



**HAL**  
open science

## First genome-wide association study of non-severe malaria in two birth cohorts in Benin

Jacqueline Milet, Anne Boland, Pierre Luisi, Audrey Sabbagh, Ibrahim Sadissou, Paulin Sonon, Nadia Domingo, Friso Palstra, Laure Gineau, David Courtin, et al.

► **To cite this version:**

Jacqueline Milet, Anne Boland, Pierre Luisi, Audrey Sabbagh, Ibrahim Sadissou, et al.. First genome-wide association study of non-severe malaria in two birth cohorts in Benin. *Hum Genet*, 2019, 138 (11-12), pp.1341-1357. 10.1007/s00439-019-02079-5 . cea-04457191

**HAL Id: cea-04457191**

**<https://cea.hal.science/cea-04457191v1>**

Submitted on 18 Jun 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **First genome-wide association study of non-severe malaria in two birth cohorts**  
2 **in Benin**

3 Jacqueline Milet<sup>1\*</sup>, Anne Boland<sup>2\*</sup>, Pierre Luisi<sup>3,4</sup>, Audrey Sabbagh<sup>1</sup>, Ibrahim Sadissou<sup>5</sup>, Paulin  
4 Sonon<sup>5</sup>, Nadia Domingo<sup>6</sup>, Friso Palstra<sup>1</sup>, Laure Gineau<sup>1</sup>, David Courtin<sup>1</sup>, Achille Massougbodji<sup>6</sup>,  
5 André Garcia<sup>1,6\*\*</sup>, Jean-François Deleuze<sup>2\*\*</sup>, Hervé Perdry<sup>7\*\*</sup>.

6 <sup>1</sup>MERIT, IRD, Université Paris 5, Sorbonne Paris Cité, Paris, 75006, France; <sup>2</sup> Centre National de  
7 Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université  
8 Paris-Saclay, F-91057, Evry, France ; <sup>3</sup>Centro de Investigación y Desarrollo en Inmunología y  
9 Enfermedades Infecciosas, Consejo Nacional de Investigaciones Científicas y Técnicas, Córdoba,  
10 Argentina ; <sup>4</sup>Facultad de Filosofía y Humanidades, Universidad Nacional de Córdoba, Argentina ;  
11 <sup>5</sup>Faculty of Medicine of Ribeirão Preto, University of São Paulo, Brazil; <sup>6</sup>Centre d'Etude et de  
12 Recherche sur le Paludisme Associé à la Grossesse et l'Enfance, Faculté des Sciences de la Santé,  
13 Cotonou, Bénin ; <sup>7</sup>Université Paris-Saclay, Centre de recherche en Epidémiologie et Santé des  
14 Populations, Institut National de la Santé et de la Recherche Médicale, Villejuif, France

15 \* Joint first authors

16 \*\* Joint last authors

17 The authors have no competing financial interests.

## 18 **Abstract**

19 Recent research efforts to identify genes involved in malaria susceptibility using genome-wide  
20 approaches have focused on severe malaria. Here we present the first GWAS on non-severe malaria  
21 designed to identify genetic variants involved in innate immunity or innate resistance mechanisms.  
22 Our study was performed on two cohorts of infants from southern Benin (525 and 250 individuals  
23 respectively) closely followed from birth to 18-24 months of age, with an assessment of a space- and  
24 time-dependent environmental risk of exposure. Both the recurrence of mild malaria attacks and the  
25 recurrence of malaria infections as a whole (symptomatic and asymptomatic) were considered. Our  
26 study highlights a role of *PTPRT*, a tyrosine phosphatase receptor involved in STAT3 pathway and  
27 several other genes whose biological functions are relevant in malaria infection. Results shows that  
28 GWAS on non-severe malaria can successfully identify new candidate genes and inform  
29 physiological mechanisms underlying natural protection against malaria.

## 30 **Introduction**

31 In spite of numerous prevention and control efforts in recent years, malaria remains a major global  
32 public health problem with 219 million cases and ~ 435,000 deaths in 2017 (1). In Africa, Bhatt et  
33 *al.* (2) have estimated that *Plasmodium falciparum* infection prevalence halved between 2000 and  
34 2015, and that the incidence of clinical disease fell by 40%, owing mainly to the large distribution  
35 of insecticide-treated nets, the most widespread intervention. However the fight against malaria is  
36 currently facing numerous challenges. A decrease in the global reduction of malaria cases and  
37 deaths is observed at present, with no significant progress made over the 2015–2017 period. The  
38 WHO African region which represents the region with the highest malaria burden (92% of malaria  
39 cases and 93% of malaria-related deaths in 2017) is also the area where medical advances have been  
40 hardest to achieve recently. Moreover control efforts are threatened by insecticide resistance and the  
41 possible spread of resistance to artemisinin (an essential component of the most efficient anti-  
42 malaria drugs, the artemisinin-based combination therapy -ACT) from Southeast Asia to Africa. In

43 absence of alternative treatments with the same efficacy and tolerability as ACT and of an efficient  
44 vaccine, fundamental malaria research continues to be essential in order to better understand the  
45 physiopathology of the disease.

46 *P. falciparum* is the most prevalent malaria parasite in sub-Saharan Africa (99.7% of malaria cases  
47 in African region). Different clinical presentations of *P. falciparum* malaria exist, from  
48 asymptomatic (parasite carriage without any clinical sign), or mild forms (parasitemia with fever),  
49 to severe forms which may ultimately lead to death. This variability of clinical presentation is  
50 thought to be attributable to environmental factors, as well as parasite and human host factors,  
51 among which genetic variation could play a major role (3,4). In 2005, Mackinnon et al. (5)  
52 estimated that 24% and 25% of the variability in the incidence of mild malaria and severe malaria,  
53 respectively, could be explained by genetic factors. Numerous genetic epidemiological studies have  
54 attempted to identify gene polymorphisms associated with susceptibility or resistance to different  
55 malaria phenotypes (see (4,6,7)).

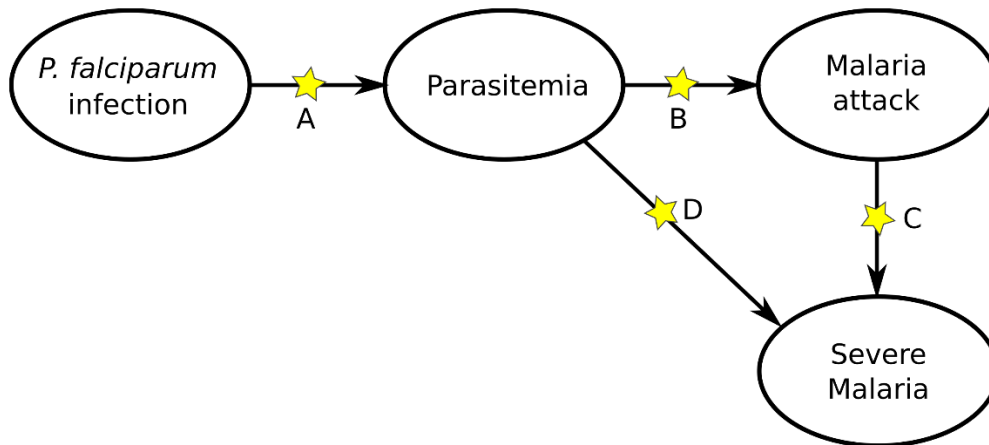
56 Most studies performed to date have focused on severe malaria, despite the fact that mild forms  
57 represent the major part of the global burden. Candidate gene studies for mild forms focused on  
58 genetic polymorphisms involved in host immune response (*IL10*, *IL3*, *LTA*), in genes possibly  
59 involved in oxidative stress (*HP*), red blood cell (RBC) invasion by parasites (*CR1*, *GRK5*) or RBC  
60 defects. Among those, sickle cell trait (haemoglobin S, *HbS*), haemoglobin C (*HbC*) at homozygote  
61 state, alpha+ thalassemia and glucose-6-phosphate dehydrogenase (*G6PD*) deficiency have been  
62 associated with a protection against parasite invasion and clinical malaria attacks (7).

63 Other association studies have focused on chromosomal regions linked with parasite infection levels  
64 and mild malaria susceptibility, in particular 5q31-33 and 6p21-23 (8–15). In this last region, *TNF*  
65 has been the most studied candidate; *NCR3*, encoding a cell membrane receptor of natural killer, has  
66 been also repeatedly associated with mild malaria (10,16). Since 2010, several genome-wide  
67 association studies (GWAS) and meta-analysis have been published on severe malaria (17–21).

68 These studies confirmed the involvement of previously known susceptibility genes (*HBB* and *ABO*)  
69 and revealed new genes (*ATP2B4*, the cluster of genes *GYP/FREM3*, among which *GYP A* and  
70 *GYP B* appear to play a central role (21)).

71 Here we present the first GWAS performed on mild malaria susceptibility. It was conducted on two  
72 cohorts of infants in southern Benin, used as discovery and replication cohorts (525 and 250  
73 individuals respectively) and genotyped with the high density Illumina® HumanOmni5 beadchip.  
74 This study intended to identify genetic factors that play an early role in disease development (Figure  
75 1), in cohorts of infants followed from birth to 18-24 months of age, thus targeting factors involved  
76 in innate immunity or innate resistance mechanisms.

77 In our study, infants were closely followed with symptomatic and asymptomatic malaria infections  
78 recorded. Furthermore geographical, climatic, behavioral and entomological data were collected in  
79 concomitant environmental risk surveys throughout the follow-up, to estimate a spatial- and time-  
80 dependent risk of exposure. Association studies were performed with the recurrence of mild malaria  
81 attacks (RMM) and the recurrence of malaria infections (RMI) including mild malaria attacks and  
82 asymptomatic infections, taking into account a time-dependent risk of exposure. We find  
83 convincing evidence supporting the involvement of *PTPRT*, *MYLK4*, *VENTX* and *UROCI* as well  
84 as strong association signals in *PTPRM*, *ACER3* and *CSMD1* all of which are relevant putative  
85 candidate genes in malaria physiopathology.



86

87 **Figure 1. The path from *P. falciparum* infection to severe malaria, through parasitemia and**  
88 **malaria attacks (parasitemia with fever). The process may stop at each transition A, B, C and D,**  
89 depending on several factors including the genetics of the host. Studies on severe malaria mainly  
90 target factors involved in transitions C and D, while studies on mild malaria focus on A and B. Our  
91 study focuses on transitions A and B in a population of infants.

## 92 **Results**

### 93 **Phenotypic and genotypic data**

94 Infants from both discovery (Tori-Bossito) and replication (Allada) cohorts come from southern  
95 Benin, from two sites located 20 km from each other, 40-50 km north-west of Cotonou, the  
96 economic capital. Ethnicity composition, based on self-reported ethnicity of the mother,  
97 significantly differed between the two study sites. The main ethnic groups were Tori (74%) and Fon  
98 (11%) in the first cohort whereas they were Aïzo (68%) and Fon (22%) in the second one. Each of  
99 the other ethnic groups accounted for less than 10% of individuals in each cohort.

