SHAOPENG GU

Email: Shaopenggu1990@gmail.com

Skill Highlights

- 8 years working of programming languages such as Python(*), Spark(*), R, C++, SQL, SAS, Perl
- 7 years of large-scale data mining and analysis with statistical, machine learning, deep learning modeling
- 5 years of mathematical & statistical algorithm development
- 5 years healthcare large-scale data (clinical, EMR, ICD 9, ICD 10, genomic) analysis experience
- 5 years of next-generation sequencing data analysis
- 3 years project leading experience
- 4 years working experience in AWS Clouds (EC2, S3, SageMaker) and Databricks
- 4 years experience with GWAS, PheWAS and their clinical applications
- 4 years experience with disease defincation, outcome prediction, clincinal report generation (PRS)
- Strong written, presentation, oral communication skills
- Capabilities writing production-level code

Experience

Biostatistics Senior Analyst 12/2023 to present
The Ohio State University Comprehensive Cancer Center

Academic Appointment Instructor (Healthcare Data Scientist) 02/2022 to present
Sanford School of Medicine, The University of South Dakota

- Lead the project: Drug-induced liver injury associated with antibiotics use in the NIH *All of Us* Research Program (about 400k patients EMR data)
- Extract useful information from EMR for SSRI discontinuation patients and analyze the genotype-phenotype association

Computational Bioinformatics Analyst 08/2019 to 12/2023
Sanford Health

- Develop machine learning models to predict the outcome of disease (COVID-19, CAD, Breast Cancer etc.) from real patients' genetic and EMR data
- Develop package for calculating polygenic risk score (PRS) for CAD, Breast Cancer, Prostate Cancer, etc. on AWS platform using real patient' genetic data and return a clinical report
- Build Next generation sequencing genotyping pipeline around Databricks' Genomic Runtime DNAseq methods and create a standards and guidelines for validating NGS pipeline
- Develop third generation (Oxford Nanopore) sequencing analytical pipeline and package that integrating all the steps of quality control, barcode demultiplexing, adapter trimming, long reads filtering, mapping, and variant calling

Research Assistant 05/2018 to 08/2019
Bioinformatics and Mathematical Biosciences Lab, South Dakota State University

- Use machine learning methods to analyze and predict the outcomes of EMR data
- Develop a robust, user-friendly feature extraction and selection Python package for DNA, RNA and Protein sequence data that integrating most of popular machine learning methods
- Research and develop RNA-seq pipeline specifically related to the quality control of read alignment process and end-stage interpretations

Selected Lead Projects

Project 1: Quantify the frequencies of DILI (Drug-induced liver injury) for each of the five most commonly used oral antibiotics within NIH All of Us Biobank

All of us is the one of the largest Electronic Health Record (EHR)-linked cohorts in the United States that contains 350,000 unique individuals. DILI frequency was quantified by applying a standardized phenotyping approach to the records of biobank participants who had been exposed to the top five most frequently prescribed oral antibiotics. More than 40 Million clinical measurements (ALT, ALP and TB) and other EMR data were extracted.

Project 2: Designing Oxford nanopore sequencing genotyping and phasing pipeline

CYP2D6 is a small gene located on the long arm of chromosome 22 and is one of the most important genes in pharmacogenetics. The accurate genotyping is hindered by 1) very polymorphic nature of the gene; 2) high homology with its pseudogene CYP2D7; 3) occurrence of structural variations. The long read sequencing (Nanopore) has potential to screen all CYP2D6 variants and to accurate genotyping of complex genes and straightforward variant phasing. Our designed pipeline successfully genotyped 135 variants and phased 125 variants in CYP2D6. The 67 variants are listed in dbSNP151.

Project 3: Developing XGBoost machine learning method with clinically measurable traits and genetic data to predict a Midwest healthcare patients' outcome from COVID-19

Patients' outcomes from COVID-19 infections vary from asymptomatic to poor prognosis, which we define as hospitalization, use of respiratory support and death. We trained XGBoost model using measurements such as common blood tests, age, BMI, sex, etc. and genetic data (6 sentinel sites from Host Genomics Initiative). Our model shows AUC 0.94-0.97 with precision 0.89-0.95.

Project 4: Predicting outcomes of chronic kidney disease from EMR data **using Random Forest Regression**

The progression of kidney disease can be predicted if the future eGFR can be accurately estimated using predictive analytics. We developed and validated a prediction model of eGFR by data extracted from regional health system. We extracted and selected 61,740 patients' demographic, clinical and laboratory information. Our developed Random Forest Regression model achieved an average R^2 of 0.95 over three years and 88% Macro Recall, 96% Marco Precision were obtained by dividing patients into different CKD stages using estimated eGFRs.

