Integrative Transcriptomic and Machine Learning Approaches to decipher Mitochondrial Gene Regulation in severe *Plasmodium vivax* Malaria

Priyanka Roy[1], Yash Aggarwal[1], Sanjay Kumar Kochar[2], Dhanpat Kumar Kochar[2], Ashis Das[1*]
[1]Molecular Parasitology and Systems Biology Laboratory, Department of Biological Sciences, Birla Institute of Technology and Science (BITS), Pilani campus, Rajasthan, India
[2]Department of Medicine, S.P. Medical College, Bikaner, Rajasthan, India
[*]Correspondence: Ashis Das adas@pilani.bits-pilani.ac.in, ashisd28@gmail.com

Abstract

Mitochondria in *Plasmodium vivax* are functionally vital despite possessing a highly reduced genome and differing substantially from the human organelle. Beyond their classical role in energy production, they dynamically coordinate processes like pyrimidine biosynthesis and heme metabolism, adapting their functions across the intra-erythrocytic development cycle (IDC). Their unique architecture and stage-specific roles enable the parasite to fine-tune mitochondrial gene expression, which operationally includes both sense and Natural Antisense Transcripts (NATs) - a class of long non-coding RNAs. The study involves the analysis of transcriptomic data to identify significant differentially expressed genes, in both sense and NATs categories, associated with severe malaria manifestations. This emphasizes the critical role of mitochondrial gene regulation in disease severity. These genes were statistically ranked and then used as input features for machine learning analysis for verification. Machine learning acted as a hypothesis-testing framework, enabling refinement of gene lists and strengthening biological interpretations. Further, a comprehensive gene enrichment analysis was performed for both sense and NATs to investigate the mitochondrial or other cellular pathways impacted during severe malaria. The findings revealed that NATs have a striking association with mitochondrial pathways and translation machinery, indicating that NATs are not merely by-products of transcription but also play a regulatory role in fine-tuning mitochondrial gene expression with severe manifestation. This work highlights mitochondrial NATs as critical regulators of parasite biology and positions the *Plasmodium* mitochondrion as a promising target for antimalarial drug development and therapeutic intervention.

Keywords: Mitochondria, *Plasmodium vivax*, Transcriptome, Machine Learning, NATs

Introduction

Malaria, a significant global health challenge, is caused by protozoan parasites of the genus *Plasmodium*, with *Plasmodium vivax* being one of the most widespread and prevalent species. According to the World Health Organization (WHO) 2024 report, an estimated 263 million malaria cases were recorded globally in 2023. Strikingly, *Plasmodium vivax* represented approximately 48.2% of malaria infections in the South-East Asia region, with India emerging as a major contributor, accounting for nearly 25% of the global *P. vivax* burden [1] , highlighting its substantial burden on public health. Historically, *P. vivax* was thought to cause benign tertian malaria, however many studies have reported that it causes significant severity & morbidity, where changes in gene expression plays a major role on the underlying mechanisms contributing to this shift. Transcriptional analyses have shown that variations in gene expression profiles can affect the parasite's ability to invade erythrocytes and modulate host immune responses [2,3].

Most eukaryotic genomes produce non-coding RNAs (ncRNAs), which integrate complex suites of gene activity and contribute to the complexity of gene regulation [4]. The lncRNA are a subclasses of non-coding RNAs (ncRNAs) that are longer than 200 nt and are typically distinguished by their inability to code for proteins. They can be categorized as intergenic, intronic, sense lncRNAs, transcribed from the sense strand of protein-coding genes, and Natural antisense transcripts (NATs), transcribed from the antisense strand of protein-coding genes, based on the genomic area of transcription [4]. The lncRNA can modulate gene expression at numerous levels, including transcription, post-transcription, RNA turnover and protein translation [6,7].

The presence of ncRNA in *Plasmodium* has the potential to regulate key events in the complex life cycle and modulate the virulence ability of the parasite, by controlling transcriptional and post-transcriptional processes [8]. Patankar et al.,2001 & Militello et al., 2005 reported the presence of natural antisense transcripts (NATs) in *Plasmodium* parasite and their role in gene regulation [9,10]. Additionally, the presence of NATs in gametocytes and ookinetes implies that these antisense RNAs could be essential for the control of gene expression and parasite development [11]. Studies from our laboratory have showed the presence of NATs using a custom designed strand specific whole genome microarray in *P. vivax* from patients showing complicated and uncomplicated disease manifestations [12].

Plasmodium mitochondria role is particularly fascinating because its function varies depending on the stage of the parasite's lifecycle. It has a minimalistic genome encoding only three proteins cytc oxidase subunit 1 (cox1) and subunit 3 (cox3), members of the cytc oxidase complex (complex IV), and cytb (cob), a member of cytochrome bc1 complex. Despite the reduced genome, *Plasmodium* mitochondria play crucial roles in metabolism, including pyrimidine biosynthesis and electron transport chain in the asexual blood stages [13, 14]. Apart from this, there is a functional citric acid cycle within the mitochondrion which is responsible for the de novo biosynthesis of heme. Despite the widespread haemoglobin breakdown in the digestive vacuole, heme production in *Plasmodium* is divided between the mitochondrion, the apicoplast, and the cytosol [13,14,15,16].

Despite, mitochondria having their own genome with transcriptional and translational machinery, majority of the proteins found in mitochondria are encoded in the nucleus due to a huge endosymbiotic gene transfer from the mitochondrial to the nuclear genome [17]. There are possible energy requirement differences between the asexual intra-erythrocytic developmental stages and the sexual stages such as gametocytes. In asexual stages, no cristae are visible in the mitochondrial membrane while in the mitochondria of gametocytes, crista-like structures accumulate in the respiratory chain complexes [18, 19]. Another interesting observation remains that there is a significantly lower level of Complex V components compared to the sexual stage of the parasites [19].

Even though the *Plasmodium* mitochondria has limited number of metabolic process & pathways, mitochondrial gene products and their expression are functionally vital, with stage-specific remodelling evident between asexual and gametocyte stages, highlighting their fundamental contribution to parasite viability and transmission potential. A wide range of drugs are designed to target mitochondrial pathways, but the emergence of resistance has become a significant concern during their development. The relaxed selective pressure for mitochondrial electron transport chain function during the blood stage, along with mitochondrial genome polyploidy seems to accelerate the emergence of atovaquone-resistance mutations [20].

Based on the data in this study, the first comprehensive evidence of mitochondrial transcripts occurring in both sense and NATs orientation in *P. vivax* derived from clinical isolates exhibiting different disease manifestation, is presented. Through integrative transcriptomic profiling and machine learning models, we uncovered key insights into mitochondrial gene regulation mediated by NATs, particularly in the context of severe disease manifestation. Notably, we observed distinct differential expression patterns of NATs across clinical conditions, with a strong association to mitochondrial pathways and the translational machinery. These findings suggest that NATs are not random transcriptional by-products but potentially function as regulatory elements that fine-tune mitochondrial gene expression during pathogenesis. Altogether, in correlation with the gene expression signatures, the mitochondrial NATs in *Plasmodium* may provide alternative targets for antimalarial drug development and disease intervention strategies.

Materials & Methods

Curating gene list with potential mitochondrial localization

To employ a systematic approach to compile a comprehensive list of *P. vivax* mitochondrial genes, a multi-layered bioinformatic strategy was employed, integrating web-based prediction tools with functional annotations. Given the absence of a universally conserved mitochondrial localization signal in *Plasmodium* species and the limitations of individual prediction tools—each trained on distinct datasets and algorithms—multiple complementary platforms were utilized, following the recommendations of Martelli et al., 2021 [21]. The four primary tools included: TargetP, which identifies mitochondrial transit peptides (mTPs) characterized by enrichment of arginine, leucine, and serine residues critical for matrix targeting [22]; Cello2GO, which leverages BLAST-based homology to assign subcellular localization [23]; MitoFates, which predicts mitochondrial pre-sequences using classical targeting features and motif analyses [24]; and Mitoprot II, a tool that employs discriminant analysis of amino acid composition, charge, hydrophobicity, and amphiphilic potential to identify mitochondrially imported proteins [25]. Additionally, the analysis was strengthened by incorporating genes annotated with mitochondrial targeting domains in PlasmoDB [26], alongside orthologous genes that have been experimentally validated as part of the *Plasmodium falciparum* mitochondrial proteome [27–31]. This combined approach ensured the inclusion of both predicted and empirically confirmed mitochondrial candidates. To enhance prediction robustness, only genes supported by at least two or more tools were retained, ensuring a high-confidence catalog of nuclear-encoded, mitochondria-targeted candidates in *P. vivax*.

Blood sample collection & processing

Venous blood samples were obtained from patients diagnosed with *Plasmodium vivax* infection at S.P. Medical College, Bikaner, India. To enrich for infected erythrocytes, the collected blood was processed using a Histopaque density gradient. All patient samples were collected by a dedicated team of clinicians under ethical approval, with informed consent obtained in accordance with hospital guidelines.The presence of *P. vivax* infection was initially confirmed by slide microscopy and rapid diagnostic tests (RDTs), followed by PCR-based validation to ensure mono-infection [32–33]. The patient's diagnosis was confirmed as hepatic dysfunction as per WHO guidelines [34] by the team of clinicians. Detailed clinical characteristics of the enrolled patients are presented in Table S1. For transcriptomic profiling, total RNA from each sample was individually hybridized onto custom-designed 8×60K single-color microarrays.

