Han Fang

Menlo Park – CA 94025
☎ (631) 413-7226 • ✉ hanfang.cshl@gmail.com • hanfang.github.io

EDUCATION

Stony Brook University (SBU) — **Stony Brook, NY**
Ph.D in Applied Mathematics and Statistics (AMS) — *2014 - 2017*
M.S. in Applied Mathematics and Statistics — *2011 - 2013*

Sun Yat-sen University (SYSU) — **Guangzhou, China**
B.S. in Optical Informatics — *2007 - 2011*

EXPERIENCE

Facebook — **Menlo Park, CA**
Data Scientist — *July 2017 - Present*

- Developed machine learning models on geo-spatial data to solve connectivity problems at a global scale

Facebook — **Menlo Park, CA**
Data Scientist Intern — *Jun 2016 - Aug 2016*

- Built machine learning models to predict user engagement with extremely high precision recall on billions of users and find effective strategies for product & infrastructure problems using big data techniques.
- Applied statistical modeling methods and developed automated pipelines in production to analyze large-scale web/mobile data; identified patterns and provided actionable recommendations.
- Developed optimization models on global geo-spatial data and solved problems which benefit 30 countries.

Cold Spring Harbor Laboratory (CSHL) — **Cold Spring Harbor, NY**
Research Assistant — *Aug 2014 - Jun 2017*

- Machine learning methods for large-scale genomics data
  - Developed Scikit-ribo, a statistical framework for analyzing large-scale genomics data in Python
    - Achieved 98% classification accuracy on millions of sequences using random forest and recursive feature selection.
    - Implemented a module using negative binomial mixture modeling to identify peaks from over-dispersed data.
    - Built a generalized linear model (GLM) with ridge penalty to robustly estimate thousands of parameters.
  - Maintaining glmnet for python (github.com/bbalasub1/glmnet_python)
    - Efficiently fits lasso, ridge, and elastic-net regularized GLM via penalized maximum likelihood for linear, logistic, multinomial, and Poisson regression
    - Supports outcome prediction, cross-validation, sparse scipy matrix, and range constraints on coefficients
  - Led a group in a data science hackathon and built classifiers to predict cancer cell types
    - Performed dimension reduction with sparse PCA, identified samples groups using hierarchical clustering.
    - Built classifiers to predict cancer types at 89% accuracy using Logistic regression with Elastic net regularization.
  - Led statistical analyses of four major studies to identify candidates of interest from massive omics data.
- Graphical algorithms for analyzing genomics data
  - Developed Topsorter for graphical assessment of structural variants (github.com/hanfang/Topsorter)
    - Traverses & finds the longest path with topological sorting of a weighted directed acyclic graph (DAG).
  - Developed Forceps, a program for comparing & choosing optimal sequences (github.com/hanfang/forceps)

Cold Spring Harbor Laboratory — **Cold Spring Harbor, NY**
Computational Science Developer — *June 2013 - June 2014*

- Computational methods for next-generation sequencing data analysis
  - Developed Scalpel, a C++ software to detect genomic mutations (scalpel.sourceforge.net)
    - Built modules, reviewed and optimized codes for de Bruijn graph assembly of millions sequences.
    - Deployed a Google cloud pipeline for analyzing and visualizing results (github.com/hanfang/scalpel-protocol).

SKILLS

Proficient: Python, SQL, R, C/C++, UNIX, Hive, Shell. **Familiar:** HPC, Hadoop, Spark, Java

Awards and fellowship

- 2017 President's Award to Distinguished Doctoral Students @ SBU (5 out of 449)
- 2017 The Woo-Jong Kim Dissertation Award @ AMS (1 out of 34)
- 2017 Excellence in Research Award @ AMS (5 out of 34)
- 2016 Research Access Project @ SBU
- 2016 Department Travel Grant @ AMS
- 2015 Reviewers' Choice @ The American Society of Human Genetics (Top 10%)
- 2015 Summer Institute in Statistics for Big Data Scholarship @ University of Washington
- 2014 Research Assistant Fellowship @ CSHL
- 2013 Research Access Project @ SBU
- 2013 Department Travel Grant @ AMS
- 2010 Outstanding Student Scholarship @ SYSU (Top 10%)