100 Principal Component Analysis (PCA) performed on the two samples did not reveal any ethnic  
101 outliers and showed a low degree of stratification (Figure S1). PC1 separated the two cohorts and  
102 PC2 likely reflected some heterogeneity in the Tori-Bossito cohort. We observed that infants did not  
103 cluster by reported ethnicity, but rather by cohort and health center, indicating that the self-reported

104 ethnicity of the mother is not a major factor of genetic heterogeneity in our samples. The PCA  
105 including 1000 Genomes Project (KGP) populations (as described in methods) confirmed that our  
106 two cohorts are homogeneous (Figure S2). As expected, our two populations overlapped and  
107 clustered with Nigerian populations (Yoruba and Esan).

108 Main characteristics of infants were compared between samples included in or excluded from  
109 GWAS (Tables S1-S2). For the discovery cohort, no significant differences were observed except  
110 for ethnic group composition which appeared to be enriched in Tori at the expense of less prevalent  
111 ethnic groups ( $p < 0.04$ ). For the replication cohort, only infants followed between 12 and 24 months  
112 of age were included in the GWAS. Thus a higher proportion of infants were excluded from  
113 analyses (infants lost to follow-up during the first year). We observed in our sample a lower  
114 proportion of infants with low birth weight and born to mothers who reported having never attended  
115 school ( $p < 0.01$  and  $p = 0.06$  respectively).

116 For the 525 infants included in the discovery sample, mean (SD) length of follow-up was 16.9  
117 months (2.83). A total of 342 infants (65.1%) experienced at least one mild malaria attack (from one  
118 to 10 episodes) and 359 infants (68.4%) were observed with at least one malaria infection (range  
119 from one to 16 malaria infections by infant). For the replication cohort of 250 infants, mean (SD)  
120 length of follow-up was 11.9 months (1.72). A proportion of 83.2% and 86.8% children experienced  
121 at least one malaria attack (range 1 to 9), or at least one malaria infection (range 1 to 14) during the  
122 follow-up, respectively.

### 123 **Adjusting on main epidemiological and environmental determinants**

124 Genome-wide association analyses were performed in two steps because computational burden  
125 inherent to the random effect Cox model (23) makes it inappropriate for the large number of tests in  
126 GWAS.

127 In the first step, the recurrence rate of malaria episodes was modeled using a random effect Cox

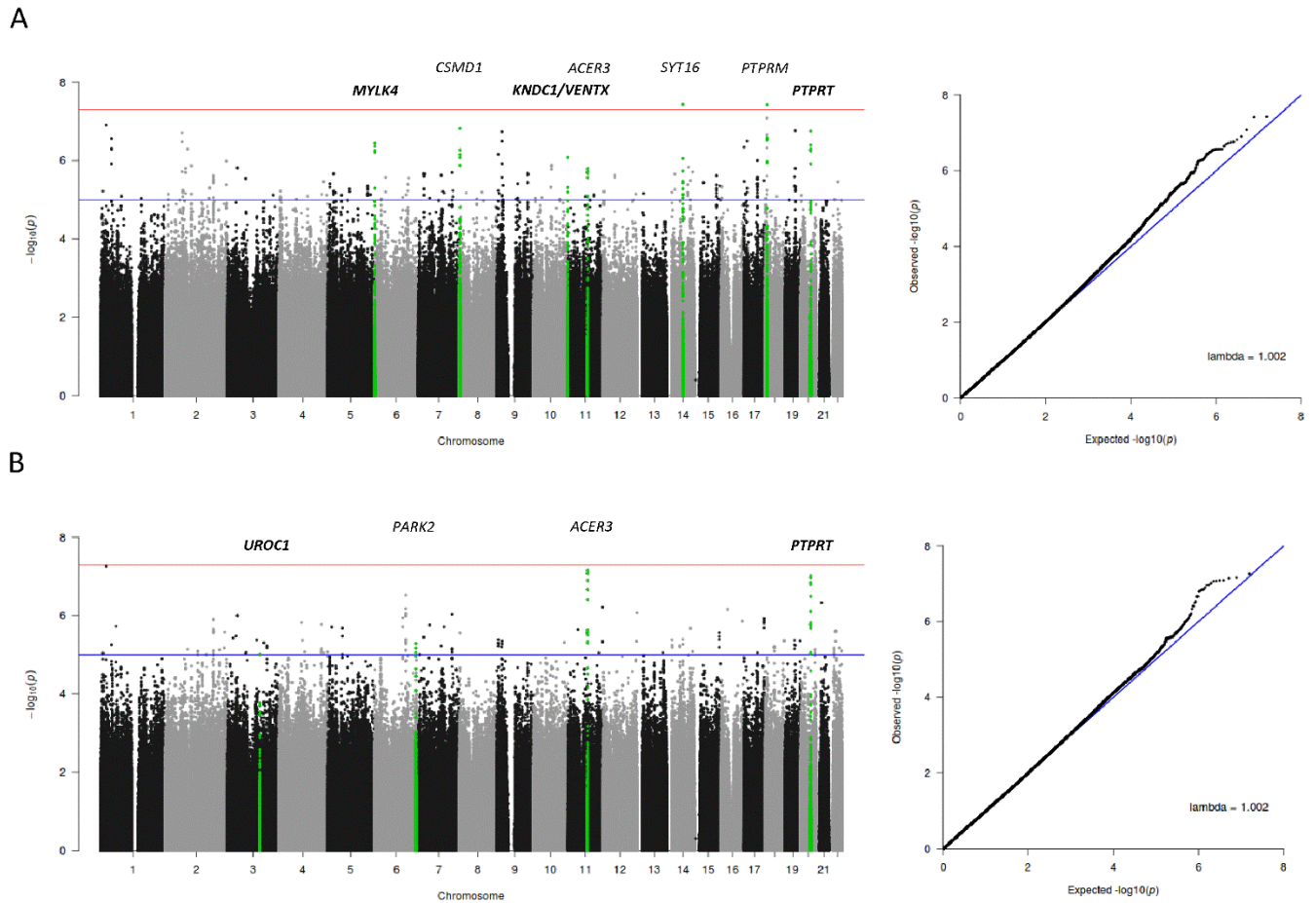
128 model for recurrent events, taking into account epidemiological and environmental factors that  
129 might influence malaria infections in infants. A few covariates among all those considered were  
130 retained in the final model (Tables S3-S4). In each cohort, the same covariates were found  
131 associated for the recurrence of mild malaria attacks (RMM) and the recurrence of malaria  
132 infections (RMI). The levels of exposure to vector bites, health centers and transmission seasons  
133 were associated with the risk of infection in both cohorts. In addition, for the Allada cohort,  
134 significant effects were observed for marital status and for maternal education level. No effects of  
135 placental infection (detected by a thick and thin placental smears), low birth weight (<2500 g) or  
136 HbS allele carriage were detected.

### 137 **Genome-wide association analyses**

138 After imputation using the KGP reference panel (phase 3), 15,566,900 high quality ( $R^2 > 0.8$ )  
139 variants with a minimum allele frequency (MAF)  $\geq 0.01$  were available for analysis. Association  
140 was tested at a genome-wide level using a linear mixed model, with confounder-adjusted  
141 phenotypes constructed from the model built at the previous step. These adjusted phenotypes  
142 correspond to individual effects which are not explained by the epidemiological covariates. The  
143 genomic inflation factor was consistent with a reliable adjustment for cryptic relatedness in our  
144 samples, for both analyses (Figure 2).

145 The analysis revealed four association signals with a  $p$ -value very close to the genome-wide  
146 association threshold of  $5 \times 10^{-8}$ : two signals for RMM (Figure 2A), located in *SYT16* (14q23.2  
147 region, lead SNP rs375961263,  $p=3.7 \times 10^{-8}$ ) and in *PTPRM*, a gene encoding a receptor-type  
148 protein tyrosine phosphatase (18p11.23 region, lead SNP rs113776891,  $p=3.77 \times 10^{-8}$ ); and two  
149 signals for RMI (Figure 2B): one located in *ACER3*, a gene encoding an alkaline ceramidase  
150 (11q13.5 region, lead SNP rs77147099,  $p=6.85 \times 10^{-8}$ ) and another one located in *PTPRT*, a gene  
151 encoding a second receptor-type protein tyrosine phosphatase (20q12 region, lead SNP  
152 rs111968843,  $p=9.70 \times 10^{-8}$ ).





153

154 **Figure 2. Manhattan plots and QQ plots for the discovery association study. A) GWAS results**

155 **for the recurrence of mild malaria attacks, B) GWAS results for the recurrence of malaria**

156 **infections.** The red line in the Manhattan plots indicates the threshold of significance ( $5 \times 10^{-8}$ ), the

157 blue line the threshold of suggestive association used for replication ( $1 \times 10^{-5}$ ). The blue line in the

158 QQ plots shows the 1:1 regression line (the expected distribution of p-value under the null

159 hypothesis). Lambda is the genomic inflation factor calculated as the ratio of the median of the

160 empirically observed distribution of the test statistic to the expected median.

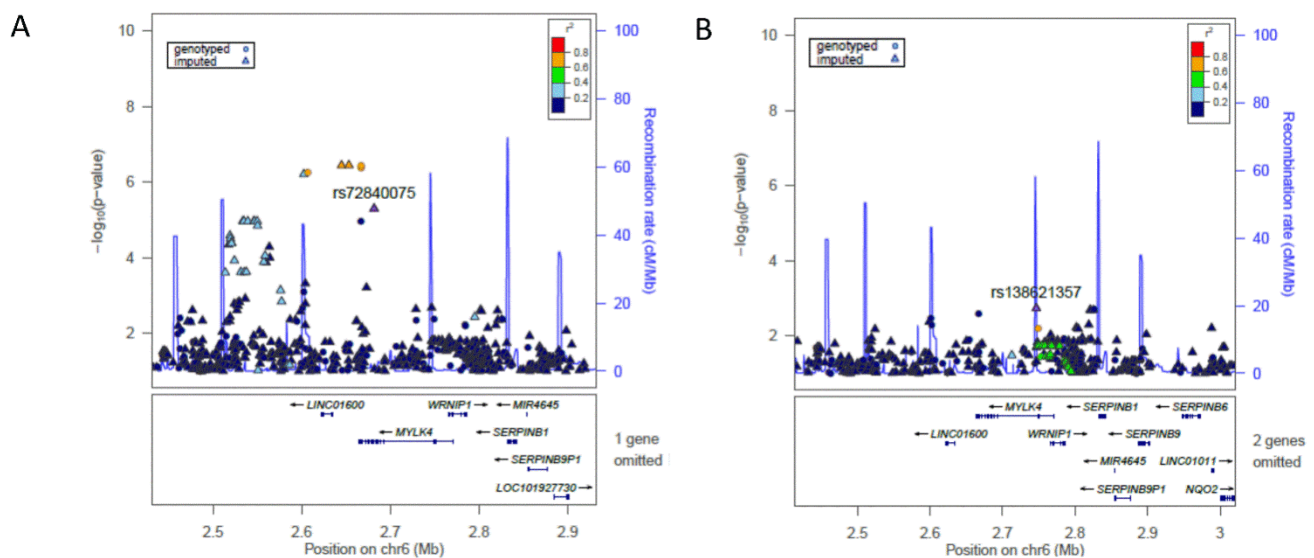
161 For these signals, data were re-analyzed in a single step with a mixed Cox model for recurrent

162 events accounting for cryptic relatedness and relevant covariates, without resorting to intermediate

163 confounder-adjusted phenotypes. Minor differences were observed in p-values and none of them

164 reached genome-wide significance.