Microarray Hybridization

Total RNA was extracted from patient blood samples following established protocols [12]. The concentration and purity of RNA were assessed using a NanoDrop 2000 spectrophotometer (Thermo Scientific) and Qubit fluorometric assay (Thermo Scientific, USA). RNA integrity was evaluated on an Agilent TapeStation system to ensure high-quality input for downstream analysis. Microarray labeling and hybridization were carried out at the Agilent-certified facility of Genotypic Technology Pvt. Ltd., Bengaluru, India. RNA samples were processed using the Agilent Quick-Amp Labeling Kit (p/n 5190-0442). In this process, total RNA was reverse transcribed at 40°C using an oligo-dT primer linked to a T7 polymerase promoter to synthesize first-strand cDNA. This was followed by second-strand synthesis to generate double-stranded cDNA, which served as a template for in vitro transcription. During in vitro transcription, cyanine-3 (Cy3) labeled cRNA was produced using Agilent Cy3 CTP. All enzymatic reactions, including cDNA synthesis and transcription, were performed at 40°C. The labeled cRNA was purified using Qiagen RNeasy columns (Cat. No. 74106), and quality metrics—yield and specific activity—were confirmed using the NanoDrop ND-1000. Prior to hybridization, labeled cRNA was fragmented at 60°C and applied to Agilent Custom *P. vivax* 8×60K Gene Expression Microarrays. Hybridization was carried out using the Agilent Gene Expression Hybridization Kit (Part No. 5190-0404) in SureHyb chambers at 65°C for 16 hours. Post-hybridization washes were performed using Agilent Gene Expression Wash Buffers (Part No. 5188-5327). Microarray slides were scanned with the Agilent Microarray Scanner (Model G2600D), and raw data were extracted using Agilent Feature Extraction software from images for further analysis.

Microarray design

A custom 60K sense-antisense microarray was meticulously designed to enable comprehensive transcriptomic profiling of *Plasmodium vivax*, with a particular emphasis on mitochondrial gene regulation. The array was constructed using genomic annotations from the *P. vivax* Sal-1 reference genome PlasmoDB v47 (www.plasmodb.org), incorporating both nuclear and mitochondrial open reading frames (ORFs). Each ORF was represented by five distinct oligonucleotide probes (60–132 mers) in both sense and antisense orientations. Additionally, probes targeting plasmodial mitochondrial sequences were cross-verified against the *Homo sapiens* mitochondrial genome, and any cross-hybridizing candidates were excluded. The array was fabricated on the Agilent microarray platform, ensuring optimal hybridization performance and data quality. Probes from our previously validated 244K microarray [12] were selectively integrated into the current 60K design to enhance coverage and reliability. In this study, we specifically report transcripts, detected through hybridization, that are either encoded by the *P. vivax* mitochondrial genome or are predicted to be mitochondrial-targeted.

Microarray Data Analysis

Feature-extracted raw image data, filtered for signals at least two-fold above background intensity, were processed using Agilent GeneSpring GX Software. Data normalization was performed using the 75th percentile shift method to ensure consistency across samples and reduce technical variability. To enhance strict stringency, only probes detecting transcripts in at least 60% of samples within each clinical group complicated (hepatic dysfunction) and uncomplicated *P. vivax* malaria were retained for further analysis. Probes exhibiting inconsistent expression for the same transcript were excluded based on Pearson correlation analysis, applying a stringent cutoff of ≥0.8 to maintain data integrity. Differential gene expression was assessed by comparing complicated samples (patients with hepatic dysfunction,

n = 9) to control samples (uncomplicated cases, n = 5). Genes showing a fold change ≥1.5 were considered biologically relevant, and their statistical significance was determined using a student t-test ($p \leq 0.05$), with multiple testing correction applied via the Benjamini-Hochberg false discovery rate. All statistical analyses were conducted using the Bioconductor package in R.

Machine Learning

Blood samples from (n=14) adult patients who had malaria due to *P. vivax* was strategically selected for machine learning based on several considerations. We use several indexes, including F1-Score, Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC), to evaluate the performance of models, each with its specific formula, significance, and application, particularly in scenarios with imbalanced data (non-uniform between uncomplicated and complicated samples).

The metric of F1-Score is the harmonic mean of precision and recall, calculated as

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1-Score, the harmonic mean of precision and recall, provides a balanced assessment of a model's ability to identify both true positives and true negatives. A score close to 1 indicates strong performance, particularly vital in datasets with unequal class distributions.

MCC is calculated using the formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. It offers a comprehensive evaluation that considers all elements of the matrix (TP, TN, FP, FN). It ranges from -1 (inverse prediction) to +1 (perfect prediction). MCC is especially useful in imbalanced datasets for its balanced insight.

AUC (Area Under the Curve) quantifies a model's ability to distinguish between classes, independent of class imbalance, AUC ranges from 0 to 1.

Unlike simple accuracy, these metrics offer nuanced insights, ensuring the model's effectiveness is not masked by skewed class ratios. A high F1-Score, MCC, and AUC collectively indicate the model's strength in accurately predicting both complicated and uncomplicated cases, providing a holistic assessment of its predictive capabilities across different class distributions.

The machine learning training was conducted on complicated samples (patients with hepatic dysfunction, n = 9) in comparison to control samples (uncomplicated cases, n = 5). Leveraging the results from differential expression analysis, we prioritized genes based on adjusted p-values—using sets of the top 92, 80, 70, 60, 50, 40, 30, and 20 transcripts from the sense transcripts, and 77, 70, 60, 50, 40, 30, and 20 from the NATs as input features for training the machine learning models. Our machine learning models, trained on these transcriptomic signatures, were further interrogated to uncover potential regulatory transcripts in sense & NATs that may act as drivers involved in the modulation of mitochondrial gene expression. The resulting models not only demonstrated high predictive performance but also guided downstream analyses to explore enriched pathways and biological functions. This integrative

approach sets the foundation for discovering novel therapeutic targets and improving clinical stratification of malaria patients.

We employed three machine learning algorithms, including XGBoost [35] with parameters n_estimators=100,max_depth=2,learning_rate=0.05,subsample=0.8,colsample_bytree=0.8,reg_alpha=0.1,reg_lambda=1, random forest [36] with parameters for max_depth = 3, min_samples_split = 5, n_estimators = 100, logistic regression [37] with parameters max_iter = 1000 to test and predict key regulators of disease severity in case of mitochondrial transcripts from sense & NATs. The Leave-One-Out (LOOCV) cross-validation technique was utilized [38].

We employed Leave-One-Out Cross-Validation (LOOCV) to address critical challenges in evaluating machine learning models trained on limited and high-dimensional biological data. For imbalanced datasets, standard cross-validation may create folds with zero minority-class samples, skewing performance metrics. Small datasets with high dimensionality are prone to overfitting. LOOCV's exhaustive resampling helps detect overfitting by testing the model's consistency across all possible training configurations. In standard k-fold cross-validation, the dataset is partitioned into k subsets, with each subset serving as the test set once. LOOCV represents the extreme case of k-fold CV, where k equals the total number of samples. LOOCV guarantees each test set contains exactly one sample, preserving class representation and providing a more reliable estimate of model sensitivity/specificity.

For each iteration of LOOCV, N-1 samples are used to train the model while the remaining single sample serves as the test set. This process is repeated N times, ensuring every sample is used as the test set exactly once.

Mathematically, the LOOCV estimate of the mean squared error (MSE) is given by:

$$M\,S\,E_{LOOCV} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{f^{-i}}\,(x_i))^2$$

where N denotes the number of iterations, and the term in the brackets denotes the model trained on all samples except the i-th observation.

LOOCV helps to ensure no data wastage, which is ideal for precious samples, and provides stability by removing randomness from fold generation, which is crucial for publishing reproducible results. Computational cost for employing this technique increases linearly with the sample size, making it prohibitive for large datasets. In our study, LOOCV ensured rigorous validation without sacrificing statistical robustness.

To gain deeper insights into model interpretability, we employed SHAP (SHapley Additive exPlanations, v0.47.2) to unravel the contribution of individual features to the prediction outcome. By computing Shapley values, we quantified the impact of each feature on the model's decision-making process, enabling a precise, feature-level understanding of predictive influence [39].

Functional enrichment analysis (FEA)

To uncover the biological relevance of the identified genes, functional enrichment analysis (FEA) was conducted, encompassing Gene Ontology (GO) predictions and pathway

enrichment through the Kyoto Encyclopedia of Genes and Genomes (KEGG: http://www.genome.jp) [40]. A total of 80 and 70 significantly expressed genes from sense and NATs, respectively, based on the overall performance of the three distinct machine learning models, were analysed using ShinyGO v0.82 [41]. This user-friendly and interactive platform enabled intuitive visualization and interpretation of the biological meaning of these large set of genes. The analysis was configured with specific parameters: *Plasmodium vivax* Sal-1 genome as the reference, a minimum gene count threshold of 20, and a significance cutoff set at $p = 0.05$.

Results

Our previous studies have also documented *P. vivax* infections presenting with severe complications such as cerebral malaria, acute respiratory distress syndrome, jaundice, acute renal failure, etc., in adult patients [12]. Building on these findings, the current study focuses on 14 adult cases of *P. vivax* malaria, encompassing both complicated (Hepatic Dysfunction) and uncomplicated manifestations, classified in accordance with WHO guidelines [34]. Detailed clinical characteristics of these patients are provided in Table S1.

