Diego Taquiri-Diaz

diego.taquiri@upch.pe • diego-taquiri • diego_taquiri

Education

2020–2024 **B.Sc in Biology**, *Universidad Peruana Cayetano Heredia (#1 program in Peru)*
GPA: 4.0/4.0, 16.3/20.0 in original scale (Top 20% of the class).

2017–2019 **B.S.E in Biomedical Engineering**, *Pontificia Universidad Católica del Perú*, Peru
GPA: 3.9/4.0, 15.5/20.0 in original scale (Top 10% of the class). Transitioned to a B.S. in Biology after 3 years of coursework.

Skills

Programming: R/RStudio, Python, Nextflow, AWS Cloud, GitHub, Bash, Unix.

Machine Learning: PyTorch, Transformers, Large Language Models (LLMs), gLMs (genomic), pLMs (protein), CNN, XGBoost, NumPy, pandas, Polars, Matplotlib, Seaborn.

Toolbox: Visual Studio Code, Cursor AI, Jupyter, SSH, Tidyverse, Docker, LaTeX, ChatGPT.

Bioinformatics: scRNAseq, RNAseq, Genomics, Metagenomics, Phylogenetics, Methylomics, NGS, Biostatistics, Molecular docking, 3D Structure Prediction.

Languages: Spanish (native speaker), English (Full professional proficiency).

Research Experience

2023–Present **Research Assistant**, *Grandjean Research Group*, University College London, UK
Principal investigator: MD, Ph.D. Louis Grandjean

- **Data integration:** Performed data normalization, transformation, batch effect correction, and dimensionality reduction to characterize the profile of pathogens and antibiotic resistance in the wastewater microbiome from multiple hospitals and temporal points.
- **Cis-regulatory element prediction:** Predicted cis-regulatory elements (promoters) of antibiotic resistance genes in bacterial strains and compared mutations across strains to explain differences in phenotypic resistance, despite identical genomic resistance profiles.
- **Metagenomic assembly:** Conducted metagenomic assembly of wastewater samples using deep sequencing with PromethION Nanopore technology to generate high-quality MAGs and achieve reproducible detection of bacterial pathogens and antibiotic resistance.
- **Phylogenetics analysis:** Annotated gene markers from metagenome-assembled genomes (MAGs), performed multiple sequence alignment, constructed phylogenetic trees, and conducted species identification. Validated candidate novel species using contamination and completeness metrics for quality control.
- **Methylomic analysis:** Performed basecalling and alignment using Nanopore sequencing modification models to identify 6mA, 5mC, and 5hmC methylation patterns in *Klebsiella pneumoniae* isolates, analyzing methylation profiles in resistance genes to explore their potential role in antibiotic resistance.

2023–Present **Research Assistant**, *Bioinformatics and Molecular Biology Lab*, Universidad Peruana Cayetano Heredia, Peru
Principal investigator: Ph.D. Mirko Zimic and Ph.D. Patricia Sheen

- **RNA-seq analysis:** Performed RNA sequencing analysis of platelets from patients with and without tuberculosis, including splice-aware mapping, unsupervised clustering, differential gene expression, gene ontology, gene set enrichment analysis, and supervised machine learning to predict TB diagnosis based on RNA profiles.

- **Annotation:** Conducted functional annotation to identify genes, metabolic pathways, plasmids, and viral elements associated with microbiome changes in metagenomic agricultural soil samples treated with organic fertilizers.
- **Differential abundance analysis:** Conducted statistical analysis to identify differentially abundant resistant genes and species associated with caries compared to healthy controls in the oral microbiomes of children, including taxonomic classification and multivariate analysis (PCoA) to effectively cluster microbiome profiles.
- **Variant calling:** Conducted reference-based mapping and variant calling on 3,932 *M. tuberculosis* whole genome sequences, followed by heteroresistance analysis to detect low-frequency mutations in bacterial subpopulations across 14 antibiotic resistance genes.
- **Molecular docking and structural modeling:** Conducted structural characterization of the PonA1 protein of *M. tuberculosis* using AlphaFold2/ESMFold to predict the full protein structure, followed by deep learning-based (Diffdock) and classical docking (Gnina) methods to identify rifampicin-binding pockets.
- **Whole-genome analysis:** Developed an end-to-end Nextflow pipeline to process 187 *M. tuberculosis* genomes sequenced in-house using Nanopore technology, generating post-sequencing metrics such as quality and mapping statistics, and conducted statistical comparisons of extraction methods using R.
- **TB detection and resistance profiling:** Developed and evaluated a nanopore sequencing protocol for direct detection of *M. tuberculosis* and antibiotic resistance from sputum samples, utilizing shotgun metagenomics, amplicon sequencing, adaptive sampling, and culture-based whole genome sequencing (WGS). Conducted read filtering, gene depth analysis, resistance profiling across 13 genes, compared enrichment methods, and performed TB lineage profiling.
- **Proteomic analysis:** Developed and validated an assay using MALDI-TOF mass spectrometry to detect pyrazinamide resistance in *M. tuberculosis* by analyzing the conversion of pyrazinamide to pyrazinoic acid in MODS culture.
- **Miscellaneuos:**
  - **Established genomics operations:** Built the lab's bioinformatics capacity from the ground up after acquiring a nanopore sequencing device, managing everything from experimental design and sequencing monitoring to data processing, analysis, visualization, and interpretation.
  - **Computational Infrastructure Management:** Managed and optimized the bioinformatics computational infrastructure for a genomics lab, including high-performance computing workstations, large-scale storage systems, and data pipelines.
  - **Team leadership and mentoring:** Led and mentored a seven-person bioinformatics team, managing multiple projects concurrently.

