Matthew Sooknah

Experienced software engineer & computational scientist working at the intersection of machine learning and biology

179 Noe Street, San Francisco CA 94114
401-588-2644 — mattsooknah@gmail.com
mattsooknah.github.io — github.com/mattsooknah

Experience

Calico Labs **South San Francisco, CA**
Senior Software Engineer, Machine Learning / Computer Vision Team *August 2019 – present*

- Applied modern deep learning techniques to a broad range of microscopy, MRI and CT datasets. Combined state-of-the-art models (e.g. U-Net, Mask-RCNN) with domain-specific post-processing to extract biologically relevant features (e.g. organ volumes, disease-specific phenotypes). Collaborated with scientists to leverage predictions for research and drug development efforts.
- Leveraged previous software and data engineering experience to wrangle messy datasets and ad-hoc scripts into streamlined training and inference pipelines with sensible interfaces, clear documentation, modular code, and well-curated data.
- Worked independently to initiate collaborations with wet-lab scientists, define project goals, run meetings and delegate responsibilities.
- Currently mentoring a summer intern doing research on deep-learning techniques for image reconstruction and denoising.
- Utilized Python, Tensorflow, Keras, scikit-image, OpenCV and ITK for image processing.

Senior Software Engineer, Data Platform Team *August 2017 – July 2019*

- Developed a web application enabling internal scientists to analyze and visualize results of RNA-seq experiments. Implemented genomic analysis pipelines and webapp backend, and worked with UI designer to develop web-based GUI. Worked with scientists to implement a lightweight system for managing sample submissions and integrate it into the application. The application has been in use for 3+ years and used to process thousands of experiments across basic research and drug development programs.
- Contributed to genome assembly and validation efforts for phased WI-38 genome assembly paper. Helped validate novel method for chromosome sorting and phasing. Developed public-facing website (in collaboration with UI designer) and custom IGV-based genome explorer to accompany the paper at wi38.research.calicolabs.com.
- Designed and co-taught a course on Python programming and data analysis for 30 internal wet-lab scientists, with the aim of empowering them to manage and analyze their own data without needing help from computational scientists. Students were able to apply what they learned to real data generated at Calico, greatly increasing scientific productivity across the organization.
- Developed pipelines using Python, R, Google Cloud Platform, Docker, Airflow, and SLURM.

10X Genomics **Pleasanton, CA**
Scientist, Computational Biology Group *January 2016 – August 2017*

- Contributed to development of Cell Ranger and Long Ranger toolkits, which are used by researchers worldwide for processing single-cell RNA-seq and genomic linked reads data generated using the 10x microfluidics platform.
- Wrote and improved analysis tools for transcript quantification, expression-based cell clustering and visualization, and internal QC and product development.
- Profiled and optimized performance of Python and Rust programs for I/O and memory-intensive applications.

The Broad Institute of MIT and Harvard **Cambridge, MA**
Software Engineer, Data Sciences & Data Engineering Group *May 2014 – December 2015*

- Developed and optimized analysis pipelines with Scala and Java for petabyte-scale genomics data from one of the largest DNA sequencing centers in the world.
- Contributed to development and support of the Picard and HTSJDK open source toolkits for processing and analyzing sequencing data.
- Collaborated with engineers and scientists to develop a backend database and data model for FireCloud, a cloud-based platform to help researchers analyze cancer genomics data.
- Developed methods for analyzing gene expression and pathway activity from customized RNA sequencing assays, to gain insight into the mechanisms of autoimmune disease.

Nabsys **Providence, RI**
Associate Scientist, Algorithms Group *June 2013 – May 2014*

- Developed tools in Java for signal processing and genome analysis to support development of the Nabsys microfluidics-based DNA mapping technology.

Education

Massachusetts Institute of Technology **Cambridge, MA**
S.B. Physics, GPA 4.9/5.0 *2009 – 2013*

Programming Languages

- **Expert:** Python
- **Proficient**: R, Java, Scala, bash
- **Familiar**: C, Rust, Scheme, Javascript

Publications

- Ilya Soifer, Nicole Fong, Nelda Yi, Andrea Ireland, Irene Lam, **Matt Sooknah**, et al. Fully Phased Sequence of a Diploid Human Genome Determined de Novo from the DNA of a Single Individual. G3 (Genes, Genomes, Genetics), September 2020. DOI: 10.1534/g3.119.400995
- Daniel O'Connell, Raivo Kolde, **Matt Sooknah**, et al. Simultaneous Pathway Activity Inference and Gene Expression Analysis Using RNA Sequencing. Cell Systems, May 2016. DOI: 10.1016/j.cels.2016.04.011

Presentations

- "Mapping, processing, and duplicate marking with Picard tools." BroadE Workshop on GATK Best Practices. Broad Institute, Cambridge, MA. March 2015.

Poster / Talk Contributions

- Florian Schmid, Georgios Koukos, Yi Liu, **Matt Sooknah** et al. "High-resolution kidney MRI in mice for longitudinal tracking of kidney volume and cyst burden." International Society for Magnetic Resonance in Medicine. Virtual Conference, May 2021. https://index.mirasmart.com/ISMRM2021/PDFfiles/0423.html
- Grace Zheng, Jessica Terry, Paul Ryvkin, **Matt Sooknah**, et al. "Single Cell RNA profiling of a Million Neurons by a Massively Parallel and Scalable Droplet Platform." Advances in Genome Biology and Technology. Hollywood Beach, FL. February 2017.
- Haynes Heaton, Patrick Marks, **Matt Sooknah**, et al. "Alignment and Variant Calling in Segmental Duplications with Linked-Read Data." Genome Informatics. Wellcome Genome Campus, Hinxton, Cambridge, UK. September 2016.