Matthew Sooknah

Computational Biologist & Software Engineer

3215 Folsom Street, San Francisco CA 94110
401-588-2644 — mattsooknah@gmail.com
mattsooknah.github.io — github.com/mattsooknah

Experience

Calico Labs **South San Francisco, CA**
Data Platform Engineer *August 2017 – present*

- Joined a new team responsible for developing a data warehouse and analysis platform, capable of handling a wide variety of data modalities and organisms.
- Interviewed scientists to determine key computing needs, and helped translate these needs into team-level goals.
- Developed data processing pipelines for multiple genomic data modalities using Google Cloud Platform, Docker, and Apache Airflow. Specified and implemented a unified API for running and accessing results of these pipelines. Worked with UI designer to develop custom data visualizations of results.
- Worked with genomics lab and UI designer to develop a lightweight, flexible LIMS system for tracking and processing of sequencing data.

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

- 10X Genomics develops a system that uses molecular partitioning and barcoding to improve traditional DNA and RNA sequencing, along with turn-key software solutions for analyzing and visualizing the resulting data.
- Applied and optimized machine learning methods (such as randomized PCA, graph-based clustering and t-SNE) for profiling and visualization of single cell RNAseq data (such as a groundbreaking dataset of 20k+ gene expression measurements across 1.3 million mouse brain cells).
- Wrote custom tool for associating genome-aligned RNAseq reads with transcripts, significantly reducing runtime and I/O usage compared to previous method.
- Developed new analysis tools and metrics to validate a customized protocol that enables better haplotype phasing from exome sequencing.
- Made improvements to standard DNA sequence analysis algorithms (e.g. short read alignment, haplotype phasing) by utilizing 10X barcoding information.

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

- Developed scalable pipelines to process petabyte-scale DNA sequencing data produced by the Broad Genomics Platform, one of the largest 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.
- Wrote high-performance tool to measure incidence of sequencing errors caused by oxidative damage to DNA during sample preparation.
- 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*

- Nabsys develops a microfluidic system that attaches tag molecules to long (∼100kb) DNA fragments at sequence-specific sites, then runs them through a nanodetector to produce a map of recognition sites.
- Implemented an improved signal processing pipeline for extracting information about molecules and recognition tags from a noisy electronic readout.
- Prototyped algorithms for assembly, validation and visualization of genomic maps and scaffolds based on Nabsys data.

Education

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

Programming Languages

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

Publications

- Daniel O'Connell, Raivo Kolde, **Matt Sooknah**, et al. 2016. Simultaneous Pathway Activity Inference and Gene Expression Analysis Using RNA Sequencing. Cell Systems 2016; 2(5): 323–334. PMID 27211859.

Presentations

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

Poster / Talk Contributions

- 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 (presenter), 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.