Keenan Graham

https://github.com/keenangraham

3165 Porter Drive
Palo Alto, CA 94304
keenansgraham@gmail.com

EXPERIENCE

Department of Genetics, Stanford University School of Medicine — *Principal Investigator: Mike Cherry, Ph.D.*

Software Developer 2 (January 2021 - Present)

- Act as **senior backend software developer** for the NIH-funded *ENCODE (Encyclopedia of DNA Elements) Project* (https://www.encodeproject.org), a follow-up to The Human Genome Project, which hosts over a petabyte of genomics data (over 25,000 experiments with different assay types), and the *IGVF (Impact of Genomic Variation on Function) Project* (https://data.igvf.org).
- Develop and maintain the open-source *encodeD*, *snovault*, *igvfd*, and *igvf-ui* repositories, **using Agile sprints and team software development methodologies, proper Git workflows, unit and integration tests, Jira tickets, and code reviews. (***Snovault* is an open-source metadata database application used in production by groups at Harvard, Broad Institute of Harvard and MIT, Stanford, UC San Diego, and Chan Zuckerberg Initiative.) Improve code quality, reduce complexity, and debug highly technical and subtle issues in application and deployment.
- Establish the data portal for IGVF and **ship autoscaling, container-based production version of submission database and frontend.** Define all infrastructure as code, including continuous deployment pipeline for releasing changes through lower (e.g. demo, dev, staging, sandbox) environments. Make infrastructure and application observable through Slack notifications of events and alarms. Automate stateful migrations (e.g. object upgrades, index creation). Use scheduled serverless functions and workflows to clean up demo infrastructure and share database snapshots automatically. Use **Docker Compose** for local development. Automate continuous integration testing and code linting.
- Deploy **vertical autoscaling Kubernetes cluster** with AWS ALB integration for hosting microservices and running **large data analysis workflows** with Spark.
- Write **high-performance queue message deduplicator in Go**. Use channels and mutexes to safely share data between goroutines.
- Participate in the IGVF consortium's **biological large language model** (LLM) working group.

Software Developer 1 (June 2018 - January 2021)

- Release the 1.5 petabytes of ***ENCODE* data as an *AWS Public Dataset*** (https://registry.opendata.aws/encode-project). Move ~600 TB of data between S3 buckets, write code on AWS Batch to sync state of portal with buckets daily, and glacierize original bucket. Write Jupyter Notebook tutorial demonstrating how to mount the S3 bucket on an EC2 instance

TECHNICAL SKILLS

Python, Go, Typescript, **Rust**, WDL, Bash

Web applications/APIs with **FastAPI**, **NextJS**, Flask, **Pyramid**, React

Git, Emacs, Jira, VScode, Conda

AWS CDK (Cloud Deployment Kit), Pulumi, Terraform

Kubernetes, Fargate, **Docker**

Modern application development and deployment methods using CI/CD, IaC, Docker containers, and Twelve-Factor App principles

Numpy, **Pandas**, Matplotlib, Seaborn, Pyspark, **Pytorch**, **Jupyter** for data analysis and machine learning

Database creation and **querying in Postgres, Elasticsearch**, Redis, TileDB, Neptune, DynamoDB, SQLite, graph/vector databases

Software design principles and patterns

AWS services including EKS, Fargate, RDS, Opensearch, Lambda, DynamoDB, CodeBuild, CodePipeline, SQS, EventBridge