PUBLICATIONS

Under review:

- Nattestad, Goodwin, Ng, Baslan, Sedlazeck, Rescheneder, Garvin, **Fang**, Gurtowski, Hutton, Tseng, Chin, Beck, Sundaravadanam, Kramer, Antoniou, McPherson, Hicks, McCombie, Schatz, "Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line", *Under review* (2017)
- Yang, Chen, Lima, **Fang**, Jimenez, Li, Lyon, He, Wang, "PennCNV-Hadoop: Accurate Detection of Copy Number Variation from Whole Genome Sequencing Data", *Under review* (2017)

Peer-reviewed:

- Sedlazeck, Rescheneder, Smolka, **Fang**, Nattestad, Haeseler, Schatz, "Accurate detection of complex structural variations using single molecule sequencing", *Nature Methods, Accepted* (2018)
- **Fang**, Huang, Radhakrishnan, Siepel, Lyon, Schatz, "Scikit-ribo Enables Accurate Estimation and Robust Modeling of Translation Dynamics at Codon Resolution", *Cell Systems* (2018)
- Vurture, Sedlazeck, Nattestad, Underwood, **Fang**, Gurtowski, Schatz, "GenomeScope: Fast reference-free genome profiling from short reads", *Bioinformatics* (2017)
- **Fang**, Wu, Yoon, Jiménez-Barrón, Mittelman, Robison, Wang, Lyon, "Whole genome sequencing of one complex pedigree illustrates challenges with genomic medicine", *BMC Medical Genomics* (2017)
- **Fang**, Bergmann, Arora, Vacic, Zody, Iossifov, O'Rawe, Wu, Jimenez Barron, Rosenbaum, Ronemus, Lee, Wang, Dikoglu, Jobanputra, Lyon, Wigler, Schatz, Narzisi, "Indel variant analysis of short-read sequencing data with Scalpel", *Nature Protocols* (2016)
- Doerfel, **Fang**, Crain, Klingener, Weiser, Lyon, "Proteomic and genomic characterization of a yeast model for Ogden syndrome", *Yeast* (2016)
- O'Rawe, Wu, Doerfel, Rope, Billie Au, Parboosingh, Moon, Kousi, Kosma, Smith, Tzetis, Schuette, Hufnagel, Prada, Martinez, Orellana, Crain, Caro-Llopis, Oltra, Monfort, Jiménez-Barrón, Swensen, Ellingwood, Smith, **Fang**, Ospina, Stegmann, Den Hollander, Mittelman, Highnam, Robison, Yang, Faivre, Roubertie, Rivière, Monaghan, Wang, Davis, Katsanis, Kalscheuer, Wang, Metcalfe, Kleefstra, Innes, Kitsiou-Tzeli, Rosello, Keegan, Lyon, "TAF1 Variants Are Associated with Dysmorphic Features, Intellectual Disability, and Neurological Manifestations", *American Journal of Human Genetics* (2015)
- Jimenez-Barron, O'Rawe, Wu, Yoon, **Fang**, Iossifov, Lyon, "Genome Wide Variant Analysis of Simplex Autism Families with an Integrative Clinical-Bioinformatics Pipeline". *Molecular Case Studies* (2015)
- Narzisi, O'Rawe, Iossifov, **Fang**, Lee , Wang, Wu, Lyon, Wigler, Schatz, "Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly", *Nature Methods* (2014)
- **Fang**, Wu, Narzisi, O'Rawe, Jimenez Barron, Rosenbaum, Ronemus, Iossifov, Schatz, Lyon, "Reducing INDEL calling errors in whole genome and exome sequencing data", *Genome Medicine* (2014)
- O'Rawe, **Fang**, Rynearson, Robison, Kiruluta, Higgins, Eilbeck, Reese, Lyon, "Integrating precision medicine in the study and clinical treatment of a severely mentally ill person", *PeerJ* (2014)

