Étude de la composante génétique de la variabilité des infections palustres simples : Approche génome entier dans deux cohortes de jeunes enfants au Bénin

Jacqueline Milet

université
PARIS-SACLAY

Étude de la composante génétique de la variabilité des infections palustres simples Approche génome entier dans deux cohortes de jeunes enfants au Bénin

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 570, Santé Publique EDSP
Spécialité de doctorat: Santé Publique – Génétique statistique
Unité de recherche : UVSQ, Inserm, CESP, 94807, Villejuif, France
Référent : Faculté de médecine

Thèse présentée et soutenue à Paris, le 20 novembre 2020, par

Jacqueline MILET

Composition du Jury

André GARCIA DR, IRD Université de Paris	Président
Aurélie COBAT CR, HDR, Université de Paris	Rapporteur & Examinatrice
Pascal RIHET PR, Université Aix-Marseille	Rapporteur & Examinateur
Céline BELLENGUEZ CR, INSERM Université Lille	Examinatrice
Alexis ELBAZ DR, INSERM Université Paris-Saclay	Examinateur
Alicia SANCHEZ-MAZAS PR, Université de Genève	Examinatrice
Hervé PERDRY MCU, Université Paris-Saclay	Directeur de thèse
Audrey SABBAGH MCU, Université de Paris	Co-Encadrante

Thèse de doctorat

NNT : 2020UPASR013

ARTICLES ET COMMUNICATIONS

Articles

Milet J, Boland A, Luisi P, Sabbagh A, Sadissou I, Sonon P, Domingo N, Palstra F, Gineau L, Courtin D, Massougbodji A, Garcia A, Deleuze JF, Perdry H. **First genome-wide association study of non-severe malaria in two birth cohorts in Benin**. Hum Genet. 2019 Dec;138(11-12):1341-1357. doi: 10.1007/s00439-019-02079-5.

Milet J, Courtin D, Garcia A, Predry H. **Mixed logistic regression in Genome-Wide Association Studies**. BMC Bioinformatics. 2020 Nov 23;21(1):536. doi: 10.1186/s12859-020-03862-2.

Posters

Milet J, Luisi P, Sabbagh A, Sadissou I, Sonon P, Boland A, Courtin D, Deleuze J-F, Garcia A, Perdry H. **Genome-wide association study of susceptibility to mild malaria in two cohorts of young Beninese children**. International Genetic Epidemiology Society meeting, Cambridge, 9-11 septembre 2017

Milet J, Perdry H. **Estimation des odds-ratios dans les études d'association génome entier basées sur le modèle logistique mixte.** Assises de Génétique Humaine et Médicale, Tours 21-24 janvier 2020

REMERCIEMENTS

En arriver à écrire cette page de remerciement est un moment longtemps attendu… et un réel soulagement. Aussi je voudrais remercier sincèrement tous ceux qui m'ont aidée, soutenue au cours de cette thèse par leurs précieux conseils, leur écoute ou quelques mots réconfortants…

Tous d'abord merci aux membres du Jury qui ont accepté de relire de manière critique ce travail. J'espère que vous arriverez sans encombre au bout de ce manuscrit.

Un grand merci à André Garcia pour m'avoir amenée à la génétique. C'est passionnant ! Et cela a été l'occasion de très belles rencontres, à l'U535 tout d'abord (où j'ai rencontré Hervé Perdry et Audrey Sabbagh), puis à l'UMR MERIT. Merci aussi de m'avoir fait confiance pour analyser ces données, et de m'avoir laissé du temps pour réaliser cette thèse.

Merci infiniment à Audrey Sabbagh et Hervé Perdry de m'avoir accompagné tout au long de cette thèse, chacun à votre manière. Cela n'a pas toujours été facile mais j'ai beaucoup appris. Merci Audrey pour ta gentillesse, je pense qu'elle a fini par déteindre un peu sur moi. Merci pour tes réponses toujours précises à mes demandes, pour tes précieux conseils. Un merci spécial à Hervé pour tout le temps passé à mes côtés, pour ces moments à manipuler les données, à réfléchir, à discuter. Merci pour ta patience aussi, notamment pour m'expliquer (et me réexpliquer) ces concepts et ces formules mathématiques qui me paraissaient bien abscons.

Un grand merci à mon informaticien préféré. Tu m'as sauvée plus d'une fois. J'ai appris grâce à toi que les *dockers* n'étaient pas seulement dans les ports et que les *containers* ne servaient pas qu'aux déménagements. C'est précieux de t'avoir à mes côtés (les aspects informatiques sont anecdotiques), et merci pour ta patience aussi… je n'ai pas tout à fait mesuré l'investissement en m'engageant dans cette thèse.

Merci à mes deux louloutes qui ont bien grandi. Merci de m'avoir de laissé travailler l'esprit tranquille et merci pour vos encouragements (je garde précieusement votre tasse). Vous êtes adorables. Eh oui, les simus elles finissent par tourner (souvent après plusieurs échecs il faut le dire) et là c'est du bonheur !

Merci à ma mère, à mon père, à mes sœurs et à mon frère. Merci pour votre soutien. Sacrée famille ! Je vous aime.

Merci à David, qui était à la supervision du second terrain à Allada avec Gilles, et qui est également un compagnon de route depuis mes débuts à l'IRD. Cela fait quelques années déjà… Merci pour ces belles aventures en Afrique, pour les aventures scientifiques aussi et pour ta présence au cours de cette thèse.

Merci à Etienne Patin pour ses précieux conseils avant de lancer les analyses de sélection naturelle ; Merci à Anaïs, David et Célia pour leur éclairage sur la fonctionnalité des gènes ; Pardon à Hervé et C. Chang pour les frayeurs.

Merci à mes collègues des soupentes durant cette thèse: Brigitte, Laure, J-C, JGG, Jean-Yves, Romain, Cornelia pour ces moments conviviaux. Et merci en particulier à Laure et à Romain, pour les goûters, pour votre bonne humeur et vos encouragements.

TABLE DES MATIERES

1 INTRODUCTION

1.1 Le paludisme

Le paludisme est une maladie infectieuse des régions tropicales et subtropicales induite par des parasites du genre *Plasmodium.* Il s'agit d'une maladie vectorielle*,* transmise par l'intermédiaire d'un moustique du genre *Anopheles*, lors d'un repas sanguin de la femelle moustique nécessaire à sa ponte. La transmission du paludisme se fait essentiellement la nuit, du crépuscule au lever du soleil, période durant laquelle les anophèles sont particulièrement actifs.

Le paludisme peut présenter une grande diversité de manifestations cliniques, allant de l'absence de symptôme (infection dite *asymptomatique*) à des formes graves pouvant conduire au décès, telles que le neuropaludisme ou l'anémie sévère. On distingue généralement les accès palustres simples, qui constituent la majorité des cas diagnostiqués, et les accès palustres graves qui nécessitent une prise en charge rapide à l'hôpital. Les accès simples sont caractérisés par des épisodes de fièvre aigus accompagnés de symptômes non spécifiques plus ou moins marqués (frissons, maux de tête, douleurs abdominales, douleurs musculaires, vomissements, etc.). S'ils sont diagnostiqués à temps et traités correctement, ils guérissent sans séquelle. Les accès graves sont liés à des complications touchant des organes vitaux ou entrainant des anomalies de la composition du sang.

