Matt Sooknah

– 1552 E Gate Way, Pleasanton, CA 94566 – mattsooknah@gmail.com –

Objective

- I have worked in R&D roles at various labs and companies doing computational biology, focused on applying computational methods to next-generation DNA sequencing data. My desire to pursue a PhD in Computational Biology stems from wanting the freedom to work on more speculative projects, to focus on learning and self-improvement rather than shipping products, and to cultivate my ability to do independent research and then communicate it to others.

Research Interests

- Algorithms for DNA sequence analysis. Methods for visualization of large biological data sets. New sequencing technologies and assays. Applying machine learning and graph / network algorithms to biological problems.

Degrees

- **S.B. Physics, Massachusetts Institute of Technology (2009-2013)**
  - GPA: 4.9/5.0
  - Relevant classes: Computational biology (graduate-level), combinatorial optimization (graduate-level), algorithms, linear algebra, statistical mechanics (including statistical methods)

Research Experience

- **10X Genomics (Jan 2016 - Present)**
  - Scientist, Computational Biology Group
  - Theme: **Algorithms that take advantage of novel barcoding schemes to improve upon short-read and single-cell sequencing**
    - 10X's technology can partition ~100kb DNA molecules (for genome / exome) or single cells (for RNAseq), attach molecular barcodes to the contents of each partition, and then construct a (mostly) standard short-read sequencing library. After sequencing, the barcodes are informatically extracted and used to reconstruct the original molecules / cell expression profiles.
  - Main research contributions:
    - **High-performance clustering + visualization of gene expression data.** Developed more efficient algorithms for dimensionality reduction, clustering and 2D visualization of single cell gene expression data of up to 1 million cells.
    - **Analyzing the efficacy of custom exome bait sets.** Custom exome capture kits were designed to leverage 10X linked-reads to allow phasing of genes across distant exons. I helped define metrics to gauge bait set performance (sequencing coverage, sensitivity / PPV of SNP and structural variant calling, gene phasing, etc), wrote tools to collect them, and subsequently used those metrics to aid in selecting the optimal bait design.
    - **Improving alignments of RNA transcripts.** I prototyped an algorithm that utilizes molecular barcode information to reconstruct RNA molecules from many individual aligned reads, increasing effective read length and accuracy.
    - Developing a method for stitching together haplotype phase blocks in an efficient manner.
- **Broad Institute of MIT and Harvard (May 2014 - Dec 2015)**
  - Software Engineer, Data Sciences & Data Engineering

  - Theme: **Analysis pipelines for novel sequencing assays and population-scale genomics**
    - I was part of two groups - investigating autoimmune diseases using RNAseq (in the Program for Medical and Population Genetics), and developing scalable tools for sequence analysis (in the Genomics Platform / Data Sciences & Data Engineering).
  - Main research contributions:
    - **Gene expression + pathway analysis of RNAseq data.** I worked on pipelines for analyzing gene expression and pathway activity from various RNAseq assays, such as a novel approach to measuring activity of transcription factors using viral insertion of reporter sequences (see Papers section). I used statistical models to link differential expression profiles with experimental conditions. The experiments were focused on probing immune responses in mouse and human cells.
    - **Quantification of sequencing artifacts.** I wrote a tool to measure the incidence of sequencing errors masquerading as low allele fraction SNPs, by testing the significance of factors like read 1 / read 2 bias. I then integrated my tool into our production pipeline, allowing the lab to quickly notice these artifacts and correlate them with sample prep techniques.
    - **Open source bioinformatics.** I contributed to development and support of the Picard and HTSJDK open source projects, which are used by the Broad and across the world to process and analyze sequencing data.
    - **Rapid QC and Cloud Analysis.** I worked on a project to move the bulk of our analysis pipelines from local compute to cloud-based platforms, such as Google Compute Engine. Most of this was software engineering, but with occasional opportunities for R&D, such as a pipeline for rapid alignment and QC of sequencing runs to check for serious errors (e.g. sample swaps, low mapping rates) before proceeding to the main cloud-based analysis.
- **Nabsys (summer 2012 internship, June 2013 - May 2014)**
  - Associate Scientist, Algorithms Group
  - Theme: **Methods for processing and analysis of DNA maps produced by a novel nanochannel-based detector.**
    - The core Nabsys technology allows short recognition sites to be mapped on ~100kb DNA molecules by attaching probes to the recognition sites and running the molecule through a nanoscale electronic sensor. The resulting data are similar to optical maps produced by e.g. BioNano Genomics, but with different properties and sources of error.
  - Main research contributions:
    - **Improving speed and accuracy of signal processing.** I implemented algorithms for noise removal, peak recognition and fitting. I also analyzed how the time-domain electronic signal maps to the length-domain of the molecule for various control samples, to understand the role of various physical processes (electromotive force, viscous drag, brownian motion, etc) and improve our error model.
    - **Algorithms & visualizations for analysis of DNA maps.** I developed an improved algorithm for seeding map alignments based on patterns of interval lengths between adjacent probes. I also contributed to algorithms for de novo assembly of single molecule maps into larger scaffolds. I explored several methods for inferring order + orientation of multiple maps. I developed a prototype tool for multiple alignment and merging of short reads by first piling them up onto scaffolds produced by our assembler. I also prototyped several visualizations of our assembly results, e.g. visual alignment of the scaffolds against ground truth sequence to identify structural errors in the assembly.
- MIT (2009-2013)
  - Main research projects:
    - **Gravitational lensing.** I analyzed how methods for characterizing weak gravitational lensing perform on large data sets of galaxies, such as the Sloan Digital Sky Survey, as well as simulated data. This required image processing and feature extraction, implementing the proposed models, and characterizing sources of error.
    - **Metagenomics.** I worked on a project at Mass General Hospital to characterize bacterial composition of samples from extreme environments (e.g. acid lakes) using a variety of published metagenomics tools.

Papers

- O'Connell DJ, Kolde R, **Sooknah M**, et al. 2016. Simultaneous Pathway Activity Inference and Gene Expression Analysis Using RNA Sequencing. Cell Systems 2, 323–334. http://dx.doi.org/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.