Supporting Information

Evidence for Additive and Interaction Effects of Host Genotype and Infection in Malaria

Youssef Idaghdour[1], Jacklyn Quinlan[1], Jean-Philippe Goulet[1], Joanne Berghout[2], Elias Gbeha[1], Vanessa Bruat[1], Thibault de Maillard[1], Jean-Christophe Grenier[1], Selma Gomez[3,4], Philippe Gros[2], Chérif M. Rahimy[4], Ambaliou Sanni[3] & Philip Awadalla[1]*

[1] Sainte-Justine Research Center, University of Montreal, Montreal, QC, H3T 1C5, Canada.
[2] Department of Biochemistry, McGill University, Montreal, QC, H3G 0B1, Canada.
[3] Biochimie et Biologie Moléculaire, Université d'Abomey-Calavi, Cotonou, RP, Benin.
[4] National Sickle Cell Disease Center, Université d'Abomey-Calavi, Cotonou, RP, Benin.

Content:

Other Supporting Information Files:

Databases S1-5 (XLS)

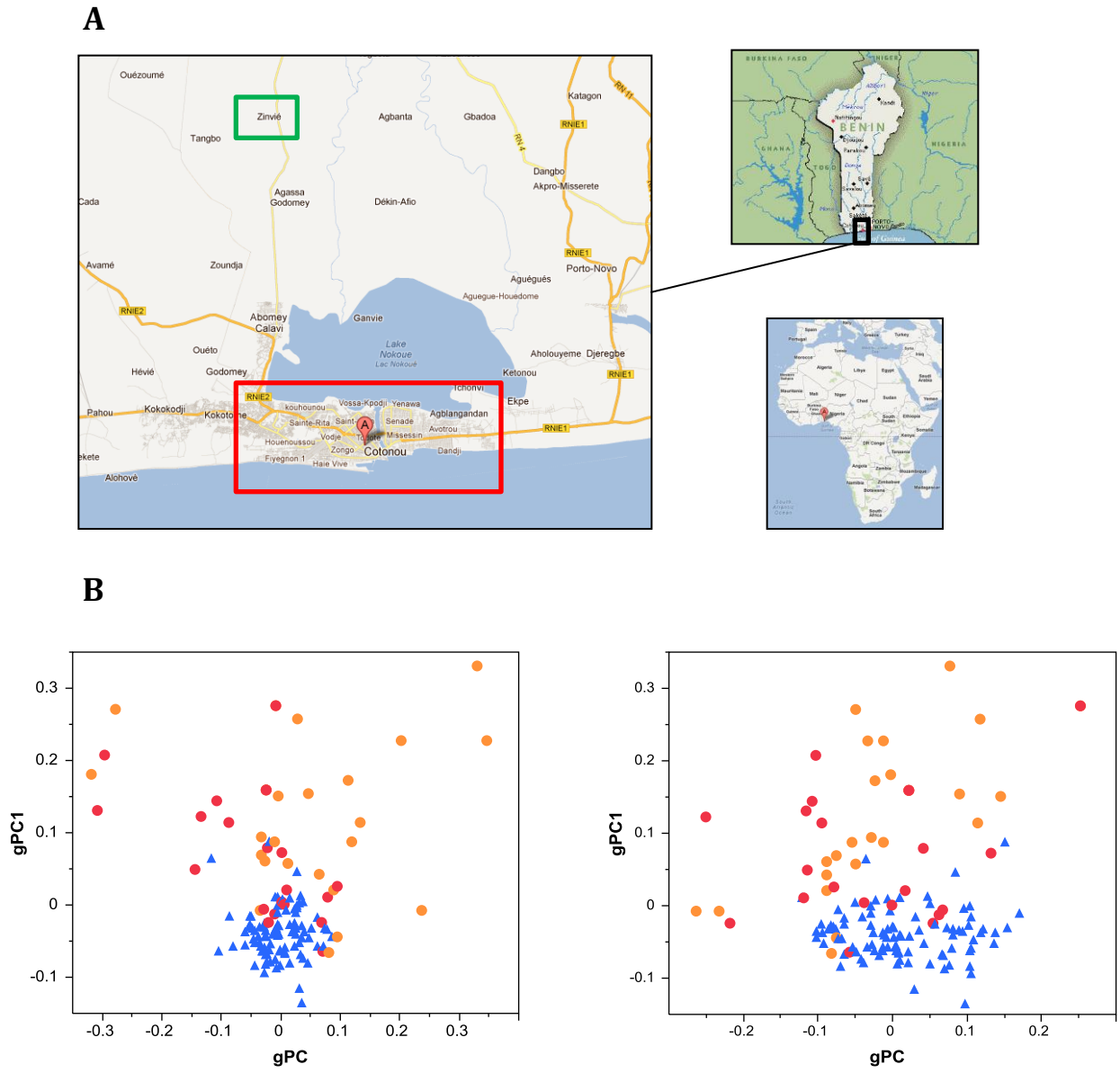


Figure S1. (**A**) Map of the southern region of Benin showing the sampling locations. The village of Zinvié (green rectangle) is located 50 km north of the coastal city of Cotonou (red rectangle). (**B**) Ancestry analysis of the entire sample using 544,672 genotypes indicated the presence of three significant genotypic principal components (gPC1-3) or eigenvectors (TW-statistic < 0.01). Scores of gPC1-3 are plotted. The individuals whose scores exerted the largest influence on the three gPCs are from the village of Zinvié. Colors in the plots indicate malaria infection status (blue, controls; red, high parasitemia group and orange, low parasitemia group). Circles and triangles indicate individuals from Zinvié and Cotonou, respectively. Ancestry analysis indicates the presence of fine population structure in the southern region of Benin and consequently we used gPC1-3 scores to account for ancestry in gene expression and eSNP analyses.