Mitochondrial associated genes

A curated list of 636 mitochondrial-associated genes was compiled, encompassing 19 genes encoded by the mitochondrial genome and 617 nuclear-encoded genes predicted to target the mitochondria. These predictions were derived using a suite of tools including TargetP [22], Cello2GO [23], MitoFate [24], MitoprotII [25], and PlasmoDB [26]. To ensure reliability, only genes predicted by at least two independent tools were retained. Additionally, orthologs from *Plasmodium falciparum* mitochondrial proteome were incorporated to strengthen the dataset. Notably, a considerable overlap was observed among the prediction platforms, supporting the robustness of the gene selection. This comprehensive list detailed in (Supplementary Data-2) includes both computationally predicted and previously validated candidates, serving as the foundation for the transcriptomic analyses presented in this study.

Differential gene expression analysis in clinical isolates

To uncover the differential gene expression patterns between complicated and uncomplicated *P. vivax* malaria, we normalized microarray data from 14 clinical samples using the 75th percentile shift method. Differential expression analysis was performed using a fold change threshold of ≥ 1.5 and a significance level of $p \leq 0.05$, with correction for multiple testing via the Benjamini-Hochberg false discovery rate method. Expression fold changes were calculated as the average signal intensity across complicated malaria samples ($n = 9$). To ensure biological relevance, only probes mapping to *P. vivax* Sal-I transcript sequences and detected in at least 9 out of 14 samples were considered for further analysis. From a total of 636 mitochondrial transcripts, a total of 626 genes were analysed for sense transcripts and 581 genes for NATs (Fig-1A, Supplementary Data-3). Out of those a total of 92 and 77 genes were differentially expressed sense and NATs in complicated *P. vivax* malaria samples as compared to the uncomplicated *P. vivax* malaria samples, respectively (Fig-1B, Supplementary Data-3). Among the 92 sense transcripts, 66 genes (71.7%) were upregulated, while 26 genes (28.3%) were downregulated (Fig-1C) and out of 77 differentially expressed NATs, 27 genes (35.1%) were upregulated, and a notable 50 genes (64.9%) were downregulated (Fig-1D). Interestingly, 17 genes exhibited differential expression in both sense and NATs transcripts (Fig-1E). Correlation analysis revealed that 13 of these gene pairs displayed a positive correlation, whereas 4 gene pairs showed inverse expression patterns (negative correlation) between sense

and NATs (Fig-1F). These findings align with previous evidence suggesting that a single gene can express both sense and NATs transcripts, with their expression either positively or negatively correlated depending on the context which may possibly modulate subsequent protein expression [12].

Evaluating Machine Learning Sensitivity Across Varying Thresholds of Significant Gene Expression

To explore the predictive potential of mitochondrial transcripts, our initial analysis focused on evaluating features derived from statistically significant differentially expressed genes (DEGs) in both sense and NATs. These features were assessed for their utility in machine learning aimed at identifying key markers or genes in mitochondrial regulation. From the differential expression analysis, 92 sense and 77 NATs were shortlisted based on their adjusted p-values (Supplementary Data-4), and subsequently prioritized for downstream predictive modelling and pathway enrichment. To gauge predictive strength, we systematically evaluated subsets of these DEGs—ranging from the top 92 to 20 for sense transcripts, and from 77 to 20 for NATs (Supplementary Data-4), and then using three different machine learning algorithms. Remarkably, performance metrics including accuracy, F1-score, and AUC consistently approached or exceeded 90% for subsets ranging from the top 80 to 20 in sense, and 70 to 20 in NATs (Fig. 2A–B). These findings underscore the robustness of these gene sets in predictive modelling and highlight their potential biological relevance. Based on the comparative evaluation, the top 80 genes in sense and top 70 in NATs offered the most reliable prediction performance across all algorithms. Accordingly, these subsets were selected for further pathway enrichment analysis to enable a deeper, systems-level understanding of the mitochondrial regulatory landscape potentially governed by these transcripts.

To establish a robust baseline, 40 genes were randomly selected from the full set of mitochondrial sense and NAT transcripts and used as machine learning features [Supplementary Data-4]. This control experiment was designed to validate the predictive strength of statistically derived features. The outcome clearly demonstrated that without prior feature selection, the machine learning models failed to yield reliable predictive performance—highlighting the critical role of informed, data-driven feature selection in enhancing model accuracy and biological interpretability.

Gene Ontology Enrichment Analysis of Top-Performing Predictive Gene Set

To investigate the functional relevance of differentially expressed mitochondrial transcripts, the top 80 sense and 70 NATs genes were subjected to enrichment analysis using ShinyGO v0.82 [41]. Interestingly, no enriched pathways were observed for the top 80 sense transcripts, suggesting limited or more dispersed functional coherence among them. In contrast, the analysis of the top 70 significantly expressed NATs revealed enrichment in key mitochondrial pathways, in the tricarboxylic acid (TCA) cycle complex (Fig-3A). Additionally, pathways associated with amino acid catabolism and mitochondrial protein complex assembly were also identified (Fig-3B). These findings strongly suggest that NATs may exert a regulatory influence on mitochondrial function—potentially modulating energy metabolism and organelle dynamics during the intraerythrocytic stages of the parasite. The absence of similar enrichment in sense transcripts underscores the unique and possibly specialized regulatory role of NATs, positioning them as critical players in parasite physiology and as promising candidates for future therapeutic targeting.

Feature Importance analysis

To identify the key contributors among the significantly expressed genes, SHAP analysis was applied to the best-fitting Random Forest model to interpret feature importance. The top-ranking genes in both sense and NATs transcripts (Fig. 4A–4B) demonstrated strong predictive relevance, from the overall 80 genes selected for sense and 70 genes for NATs, providing the most reliable prediction performance across all algorithms tested.

Genes with SHAP values were considered to have a greater influence on the model's predictions and are potential candidates for involvement in mitochondrial regulatory processes. Pathway enrichment analysis using ShinyGO v0.82 [41] revealed that the top-ranking sense genes did not map to any significantly enriched mitochondrial pathways. In contrast, the high-impact NATs genes with mean SHAP values were found to be enriched in several key mitochondrial pathways and translation regulator activity.

By providing gene-level feature importance scores, SHAP helped to distinguish which significantly expressed genes were truly influential in driving the model's predictions, rather than being statistical noise. In this case, SHAP facilitated the prioritisation of candidate genes that may play regulatory roles in mitochondrial pathways, especially among NATs, thereby generating biologically meaningful insights that could guide future functional validation and mechanistic studies.

Discussion

This study represents the first comprehensive investigation into mitochondrial gene regulation in *P. vivax* using clinical isolates from patients with severe malaria, focusing on both sense and NATs. By integrating transcriptomic profiling with advanced machine learning approaches to understand the reliability and rationality when used on these clinical isolates, we have uncovered novel insights into the regulatory dynamics of mitochondrial function, particularly highlighting the regulatory role of NATs. These findings unveil gene regulatory mechanisms within the mitochondria that may drive the pathogenesis of severe *P. vivax* malaria, highlighting key molecular players that could serve as potential regulators in disease severity.

The mitochondrion of *Plasmodium* species, including *P. vivax*, is a highly reduced organelle with a 6 kb linear genome encoding only three protein-coding genes: cytochrome c oxidase subunits 1 and 3, and cytochrome b along with scrambled fragments of ribosomal RNAs [42]. Despite its simplicity, this organelle is critical for metabolic processes such as pyrimidine biosynthesis, heme metabolism and the electron transport chain (ETC), which are essential for parasite survival during the intraerythrocytic developmental cycle (IDC) [43]. Unlike typical eukaryotic mitochondria, *Plasmodium* mitochondria do not primarily produce ATP via the TCA cycle; instead, TCA cycle enzymes support biosynthetic pathways, including amino acid and heme synthesis, highlighting the need for precise gene regulation to meet the parasite's metabolic demands under host-induced stress [44].

Non-coding RNAs (ncRNAs), particularly NATs, have emerged as key regulators of gene expression in *P. vivax* [12]. Emerging studies in *P. falciparum* suggest that NATs play a pivotal role in modulating stage-specific gene expression during the intraerythrocytic developmental cycle (IDC), orchestrating regulatory mechanisms such as transcriptional interference and chromatin remodeling to help the parasite dynamically adapt to changing host environments [45]. These regulatory mechanisms are crucial for *P. vivax*, which faces unique challenges due to its ability to form dormant hypnozoites and cause relapsing infections [46]. Gene regulation, particularly in the mitochondrion, is vital for balancing replication, metabolic efficiency, and

immune evasion, especially in severe malaria where physiological stresses are amplified [47]. Our study builds on this foundation by incorporating a hypothesis-driven statistical framework to identify genes of interest, which were subsequently used as input features for machine learning. While machine learning is conventionally employed for prediction or classification, we took a novel approach, we began by formulating a biologically relevant hypothesis, then selected data that aligned with this hypothesis, and only then applied machine learning. The rationale was that if the model achieved high predictive accuracy using these features, it would lend logical support to the initial hypothesis—thus enabling us not just to predict, but also to interpret the biological significance underlying the prediction. This integrative strategy bridges hypothesis testing and data-driven modelling, offering a more interpretable path to uncover meaningful patterns in complex datasets.