Google Cloud and Azure

INTERESTS

Software architecture and message/event-based integration patterns

Serverless infrastructure

using Goofys and interact with ENCODE data in the cloud using pyBigWig. **Save project over $4 million on storage and egress.**
- Implement the uniform bioinformatics **processing pipeline for ENCODE DNase-seq experiments** in WDL (Workflow Description Language) (https://github.com/encode-dcc/dnase-seq-pipeline) used on all **~1,800 DNase-seq experiments** on the ENCODE portal. Adjust resource parameters for big DNase experiments.
- Participate in the **ENCODE Imputation Challenge**. Research state-of-the-art deep learning techniques and lead group journal discussion. **Train a deep neural network model** to predict over 6,000 tracks of experimental data from different cells, assays, and chromosomes. Write Cython for efficient processing of genomic data.
- Write Rustybot, a **multithreaded chatbot in Rust**, for controlling and monitoring AWS EC2 instances from Slack (https://github.com/keenangraham/rustybot).
- **Refactor backend search APIs using modular architecture and test-driven development**. Add over 5,000 lines of tests and promote component reuse. Improve search relevance using feedback from consortium working groups.
- Add a custom embedding mechanism to speed up indexing by 20% and prevent reindexing unnecessary objects. Stream UUIDs in different order to **better utilize CPUs and avoid memory spikes**.
- Devise and implement frontend (Javascript/React) and backend (Python) code for a **search-as-you-type** feature on ENCODE portal. Add gene search for genome browser. Fix debounce issue. **Improve search relevancy** in Elasticsearch using feedback from ENCODE consortium working groups.
- Implement GA4GH's *RNAget API* and generate RNA expression matrices from JSON. Add newline-delimited JSON generator for streaming. **Use Redis for caching aggregations**. Create a prototype of region search using tileDB backend
- Design and implement **a queue-based microservice to reindex invalidated documents into Opensearch** with zero downtime.
- **Extend JSONSchema** specification with custom keywords. Refactor metadata submission to use the latest JSONSchema version in validator.
- Tag and build ENCODE releases and run **production deployments**.
- Mentor summer software intern, interview software engineer and data wrangler candidates, give presentations on technical topics. **Give a guest lecture** on Jupyter notebooks and *pandas* to the *Introduction to Python for Genomics* class at Stanford.

Research Data Analyst (May 2017 - June 2018)

- Design and implement *QANCODE*, an **automated image-based comparison tool for regression testing in the browser**.
- Modify genomic analysis pipeline accessioning code to support a wider variety of assay types. Introduce documentation tools to pipeline group.
- Develop a general quality metric reporting tool to perform complex data scraping and manipulation and create reproducibly formatted Google Sheets for use by consortium working groups.
- Guide new group members in setting up a development environment.
- Analyze biosample, experiment, file and quality metric metadata.
- Present tutorial on JSON objects and REST API requests to the portal.

Machine/deep learning

Data visualization/WebGL

Bioinformatics

Bayesian analysis/statistical modeling

Department of Microbiology and Parasitology, Genova Diagnostics — *Laboratory Director: James Kelton, Ph.D.*

2015 - April 2017

- Write scripts to extract data from the laboratory information system.
- Calculate statistics and create visualizations from patient data with *pandas* and Jupyter notebooks.
- Use Bayesian analysis to answer epidemiological questions.
- Build a parasitology monitoring tool and interactive visualization in D3.js, JavaScript, HTML, and CSS.

2013 - April 2017

- Identify intestinal protozoa and parasitic helminths/ova in patient samples submitted for clinical diagnostics.

Department of Clinical Genetics, Fullerton Genetics Center, Mission Hospital — *Laboratory Director: Jack Tarleton, Ph.D., FACMG*

2013 **(Volunteer Project)**

- Identify novel mutations in the *CLCN1* gene leading to myotonia congenita.
- Correct inconsistencies between lookup table of point mutations and raw DNA exon sequences of the gene.

Department of Epidemiology, UNC Chapel Hill Gillings School of Global Public Health — *Principal Investigator: Carla Cerami Hand, M.D./Ph.D.*

2010 - 2011

- Determine parasitemia in blood samples infected with the malaria parasite *Plasmodium falciparum*.
- Isolate the non-erythropoietic, tissue-protective erythropoietin heteroreceptor.

EDUCATION

University of North Carolina at Chapel Hill, Chapel Hill, NC — *Bachelor of Science in Biology*

2011

Coursework in the Biology of Blood Diseases, Molecular Biology, Genetics, Organic Chemistry, Biochemistry, Modern Physics, Ordinary Differential Equations, Linear Algebra, Advanced Mathematical Methods.

RECENT PROJECTS

Semantic search of ENCODE metadata — *Using OpenAI embeddings from the text-embedding-ada-002 model.*

This includes a general Python library that calculates and stores embeddings for JSON documents, queries and ranks embeddings with cosine similarity, and summarizes the relevancy of results given a query using *gpt-3.5-turbo*. The backend

is FastAPI, the frontend is NextJS. It runs locally with Docker Compose and is deployed in the cloud on AWS ECS Fargate using AWS CDK.

More at https://semantic-search.demo.igvf.org