1.1.1 Répartition géographique

Le paludisme reste aujourd'hui un problème de santé publique majeur avec environ 228 millions de cas dans le monde et 405 000 décès recensés en 2018, d'après le dernier rapport de l'Organisation Mondiale de la Santé (*WHO | World malaria report*, 2019). La maladie est encore endémique, c'est-à-dire présente de manière permanente, dans 87 pays. Les régions défavorisées des zones tropicales et subtropicales sont les plus touchées, en particulier l'Afrique subsaharienne, et certains pays de l'Océanie comme la Papouasie-Nouvelle-Guinée (Figure 1.1).

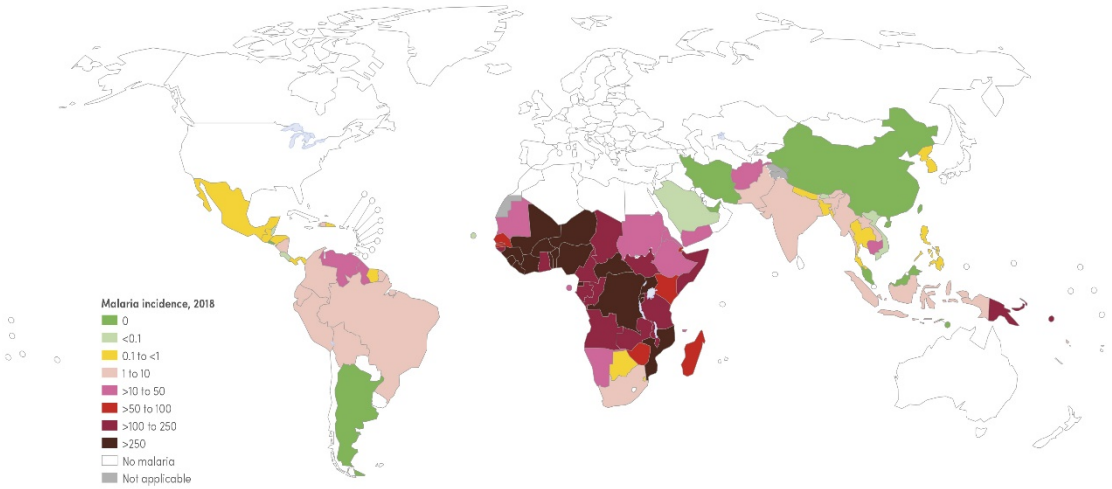


Figure 1.1 Incidence des cas de paludisme dans le monde, en 2018.
L'incidence correspond au nombre de nouveaux cas de paludisme sur une année pour 1000 personnes. Source : Rapport OMS sur le paludisme 2019

Cinq espèces de parasites du genre *Plasmodium* peuvent infecter l'Homme. Quatre espèces infectent presque exclusivement l'Homme comme hôte intermédiaire (l'hôte primaire étant le moustique) : *Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale* et *Plasmodium malariae.* La cinquième espèce, *Plasmodium knowlesi,* est connue à l'origine pour infecter le singe. Elle est considérée aujourd'hui également comme un parasite humain (White, 2008), plusieurs milliers de cas chez l'Homme ayant été rapportés en Asie du sud-est depuis une quinzaine d'années.

Ces cinq espèces diffèrent par leur répartition géographique et par la pathologie engendrée. *P. falciparum* et *P. vivax* sont les espèces les plus répandues et co-existent dans de nombreuses régions du monde. La Figure 1.2 représente le nombre de cas de paludisme attribuables à ces deux espèces dans les cinq régions définies par l'OMS. *P. falciparum* est l'espèce dominante en Afrique (responsable de plus de 99 % des cas) et la plus pathogène, responsable de la majorité des cas de paludisme graves et des décès. *P. vivax* est répandue en Asie du Sud-Est (37,2%) et est majoritaire en Amérique Latine (74%). Les infections par *P. vivax* sont souvent considérées comme bénignes bien que des complications graves aient été également observées.

Comparées aux deux premières espèces, les prévalences de *P. malariae et de P. ovale* sont faibles, et peu de données précises sont disponibles. *P. malariae* est présente de manière sporadique dans

l'ensemble des régions alors que *P. ovale* est localisée essentiellement en Afrique de l'Ouest. Les infections engendrées par ces espèces se caractérisent par des niveaux de densité parasitaire faibles qui peuvent passer inaperçues sur des longues périodes. L'espèce *P. knowlesi* est rencontrée principalement dans les pays d'Asie du Sud-Est et en particulier à l'est de la Malaisie. Comme pour *P. malariae*, les infections sont généralement bénignes mais peuvent entrainer dans une minorité de cas des complications et conduire au décès.

Figure 1.2 Répartition des cas de paludisme liés à P.falciparum et à P.vivax dans le monde
Estimations en millions de cas. AFR: région Afrique; AMR: région Amérique; EMR: région Méditerranée orientale; SEAR: région Asie du sud-est; WPR: région Pacifique occidental. ***Source :*** *Rapport OMS sur le paludisme 2018*

En résumé, l'Afrique est le continent qui paie le plus lourd tribut au paludisme. C'est le continent le plus touché en terme de nombre de cas, et où sévit le parasite le plus virulent *P. falciparum*. En 2018, l'Afrique subsaharienne enregistrait environ 90% des cas de paludisme et 95 % des décès.

1.1.2 Evolution depuis les années 2000

Depuis les années 2000, la mobilisation internationale et les investissements financiers importants de la part des acteurs publics ou privés ont permis de faire reculer la maladie. Entre 2000 et 2015 on estime que les nombres de cas et de décès ont été divisés par deux dans le monde. Ces succès sont liés à la large distribution de moustiquaires imprégnées d'insecticide à longue durée d'action et à la mise en place des traitements antipaludéens très efficaces à base d'artémisinine (les CTA pour Combinaison Thérapeutique à base d'Artémisinine). En Afrique, sur cette période, Bhatt *et al.* (Bhatt et al., 2015) rapportent une diminution de moitié de la prévalence de l'infection à *P. falciparum* chez

les enfants de 2 à 10 ans et de 40% de l'incidence des cas cliniques (désignant les infections associées à des symptômes cliniques, incluant à la fois les accès simples et graves).

Cependant, depuis 2010, le rythme de diminution du nombre de cas de paludisme dans le monde s'est considérablement ralenti, avec même une légère ré-augmentation du nombre de cas depuis 2016 (*WHO | World malaria report*, 2019). Les situations sont contrastées suivant les régions. En ce qui concerne l'Afrique, l'incidence des cas de paludisme est stable depuis les années 2014-2015.

1.1.3 Les défis actuels de la lutte contre le paludisme

Les deux principaux outils de lutte mis en place dans les années 2000, les moustiquaires imprégnées et les CTA, sont menacés par des résistances aux insecticides et aux médicaments (malERA, 2017). La résistance aux pyréthrénoïdes, la seule classe d'insecticides utilisée actuellement pour imprégner les moustiquaires, est maintenant largement répandue (Knox et al., 2014). Par ailleurs, des résistances à l'artémisinine et aux autres composants des ACT sont apparues dans la Région du Grand Mékong, en Asie du Sud-Est, à la fin des années 2000 (Ashley et al., 2014). Historiquement, c'est aussi dans cette région qu'ont émergé les résistances aux autres antipaludiques, notamment à la chloroquine à la fin des années 1950, puis à la sufadoxine-pyriméthamine, avant de se diffuser à travers l'Asie et de gagner l'Afrique.