Towards this, analysing the transcriptomic data from 14 clinical isolates (9 complicated and 5 uncomplicated), revealed a dynamic landscape of differentially expressed genes (DEGs), with 92 sense transcripts and 77 natural antisense transcripts (NATs) showing significant changes in severe malaria (fold change ≥ 1.5, $p \leq 0.05$, Benjamini-Hochberg corrected). The approach allowed identification of 92 differentially expressed sense transcripts, with 71.7% of these upregulated. The predominant upregulation of mitochondrial genes during severe malaria likely reflects the parasite's adaptive response to the physiological stress imposed by host immune responses and the changing erythrocytic environment. Notably, our findings contrast with the expression patterns observed in NATs, where 64.9% were downregulated in complicated malaria. This downregulation of NATs differs significantly from the predominantly upregulated sense transcripts, indicating distinct regulatory mechanisms governing sense and antisense transcript expression. When examining individual gene pairs that express both sense and NATs, our correlation analysis revealed that only 17 genes exhibited differential expression in both sense & NATs, with 13 gene pairs showing positive correlation and 4 pairs displaying negative correlation between their sense and corresponding NATs. This observation aligns with earlier findings in *P. vivax*, which demonstrated dynamic shifts in the expression of sense–antisense gene pairs throughout the intraerythrocytic developmental cycle, highlighting their potential role in stage-specific regulation [12]. Furthermore, studies across eukaryotic systems have demonstrated that NATs exhibit diverse regulatory relationships with their sense counterparts, with mechanisms ranging from transcriptional interference leading to negative correlations, to RNA stabilization and chromatin remodeling resulting in positive correlations [48,49]. The complexity of antisense-mediated gene regulation is further evidenced by genome-wide studies showing that the majority of sense-antisense pairs demonstrate positive correlations, while a smaller fraction exhibits negative correlations, suggesting that the regulatory outcome depends on the specific molecular mechanisms involved [50,51].

The exceptional performance of our machine learning models, achieving F1-scores, MCC, and AUC values consistently $\geq 90\%$ for optimal gene sets, validates the biological relevance of the identified transcriptomic signatures. The superior performance of the top 80 sense genes and top 70 NATs compared to randomly selected controls demonstrates that our statistically-driven feature selection successfully captured biologically meaningful signals. This approach aligns with recent advances in machine learning applications, where transcriptomic biomarkers have shown promise for disease classification [52].

The implementation of LOOCV was particularly appropriate for our small, high-dimensional dataset, addressing the challenge of imbalanced sample sizes between complicated (n=9) and uncomplicated (n=5) cases. This rigorous validation approach ensures the robustness of our

findings and provides confidence in the predictive capabilities of the identified gene signatures. Similar validation strategies have proven essential in other gene expression studies with limited sample sizes [53].

The most striking finding of our study is the selective enrichment of NATs in critical mitochondrial pathways, particularly the tricarboxylic acid (TCA) cycle complex, amino acid catabolism, and mitochondrial protein complex assembly. This selective enrichment contrasts sharply with the absence of significant pathway enrichment among sense transcripts, positioning NATs as specialized regulators of mitochondrial function during severe malaria. The enrichment of NATs in TCA cycle pathways is particularly significant given the established importance of mitochondrial metabolism in *Plasmodium* biology. Importantly, our analysis revealed that two genes encoding components of the branched-chain alpha-keto acid dehydrogenase (BCKD) complex- PVX_119310 (lipoamide acyltransferase component of branched-chain alpha-keto acid dehydrogenase complex, putative) and PVX_122460 (2-oxoisovalerate dehydrogenase subunit alpha, mitochondrial, putative) showed downregulation in NATs, though these genes themselves were not differentially expressed in the sense. The downregulation of NATs targeting these BCKD complex components represents a sophisticated regulatory mechanism with important metabolic implications. The BCKD complex catalyzes the oxidative decarboxylation of branched-chain amino acids (leucine, isoleucine, and valine), converting them to their corresponding acetyl-CoA derivatives that can subsequently fuel cellular energetics through the TCA cycle [54]. This complex functions as a critical metabolic hub, comprising multiple enzyme subunits including the E1 component (containing the 2-oxoisovalerate dehydrogenase subunit alpha encoded by PVX_122460) and the E2 component (lipoamide acyltransferase encoded by PVX_119310). The selective downregulation of NATs for these BCKD components, in the absence of corresponding sense transcript differential expression, could potentially regulate adjustment during severe malaria when the parasite faces heightened metabolic demands and host-induced stress, as branched-chain amino acids serve as essential building blocks for protein synthesis and can function as alternative fuel sources when glucose availability becomes limiting.

Studies have demonstrated that NATs frequently act through diverse mechanisms including RNA masking, chromatin modification, and fine-tuning of gene expression in response to cellular stress conditions [55]. In the context of severe malaria, this NAT-mediated modulation of BCKD complex components could enhance the parasite's metabolic flexibility by promoting the capacity for branched-chain amino acid catabolism without the energetic cost of continuously maintaining high enzyme expression levels. Our findings suggest that NATs may fine-tune the expression of both classical TCA cycle enzymes and amino acid catabolic enzymes, potentially allowing the parasite to rapidly adjust metabolic flux in response to changing host conditions.

The unexpected enrichment of NATs in translation regulator activity represents a novel finding with significant implications for our understanding of mitochondrial gene regulation. This suggests that NATs may influence not only transcriptional regulation but also post-transcriptional processes, including protein synthesis and mRNA stability. Recent studies have highlighted the importance of RNA methylation and alternative splicing in malaria parasites, demonstrating multiple regulatory layers that control the post-translational proteome [56]. Our findings extend this concept by implicating NATs in translational control, potentially providing an additional mechanism for rapid adaptation to changing cellular conditions. The dual role of NATs in both metabolic pathway regulation and translational control suggests a sophisticated

regulatory network that enables precise coordination of mitochondrial function during severe malaria.

The application of SHAP analysis provided unprecedented insights into individual gene contributions by quantifying feature-level importance, SHAP analysis distinguished truly influential genes from statistical noise, a critical advancement given the high-dimensional nature of transcriptomic data. This interpretable machine learning approach has gained recognition in biomedical research for its ability to provide mechanistic insights into complex biological systems [57]. Our SHAP analysis revealed that while top-ranking sense genes showed no pathway enrichment, high-impact NATs genes demonstrated significant enrichment in mitochondrial pathways, highlighting the specialized regulatory role of antisense transcripts.

Our study introduces several methodological innovations that enhance the reliability and interpretability of transcriptomic analyses in malaria research. The integration of statistical differential expression analysis with machine learning validation provides a robust framework for discovery of critical molecular targets. The use of SHAP analysis for feature importance interpretation represents a significant advancement over traditional gene ranking approaches, providing mechanistic insights into regulatory networks. The custom 60K sense-antisense microarray design enabled comprehensive profiling of both sense and antisense transcripts.

The identification of NATs as key modulators of mitochondrial regulators of both metabolic pathways and translational control suggests sophisticated regulatory networks that could be utilized for drug development. The robust predictive performance of our machine learning models demonstrates the potential for molecular players in clinical applications. Functional validation studies using gene knockout or RNA interference approaches could directly demonstrate the regulatory roles of specific NATs. Certain limitations should be acknowledged in interpreting our findings. The relatively small sample size, while adequate for the analytical approaches employed, limits the statistical power for detecting subtle effects. Future studies should include larger cohorts with diverse clinical presentations to validate and extend our findings. The use of microarray technology, while providing robust detection of known transcripts, may miss novel regulatory elements that could be identified through RNA sequencing approaches.

The positioning of the *Plasmodium* mitochondrion as a promising target for antimalarial drug development is strongly supported by our findings. The unique architecture and regulatory mechanisms identified in this study provide multiple avenues for therapeutic intervention, from targeting specific metabolic enzymes to disrupting NAT-mediated regulation.

Acknowledgments

We thank all the patients and technical workers for participating in and supporting this project. This work is funded by the Indian Council of Medical Research (ICMR) under the Extramural Ad-hoc Scheme (PID: 2019-1121, a multicentric project with AD as PI and DG and SKK as Co-PI). We thank Birla Institute of Technology (BITS), Pilani (Pilani campus), International Centre for Genetic Engineering and Biotechnology (ICGEB), New Delhi, and SP Medical College, Bikaner, for providing the necessary facilities required for this work. We also thank Genotypic Technology Pvt. Ltd., Bangalore, India, for the microarray hybridization and consultation regarding data analysis.

Availability of data and materials

All microarray results have been deposited in the Gene Expression Omnibus database (GSE 300144). All data generated are provided in the main text and supplementary data files. All patient meta-data included in this article are included in Table S1. The codes are available at https://github.com/MPSB-Lab/ML-Microarray.git. Reprint requests should be addressed to AD. All Venn diagram illustrations have been made using Venny [58]. Heatmaps were generated using the Seaborn and Matplotlib libraries, while SHAP plots were created using the SHAP and Matplotlib libraries in python v3.11.13.

Declaration of interest

The authors declare no competing interests.

Ethics statement

Ethics Approval - Sample collection was earlier approved by SP Medical College Hospital's Ethics Committee (No. F (Acad)SPMC/2003/2395). Permission to use these samples for further studies was given through IERB approval No. F29(Acad)SPMC/2020/3151 dated 05.09.2020. Informed consent was obtained from all individual participants included in the study. We sincerely thank all the patients for their voluntary consent and participation in the study.