Education

Master of Science: Mathematics – Statistics Specialization (2019) *South Dakota State University, SD*
Bachelor of Science: Electrical Engineering (2018) *South Dakota State University, SD*

Selected Publications

Shaopeng Gu, Govarthanan Rajendiran, Kennedy Forest, Tam C. Tran, Joshua C. Denny, Eric A. Larson, Russell A. Wilke. Drug-Induced Liver Injury with Commonly Used Antibiotics in the All of Us Research Program. ***Clinical Pharmacology & Therapeutics.*** doi: 10.1002/cpt.2930

Anjun Ma, Xiaoying Wang, Cankun Wang, Jingxian Li, Tong Xiao, Juexing Wang, Yuzhou Chang, Yang Li, Yutao Liu, **Shaopeng Gu**, Duolin Wang, Yuexu Jiang, Jinpu Li, Li Su, Zihai Li, Dong Xu, Qin Ma. Single-cell biological network inference using a heterogeneous graph transformer***. Nature Communications***. doi: 10.1038/s41467-023-36559-0

Shaopeng Gu, Sydney Lovrien, Mohammad Z Qammar, Russell A Wilke. Rural Land Management and Kidney Health. *S D Med*. PMID: 36898195

Adam McDermaid, Xin Chen, Yiran Zhang, Cankun Wang, **Shaopeng Gu,** Juan Xie, Qin Ma. A new machine learning-based framework for mapping uncertainty analysis in RNA-Seq read alignment and gene expression estimation. ***Frontiers in Genetics.*** doi: 10.3389/fgene.2018.00313

Praveen Cherukuri, Melissa Soe, David Condon, Shubhi Bartaria, Kaitlynn Meis, **Shaopeng Gu**, Frederick Frost, Lindsay Fricke, Krzysztof Lubieniecki, Joanna Lubieniecka, Robert Pyatt, Catherine Hajek, Cornelius Boerkoel, Lynn Carmichael. Establishing analytical validity of BeadChip array genotype data by comparison to whole-genome sequence and standard benchmark datasets***. BMC Medical Genomics***. doi: 10.1186/s12920-022-01199-8

Zhao Jing, **Shaopeng Gu**, Adam McDermaid. Predicting outcomes of chronic kidney disease from EMR data based on Random Forest Regression***. Mathematical Biosciences.*** doi: 10.1016/j.mbs.2019.02.001

Russel Wilke, Mohammad Qamar, Roxana Lupu, **Shaopeng Gu**, Jing Zhao, Chronic Kidney Disease in Agricultural Communities***. The American Journal of Medicine.*** doi: 10.1016/j.amjmed.2019.03.036

David Condon, **Shaopeng Gu**, Murat Sincan, Lynn Carmichael, Tapati Mazumdar, Jerome Rotter, Catherine Hajek. Addition of Patient Genetic Data Is Clinically useful in Prediction of Patient Outcome after COVID-19 Infection**.** (Accepted for publication by ***PLOS ONE***)

Ashish Dubey, Nirmal Adhikari, Swaminathan Venkatesan, **Shaopeng Gu**, Devendra Khatiwada, Qi Wang, Lal Mohammad, Mukesh Kumar, Qiquan Qiao*, Solution processed pristine PDPP3T polymer as hole transport layer for efficient perovskite solar cells with slower degradation. ***Solar Energy Materials and Solar Cells.*** doi: 10.1016/j.solmat.2015.10.008

Nirmal Adhikari, Ashish Dubey, Eman A. Gaml, Bjorn Vaagensmith, Khan Mamun Reza, Sally Adel Abdelsalam Mabrouk, **Shaopeng Gu**, Jiantao Zai*, Xuefeng Qian*, Qiquan Qiao*, Crystallization of Perovskite Film for Higher Performance Solar Cells by Controlling Water Concentration in Methyl Ammonium Iodide Precursor Solution. ***Nanoscale.*** doi: 10.1039/C5NR06687E

Dubey A, Adhikari N, Venkatesan S, **Gu S**, Khatiwada D, Wang Q, Mohammad L, Kumar M, Qiao Q. Shelf life stability comparison in air for solution processed pristine PDPP3T polymer and doped spiro-OMeTAD as hole transport layer for perovskite solar cell. ***Data Brief.*** doi: 10.1016/j.dib.2016.02.021.

Community Service

Manager, Sioux Falls Lion Dance Team 05/2022-12/2023

President, Chinese Student & Scholar Association, South Dakota State University 09/2012-09/2014