La recherche depuis 50 ans sur le développement d'un vaccin contre le paludisme s'est heurté à la complexité du parasite et à ses stratégies d'évasion du système immunitaire (Gomes et al., 2016; Lyke, 2017). Le vaccin RTS,S est le premier et le seul vaccin pour le moment à avoir montré une efficacité protectrice contre le paludisme dans un essai de phase III (RTS,S Clinical Trials Partnership, 2015). Cependant l'efficacité de ce vaccin dirigé contre *P. falciparum* reste limitée. L'essai clinique a montré que la protection pour les cas cliniques, après une durée de suivi moyenne de 48 mois était seulement de 26 % chez les enfants ayant reçu trois doses de vaccin et de 39% chez les enfants ayant reçu une 4ème dose, 15-18 mois après la première injection. Ce vaccin fait actuellement l'objet d'une étude pilote à grande échelle dans trois pays africains (Ghana, Malawi et Kenya) pour évaluer son utilisation en tant qu'outil complémentaire aux mesures déjà recommandées par l'OMS.

La lutte contre le paludisme, après des progrès importants dans les années 2000-2015, est aujourd'hui menacée à la fois par des résistances aux insecticides et de par la propagation possible de la résistance à l'artémisinine. Aussi, en l'absence de vaccin efficace et de traitement alternatif présentant la même efficacité et la même tolérance que les ACT, la recherche sur le paludisme reste indispensable afin d'améliorer la compréhension de la maladie et de développer des nouveaux outils ou stratégies de lutte.

1.2 Infection à Plasmodium falciparum

1.2.1 Cycle du parasite

Le cycle de vie complet de *P. falciparum* implique deux hôtes : l'anophèle et l'Homme. La description qui suit concerne uniquement le cycle chez l'Homme, qui nous intéresse ici pour la compréhension de la physiopathologie de la maladie et des différentes formes de paludisme. Durant un repas sanguin, la femelle anophèle infectée inocule à l'Homme des parasites sous forme de sporozoïtes, qui vont gagner le foie *via* la circulation sanguine (Figure 1.3). Une première phase de réplication des parasites a lieu à l'intérieur des cellules du foie : la phase hépatique. A la fin de cette phase qui dure environ 10 jours, les hépatocytes infectés éclatent libérant chacun dans la circulation sanguine, des milliers de parasites appelés mérozoïtes (20 000 à 30 000 par hépatocyte).

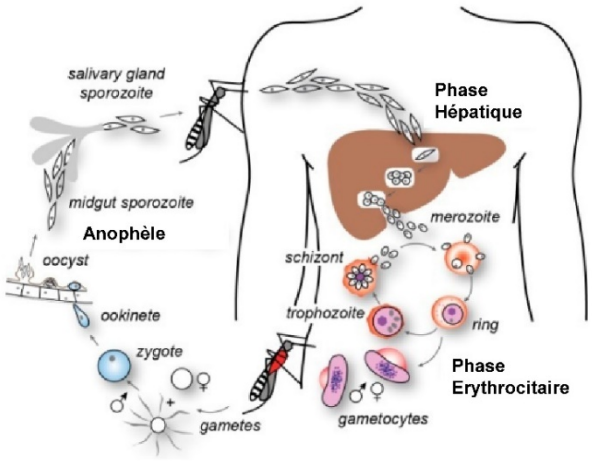


Figure 1.3 Le cyle de vie de Plasmodium falciparum
D'après Cowman et al., 2012

Commence alors la phase érythrocytaire : les mérozoïtes infectent les globules rouges où ils se multiplient par cycles successifs. Chaque cycle de réplication dure 48h pour *P. falciparum,* produit de 20 à 30 parasites et conduit à la destruction du globule rouge. Les parasites libérés dans la circulation sanguine envahissent alors d'autres globules rouges et continuent leur réplication conduisant à une augmentation rapide de la charge parasitaire d'un individu, celle-ci pouvant atteindre plusieurs milliards de parasites. Durant la phase érythrocytaire, certains parasites vont évoluer en gamétocytes, stade sexué du parasite, qui restent dans la circulation sanguine et pourront être ingérés par une femelle anophèle lors d'un repas sanguin, perpétuant ainsi la transmission.

1.2.2 Evolution naturelle de la maladie

La phase hépatique, qui suit l'inoculation des parasites, est asymptomatique. Les premiers symptômes cliniques apparaissent lors de la phase érythrocytaire. Lorsque les hématies infectées sont lysées, elles libèrent dans le sang les parasites ainsi que des substances toxiques (dont l'hémozoïne) qui sont en partie responsables de la fièvre et des symptômes. C'est aussi durant cette phase que la maladie peut être diagnostiquée, lorsque les parasites circulent dans le sang périphérique. Le diagnostic est généralement effectué au moyen d'une goutte épaisse (une goutte de sang étalée sur une lame de verre, examinée par microscopie) ou par un test de diagnostic rapide (TDR), qui permet de détecter la présence d'antigènes spécifiques du parasite.

Les symptômes surviennent une dizaine de jours après l'infection pour les individus non immuns, cependant ils ne sont pas systématiques. Dans les régions endémiques, beaucoup d'individus, majoritairement des adultes mais également des enfants, peuvent être porteurs de parasites tout en restant asymptomatiques.

Les mécanismes conduisant ensuite aux formes graves ne sont pas totalement élucidés. Les accès graves, tels que définis par l'OMS (« WHO | Severe Malaria », 2014), représentent un ensemble hétérogène de syndromes cliniques incluant le neuropaludisme, l'anémie sévère, la détresse respiratoire ou l'acidose pour les plus fréquents. La sévérité de l'infection à *P. falciparum* est liée à

plusieurs phénomènes qui peuvent se combiner : une multiplication excessive des parasites (hyperparasitémie), la destruction des globules rouges infectés (anémie grave), la cytoadhérence des hématies infectées et une dérégulation du système immunitaire entrainant une réponse pro-inflammatoire excessive. Le phénomène de cytoadhérence a été décrit plus particulièrement dans le cadre du neuropaludisme (Wassmer & Grau, 2017). Lors de l'infection, les parasites modifient la membrane des globules rouges en exprimant à leur surface plusieurs protéines. Les hématies infectées acquièrent ainsi la capacité à adhérer à l'endothélium des vaisseaux sanguins et aux autres globules (phénomène de *rosetting*) conduisant à la séquestration des parasites dans différents organes et tissus et en particulier dans le cerveau.