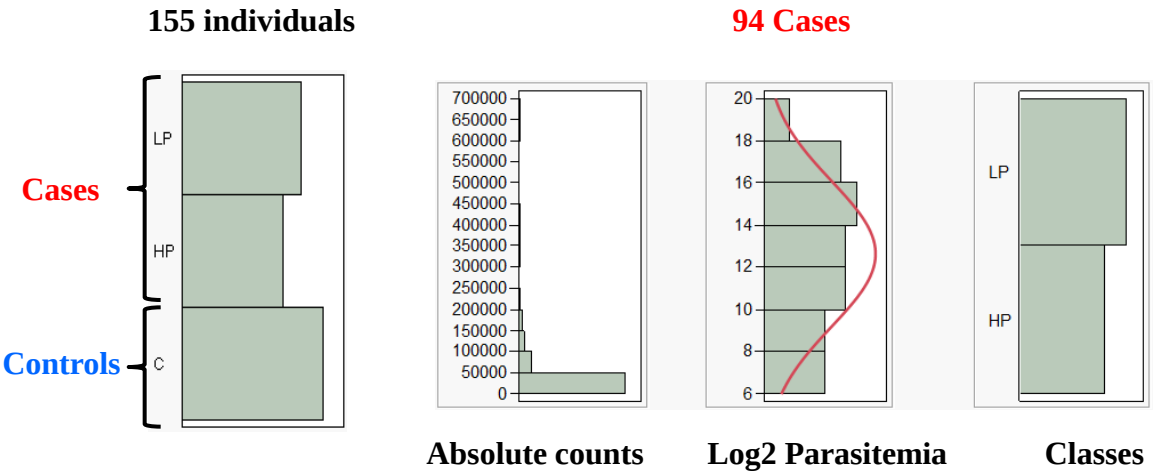


Figure S2. Malaria infection status and parasite load. A total of 94 children undergoing the symptomatic phase of *Plasmodium falciparum* malaria infection and 61 controls were sampled. Thick blood smears and the Parascreen™ malaria rapid detection test were used to determine malaria infection status and parasite laod. Absolute counts of the number of parasites per microlitter of whole blood were log2 transformed. Parasite load was also considered as a discrete categorical variable where infected individuals are classified as high or low parasitemia based on the median value as a cut-off.

