Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sample composition shapes genome-wide classification of Plasmodium vivax infections in neighbouring Amazonian communities

Christopher Delgado-Ratto

christopher.delgadoratto@uantwerpen.be

University of Antwerp https://orcid.org/0000-0003-2300-5181

Mahdi Safarpour

University of Antwerp

Kristel Van Steen

University of Liège https://orcid.org/0000-0001-9868-5033

Dionicia Gamboa

Universidad Peruana Cayetano Heredia

Joseph Vinetz

Yale University

Jean-Pierre Van geertruyden

https://orcid.org/0000-0001-5006-6364

Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11785416/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Sample composition shapes genome-wide classification of *Plasmodium vivax* infections in neighbouring Amazonian communities

Mahdi Safarpour [1 *], Kristel Van Steen [2,3], Dionicia Gamboa [4,5], Joseph M Vinetz [5,6], Jean-Pierre Van geertruyden [1], Christopher Delgado-Ratto [1 *]

[1] Malaria Research Group (MaRch), Family Medicine and Population Health Department, Faculty of Medicine and Health Sciences, Global Health Institute, University of Antwerp, Antwerp, Belgium

[2] BIO3—GIGA-R Molecular and Computation Biology, University of Liege, Liège, Belgium

[3] BIO3—Department of Human Genetics, KU Leuven, Leuven, Belgium

[4] Laboratorio de Malaria: Parásitos y Vectores, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias e Ingeniería, Universidad Peruana Cayetano Heredia, Lima, Peru

[5] Laboratorio ICEMR-Enfermedades Emergentes, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias e Ingeniería, Universidad Peruana Cayetano Heredia, Lima, Peru

[6] Section of Infectious Diseases, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, United States

* These authors share corresponding authorship
Correspondence: Mahdi Safarpour (Mahdi.Safarpour@uantwerpen.be) and Christopher Delgado-Ratto (Christopher.DelgadoRatto@uantwerpen.be)

Abstract

As malaria transmission declines, determining where an infection was acquired becomes central to elimination, yet human mobility moves parasites between communities and makes their genetic profiles more similar. We asked whether genome-wide variation could assign *Plasmodium vivax* infections to one of two neighboring communities with residual transmission in the Peruvian Amazon, and how far this depended on sample composition. We analyzed 47 infections, 12 of them from people reporting recent local travel, using a machine-learning classifier (penalized logistic regression) with repeated nested cross-validation. Discrimination was modest (area under the curve 0.625, 95% interval 0.522–0.724) and close to chance once travelers were excluded (0.529, 0.402–0.657), although the two analyses used different, overlapping sets of infections. The most informative variants also differed, mapping to *msp7*-like genes and a putative multidrug resistance protein when travelers were included, and to merozoite surface protein 3 when they were not. Fine-scale genomic assignment in connected endemic communities is therefore limited, and sensitive to sample composition.

Introduction

Understanding the origin of malaria infections is critical for designing effective control and elimination strategies, particularly in settings where transmission is heterogeneous and influenced by human mobility [1,2]. In many endemic regions, infections may arise locally or be imported from nearby or distant locations, making it challenging to distinguish between ongoing transmission and reintroduction events [3,4]. In this context, genomic data may serve as a powerful tool for inferring the origin of infections and supporting malaria surveillance by describing parasite population structure, connectivity, and importation patterns [5,6]. In elimination settings, this information could help identify infections requiring further epidemiological investigation and guide the allocation of limited surveillance resources [6].

Genome-wide single nucleotide polymorphisms (SNPs) are widely used to study population structure, genetic differentiation, and parasite connectivity in *Plasmodium vivax* (*P. vivax*) [7-9]. These applications are particularly important in malaria-endemic settings where genetically related parasite populations may circulate across nearby communities, making it difficult to determine whether infections are locally acquired or linked to parasite movement. Beyond describing population structure, SNP data can also be used to predict the likely geographic

origin of infections. For example, a 33-SNP molecular barcode combined with likelihood-based classification has been used to assign *P. vivax* isolates to their country or region of origin, as has a second barcode panel [10], although performance depends on how well the reference data represent the relevant source populations [11]. However, it remains unclear whether these approaches can also distinguish infections sampled in nearby, genetically connected communities [9].

In this context, genome-wide analytical frameworks offer an opportunity to move beyond fixed barcode panels and evaluate whether broader parasite genomic variation contains sufficient information to distinguish infections across closely connected populations. Although genome-wide association studies (GWAS) are traditionally used to identify genetic variants associated with biological or clinical phenotypes, the same association-based logic can be adapted to a population-discrimination question [12]. Here, community or location can be treated as the outcome, and genome-wide SNP variation can be tested or modeled to determine whether parasite genotypes carry signals that distinguish one population from another [11]. The performance of such frameworks depends on several factors, including sequencing quality, sample size, population structure, and other confounders that may influence the relationship between SNPs and geographic origin [13,14].

Among these factors, human mobility is particularly important because it can reshape local parasite population structure and alter the relationship between parasite genotypes and geographic origin [2-4]. Travelers may carry parasites genetically distinct from those circulating in their community of residence, either because infections were acquired elsewhere or because mobility links multiple transmission networks [4]. Therefore, including or excluding travelers may change sample composition, which may affect observed genetic differentiation and predictive performance. Understanding this sensitivity is essential for interpreting genome-wide signals in malaria molecular surveillance, especially in settings where parasite movement, importation, and local transmission occur together.

Previous population-genetic analyses of the same whole-genome sequencing (WGS) dataset showed low genetic differentiation and substantial parasite connectivity between Libertad and Gamitanacocha, while traveler-associated infections generally remained embedded within community-level genetic clusters [15]. These findings indicated that the two communities are genetically connected, making this a challenging setting for fine-scale parasite assignment. Building on this evidence, we investigated whether genome-wide SNP data contain sufficient predictive information to distinguish *P. vivax* infections detected in Libertad and Gamitanacocha, and whether this ability is influenced by traveler-associated infections.

To address this question, we employed Elastic Net logistic regression, a penalized approach well suited to high-dimensional genomic data, and evaluated model performance using repeated cross-validation [16]. Models were fitted to the full dataset and to a dataset restricted to non-travelers. Beyond overall model performance, we compared predicted probabilities, SNP selection frequencies, and model coefficients between the two datasets to assess differences in classification patterns and the SNPs contributing to prediction. Finally, selected SNPs were functionally annotated to identify their genomic context and explore whether predictive variants were located in or near genes with potentially relevant biological functions. Overall, this exploratory study examined how well parasite genomic data could distinguish infections detected in nearby communities and highlighted the limitations of this approach. The findings may help guide the interpretation of genomic data alongside travel history to investigate malaria cases and better understand transmission between closely connected communities.

Methods

Study setting and samples

This study used the whole-genome sequencing (WGS) data described previously [15], and restricted to infections detected in Libertad and Gamitanacocha. These communities are located along the Mazán River in the Loreto region of the Peruvian Amazon and are connected through the same river system and shared travel routes. The analysis included 47 *P. vivax* infections: 24 from Libertad and 23 from Gamitanacocha. Of these, 12 were detected in individuals with recent local travel history. Travelers were defined as individuals who reported spending at least one night outside their community during the past month, based on a self-reported travel history questionnaire. Travel history therefore indicated recent mobility but did not establish where an infection was acquired. Mazán and Iquitos were the most frequently reported destinations, while direct travel between Libertad and Gamitanacocha was uncommon [15].

Ethics statement

The study was approved by the Ethics Committee of Antwerp University Hospital and the University of Antwerp (Project ID 3343; BUN B3002022000058), the Ethics Committee of Universidad Peruana Cayetano Heredia (SIDISI codes 208631 and 101518), and the Institutional Review Board of the Institute of Tropical Medicine, Antwerp (reference 1818/24). All methods were carried out in accordance with the Declaration of Helsinki and the Belgian law of 7 May 2004 on experiments on human persons. Written informed consent was obtained from all adult participants and from the parents or legal guardians of participants under 18 years of age; participants aged 12 to 17 years also provided written assent. Samples and epidemiological data were pseudonymised before analysis.

Genome-wide comparisons between *P. vivax* populations

Sequencing, read processing, variant calling, and quality filtering followed the workflow described previously [15]. After sample- and variant-level filtering, 19,297 biallelic SNPs were retained for downstream analyses. Genome-wide comparisons were conducted to assess genetic differentiation between *P. vivax* populations from Libertad and Gamitanacocha. Rather than a classical GWAS with a phenotype, this analysis represents a genome-wide allele frequency comparison between populations.

Samples were grouped by community, and analyses were conducted on three datasets: all samples, non-travelers only, and travelers only. For each comparison, we created a dataset containing samples from the two communities and coded community membership as 1 for Gamitanacocha and 2 for Libertad. Genome-wide association analyses were performed using PLINK v1.9 [17] to test for differences in allele frequencies between Gamitanacocha and Libertad at each SNP. Association results were visualized using Manhattan plots, with SNPs plotted by chromosomal position against $-\log_{10}(p\text{-value})$. For comparison across datasets, a fixed threshold of $-\log_{10}(P) = 5$ ($P = 1 \times 10^{-5}$) was used to highlight SNPs showing evidence of allele-frequency differences between Gamitanacocha and Libertad. This cutoff was more permissive than the Bonferroni-corrected threshold for 19,297 tests, calculated as $P = 0.05/19{,}297 = 2.59 \times 10^{-6}$, corresponding to $-\log_{10}(P) = 5.59$. The lower cutoff of 5 was used to explore SNPs that might not meet the stricter threshold.

Assessment of population stratification

Population stratification and potential inflation of association statistics were assessed using the genomic inflation factor (λ_{GC}) and quantile–quantile (QQ) plots. SNP association results were obtained from PLINK separately for all samples (full model) and for non-travelers. The genomic inflation factor (λ_{GC}) was calculated as the ratio of the median of the observed chi-square test statistics to the expected median under the null distribution, with the expected median corresponding to a chi-square distribution with one degree of freedom [18]. QQ plots were generated by comparing the observed distribution of SNP association p-values with the distribution expected under the null hypothesis, with systematic deviation above the diagonal interpreted as inflation of the association statistics.