Author Contributions

AD conceived the study, acquired funding from ICMR, India, and networked to facilitate the outcome. PR was involved in performing the experimentation, analysis, and data interpretation, designing the figures, and drafting the manuscript under the constant guidance of AD. YG was involved in designing, training, and evaluating the machine learning models used in this study. DKK and SKK supervised sample collection and clinical characterization. All authors read and approved the manuscript.

References

1. WHO. World malaria report 2024: addressing inequity in the global malaria response. Geneva: World Health Organization. 2024 (Accessed on Feb 27, 2025).
2. Kepple, D., Ford, C. T., Williams, J., Abagero, B., Li, S., Popovici, J., Yewhalaw, D., & Lo, E. (2023b). Comparative transcriptomics reveal differential gene expression in Plasmodium vivax geographical isolates and implications on erythrocyte invasion mechanisms. bioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2023.02.16.528793
3. Boopathi, P. A., Subudhi, A. K., Middha, S., Acharya, J., Mugasimangalam, R. C., Kochar, S. K., ... & Das, A. (2016). Design, construction and validation of a Plasmodium vivax microarray for the transcriptome profiling of clinical isolates. Acta Tropica, 164, 438-447.
4. Mattick, J. S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. EMBO Reports, 2(11), 986–991. https://doi.org/10.1093/embo-reports/kve230.
5. Ma, L., Bajic, V. B., & Zhang, Z. (2013). On the classification of long non-coding RNAs. RNA Biology, 10(6), 924–933. https://doi.org/10.4161/ rna.24604
6. Salzman, J. (2016). Circular RNA expression: Its potential regulation and function. Trends in Genetics, 32(5), 309–316. https://doi.org/10. 1016/j.tig.2016.03.002
7. Yu, C. Y., & Kuo, H. C. (2019). The emerging roles and functions of circular RNAs and their generation. Journal of Biomedical Science, 26(1), 1–12. https://doi.org/10.1186/s12929-019-0523-z

8. Lodde, V., Floris, M., Muroni, M. R., Cucca, F., & Idda, M. L. (2021). Non-coding RNAs in malaria infection. Wiley Interdisciplinary Reviews. RNA, 13(3). https://doi.org/10.1002/wrna.1697
9. Patankar, S., Munasinghe, A., Shoaibi, A., Cummings, L. M., & Wirth, D. F. (2001). Serial analysis of gene expression in Plasmodium falciparum reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. Molecular biology of the cell, 12(10), 3114-3125.
10. Militello, K. T., Patel, V., Chessler, A., Fisher, J. K., Kasper, J. M., Gunasekera, A., & Wirth, D. F. (2005). RNA polymerase II synthesizes antisense RNA in. Synthesis, 11(4), 365–370. https://doi.org/10.1261/rna.7940705.strongly
11. Lopez-Barragan, M. J., Lemieux, J., Quiñones, M., Williamson, K. C., Molina-Cruz, A., Cui, K., Barillas-Mury, C., Zhao, K., & Zhuan, S. X. (2011). Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum. BMC Genomics, 12, 587. https://doi.org/10.1186/1471-2164-12-587
12. Boopathi, P., Subudhi, A. K., Garg, S., Middha, S., Acharya, J., Pakalapati, D., Saxena, V., Aiyaz, M., Chand, B., Mugasimangalam, R. C., Kochar, S. K., Sirohi, P., Kochar, D. K., & Das, A. (2013). Revealing natural antisense transcripts from Plasmodium vivax isolates: Evidence of genome regulation in complicated malaria. Infection Genetics and Evolution, 20, 428–443. https://doi.org/10.1016/j.meegid.2013.09.026
13. Vaidya, A. B. & Mather, M. W. Mitochondrial evolution and functions in malaria parasites. Annu. Rev. Microbiol. 63, 249–267 (2009).
14. Van Dooren, G. G., Stimmler, L. M., & McFadden, G. I. (2006). Metabolic maps and functions of thePlasmodiummitochondrion. FEMS Microbiology Reviews, 30(4), 596–630. https://doi.org/10.1111/j.1574-6976.2006.00027.x
15. Nagaraj VA, Arumugam R, Prasad D, Rangarajan PN, Padmanaban G. Protoporphyrinogen IX oxidase from Plasmodium falciparum is anaerobic and is localized to the mitochondrion. Mol Biochem Parasitol. 2010a;174(1):44–52.
16. Nagaraj VA, Prasad D, Arumugam R, Rangarajan PN, Padmanaban G. Characterization of coproporphyrinogen III oxidase in Plasmodium falciparum cytosol. Parasitol Int. 2010b;59(2):121–7.
17. Gross J, Bhattacharya D. Mitochondrial and plastid evolution in eukaryotes: an outsiders' perspective. Nat Rev Genet. 2009;10(7):495–505.
18. Sturm A, Mollard V, Cozijnsen A, Goodman CD, McFadden GI. 2015. Mitochondrial ATP synthase is dispensable in blood-stage Plasmodium berghei rodent malaria but essential in the mosquito phase. Proc Natl Acad Sci U S A 112:10216–10223. https://doi.org/10.1073/pnas.1423959112.
19. Evers F, Cabrera-Orefice A, Elurbe DM, Lindert MKT, Boltryk SD, Voss TS, Huynen MA, Brandt U, Kooij TWA. 2020. Composition and stage dynamics of mitochondrial complexes in Plasmodium falciparum. bioRxiv https://doi.org/2020:2020.10.05.326496.
20. Goodman, C. D., Buchanan, H. D., & McFadden, G. I. (2017). Is the Mitochondrion a Good Malaria Drug Target? Trends in Parasitology, 33(3), 185–193. https://doi.org/10.1016/j.pt.2016.10.002
21. Martelli, P. L., Savojardo, C., Fariselli, P., Tartari, G., & Casadio, R. (2021). Computer-aided prediction of protein mitochondrial localization. In Mitochondrial Medicine: Volume 1: Targeting Mitochondria (pp. 433-452). New York, NY: Springer US.
22. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, et al. Detecting sequence signals 555 in targeting peptides using deep learning. Life Sci Alliance. 2019; 2(5):e201900429. doi: 556 10.26508/lsa.201900429.

23. Yu CS, Cheng CW, Su WC, et al. CELLO2GO: a web server for protein subCELlular 558 LOcalization prediction with functional gene ontology annotation. PLoS One. 2014; 559 9(6):e99368. doi: 10.1371/journal.pone.0099368.
24. Fukasawa Y, Tsuji J, Fu SC, et al. MitoFates: Improved prediction of mitochondrial 561 targeting sequences and their cleavage sites. Mol Cell Proteomics. 2015; 14(4):1113-26. doi: 562 10.1074/mcp.M114.043083.
25. Claros' MG, Vincens P. Computational method to predict mitochondrially imported 566 proteins and their targeting sequences. Eur. J. Biochem. 1996; 241(3):779-86. doi: 567 10.1111/j.1432-1033.1996.00779.x.
26. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 2009;37:D539–43
27. Evers, F., Cabrera-Orefice, A., Elurbe, D. M., Kea-te Lindert, M., Boltryk, S. D., Voss, T. S., ... & Kooij, T. W. (2021). Composition and stage dynamics of mitochondrial complexes in Plasmodium falciparum. Nature communications, 12(1), 3820.
28. Lamb, I. M., Rios, K. T., Shukla, A., Ahiya, A. I., Morrisey, J., Mell, J. C., ... & Vaidya, A. B. (2022). Mitochondrially targeted proximity biotinylation and proteomic analysis in Plasmodium falciparum. PLoS One, 17(8), e0273357.
29. van Esveld, S. L., Meerstein-Kessel, L., Boshoven, C., Baaij, J. F., Barylyuk, K., Coolen, J. P., ... & Huynen, M. A. (2021). A prioritized and validated resource of mitochondrial proteins in Plasmodium identifies unique biology. Msphere, 6(5), 10-1128.
30. Van Dooren, G. G., Stimmler, L. M., & McFadden, G. I. (2006). Metabolic maps and functions of the Plasmodium mitochondrion. FEMS microbiology reviews, 30(4), 596-630.
31. Deponte, M. (2013). Mitochondrial protein import in malaria parasites. In Encyclopedia of malaria (pp. 1-13). Springer, New York, NY.
32. Pakalapati D, Garg S, Middha S, Kochar A, Subudhi AK, Arunachalam BP, Kochar SK, 669 Saxena V, Pareek RP, Acharya J, Kochar DK, Das A. Comparative evaluation of microscopy, 670 OptiMAL(®) and 18S rRNA gene based multiplex PCR for detection of Plasmodium 671 falciparum & Plasmodium vivax from field isolates of Bikaner, India. Asian Pac J Trop Med. 672 2013 May 13;6(5):346-51. doi: 10.1016/S1995-7645(13)60037-1. PMID: 23608372.
33. Pakalapati D, Garg S, Middha S, Acharya J, Subudhi AK, Boopathi AP, Saxena V, Kochar 674 SK, Kochar DK, Das A. Development and evaluation of a 28S rRNA gene-based nested PCR 675 assay for P. falciparum and P. vivax. Pathog Glob Health. 2013 Jun;107(4):180-8. doi: 676 10.1179/2047773213Y.0000000090. PMID: 23816509; PMCID: PMC4001468.
34. WHO guidelines for malaria, 30 November 2024. Geneva: World Health Organization; 2024. https://doi.org/10.2471/B09146.
35. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
36. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
37. Boateng, E. Y., & Abaye, D. A. (2019). A review of the logistic regression model with emphasis on medical research. Journal of data analysis and information processing, 7(04), 190.

