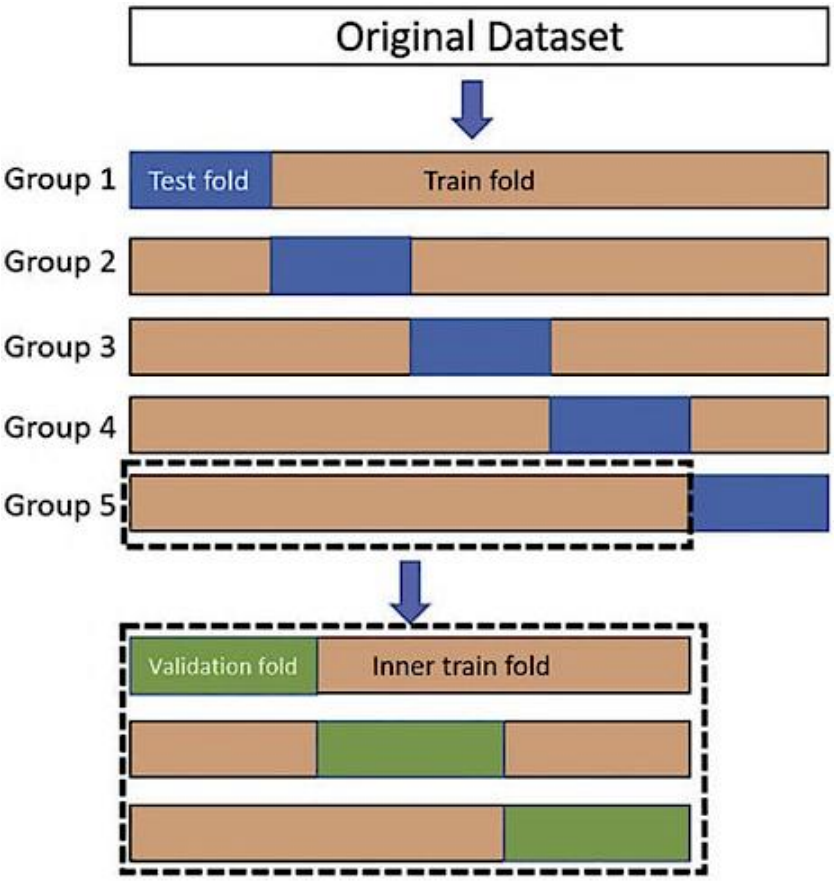


Figure S1. Nested cross-validation framework for developing an Elastic Net model to predict the origin of infection. Schematic representation of the analytical pipeline used to develop a prediction model classifying infections as originating from Libertad or Gamitanacocha communities. The original dataset is partitioned into multiple outer folds: in each iteration, one fold (blue) is held out as the test set, and the remaining folds (brown) are used for model training. Within each training set, an inner cross-validation procedure (dashed box) is used to optimize model hyperparameters, including the Elastic Net parameters α (mixing parameter) and λ (regularization strength), as well as K, the number of neighbors used in the k-nearest neighbors (KNN) imputation of missing values. The inner loop splits the data into inner training folds (brown) and validation folds (green), and model performance is evaluated across combinations of these parameters to identify the optimal set. The tuned model is then applied to the outer test fold, and performance is assessed using the area under the receiver operating characteristic curve (AUC). This process is repeated across all folds and iterations (20 iterations × 5 folds), yielding 100 AUC estimates that are used to summarize overall model performance. Two models were developed using this framework: one including all samples (travelers and local infections), and one restricted to local infections only.

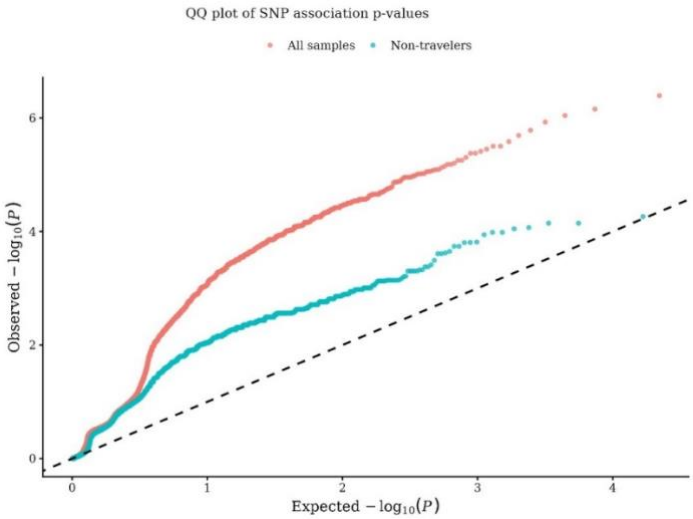


Figure S2. Quantile–quantile (Q–Q) plot of SNP association p-values. Each point represents a SNP. The red points correspond to all samples (including travelers), whereas the blue points represent non-travelers (local infections only). The dashed diagonal line indicates the expected null distribution where observed and expected p-values are equal.

