly efficient screens for regulatory circuits using random composite perturbations. *Nature Biotechnology*, in revision. *Co-supervised

Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics*. **52**, 626–633 (2020).

Yao DW, Balanis NG, Eskin E, Graeber TG. A linear mixed model approach to gene expression-tumor aneuploidy association studies. *Scientific Reports*. **9**, 1-8 (2019).

OTHER PUBLICATIONS

Siewert-Rocks KM, Kim SS, **Yao DW**, Shi H, Price AL. Leveraging gene co-regulation to identify gene sets enriched for disease heritability. *The American Journal of Human Genetics*. **109**, 393–404 (2022).

Mandric I, Yang HT, Strauli N, Montoya D, Rotman J, Van Der Wey W, Ronas JR, Statz B, **Yao DW**, Zelikovsky A, Spreafico R, Shifman S, Zaitlen N, Rossetti M, Ansel M, Eskin E, Mangul S. Profiling immunoglobulin repertoires across multiple human tissues by RNA sequencing. *Nature Communications*. **11**, 3126. (2020).

Mitchell K, Brito JJ, Mandric I, Wu Q, Knyazev S, Chang S, Martin LS, Karlsberg A, Gerasimov E, Littman R, Hill BL, Wu NC, Yang HT, Hsieh K, Chen L, Littman E, Shabani T, Enik G, **Yao DW**, Sun R, Schroeder J, Eskin E, Zelikovsky A, Skums P, Pop M, Mangul S. Benchmarking of computational error-correction methods for next-generation sequencing data. *Genome Biology*. **21**, 71 (2020).

PRESENTATIONS

Yao DW, Binan L, Bezney J, Simonton B, Freedman J, Frangieh CJ, Dey K, Gusev A*, Regev A*, Cleary B*. A new framework for efficient Perturb-seq enables cheap large-scale dissection of the innate immune response and provides insight into regulatory eQTL relationships. *American Society of Human Genetics Annual Meeting*. October 2022. Los Angeles, CA. *Co-supervised.

Yao DW, Frangieh C, Simonton B, Bezney J, Cleary B, Gusev A. Dissecting regulatory disease networks via integration of Perturb-seq, eQTL, and GWAS. *American Society of Human Genetics Annual Meeting*. October 2020. Virtual.

Yao DW, Gusev A, Cleary B. Integrating Perturb-seq with eQTL and GWAS data to dissect regulatory disease networks. *Broad Institute V2F meeting*. June 2020. Boston, MA.

Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by gene expression levels. *American Society of Human Genetics Annual Meeting*. October 2019. San Diego, CA.

Yao DW, O'Connor LJ, Price AL, Gusev A. Estimating heritability mediated by gene expression levels for complex traits. *14th Annual Broad Institute Scientific Retreat*. December 2018. Boston, MA.

Yao DW, Balanis NG, Eskin E, Graeber TG. Identifying transcriptional programs associated with tumor aneuploidy using the linear mixed model. *UCLA Research Poster Day*. May 2018. Los Angeles, CA.

Yao DW, Sheu KM, Balanis NG, Graeber TG. Concordant molecular signatures across prostate neuroendocrine and lung small cell cancers. *UCLA Small Cell Cancer Group Collaborative Meeting*. October 2016. Los Angeles, CA.

Yao DW, Balanis NG, Graeber TG. Bioinformatic analysis of pan-cancer genomic instability signatures. *UCLA Biomedical Research Summer Poster Symposium*. August 2016. Los Angeles, CA.

AWARDS & SCHOLARSHIPS

American Society of Human Genetics, Reviewer's Choice Abstract — Sep 2019, Oct 2020, Oct 2022
- Awarded to top 10% of submitted abstracts

National Science Foundation, Graduate Research Fellowship Program — Mar 2020
- Awards an annual stipend of $34,000 for three years to pursue independent research

University of California, Los Angeles, Undergraduate Research Scholars Program — Sep 2017
- Awards $5,000 to undergraduates with outstanding research

Amgen, Amgen Scholars Program — Jun 2017
- Awards $4,000 to pursue research over the summer at UCLA

University of California, Los Angeles, Summer Biomedical Research Scholarship — Jun 2016
- Awards $4,000 to pursue research over the summer at UCLA

ACTIVITIES

Clinical and Translational Science Institute, UCLA

Clinical Research Associate — Dec 2015 – Sep 2017
- Assisted doctors with collecting and processing clinical data for clinical studies on various disorders including atherosclerosis and polycystic ovary syndrome