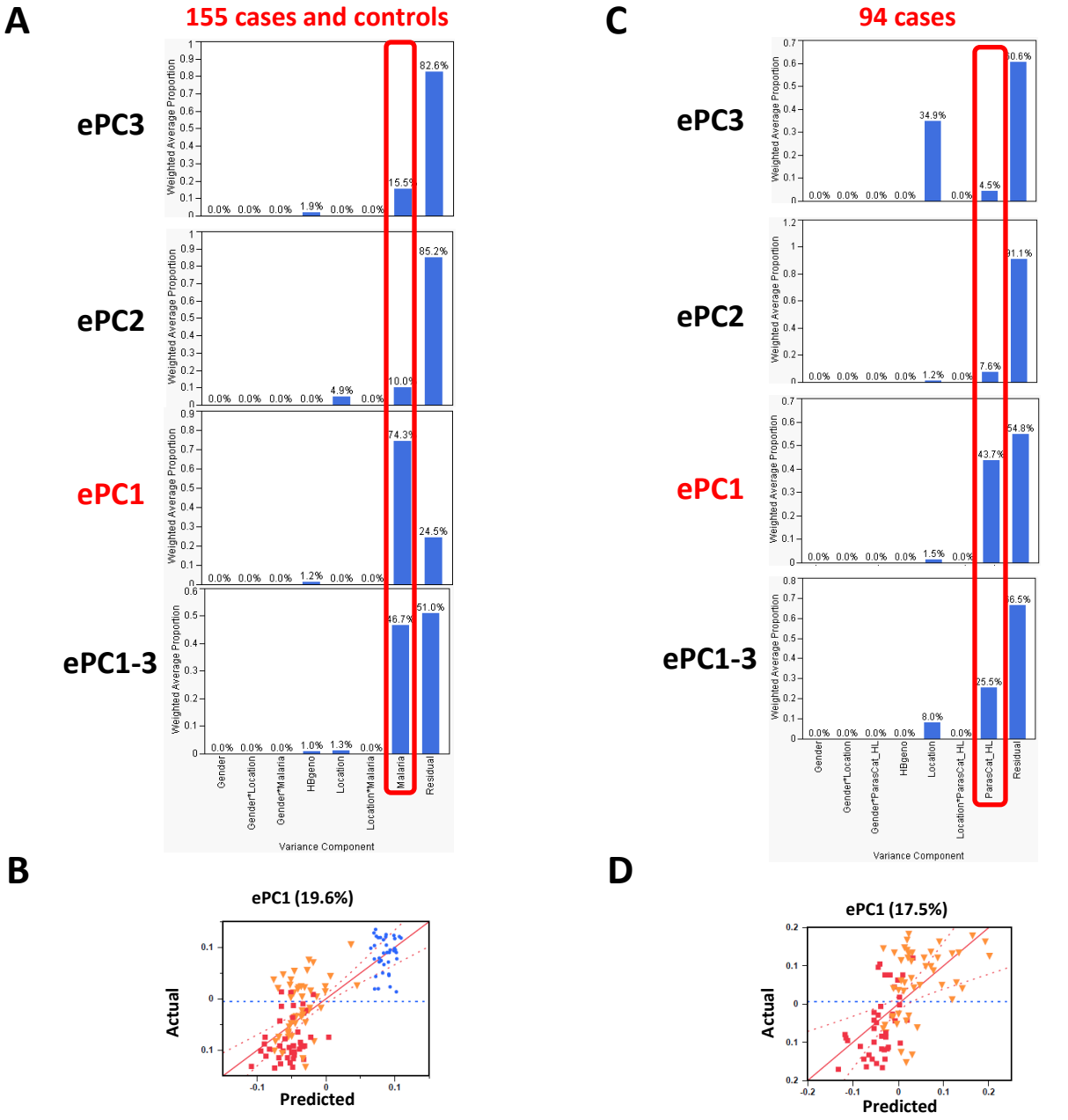


Figure S3. Variance principal component (VPC) analysis of whole-transcriptome data (**A** and **C**) in the 155 cases and controls and in the 94 infected individuals, respectively. Expression principal components 1-3 (ePC1-3) were modeled simultaneously or individually in the VPC analysis as a function of all fixed effects in the data indicated in the X axis. The bars indicate the weighted average proportion of the explained variance for each of the effects. **B** and **D** show the actual vs predicted plots from multiple regression analysis of the major expression Principal Component ePC1 in the 155 cases and controls and in the 94 infected individuals, respectively. This analysis accounts for all fixed in addition to total blood cell counts and ancestry (genotypic PC1-3, see Figure S1). Colors in plots **B** and **D** indicate malaria infection status (blue, controls; red, high parasitemia class and orange, low parasitemia class).