38. Geroldinger, A., Lusa, L., Nold, M., & Heinze, G. (2023). Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures—a simulation study. Diagnostic and Prognostic Research, 7(1), 9.
39. Marcílio, W. E., & Eler, D. M. (2020, November). From explanations to feature selection: assessing SHAP values as feature selection mechanism. In 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI) (pp. 340-347). Ieee.
40. Kanehisa, M. (2001). Kegg: Kyoto encyclopedia of genes and genomes. Proc. Natl. Acad. Sci., 98, 4569-4574.
41. Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics, 36(8), 2628-2629.
42. Painter, H. J., Morrisey, J. M., Mather, M. W., & Vaidya, A. B. (2007). Specific role of mitochondrial electron transport in blood-stage Plasmodium falciparum. Nature, 446(7131), 88–91. https://doi.org/10.1038/nature05572.
43. Krungkrai J. (2004). The multiple roles of the mitochondrion of the malarial parasite. Parasitology, 129(Pt 5), 511–524. https://doi.org/10.1017/s0031182004005888.
44. MacRae, J. I., Dixon, M. W., Dearnley, M. K., Chua, H. H., Chambers, J. M., Kenny, S., Bottova, I., Tilley, L., & McConville, M. J. (2013). Mitochondrial metabolism of sexual and asexual blood stages of the malaria parasite Plasmodium falciparum. BMC biology, 11, 67. https://doi.org/10.1186/1741-7007-11-67.
45. Siegel, T. N., Hon, C. C., Zhang, Q., Lopez-Rubio, J. J., Scheidig-Benatar, C., Martins, R. M., Sismeiro, O., Coppée, J. Y., & Scherf, A. (2014). Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. BMC genomics, 15(1), 150. https://doi.org/10.1186/1471-2164-15-150.
46. White, N. J., & Imwong, M. (2012). Relapse. Advances in parasitology, 80, 113–150. https://doi.org/10.1016/B978-0-12-397900-1.00002-5.
47. Bozdech, Z., Llinás, M., Pulliam, B. L., Wong, E. D., Zhu, J., & DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS biology, 1(1), E5. https://doi.org/10.1371/journal.pbio.0000005.
48. Das, S., Zea Rojas, M. P., & Tran, E. J. (2024). Novel insights on the positive correlation between sense and antisense pairs on gene expression. Wiley interdisciplinary reviews. RNA, 15(4), e1864. https://doi.org/10.1002/wrna.1864.
49. Lapidot, M., & Pilpel, Y. (2006). Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. EMBO reports, 7(12), 1216–1222. https://doi.org/10.1038/sj.embor.7400857.
50. Ling, M. H., Ban, Y., Wen, H., Wang, S. M., & Ge, S. X. (2013). Conserved expression of natural antisense transcripts in mammals. BMC genomics, 14, 243. https://doi.org/10.1186/1471-2164-14-243.
51. Latgé, G., Poulet, C., Bours, V., Josse, C., & Jerusalem, G. (2018). Natural Antisense Transcripts: Molecular Mechanisms and Implications in Breast Cancers. International journal of molecular sciences, 19(1), 123. https://doi.org/10.3390/ijms19010123.
52. Li, J. X., Liao, W. Z., Huang, Z. M., Yin, X., Ouyang, S., Gu, B., & Guo, X. G. (2023). Identifying effective diagnostic biomarkers for childhood cerebral malaria in Africa integrating coexpression analysis with machine learning algorithm. European journal of medical research, 28(1), 76. https://doi.org/10.1186/s40001-022-00980-w.
53. Glaab, E., Bacardit, J., Garibaldi, J. M., & Krasnogor, N. (2012). Using rule-based machine learning for candidate disease gene prioritization and sample classification of

cancer gene expression data. PloS one, 7(7), e39932. https://doi.org/10.1371/journal.pone.0039932.

54. Oppenheim, R. D., Creek, D. J., Macrae, J. I., Modrzynska, K. K., Pino, P., Limenitakis, J., Polonais, V., Seeber, F., Barrett, M. P., Billker, O., McConville, M. J., & Soldati-Favre, D. (2014). BCKDH: the missing link in apicomplexan mitochondrial metabolism is required for full virulence of Toxoplasma gondii and Plasmodium berghei. PLoS pathogens, 10(7), e1004263. https://doi.org/10.1371/journal.ppat.1004263.
55. Wanowska, E., Kubiak, M. R., Rosikiewicz, W., Makałowska, I., & Szcześniak, M. W. (2018). Natural antisense transcripts in diseases: From modes of action to targeted therapies. Wiley interdisciplinary reviews. RNA, 9(2), e1461. https://doi.org/10.1002/wrna.1461.
56. Roy, P., Gujarati, S., Gupta, P., Gupta, I., Mahapatra, T., Gupta, D., Kochar, S. K., Kochar, D. K., & Das, A. (2025). A tale of two parasites: a glimpse into the RNA methylome of patient-derived Plasmodium falciparum and Plasmodium vivax isolates. Malaria journal, 24(1), 139. https://doi.org/10.1186/s12936-025-05376-9.
57. Chhatbar, K., Bird, A., & Sanguinetti, G. (2025). Unravelling epigenetic regulation of gene expression with explainable AI - a case study leveraging degron data. bioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2025.01.30.635704.
58. Oliveros, JC. Venny. An interactive tool for comparing lists with Venn's diagrams. 2015. https:// bioin fogp. cnb. csic. es/ tools/ venny/ index. html. Accessed 14 May 2025.

FIGURES

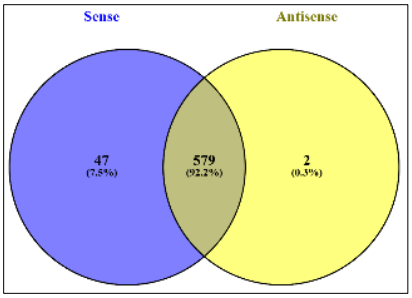

Figure 1A. Summary of mitochondrial transcript analysis: 626 sense and 581 NAT transcripts

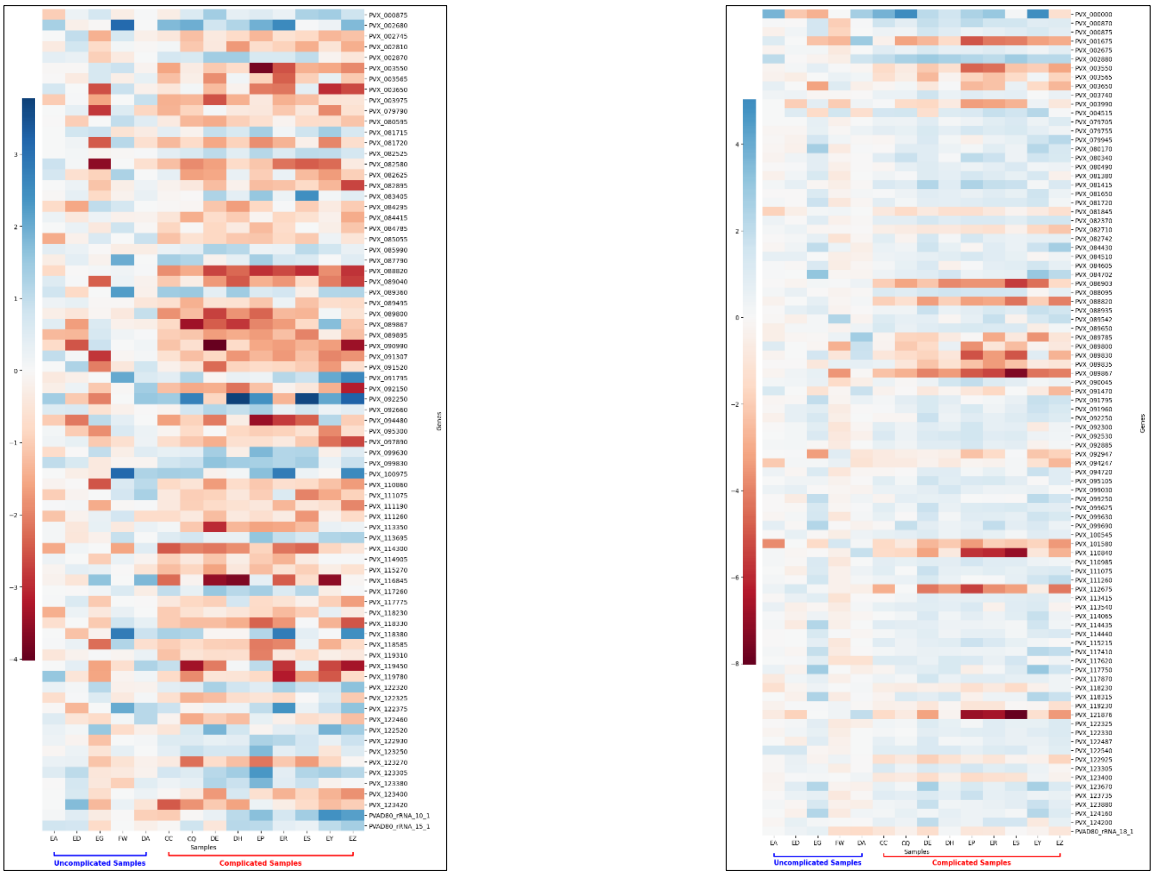

Figure 1B. Heatmap showing differential expression of mitochondrial transcripts: 92 sense and 77 NAT genes

