Beri Shifaw

Cloud Bioinformatics Engineer | Genomic Workflows & Infrastructure
Boston, Massachusetts 02108 • bshif28@gmail.com • 617-903-8109 • www.linkedin.com/in/beri-shifaw

Computational engineer with several years of experience developing tools and workflows in genomics research. Additionally, supported scientific computing across cloud and HPC environments. Skilled in building, optimizing, and deploying NGS workflows using cloud engineering, containerization, and CI/CD practices. Experienced in supporting bench scientists and integrating machine learning methods for genome assembly. Passionate about using data engineering to advance genomic research and human health.

SKILLS AND TECHNICAL EXPERTISE

- Analysis & Bioinformatics: Jupyter Notebooks, GATK, long-read genomics tools, IGV
- Software Development: Python, Github, Github Actions, Pytest, CI/CD, WDL, Git
- Computing Infrastructure: Google Cloud (GCP), Docker, HPC
- Data Management: BigQuery, PostgreSQL, Pub/Sub
- Machine Learning: PyTorch, Neural Networks, Pandas, XGBoost

PROFESSIONAL EXPERIENCE

Broad Institute of MIT and Harvard**,** Cambridge, Massachusetts Sep 2017 - Jul 2025

Senior Software Engineer [Bioinformatics] (Jun 2022- Jul 2025)

Developed and deployed scalable genomic workflows in the cloud, automated testing and monitoring systems, and built tools supporting deep learning, QC analysis, and scientific decision-making.

- Long-Read Genomics Pipeline Automation & Deployment to *All of Us* and *Gabriella Miller Kids First* [1]
  - Maintained end-to-end pipelines used for data processing of ~15,000 long-read WGS samples.
  - Built, optimized, and deployed Docker containers for use in cloud-based workflows, ensuring seamless integration and scalability.
  - Designed a system for automated testing of Github-hosted WDL workflows, ensuring reliability.
  - Developed a framework to track and visualize resource usage of cloud workflows by streaming performance metrics to BigQuery for real-time monitoring and post-run analysis, leading to a 40% cost reduction in workflow execution enabling additional downstream scientific investigations.[2]
- Explore Deep Learning Model Techniques for Read Error Classification[3]
  - Built a framework to filter out erroneous long-read kmers with short-read WGS data, helping improve haplotype-resolved assembly in highly diverse genomic regions (e.g. MHC/HLA, CYP2D6).
  - Achieved a 10% boost in classification accuracy by replacing XGBoost with a neural network model.
- Sequencing Quality Control & Data Visualization
  - Developed scripts for long-read (PacBio/ONT) sequencing QC analysis; generating key metrics (read length, quality, coverage) and custom visualizations in Python to assess run quality.
  - Assessed new wet lab procedures by creating coverage and quality visualizations that directly informed the lab's project planning.
  - Presented findings to stakeholders to guide data reuse and re-sequencing decisions.

Lead Support Engineer (Jun 2021 - Jun 2022)
Guided bench scientists' transition from on-prem scripts to cloud-native pipelines written in WDL with containerized tools. Troubleshooting escalated cloud infrastructure issues, including stalled virtual machines, billing discrepancies, and user account management.

- Cromshell: Porting and Enhancing Workflow Execution Tool[4]
  - Rebuilt Cromshell by migrating from a monolithic bash script to a modular Python codebase, enhancing its robustness, stability, and maintainability.
  - Designed an extensible architecture to support rapid development of new commands and features.
  - Established continuous integration using PyTest and GitHub Actions to enforce code quality and catch regressions early.
  - Deployed the tool to production on the All of Us Researcher Workbench.

Senior Product Support Specialist (Sep 2017 - Jun 2021)
Supported the Terra platform by leading initiatives to improve user workflows, automating internal processes, and delivering data-driven insights to guide product development and user enablement.

- Designed and implemented an automation system connecting Smartsheet and Jira, improving the organization and consistency of managing and updating bioinformatics workspaces and tools on Terra.
- Created and analyzed Mixpanel dashboards visualizing user behavior and workspace utilization, delivering actionable insights used by support teams for resource allocation and by the product team to inform business decisions and development roadmaps.
- Led tutorials on GATK, WDL, and Terra, and provided user support via forums and office hours; accelerating onboarding and productivity for new users.

Intel Corporation, Hudson, Massachusetts — Sep 2015 - Oct 2017
Life Science Architect

- Secure HPC Infrastructure Deployment for Clinical Research Collaboration (CCC)
  - Built and configured secure server clusters, enabling data sharing between hospitals and research institutions. Installed core infrastructure components including linux operating systems and account management systems, and deployed systems at partner sites such as Dana-Farber and the Ontario Institute for Cancer Research.

EDUCATION

University of Maryland University College — College Park, Maryland
Master of Science, Major in Biotechnology - Specialization Bioinformatics — Sep 2011 - Jun 2015

University of Maryland — College Park, Maryland
Bachelor of Science, Major in Biological Science — Sep 2006 - Jun 2010

GIT REPOSITORIES

- GitHub Profile: github.com/bshifaw
- Project 1: Long-Read Pipelines for AoU - https://github.com/broadinstitute/long-read-pipelines
- Project 2: WDL Resource Monitor - https://github.com/broadinstitute/cromwell-task-monitor-bq-vis
- Project 3: K-mer Error Classification - https://github.com/broadinstitute/hidive
- Project 4: Cromshell - https://github.com/broadinstitute/cromshell