

# Douglas Yao

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

## EDUCATION

### Harvard University

- Ph.D. in Computational Biology

Jul 2018 – Jun 2022 (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

#### Statistical Genetics Researcher

- PIs: Alexander Gusev and Alkes Price
- Developed statistical methods to characterize the genetic basis of complex diseases including diabetes, schizophrenia, and cancer in human populations
- Derived a novel method of moments estimator of a quantity known as “heritability mediated by gene expression levels.” Established unbiasedness and calibration of estimator via derivations and simulations. Applied method to real human disease genetics data from UK Biobank and RNA-seq data from the Genotype-Tissue Expression Consortium.
- Wrote first author publication in *Nature Genetics* (top genetics journal in the world by impact factor, 27.6). Paper cited 40 times in half a year since publication.

### nference, Inc.

Aug 2020 – Present

#### 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
- As an undergraduate, 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**, 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).

## PRESENTATIONS

**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**

- National Science Foundation**, Graduate Research Fellowship Program Mar 2020
- Awards an annual stipend of \$34,000 for three years to pursue independent research
- American Society of Human Genetics**, Reviewer's Choice Abstract Sep 2019
- Awarded to top 10% of submitted abstracts
- 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