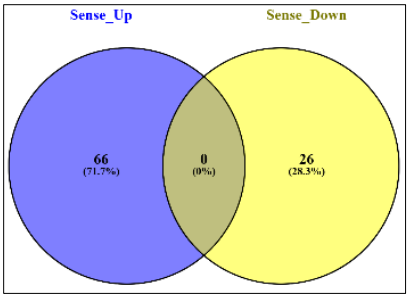

Figure 1C. Venn diagram illustrating the distribution of differentially expressed sense transcripts: 66 upregulated and 26 downregulated genes

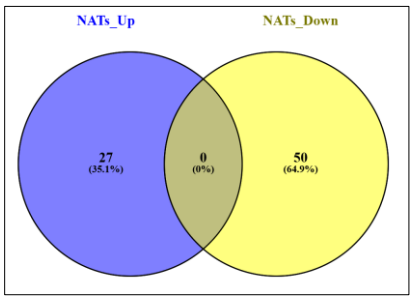

Figure 1D. Venn diagram illustrating the distribution of differentially expressed NATs transcripts: 27 upregulated and 50 downregulated genes

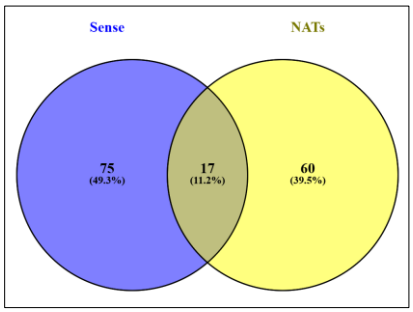

Figure 1E. Venn diagram showing overlapping genes between differentially expressed sense and NATs

.

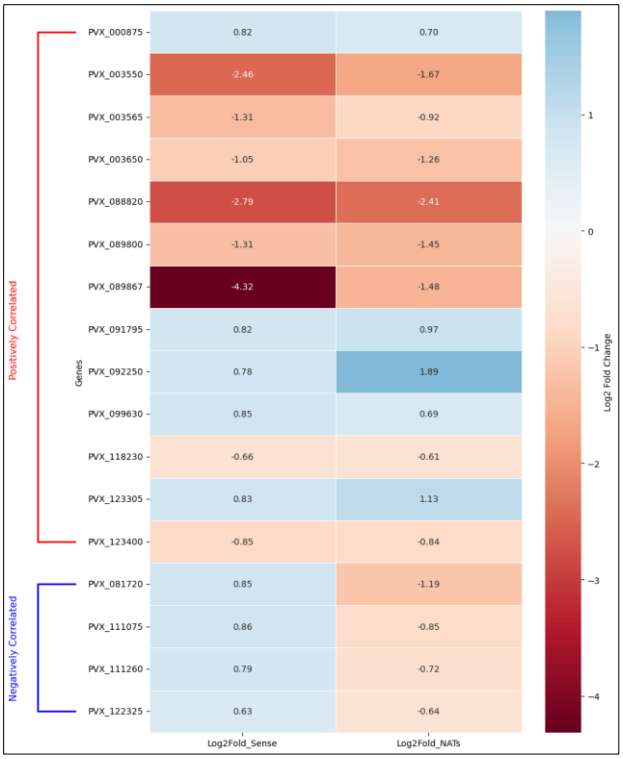

Figure 1F. Heatmap depicting correlation patterns among overlapping sense–NAT gene pairs, with 13 showing positive and 4 showing negative correlations

Machine learning results of the top 92, 80, 70, 60, 50, 40, 30, 20 significantly expressed genes, and randomly selected 40 genes										
Model	**Metric**	**92 genes**	**80 genes**	**70 genes**	**60 genes**	**50 genes**	**40 genes**	**30 genes**	**20 genes**	**Random 40 genes**
Random Forest	Accuracy	0.867	0.933	0.933	0.933	0.933	0.933	0.933	0.933	0.800
	Precision	0.889	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.750
	Sensitivity	0.889	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Specificity	0.833	0.833	0.833	0.833	0.833	0.833	0.833	0.833	0.500
	F1 Score	0.889	0.947	0.947	0.947	0.947	0.947	0.947	0.947	0.857
	MCC	0.722	0.866	0.866	0.866	0.866	0.866	0.866	0.866	0.548
	AUC	0.981	0.981	0.981	0.981	1	0.963	1	1	0.954
Logistic Regression	Accuracy	0.933	0.933	0.933	0.933	0.933	0.933	0.933	0.933	0.6
	Precision	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.667
	Sensitivity	1	1	1	1	1	1	1	1	0.667
	Specificity	0.833	0.833	0.833	0.833	0.833	0.833	0.833	0.833	0.5
	F1 Score	0.947	0.947	0.947	0.947	0.947	0.947	0.947	0.947	0.667
	MCC	0.866	0.866	0.866	0.866	0.866	0.866	0.866	0.866	0.167
	AUC	0.926	0.926	0.926	0.944	0.944	0.963	0.981	1	0.759
XGBoost	Accuracy	0.8	0.8	0.867	0.867	0.8	0.8	0.8	0.8	0.867
	Precision	0.692	0.692	0.692	0.692	0.692	0.692	0.692	0.692	0.818
	Sensitivity	1	1	1	1	1	1	1	1	1
	Specificity	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.667
	F1 Score	0.818	0.818	0.818	0.818	0.818	0.818	0.818	0.818	0.9
	MCC	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.739
	AUC	1	1	1	1	1	1	1	1	1

Figure 2A. Machine learning results of the top 92, 80, 70, 60, 50, 40, 20 significantly expressed genes, and randomly selected 40 genes in Sense

Machine learning results of the top 77, 70, 60, 50, 40, 30, 20 significantly expressed genes, and randomly selected 40 genes									
Model	**Metric**	**77 genes**	**70 genes**	**60 genes**	**50 genes**	**40 genes**	**30 genes**	**20 genes**	**Random 40 genes**
Random Forest	Accuracy	0.8	0.933	0.933	0.933	0.933	0.867	0.933	0.867
	Precision	0.75	0.9	0.9	0.9	0.9	0.818	0.9	0.818
	Sensitivity	1	1	1	1	1	1	1	1
	Specificity	0.5	0.833	0.833	0.833	0.833	0.667	0.833	0.667
	F1 Score	0.857	0.947	0.947	0.947	0.947	0.9	0.947	0.9
	MCC	0.612	0.866	0.866	0.866	0.866	0.739	0.866	0.739
	AUC	0.981	0.981	1	1	1	1	1	0.907
Logistic Regression	Accuracy	0.933	0.933	0.933	0.933	0.933	0.933	0.933	0.6
	Precision	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.667
	Sensitivity	1	1	1	1	1	1	1	0.667
	Specificity	0.833	0.833	0.833	0.833	0.833	0.833	0.833	0.5
	F1 Score	0.947	0.947	0.947	0.947	0.947	0.947	0.947	0.667
	MCC	0.866	0.866	0.866	0.866	0.866	0.866	0.866	0.167
	AUC	1	1	0.981	0.981	1	0.981	0.981	0.759
XGBoost	Accuracy	0.867	0.867	0.867	0.867	0.867	0.867	0.867	0.867
	Precision	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.818
	Sensitivity	1	1	1	1	1	1	1	1
	Specificity	0.833	0.833	0.833	0.833	0.833	0.833	0.833	0.667
	F1 Score	0.947	0.947	0.947	0.947	0.947	0.947	0.947	0.9
	MCC	0.866	0.866	0.866	0.866	0.739	0.739	0.866	0.739
	AUC	1	1	1	1	1	1	1	1

Figure 2B. Machine learning results of the top 77, 70, 60, 50, 40, 30, 20 significantly expressed genes, and randomly selected 40 genes in NATs