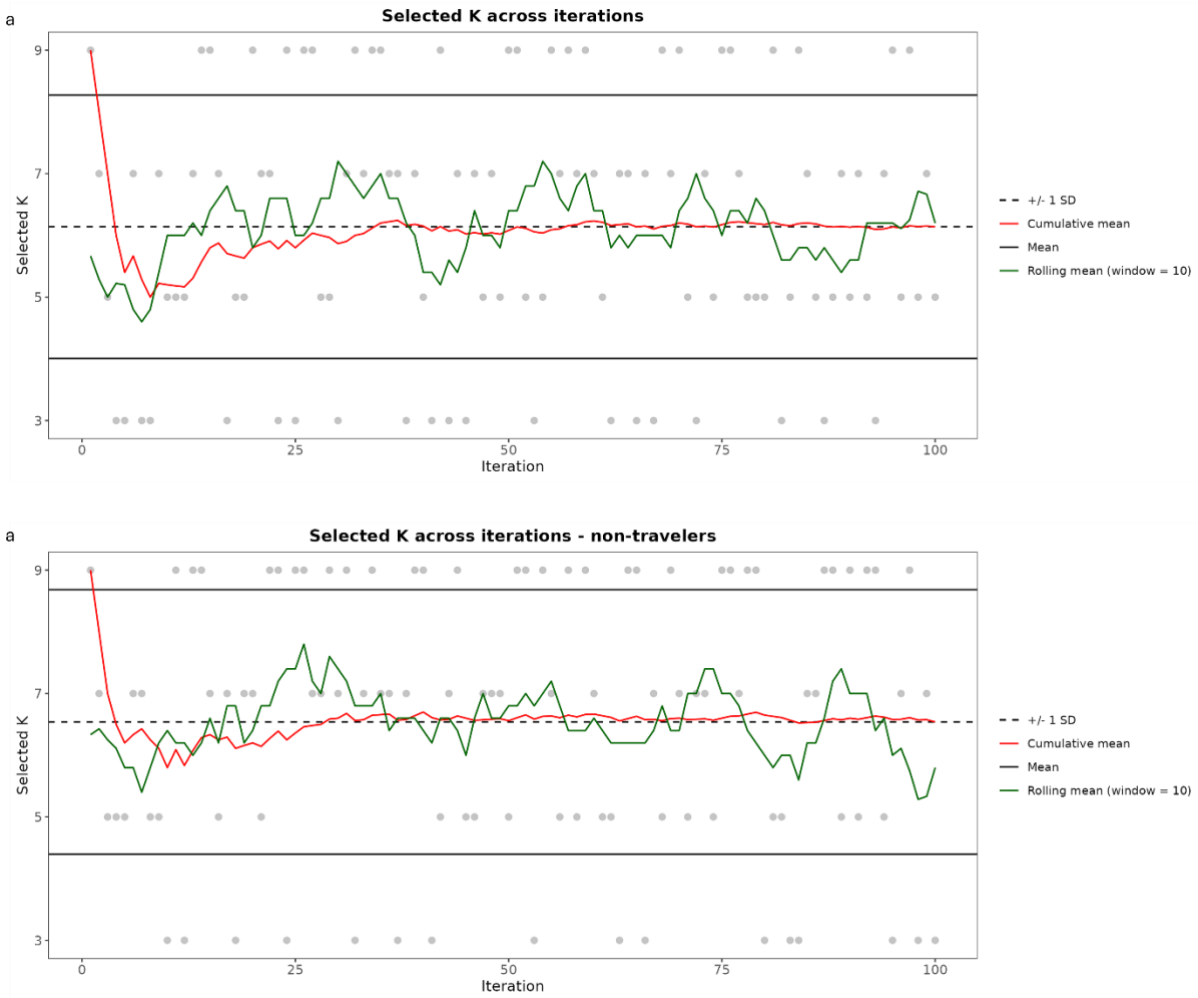


Figure S3. Stability of selected K across iterations. Plot showing the selected number of neighbors (K) across iterations of the modeling process. Each point represents the optimal K chosen in each iteration. The red line indicates the cumulative mean, the solid black line represents the overall mean, and the green line shows the rolling mean (window = 10). The dashed

horizontal lines denote the mean ± 1 standard deviation. Results are shown for the full model (panel a) and for the model restricted to non-travelers (panel b).

