Ryan Pontius

Hanover, New Hampshire | ryanpontius@gmail.com | LinkedIn | Github

Education

Dartmouth College (Geisel School of Medicine), Hanover, NH **June 2026**
Master of Science, Health Data Science

- Key Courses: Python for Data Scientists, Epidemiology, Data Wrangling, Biostatistics I & II, Applied Biostatistics, Machine Learning, Bioinformatics, Advanced Biomedical Data Science

University of California, Santa Cruz, Santa Cruz, CA **Sep 2019 - June 2023**
Bachelor of Science, Major in Molecular, Cell, and Developmental Biology **GPA 3.70/4.0**

Work Experience

Augmented Health Lab, **Emory University,** Atlanta, GA **October 2025-Present**
Research Data Engineer
Develop infrastructure for Glucose-ML, an open-source CGM dataset collection, by engineering data pipelines and preprocessing tools to standardize large-scale health data for machine learning applications. (Advisor: Dr. Temiloluwa Prioleau)

- **Design and implement scalable Python and Bash pipelines** to build the Glucose-ML repository from the ground up, ingesting, cleaning, and standardizing 49M+ continuous glucose monitoring (CGM) records.
- **Lead development of statistical and analytical tools** to compute clinically relevant glucose metrics and generate visualizations to support the interactive web-interface and feature engineering for downstream machine learning workflows.
- **Develop and optimize data processing workflows** to streamline data retrieval, cleaning, and standardization for internal teams and public users, improving efficiency and accessibility of large-scale health datasets.
- **Built and evaluated machine learning models** to classify diabetes status using the Glucose-ML dataset collection to demonstrate the platform's ability to support scalable, real-world machine learning workflows; **First author** on the peer-reviewed manuscript introducing Glucose-ML as a large-scale, evolving CGM dataset collection for advancing machine learning and AI research in diabetes.

California Conservation Genomics Project, UCSC & UCLA, Santa Cruz & Los Angeles, CA **June 2023-Present**
Data Wrangler/Junior Specialist I (6/23-6/24)*, II* (6/24-6/25) → *Research Associate II (6/25-present)*
Support the California Conservation Genomics Project's (CCGP) mission to produce the most comprehensive multispecies genomic dataset ever assembled by developing bioinformatic pipelines, managing genomic data, and collaborating across institutions to support downstream population genomic analysis. (Advisors: Dr. Brad Shaffer (UCLA), Dr. Erik Enbody (Cornell), Dr. Russell Corbett-Detig (UCSC)).

- **Bioinformatics Engineering:** Engineer and deploy Snakemake workflows, using Python/R/Bash, to automate large-scale bioinformatic tasks on high-performance computing (HPC) systems using Slurm-based job scheduling, including running the CCGP Variant Calling pipeline across 150+ population-level whole genome sequencing (WGS) datasets.
- **Manage intake, validation, and quality control of large-scale** WGS datasets and associated sample metadata using Python-based workflows, ensuring data integrity and consistency across tens of thousands of samples to support downstream statistical and genomic analyses.
- **Database Management:** Oversee and maintain a MongoDB database of 20,000+ samples and manage the storage and organization of 140,000+ FASTQ files across AWS S3 and Google Cloud infrastructure, supporting efficient querying and large-scale data access.
- **Project Management:** Coordinate submission of genomic datasets to NCBI repositories (SRA, BioProject, BioSample) and deliver processed genomic datasets and analytical reports to 70+ consortium collaborators, providing ongoing technical and research consultation as needed.

Key Coding Projects

Diabetes Classification Workflows (Glucose-ML Case Studies) **April 2026**

- Designed and implemented two machine learning case studies demonstrating practical applications of the Glucose-ML collection:
  - **Case Study 1**: Built classification workflows using logistic regression, random forest, and XGBoost to predict diabetes status across 13 open-source CGM datasets from the Glucose-ML collection.
  - **Case Study 2**: Conducted comparative analysis evaluating model classification performance when trained on a single dataset versus multiple datasets, highlighting the benefits of multi-dataset training.

Glucose-ML Public Repository **October 2025 - Present**

- Develop Python-based tools to standardize continuous glucose monitoring (CGM) data and structure metadata for integration into the Glucose-ML dataset collection and public-facing platform. Engineer scalable data pipelines to support downloading

and preprocessing of 20+ datasets, enabling users to access and apply standardized CGM data for machine learning workflows.

GONE Workflow - Effective Population Size Analysis **March 2024**

- Developed a Snakemake workflow that estimates recent effective population sizes (~100–200 generations in the past) at the per-sample level across CCGP species projects. The workflow generates visuals and analytical outputs that supports the CCGP Landscape Genomics teams downstream population genomics research.

Stairway Workflow **December 2023**

- Developed a bioinformatics pipeline leveraging PAV and MSMC2 to analyze phased genome haplotypes, detect structural and sequence variation, and estimate effective population size changes over time. The pipeline models demographic history from genetic variation and generates analytical reports on heterozygosity and mutation patterns.

Publications

[1] **Ryan Pontius**, Worayada Pitakanonda, Zimo Li, Kultum Lhabaik, Fengran Wang, Baiying Lu, Yanjun Cui, and Temiloluwa Prioleau. 2026. Glucose-ML: An evolving collection of continuous glucose datasets to accelerate data-centric AI for diabetes. Submitted to *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2026). *Under review*

[2] Baiying Lu, Zhaohui Liang, **Ryan Pontius**, Shengpu Tang, Temiloluwa Prioleau. GlucoFM-Bench: Benchmarking Time-Series Foundation Models for Blood Glucose Forecasting. 2026. **In preparation*

[3] Enbody, E.D., A. Nakamoto, C. Mirchandani, M. Baylis, A. Chambers, C. Miller, **R. Pontius**, E. Toffelmier, CCGP Consortium, B. Shaffer, R. Corbett-Detig. Regional conservation genomics highlights factors affecting population health. **In preparation.*

Technical Skills & Certifications

Programming Languages: Python, Bash, R, SQL
Workflow Languages: Snakemake
Packages/Libraries: pandas, pymongo, pygsheets, numpy
Bioinformatics: Whole Genome Sequencing, WGS dataset analysis, variant calling, SAMtools, BCFtools, Illumina short-read data, plink, MSMC, Next Generation Sequencing (NGS)
Computing: AWS S3, Google Cloud, GitHub, High-Performance Computing (HPC), Excel, Tableau, Prompt Engineering, AI-assisted data analysis and development workflows
Databases: MongoDB, NIH NCBI Sequence Read Archive (SRA), BioSample, & BioProject.
Certifications: Collaborative Institutional Training Initiative (CITI) Program - Human Research - (1) Group 1: Biomedical Research, (2) Group 2: Social/Behavioral Research, (3) Conflict of Interest

Leadership

Geisel School of Medicine at Dartmouth College, Hanover, NH **October 2025-Present**

MS Class Representative - Treasurer

- Elected to represent the MS cohort and advocate for student priorities in Student Government.
- Manage student activity budget and coordinate events across Dartmouth/Geisel organizations.