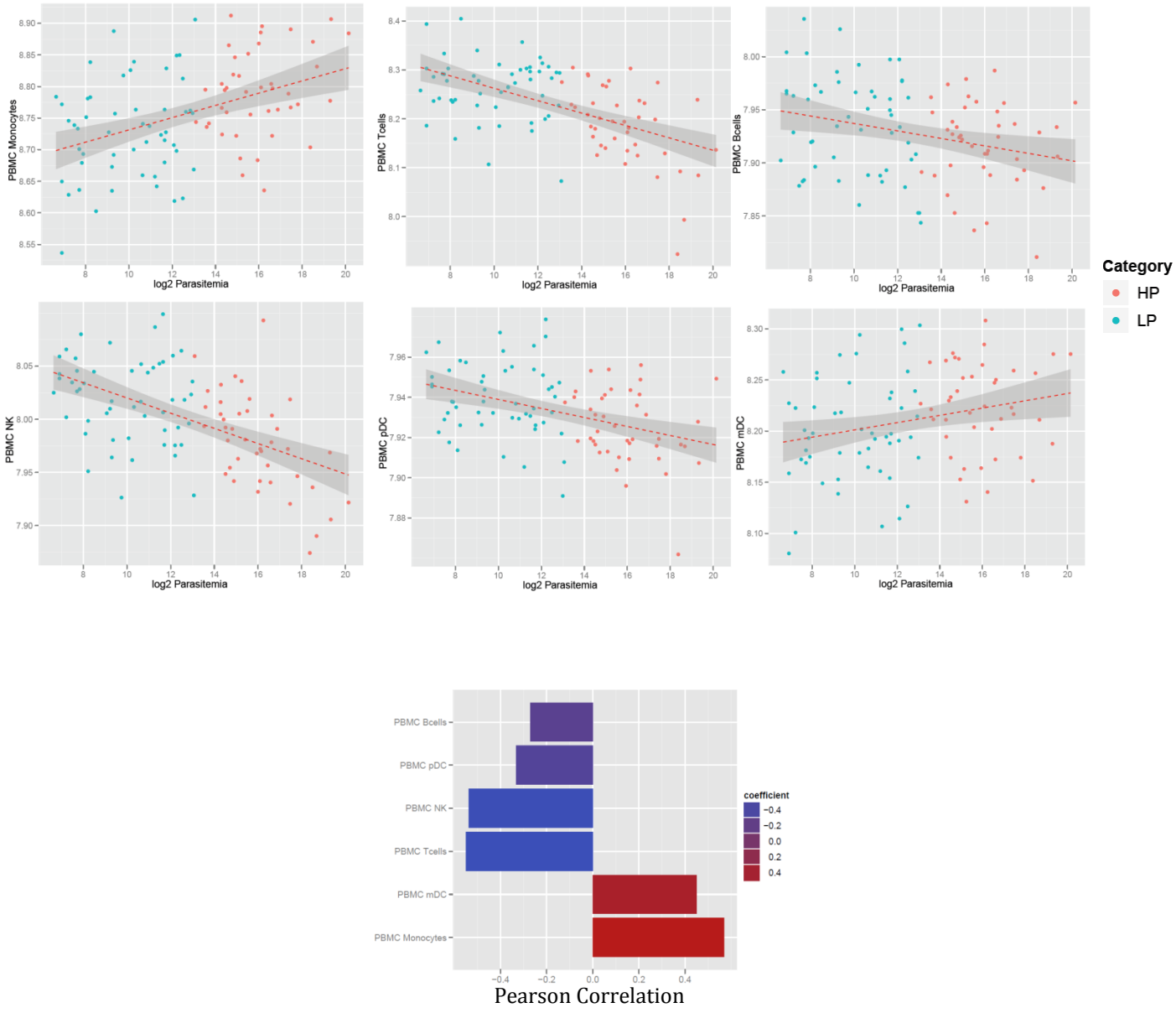


Figure S4. Correlation between parasitemia and average transcript abundance of each cell type-specific modules across all 94 infected individuals. Cell type-specific modules are constructed based on the level of expression levels of each gene relative to each other cell type in peripheral blood mononuclear cells (PBMCs) mixture as described by Nakaya et al. (2010). This analysis shows a significant effect of parasitemia on the six cell type-specific expression profiles investigated (B cells; T cells; mDC, myeloid dentritic cells; pDC, plasmacytoid ; NK, natural killer cells, and monocytes, $P < 10^{-7}$). The two parasitemia groups are shown in different colours as indicated

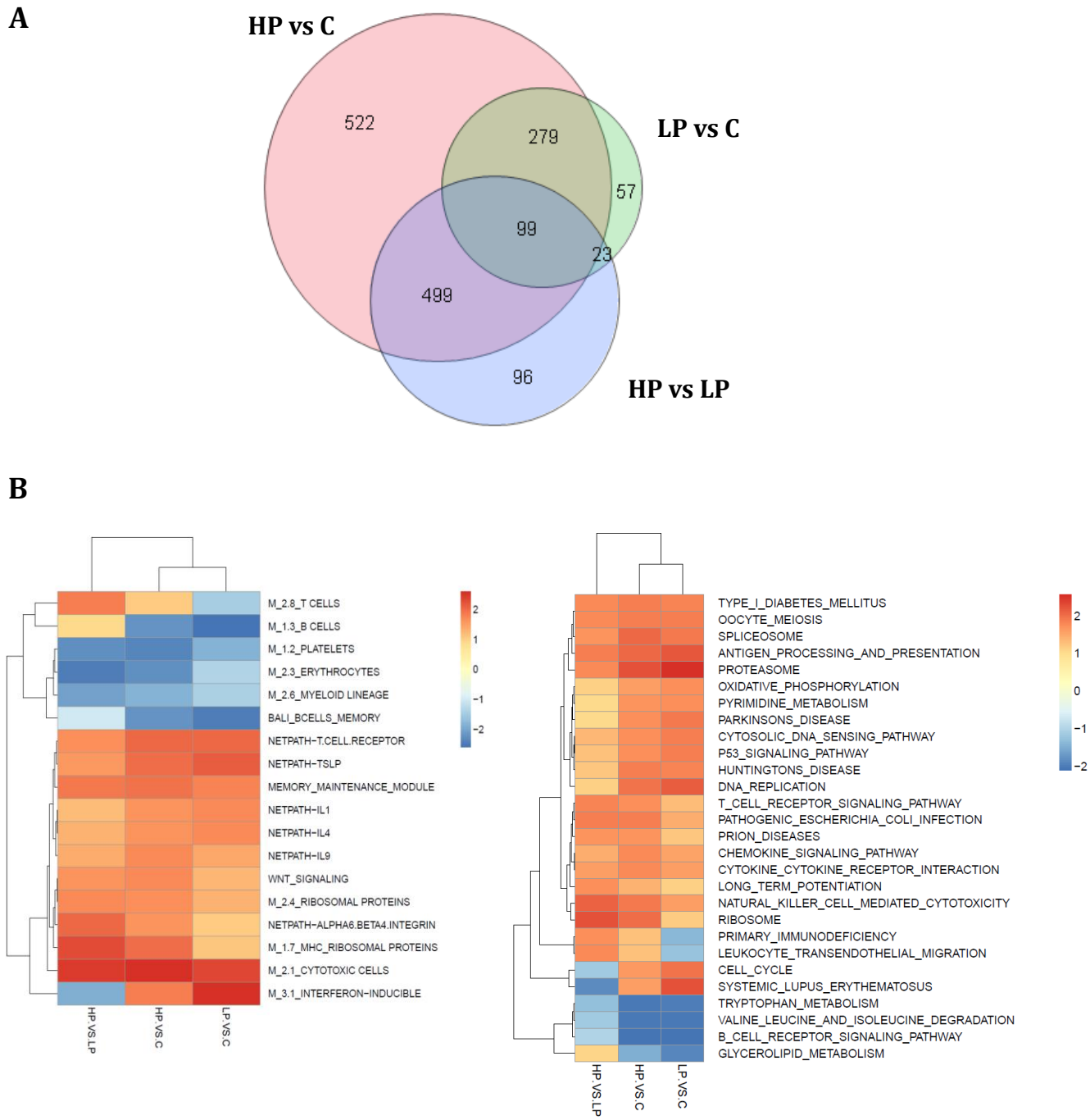
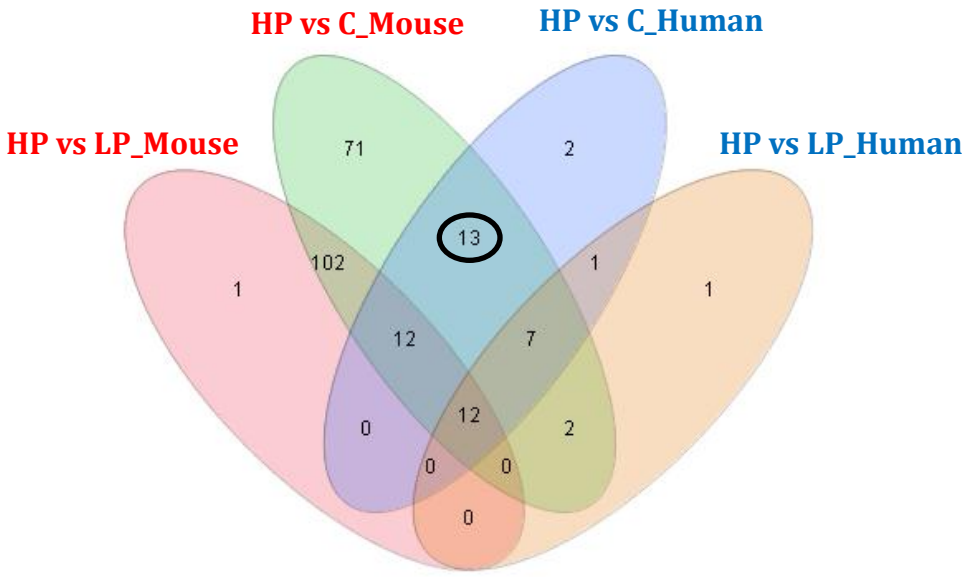