Missing data imputation

Missing SNP genotypes were imputed using a k-nearest neighbors (KNN) approach with mode imputation, implemented within a nested cross-validation framework to avoid information leakage [19]. The number of neighbors (K) was evaluated on a predefined grid (K = 3, 5, 7, and 9). For each iteration, the dataset was randomly partitioned into 5 outer folds, and this process was repeated 20 times (for a total of 100 outer folds). Within each outer fold, one subset was held out as the test set, and the remaining data were used for training. All imputation steps were performed independently within each training set, and the same imputation parameters were then applied to the corresponding test set.

Within each outer training set, an inner 3-fold cross-validation loop was used to jointly select the optimal K for imputation. For each candidate K, missing values in the training data were imputed using the mode of the K nearest neighbors, with neighbors defined by pairwise genetic distance calculated as the proportion of mismatched alleles across shared non-missing SNPs. If no valid neighbors were available for a given SNP, the column-wise mode (calculated from the training data) served as a fallback. Once the optimal K was selected in the inner loop, missing values in the full outer training set were imputed using this K. The same trained imputation structure (including selected SNPs and column modes) was then applied to the outer test set, where missing values were imputed using the K nearest neighbors identified from the training data only.

Elastic Net modeling

To evaluate whether parasite SNP data could predict the community in which an infection was detected (Libertad vs. Gamitanacocha), we developed penalized logistic regression models using Elastic Net regularization, which combines L1 (LASSO) and L2 (ridge) penalties to enable both variable selection and stable coefficient estimation in

the presence of correlated predictors [16]. Prior to model fitting, SNP predictors were standardized to have mean=0 and standard deviation=1 using statistics derived from the training data, and the same transformation was applied to the corresponding validation data to ensure consistency and prevent data leakage.

The Elastic Net hyperparameters, including the mixing parameter (α) and the regularization strength (λ), were optimized using an inner cross-validation procedure within each training fold. The α parameter was evaluated on a grid from 0 to 1 (in increments of 0.1), while λ was selected via cross-validated glmnet to maximize the area under the receiver operating characteristic curve (AUC) [20]. For each combination of α (and K for imputation), model performance was assessed using the inner folds, and the optimal parameter set was chosen based on the highest mean AUC (Supplementary Fig. S1).

Evaluation of K selection and its impact on model performance

To assess the stability of the K parameter used in KNN imputation, the optimal K selected in each cross-validation iteration was recorded across all 100 iterations. Summary statistics, including the cumulative mean, overall mean, and rolling mean, were used to evaluate the consistency of K selection over iterations.

To examine the impact of K on model performance, AUC values obtained from each iteration were stratified by the selected K. The distribution of AUC across iterations was then compared for different K values using boxplots, allowing evaluation of how the choice of K influences predictive performance. These analyses were performed separately for the model developed using all samples (full model) and the model restricted to non-travelers.

Model performance evaluation

Model performance was evaluated using repeated nested cross-validation, with 20 repetitions and five outer test folds per repetition. Each infection received one held-out prediction per repetition, giving 20 predictions per infection. Predictions from all 100 outer test folds were pooled to calculate AUC separately for the full dataset and the non-traveler subset. Gamitanacocha was treated as the positive class.

Exploratory 95% bootstrap uncertainty intervals were calculated using 100,000 infection-level resamples. Infections were sampled with replacement within each combination of community and travel status, preserving the observed group sizes. All 20 predictions for each selected infection were retained together. When a shared non-traveler infection was sampled, its predictions from both analyses were included the same number of times. Both AUCs and their difference, calculated as full minus non-traveler AUC, were recalculated for each resample. The 2.5th and 97.5th percentiles defined the interval limits. The intervals were conditional on the saved predictions, as models were not refitted.

In addition to Receiver Operating Characteristic (ROC)-based evaluation, precision–recall (PR) curves were generated to summarize the trade-off between precision and recall across classification thresholds [21]. To further evaluate classification behavior, threshold-dependent metrics were examined, including F1-score, recall (sensitivity), and precision, each calculated across a range of probability thresholds. These analyses allowed assessment of how model performance varies with the choice of classification threshold and facilitated identification of an optimal threshold for prediction.

Analysis of predicted probabilities

To evaluate model discrimination, predicted probabilities from the Elastic Net models were analyzed. For each sample, the predicted probability corresponded to the likelihood of belonging to Gamitanacocha (class 1). The distribution of predicted probabilities was examined separately for models developed using all samples (full model) and non-travelers only. For each model, predicted probabilities were visualized using both boxplots and density distributions, stratified by the true community label, to assess the degree of separation between Libertad and Gamitanacocha.

To further describe patterns by travel history, predicted probabilities from the full model were stratified by travel status within each community. Both sample-level distributions and aggregated summaries (boxplots) were used to compare patterns between travelers and non-travelers, allowing assessment of how classification patterns vary with reported travel history.

Identification of informative SNPs

To identify SNPs contributing to classification, model coefficients were extracted from each cross-validation iteration. SNPs were considered selected if their Elastic Net coefficients were non-zero. Two complementary ranking approaches were used. First, SNPs were ranked based on their selection frequency, defined as the number

of times a SNP was selected across all 100 iterations, to identify SNPs consistently contributing to the model. Second, SNPs were ranked based on their mean penalized model coefficient across 100 iterations. For coefficient-based ranking, the top 30 SNPs were selected, including the 15 most positive and 15 most negative coefficients, representing SNPs most strongly associated with classification toward Gamitanacocha (class 1) or Libertad (class 0). These analyses were performed separately for the full model and the non-traveler model to compare SNP selection frequencies and mean coefficients between the two datasets.

Comparison of SNP selection between full and non-traveler models

To assess the consistency of SNP selection between two models, SNPs identified in the full dataset (all samples) were compared with those from the non-traveler subset. For each model, SNPs were ranked based on their selection frequency across cross-validation iterations, and the top 30 SNPs were retained. Overlap between the two sets was evaluated to identify SNPs consistently selected across both models, as well as those specific to either top-ranked set. For SNPs shared between the two top-ranked sets, differences in mean penalized coefficients were then quantified using the mean model coefficient across iterations. The difference (Δ) was calculated as the mean coefficient in the full model minus that in the non-traveler model. This approach allowed assessment of both the stability of SNP selection and differences in signed coefficients between the two fitted models.

Comparison of SNP coefficients between full and non-traveler models

To compare SNP importance between the full dataset and the non-traveler subset, SNPs were ranked separately in each dataset by their mean model coefficients across cross-validation iterations. For each model, the top 15 SNPs with the most positive coefficients and the top 15 SNPs with the most negative coefficients were selected, representing SNPs most strongly associated with classification toward Gamitanacocha (class 1) and Libertad (class 0), respectively. Overlap between the two datasets was assessed independently for the positive and negative SNP sets to evaluate the consistency of coefficient-based selection. For SNPs present in both models within each group, differences in mean coefficients were calculated as the mean coefficient in the full model minus that in the non-traveler model (Δ). This approach enabled comparison of both the stability of SNP selection based on mean coefficient and differences in signed coefficients between models fitted to the two datasets.

Functional annotation of selected SNPs

Functional annotation of selected SNPs was performed using snpEff based on the *P. vivax* PAM reference annotation database available in the snpEff database directory (PlasmoDB-68_PvivaxPAM) [22]. The input files for annotation were the filtered joint VCF files generated after variant calling and filtering. The annotated VCF files were then used for downstream SNP-level functional interpretation. In addition, GATK VariantsToTable was used to extract variant information, including chromosome, position, variant type, and genotype information.

For downstream functional interpretation, SNPs were selected from the machine-learning output files based on their mean model coefficients. Specifically, the top 15 SNPs with the highest positive mean coefficients and the top 15 SNPs with the most negative mean coefficients were identified separately in models trained using all samples (full model) and in models trained using non-travelers only. SNP identifiers were parsed into chromosome and position coordinates and matched to the snpEff-annotated VCF by exact genomic position.

The selected SNPs were then compared between the full and non-travelers models to identify SNPs unique to each model and SNPs shared between models. SNPs classified as model-specific represented variants present in the top coefficient-ranked list of one model but absent from the corresponding top list of the other model. For shared SNPs, the difference in mean coefficient was calculated as Δ = mean coefficient in all samples minus mean coefficient in non-travelers. Shared variants were grouped according to whether their mean coefficient was more positive in the full model, more positive in the non-traveler model, or equal in both models.

Functional annotations were extracted from the ANN field generated by snpEff. This structured field in the annotated VCF reports the predicted consequence of each variant, its impact category, the affected gene and transcript, and any associated coding or protein change. The snpEff ANN field was parsed to extract the predicted variant effect, predicted impact, gene identifier, transcript identifier, transcript biotype, coding change, and protein change. Variant effects were grouped into functional classes, including synonymous, non-synonymous/missense, high-impact coding, frameshift, splice-related, intronic, upstream/downstream, and intergenic variants. Variants with HIGH or MODERATE predicted impact were considered the most likely to have functional relevance, while upstream and downstream variants were interpreted as possible regulatory candidates.

Because individual SNPs may receive multiple annotations, two complementary summaries were generated. A complete annotation table retained all snpEff annotations per SNP, while a simplified table retained one primary

annotation per SNP based on predicted impact priority, with HIGH prioritized over MODERATE, LOW, and MODIFIER effects. Gene descriptions were added by linking snpEff gene identifiers to the corresponding entries in the *P. vivax* GFF annotation file using the gene-level description field.

Data Availability

The raw sequence data generated and analysed during this study are deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA1492536 (https://dataview.ncbi.nlm.nih.gov/object/PRJNA1492536?reviewer=ikl9kn8dnem3rjbnk3357edsjp&archive=biosample).

Code availability

R scripts used for data processing, genome-wide association (GWAS) studies, and figure generation are available at the GitHub repository: https://github.com/pv-wgs-peru-mobility/MetaVivax.

Results

Genetic differentiation between communities

Genome-wide comparisons showed limited differentiation between Libertad and Gamitanacocha communities, consistent with the low pairwise Fixation Index (Fst) estimates previously obtained from the same WGS data (Fst=0.056 overall and Fst=0.043 among non-travelers) [15]. In the analysis including all samples, a small number of loci (n = 31) exceeded the threshold, appearing as scattered peaks across multiple chromosomes rather than forming concentrated regions of differentiation (Fig. 1). When comparing infections in non-travelers from Libertad and Gamitanacocha, no variants exceeded the threshold, although some loci showed modestly elevated signals. In the traveler-only analysis, one locus crossed the threshold.