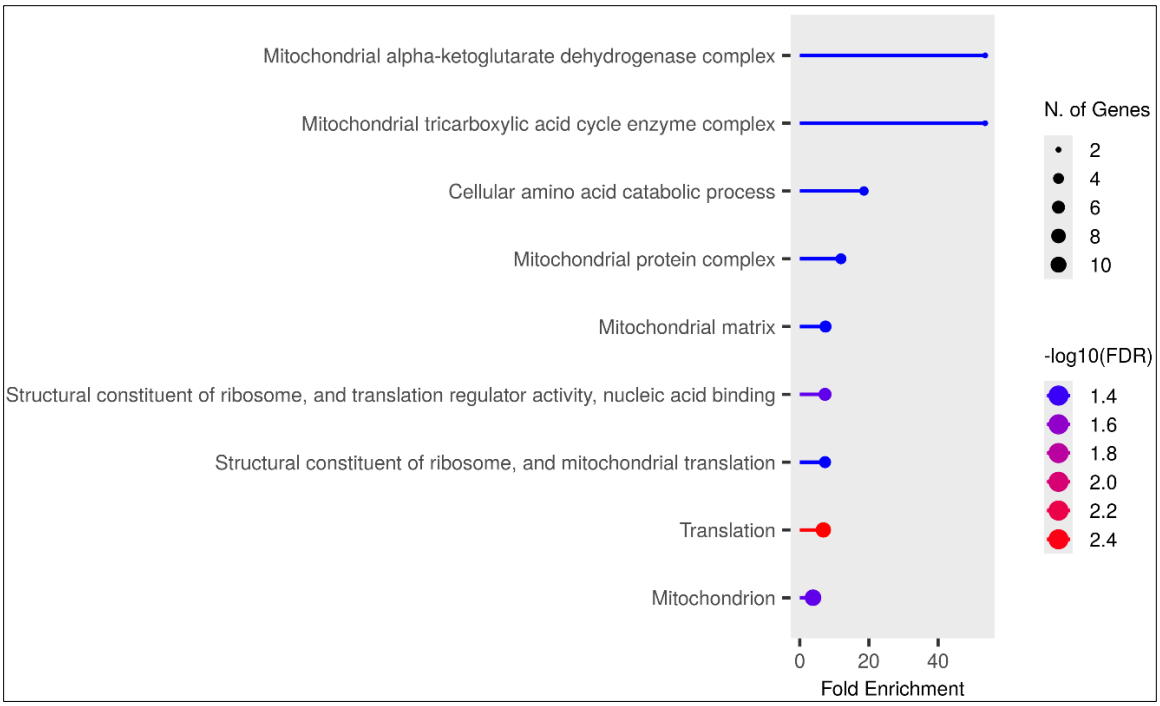

Figure 3A. Pathway analysis of the top 70 significantly expressed genes in NATs

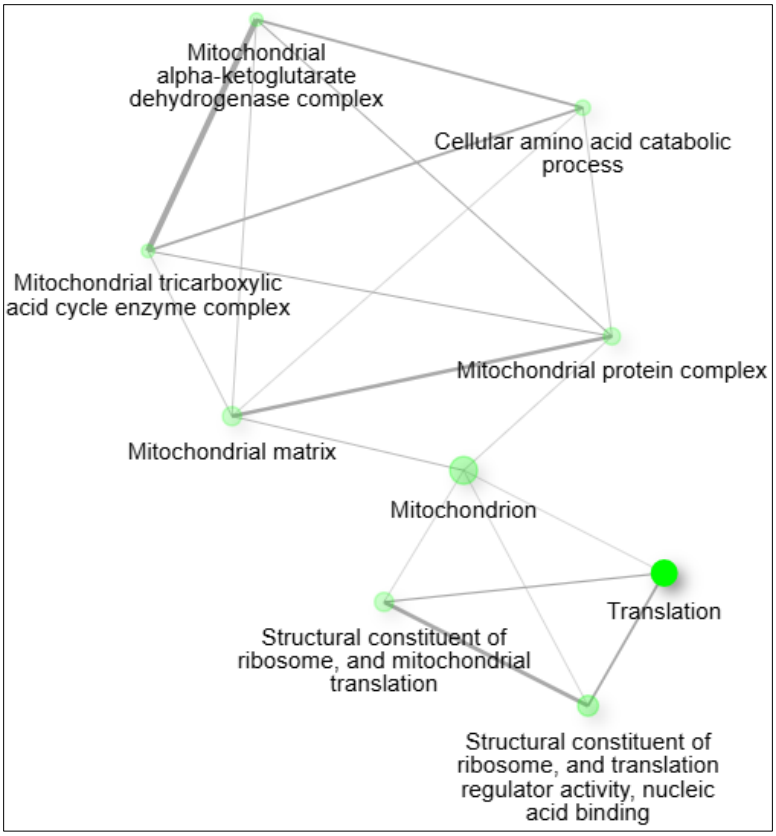

Figure 3B. Network of pathways of top 70 significantly expressed genes in NATs

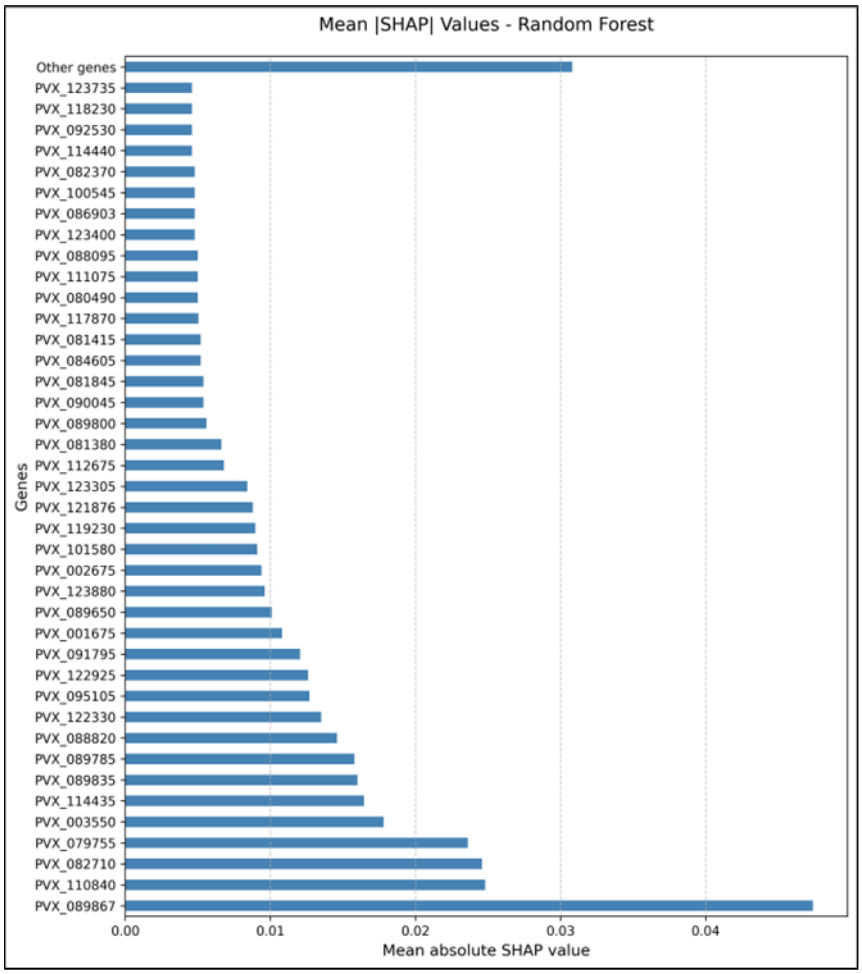

Figure 4A. SHAP analysis highlighting key feature importance genes identified by the Random Forest model for sense transcripts

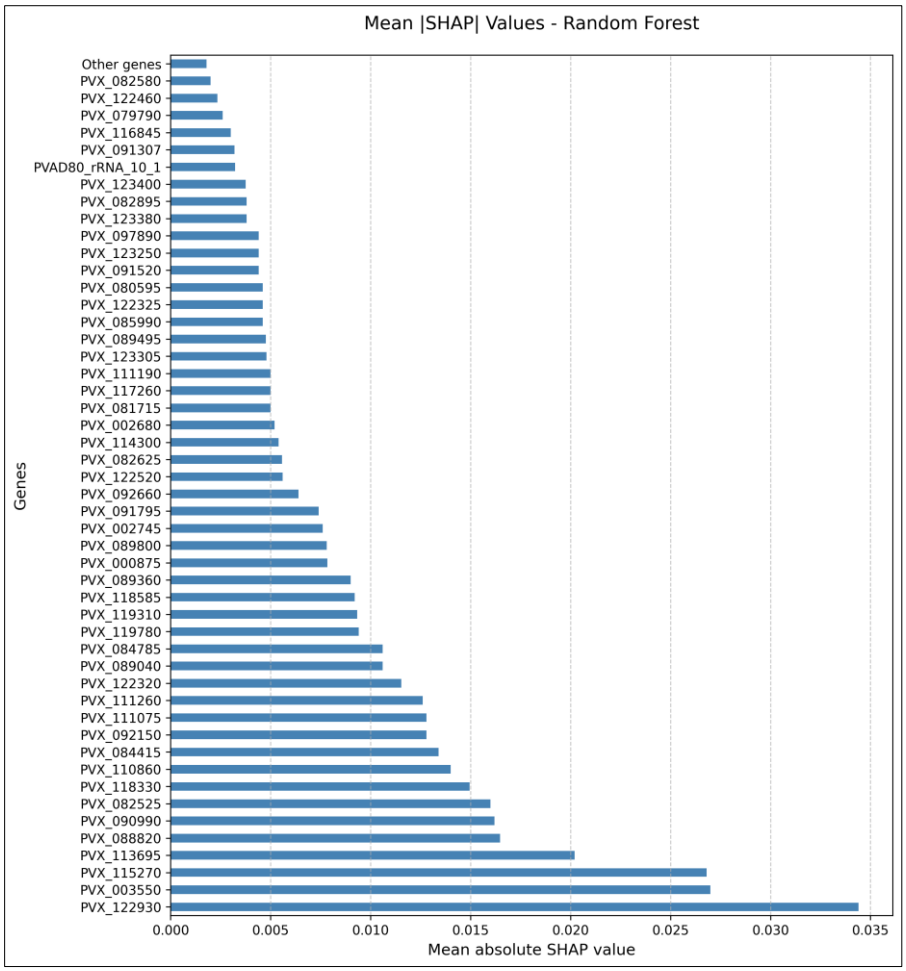

Figure 4B. SHAP analysis highlighting key feature importance genes identified by the Random Forest model for NATs