Feature engineering for malaria genomic epidemiology

This M2 or equivalent internship is part of a French National Research Agency (ANR) funded project, UniGEM (Unified inference for Genomic Epidemiology of Malaria). Using simulation-based inference with deep-learning, UniGEM aims to build a neural network to estimate epidemiological parameters of *P. falciparum,* the deadliest malaria parasite species (read [1] to learn more about the ideas behind UniGEM). Within UniGEM, this internship aims to engineer the features input into the neural network and to compute their values using real malaria parasite data. These values will be used in a subsequent phase of the UniGEM project (beyond the scope of the internship) to assess the realism of simulated data, and to estimate epidemiological parameters globally.

***P. falciparum* data** are sourced from MalariaGEN, a network of data contributors coordinated by a team of sequencing experts at the Liverpool School of Tropical Medicine; see www.malariagen.net. The newest and largest *P. falciparum* data release, Pf8, includes whole-genome sequence (WGS) data for over 33,000 samples collected from 122 admin-level-one locations in 34 countries across South America, Africa, Asia, and Oceania between 1966 and 2022 [2]. Ninety-nine studies contributed samples to Pf8. They are summarised in an online application. Most studies sampled from malaria-positive patients in clinical field studies. The WGS data were generated centrally at the Wellcome Sanger Institute and are thus highly standardised; 24,409 analysis-ready samples passed central quality control. Among these, data on 1,544,381 biallelic single nucleotide polymorphisms in coding regions have been used to partition samples into 10 populations worldwide, and to estimate FWS (a per-sample intra-infection inbreeding coefficient) [2].

Internship work programme: The internship will start with training, before Pf8 data are partitioned and summarised. The intern will be encouraged to package their workflow into a reproducible and modular pipeline. These steps are described in more detail below.

Training: The internship will start with MalariaGEN online training, and training through tutorials provided by PGEforge, a community-driven platform designed to simplify *Plasmodium* genomic data analysis; see https://mrc-ide.github.io/PGEforge/.

Data partitioning: The 24,409 analysis-ready samples will be extracted and partitioned into subsets by country (since this is the most programmatically relevant geographic division) and, if necessary, by genetic populations within countries, i.e., on a more granular level than that reported in [2]. To do this, the intern will apply principal coordinates analysis to per-country samples. If evidence of population structure aligns with admin-level-one geography, we will further partition per-country samples geographically. If evidence of population structure aligns with collection dates, we will further partition per-country samples temporally. If evidence of population structure aligns with neither space nor time, or if there is no evidence of population structure among the per-country samples, we will not partition per-country samples. All subsets will be labelled as either having evidence of residual population structure (i.e., structure that aligns with neither space nor time) or not.

Summarise P. falciparum populations in a vectorised format, where each statistic is a candidate node in layer zero of a neural network: For all subsets with more than 20 samples a suite of summary statistics will be generated. Statistics will include summaries of distributions over multiplicities of infection (i.e., numbers of genetically distinct groups of parasites per infection), inbreeding coefficients, allele frequencies, linkage disequilibrium, pairwise relatedness, identity-by-descent segment lengths, etc. Some statistics will be generated using software designed specifically to analyse malaria parasite genetic data; see [3]. Although the interpretation of genetic summaries is an essential part of the intern's training, the intern is not expected to scrutinise their epidemiological meaning: the plan is to plug statistics into a neural network and let it interpret them as an abstract ensemble. A summary of symptomatic status will also be added to the vector of per-population genetic statistics: for each of the partner studies whose samples can be found among the subsets with sufficient samples, the average

symptomatic status of the contributed samples will be estimated from the literature, noting that subsets may include samples from different studies.

Engineering: The intern will be encouraged to package their workflow into a well-documented, modular, and parallelisable pipeline that can be used to efficiently generate vectors of per-population statistics given any set of samples. This pipeline will be used in a subsequent phase of the UniGEM project (beyond the scope of the internship) to generate summary statistics for synthetic datasets used to train, develop and test the neural network.

Location: This internship is supervised by Aimee Taylor, chargée de recherche, Infectious Disease Epidemiology and Analytics (IDEA) Unit, Department of Global Health, Institut Pasteur, Paris. Michael White directs IDEA. It is a highly international and multidisciplinary unit, with expertise in mathematical modelling and machine learning, wet-lab expertise in multiplex serological and genetic assays, expertise in diagnostic development and production, and expertise coordinating field studies and clinical trials. Major projects the intern will learn about through group meetings include PvSTATEM, a large cluster-randomised clinical trial with collaborators in Ethiopia and Madagascar, and ULTIMASero , a large cohort study with collaborators in Senegal, Cameroon, and Madagascar.

Required skills:

- Strong programming skills (R or Python or other)
- Familiarity with Git / GitHub for version control
- English proficiency (B2 level or higher)
- Experience handling large genomic datasets is desirable, e.g., Variant Call Format (VCF) files
- Basic understanding of population genetics is desirable, e.g., fixation index, F_{ST}

What the intern will gain: experience working with the largest open dataset of genome variation in any non-human eukaryotic species, conducting a broad range of population genetic analyses, engineering deep-learning inputs, pipeline engineering, and an enhanced open-science profile.

Publication prospects: The intern will be invited to co-author all publications of UniGEM activities that make use of the internship's results. Depending on the internship's progress, there is also potential for a software note describing the pipeline and its application to Pf8.

PhD prospect: This internship could potentially progress into a PhD funded through PRAIRIE (Paris Artificial Intelligence Research Institute). As such, candidates interested in PhD studies at the artificial intelligence / epidemiology interface are encouraged to apply.

Recommended reading

[1] F. Camponovo *et al.*, "Measurably recombining malaria parasites," *Trends Parasitol.*, vol. 39, no. 1, Jan. 2023.

[2] Malaria Genomic Epidemiology Network (MalariaGEN) *et al.*, "Pf8: an open dataset of Plasmodium falciparum genome variation in 33,325 worldwide samples," *Wellcome Open Res.*, vol. 10, June 2025.

[3] A. R. Taylor *et al.*, "Review of MrsFreqPhase methods: methods designed to estimate statistically malaria parasite multiplicity of infection, relatedness, frequency and phase," *Malar. J.*, vol. 23, no. 1, Oct. 2024.