Paul Sample, PhD

pjsample@gmail.com
linkedin.com/in/paul-sample-211404149
github.com/pjsample

Computational biologist with experience in **language models** applied to **biological systems,** machine learning, and **experiment design** for generating genome-scale datasets and synthetic biology applications

Skills

- **Machine learning:** DNA and protein language models (GPT, BERT, XLNet), computer vision (CNNs, ViTs), Bayesian statistics
- **Data Experience:** Next-generation sequencing data (NGS), ATAC-seq, RNA-seq, CUT&RUN, CHiP-seq
- **Cloud Computing:** AWS, DataBricks, GCP
- **Assay Development:** High-throughput assays to generate data for modeling translation regulation and transcription regulation, single-cell sequencing optimization (SPLiT-Seq)
- **Domain knowledge:** Background in biology and wet lab experience
- **Programming / Libraries:** Python, Pytorch, ML and stats libraries

Work Experience

Senior Data Scientist – Outpace Bio 2021 – present

- **Pretrained a GPT DNA model for generating enhancers and promoters for specific activity in T cells.** GPT model that incorporates numeral embeddings and conditional tokens to guide DNA sequence generation for specific regulatory activity.

- **Developed an assay to measure ~1 million enhancer-promoter pairs in T cells.** Cell type-specific transcription regulation depends on the interaction of enhancers with promoters. I designed a massively parallel reporter assay (MPRA) to collect the data needed for model training.

- **Developed a protein language model (pLM) to create monobody binders.** Used an XLNet and prefix language modeling to infill binding loops of monobodies (similar to antibodies) to make high affinity protein-protein binders.

- **Fine-tuned SWIN transformer to interpret live-cell microscopy images.** Designed an experiment to generate data that I then used to fine-tune a vision model for improved accuracy of cell behavior metrics.

Senior Bioinformatics Scientist – Guardant Health 2019 – 2021

- **Worked on the LUNAR-2 product for early detection of cancer.** Used machine learning with epigenetic data to build predictive models for early colorectal cancer detection.

- **Built statistical models and used data visualization** to better understand the complexity of biological signals and to address technical confounders.

Senior Scientific Programmer – Bellwether Bio 2018 – 2019

- **Developed predictive models for cancer diagnosis via genomic analysis.** Improved model to accurately predict multiple cancer types.

- **Feature engineering from terabytes of high-dimension data.** Used Google BigQuery to identify predictive features from the human genome.

Research Experience

Postdoctoral Researcher – University of Washington 2014 – 2018
Advisor: Dr. Georg Seelig

- **Trained a CNN to predict protein expression from 5' UTR sequence and design sequences *de novo*.** A CNN was used in combination with a genetic algorithm to engineer new 5' UTRs for maximal protein expression and to accurately predict the effect of human 5' UTR variants for disease prediction.
- **Optimized single-cell sequencing technology (SPLiT-Seq).** Worked with a team to successfully increase the number of cells and RNA sequence reads per cell. Over 500 experiment variables were tested in a single high-throughput experiment to find optimal protocol conditions.
- **Developed a massively parallel assay to measure translation of millions of 5' UTR sequence variants in human cells in a single experiment.** Combined random 5' UTR library construction and polysome profiling of library mRNA to generate large datasets applicable for machine learning.
- **Mentored graduate students in data science and molecular biology**.

Ph. D. – The Ohio State University 2009 – 2014
Advisor: Dr. Juan D. Alfonzo

- **Designed, led a collaborative team, and programmed 'RoboOligo'**, a software application that interprets tandem mass spectrometry data of complex, modified nucleotide-containing RNA oligomers.
- **Identified the hypermodified ribonucleosides hydroxywybutosine and wyosine in the trypanosome mitochondrion using HPLC-MS/MS.** This was the first observation of any wybutosine-derivative within an organelle.

Education

Ph. D. Microbiology. The Ohio State University 2014
B.S. Microbiology. The Ohio State University 2009

Select Publications

Sample PJ, Wang B, Reid DW, Morris DR, Seelig G. **Human 5′ UTR design and variant effect prediction from a massively parallel translation assay.** *Nature Biotechnology* (2019)

Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample PJ, Mukherjee S, Chen W, Peeler DJ, Yao Z, Gray L, Nguyen T, Tasic B, Sellers DL, Pun SH, Seelig G. **Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding.** *Science* 360.6385 (2018): 176-182

Sample PJ, Gaston KW, Alfonzo JD, Limbach PA. **RoboOligo: software for mass spectrometry data to support manual and *de novo* sequencing of post-transcriptionally modified ribonucleic acids.** *Nucleic Acids Research* 43.10 (2015): e64-e64