Figure S5. Differential expression in mouse whole blood transcriptome (**A**) The Venn diagram shows the numbers of differentially expressed transcripts for for the 2-way contrasts between the controls (C), the high parasitemia (HP) and the low parasitemia (LP) groups and the overlaps between them. (**B**) For each contrast, gene set enrichment analysis (GSEA) was performed for the C2, C3 and C5 collections of the MsigDB database. Only pathways and modules significantly enriched (5% FDR) from at least one contrast are shown. Colours in the heat map indicate the enrichment score from the GSEA analysis.

A



B

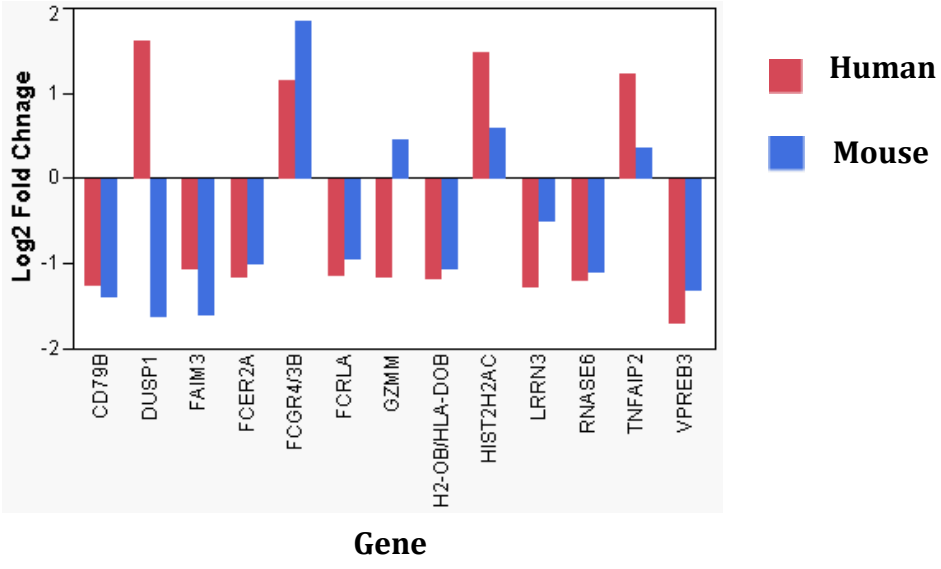


Figure S6. Differential expression in mouse whole blood transcriptome (**A**) The Venn diagram shows the overlap between differentially expressed genes for the 2-way contrasts between the controls (C) and the high parasitemia (HP) group for both hosts. Only genes with fold change > 2 in the human dataset were considered for this contrast. Thirteen genes were significantly regulated in both hosts with fold change > 2 in the human dataset when specifically comparing the high parasitemia group to controls (**B**) Fold change and direction of regulation for the 13 genes highlighted in the Venn diagram.