Quantile–quantile (QQ) plots revealed substantial deviation of observed p-values from the expected null distribution in both datasets (Supplementary Fig. S2). In the non-traveler dataset, observed values deviated above the diagonal. This deviation was more pronounced in the full dataset including travelers, where the distribution was shifted upward relative to the expected line. The genomic inflation factor was higher in the full dataset (λ_{GC} = 3.72) than in the non-traveler dataset (λ_{GC} = 3.37).

Model tuning and imputation

Across the 100 outer cross-validation folds, the mean selected K for KNN imputation was 6.14 (median = 7, SD = 2.13) in the full model and 6.54 (median = 7, SD = 2.14) in the non-traveler model (Supplementary Fig. S3).

Overall, K = 7 was most frequently selected in the full model (29 of 100 folds), whereas K = 9 was most frequently selected in the non-traveler model (32 of 100 folds). The selected K changed between consecutive folds in 83 of 99 comparisons in the full model and 74 of 99 comparisons in the non-traveler model. Across all 100 folds, the cumulative mean ranged from 5.00 to 9.00 in the full model and from 5.80 to 9.00 in the non-traveler model. The final cumulative mean was 6.14 in the full model and 6.54 in the non-traveler model. Rolling means calculated with a window of 10 folds ranged from 4.6 to 7.2 in the full model and from 5.3 to 7.8 in the non-traveler model (Supplementary Fig. S3).

In the full model, median AUC was 0.70 (IQR 0.50–0.84) for K = 7, followed by 0.69 (0.56–0.77) for K = 5, 0.64 (0.57–0.75) for K = 3, and 0.60 (0.50–0.71) for K = 9. In the non-traveler model, median AUC was approximately 0.53 (IQR 0.50–0.67) for K = 3, 0.56 (0.50–0.68) for K = 5, and 0.56 (0.50–0.67) for both K = 7 and K = 9. The interquartile ranges overlapped across all four K values in both models (Supplementary Fig. S4).

Classification performance and predicted probabilities

The full model achieved an AUC of 0.625 (95% bootstrap interval: 0.522–0.724), indicating modest discriminatory ability between Gamitanacocha and Libertad. When restricting the analysis to non-travelers, the AUC was 0.529 (95% bootstrap interval: 0.402–0.657), approaching random classification (Fig. 2). The AUC difference between the two models was 0.096 (95% bootstrap interval: 0.019–0.168), indicating higher observed discrimination between communities in the full dataset.

The precision–recall analysis showed a PR-AUC of 0.562 for the full model and 0.493 for the non-traveler model. The reported F1 scores were almost identical (0.678 and 0.677, respectively), despite the difference in PR-AUC (Supplementary Fig. S5).

The relationship between classification threshold and performance metrics differed between the full and non-traveler models (Supplementary Fig. S6). For the F1-score, both models showed a plateau at lower thresholds, followed by a sharp decline as the threshold increased. The optimal threshold was higher in the full model (0.47) compared to the non-traveler model (0.23). Additionally, the full model maintained slightly higher F1 values over a broader range of thresholds, whereas the non-traveler model showed a more rapid decline after the optimal point.

Recall decreased as the classification threshold increased in both models. At the selected thresholds of 0.467 for the full model and 0.226 for the non-traveler model, recall was 0.90 and 0.99, respectively, while precision was 0.54 and 0.52. At a threshold of 0.60, recall had declined to 0.17 in the full model and 0.16 in the non-traveler model.

In the full model, precision increased to 0.60 at a threshold of 0.55, then decreased to 0.46 at a threshold of 0.75 before fluctuating at higher thresholds. In the non-traveler model, precision reached 0.54 at a threshold of 0.52 and decreased to 0.22 at a threshold of 0.75 (Supplementary Fig. S6).

In the full model, the median predicted probability of belonging to Gamitanacocha was 0.54 (IQR 0.52–0.58) for Gamitanacocha samples and 0.48 (0.42–0.56) for Libertad samples (Fig. 3a). In the non-traveler model, the corresponding medians were 0.54 (IQR 0.51–0.55) for Gamitanacocha and 0.49 (0.45–0.58) for Libertad (Fig. 3b). The narrower IQR in Gamitanacocha indicated less variation in predicted probabilities than in Libertad. The density curves peaked near 0.50 for both communities in the full dataset and the non-traveler datasets (Fig. 3).

Predicted probabilities of belonging to Gamitanacocha overlapped between travelers and non-travelers in both communities. In Gamitanacocha, probabilities ranged from 0.40 to 0.61 among non-travelers and from 0.46 to 0.60 among travelers. Libertad showed wider observed ranges in both groups: 0.22–0.64 among non-travelers and 0.27–0.60 among travelers. In both communities, travelers had a narrower range of predicted probabilities than non-travelers, and both travelers and non-travelers had values above and below 0.50. (Fig. 4a). Across both communities, travelers had a lower median predicted probability (0.46; IQR 0.43–0.57) than non-travelers (0.54; IQR 0.46–0.57), although their interquartile ranges overlapped (Fig. 4b).

Identification and comparison of informative SNPs

Among the top 30 SNPs ranked by selection frequency, selection counts ranged from 59 to 88 of the 100 cross-validation folds in the full model and from 41 to 68 in the non-traveler model (Figs. 5 and 5. The most frequently selected SNP in the full model, CASCJQ010000014:2575209, was selected in 88 folds, compared with 59 folds in the non-traveler model. Four SNPs shared the highest selection count in the non-traveler model, each appearing in 68 folds: CASCJQ010000014:2575210, CASCJQ010000013:2199774, CASCJQ010000005:564355 and CASCJQ010000003:93468.

Positive coefficients, corresponding to classification toward Gamitanacocha, ranged from 0.029 to 0.171 in the full model and from 0.011 to 0.054 in the non-traveler model. Negative coefficients, corresponding to classification toward Libertad, ranged from −0.113 to −0.025 and from −0.068 to −0.016, respectively (Figs. 5b and 5d). The largest positive coefficient in the full model was observed for CASCJQ010000013:1020569 (0.171), whereas CASCJQ010000012:1943763 ranked highest in the non-traveler model (0.054). The most negative coefficients were observed for CASCJQ010000012:800711 in the full model (−0.113) and CASCJQ010000014:2575209 in the non-traveler model (−0.068).

The two models shared 16 of their top 30 frequency-ranked SNPs (53.3%), while 14 SNPs (46.7%) were specific to each ranked list (Supplementary Fig. S7a). Across these 16 shared SNPs, differences in mean coefficient (Δ = full model minus non-traveler model) ranged from −0.018 to 0.021. Eleven SNPs had more positive coefficients in the full model, whereas five had more negative coefficients (Supplementary Fig. S7b). The largest positive differences were observed for CASCJQ010000014:2575210 (Δ = 0.021) and CASCJQ010000009:664697 (Δ = 0.020). The largest negative differences occurred for CASCJQ010000005:661739 (Δ = −0.018) and CASCJQ010000009:279513 (Δ = −0.011).

Coefficient-based rankings showed overlap for both positive and negative coefficients. The full and non-traveler models shared eight of their top 15 positive-ranked SNPs (53.3%), with seven SNPs (46.7%) specific to each top-ranked set. The same pattern was observed for negative coefficients: eight SNPs were shared, and seven were specific to each top-ranked set (Supplementary Figs. S8a and S8c).

All eight shared positive-ranked SNPs had larger positive mean coefficients in the full model, with differences of 0.013–0.148 (Supplementary Fig. S8b). The largest difference was observed for CASCJQ010000013:1020569, with a mean coefficient of 0.171 in the full model and 0.023 in the non-traveler model. Among the eight shared negative-ranked SNPs, six had more negative coefficients in the full model, whereas two, CASCJQ010000014:2575209 and CASCJQ010000014:2575210, had more negative coefficients in the non-traveler model (Supplementary Fig. S8d).

Genomic annotation profiles of selected SNPs

Among the 14 SNPs specific to the full-model top-ranked sets, six (42.9%) had MODERATE predicted impact and eight (57.1%) had MODIFIER predicted impact (Fig. 6). The MODERATE-impact annotations mapped to two MSP7-like genes (PVPAM_120024800 and PVPAM_120024900), a putative multidrug resistance protein 2 (PVPAM_120064400), a putative transcription factor (PVPAM_030011200) and two conserved *Plasmodium* proteins of unknown function (PVPAM_110023100 and PVPAM_110006500). The annotation for PVPAM_120024800 occurred among SNPs with positive coefficients, corresponding to classification toward Gamitanacocha, whereas PVPAM_120024900 occurred among SNPs with negative coefficients, corresponding to classification toward Libertad.

Among the 14 SNPs specific to the non-traveler top-ranked sets, two (14.3%) had MODERATE predicted impact, two (14.3%) had LOW predicted impact and ten (71.4%) had MODIFIER predicted impact (Fig. 7). MODERATE-impact annotations mapped to merozoite surface protein 3 (PVPAM_100038000) and a conserved *Plasmodium* protein of unknown function (PVPAM_030007100). A second MSP3 gene, PVPAM_100037600, appeared among the MODIFIER-impact annotations. LOW-impact annotations mapped to a putative zinc finger protein (PVPAM_030013700) and a conserved *Plasmodium* protein of unknown function (PVPAM_120020700). Other MODIFIER-impact annotations mapped to genes annotated as a putative G-patch domain-containing protein, vacuolar protein sorting-associated protein VTA1, apicoplast ribosomal protein L33, a lipase maturation factor and DNA damage-inducible protein 1.

Among the shared SNPs, annotations associated with more positive mean coefficients in the full model included a conserved *Plasmodium* protein of unknown function with MODERATE predicted impact (PVPAM_090052100) and a putative sphingomyelin phosphodiesterase with LOW predicted impact (PVPAM_140065200). MODIFIER-impact annotations in this group included a putative *P. vivax* Vir protein (PVPAM_040035200), a 6-cysteine protein (PVPAM_080035400), a putative ubiquitin domain-containing protein DSK2 (PVPAM_090018300), and a putative Golgi apparatus membrane protein TVP23. By comparison, annotations associated with more positive mean coefficients in the non-traveler model included a putative polyribonucleotide 5′-hydroxyl-kinase Clp1 with MODERATE predicted impact (PVPAM_050020200) and a conserved *Plasmodium* protein of unknown function with LOW predicted impact (PVPAM_050036500). MODIFIER-impact annotations in this group included a putative serine/threonine protein kinase STK2 and a CPW-WPC family protein (Fig. 8).