A total of 356 and 214 variants were associated at  $p < 1 \times 10^{-5}$  with RMM and RMI respectively (Tables S5-S6), and were tested for replication in the second cohort. SNPs that showed evidence of replication ( $p < 0.05$ ) are presented in Tables 1 and 2. Highest statistical support was observed for SNPs located in the 6p25.2 locus with RMM (3 SNPs, min- $p = 0.0056$ ). The most strongly associated SNP in the replication cohort, rs72840075, is located in an intron of *MYLK4* which encodes a myosin light chain kinase (Figure 3). A meta-analysis of these SNPs showed a borderline significant association for SNP rs142480106 ( $p = 5.29 \times 10^{-8}$ ) (Table 1).



165

166 **Figure 3. Annotated regional association plot for the 6p25.2 locus: A) Association results with**  
167 **the recurrence of mild malaria attacks, pairwise LD are shown with rs72840075, the best**  
168 **associated SNP in replication analysis B) Association results in conditional analysis on**  
169 **rs72840075. Plots show LD calculated on Nigerian Population of KGP dataset (ESN and YRI**  
170 **populations) in the 250 kb region.**

171  
172

**Table 1. Association signals which replicated at  $p < 0.05$  for the recurrence of mild malaria attacks.**

Locus	SNP	Risk allele	MAF	$p$ initial GWA <sup>a</sup>	$p$ Cox model Tori-bossito <sup>b</sup>	HR (CI 95%) Tori-Bossito	$p$ Cox model Allada <sup>b</sup>	HR (CI 95%) Allada	Nearest Genes	$p$ Meta-analysis
6p25.2	rs76088706	T	0.025	$5.56 \times 10^{-7}$	$2.68 \times 10^{-6}$	2.22 (1.59-3.11)	0.036	1.75 (1.03-2.96)	C6orf195(Intergenic)	$3.65 \times 10^{-7}$
	rs142480106	T	0.019	$4.35 \times 10^{-7}$	$1.14 \times 10^{-6}$	2.52 (1.73-3.66)	0.013	2.08 (1.17-3.70)	MYLK4 (UTR3)	$5.29 \times 10^{-8}$
	rs72840075	A	0.030	$5.04 \times 10^{-6}$	$1.05 \times 10^{-5}$	2.03 (1.48-2.79)	0.0056	1.88 (1.20-2.95)	MYLK4 (intronic)	$2.01 \times 10^{-7}$
10q26.3	rs182416945	A	0.12	$8.31 \times 10^{-7}$	$5.56 \times 10^{-6}$	1.55 (1.28-1.87)	0.02	1.28 (1.04-1.58)	KNDC1 (intronic)	$7.94 \times 10^{-7}$
20q12	rs6124419	A	0.12	$1.21 \times 10^{-6}$	$2.51 \times 10^{-6}$	1.53 (1.28-1.83)	0.04	1.26 (1.01-1.58)	PTPRT (intronic)	$6.82 \times 10^{-7}$

173

HR, hazard ratio; MAF, minor allele frequency

174

<sup>a</sup>p-value in GWA analysis performed in two steps using a confounder-adjusted phenotype.

175

<sup>b</sup>p-value of the association calculated with a random effect Cox model accounting for cryptic relatedness and relevant covariates (health center, risk of exposure, transmission season for both cohorts, marital status and education of women in addition for Allada cohort)

176

177

**Table 2. Association signals which replicated at  $p < 0.05$  for the recurrence of malaria infections.**

178

Locus	SNP	Risk allele	MAF	$p$ initial GWA <sup>a</sup>	$p$ Cox model Tori-bossito <sup>b</sup>	HR (CI 95%) Tori-Bossito	$p$ Cox model Allada <sup>b</sup>	HR (95%) Allada	Nearest Genes	$p$ Meta-analysis
3q21.3	rs9871671	A	0.11	$9.79 \times 10^{-6}$	$5.00 \times 10^{-5}$	1.49 (1.22-1.81)	0.04	1.30 (1.01-1.66)	UROCI (exonic)	$8.25 \times 10^{-6}$
20q12	rs6124419	A	0.12	$3.19 \times 10^{-7}$	$1.35 \times 10^{-6}$	1.58 (1.31-1.91)	0.04	1.28 (1.00-1.64)	PTPRT (intronic)	$4.10 \times 10^{-7}$

179

HR, hazard ratio; MAF, minor allele frequency

180

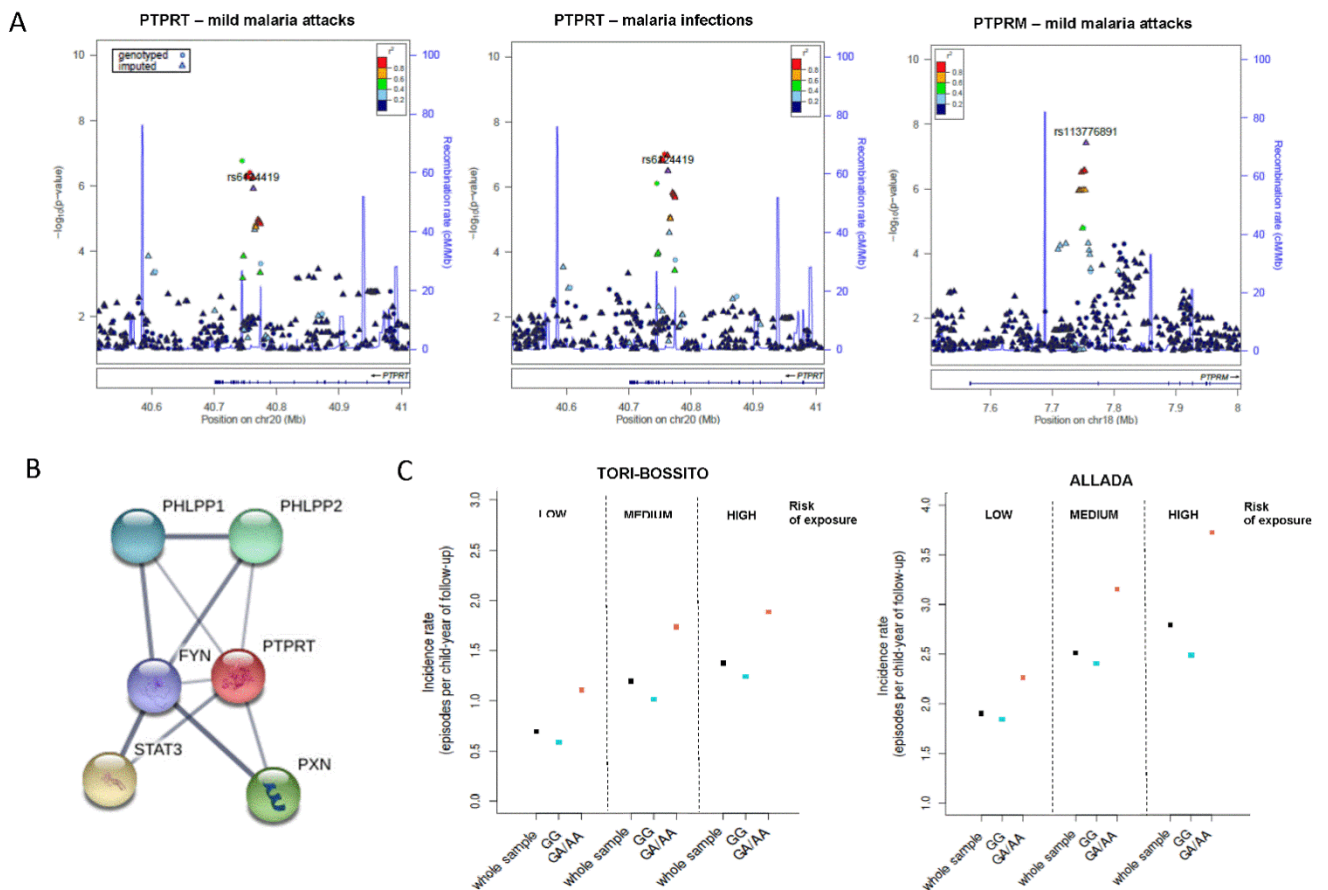
<sup>a</sup>p-value in GWA analysis performed in two steps using a the confounder-adjusted phenotype

181

<sup>b</sup>p-value of the association calculated with a random effect Cox model accounting for cryptic relatedness and relevant covariates (health center, risk of exposure, transmission season for both cohorts, marital status and education of women in addition for Allada cohort)

182

183 The same SNP, rs6124419, located in intron 18 of *PTPRT*, replicates for both phenotypes. The  
 184 regional linkage disequilibrium (LD) plot showed a narrow signal of around 20 kb width in both  
 185 cases which encompasses exons 20 to 22 (Figure 4A). An estimation of the incidence rate of mild  
 186 malaria attacks in three risk groups based on exposition levels showed that the effect of rs6124419  
 187 is consistent regardless of the exposition level and the cohort (Figure 4C). Interestingly *PTPRT* is a  
 188 paralogue of *PTPRM*, for which a highly suggestive association peak was observed with mild  
 189 malaria attacks in the 18p11.23 chromosomal region (Figure 4A).



190  
 191 **Figure 4. A) Annotated regional association plots for *PTPRT* and *PTPRM* loci. B) Interaction**  
 192 **network for *PTPRT* C) Crude incidence rate of mild malaria attacks for rs6124419 genotypes**  
 193 **calculated for three classes of environmental exposure. Plots show LD calculated on Nigerian**  
 194 **Population of KGP dataset (ESN and YRI populations). Replicating SNP are highlighted for**

195 PTPRT. The three classes were defined from the mean environmental exposure risk across all the  
196 follow-up calculated for each infant. Low, medium and high correspond to the three tertiles of the  
197 distribution.

198 Other associations which replicated in the Allada cohort include rs182416945 in *KNDC1* and  
199 rs9871671 in *UROCI* (Table 2). This last one, a missense SNP, is predicted to be probably  
200 damaging with a score of 0.99 in PolyPhen-2(24) and has a Combined Annotation Dependent  
201 Depletion (CADD)(25,26) score of 26.5, indicating that this variant is among the 1% most  
202 deleteriousness substitutions of the human genome.

203 For each signal aforementioned, a regional LD plot for the nearby region ( $\pm$  200 kb) was realized  
204 and a conditional analysis was run (Figure 4A and Figures S3 and S4). For none of them, a  
205 significant residual signal was found.

