Han Fang

Westbury – NY 11590

☎ (631) 413-7226 • ✉ hanfang.cshl@gmail.com • 🖰 hanfang.github.io

EDUCATION

Stony Brook University (SBU)   **Stony Brook, NY**

Ph.D in Applied Mathematics & Statistics (AMS)   *2014 - 2017(expected)*
M.S. in Applied Mathematics & Statistics   *2011 - 2013*

Awards:
○ President's Award to Distinguished Doctoral Students
○ 2016 ASHG Reviewers' Choice
○ Scholarships: Summer Institute in Statistics for Big Data Scholarship, CSHL Fellowship, Research Access Project

Sun Yat-sen University (SYSU)   **Guangzhou, China**

B.S. in Optical Informatics   *2007 - 2011*

Scholarships: SYSU 2010 Outstanding Student Scholarship

EXPERIENCE

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 daily pipelines in production to analyze large-scale web/mobile data; identified patterns and provided actionable recommendations.
○ Developed optimization models on global spatial data and solved problems that benefit over 30 countries.

Cold Spring Harbor Laboratory (CSHL)   **Cold Spring Harbor, NY**

Research Assistant   *2014 - Present*

○ Machine learning methods for large-scale genomics data
  - Developed Scikit-ribo, an statistical learning framework in python
    · Improved classification accuracy on millions of sequences by 24% using a RBF kernel SVM classifier with cross validation. Increased speed by 10 fold by using a random forest classifier coupled with feature selection.
    · Implemented a module using negative binomial mixture modeling to identify peaks from over-dispersed data.
    · Built a generalized linear model with ridge penalty using glmnet to accurately infer coefficients of interest.
  - 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 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   *2013 - 2014*

○ Computational methods for next-generation sequencing data analysis
  - Co-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).
  - Developed a novel classification scheme using Chi-Square statistics for benchmarking and error reduction.
  - Created a parallel bioinformatics management workflow with MapReduce for fast and scalable analysis of large-volume data in a high performance cluster environment.

SKILLS

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

PUBLICATIONS

In preparation:

○ **Fang**, Huang, Buskirk, Radhakrishnan, Doerfel, Green, Lyon, Schatz, "Scikit-ribo: Accurate inference and robust modelling of translation dynamics at codon resolution", *In preparation* (2017)

○ **Fang**, Sedlazeck, Ramakrishnan, Luo, Schatz, "Topsorter: graphical assessment of structural variants with 10X genomics data", *In preparation* (2017)

○ Sedlazeck, Rescheneder, Smolka, **Fang**, Nattestad, Haeseler, Schatz, "Accurate detection of Structural Variations using long reads", *In preparation* (2017)

○ 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 highly rearranged cancer cell line", *In preparation* (2017)

Peer-reviewed:

○ Yang, Chen, Lima, **Fang**, Jimenez, Li, Lyon, He, Wang, "PennCNV-Hadoop: Accurate Detection of Copy Number Variation from Whole Genome Sequencing Data", *In revision - Nucleic Acids Research* (2017)

○ Vurture, Sedlazeck, Nattestad, Underwood, **Fang**, Gurtowski, Schatz, "GenomeScope: Fast reference-free genome profiling from short reads", *Accepted - 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)

CONFERENCE ABSTRACTS

Platform Talk Presentations:

○ *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 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