GitHub Repositories

2024 **nanoGPT-DNA**, *Developed a lightweight GPT-like large language model (LLM) transformer in PyTorch to learn the regulatory language of DNA sequences. Adapted nanoGPT to train on the human genome with nucleotide-level tokenization, focusing on regulatory motif classification evaluation as a genomic language model (gLM) (Work in progress).*

2024 **HLCA-scAtlas-Exploration**, *Implemented the integration of 13 single-cell RNA sequencing (scRNAseq) datasets from the Human Lung Cell Atlas (HLCA), encompassing 600,000 cells. Focused on evaluating the effects of normalization, batch-effect correction, and clustering using Scanpy, scVI-tools and Polars.*

2024 **EGFR-pLM-Pathogenicity**, *Analyzed EGFR mutations and their impact on therapeutic antibody resistance using structural modeling and protein language models (pLM). Integrated patient-specific EGFR mutations with ESM1b predictions to explore variant pathogenicity and their roles in therapeutic evasion.*

2024 **AWS-scRNAseq-Nextflow**, *Deployed and ran Nextflow-based `nf-core` pipelines for single-cell RNA sequencing on AWS using AWS Batch and Tower. Automated quality control, integration, clustering, and cell type annotation.*

2024 **HeartDiseaseMLInterpretation**, *Trained a machine learning model using XGBoost with hyperparameter tuning and implemented SHAP for interpretability, highlighting feature importance in heart disease prediction.*

2024 **AutismSketchClassifier**, *Pre-trained a ResNet CNN neural network and used its feature vectors for KNN classification to detect autism-specific features in children's sketches.*

2024 **ECGMortalityPredictor**, *Processed ECG, EEG, and EMG signals, applying wavelet transforms for denoising and feature extraction. Developed a TinyML-based early warning system using XGBoost to predict cardiovascular risk in Chagas patients.*

2024 **SimpleGenomicNextflow**, *Developed and maintained a suite of user-friendly and flexible Nextflow scripts for genomic and metagenomic analysis.*

2023 **ONT-tb-extraction**, *Developed R scripts for statistical analysis and plotting, along with a Nextflow pipeline for comparative analysis of Nanopore sequencing of TB.*

Publications

Under review

K. Vallejos-Sánchez, **D.A. Taquiri-Díaz**, O. Romero, A.P. Vargas, J. Coronel, A. Torres, J.L. Perez, A. Ochoa, R.H. Gilman, L. Grandjean, M. Cohen-Gonsaud, M. Zimic, P. Sheen. "Identifying heteroresistant tuberculosis infection from whole genome analysis of Peruvian isolates". **Contribution: I designed and performed computational analyses.**

To be submitted in early 2025

C. León, A. Osmaston, **D.A. Taquiri-Díaz**, O. Romero, J. Perez, B. Sobkowiak, J. Hatcher, A. Torres, R. Gilman, S. Huaman, J. Coronel, M. Zimic, P. Sheen, L. Grandjean. "A comparison of methods for *Mycobacterium tuberculosis* DNA extraction optimised for long-read Nanopore sequencing." **Contribution: I designed and performed computational analyses.**

G. Lawson, G. Tan, C. Leon Palomino, A. Osmaston, O. Romero, **D.A. Taquiri-Díaz**, L. Mascaro Rivera, L. Merino Castaneda, I. Baltas, A. Torres Ortiz, B. Sobkowiak, A. Mendoza Ticona, R. Gavilan, L. Alvarado Ruis, D. Gómez de la Torre, R.H. Gilman, J. Hatcher, P. Sheen Cortavarria, M. Zimic Peralta, M. Pajuelo Travezaño, L. Grandjean. "Identifying Novel Mechanisms of Carbapenem Resistant Enterobacterales in Lima, Peru". **Contribution: I designed and performed computational analyses, and contributed to manuscript writing.**

Relevant Coursework

2024 **Natural language processing**, *Undergradutate Course (+60 hours)*, Universidad Peruana Cayetano Heredia, Peru.

2024 **Computer Vision**, *Undergradutate Course (+60 hours)*, Universidad Peruana Cayetano Heredia, Peru.

2024 **Population genomics**, *International Workshop (18 hours)*, Universidad San Martín de Porres, Peru., In collaboration with the Barreiro Lab, University of Chicago.

2023 **Bioinformatics and Artificial Intelligence**, *International Training (20 hours)*, Peruvian Society of Bioinformatics and Computational Biology (SPBBC).