## 206 **Functional annotation and gene mapping of GWAS results**

207 To identify the genes and variants most likely associated with these four replicating association  
208 signals, and to select potential genes of interest pointed out by non-replicating signals, we used the  
209 FUMA web platform (27). FUMA identifies independent associated genomic regions and prioritizes  
210 genes based on functional consequences of SNPs in each of them.

211 Thirty independent associated regions were identified for RMM and 15 for RMI, based on the  
212 presence of at least 3 SNPs below the  $p$ -value threshold of  $10^{-5}$  in a region (LD threshold to  
213 aggregate SNPs in one signal,  $r^2 = 0.1$ ). Candidate SNPs (among SNPs present either in our data or  
214 in the AFR population of the KGP reference panel) were then defined based on LD ( $r^2 > 0.6$ ) with  
215 the lead SNP in each of these regions, or with one of the replicating SNPs.

216 These candidate SNPs were mapped to genes, by positional mapping (deleterious coding SNPs -  
217 CADD score  $> 12.37$  - or regulatory elements likely to affect binding in non-coding regions -  
218 Regulome database (RDB, (28)) score  $\leq 2$ ), and by eQTL mapping based on expression data from  
219 relevant tissue types for malaria (whole blood, cell-EBV transformed lymphocytes, liver, skin, cells  
220 transformed fibroblasts and spleen). These mappings identified 20 genes for RMM and 8 for RMI

221 (Tables S7-S8). Results for the strongest signals of association with both phenotypes are  
222 summarized in Figure 5. We additionally performed chromatin interaction mapping with high  
223 chromosome contact map (Hi-C) data (29) with FUMA, allowing to identifying genes located in  
224 regions with significant chromatin interaction with the candidate SNPs (Figure 5).  
225 For each of the four replicating loci FUMA identified a single gene, suggesting that these genes are  
226 very likely to be the “causal gene” in these regions. Three of these highlighted genes correspond to  
227 the gene in which the replicating SNP is located: *MYLK4* in 6p25.2 region, *PTPRT* in 20q12 region,  
228 *UROCI* in 3q21.3 region. The last highlighted gene, *VENTX* in 10q26.3 region, is located 17kb  
229 from the lead SNP. In *MYLK4* locus, the replicating SNP, rs72840075, is an eQTL in whole blood  
230 (eQTLGen and BIOS eQTL (30) database, FDR<0.05). Allele at risk in our data is associated with a  
231 higher expression of this gene. One deleterious SNP is observed in *PTPRT* ( $r^2=0.78$  with the  
232 replicating SNP, CADD = 14.39), and in *UROCI* (the aforementioned deleterious SNP). In *VENTX*  
233 locus, two SNPs, rs182416945 (the replicating one) and rs138609386, in complete DL ( $r^2 = 1$ ) with  
234 the first one, were identified as a significant eQTL in whole blood (eQTLGen database). The allele  
235 at risk in our data is associated with a lower expression of *VENTX*, a gene coding for the VENT  
236 Homeobox.  
237 FUMA highlighted also *PTPRM* as the most probable gene associated with the prominent signal for  
238 RMM in 18p11.23 region. Two candidate SNPs, rs138057394 and rs112288193 ( $r^2=0.74$  with the  
239 lead SNP, for both SNPs) are annotated as deleterious coding SNPs (CADD score of 19.58 and  
240 14.64 respectively).  
241 Furthermore, among the other genes pointed out by FUMA (Figure 5) in non-replicating signals,  
242 *ACER3* in 11q35.5 region is identified by the three mapping methods performed. Numerous  
243 significant eQTL associations were observed in the cells transformed fibroblast tissue (GTEx v7  
244 database). For these eQTL, alleles at risk in our data were associated with a lower expression of the  
245 gene. This association signal is the strongest signal for RMI, with a  $p$ -value approaching the  
246 genome-wide significance threshold, ( $p = 6.85 \times 10^{-8}$ , figure 2) and it was also found associated



247 with RMM ( $p = 1.60 \times 10^{-6}$ ).

Phenotype	Locus	Location (start - end)	Lead SNP	MAF	P initial GWAS	Gene	functional consequences on gene							
							pos.	eQTL	CI	positional mapping		eQTL mapping		
										rsID	CADD score	RDB score	nSNPs	eQTL dir.
Mild malaria attacks RMM	6p25.2	2602462-2681733	rs547331171	0.03	$3.61 \times 10^{-7}$	<b>MYLK4</b>							1	+
	10q26.3	135034267-135052777	rs182416945	0.12	$8.31 \times 10^{-7}$	<b>VENTX</b>							2	-
	20q12	40745108-40774357	rs111968843	0.11	$1.21 \times 10^{-6}$	<b>PTPRT</b>				rs144104706	14.39			
	18p11.23	7742657-7754654	rs113776891	0.08	$3.78 \times 10^{-8}$	PTPRM				rs138057394 rs112288193	19.58	2b		
	18p11.31	7012462-7034932	rs143310084	0.02	$8.32 \times 10^{-8}$	LAMA1				rs566655	14.64			
	1p34.2	41902797-41932682	rs75470436	0.01	$2.76 \times 10^{-7}$	CTPS1							1	+
	2p12	80290041-80298707	rs6708548	0.08	$1.58 \times 10^{-6}$	CTNNA2				rs1484465 rs1484464	19.08			
	11q13.5	76570145-76654016	rs77147099	0.02	$1.60 \times 10^{-6}$	ACER3							7	-
	5p14.2	24307924-24461650	rs12519312	0.01	$2.15 \times 10^{-6}$	C5orf17							11	-
	5p14.3	22318243-22620207	rs141690513	0.03	$2.16 \times 10^{-6}$	CDH12				rs141690513	14.03			
Malaria infections RMI	20q12	40745108-40774357	rs111968843	0.11	$9.70 \times 10^{-8}$	<b>PTPRT</b>				rs144104706	14.39			
	3q21.3	126218211-126218211	rs9871671	0.11	$9.79 \times 10^{-6}$	<b>UROC1</b>				rs9871671	26.50			
	11q13.5	76418359-76739604	rs77147099	0.02	$6.85 \times 10^{-8}$	ACER3				rs141181984 rs11608202	13.95	2b	21	-
	6q22.31	122845480-123163778	rs146900853	0.11	$3.00 \times 10^{-7}$	SMPDL3A							1	+
	11q25	133411892-133432324	rs7929384	0.05	$6.14 \times 10^{-7}$	OPCML				rs7126528 rs7126348	17.04			
	2q32.1	188435603-188473204	rs7583251	0.16	$1.25 \times 10^{-6}$	TFPI				rs115976332		2b		
	6q26	161851951-162045282	rs189683911	0.08	$5.16 \times 10^{-6}$	PRKN				rs80324971		2a		
20q13.32	58143026-58143026	rs144135811	0.15	$6.83 \times 10^{-6}$	MED15				rs114819925 rs5758468		1a	1	-	

248

249 **Figure 5. Genes identified by FUMA based on functional consequences of SNPs, in**  
250 **replicating association signals and in the strongest non-replicating ones.** Genomic regions  
251 reported are those with at least one gene identified based on positional mapping or eQTL mapping.  
252 Genes highlighted in bold correspond to replicating signals. For each genomic region the lead SNP  
253 was identified as the replicating SNP or the SNP with the lowest  $p$ -value. The  $p$ -value of the initial  
254 GWA was calculated with a linear mixed model on the adjusted phenotypes. Filled red boxes  
255 indicate whether the gene was identified by positional mapping (pos.), eQTL mapping (eQTL) or  
256 chromatin interaction mapping (CI). For positional mapping, deleterious SNP names are given  
257 (rsID) with the maximum CADD score and RegulomDB score observed for these SNPs. For eQTL  
258 mapping, the number of significant eQTL associations (nSNPs) are reported, together with the  
259 direction of effect allele to gene expression, for the risk increasing allele of GWAS (eQTL  
260 direction).

261

## 262 **Candidate associations**

263 We also tested the statistical significance of genes and variants previously reported in literature as  
264 associated with malaria (Table 3). Only SNP rs40401 in *IL3* was found marginally associated with  
265 both phenotypes ( $p = 0.01$  and  $0.02$  respectively). Moreover, two suggestive peaks are observed in  
266 the candidate regions: one in the *IL3* locus with both phenotypes (10 SNPs encompassing *IL3* and  
267 exon 1 to 3 of *ACSL6*,  $\text{min-p} = 8.89 \times 10^{-5}$  with RMM and  $\text{min-p} = 1.66 \times 10^{-4}$  with RMI, rs7714191)  
268 and one in the *IL10* locus with RMM, 26 kb upstream of the gene (4 SNPs,  $\text{min-p} = 2.04 \times 10^{-4}$ ,  
269 rs116126622).



270 **Table 3. Candidate SNP associations**

Gene	SNP	Variant	Location (GRCh37)	MAF	Change	Malaria attacks		Malaria infections	
						$\beta$ (CI 95%)	p	$\beta$ (CI 95%)	p
<i>HBB</i>	rs334	HbS	11:5248232	0.10	A > T	-0.040	0.24	0.05	0.34
	rs33930165	HbC	11:5248233	0.04	G > A	-0.011	0.83	-0.01	0.95
<i>TNF</i>	rs1799964	3'UTR, <i>TNF</i> -857	6:31574531	0.19	T > C	0.002	0.92	0.02	0.59
	rs1800629	3'UTR, <i>TNF</i> -308	6:31543031	0.11	G > A	0.018	0.11	-0.011	0.80
<i>IL10</i>	rs1800871	3'UTR, <i>IL10</i> -819	1:206946634	0.38	C > T	0.023	0.23	0.02	0.51
<i>IL3</i>	rs40401	Exon 1, missense	5:131396478	0.71	C > T	0.050	0.01	-0.08	0.02
<i>ABO</i>	rs8176746	Exon 10, missense	9:136131322	0.15	C > A	-0.022	0.42	-0.062	0.13
	rs8176719	Exon 9, frameshift	9:136132908	0.27	- > G	-0.027	0.18	-0.038	0.22
<i>ATP2B4</i>	rs1541255	Exon 1	1:203652141	0.41	A > G	-0.006	0.75	-0.035	0.24
	rs10900585	Intron 2	1:203654024	0.46	G > T	0.000	0.96	0.014	0.65

271 MAF, minor allele frequency;

272 Results of association tests performed in the discovery cohort using the confounder-adjusted phenotype.  $\beta$  is the estimate of the SNP effect under an

273 additive model. A negative value indicates a protective effect of the alternative allele and conversely.

## 274 **Discussion**

275 Genetic factors of resistance to malaria may be involved at different stages during the course of the  
276 disease: inoculation of parasite, development of asymptomatic parasitemia, appearance of clinical  
277 symptoms including fever, rapid progress to complications and manifestation of severe malaria (4).  
278 However research efforts to identify host genetic factors have so far mainly focused on severe  
279 malaria, the life threatening form of the disease. While some genetic factors may be in common  
280 between severe and mild forms, studies on severe malaria are not designed to discover genes  
281 affecting specifically the risk of malaria infection and the development of mild malaria. For this  
282 reason, studies based on non-severe malaria forms may help to complete our understanding of the  
283 disease.