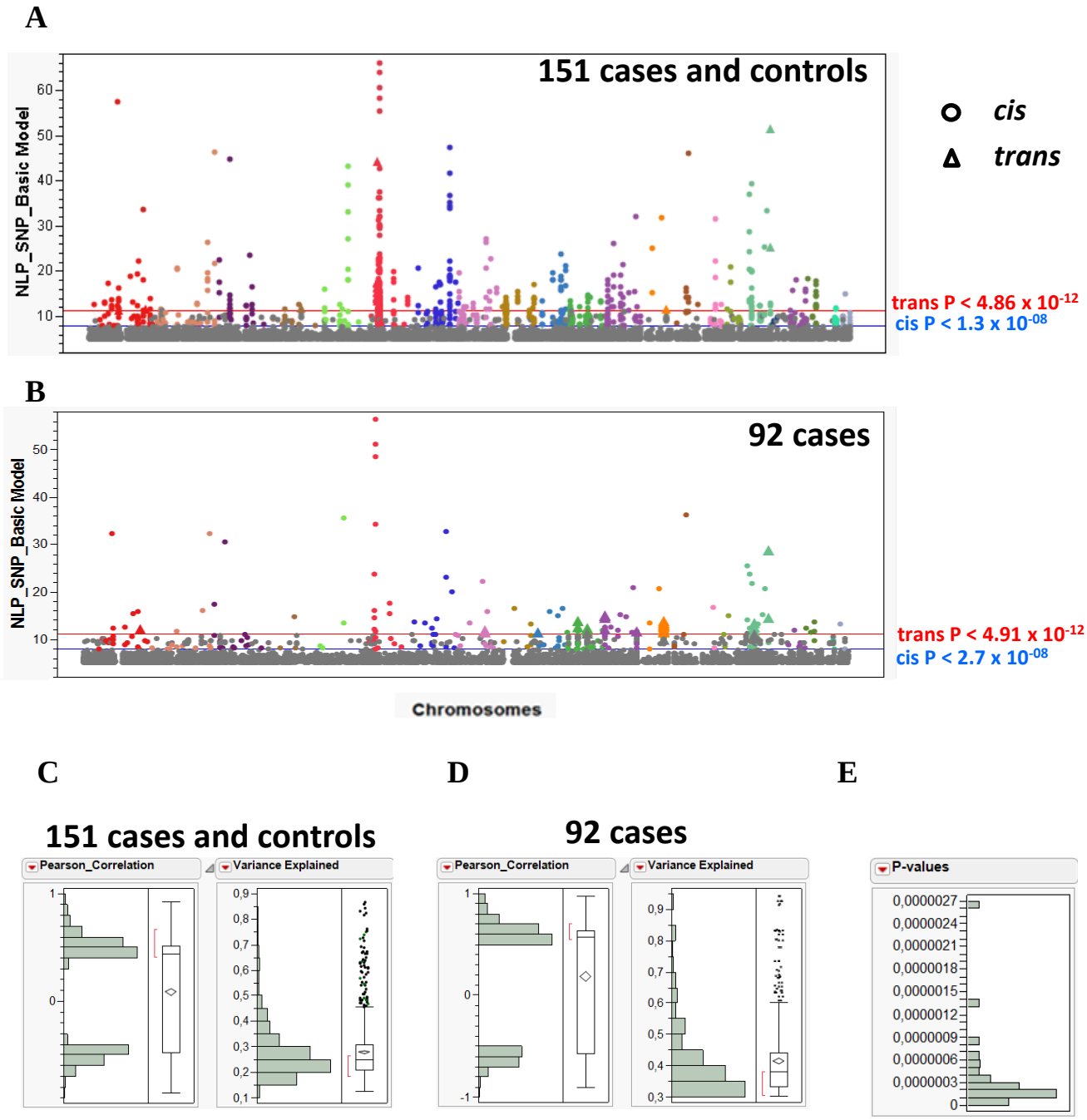


Figure S7. Manhattan plot of all genome-wide significant associations for the basic model for the combined dataset of cases and controls (**A**) and for the cases alone (**B**). Genome-wide significance thresholds for local (circles) and distal (triangles) associations are shown to the right of the plot in blue and red, respectively. Each chromosome is indicated by a different color. Non-significant associations are shown in gray. The distribution of the variance explained and Pearson correlation for the genome-wide significant associations are shown in **C** and **D**. Distribution of statistical thresholds applied to local association depending on the number of SNPs tested against each and accounting for the number of loci tested is shown in **E**. The uncovered associations explain on average 31% and 45% of transcript abundance variance for local and distal associations, respectively.