Dissertation:

- **Fang**, "Graphical and machine learning algorithms for large-scale genomics data" (2017)

CONFERENCE

Platform Talk Presentations:

- *Machine Learning for Facebook Data Warehouse*
  The Data Science Conference, Chigaco, IL 2018
- *Scikit-ribo reveals precise codon-level translational control by dissecting ribosome pausing and codon elongation.*
  Biological Data Science Meeting, Cold Spring Harbor, NY 2016
- *Scikit-ribo reveals precise codon-level translational control by dissecting ribosome pausing and codon elongation.*
  Advances in Genome Biology and Technology(AGBT) Meeting, Orlando, FL 2016
- *Scikit-ribo: Accurate A-site prediction and robust modeling of translation control from Riboseq & RNAseq data.*
  Genome Informatics Meeting, Cold Spring Harbor, NY 2015
- *Reducing INDEL calling errors in whole genome and exome sequencing data.*
  Biological Data Science Meeting, Cold Spring Harbor, NY 2014

First-author Poster Presentations:

- *Scikit-ribo: Accurate estimation and robust modelling of translation dynamics at codon resolution*
  Biology of Genome Meeting, Cold Spring Harbor, NY 2017
- *Scikit-ribo reveals precise codon-level translational control by dissecting ribosome pausing and codon elongation.*
  Advances in Genome Biology and Technology(AGBT) Meeting, Hollywood, FL 2017
- *Scikit-ribo reveals precise codon-level translational control by dissecting ribosome pausing and codon elongation.*
  Genome Informatics Meeting, Cambridge, UK 2016
- *Scikit-ribo reveals precise codon-level translational control by dissecting ribosome pausing and codon elongation.*
  Translational control Meeting, Cold Spring Harbor, NY 2016
- *Scikit-ribo: Accurate A-site prediction and robust modeling of translation control from Riboseq & RNAseq data.*
  Probabilistic Modeling in Genomics Meeting, Cold Spring Harbor, NY 2015
- *Indel variant analysis of short-read sequencing data with Scalpel. (Reviewers' Choice)*
  American Society of Human Genetics Annual Meeting, Baltimore, MD 2015
- *Reducing INDEL calling errors in whole genome and exome sequencing data.*
  Personal Genomes Meeting, Cold Spring Harbor, NY 2014
- *Whole genome analysis of a pedigree with Prader-Willi syndrome, hereditary hemochromatosis, and dysautonomia.*
  Personal Genomes Meeting, Cold Spring Harbor, NY 2014
- *Reducing INDEL calling errors in whole genome and exome sequencing data.*
  American Society of Human Genetics Annual Meeting, San Diego, CA 2014
- *Complexities of INDEL detection based on micro-assembly methods; WGS and WES comparisons.*
  Biology of Genome Meeting, Cold Spring Harbor, NY 2014
- *Whole genome sequencing analysis of a family with familial dysautonomia and neuropsychiatric symptoms.*
  Personal Genomes Meeting, Cold Spring Harbor, NY 2013
- *The statistical properties of longitudinal phenotypes determined by trajectory models in linkage analysis*
  Genetics Analysis Workshop 18, Stevenson, WA 2012

Seminars:

- *Methods for analyzing Riboseq and 10X Genomics data*
  Quantitative Biology Seminar, Cold Spring Harbor, NY 2017
- *Scikit-ribo reveals precise codon-level translational control by dissecting ribosome pausing and codon elongation.*
  Quantitative Biology Seminar, Cold Spring Harbor, NY 2016
- *Reducing INDEL calling errors in whole genome and exome sequencing data.*
  Quantitative Biology Seminar, Cold Spring Harbor, NY 2014
- *Complexities of INDEL detection based on micro-assembly methods; WGS & WES comparisons.*
  CSHL Genome Center Seminar, Cold Spring Harbor, NY 2014

Journal Reviewing

Nucleic Acids Research, RECOMB, WABI, MCBIOS