284 Our study represents the first screening of genetic variations associated with non-severe forms of  
285 malaria at a genome-wide level. It was conducted on two birth cohorts closely followed with a  
286 protocol of surveillance which allowed to ascertain malaria infections as a whole. Infants were  
287 visited at home weekly or bi-monthly during all the follow-up for the systematic detection of fever  
288 and symptomatic malaria. Moreover, a TBS was performed monthly to detect asymptomatic  
289 infections (23,24). Another strength of our study is the assessment of environmental risk of  
290 exposure at an individual level all along the follow-up. Although there is consistent evidence for  
291 local variation in exposure to *P. falciparum*, which may partly explain the heterogeneity in malaria  
292 infection incidence observed in children (31,32), this factor has been seldom considered in genetic  
293 epidemiology studies. Here, environmental risk of exposure was taken into account in an  
294 unprecedented manner. Entomological, climatic, environmental (information on house  
295 characteristics and on its immediate surrounding) as well as geographical data (soil type,  
296 watercourse nearby, vegetation index, rainfall, etc) were collected all along the follow-up.  
297 Altogether these data allowed modeling, for each child included in the follow-up, an individual risk  
298 by means of a time-dependent variable (33). The benefit of using this estimated risk has been

299 demonstrated in two studies on the impact of placental malaria on infant susceptibility (34,35). This  
300 risk turned out to have a highly significant effect on the recurrence of events in the present study  
301 ( $p < 1 \times 10^{-8}$  and  $p < 1 \times 10^{-6}$  for discovery and replication cohort respectively).

302 We find strong support for the involvement of protein tyrosine phosphatase (PTP) receptors. Highly  
303 suggestive association peaks were observed for *PTPRM* with RMM and association peaks observed  
304 for *PTPRT* replicated for both phenotypes. Interestingly, both genes are paralogues and in each  
305 signal a deleterious coding SNP which could be the causal SNP is identified. Although the  
306 replication of *PTPRT* was weak in the second cohort ( $p = 0.01$ ), this can be explained by the fact that  
307 our replication cohort is relatively small in size, that infants in this cohort were not actively  
308 followed during the first year of life, which limited our analysis to the 12-24 months period.  
309 However, assessment of incidence rate of mild malaria attacks in function of groups of different  
310 level of exposure, showed very consistent results among exposure groups and cohorts which is an  
311 additional argument for a true association. *PTPRT* and *PTPRM* are proteins belonging to the PTP  
312 superfamily, which regulates diverse signaling pathways by catalyzing the removal of a phosphate  
313 group from specific signaling proteins. *PTPRT* and *PTPRM* encodes two PTP receptors of the same  
314 sub-family type IIb (36). These transmembrane proteins have an extracellular domain involved in  
315 cell-cell aggregation and a phosphatase domain which allows intra-cellular signaling. Interestingly  
316 *PTPRT* protein directly interacts with *STAT3* as shown by the STRING interaction Network  
317 (Figure 4 B) (37) and is the only interactor for which a high confidence ( $> 0.70$ ) is reported in the  
318 STRING database. The role of *PTPRT* in regulating *STAT3* pathways has been demonstrated, with  
319 *STAT3* identified as a direct substrate of *PTPRT* (38,39). *STAT3* is a signal transducer and  
320 activator of transcription which behaves similarly to *NF- $\kappa$ B*, the role of which in malaria infection  
321 is well-established. Moreover, *STAT3* has been reported as associated with cerebral malaria  
322 severity in experimental studies (40,41). Liu et al. have demonstrated using murine models that  
323 *STAT3* is activated by *P. berghei* infection. The *STAT3* pathway is thus very likely to be involved  
324 in the first stages of malaria infection and represents a potential drug target.

325 Other genes replicated with  $p < 0.05$  in the validation cohort. *MYLK4* belongs to the family of  
326 myosin lighth chain kinases (MYLKs) that catalyse myosin interaction with actin filaments. Since  
327 members of MYLKs family play an essential role in the organization of the actin/myosin  
328 cytoskeleton, and in cell motility (42), *MYLK4* may play a role in RBC membrane structure. At  
329 10q26.3 locus, replication is observed for a SNP located in *KNDC1*, but two eQTL associations  
330 identified *VENTX* as the most probable effector for this association signal. *VENTX* is a homeobox  
331 transcriptional factor that controls proliferation and differentiation of hematopoietic and immune  
332 cells (43–45). Wu et al. works (44,45) have shown that *VENTX* is a key regulator of macrophage  
333 terminal differentiation and of dendritic cells differentiation and maturation. *VENTX* is up-regulated  
334 during monocyte-to-macrophage differentiation and overexpression of *VENTX* accelerates the  
335 differentiation towards dendritic cells. Conversely, *VENTX* deficiency impairs dendritic cells  
336 differentiation and affects various aspects of macrophage differentiation and function, including a  
337 downregulation of multiple membrane receptors critical for innate and adaptive immunity. Both  
338 macrophages and dendritic cells are innate immune cells derived from monocytes that play an  
339 essential role in the response to malaria infections (Chua et al., 2013). Allele at risk of mild malaria  
340 attacks in our data was associated with a lower expression of *VENTX*, which would negatively  
341 impact the macrophages and/or dendritic cells response to malaria infection in the most susceptible  
342 children. Finally, a high deleterious missense SNP in *UROCI* (Urocanate Hydratase 1), a gene  
343 expressed in the liver (GTEx RNA-seq), is also associated. Urocanate Hydratase or urocanase is an  
344 enzyme that catabolizes urocanic acid to 4-imidazolone-5-propionic acid in liver. Urocanic acid is  
345 found in the skin and the sweat of humans, and has been shown to protect the skin from ultra violet  
346 radiation. It has also been demonstrated to be a major chemo-attractant for a skin-penetrating  
347 parasitic nematode (46). Thus, urocanic acid and urocanase could be hypothesized to play a role in  
348 the attractiveness of humans to mosquitoes.

349 Several strong association signals in the discovery cohort were not replicated in the validation  
350 cohort. Given the limits of the replication cohort mentioned above, we think that the most

351 promising ones are still worth consideration and functional annotation and gene mapping at these  
352 loci were performed with FUMA. *ACER3* identified at 11q13.5 locus appears of particular  
353 relevance. It is associated both with a protection against asymptomatic and symptomatic malaria  
354 infections. *ACER3* is an alkaline ceramidase, involved in the degradation of ceramides into  
355 sphingosine-1-phosphatase (S1P). Ceramides have an anti-plasmodium action (47,48), which is  
356 inhibited by S1P; and it has been hypothesized that antimalarial drug artemisinin and mefloquine  
357 have an antiparasitic action through the activation of sphingomyelinase producing ceramide (49). In  
358 our data allele at risk is associated with a decreased *ACER3* expression, thus indicating a lower  
359 ceramidase activity in the children with higher recurrence of malaria infections. This is not in  
360 accordance with the anti-plasmodium action of ceramides as mentioned above as a lower *ACER3*  
361 expression would be rather associated with a higher availability of ceramides. However *ACER3* has  
362 been also involved in the activation of innate immune in mice and this could also to explain the role  
363 of *ACER3* observed in malaria protection (50).

364 A second gene of interest highlighted by FUMA in malaria infections analyses is *PRKN* (or  
365 *PARK2*) coding for the parkin RBR E3 Ubiquitin Protein Ligase. Mutations in this gene are known  
366 to cause Parkinson disease and this gene has also been involved in innate immune response with a  
367 role in ubiquitin-mediated autophagy of intracellular pathogens, specifically *M.tuberculosis* (51).

368 Finally, a strong association signal in *CSMD1* gene on chromosome 8 was evidenced in mild  
369 malaria attacks analysis. No functional evidence was found by FUMA for this association signal;  
370 however, an association signal in *CSMD1* was recently found in a GWAS of severe malaria in  
371 Tanzania (20); it is a transmembrane protein, assumed to be a tumor suppressor gene (52), and is  
372 also associated with phenotypes as diverse as blood pressure (53) and schizophrenia (54) but there  
373 is as of yet no clear scenario to explain its potential role in malaria.

374 We also examined association results for the most associated candidate variants with malaria-  
375 related phenotypes in the literature (Table 3). There was no evidence of replication for any variant,  
376 except rs40401 located in *IL3* ( $p = 0.01$  and  $p = 0.02$  for malaria attacks and malaria infections,

377 respectively). The fact that our study target population differs substantially from those used in most  
378 published studies on mild malaria may partly explain the limited evidence of replication observed.  
379 *HBB* was not significantly associated with non-severe malaria in our sample. There are  
380 accumulating evidence for an age-related protective effect of HbS (55–57) with an acquired  
381 mechanism of protection from both asymptomatic and symptomatic *P. falciparum* infections in  
382 children. Our results are in line with two of these studies (55,56) which did not detect a protective  
383 effect before the age of two.

384 Our study has identified several genes whose biological function is relevant to malaria  
385 physiopathology and which could play a role in the control of malaria infection. Naturally, future  
386 studies are required to further validate these findings. The difficulty of setting up such longitudinal  
387 field studies have indeed limited the number of individuals that have been included in malaria  
388 cohorts. However our results show that GWAS on non-severe malaria can successfully identify new  
389 candidate genes and inform physiological mechanisms underlying natural protection against  
390 malaria. Improving our understanding of the disease course is crucial for the development of  
391 effective control measures.

## 392 **Materials and Methods**

### 393 **Study population – discovery cohort**

394 The discovery cohort was composed of 525 infants followed-up from birth until 18 months (58).  
395 This study took place from June 2007 to January 2010 in 9 villages of Tori-bossito district located  
396 40 km North-East of Cotonou. Southern Benin is characterised by a subtropical climate, with 2  
397 rainy seasons (a long rainy season from April to July and a short one in August and September).  
398 Clinical incidence of malaria mainly due to *P. falciparum* (97%) was estimated to be 1.5 (95%CI  
399 1.2-1.9) malaria episodes per child (0-5 years) per year in this area (59) and entomological  
400 inoculation rate on average 15.5 infective bites per human per year in studied villages (33).

401 The design and the flow-chart of the study have been published elsewhere (58). Briefly, 656 infants  
402 were included at birth and followed with a close parasitological and clinical survey until the age of  
403 18 months. During the entire follow-up, infants were weekly visited at home by a nurse of the  
404 program and temperature was systematically controlled. In case of axillary temperature higher or  
405 equal to 37°5 (or a history of fever in the preceding 24 hours), child was referred to health center  
406 for medical screening where both a TBS and a rapid diagnostic test (RDT) were performed. Once a  
407 month a systematic TBS was performed to detect asymptomatic parasitemia. At any time in case of  
408 suspicious fever or clinical signs, related or not to malaria, mothers were invited to bring their  
409 infants to the health center where the same protocol (temperature, TBS and RDT) was applied.  
410 Symptomatic malaria infection was treated with artemether-lumefantrine combination therapy as  
411 recommended by the National Malaria Control Program. A total of 10589 TBS were performed  
412 along the follow-up which were read by two independent technicians.