Discussion

This study investigated whether genome-wide SNP data could discriminate *P. vivax* infections sampled in two geographically close communities in the Peruvian Amazon and whether including traveler-associated infections changed the observed genomic classification signal. Overall, the findings showed that discrimination between Libertad and Gamitanacocha is modest. When all samples were included, the Elastic Net model achieved an AUC of 0.625, whereas performance was close to random classification when the analysis was restricted to non-travelers (AUC = 0.529). This pattern indicated that observed discrimination varied with sample composition. Previous population-genetic analyses of the same WGS data by Safarpour et al. showed low genetic differentiation, partial clustering overlap, and substantial parasite connectivity between Libertad and Gamitanacocha communities [15]. The present findings extend these earlier results by showing that the similar community-level genomic structure provided only modest predictive information in the sampled infections. These findings should therefore be viewed as a proof of concept for fine-scale assignment to the community of sampling.

This work applied a genome-wide predictive framework at a fine spatial scale within an endemic region, while explicitly examining how the inclusion of travelers changed model performance and SNP ranking. Several previous applications of parasite genotyping for origin assignment have focused on broader geographic scales, such as country-level or regional classification. For example, *P. vivax* barcode studies have shown that carefully selected SNP panels can predict broad geographic origin and support the identification of imported infections [10,23]. Similarly, machine-learning approaches using whole-genome sequence data have demonstrated that parasite genomes can be used for geographic classification when sufficient reference diversity is available [11]. In this study, modest discrimination highlighted the difficulty of distinguishing infections sampled in nearby communities with closely related malaria parasite populations. These findings helped clarify the limitations of genomic assignment for case investigation in connected endemic settings.

Consistent with the modest classification performance, the limited allele-frequency differences suggest that parasite populations in Libertad and Gamitanacocha are not strongly genetically differentiated. No SNPs exceeded the reporting threshold among non-travelers. This does not necessarily indicate the absence of population structure. Rather, it is compatible with shared parasite ancestry and weak allele-frequency differences at individual loci, although the small sample also limits power to detect those differences. This interpretation is compatible with the low differentiation and overlapping population structure reported for these communities using the same WGS data [15]. Earlier population genetic work in the Peruvian Amazon also reported clonal structure and parasite movement connecting transmission areas, highlighting that *P. vivax* populations can be locally structured yet epidemiologically connected [7,24].

The allele-frequency comparisons also showed inflation in the association test statistics, with genomic inflation factors of 3.72 in the full dataset and 3.37 among non-travelers. In classical human GWAS, inflation can arise from uncontrolled ancestry, relatedness, technical effects, or widespread associations [13,18,25]. In this study, the tested outcome was community membership rather than a biological disease phenotype; therefore, deviation from the null may partly reflect the population structure that the analysis aimed to detect. However, the higher inflation in the full dataset does not by itself demonstrate that traveler-associated infections increase genetic diversity. Relatedness, linkage disequilibrium, sample composition, and technical variation may also influence association statistics [13,26]. These potential sources of inflation should be assessed before drawing conclusions about individual SNPs. The observed associations therefore require confirmation in independent samples and further population-genetic analyses. Therefore, they do not, on their own, show that the SNPs have causal effects or are under natural selection.

The lower observed AUC after excluding travelers suggests that the genomic classification signal depends on which infections are included. The AUC differed by 0.096 (95% bootstrap interval: 0.019–0.168) between the full dataset and the non-traveler subset. The present comparison showed how observed performance differed when travelers were included or excluded and provided a basis for larger studies examining the role of travel history in genomic classification. This is aligned with studies showing that integrating parasite genetic data with human mobility information can improve the investigation of malaria connectivity and imported infections [4]. It also fits with the broader genetic epidemiology framework in which connectivity, gene flow, relatedness, and imported cases are key use cases for malaria surveillance and elimination programs [6].

The predicted probability analysis provided additional context for the limited discrimination between communities. In Gamitanacocha, travelers and non-travelers showed overlapping predicted probabilities. In Libertad, both groups covered a broader range than their counterparts in Gamitanacocha. Within each community, non-travelers had a wider observed range than travelers. These patterns indicated that travelers did not form a clearly distinct group based on their classification scores. Interpreting these patterns using travel history is complicated by possible differences in travel routes, destinations, and exposure settings between the two communities. Moreover, infections detected in travelers may have been acquired before travel, during travel or after return, or may reflect relapse. This is particularly relevant for *P. vivax*, because dormant hypnozoites can reactivate, making the location of infection acquisition difficult to determine [27]. Similarly, no reported travel during the preceding month does not exclude earlier exposure elsewhere or infection with parasites circulating between connected communities.

The comparison of informative SNPs between the full and non-traveler models also provided insight into how sample composition changes the features used for classification. The partial overlap between the ranked SNP sets suggested that some SNPs remained informative when travelers were included or excluded, while others ranked differently between the two models. These findings suggest that sample composition may influence which markers rank most highly.

Comparing coefficients between models also requires care because their magnitude depends partly on regularization and predictor scaling [16,20]. Differences between models should therefore be interpreted as changes in how SNPs contribute to classification, rather than evidence that traveler-associated infections strengthen biological effects. Beyond these modelling considerations, population structure, linkage disequilibrium, sample composition and local demography can also influence which SNPs show genomic associations in malaria studies [8,28]. The selected SNPs may thus capture genetic differences between parasite populations, but their predictive contribution alone does not establish a functional role.

Functional annotation helped assess whether including or excluding traveler-associated infections changed the biological context of the top-ranked SNPs. SNPs ranked highly only in the full model included MODERATE-impact annotations in two *msp7*-like genes and a putative multidrug resistance protein 2. MSP7-like proteins are encoded by a family of related genes and have been studied for their roles during blood-stage malaria infection and their interactions with the host immune system [29,30]. The SNP annotated to a putative multidrug resistance protein 2 also needs further investigation, because the protein's role in drug transport and parasite biology remains unclear [31]. Neither the annotation nor its predicted impact demonstrates differences in drug susceptibility in these infections. Including or excluding travelers highlighted different genes for follow-up, but further studies are needed to determine whether the selected variants affect parasite function.

SNPs found only among the top-ranked features in the non-traveler model included a MODERATE-impact annotation in a gene annotated as merozoite surface protein 3. The MSP3 family includes merozoite-associated proteins [32]. Antibody responses to PvMSP3 antigens have been observed in people with malaria [33,34]. The SNP annotated to PVPAM_100038000 had a MODERATE predicted impact and favored classification as Libertad, whereas the SNP annotated to PVPAM_100037600 had a MODIFIER predicted impact and favored classification as Gamitanacocha. Surface-protein annotations therefore appeared among markers predicting either community in the non-traveler model. These annotations provide biological context for the selected SNPs, but further research is needed to determine whether the SNPs affect protein function.

The shared SNPs provided further biological context across both model settings. These SNPs mapped to genes annotated as a putative *P. vivax* VIR protein, a 6-cysteine protein, a putative sphingomyelin phosphodiesterase, a putative ubiquitin domain-containing protein DSK2, and conserved *Plasmodium* proteins of unknown function. VIR proteins form a large *P. vivax* multigene family (*vir*), investigated in relation to immune recognition and host–parasite interaction, although the functions of many members remain incompletely characterized [35,36]. Similarly, *Plasmodium* 6-cysteine proteins have diverse roles across the parasite life cycle [37]. Studies using recombinant proteins have shown that the *P. vivax* proteins P12 and P41 interact with each other [38]. Annotations shared by the full and non-traveler models highlight genes that remained relevant to prediction when travelers were included or excluded. Further studies are needed to confirm whether these SNPs consistently help distinguish infections from the two communities and whether they affect gene function.

Functional annotation provided biological context for the selected SNPs, but it did not establish that these SNPs are causal, under selection, or directly involved in the annotated biological processes. Elastic Net selected SNPs according to their contribution to prediction, and correlated SNPs may act as markers for broader haplotypes or population structure [16]. MODERATE, LOW, and MODIFIER are predicted annotation categories, not experimentally measured effects on parasite fitness or transmission [22]. The gene-level findings should therefore be considered hypothesis-generating and require replication in larger independent datasets.

Although this study was not a classical GWAS of parasite phenotype, the approach is conceptually related to parasite association studies that use genome-wide SNPs to identify loci linked to biological or epidemiological traits. For example, parasite GWAS has been used to identify loci associated with antimalarial drug response in *P. falciparum*, while controlling for population structure and linkage disequilibrium [39]. Host malaria GWAS have also shown that strong signals can occur in biologically plausible regions, such as erythrocyte-related genes, but require large sample sizes, replication, and careful handling of population structure [40]. These examples illustrate the need to distinguish predictive markers from biological mechanisms. In the present analysis, the outcome was community of sampling, so an allele-frequency difference or a model coefficient cannot be interpreted as an effect on drug response, virulence, or transmission. Therefore, further epidemiological and experimental studies are needed to determine whether the selected SNPs affect parasite function.

An important contribution of this study is that it evaluated travel history as epidemiological context for interpreting genomic classification. In malaria elimination settings, imported infections can contribute to parasite movement and reintroduce transmission into areas where transmission has declined [2,4,10]. The present results showed that

excluding travelers changes the observed classification performance and SNP ranking. These findings support recording travel history alongside genomic data to help interpret classification results. Travelers and non-travelers had overlapping predicted probabilities, meaning that the model did not clearly distinguish infections by travel status. To investigate where an infection may have been acquired, genetic findings should be considered together with the travel history data and the time and place of infection detection [4,6].