1.2.3 Développement de l'immunité protectrice

Dans les régions de transmission continue, une immunité protectrice partielle contre *P. falciparum* se développe durant l'enfance en réponse à des infections répétées (Doolan et al., 2009). Cette immunité, dite acquise, est complexe et lente à se mettre en place. Les jeunes enfants acquièrent d'abord une immunité clinique, qui leur permet de contrôler la maladie, c'est-à-dire de réduire les risques de complications et de décès pour un niveau de parasitémie donné. Une immunité dite anti-parasitaire se met ensuite en place conduisant au contrôle du niveau de parasitémie. A l'âge adulte, les hauts niveaux d'immunité acquise permettent de contrôler la parasitémie à des densités relativement faibles. Aussi dans ces régions, la majorité des adultes, quand ils sont infectés, sont asymptomatiques et les cas de paludisme graves concernent essentiellement les jeunes enfants de moins de cinq ans.

Il n'existe pas à l'heure actuelle de consensus sur les déterminants majeurs de cette immunité ni sur les mécanismes à l'œuvre (Crompton et al., 2014; Doolan et al., 2009). Les anticorps y jouent un rôle prépondérant ainsi que d'autres cellules immunitaires telles que les cellules T et les cellules NK (*Natural Killer*). L'immunité acquise peut intervenir à différents niveaux du cycle de *P. falciparum*. L'immunité acquise naturelle au stade hépatique semble relativement faible mais elle peut être induite, par exemple par l'inoculation de sporozoïtes atténués, ayant la capacité d'infecter le foie

mais pas de s'y répliquer. Le vaccin RTS,S, également, vise à stimuler la production d'anticorps à ce stade et à empêcher les parasites d'envahir et de se développer dans les hépatocytes. Il cible une protéine de surface du sporozoïte, la CSP (pour *circumsporozoite protein*). L'importance de l'immunité acquise naturelle au stade érythrocytaire et le rôle clé joué par les anticorps à ce stade ont été démontrés dans les années 1960 par Cohen et ses collaborateurs, qui ont montré que le transfert d'immunoglobulines G purifiées provenant d'adultes immuns à des enfants ayant développé un paludisme grave conduisait à une réduction rapide de la parasitémie et à une disparition de la fièvre. A ce stade, les anticorps ont la capacité de se fixer à la fois aux mérozoïtes qui sont libérés dans la circulation sanguine (les empêchant d'infecter de nouveaux globules rouges) et aux antigènes à la surface des hématies infectées, entrainant ensuite leur phagocytose. Cette réponse anticorps n'est efficace qu'après plusieurs années d'infections répétées, du fait de la diversité génétique des protéines de *P. falciparum* et de la capacité du parasite à varier les protéines exprimées à la surface des globules rouges infectés. Les individus doivent être exposés de manière répétée à un nombre suffisant de clones de parasites avant de développer un répertoire d'anticorps permettant de les protéger.

Du fait de l'acquisition d'une immunité partielle protectrice au cours de l'enfance, les jeunes enfants apparaissent comme un groupe de population à privilégier pour identifier les facteurs de résistance innée au paludisme. En effet, chez les enfants ayant acquis un certain niveau d'immunité, le rythme d'acquisition de l'immunité ayant pu être différent d'un individu à l'autre, l'identification des facteurs innés de résistance est rendue plus complexe. Pour les formes graves de paludisme la plupart des études génétiques dans les régions fortement endémiques ont ciblé des enfants de moins de 5 ans. Pour les formes simples, quelques études ont porté sur des échantillons de très jeunes enfants uniquement, mais ces études ont souvent inclus des échantillons plus larges, incluant des adolescents ou des jeunes adultes, voire l'ensemble de la population.

1.2.4 Facteurs influençant la variabilité de présentation clinique

En résumé, l'infection à *P. falciparum* peut conduire à trois grandes formes de présentation clinique :

- Une infection asymptomatique (les individus sont porteurs de parasites mais ne présentent pas de symptôme clinique);
- Un accès palustre simple (caractérisé par la présence de parasites dans le sang accompagnée de symptômes cliniques dont la fièvre) ;
- Un accès palustre grave (caractérisé par la présence de parasites dans le sang accompagnée de symptômes cliniques et d'un ou plusieurs critères de sévérité).

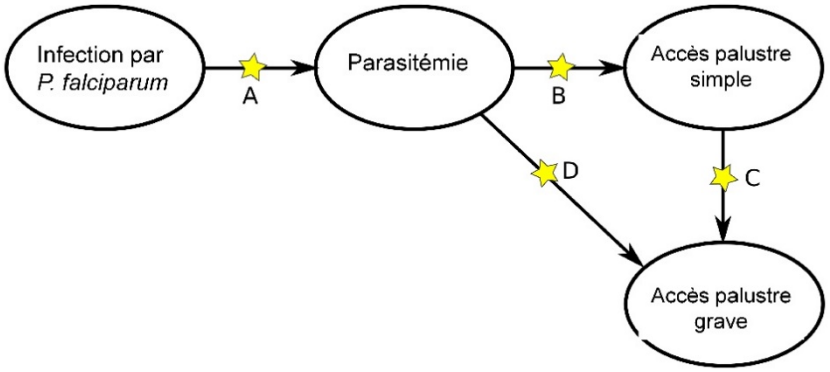


Figure 1.4 Progression de l'infection à P. falciparum de l'inoculation du parasite aux formes graves de paludisme

Ces trois formes correspondent à différentes étapes dans la séquence de l'évolution naturelle de la maladie (Figure 1.4). Il n'est pas établi aujourd'hui si l'ensemble des accès graves surviennent à la suite d'un accès simple (transition C) ou si certaines infections évoluent directement en accès graves (transition D). Comme pour les maladies infectieuses de manière générale, de multiples facteurs peuvent influencer le risque d'infection, l'évolution de la maladie ainsi que sa présentation clinique. On peut distinguer deux grandes catégories de facteurs pour le paludisme : les facteurs environnementaux (niveau d'exposition au vecteur, virulence des parasites) et les facteurs de l'hôte humain (facteurs comportementaux, niveau d'immunité et facteurs génétiques). Ces facteurs peuvent intervenir au niveau des différentes transitions (flèches A à D de la Figure 1.4), de l'inoculation des parasites aux formes graves.

Les facteurs environnementaux

Le niveau d'exposition au vecteur. Il joue un rôle essentiel dans le risque d'infection et dans le risque de développer la maladie. Le développement du moustique dépend des conditions climatiques (précipitations, température, humidité). Dans beaucoup de régions endémiques, même si la transmission est continue, l'intensité de celle-ci varie en fonction des saisons, avec un pic de transmission au moment de, ou juste après, la (ou les) saison(s) des pluies. En plus de cette variation saisonnière, il existe également de plus en plus d'évidence d'une hétérogénéité locale de l'exposition au vecteur entre des villages voisins et au sein d'un même village, qui explique en partie les variations du nombre d'infections chez les enfants (Bousema et al., 2010; Cottrell et al., 2012; Greenwood, 1989). Cette hétérogénéité locale apparait liée à des facteurs tels que l'environnement immédiat de la maison, la proximité d'un cours d'eau, le type de sol et le type de végétation (Cottrell et al., 2012).