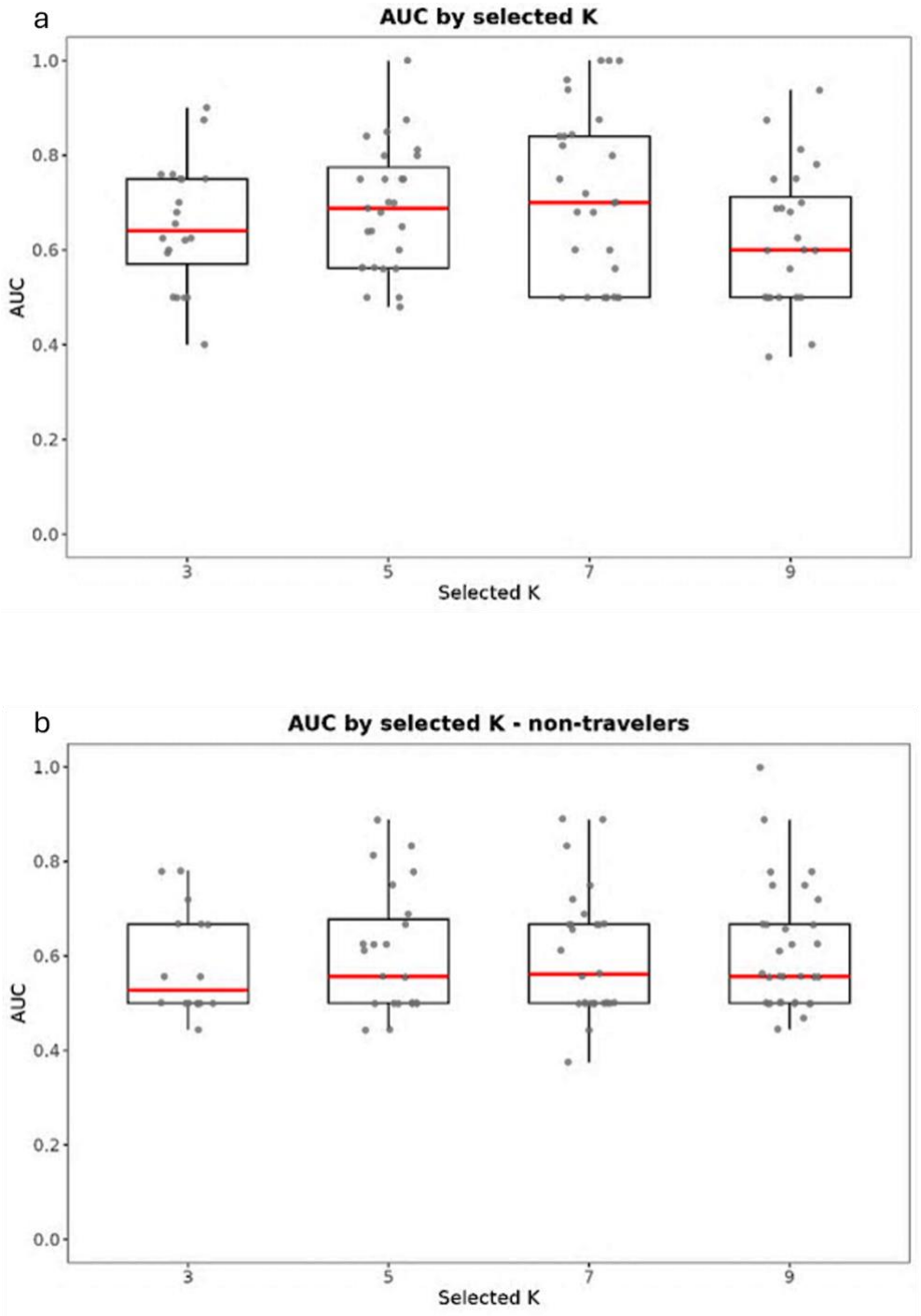


Figure S4. Model performance (AUC) by selected K. Boxplots show the distribution of AUC values across 100 iterations for different selected values of K. Each point represents the AUC obtained in a given iteration, and the red horizontal line within each box indicates the median AUC. Results are shown for the full model (a) and for the model restricted to non-travelers (b).