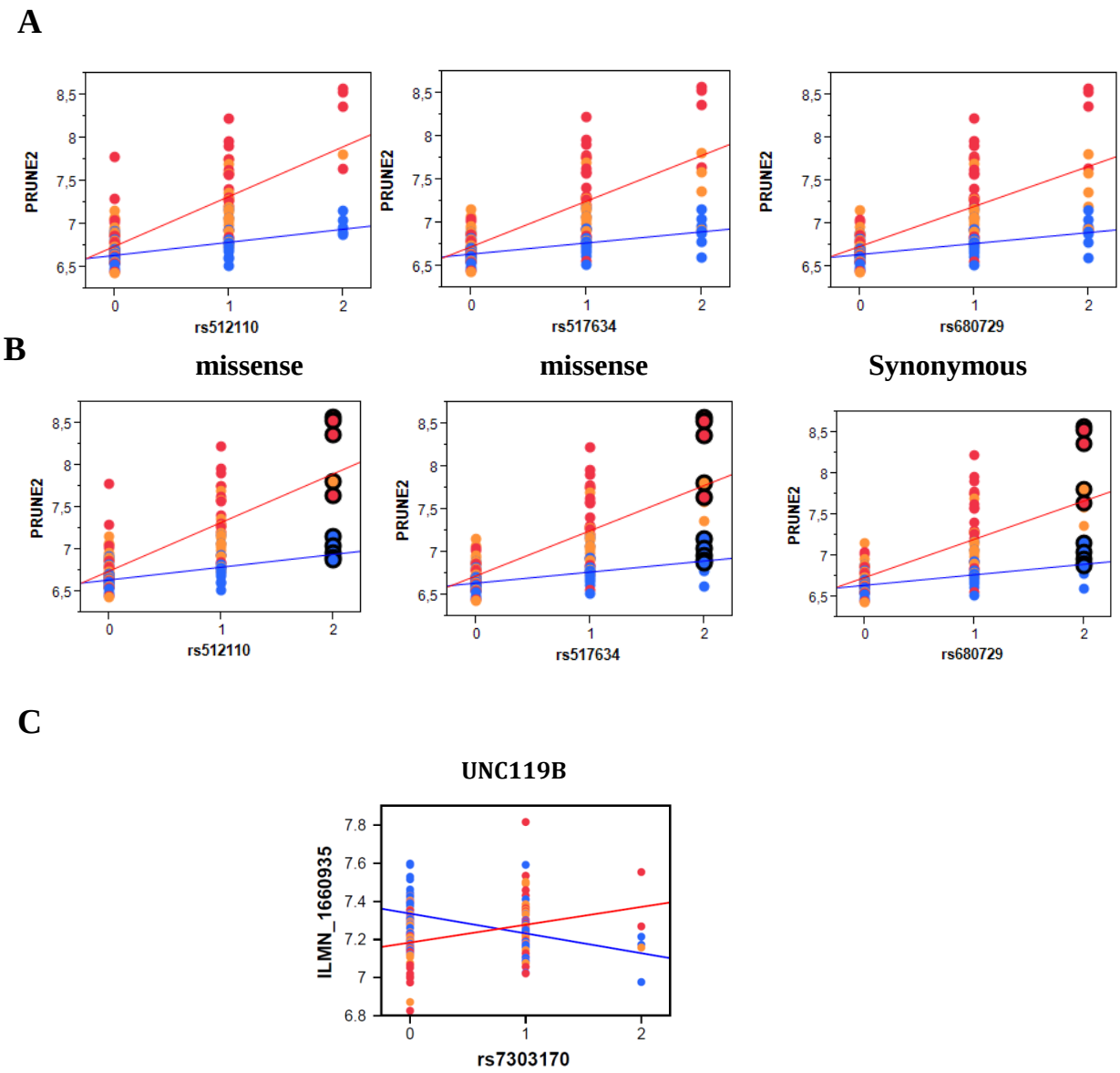


Figure S8. Examples of significant interactions. A robust interaction involves the pro-apoptotic gene *PRUNE2* (**A**) and implicates three coding genotypes, two of which are missense mutations. Stratification of individuals according to haplotype classes shows the presence of haplotypes with the two missense variants in perfect linkage disequilibrium (*D'*=1) as indicated by the outlined circles shown in (**B**). The interaction involving the gene *UNC119B* is shown in **C**. All genotypes implicated in these associations have a MAF > 10% within both the controls and the infected group. Colors indicate malaria infection status (blue, controls; red, high parasitemia class and orange, low parasitemia class).