La virulence du parasite. Le processus d'infection et les symptômes cliniques résultent d'une interaction complexe entre l'hôte et le parasite. La virulence de *P. falciparum* est liée à sa capacité à envahir et à proliférer à l'intérieur des globules rouges ainsi qu'à la capacité des globules rouges infectés à adhérer à la paroi des vaisseaux sanguins. La virulence de *P. falciparum* a été associée en particulier à l'expression de la protéine PfEMP1 (*Plasmodium falciparum Erythrocyte Membrane Protein 1*), protéine parasitaire exprimée à la surface des globules rouges infectés, qui intervient dans la séquestration des parasites (Smith et al., 1995). Cette protéine est codée par de multiples gènes *var* (une soixantaine) qui sont classés en plusieurs sous-catégories : groupes A, B, C, var1, var2csa et var3. Les formes grave de paludisme ont été associées chez les enfants en Afrique, à l'expression de sous-catégories spécifiques de gènes (groupe A et groupe B) ainsi qu'à l'expression de gènes *var* codant des motifs conservés de PfEMP1, appelés domaines cassettes (Berger et al., 2013; Lavstsen et al., 2012). Des variants au sein des gènes *MSP-1* et *MSP-2* (*Merozoïte Surface Protein 1 et 2*) de *P. falciparum,* qui jouent un rôle important dans l'invasion des globules rouges, ont été aussi impliqués dans les formes sévères (Ghanchi et al., 2015; Sahu et al., 2008).

Les facteurs de l'hôte

Les facteurs comportementaux. L'utilisation de mesures de protection contre le vecteur, telles que l'utilisation de moustiquaires ou la pulvérisation d'insecticide intra-domiciliaire au niveau du foyer, permet de réduire le risque d'infection. L'efficacité de ces deux outils a été mise en évidence au travers de nombreuses études. Dans les années 1990, une méta-analyse, considérant 10 études en Afrique subsaharienne, a estimé que l'utilisation d'une moustiquaire imprégnée d'insecticide à longue durée d'action permettait une réduction de 24% des infections par rapport à l'utilisation d'une moustiquaire non-imprégnée, et d'environ 50% comparé à des individus n'utilisant pas de moustiquaires (Choi et al., 1995). Plus récemment, une étude similaire pour la pulvérisation d'insecticides intra-domiciliaire, basée sur 13 études publiées en Afrique subsaharienne, indique une réduction du risque de 62% (Kim et al., 2012). D'autres comportements liés à des pratiques sociales ou culturelles peuvent intervenir sur le risque d'infection. Il est à noter également que le délai de prise en charge du paludisme, après l'apparition des premiers symptômes, a un impact sur la sévérité des symptômes.

Le niveau d'immunité. Comme nous l'avons déjà dit, une immunité protectrice partielle se met en place au cours de l'enfance, avec la répétition des infections. Aussi le risque de développer un accès grave, puis un accès simple diminue avec l'âge, et la majorité des infections chez les adultes sont asymptomatiques. Au niveau des populations, l'intensité de la transmission du paludisme va influencer la rapidité de mise en place de l'immunité acquise (Okiro et al., 2009; Snow et al., 1997).

Paludisme associé à la grossesse. Le paludisme pendant la grossesse, et en particulier l'infection placentaire, a été associé à une sensibilité accrue des enfants au paludisme pendant leur première année de vie**.** Plusieurs études ont trouvé de manière cohérente un effet de l'infection placentaire par *P. falciparum* sur le délai de survenue de la première infection palustre chez les jeunes enfants. Une des hypothèses qui ont été émises pour expliquer la sensibilité accrue aux infections des enfants nés de mère avec un placenta infecté, est un mécanisme de tolérance immunitaire chez les enfants en contact avec des antigènes parasitaires durant la vie *in utero*. Mais plusieurs autres facteurs,

parmi lesquels l'intensité de la transmission du paludisme, peuvent être responsables de cette sensibilité plus importante. L'effet du paludisme associé à la grossesse sur la sensibilité de l'enfant au paludisme n'est cependant pas encore établi aujourd'hui. Dans une revue récente sur le sujet, Kakuru *et al.* (Kakuru et al., 2019) estiment que les preuves en faveur d' un mécanisme de tolérance immunitaire sont limitées, une seule étude ayant pris en compte l'intensité de la transmission dans les analyses (Le Port et al., 2011). Cette étude a été réalisée au sein de notre unité de recherche et a porté sur une des deux cohortes (celle de Tori-Bossito) inclues dans les analyses de ce manuscrit. Elle montre un effet significatif de l'infection placentaire sur le délai de survenue de la première infection (rapport des risques instantanés, RRI = 2,13 [1.24-3.67], $p < 0.01$). Cependant nous avons montré sur cette même cohorte que l'infection placentaire n'apparaissait pas liée à un risque d'infection plus élevé d'accès palustre, après ce premier accès (Bouaziz et al., 2018).

Les facteurs génétiques de l'hôte qui nous intéressent plus particulièrement sont développés dans la partie suivante. Ces facteurs peuvent intervenir au niveau des différentes transitions du schéma de la Figure 1.4. Ils peuvent être impliqués dans des mécanismes de résistance innée au paludisme ou dans le développement de l'immunité protectrice.

1.3 Facteurs génétiques de l'hôte

1.3.1 Un peu d'histoire

La drépanocytose et autres anomalies du globule rouge

On s'est intéressé très tôt aux facteurs génétiques de l'hôte humain impliqués dans la résistance au paludisme. Les généticiens des populations J. B. S. Haldane et G. Montalenti, dès 1949, constatant la superposition de la distribution de la thalassémie et du paludisme dans le sud de l'Europe (Italie, Sicile et Grèce), proposent que le fait d'être hétérozygote (porteur d'un seul allèle délétère) pour la thalassémie confère un avantage sélectif, et que le paludisme ait été la pression de sélection à l'origine des fréquences élevées de la thalassémie dans cette région (Haldane, 1949). Avec cette hypothèse, bien connue sous le nom de Malaria Hypothesis, J. B. S Haldane suggère pour la première

fois le concept de résistance génétique aux infections (Weatherall, 2004). Presque simultanéement, A. C. Allison observe au Kenya que les individus porteurs du trait drépanocytaire (non malades mais qui présentent des caractéristiques intermédiaires au niveau des globules rouges) sont présents à une fréquence anormalement élevée uniquement dans les régions où sévit le paludisme. A cette époque, le gène responsable de la drépanocytose n'est pas connu, mais deux articles (Beet, 1949; Neel, 1949) ont montré de manière indépendante que la drépanocytose était une maladie héréditaire et que les porteurs du trait drépanocytaire étaient hétérozygotes pour l'allèle délétère, et les individus malades homozygotes (porteurs de deux allèles). Les travaux d'A. C. Allison, publiés un peu plus d'une dizaine d'années plus tard, apporteront un ensemble d'éléments confirmant l'hypothèse de protection des individus hétérozygotes. Ils montrent notamment que les individus porteurs du trait drépanocytaire sont moins infectés par *P. falciparum* (Allison, 1954).

Il a été établi par la suite que la drépanocytose était due à un variant du gène *HBB*, codant pour la chaîne β de l'hémoglobine conduisant à la production d'une hémoglobine anormale (hémoglobine S, HbS). Les individus homozygotes pour HbS souffrent de la drépanocytose, maladie qui entraine une déformation des globules rouges avec des conséquences physiologiques très importantes et associée longtemps à une forte probabilité de décès avant l'âge de cinq ans dans les régions endémiques. Les individus hétérozygotes par contre, présentent très peu de symptômes et ont un risque dix fois moindre de faire un accès grave (Ackerman et al., 2005; Hill et al., 1991). L'hémoglobine S est un cas d'école de sélection balancée dans les populations humaines sur lequel nous reviendrons dans la partie concernant la sélection naturelle.