This study has several limitations. First, the sample size was relatively small for genome-wide prediction and SNP-level comparison, particularly after stratification by travel status. This reduces statistical power, increases uncertainty about individual SNP coefficients, and limits interpretation of the allele-frequency differences. Repeated cross-validation does not overcome the small sample size because it repeatedly uses the same infections. The 100 outer-fold evaluations therefore provided related estimates of performance, rather than 100 independent validations [41]. Uncertainty intervals must account for this reuse of samples, and the reported AUC values alone do not establish whether performance reliably exceeds chance [42]. Second, *P. vivax* biology complicates inference of infection origin. Relapses from hypnozoites may produce blood-stage infections weeks or months after the original mosquito bite, so the location where infection was detected may differ from the location where it was acquired [27]. The one-month travel definition also does not capture all earlier exposure. Third, the models were internally validated using repeated nested cross-validation but were not tested on an independent external dataset. Therefore, the reported performance reflects predictive ability within the current study setting and should not be interpreted as confirmed accuracy in other communities or time periods. In addition, randomly dividing samples into training and test folds may place closely related parasites in both groups. This can make the model appear more accurate than it would be when classifying infections from genetically different parasite lineages [14]. Therefore, external validation using independently sampled infections is required before this approach can support surveillance decisions. Fourth, missing genotype imputation, SNP filtering, and hyperparameter choices may influence model performance and SNP selection. Fold-to-fold variation in selected K indicated that the imputation setting also depends on the resampled data. Even if the average selected K stabilizes as more folds are evaluated, this does not show that the model classifies infections accurately. Finally, functional interpretation of selected SNPs remains limited because the analysis identified predictive markers rather than experimentally validated biological mechanisms.

Future studies should include larger and temporally structured sampling from multiple communities across the Peruvian Amazon. This would allow the current binary comparison to be extended into a multi-source classification framework and would help evaluate plausible source populations and parasite mixing. Reference sampling should include relevant travel destinations, and infections should remain unassigned when the available data do not clearly support a source population. Therefore, the current two-community classifier cannot establish an origin outside those communities. Integration of travel surveys, mobility data, entomological information, and parasite genomic data would also improve investigation of routes of parasite gene flow and sources of transmission [4]. More detailed information on travel dates, destinations, symptom onset, and previous malaria episodes would also help distinguish alternative exposure histories.

Future analyses should also compare Elastic Net with complementary methods, including random forests, gradient boosting, support vector machines, and probabilistic assignment models, while prioritizing interpretability and external validation. Such comparisons should use the same held-out infections and include simple baseline models, probability calibration, and validation across sampling periods or parasite-relatedness groups. Finally, selected SNPs should be evaluated in independent genomic datasets and, where justified, experimental studies to determine whether they mark population structure or have measurable biological consequences. Together, these extensions would strengthen evaluation of genome-wide data as a surveillance tool for understanding human mobility and malaria transmission.

Conclusion

Genome-wide SNP data showed a modest signal for assigning *P. vivax* infections to the two communities where they were sampled, and the observed AUC was lower when traveler-associated infections were excluded. These findings indicated that sample composition affected the classification results and selected genomic features. The substantial overlap in predicted probabilities further highlighted the limited separation between the sampled communities. This framework should therefore be viewed as a proof of concept that may support, rather than replace, travel history and epidemiological case investigation. Larger datasets, external validation, and broader geographic sampling will be required before fine-scale genomic assignment can inform malaria-elimination programs.

Overall, the findings support the value of interpreting parasite genomic data alongside human mobility information, while recognizing the limits of classification in closely connected endemic communities. The partially shared SNP rankings and differences between the fitted models illustrated the importance of reporting sample composition when interpreting predictive genetic markers. With larger datasets, external validation, and integration of mobility and epidemiological information, this framework could help prioritize unusual genomic observations for investigation and generate hypotheses about parasite connectivity and transmission.

References

1 Cotter, C. *et al.* The changing epidemiology of malaria elimination: new strategies for new challenges. *Lancet* **382**, 900-911 (2013). https://doi.org:10.1016/S0140-6736(13)60310-4

2 Tatem, A. J. & Smith, D. L. International population movements and regional Plasmodium falciparum malaria elimination strategies. *Proc. Natl Acad. Sci. USA.* **107**, 12222-12227 (2010). https://doi.org:10.1073/pnas.1002971107

3 Wesolowski, A. *et al.* Quantifying the impact of human mobility on malaria. *Science* **338**, 267-270 (2012). https://doi.org:10.1126/science.1223467

4 Chang, H.-H. *et al.* Mapping imported malaria in Bangladesh using parasite genetic and human mobility data. *eLife* **8**, e43481 (2019). https://doi.org:10.7554/eLife.43481

5 Neafsey, D. E., Taylor, A. R. & MacInnis, B. L. Advances and opportunities in malaria population genomics. *Nat. Rev. Genet.* **22**, 502-517 (2021). https://doi.org:10.1038/s41576-021-00349-5

6 Dalmat, R., Naughton, B., Kwan-Gett, T. S., Slyker, J. & Stuckey, E. M. Use cases for genetic epidemiology in malaria elimination. *Malar. J.* **18**, 163 (2019). https://doi.org:10.1186/s12936-019-2784-0

7 Kattenberg, J. H. *et al.* Plasmodium vivax genomic surveillance in the Peruvian Amazon with Pv AmpliSeq assay. *PLoS Negl. Trop. Dis.* **18**, e0011879 (2024). https://doi.org:10.1371/journal.pntd.0011879

8 Ibrahim, A. *et al.* Population-based genomic study of Plasmodium vivax malaria in seven Brazilian states and across South America. *Lancet Reg. Health Am.* **18**, 100420 (2023). https://doi.org:10.1016/j.lana.2022.100420

9 Safarpour, M., Cabrera-Sosa, L., Gamboa, D., Van Geertruyden, J.-P. & Delgado-Ratto, C. Detecting imported malaria infections in endemic settings using molecular surveillance: current state and challenges. *Front. Epidemiol.* **5**, 1490141 (2025).

10 Trimarsanto, H. *et al.* A molecular barcode and web-based data analysis tool to identify imported Plasmodium vivax malaria. *Commun. Biol.* **5**, 1411 (2022). https://doi.org:10.1038/s42003-022-04352-2

11 Deelder, W. *et al.* Geographical classification of malaria parasites through applying machine learning to whole genome sequence data. *Sci. Rep.* **12**, 21150 (2022). https://doi.org:10.1038/s41598-022-25568-6

12 Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primers.* **1**, 59 (2021). https://doi.org:10.1038/s43586-021-00056-9

13 Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **27**, e1608 (2018). https://doi.org:10.1002/mpr.1608

14 Scutari, M., Mackay, I. & Balding, D. Using genetic distance to infer the accuracy of genomic prediction. *PLoS Genet.* **12**, e1006288 (2016). https://doi.org:10.1371/journal.pgen.1006288

15 Safarpour, M. *et al*. Genome-wide characterization of Plasmodium vivax infections in local travelers and non-travelers from the Peruvian Amazon. *Sci. Rep.* (2026). https://doi.org:10.1038/s41598-026-80071-6

16 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301-320 (2005). https://doi.org:10.1111/j.1467-9868.2005.00503.x

17 Chang, C. C. *et al*. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015). https://doi.org:10.1186/s13742-015-0047-8

18 Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999). https://doi.org:10.1111/j.0006-341X.1999.00997.x

19 Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91 (2006). https://doi.org:10.1186/1471-2105-7-91

20 Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1-22 (2010). https://doi.org:10.18637/jss.v033.i01

21 Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432 (2015). https://doi.org:10.1371/journal.pone.0118432

22 Cingolani, P. *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92 (2012). https://doi.org:10.4161/fly.19695

23 Diez Benavente, E. *et al*. A molecular barcode to inform the geographical origin and transmission dynamics of Plasmodium vivax malaria. *PLoS Genet.* **16**, e1008576 (2020). https://doi.org:10.1371/journal.pgen.1008576

24 Delgado-Ratto, C. *et al*. Population genetics of Plasmodium vivax in the Peruvian Amazon. *PLoS Negl. Trop. Dis.* **10**, e0004376 (2016). https://doi.org:10.1371/journal.pntd.0004376

25 Bulik-Sullivan, B. K. *et al*. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291-295 (2015). https://doi.org:10.1038/ng.3211

26 Taylor, A. R. *et al*. Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. *PLoS Genet.* **13**, e1007065 (2017). https://doi.org:10.1371/journal.pgen.1007065

27 White, N. J. Determinants of relapse periodicity in Plasmodium vivax malaria. *Malar. J.***10**, 297 (2011). https://doi.org:10.1186/1475-2875-10-297

28 Ruybal-Pesántez, S. *et al*. Molecular markers for malaria genetic epidemiology: progress and pitfalls. *Trends Parasitol.* **40**, 147-163 (2024). https://doi.org:10.1016/j.pt.2023.11.006

29 Castillo, A. I., Pacheco, M. A. & Escalante, A. A. Evolution of the merozoite surface protein 7 (msp7) family in Plasmodium vivax and P. falciparum: a comparative approach. *Infect. Genet. Evol.* **50**, 7-19 (2017). https://doi.org:10.1016/j.meegid.2017.01.024

30 Cheng, C. W., Jongwutiwes, S., Putaporntip, C. & Jackson, A. P. Clinical expression and antigenic profiles of a Plasmodium vivax vaccine candidate: merozoite surface protein 7 (PvMSP-7). *Malar. J.* **18**, 197 (2019). https://doi.org:10.1186/s12936-019-2826-7

31 Koenderink, J. B., Kavishe, R. A., Rijpma, S. R. & Russel, F. G. M. The ABCs of multidrug resistance in malaria. *Trends Parasitol.* **26**, 440-446 (2010). https://doi.org:10.1016/j.pt.2010.05.002

32 Rice, B. L. *et al.* The origin and diversification of the merozoite surface protein 3 (msp3) multigene family in Plasmodium vivax and related parasites. *Mol. Phylogenet. Evol.* **78**, 172-184 (2014). https://doi.org:10.1016/j.ympev.2014.05.013

33 Bitencourt, A. R. *et al.* Antigenicity and immunogenicity of Plasmodium vivax merozoite surface protein-3. *PLoS One* **8**, e56061 (2013). https://doi.org:10.1371/journal.pone.0056061

34 Kuamsab, N. *et al.* Anti-Plasmodium vivax merozoite surface protein 3 Y (PvMSP3 Y) antibodies upon natural infection. *Sci. Rep.* **14**, 9595 (2024). https://doi.org:10.1038/s41598-024-59153-w

35 Goo, Y.-K. Vivax malaria and the potential role of the subtelomeric multigene vir superfamily. *Microorganisms* **10**, 1083 (2022). https://doi.org:10.3390/microorganisms10061083

36 Bernabeu, M. *et al.* Functional analysis of Plasmodium vivax VIR proteins reveals different subcellular localizations and cytoadherence to the ICAM-1 endothelial receptor. *Cell. Microbiol.* **14**, 386-400 (2012). https://doi.org:10.1111/j.1462-5822.2011.01726.x