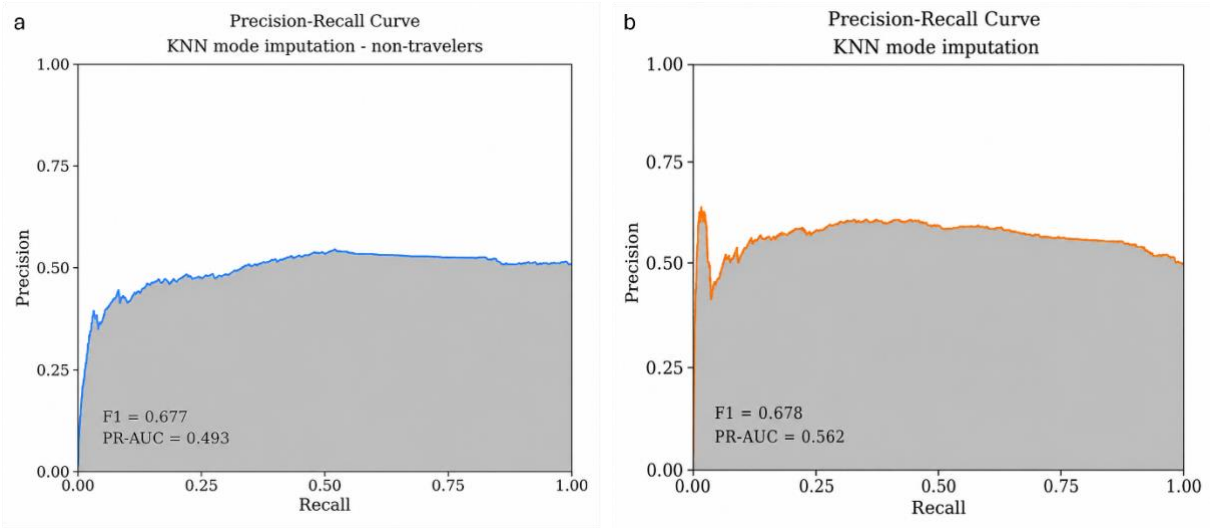


Figure S5. Precision–recall curves. The blue curve (a) shows the performance of the model trained on local infections (only non-travelers), while the orange curve (b) shows the performance of the model trained on all samples, including travelers (the full model).