A la suite des découvertes d'Allison, d'autres anomalies du globule rouge présentant des fréquences anormalement élevées dans des régions endémiques sont étudiées (Nagel & Roth, 1989). Il s'agit de deux autres variants du gène *HBB*, HbE (Flatz, 1967) et HbC (Labie et al., 1984), de variants impliqués dans la régulation de l'expression des gènes *HBB* et *HBA* (codant pour la chaîne α de l'hémoglobine) impliqués respectivement dans la β-thalassémie et l'α-thalassémie, d'un variant dans la protéine

SLC4A1 (ou protéine Bande 3) codant pour le groupe sanguin Diego et responsable de l'ovalocytose (Amato & Booth, 1977) et des variants dans le gène *G6PD* (Glucose-6-Phosphate Déshydrogénase) entrainant un déficit de cette enzyme impliquée dans la résistance des globules rouges au stress oxydatif (Bienzle et al., 1972).

Caractérisation des facteurs génétiques

Pour identifier le rôle des facteurs génétiques, les premières études se basent sur des méthodes épidémiologiques. L'association entre les variants génétiques et le(s) trait(s) relatif(s) à la maladie, appelé(s) phénotype(s) en génétique, est testée dans des études cas-témoins. Ces études contrastent les fréquences alléliques entre un échantillon de cas (individus malades, ou présentant des caractères de gravité, par exemple une forte densité parasitaire) et des individus sains (non malades, ou avec des symptômes modérés, appelés témoins).

A partir des années 1990, d'autres approches, spécifiques à la génétique, telles que l'analyse de ségrégation et l'étude de jumeaux sont utilisées. L'analyse de ségrégation étudie la transmission du phénotype au sein des familles et permet de déterminer s'il existe un gène avec un effet majeur parmi les différentes sources de corrélation familiale (Demenais et al., 1996). Les trois études réalisées ont montré l'effet d'un gène majeur dans le contrôle de la densité parasitaire dans le sang au Cameroun (Abel et al., 1992; Garcia, Cot, et al., 1998) et au Burkina Faso (P. Rihet, Abel, et al., 1998) et dans différents contextes de transmission, en milieu rural et urbain (P. Rihet, Abel, et al., 1998). Si la première étude met en évidence le rôle d'un facteur génétique avec un mode de transmission récessif, les deux études suivantes (Garcia, Cot, et al., 1998; P. Rihet, Abel, et al., 1998) s'accordent sur le fait que les résultats obtenus ne sont pas compatibles avec la transmission mendélienne d'un seul gène, et concluent à l'existence d'un contrôle génétique complexe pour les niveaux d'infection dans le sang. Les études de jumeaux comparent les similitudes de phénotype entre des jumeaux monozygotes (ou vrais jumeaux qui sont identiques au niveau de leur séquence ADN) et les jumeaux dizygotes (ou faux-jumeaux, qui partagent comme des frères et sœurs, en moyenne 50% de leurs gènes). Elles sont utilisées afin de distinguer les rôles respectifs des facteurs

génétiques et des facteurs environnementaux, en faisant l'hypothèse que l'environnement partagé entre jumeaux monozygotes et jumeaux dizygotes est le même. Jepson et al. (A. P. Jepson et al., 1995) ont réalisé un suivi longitudinal (suivi des individus en population sur une certaine période de temps) chez des paires de jumeaux en Gambie durant une saison de transmission. Cette étude a montré une corrélation significativement plus importante chez les jumeaux monozygotes que chez les jumeaux dizygotes pour le développement d'un accès palustre, indiquant que le développement d'une infection fébrile était au moins en partie génétiquement déterminé. L'ensemble de ces études familiales ont confirmé le rôle des facteurs génétiques dans le contrôle de la densité parasitaire et dans la survenue des accès simples.

Enfin, en 2005, Mackinnon et al. (Mackinnon et al., 2005) ont quantifié la part des facteurs génétiques (ou héritabilité) dans le risque de survenue des accès simples et graves. L'héritabilité, pour un trait quantitatif, est définie comme la proportion de variance d'un trait expliquée par les facteurs génétiques. Ils ont utilisé pour cela les données de deux suivis longitudinaux (640 et 2914 enfants respectivement pour les accès simples et graves) pour estimer l'incidence des deux formes. Dans ces études, plusieurs enfants ont été suivis par foyer (partageant donc un même environnement) et les relations de parenté entre les individus au sein et entre les foyers ont été identifiées, ce qui a permis de distinguer la part respective des facteurs génétiques et des facteurs environnementaux. Ils ont estimé que 24 % et 25 % de la variation totale de l'incidence des accès palustres simples et des accès graves respectivement, chez les enfants, pouvaient s'expliquer par des facteurs génétiques et que l'hémoglobine S, le facteur génétique le plus important connu jusqu'alors, en expliquait seulement 2 %.

1.3.2 La région 5q31-q33 et la région HLA

Les recherches utilisant des données familiales ont été poursuivies par des études de liaison. Le but de ces études est de localiser sur le génome les facteurs génétiques, en analysant la co-transmission des marqueurs et de la maladie (ou d'un trait) au sein des familles (Morton, 1955). Ces études ont mis en évidence l'implication de la région 5q31-5q33 dans le contrôle de la parasitémie, quasi-

simultanément au Cameroun et au Burkina Faso (Garcia, Marquet, et al., 1998; P. Rihet, Traoré, et al., 1998) et de la région HLA (6p21-p23) dans le contrôle des accès palustres simples en Gambie (Jepson et al., 1997). Le rôle de ces deux régions a été confirmé par la suite par plusieurs autres analyses de liaison (détaillées dans la revue de la littérature réalisée par S. Marquet (Marquet, 2017)).

1.3.3 Les études d'association gènes candidats

De nombreuses études d'association gènes candidats ont été conduites depuis les premières découvertes sur les polymorphismes des globules rouges, pour identifier les gènes impliqués dans la résistance aux différentes formes de paludisme. Elles ont porté en premier lieu sur les gènes jouant un rôle dans la physiologie des globules rouges, dans le phénomène de cytoadhérence des globules rouges infectés et dans la réponse immunitaire à l'infection.