37 Lyons, F. M. T., Gabriela, M., Tham, W.-H. & Dietrich, M. H. Plasmodium 6-cysteine proteins: functional diversity, transmission-blocking antibodies and structural scaffolds. *Frontiers in Front. Cell. Infect. Microbiol.* **12**, 945924 (2022). https://doi.org:10.3389/fcimb.2022.945924

38 Hostetler, J. B. *et al.* A library of Plasmodium vivax recombinant merozoite proteins reveals new vaccine candidates and protein-protein interactions. *PLoS Negl. Trop. Dis.* **9**, e0004264 (2015). https://doi.org:10.1371/journal.pntd.0004264

39 Wang, Z. *et al.* Genome-wide association analysis identifies genetic loci associated with resistance to multiple antimalarials in Plasmodium falciparum from China-Myanmar border. *Sci. Rep.* **6**, 33891 (2016). https://doi.org:10.1038/srep33891

40 Jallow, M. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat. Genet.* **41**, 657-665 (2009). https://doi.org:10.1038/ng.388

41 Bengio, Y. & Grandvalet, Y. No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.* **5**, 1089-1105 (2004).

42 LeDell, E., Petersen, M. & van der Laan, M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron. J. Stat.* **9**, 1583-1607 (2015). https://doi.org:10.1214/15-EJS1035

Acknowledgements

We thank the study participants in Peru and the UPCH team for acquiring the samples, and L. Cabrera-Sosa in particular for organising their preparation and shipment to Antwerp. We thank A. Rosanas-Urgell and E. Kattenberg for their support with the experimental work, and the Institute of Tropical Medicine, Antwerp, for the laboratory facilities. This work was supported by the Research Foundation - Flanders (FWO) through the project "Unraveling the contribution of Plasmodium vivax metapopulation to the persistence of malaria transmission in residual areas" (MetaVivax; G0A4222N to J.-P.V.G. and C.D.-R.). Research reported in this publication was also supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number U19AI089681 (to J.M.V. and D.G.), which funded sample collection and field activities. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The computational resources and services used in this work were provided by the HPC core facility CalcUA of the Universiteit Antwerpen, and VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

Author contributions

M.S. developed the study concept and analytical framework, performed the analyses and wrote the original draft. K.V.S. contributed to the study concept and advised refinements to the analytical framework and additional analyses. D.G. and J.M.V. led the acquisition of the samples and epidemiological data in Peru. C.D.-R. conceived the study, and C.D.-R. and J.-P.V.G. supervised it. K.V.S. and C.D.-R. provided scientific feedback on the manuscript. J.-P.V.G., C.D.-R., J.M.V., and D.G. obtained funding. All authors revised the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the Research Foundation – Flanders (FWO) through the project "Unraveling the contribution of Plasmodium vivax metapopulation to the persistence of malaria transmission in residual areas" (MetaVivax; G0A4222N to J.-P.V.G. and C.D.-R.). Research reported in this publication was also supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number U19AI089681 (to J.M.V. and D.G.), which funded sample collection and field activities. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

Figures (Figs. 1 - 8)

Fig. 1. Genome-wide association results comparing *P. vivax* isolates between the communities Libertad and Gamitanacocha. Results are shown for all samples including travelers and non-travelers (A), samples excluding travelers (B), and traveler-associated infections only (C). Each point represents a SNP, plotted against the $-\log_{10}(p\text{-value})$ by chromosomal position. Alternating colors (blue and orange) distinguish different chromosomes. The horizontal line indicates the reported cutoff of $-\log_{10}(p\text{-value}) = 5$.

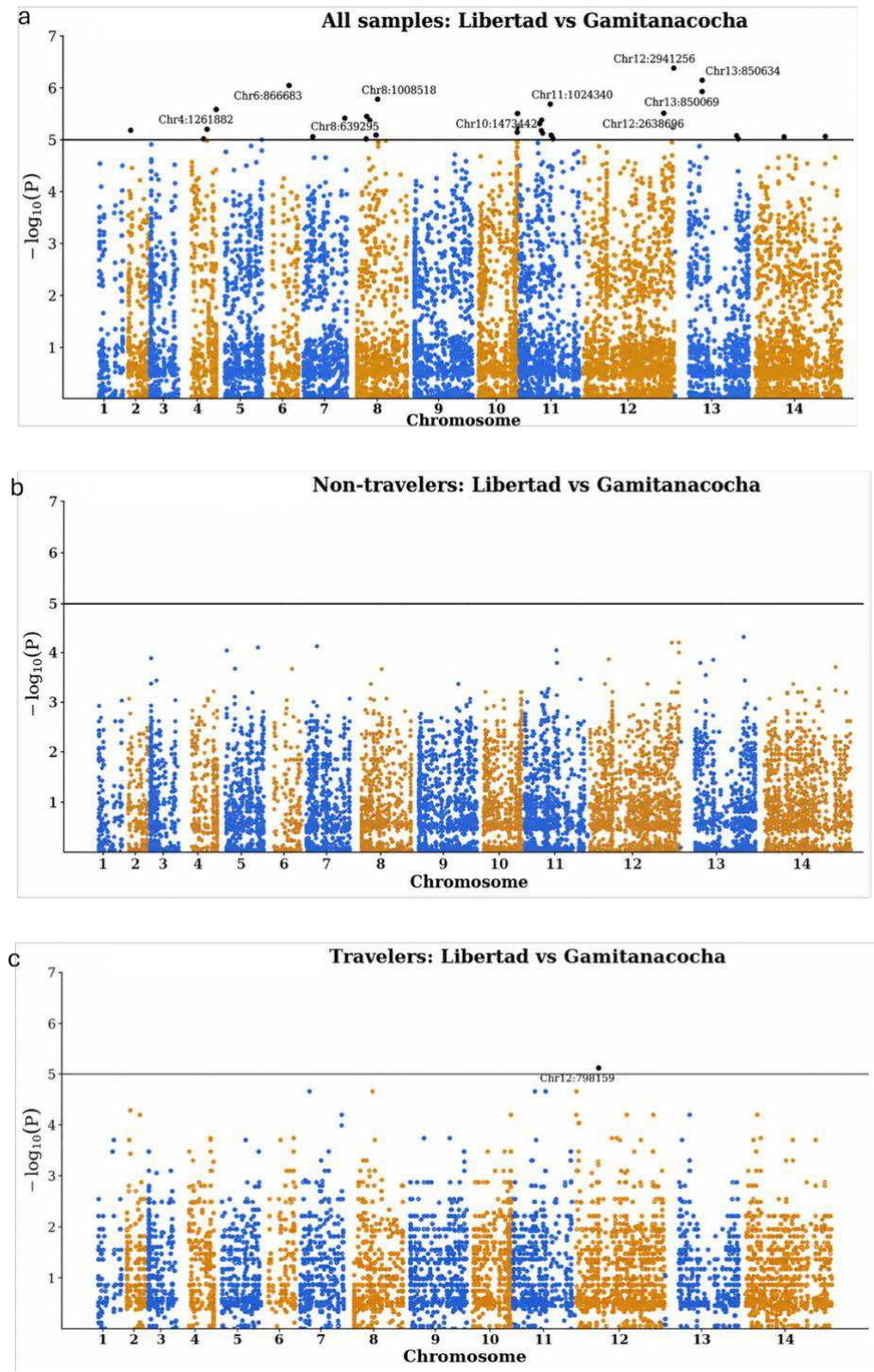


Fig. 2. Receiver operating characteristic (ROC) curves comparing model performance. ROC curves showing the performance of prediction models trained on all samples (travelers and non-travelers) and non-travelers only.

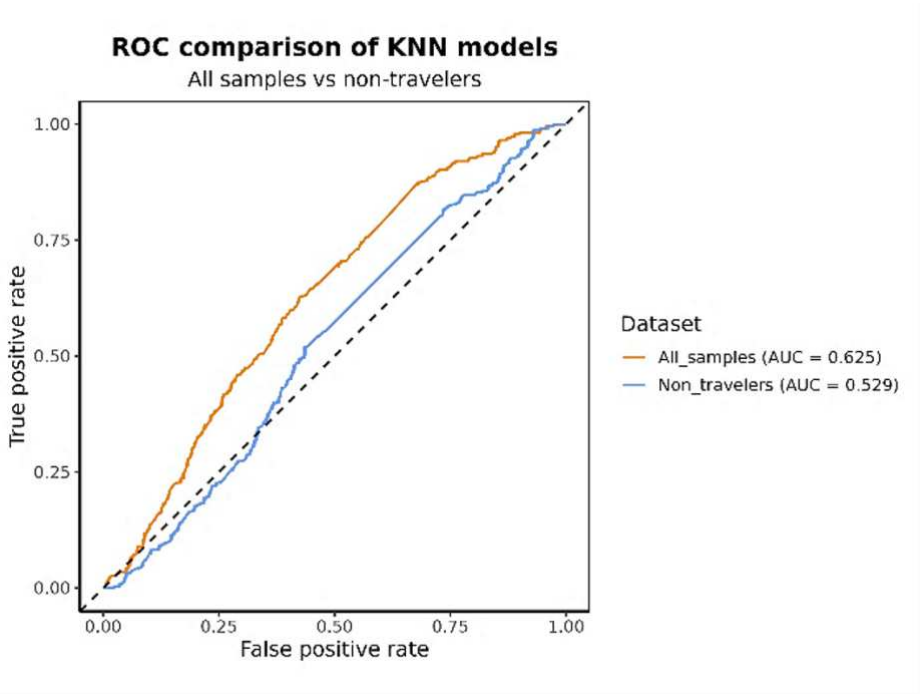


Fig. 3. Predicted probabilities by community. The top panel (a and b) show predicted probabilities for the model developed using all samples (travelers and non-travelers), including a boxplot and the corresponding distribution of predicted probabilities. The bottom panels (c and d) show the same analyses for the model restricted to non-travelers only. In all panels, the predicted probabilities represent the likelihood of belonging to Gamitanacocha (class 1), enabling comparison of class separation between the two modeling approaches.