413 To assess the risk of exposure to *Anopheles* bites, environmental (information on house  
414 characteristics and on its immediate surrounding) and geographical data (satellite images, soil type,  
415 watercourse nearby, vegetation index, rainfall, etc) were recorded. Throughout the study, every six  
416 weeks, human landing catches were performed in several points of the villages to evaluate spatial  
417 and temporal variations of *Anopheles* density. Altogether these data allowed modeling, for each  
418 child included in the follow-up, an individual risk of exposure by means of a space- and time-  
419 dependent variable (33).

#### 420 **Study population – replication cohort**

421 The replication cohort was composed of 250 infants who were part of a mother-child study  
422 conducted from 2009 to 2013 in the district of Allada, a southern semi-rural area located 55  
423 kilometres north of Cotonou and 20 km from the first study site (60). Infants were born to mothers  
424 who participated in a multi-country clinical trial for the prevention of malaria in pregnancy  
425 (MiPPAD trial, “Malaria in Pregnancy Preventive Alternative Drugs, NCT00811421,(61)). The first



426 400 newborns from MiPPAD participants (62) were enrolled in the present study. In brief, 400  
427 offsprings were included at birth between January 2010 and June 2012 and among them 306  
428 integrated the second year follow-up. During the first year, detection of malaria cases was passive  
429 Children were seen three times at 6, 9 and 12 months of age for scheduled visit. At these occasions  
430 a TBS was performed for malaria screening. Between 12 and 24 months the same protocol of  
431 follow-up as in the first study was set up. Infants were visited at home every two weeks by a nurse  
432 and temperature was taken. As before in case of axillary temperature greater than or equal to  
433 37.5°C, child was referred to health center where a RDT and a TBS were performed. Each month a  
434 systematic TBS was made to detect asymptomatic malaria. During the whole 24 months follow-up  
435 period, in case of suspicious fever or any health problems, mothers were invited to visit health  
436 centers where the same protocol for malaria screening was applied (temperature, TBS and RDT).  
437 Clinical malaria infections were treated as recommended by the National Malaria Control Program.  
438 Environmental and geographical data were collected as well, to estimate risk of exposure for each  
439 infant. The methodology detailed in Cottrell et al., was used again except that this time *Anopheles*  
440 density was measured in the children's room, using a CDC light trap.

441 In both cohorts, children included in analysis had a minimum of three months follow-up. For the  
442 replication study, we choose to consider only the second year follow-up because i) the differences  
443 of malaria follow-up protocols in the first year lead to the detection of a lower number of malaria  
444 infections, ii) environmental and entomological data needed to estimate the risk of exposure were  
445 missing for children who did not enter the second year follow-up. A description of main  
446 characteristics of infants included or not in the GWAS are presented in supplementary table S1 and  
447 S2.

#### 448 **Ethics Approval**

449 For the two cohort studies, both oral and written communications were provided to parent's children  
450 interested in participating. An informed consent, written in French and in Fon, was presented to, and



451 signed by parents who agreed to participate. The protocols of these studies were approved by both  
452 the Beninese Ethical Committee of the Faculté des Sciences de la Santé (FSS) and the IRD  
453 Consultative Committee on Professional Conduct and Ethics (CCDE).

#### 454 **Genotyping and data quality control**

455 Genotyping was conducted at the Centre National de Recherche en Génomique Humaine (CNRGH,  
456 CEA, Evry, France). Before genotyping, a quality control was systematically performed on each  
457 DNA sample. All samples were quantified by fluorescence, in duplicate, using the Quant-It kits  
458 (Thermo Fischer Scientific). The lowest values systematically underwent a second measurement  
459 before any sample was excluded. The quality of material was estimated using about 10% of the total  
460 samples received (selected randomly throughout the collection) by performing: i) a quality check by  
461 migration on a 1% agarose gel to ensure the samples were not degraded, ii) a standard PCR  
462 amplification reaction on the samples to ensure that the genomic DNA was free of PCR inhibitors,  
463 iii) a PCR test to verify the gender of the individual (63). All samples with concentrations below 20  
464 ng/ $\mu$ L, or a major quality problem (degradation and/or amplification problems) were systematically  
465 excluded from the study.

466 After quality control, DNA samples were aliquoted in 96-well plates (JANUS liquid handling robot,  
467 Perkin Elmer) for genotyping; sample tracking was ensured by a systematic barcode scanning for  
468 each sample. Two DNA positive controls were systematically inserted in a random fashion into the  
469 plates. Genotyping was performed on Illumina HumanOmni5-4v1 chips, on a high throughput  
470 Illumina automated platform, in accordance with the standard automated protocol of Illumina ®  
471 'Infinium HD Assay (Illumina ®, San Diego, USA). The PCR amplification of the genomic DNA  
472 was performed in a dedicated (« pre-PCR ») laboratory. Fragmentation and hybridization of the  
473 DNA on the chips were performed in a dedicated (« post-PCR ») laboratory. Several quality  
474 controls were systematically included during the process, such as visual inspection of the DNA  
475 pellets after precipitation, visual inspection of the deposited cocktail of reagents for hybridization,

476 systematic verification of the temperature of the heating block during the extension and imaging  
477 steps. Reading of the chips was performed on iScan+ scanners (Illumina®, San Diego,  
478 USA). Primary analysis of the genotyping results was done using the GenomeStudio software  
479 (Illumina®, San Diego, USA). The analysis of the internal controls provided by Illumina and the  
480 randomly distributed positive controls allowed the validation of the technological process.

481 Standard control steps and criteria (64) in GWAS were then applied to data. For DNA samples,  
482 individuals with discordant genotypic and reported sex were removed (n=5); heterozygosity versus  
483 sample call rate were examined, leading to removal of one clear outlier; all remaining samples had a  
484 call rate >0.97 (mean = 0.998) and were kept for analysis. A genetic relationship matrix (GRM)  
485 (65) was calculated on common variants (MAF > 0.05) and thinned data ( $r^2 < 0.2$ ) to identify  
486 duplicates and examine relationships between individuals. A relatively high relatedness was  
487 observed in both cohorts which was expected as most of the recruitment was made in rural villages.  
488 One pair of individuals was removed because of an unexpectedly high kinship coefficient ( $\Phi =$   
489 0.27, corresponding to siblings). All others were kept for association analysis (maximum kinship  
490 coefficient  $\Phi = 0.16$ ). Markers with call rate <0.98, MAF <0.01, Hardy-Weinberg equilibrium  
491 (HWE) test  $P < 10^{-8}$  (calculated on unrelated individuals) were filtered out, as well as all non-  
492 autosomal SNPs. Furthermore, a set of 21 pairs of samples (duplicated DNA samples) were used to  
493 identify and remove SNPs with poor reproducibility (4986 SNPs showing at least one discordance).  
494 After these quality control steps, the total sample contained 2,363,703 markers for 775 children.

#### 495 **Population stratification**

496 A principal component analysis (PCA) was performed to evaluate potential population stratification  
497 in our samples. The analysis were performed on 624 pairwise unrelated individuals ( $\Phi < 0.05$ ), after  
498 filtering on allele frequency (MAF >0.05) and LD thinning ( $r^2 < 0.2$ ). The 151 remaining individuals  
499 were then projected onto the principal components.

500 In order to compare our two Beninese populations with other West African populations, a second

501 PCA was performed including West African populations of the KGP. Genetic data were merged on  
502 common positions, filtered and thinned as for the first PCA. Only 100 unrelated individuals from  
503 each of the two study cohorts were selected, to give them the same weight as KGP samples in PCA.  
504 The remaining individuals were projected on the factorial plan thus obtained.

## 505 **Imputation**

506 Imputation of the two cohorts together was performed on the Michigan Imputation Server (66) . The  
507 reference allele was homogenized to the reference genome by converting the genotypes to the  
508 positive strand when necessary; duplicated SNPs, indels, monomorphic SNPs, and SNPs with  
509 inconsistent reference allele with the reference genome were removed. The resulting genotype data  
510 containing 2,539,302 SNPs were uploaded to the Michigan Imputation Server. A last QC step was  
511 performed on the server by comparing allele frequencies observed in the uploaded data and in the  
512 African reference panel from 1000 Genomes v5(67) , and flipping strand for variants showing  
513 significant difference in allele frequency ( $\chi^2$  statistics greater than 300). After this step, there was a  
514 high matched between genotype data and the reference panel (S5 Fig).

515 Genotype data were finally imputed using the minimac3 algorithm provided by the Michigan  
516 Imputation Server. We selected SHAPEIT v2 for prephasing (68,69) and all haplotypes from 1000  
517 Genomes v5 (67) as the reference panel. We filtered out imputed SNPs with a squared correlation  
518 ( $R^2$ ) between input genotypes and expected continuous dosages below 0.8, or with a MAF below  
519 0.01, which yielded an imputed genotype data set of 15,566,900 variants.

## 520 **Adjusting on relevant epidemiological and environmental factors**

521 Association studies were performed with RMM and RMI. A mild malaria attack was defined as a  
522 positive RDT or TBS along with fever (axillary temperature  $\geq 37.5^\circ\text{C}$ ) or a history of fever in the  
523 preceding 24 hours. Each mild malaria attack was recorded with its precise date of event and  
524 children were considered not to be at risk of malaria within the 14 days after receiving an anti-

525 malarial treatment. Malaria infections included both symptomatic and asymptomatic infections. A  
526 positive malaria diagnostic TBS without any fever or history of fever in the preceding 24 hours was  
527 considered as an asymptomatic infection if the child did not experience a clinical malaria in the  
528 following three days.

529 The risk of infection greatly varies over the year depending on the season and malaria transmission  
530 level. This variation can be accounted for using the time dependent risk of exposure which was  
531 defined for each infant in a previous work (Cottrell et al. 2012). To do so, we used a Cox model for  
532 recurrent events, with a rate of events at time  $t$  (Therneau et Grambsch 2000)

$$533 \quad \lambda(t) = \lambda_0(t)e^{X_t\beta + Zb} \quad (1)$$

534 where  $X_t$  is a design matrix for (possibly time dependent) covariates with a fixed effect,  $Z$  is a  
535 matrix of indicator variables designed for including a vector of random individual effects  $b$  with  
536 variance proportional to the GRM  $2\Phi$ , allowing to account for cryptic relatedness and population  
537 structure. The matrix  $X_t$  includes covariates (sex, birthweights, risk of exposure at time  $t$ , etc). This  
538 model has the advantage of both taking into account incomplete follow-up and all the malaria  
539 infections, while incorporating time-dependent susceptibility variables.

540 To fit this model, the follow-up was divided in month intervals following scheduled visit. For each  
541 event we considered the risk of exposure estimated at the previous scheduled visit. A stepwise  
542 backward strategy was used to select relevant covariates. Factors considered were individual factors  
543 (sex, birth weight), health center, factors related to malaria during pregnancy (placental malaria,  
544 infection during pregnancy - replication cohort only-, intake of intermittent preventive treatment –  
545 discovery cohort only-, arm of clinical trials - replication cohort only-), parity (primigravida vs  
546 multigravida), maternal anemia at delivery, use of bednet during the follow-up (categorical variable  
547 based on mother declaration during the follow-up), risk of exposure and transmission season (a  
548 time-dependent variable in four categories, both follow-up spanning over 3 years: dry season, rainy  
549 season 2007, rainy season 2009 and rainy season 2010 for example for the discovery cohort) and

550 socio-economic factors (mother educational level, marital status). The final model included only  
551 covariates significant at  $p \leq 0.05$ ).