Plusieurs revues de la littérature ont fait la synthèse des résultats obtenus (Driss et al., 2011; Kwiatkowski, 2005; Marquet, 2017; Verra et al., 2009). En ce qui concerne les formes simples de paludisme, plusieurs gènes impliqués dans la physiologie des globules rouges ont été associés à la fois au contrôle de la parasitémie et dans la protection contre les accès palustres simples. Il s'agit du gène *HBB* (allèles HbS et HbC) ainsi que du gène *G6PD*. L'haptoglobine (*HP*), qui protège contre le stress oxydatif engendré par la lésion des globules rouges a été également associé à la survenue des accès simples. Les études gènes candidats ont porté également sur les régions 6p21-p23 et 5q31-q33, afin d'identifier plus précisément le(s) gène(s) responsable(s) des pics de liaison dans ces régions. Bien que plusieurs associations aient été mises en évidence dans ces régions, les gènes et polymorphismes impliqués n'ont pas encore été précisément identifiés. Dans la région 6p21-p23 liée aux accès simples, les gènes *TNF* et *NCR3* ont tous les deux été trouvés associés à plusieurs reprises. *TNF* a été le candidat le plus étudié ; plusieurs polymorphismes ont été associés avec les accès palustres simples (dont *TNF*-308A, *TNF*-1031, *TNF*-851 et *TNF*-1304). La protéine TNF est une cytokine pro-inflammatoire impliquée dans la destruction de *P. falciparum* par les neutrophiles et les monocytes (Kumaratilake et al., 1990). *NCR3* a été a été mis en évidence dans une étude familiale au

Burkina Faso (Delahaye et al., 2007) et a été répliqué récemment en République Démocratique du Congo (Baaklini et al., 2017). Le gène *NCR3* code pour un récepteur/activateur des cellules Natural killer (NK), qui joue un rôle dans la réponse à l'infection, en interagissant directement avec les globules rouges infectés et en induisant la production d'INFγ. Dans la région 5q31-q33, plusieurs gènes candidats ont été trouvés associés dans une étude mais ces associations n'ont pour le moment pas été répliquées (*IRF1* et *ARHGAP26* dans des populations en Afrique de l'Ouest, *ADRB2* en Inde).

Durant cette thèse, j'ai compilé l'ensemble des associations publiées avec une des trois formes cliniques du paludisme. L'objectif était d'établir une liste de gènes pour chaque phénotype avec le nombre de fois où ils ont été retrouvés, permettant d'obtenir d'une manière pragmatique un niveau de confiance dans les différentes associations identifiées. En effet, les tests statistiques comportent une part d'incertitude liée au fait que l'association peut être due au hasard ; la réplication des résultats dans une ou plusieurs populations indépendantes est donc essentielle pour confirmer qu'une association est bien réelle et qu'elle n'est pas un artefact statistique. La recherche bibliographique a été réalisée sur la période allant des premières associations génétiques à avril 2018 en utilisant des revues de la littérature sur le sujet (Driss et al., 2011; Kwiatkowski, 2005; Verra et al., 2009) et des recherches sur Pubmed (cf. encadré de la Figure 1.5 pour les détails de la méthodologie).

Recherche bibliographique sur les gènes associés au paludisme à *P. falciparum*

Une recherche des gènes trouvés associés aux phénotypes palustres dans la littérature a été effectuée à partir de revues bibliographiques déjà existantes sur les gènes impliqués dans le paludisme : Kwiatkowski *et al.*, 2005, Verra *et al*., 2009 et Driss *et al.*, 2011. Celle-ci a été complétée par une recherche exhaustive dans Pubmed avec les termes ((malaria) AND gene AND associat*) sur la période après 2010. Elle s'arrête en avril 2018.

Cette revue de la littérature a porté sur les **études d'association génétique** réalisées **chez l'Homme**. Elle a consisté à répertorier les gènes présentant une association significative avec l'une des trois formes cliniques du paludisme : infections asymptomatiques, accès simples et accès graves. Elles incluent à la fois les études gènes candidats et les approches génome entier.

En revanche, cette revue ne prend pas en compte :

- Les études sur des phénotypes liés au paludisme gestationnel ou placentaire ; sur l'issue des accès graves ; sur des phénotypes immunologiques,
- Les études de liaison génétique,
- Les études *in vitro* ou conduites chez l'animal (modèles souris et singe)
- Les études utilisant des données d'expression de gènes ou de protéines

Une association a été considérée comme significative si elle était présentée comme telle par les auteurs de l'article. Pour les premiers articles incluant peu de polymorphismes, la *p-valeur* seuil est généralement 0.05. Le nombre de gènes par étude ainsi que le nombre de polymorphismes testés par gène augmentant de manière générale au cours du temps, des corrections pour les tests multiples plus ou moins stringentes ont été appliquées. Nous nous sommes basés sur les interprétations des auteurs pour considérer que l'association était significative ou non.

L'ensemble des associations relevées ont été compilées dans une base de données. L'objectif étant de recenser les gènes trouvés associés au paludisme avec le nombre de fois où l'association a été répliquée, l'unité d'enregistrement a été définie par « un gène associé à une forme clinique donnée, dans une population (ou groupe ethnique) donnée »
Ainsi deux articles portant sur l'étude d'un même gène dans un même échantillon d'individus mais avec des polymorphismes différents (densification des marqueurs ou portant sur des régions différentes du gène) ne constituent qu'un seul enregistrement dans la base de données (une exception concerne le gène *HBB* pour lequel nous avons distingué les associations concernant l'allèle HbS de celles concernant l'allèle HbC). A l'inverse, un seul article peut donner lieu à plusieurs enregistrements. C'est le cas des études incluant à la fois des cas graves, des cas simples et des cas asymptomatiques (si l'association est trouvée avec plusieurs formes), des études montrant des associations avec plusieurs gènes, ou des études incluant plusieurs populations ou groupes ethniques et présentant des résultats par population.

Certains échantillons ayant été utilisés pour de nombreuses études successives, nous avons examiné de plus près *a posteriori* les doublons (même gène trouvé associé plusieurs fois dans le même pays) et conservé qu'un seul résultat lorsque ceux-ci portaient sur le même échantillon (ou avec une proportion importante d'individus en commun dans les deux échantillons). Ce cas de figure se présente essentiellement (mais pas uniquement) pour l'allèle HbS (qui a fait souvent l'objet d'un premier article et dont les résultats d'association sont publiés à nouveau dans les études suivantes sur d'autres gènes) et pour les cohortes qui ont été inclues dans les études du consortium *Malaria Genomic Epidemiology Network*. Pour un certain nombre de gènes (*HBB*, *G6PD, ABO, CD40LG* par exemple), des associations significatives ont déjà été publiées dans plusieurs populations, avant d'être confirmées par les études multicentriques et les GWAS.

Les analyses présentées dans cette partie du manuscrit ne concernent que les **études d'association gènes candidats** ayant porté sur le **paludisme à *P. falciparum***. Les régions utilisées dans la Figure 1.5 et dans les Tables des gènes correspondent aux régions géographiques. Le continent africain a été divisé en cinq régions d'après les régions définies par l'ONU (Annexe 1, Figure 1).

Figure 1.5 Recherche bibliographique sur les gènes associés au paludisme à P.falciparum

Nous avons recensé 202 articles avec au moins une association significative, et un total de 351 associations gène/phénotype distinctes. Lorsque l'on regarde la distribution des articles en fonction de leur année de publication (Figure 1.6), on remarque que leur nombre augmente fortement depuis les années 1995-1999. Il faut noter que la dernière barre du graphique, moins élevée que les précédentes, couvre une période plus courte (un peu plus de trois ans jusqu'en avril 2018 contre 5 ans pour les autres) mais ne reflète pas pour autant une diminution importante du nombre de publications (un peu moins de 13 en moyenne sur la période 2010-2014 contre un peu plus de 9 sur la période 2015-2018).