2023 **Bioinformatics I**, *Undergradutate Course (+70 hours)*, Universidad Peruana Cayetano Heredia, Peru.

2023 **Introduction to Machine Learning**, *Undergradutate Course (+50 hours)*, Universidad Peruana Cayetano Heredia, Peru.

2022 **Neuromatch Academy: Deep Learning**, *International Summer School (+60 hours)*

Leadership Experience

2022 **Directive Board Member**, *Journal Club*, Student Club, UPCH

Directed communications and club leader recruitment, successfully establishing 20 specialized journal clubs and guiding over 100 new participants.

2022 **Research Secretary**, *Student Center for Sciences CEC*, Student council, UPCH

Orchestrated the faculty-wide university Science Week, managing over $3,000 in funding and achieving an engagement of 1,000+ attendees. Additionally, organized a series of science webinars and career guidance sessions, featuring insights from invited speakers.

2019 **Vice-President**, *IEEE Student Branch UPCH*, UPCH

Coordinated multiple university, inter-university, and national congresses, meetings and events, engaging over 100 participants per event.

2019 **Founding Leader**, *Biomedical Engineering Association*, Student council, PUCP

Led the foundational efforts to establish the Biomedical Engineering Association, managed a core group of 10 members in the structuring and drafting of the association's statutes.

Grant Writing

2023 **Research grant 82878, $100,000**, *National Council for Science, Technology, and Technological Innovation (CONCYTEC)*, Peru

Conceptualized and authored the bioinformatics section of the grant proposal: "Development and evaluation of a MinION (Nanopore) sequencing-based protocol for determining Heteroresistance in tuberculosis patients directly from sputum samples".

Posters & Presentations

2024 **HeLa Project, Data Science Division, Universidad Peruana Cayetano Heredia**, *Invited Seminar*, Single-cell RNA-seq 101: Fundamentals of Bioinformatics Analysis of Cellular Transcriptomics. **D.A. Taquiri-Díaz**.

2024 **American Society of Tropical Medicine and Hygiene (ASTMH)**, *Accepted Poster: Identifying Novel Mechanisms of Carbapenem Resistant Enterobacterales in Lima, Peru*, New Orleans, LA, USA (Accepted, not presented due to funding constraints). C. Tan, **D.A. Taquiri-Díaz**, O. Romero.

2024 **V International Congress of the Peruvian Society of Bioinformatics and Computational Biology:**, *Oral Presentation (online): Exploring the Computational Resources of the Bioinformatics and Molecular Biology Laboratory of UPCH and its Applications in Research*, Sociedad Peruana de Bioinformática y Biología Computacional, Lima, Peru. **D.A. Taquiri-Díaz**.

2019 **International Conference on Electronics, Electrical Engineering and Computing (XXVI INTERCON)**, *Tech Fair Stand: Customized Glove for De Quervain's Tenosynovitis Prevention*, Universidad Autonoma del Peru, Peru. **D.A. Taquiri-Díaz**, A. Tecse.

Honors & Awards

2019 **International Conference on Electronics, Electrical Engineering and Computing (XXVI INTERCON)**, *Best Applied Technological Development*, Universidad Autonoma del Peru, Peru

2019 **Institute of Electrical and Electronics Engineers (IEEE)**, *Best new student branch*, Universidad Peruana Cayetano Heredia, Peru.

Teaching Experience

2023 **Teaching Assistant**, *Course: Bioinformatics I: Sequence Analysis*, Master's Program, Universidad Peruana Cayetano Heredia, Peru
Led practical workshops (12 hours) for approximately 30 students, covering genomics assembly, molecular modeling with AlphaFold2/3 and molecular docking.

2022 **Academic Tutor**, *Course: Molecular Biology of the Cell*, Undergraduate's Program, Universidad Peruana Cayetano Heredia, Peru
Lectured sessions for the Peer Academic Mentoring Program (37 hours).

Extracurriculars

2024 **Journal Club Coordinator**, *Bioinformatics*, Sociedad Peruana de Bioinformática y Biología Computacional

2022 **Journal Club Participant**, *Structural Biology*, Sociedad Peruana de Bioinformática y Biología Computacional

2022 **Journal Club Coordinator**, *Artificial Intelligence*, Journal Club UPCH

2021 **Writer**, *University Journal*, The Novice Scientist UPCH

2020 **Journal Club Participant**, *Cell Biology*, Journal Club UPCH

2020 **Journal Club Participant**, *Cancer Biology*, Journal Club UPCH

References

MD, Ph.D. Louis Grandjean, *Professor of Infectious Diseases at University College London, UK.*, Email: l.grandjean@ucl.ac.uk

MD, Ph.D. Robert H. Gilman, *Professor of International Health at Johns Hopkins, USA*, Email: rgilman1@jhmi.edu

Ph.D. Mirko Zimic, *Professor of Bionformatics at Universidad Peruana Cayetano Heredia, Peru*, Email: mirko.zimic@upch.pe

Ph.D. Patricia Sheen, *Professor of Molecular Biology at Universidad Peruana Cayetano Heredia, Peru*, Email: patricia.sheen@upch.pe