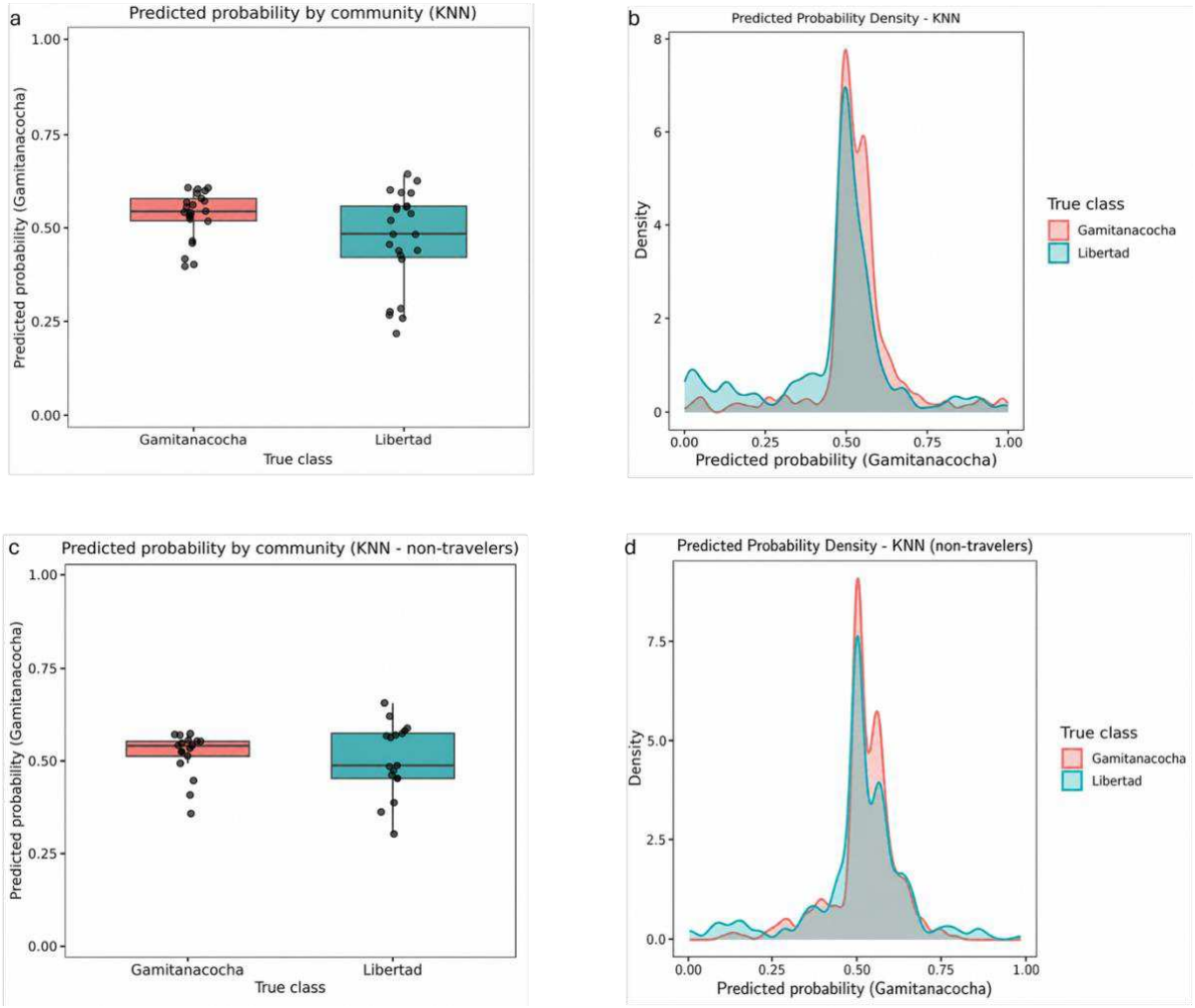


Fig. 4. Predicted probabilities by travel history and community for the full model. The left panel (A) shows sample-level predicted probabilities stratified by travel history within each community (Gamitanacocha and Libertad), while the right panel (B) summarizes these patterns using boxplots across all samples. Predicted probabilities correspond to the likelihood of belonging to Gamitanacocha (class 1).

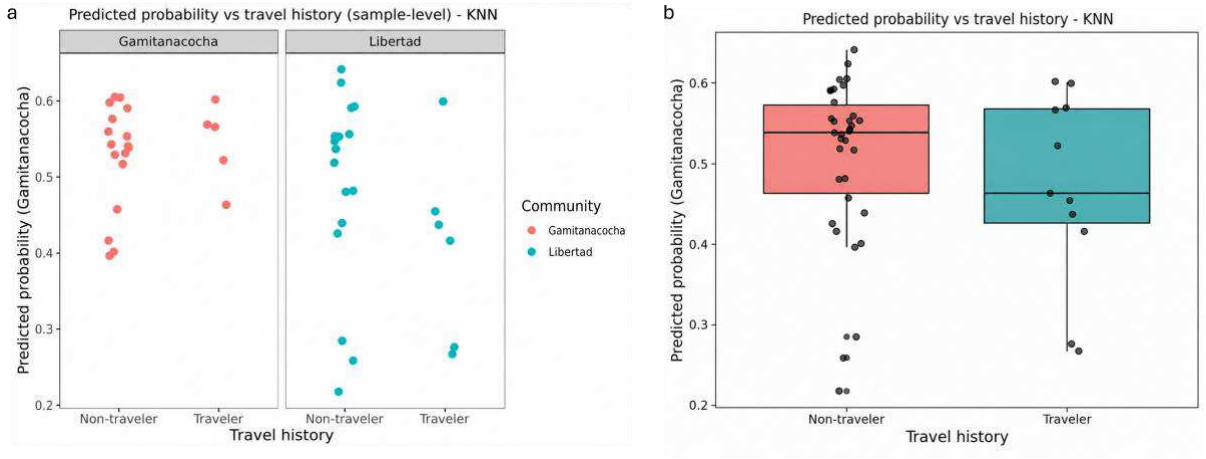


Fig. 5. Top-ranked SNPs identified by the predictive models. The top panels show the top 30 SNPs identified by the full model. The left panel (A) ranks SNPs by their selection frequency, highlighting variants most consistently selected by the model. The right panel (B) ranks SNPs by mean penalized coefficient across 100 iterations, including the 15 most positive (blue) and 15 most negative (red) coefficients, reflecting their contribution to classifying infections (with Gamitanacocha as class 1). The bottom panels show the corresponding results for the non-traveler model: selection frequency (C) and mean penalized coefficient (D).

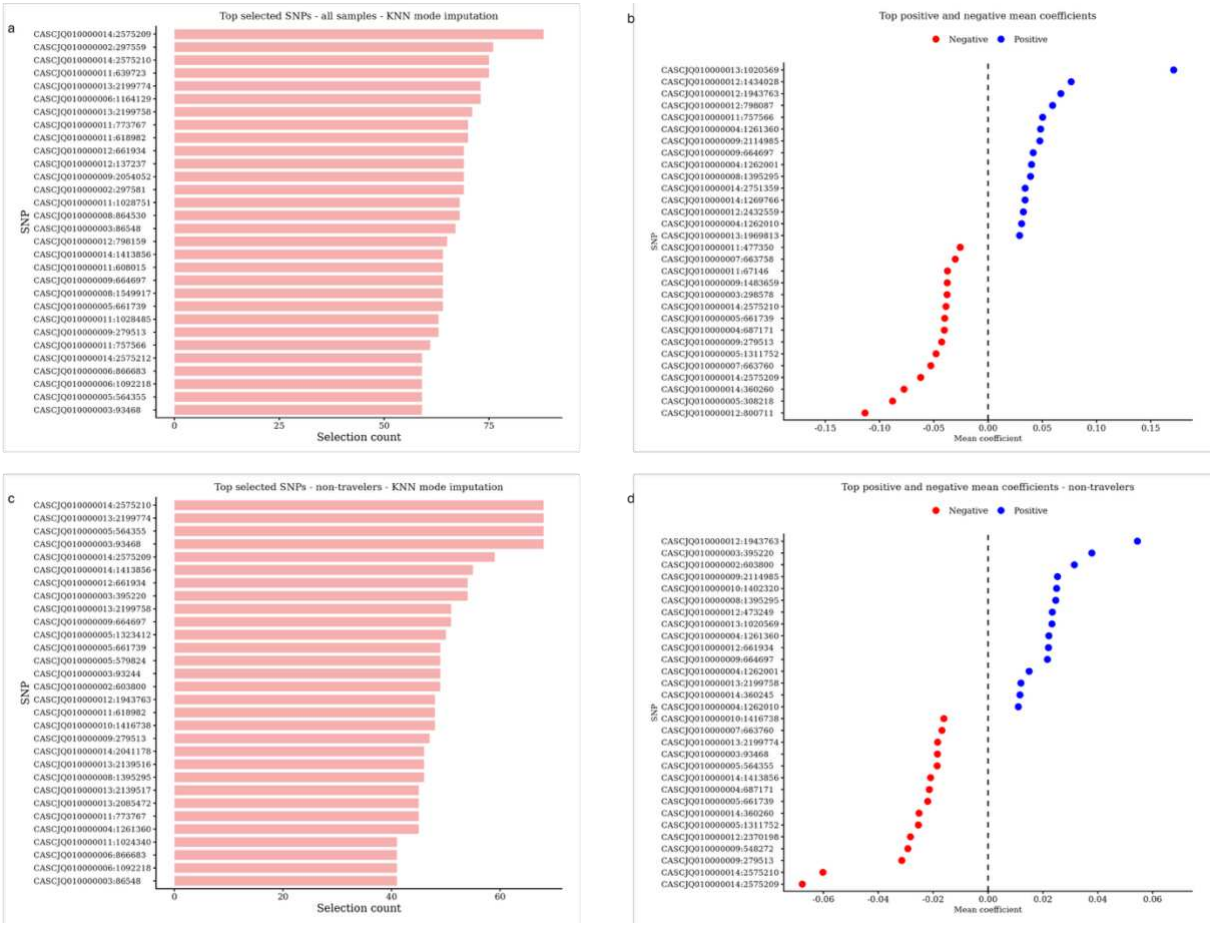


Fig. 6. Gene-level annotation of SNPs specific to the all-samples (full model) top coefficient-ranked sets. Genes are shown on the y-axis and coefficient direction is shown on the x-axis: positive coefficients indicate classification toward Gamitanacocha, and negative coefficients indicate classification toward Libertad. Bubble size represents the absolute mean coefficient, summarizing coefficient magnitude, while color indicates the predicted snpEff impact.