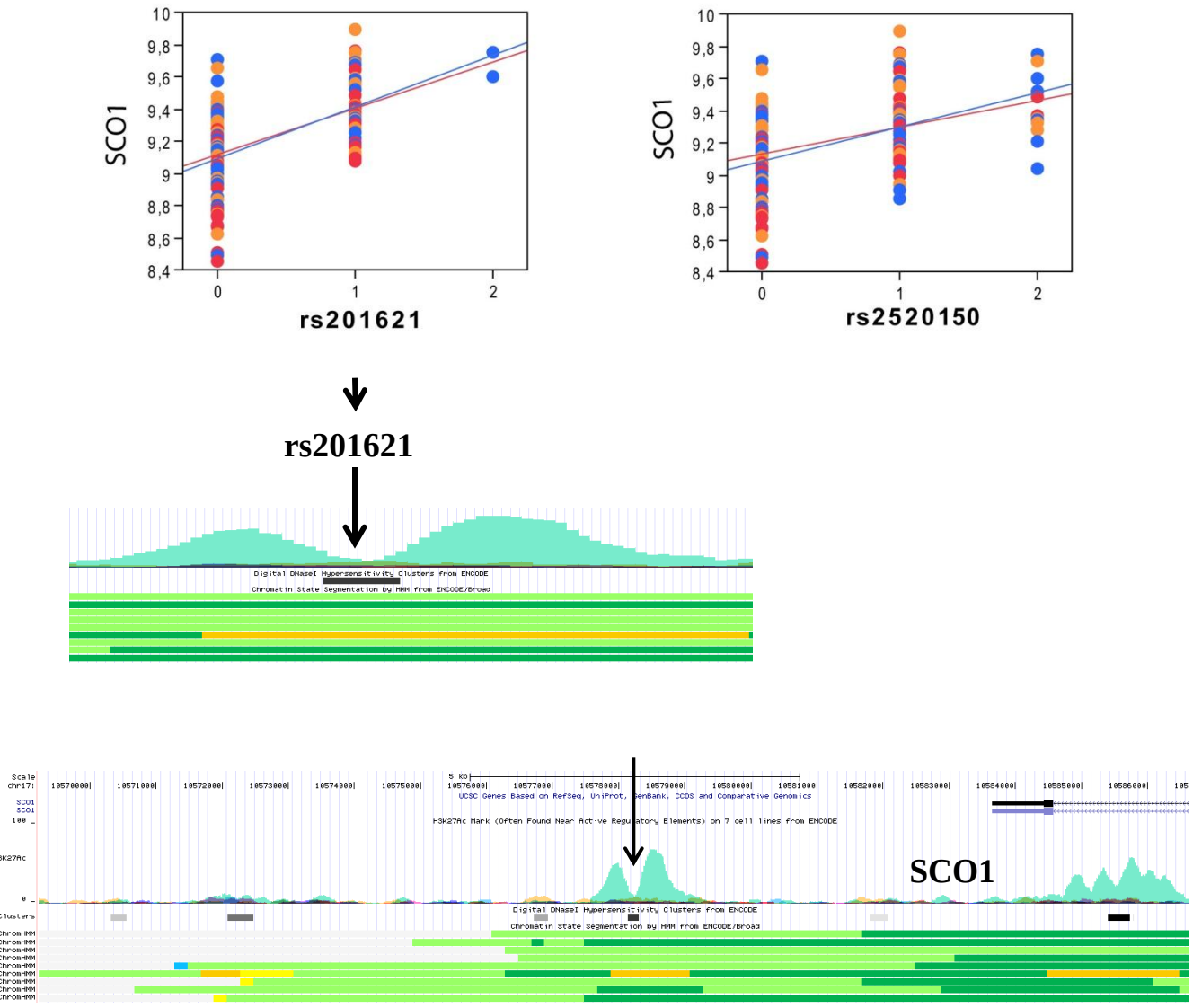


Figure S9. Two genome-wide significant eQTL associations associations implicating the *SCO1* gene previosely identified as a candidiate in a malaria GWAS by Jallow et al. (2009). The strongest eQTL we detected (rs201621, $P = 8.91 \times 10^{-14}$) is located 4Kb upstream of the *SCO1* transcription start site in a functional non-coding element as evidenced by the ENCODE data. Chromatin state analysis (Ernst & Kellis 2010, Ernst et al 2011) classifies the element as a strong enhancer. This finding implicates allelic variation of rs201621 in the effect captured by the malaria GWAS likely through contribution to differential expression of *SCO1*. Differential expression of this gene, although marginal (FDR 8%), detected between the high parasitemia group and both the controls and the low parasitemia group hints to a protective effect conferred by higher expression levels. This effect is likely implicating detoxification pathways of reactive oxygen species . The GWAS SNP rs6503319 was interrogated in our sample but is not in LD with our top eSNP hit ($D' = 0.17$) explaining its lack of association with the regulatory effect robustly captured by eSNP rs201621. The latter was not interrogated in the platform (Affymetrix 500k Array) used in the GWAS.

Table S1. Significant genotype-by-infection interactions. These results were obtained from the analysis of the combined dataset of 151 cases and controls (Model 2). The minor allele frequency (MAF) is reported for each eSNP and for each tested group. Only peak associations are shown and significance is shown as –log10 *P* value (NLP).

Gene	Gene Chr	eSNP	eSNP Chr	eSNP Status	MAF Cases	MAF Controls	MAF Combined dataset	Pearson_CorrModel1	NLP_eSNP Model 1	NLP_eSNP Model 2	NLP_eSNP_by_infection Model 2
PRUNE2	9	rs517634	9	MISSENSE	0.30	0.29	0.29	0.58	14.36	16.96	8.37
SLC39A8	4	rs9331	4	UTR	0.47	0.47	0.47	0.55	12.49	13.65	6.07
C3AR1	12	rs2072449	12	INTERGENIC	0.46	0.47	0.46	0.37	5.50	10.04	5.75
PADI3	1	rs2501799	1	INTRON	0.32	0.28	0.31	<0.01	0.31	0.32	5.79
UNC119B	12	rs7303170	12	INTERGENIC	0.25	0.21	0.23	<0.01	0.05	0.05	5.66

Other Supporting Information Files

Dataset S1. List of significant probes and the results of covariance of the human dataset using the model ANCOVA II.

Dataset S2. List of significant probes and the results of analysis of variance of the mouse dataset.

Dataset S3. All 268 eSNP local and distal peak associations from the analysis of the combined dataset of cases and controls using Model 1. Gene differentially expressed for the malaria infection effect are indicated.

Dataset S4. All 155 eSNP peak associations from the analysis of cases using Model 1. Gene differentially expressed for the malaria infection effect are indicated.

Dataset S5. Characteristics of the samples included in this study.