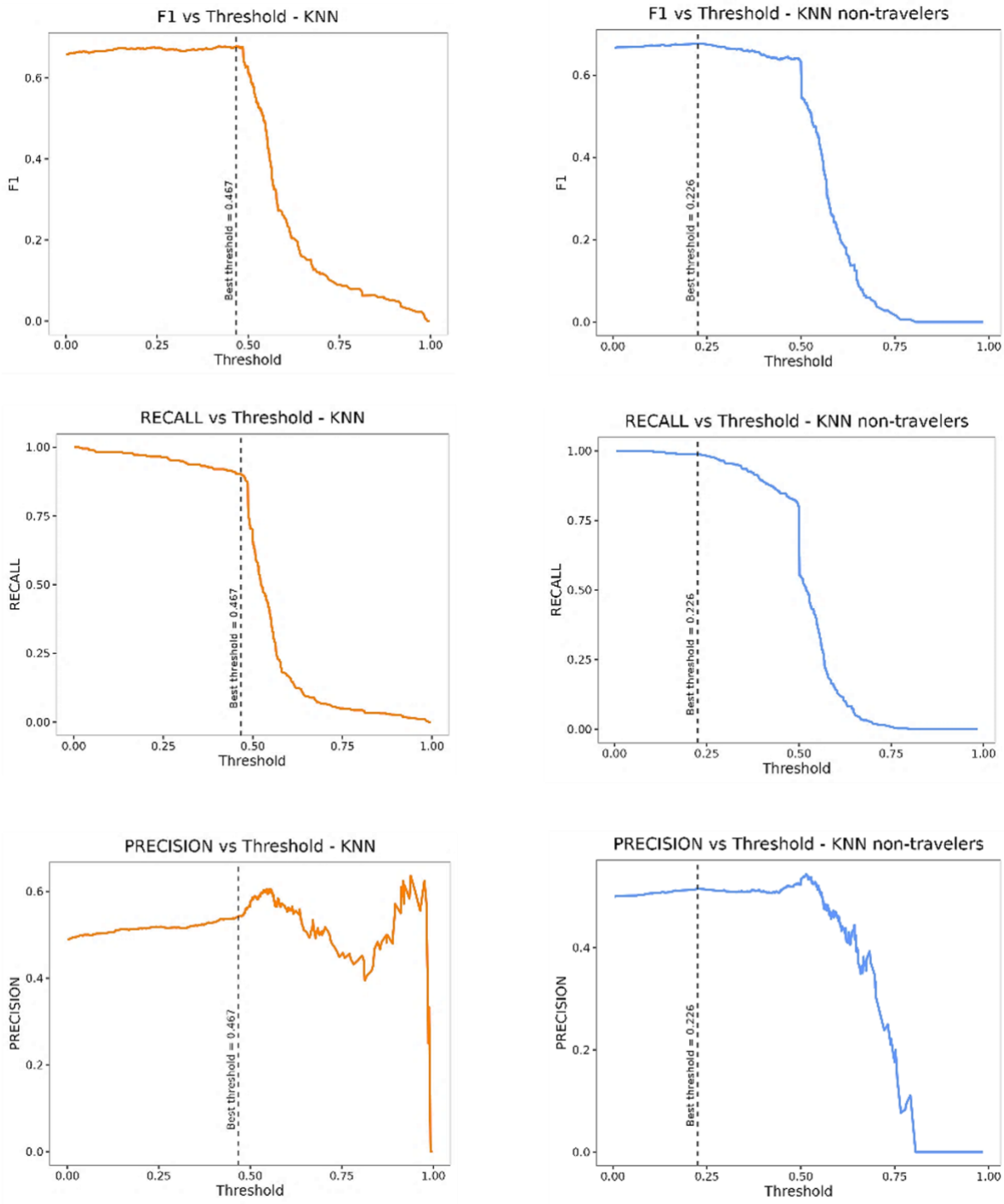


Figure S6. Performance metrics across classification thresholds. The top panels (blue) show F1-score, recall, and precision as functions of the classification threshold for the model developed using non-travelers only, while the bottom panels (orange) show the corresponding metrics for the full model (all samples: travelers and local infections). The vertical dashed line in each panel indicates the selected optimal threshold used for classification.

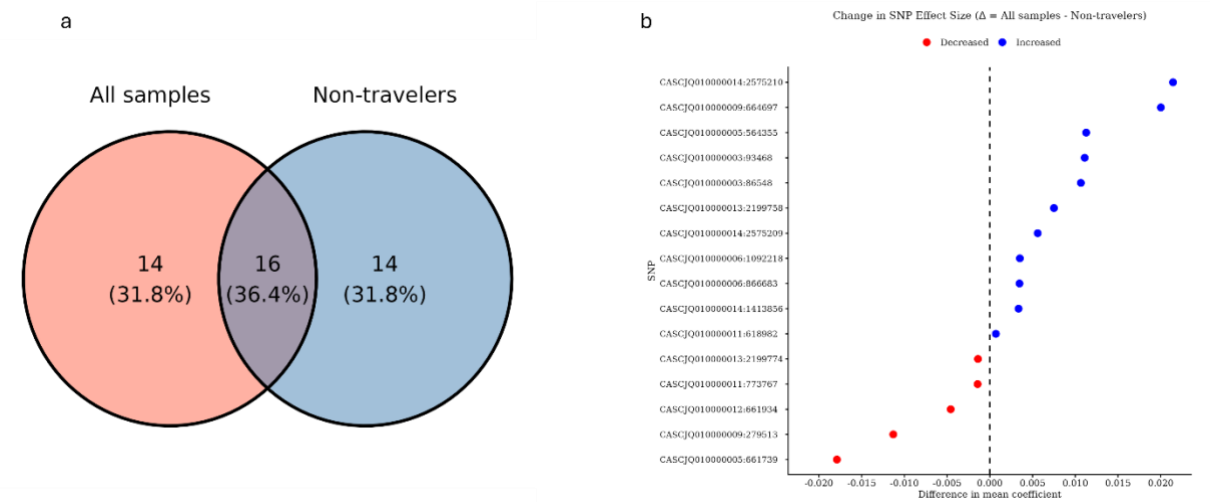


Figure S7. Comparison of selected SNPs between the full and non-traveler models. The left panel (a) shows the overlap between the top 30 SNPs ranked by selection frequency in each model. The right panel (b) presents the difference in mean coefficient (Δ) for all 16 shared SNPs, calculated as the mean coefficient in full model minus that in non-travelers. Blue points indicate SNPs with stronger effects in the full dataset, while red points indicate stronger effects in non-travelers.