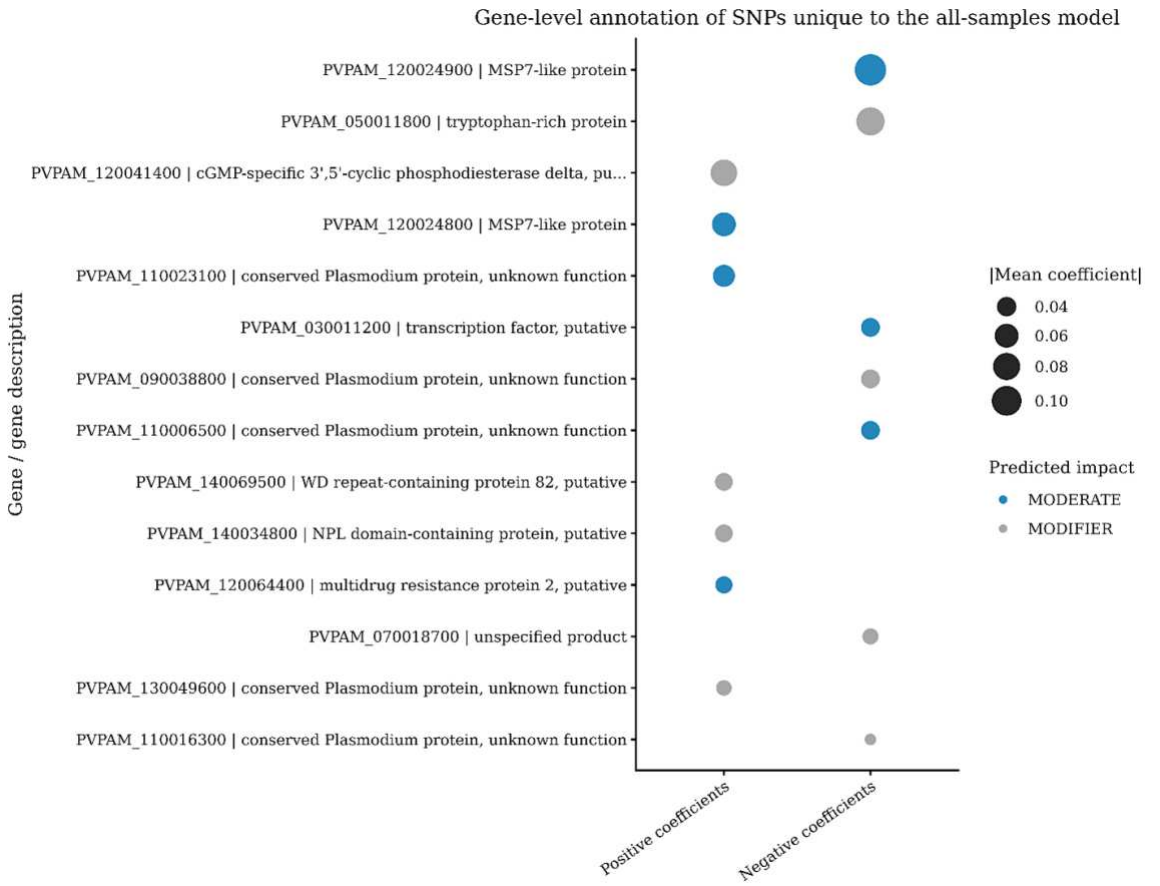


Fig. 7. Gene-level annotation of SNPs specific to the non-traveler top coefficient-ranked sets. Genes are shown on the y-axis and coefficient direction is shown on the x-axis: positive coefficients indicate classification toward Gamitanacocha, and negative coefficients indicate classification toward Libertad. Bubble size represents the absolute mean coefficient, summarizing coefficient magnitude, while color indicates the predicted snpEff impact.

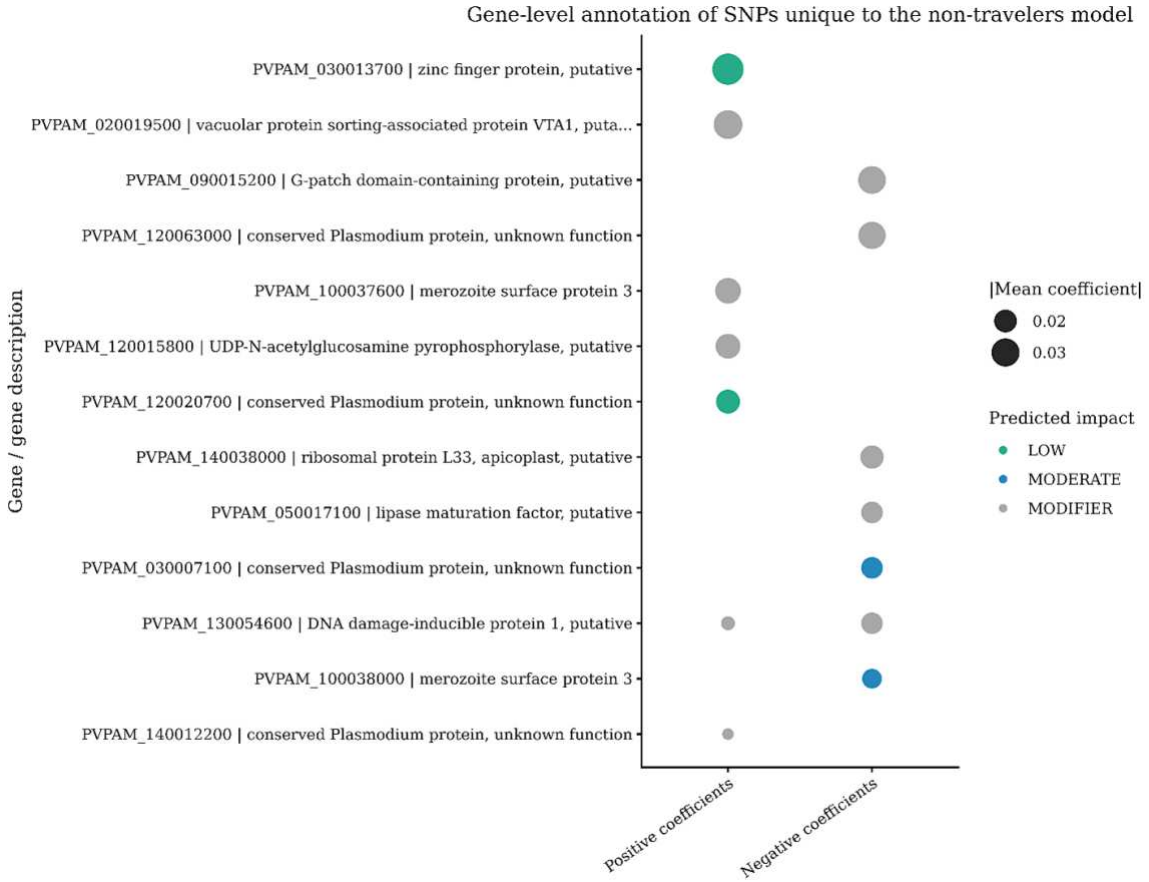
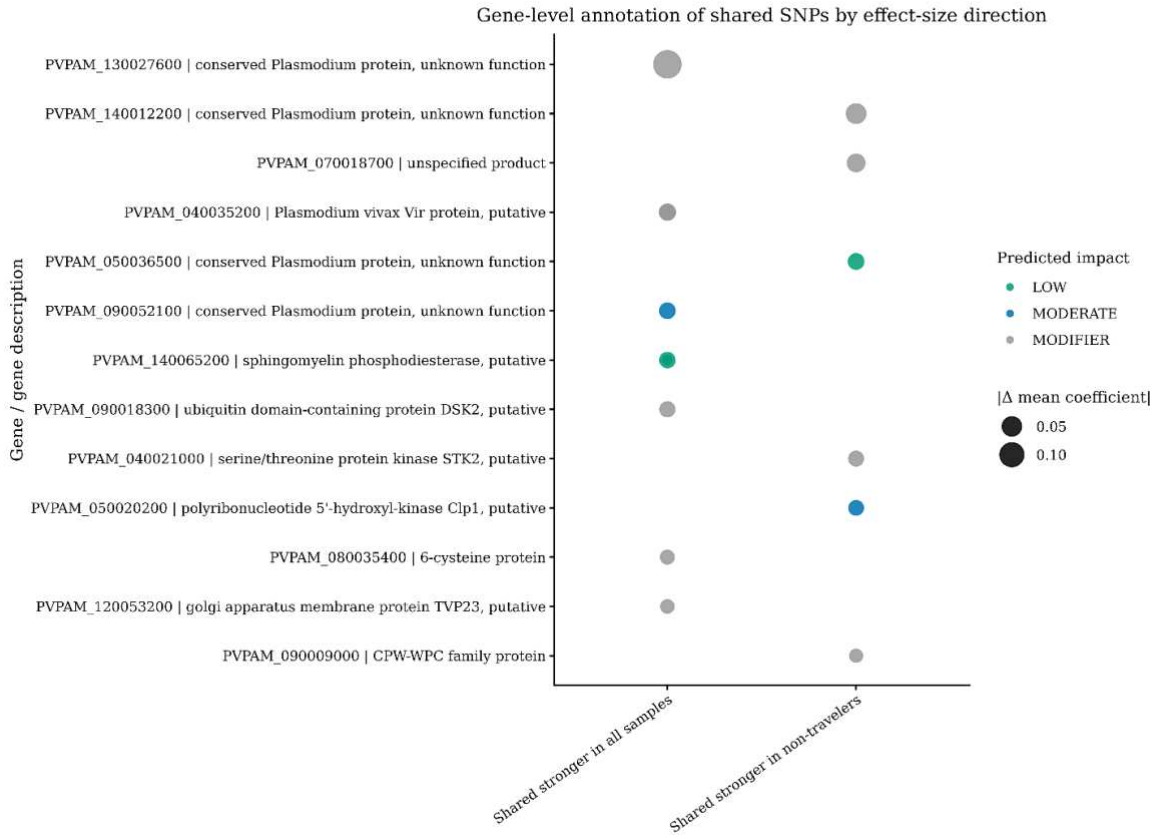


Fig. 8. Gene-level annotation of shared SNPs grouped by the signed difference in mean coefficients (Δ = all samples minus non-travelers). Genes are shown on the y-axis. The x-axis separates SNPs with more positive coefficients in the full model (Δ > 0) from those with more positive coefficients in the non-traveler model (Δ < 0). Bubble size represents |Δ|, and color indicates predicted snpEff impact.



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.pdf