Ryan P Smith

Salt Lake City, UT
linkedin.com/in/rpseq
ryan.smith.p@gmail.com
github.com/rpseq
+1-319-899-0190

Profile

Bioinformatics scientist turned data infrastructure engineer. Strong background with 7 years of academic bioinformatics work and 5 years of professional data engineering at Amazon and Recursion. I like building reliable and scalable data systems on-prem and in the cloud.

Technical Skills

- **Languages**: Python, Bash, R (ggplot), awk, SQL, C++, golang, LaTeX, Mathematica, Java, Scala
- **Technologies**: Linux/Unix, K8s, Slurm, HPC, Helm, AWS, GCP, Grafana, Terraform, Prometheus, Postgres, Docker, Airflow, Spark, HBase, Pinot, Druid, Avro/Parquet, Kafka, GATK, bedtools, bwa-mem

Experience

- **Recursion** Salt Lake City, UT
  Senior Software Engineer *September 2020 – Present*
  - **Inception Labs**: Work with senior leadership to develop novel product ideas and take them through proof-of-concept. Rapidly design and build a large-scale Perturbseq sample processing pipeline using our on-prem HPC cluster.
  - **Transcriptomics Data Processing**: Workflow orchestration. Design and build out automated RNAseq data processing systems in GCP. This includes a metadata tracking service and cloud computing pipelines to process millions of cell samples from an automated laboratory process.
- **eero, an Amazon company** San Francisco, CA
  Software Development Engineer, Data *January 2019 – September 2020*
  - **Data Infrastructure and Operations**: Design, provision, maintain, and scale k8s clusters to power our data systems. Provisioned with Terraform and automated deployments. Spark for custom batch and streaming transformations. Metrics, monitoring and alerting using Prometheus and Grafana. Manage a diverse collection of k8s data services.
  - **Real-time Data Systems**: Build and maintain data systems powering customer-facing features and internal tools. Visual dashboards for observing device bandwidth usage as well as security and content filtering events across a fleet of more than 6 million routers. Ingesting 60K records per second to a large HBase cluster, serving up to 100K customer queries per second.
  - **Data Analytics Platform**: Design and provision AWS infrastructure for a data platform built around Pinot in k8s. Powers sub-second queries over very large datasets with continuous realtime ingestion of fleet data from Kafka, using Superset for data visualization and exploration.
- **Girihlet, Inc., Oakland Genomics Center** Oakland, CA
  Bionformaticist + IT Engineer *Aug 2018 – Oct 2018*
  - **Bioinformatics and IT**: Provide a QC dashboard for Illumina data. Bash, Python, and Perl scripting to automate mitochondrial genome variant calling. Upgrade and maintain the Oakland Genomics Center networks. Maintain two on-premises CentOS servers for data processing and web hosting (Apache, NGINX)
- **The McDonnell Genome Institute, Washington University** St Louis, MO
  Graduate Research Scientist, Ira Hall Lab, Computational Genetics *July 2014 – May 2018*
  - **Distributed Computing, Bioinformatics**: Process and analyze population-level whole-genome sequencing data to investigate structural variation in human genomes. Computation in a large HPC environment. Pipelines often use tools combining Bayesian statistics and machine learning approaches.
  - **Single-cell Sequencing**: Develop novel computational and molecular biology methods for whole-genome sequencing of single mammalian neurons. Process up to 10 TB of raw Illumina sequencing data in 36 hours.
  - **Teaching Assistant**: Advise students sequencing the genomes of unknown bacteriophages. Lab course using bioinformatics tools to identify genes, predict their function, and classify into phylogenetic groups.

Education

- **Washington University** St Louis, MO
  MA, Molecular Genetics and Genomics *Aug. 2014 – May 2018*
- **University of Iowa** Iowa City, IA
  BS, Microbiology and Informatics *Aug. 2009 – May 2014*