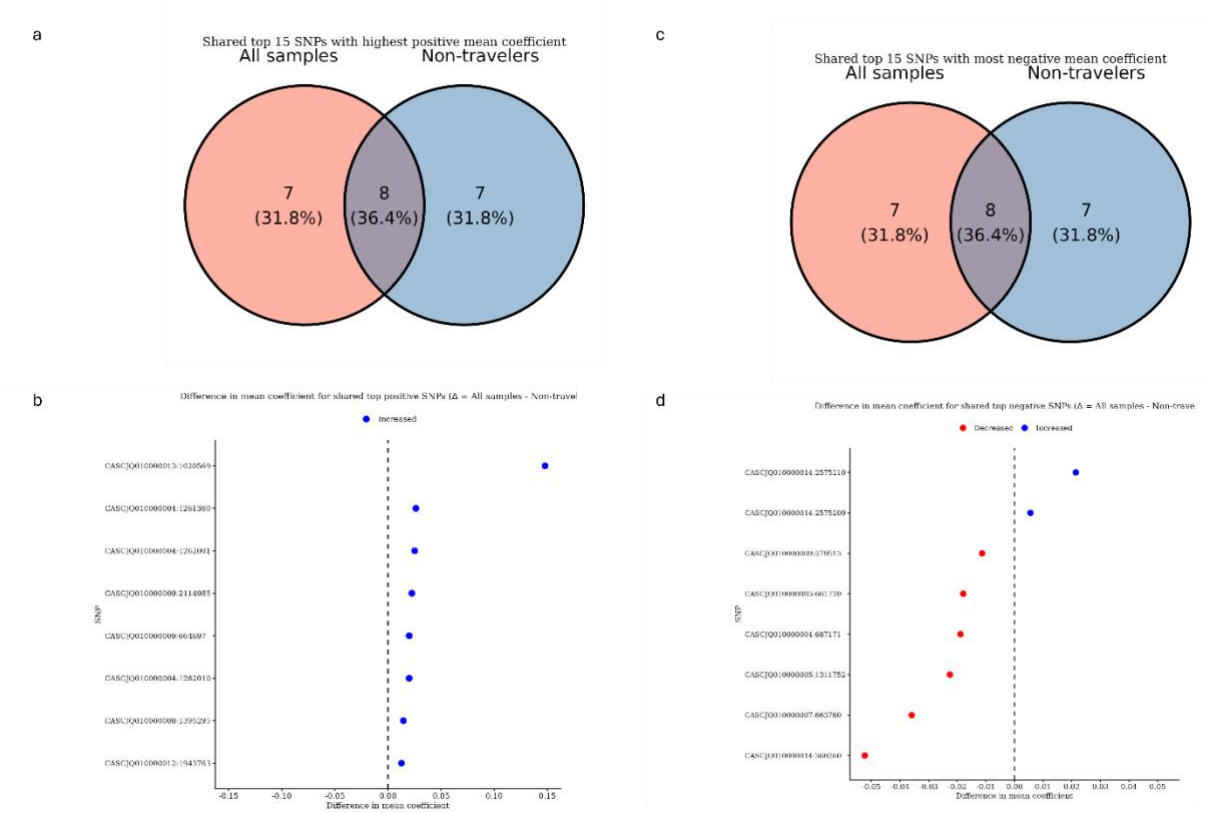


Figure S8. Comparison of SNPs based on effect size between all samples and non-travelers. Panel a shows the overlap of the top 15 SNPs with the highest positive mean coefficients, while Panel c shows the overlap of the top 15 SNPs with the most negative mean coefficients. Panels b and d show the difference in mean coefficient for shared positive- and negative-effect SNPs, respectively, calculated as all samples minus non-travelers. Blue points indicate stronger effects in the full dataset, while red points indicate stronger effects in non-travelers.