## 552 **Genome-wide association analyses**

553 It would have been appealing to use the Cox model (1) for the genome-wide analysis, including the  
554 genotype of each SNP to be tested for association in the covariates matrix  $X_t$ . However the  
555 computational burden inherent to this modeling (with a non-sparse GRM  $2\Phi$ ) makes it  
556 inappropriate for the large number of tests in GWAS.

557 We thus used the Best Linear Unbiased Predictor (BLUP)  $\hat{b}$  of  $b$  obtained from the fitted model  
558 (1), including all covariates selected by the stepwise backward procedure. This value  $\hat{b}$  is an  
559 “individual frailty”, corresponding to individual effects which are not explained by the  
560 epidemiological covariates. To investigate genetic factors involved in these individual effects, they  
561 were analyzed using a linear mixed model  $\hat{b} \sim MVN(G\gamma, 2\tau\Phi + \sigma^2 Id)$  where  $G$  is the vector of  
562 genotypes at the SNP to be tested, coded as 0, 1 or 2 according the number of alternate alleles,  
563 which corresponds to an additive model on the log hazard-ratio scale.

564 This analysis was performed for each genotyped SNP with a MAF  $> 0.01$ , as well as with dosages  
565 of imputed SNPs. The GRM matrix  $2\Phi$  used was computed as described in the quality control  
566 section. The quantile-quantile plot of the  $p$ -values was visually inspected and the genomic inflation  
567 factor  $\lambda$  (ratio of the median of the observed distribution of the test statistic to the expected median)  
568 was calculated to verify the absence of inflation of test statistics due to relatedness and population  
569 structure.

570 SNPs that showed an association at  $p \leq 10^{-5}$  in the discovery cohort were selected for replication  
571 where the same two steps approach was applied. Gene-based annotations given by ANNOVAR (71)  
572 and CADD score (25) were added for these variants in supplementary tables. All highly suggestive  
573 signals (around the significant threshold of  $p \leq 5 \times 10^{-8}$  in the discovery cohort, or which

574 replicated at  $p \leq 0.05$ ) were re-analysed with the mixed Cox model of equation (1) above. For  
575 these signals we additionally performed a meta-analysis combining the results obtained by mixed  
576 Cox models, using classical method implemented in METAL software (72).  
577 All statistical analyses were done using R software (73), with the package gaston (74) for data  
578 quality control and linear mixed models, and the package coxme (23) for the random effects Cox  
579 model. Manhattan plots were obtained with the package qqman (75), and regional linkage  
580 disequilibrium plot for association signals with Locuszoom Standalone v1.4 (76).

### 581 **Gene mapping and biological prioritization**

582 Loci that showed evidence of association at  $p \leq 10^{-5}$  with one of the two phenotypes in the  
583 discovery cohort, were mapped to gene using the FUMA web platform (FUMAGWAS v1.3.3;  
584 <http://fuma.ctglab.nl/>) (27), designed for post-processing of GWAS results and prioritizing of genes.  
585 FUMA annotates candidate SNPs in genomic risk loci and subsequently maps them to prioritized  
586 genes based on (i) physical position mapping on the genome, (ii) expression quantitative trait loci  
587 (eQTL) mapping and (iii) 3D chromatin interactions (chromatin interaction mapping). It  
588 incorporates the most recent bioinformatics databases, such as Combined Annotation Dependent  
589 Depletion score (CADD score) (25,26), regulatory elements in the intergenic region of the genome  
590 (RegulomeDB) (28), Genotype-Tissue Expression (GTEx) and 3D chromatin interactions from HI-  
591 C experiments (29).

592 We considered as genomic risk loci, loci with evidence of replication (table 1) and regions with at  
593 least 3 SNPs below the  $p$ -value threshold of  $10^{-5}$ . In the initial step, SNPs in LD ( $r^2 > 0.01$ , estimated  
594 from AFR reference panel of KGP phase 3) in a 250 Kb windows were considered as belonging to  
595 the same risk locus. Thereafter, a set of candidate SNPs were selected, based on LD ( $r^2 > 0.6$ ) with  
596 the lead SNP (defined as SNP with the lowest  $p$ -value) or with replicating SNPs from our dataset  
597 and from AFR population data of KGP phase 3.

598 Positional mapping was performed using maximum distance from SNPs to gene of 10 kb. Criteria

599 to define deleterious SNPs was a CADD score  $>12.37$  which indicates potential pathogenicity or a  
600 RegulomeDB score  $\leq 2$  which indicates that the SNP likely lies in a functional location (categories  
601 1 and 2 of Regulome classification identified as “likely to affect binding”). eQTL mapping was  
602 performed using eQTL data from tissue types relevant for non-severe malaria. Search for eQTL  
603 associations were carried out in Blood eQTL (77), BIOS QTL (30), eQTLGen, three databases of e-  
604 QTL associations identified in blood, and in the following tissue types of GTEx v7 database: whole  
605 blood, cells-transformed fibroblast, cells-EBV-transformed lymphocytes, liver, skin (sun exposed or  
606 not) and spleen. Chromatin interaction mapping was performed from Hi-C experiments data of five  
607 tissues/cell types identified as relevant for non-severe malaria: liver, spleen GM12878, IMR90 and  
608 trophoblast-like-cell. Hi-C experiments identified chromatin interaction between small  
609 chromosomal regions two by two. In this analysis a candidate SNP is mapped to a gene if a  
610 significant interaction is observed between a first region containing the candidate SNP and the  
611 second one where the gene is located. FDR was used to correct for multiple testing (FDR  $<0.05$  for  
612 eQTL association and FDR  $<1 \times 10^{-6}$  for chromatin interaction, the default values in FUMA).

### 613 **Candidate associations**

614 We report associations observed in the discovery cohort for genes ( $\pm 200$  kb 3' and 5') showing the  
615 greatest evidence of association with malaria in the literature. Genes previously associated with  
616 mild malaria phenotype were selected from the review by Marquet (7): *HbS* and *HbC* alleles of  
617 *HBB*, *IL3* from the 5q31-q33 region, *TNF*, *NCR3* from 6p21-p23 region, and *IL10*. Association was  
618 not assessed for the *HP* gene and alpha-thalasemia condition, because the structural variants  
619 involved are not directly accessible through high density genotyping array (78). We also checked  
620 significance for genes previously associated with severe malaria that were confirmed by recent  
621 GWA and multi-center studies: *ABO*, *ATP2B4*, and *FREM3/GYP* locus (17–19,22).

### 622 **Acknowledgements**

623 This research is a collaboration between the CEA/ Jacob/CNRGH and the IRD/UMR216. We wish



624 to thanks the collaborators of the CERPAGE who participated actively in the longitudinal follow-  
625 ups. Longitudinal follow-ups were funded by the ANR grants (ANR-SEST 2006 040-01 and ANR-  
626 PRSP 2010 012-001); the Ministère de la Recherche et des Technologies, France (REFS Nu2006-  
627 22) and the IRD. We made use of data previously generated in the MiPPAD study (EDCTP-  
628 IP.07.31080.002). The genome-wide scan was supported by the CNRGH. We are grateful to the  
629 Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing  
630 computing and storage resources.

## 631 **References**

1. World Health Organization. WHO | World malaria report 2018. 2018.
2. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*. 2015 Oct 8;526(7572):207–11.
3. Verra F, Mangano VD, Modiano D. Genetics of susceptibility to *Plasmodium falciparum*: from classical malaria resistance genes towards genome-wide association studies. *Parasite Immunol*. 2009 May;31(5):234–53.
4. Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet*. 2005 Aug;77(2):171–92.
5. Mackinnon MJ, Mwangi TW, Snow RW, Marsh K, Williams TN. Heritability of malaria in Africa. *PLoS Med*. 2005 Dec;2(12):e340.
6. Driss A, Hibbert JM, Wilson NO, Iqbal SA, Adamkiewicz TV, Stiles JK. Genetic polymorphisms linked to susceptibility to malaria. *Malar J*. 2011 Sep 19;10:271.
7. Marquet S. Overview of human genetic susceptibility to malaria: From parasitemia control to severe disease. *Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis*. 2017 Jun 1;
8. Garcia A, Marquet S, Bucheton B, Hillaire D, Cot M, Fievet N, et al. Linkage analysis of blood *Plasmodium falciparum* levels: interest of the 5q31-q33 chromosome region. *Am J Trop Med Hyg*. 1998 Jun;58(6):705–9.
9. Rihet P, Traoré Y, Abel L, Aucan C, Traoré-Leroux T, Fumoux F. Malaria in humans: *Plasmodium falciparum* blood infection levels are linked to chromosome 5q31-q33. *Am J Hum Genet*. 1998 Aug;63(2):498–505.
10. Flori L, Kumulungui B, Aucan C, Esnault C, Traoré AS, Fumoux F, et al. Linkage and association between *Plasmodium falciparum* blood infection levels and chromosome 5q31-q33. *Genes Immun*. 2003 Jun;4(4):265–8.
11. Sakuntabhai A, Ndiaye R, Casadémont I, Peerapittayamongkol C, Peerapittayamonkol C, Rogier C, et al. Genetic determination and linkage mapping of *Plasmodium falciparum* malaria



related traits in Senegal. *PLoS One*. 2008;3(4):e2000.

12. Jepson A, Sisay-Joof F, Banya W, Hassan-King M, Frodsham A, Bennett S, et al. Genetic linkage of mild malaria to the major histocompatibility complex in Gambian children: study of affected sibling pairs. *BMJ*. 1997 Jul 12;315(7100):96–7.
13. Brisebarre A, Kumulungui B, Sawadogo S, Atkinson A, Garnier S, Fumoux F, et al. A genome scan for *Plasmodium falciparum* malaria identifies quantitative trait loci on chromosomes 5q31, 6p21.3, 17p12, and 19p13. *Malar J*. 2014 May 28;13:198.
14. Timmann C, Evans JA, König IR, Kleensang A, Rüschemdorf F, Lenzen J, et al. Genome-wide linkage analysis of malaria infection intensity and mild disease. *PLoS Genet*. 2007 Mar 23;3(3):e48.
15. Milet J, Nuel G, Watier L, Courtin D, Slaoui Y, Senghor P, et al. Genome wide linkage study, using a 250K SNP map, of *Plasmodium falciparum* infection and mild malaria attack in a Senegalese population. *PLoS One*. 2010;5(7):e11616.
16. Baaklini S, Afridi S, Nguyen TN, Koukouikila-Koussounda F, Ndounga M, Imbert J, et al. Beyond genome-wide scan: Association of a cis-regulatory NCR3 variant with mild malaria in a population living in the Republic of Congo. *PLoS One*. 2017;12(11):e0187818.
17. Timmann C, Thye T, Vens M, Evans J, May J, Ehmen C, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*. 2012 Sep 20;489(7416):443–6.
18. Band G, Le QS, Jostins L, Pirinen M, Kivinen K, Jallow M, et al. Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genet*. 2013 May;9(5):e1003509.
19. Malaria Genomic Epidemiology Network, Band G, Rockett KA, Spencer CCA, Kwiatkowski DP. A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature*. 2015 Oct 8;526(7572):253–7.
20. Ravenhall M, Campino S, Sepúlveda N, Manjurano A, Nadjm B, Mtove G, et al. Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. *PLoS Genet*. 2018 Jan;14(1):e1007172.
21. Malaria Genomic Epidemiology Network, Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. *Nat Genet*. 2014 Nov;46(11):1197–204.
22. Leffler EM, Band G, Busby GBJ, Kivinen K, Le QS, Clarke GM, et al. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*. 2017 16;356(6343).
23. Terry M. Therneau. *coxme: Mixed Effects Cox Models*. [Internet]. Available from: <https://CRAN.R-project.org/package=coxme>
24. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010 Apr;7(4):248–9.
25. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014 Mar;46(3):310–5.

26. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D886–94.
27. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017 28;8(1):1826.
28. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 2012 Sep;22(9):1790–7.
29. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* 2016 15;17(8):2042–59.
30. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet.* 2017;49(1):139–45.
31. Greenwood BM. The microepidemiology of malaria and its importance to malaria control. *Trans R Soc Trop Med Hyg.* 1989;83 Suppl:25–9.
32. Bousema T, Drakeley C, Gesase S, Hashim R, Magesa S, Mosha F, et al. Identification of hot spots of malaria transmission for targeted malaria control. *J Infect Dis.* 2010 Jun 1;201(11):1764–74.
33. Cottrell G, Kouwaye B, Pierrat C, le Port A, Bouraïma A, Fonton N, et al. Modeling the influence of local environmental factors on malaria transmission in Benin and its implications for cohort study. *PloS One.* 2012;7(1):e28812.
34. Le Port A, Cottrell G, Chandre F, Cot M, Massougbojji A, Garcia A. Importance of adequate local spatiotemporal transmission measures in malaria cohort studies: application to the relation between placental malaria and first malaria infection in infants. *Am J Epidemiol.* 2013 Jul 1;178(1):136–43.
35. Bouaziz O, Courtin D, Cottrell G, Milet J, Nuel G, Garcia A. Is Placental Malaria a Long-term Risk Factor for Mild Malaria Attack in Infancy? Revisiting a Paradigm. *Clin Infect Dis Off Publ Infect Dis Soc Am.* 2018 Mar 5;66(6):930–5.
36. Nikolaienko RM, Agyekum B, Bouyain S. Receptor protein tyrosine phosphatases and cancer. *Cell Adhes Migr.* 2012 Jul 1;6(4):356–64.
37. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 2017 Jan 4;45(D1):D362–8.
38. Zhang X, Guo A, Yu J, Possemato A, Chen Y, Zheng W, et al. Identification of STAT3 as a substrate of receptor protein tyrosine phosphatase T. *Proc Natl Acad Sci U S A.* 2007 Mar 6;104(10):4060–4.
39. Peyser ND, Freilino M, Wang L, Zeng Y, Li H, Johnson DE, et al. Frequent promoter hypermethylation of PTPRT increases STAT3 activation and sensitivity to STAT3 inhibition in head and neck cancer. *Oncogene.* 2016 Mar 3;35(9):1163–9.

40. Liu M, Amodu AS, Pitts S, Patrickson J, Hibbert JM, Battle M, et al. Heme mediated STAT3 activation in severe malaria. *PLoS One*. 2012;7(3):e34280.
41. Liu M, Solomon W, Cespedes JC, Wilson NO, Ford B, Stiles JK. Neuregulin-1 attenuates experimental cerebral malaria (ECM) pathogenesis by regulating ErbB4/AKT/STAT3 signaling. *J Neuroinflammation*. 2018 Apr 10;15(1):104.
42. Tan I, Leung T. Myosin light chain kinases: division of work in cell migration. *Cell Adhes Migr*. 2009 Sep;3(3):256–8.
43. Gao H, Wu X, Sun Y, Zhou S, Silberstein LE, Zhu Z. Suppression of homeobox transcription factor VentX promotes expansion of human hematopoietic stem/multipotent progenitor cells. *J Biol Chem*. 2012 Aug 24;287(35):29979–87.
44. Wu X, Gao H, Ke W, Giese RW, Zhu Z. The homeobox transcription factor VentX controls human macrophage terminal differentiation and proinflammatory activation. *J Clin Invest*. 2011 Jul;121(7):2599–613.
45. Wu X, Gao H, Bleday R, Zhu Z. Homeobox transcription factor VentX regulates differentiation and maturation of human dendritic cells. *J Biol Chem*. 2014 May 23;289(21):14633–43.
46. Safer D, Brenes M, Dunipace S, Schad G. Urocanic acid is a major chemoattractant for the skin-penetrating parasitic nematode *Strongyloides stercoralis*. *Proc Natl Acad Sci U S A*. 2007 Jan 30;104(5):1627–30.
47. Labaied M, Dagan A, Dellinger M, Gèze M, Egée S, Thomas SL, et al. Anti-Plasmodium activity of ceramide analogs. *Malar J*. 2004 Dec 10;3:49.
48. Heung LJ, Luberto C, Del Poeta M. Role of Sphingolipids in Microbial Pathogenesis. *Infect Immun*. 2006 Jan;74(1):28–39.
49. Pankova-Kholmyansky I, Dagan A, Gold D, Zaslavsky Z, Skutelsky E, Gatt S, et al. Ceramide mediates growth inhibition of the *Plasmodium falciparum* parasite. *Cell Mol Life Sci CMLS*. 2003 Mar;60(3):577–87.
50. Wang K, Xu R, Snider AJ, Schrandt J, Li Y, Bialkowska AB, et al. Alkaline ceramidase 3 deficiency aggravates colitis and colitis-associated tumorigenesis in mice by hyperactivating the innate immune system. *Cell Death Dis*. 2016 Mar 3;7:e2124.
51. Manzanillo PS, Ayres JS, Watson RO, Collins AC, Souza G, Rae CS, et al. PARKIN ubiquitin ligase mediates resistance to intracellular pathogens. *Nature*. 2013 Sep 26;501(7468):512–6.
52. Sun PC, Uppaluri R, Schmidt AP, Pashia ME, Quant EC, Sunwoo JB, et al. Transcript map of the 8p23 putative tumor suppressor region. *Genomics*. 2001 Jul;75(1–3):17–25.
53. Hong K-W, Go MJ, Jin H-S, Lim J-E, Lee J-Y, Han BG, et al. Genetic variations in ATP2B1, CSK, ARSG and CSMD1 loci are related to blood pressure and/or hypertension in two Korean cohorts. *J Hum Hypertens*. 2010 Jun;24(6):367–72.
54. Håvik B, Le Hellard S, Rietschel M, Lybæk H, Djurovic S, Mattheisen M, et al. The complement control-related genes CSMD1 and CSMD2 associate to schizophrenia. *Biol Psychiatry*. 2011 Jul 1;70(1):35–42.

55. Williams TN, Mwangi TW, Roberts DJ, Alexander ND, Weatherall DJ, Wambua S, et al. An immune basis for malaria protection by the sickle cell trait. *PLoS Med.* 2005 May;2(5):e128.
56. Gong L, Maiteki-Sebuguzi C, Rosenthal PJ, Hubbard AE, Drakeley CJ, Dorsey G, et al. Evidence for both innate and acquired mechanisms of protection from *Plasmodium falciparum* in children with sickle cell trait. *Blood.* 2012 Apr 19;119(16):3808–14.
57. Lopera-Mesa TM, Doumbia S, Konaté D, Anderson JM, Doumbouya M, Keita AS, et al. Effect of red blood cell variants on childhood malaria in Mali: a prospective cohort study. *Lancet Haematol.* 2015 Apr;2(4):e140-149.
58. Le Port A, Cottrell G, Martin-Prevel Y, Migot-Nabias F, Cot M, Garcia A. First malaria infections in a cohort of infants in Benin: biological, environmental and genetic determinants. Description of the study site, population methods and preliminary results. *BMJ Open.* 2012;2(2):e000342.
59. Damien GB, Djènontin A, Rogier C, Corbel V, Bangana SB, Chandre F, et al. Malaria infection and disease in an area with pyrethroid-resistant vectors in southern Benin. *Malar J.* 2010 Dec 31;9:380.
60. d'Almeida TC, Sadissou I, Milet J, Cottrell G, Mondière A, Avokpaho E, et al. Soluble human leukocyte antigen -G during pregnancy and infancy in Benin: Mother/child resemblance and association with the risk of malaria infection and low birth weight. *PloS One.* 2017;12(2):e0171117.
61. González R, Mombo-Ngoma G, Ouédraogo S, Kakolwa MA, Abdulla S, Accrombessi M, et al. Intermittent preventive treatment of malaria in pregnancy with mefloquine in HIV-negative women: a multicentre randomized controlled trial. *PLoS Med.* 2014 Sep;11(9):e1001733.
62. Accrombessi M, Ouédraogo S, Agbota GC, Gonzalez R, Massougboji A, Menéndez C, et al. Malaria in Pregnancy Is a Predictor of Infant Haemoglobin Concentrations during the First Year of Life in Benin, West Africa. *PloS One.* 2015;10(6):e0129510.
63. Wilson JF, Erlandsson R. Sexing of human and other primate DNA. *Biol Chem.* 1998 Oct;379(10):1287–8.
64. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc.* 2010 Sep;5(9):1564–73.
65. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011 Jan 7;88(1):76–82.
66. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016;48(10):1284–7.
67. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature.* 2015 Oct 1;526(7571):75–81.
68. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 2011 Dec 4;9(2):179–81.
69. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods.* 2013 Jan;10(1):5–6.

70. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model [Internet]. New York: Springer-Verlag; 2000 [cited 2018 Apr 26]. (Statistics for Biology and Health). Available from: [//www.springer.com/us/book/9780387987842](http://www.springer.com/us/book/9780387987842)
71. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010 Sep;38(16):e164.
72. Willer CJ, Li Y. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinforma Oxf Engl.* 2010 Sep 1;26(17):2190–1.
73. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing [Internet]. R Foundation for Statistical Computing; 2017. Available from: <https://www.R-project.org>
74. Hervé Perdry & Claire Dandine-Roulland. *gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models.* 2018.
75. Stephen Turner. *qqman: Q-Q and Manhattan Plots for GWAS Data.* [Internet]. 2014. Available from: <https://CRAN.R-project.org/package=qqman>
76. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* 2010 Sep 15;26(18):2336–7.
77. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet.* 2013 Oct;45(10):1238–43.
78. Higgs DR. The Molecular Basis of  $\alpha$ -Thalassemia. *Cold Spring Harb Perspect Med.* 2013 Jan;3(1).