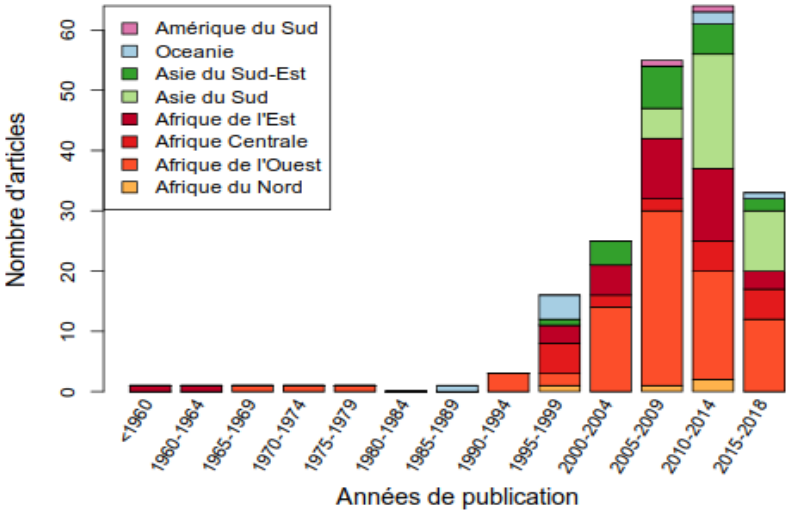


Figure 1.6 Nombre d'études gènes candidats publiées incluant au moins une association significative par période de 5 ans
Sont considérées les études portant sur P. falciparum montrant une association avec l'un des trois phénotypes palustres étudiés

La localisation des populations d'étude tend aussi à se diversifier au cours du temps. Si les premières études ont été réalisées essentiellement en Afrique, où la population est la plus touchée par le paludisme, et où se concentre la majorité des infections à *P. falciparum,* elles ont été étendues depuis les années 2005-2009 aux autres continents (Asie, Océanie et Amérique du Sud). Depuis cette période, une part importante des études (37%) sont réalisées en Asie, plus particulièrement en Inde pour la région Asie du Sud et en Thaïlande pour la région Asie du Sud-Est. Cette représentation ne

concerne que les infections à *P. falciparum* et n'inclut donc pas les études génétiques réalisées en Asie et en Amérique du Sud sur les infections à *P. vivax*.

La répartition des études est très inégale suivant les trois formes. La majorité des études (n=137 soit 68%) a porté sur les accès graves. Dans ces études, 90 gènes ont été trouvés associés au moins une fois, dont 41 associés dans au moins deux échantillons indépendants. Les gènes qui ont été trouvés associés le plus grand nombre de fois sont *HBB*, *ABO* (codant pour le système de groupe sanguin du même nom) et *G6PD*. En comparaison, on dénombre 44 (22%) et 21 (10%) études pour les accès simples et les infections asymptomatiques respectivement. Elles ont permis d'identifier 45 gènes (10 pour lesquels l'association a été répliquée au moins une fois) pour les accès simples et 17 gènes (5 pour lesquels l'association a été répliquée au moins une fois) pour les infections asymptomatiques. Les gènes qui ont été trouvés associés dans au moins deux échantillons indépendants sont présentés dans les Tables 1 et 2, pour les infections asymptomatiques et les accès palustres respectivement. Une table similaire concernant les associations avec les accès palustres graves figure en Annexe (Annexe 1 Table 1).

Les gènes qui ont été retrouvés le plus souvent associés correspondent à ceux cités par les revues de la littérature et mentionnés au début de cette section. Pour le phénotype accès palustre simple, on note cependant que *NOS2 (Nitric oxide synthase2)*, qui est cité dans la plupart des revues comme uniquement associé aux accès graves, fait partie des quatre gènes pour lesquels on observe le plus grand nombre de réplications avec *HBB*, *TNF* et *HP* (4 études distinctes; Table 2). *NOS2* code pour une enzyme produisant de l'oxyde nitrique, un radical libre qui a notamment des propriétés antiparasitaires. On note également des disparités au niveau de la répartition géographique des associations identifiées : certaines sont répliquées dans un grand nombre de régions distinctes (allèle HbS du gène *HBB*, *TNF*, *HP*) alors que d'autres apparaissent plus spécifiques de certaines régions : allèle HbC ou gène *LTA* trouvés associés uniquement en Afrique de l'Ouest, *NOS2* pour lequel à l'inverse aucune association n'a été trouvée en Afrique de l'Ouest. Ces disparités peuvent

être liées à des différences de fréquences des variants conférant une protection, comme c'est le cas pour l'allèle HbC qui est présent uniquement en Afrique de l'Ouest dans une région allant du Nord du Mali, au Ghana et au Bénin. Dans la dernière étude multicentrique sur les accès graves réalisée récemment (Malaria Genomic Epidemiology Network, 2019), les auteurs suggèrent aussi une hétérogénéité des effets génétiques entre les populations (liée à des interactions gène-gène ou des interactions gène-environnement).

Gene	**N. assos.**	**Populations d'étude : Pays (groupe ethnique)**	**Références**	**Nombre d'associations par région**
HBB (HbS)	8	Ouganda (Luganda) Tanzanie (Bondei) Ghana Gambie Cameroun Tanzanie Senegal (Serere) Burkina Faso	Allison, 1954 Allison & Clyde, 1961 Ringelhann et al., 1976 Allen et al., 1992 Le Hesran et al., 1999 Stirnadel et al., 1999 Sokhna et al., 2000 Mangano et al., 2015	Afr. de l'Ouest : 4 Afr. de l'Est : 3 Afr. Centrale : 1
HBB (HbC)	3	Ghana Burkina Faso Burkina Faso	Ringelhann et al., 1976 Rihet et al., 2004 Mangano et al., 2015	Afr. de l'Ouest : 3
G6PD	3	Tanzanie (Bondei) Gambie Mali (Dogon)	Allison & Clyde, 1961 Okebe et al., 2014 Maiga et al., 2014	Afr. de l'Ouest : 2 Afr. de l'Est : 1
LTA	2	Burkina *(2 échantillons indépendants)*	Barbier et al., 2008	Afr. de l'Ouest : 2
MBL2	2	Ghana (Dagomba) Gabon	Holmberg et al., 2008 Boldt et al., 2009	Afr. de l'Ouest: 1 Afr. Centrale : 1

Table 1.1 Gènes trouvés associés au moins deux fois avec les infections asymptomatiques dans les études d'association gènes candidats

Pour *HBB*, nous avons distingué les associations concernant HbS de celles impliquant HbC. *N. assos, nombre de fois où l'association a été retrouvée ; Populations d'étude, pays où se sont déroulées les études, avec entre parenthèses le nom du groupe ethnique majoritaire (quand celui-ci est précisé).*

Gene	N. assos.	Populations d'étude : Pays (groupe ethnique)	Références	Nombre d'associations par région
HBB HbS	10	Gambie Burkina Faso (Mossi), Kenya (Luo) Kenya (Giriama) Senegal (Serere) Mali Ghana Soudan (Hausa) Soudan (Massalit) Angola	Allen et al., 1992 Modiano et al., 2001 Aidoo et al., 2002 Williams et al., 2005 Migot-Nabias et al., 2006 Crompton et al., 2008 Kreuels et al., 2009 Salih et al., 2010 Salih et al., 2010 do Sambo et al., 2015	Afr. du Nord : 2 Afr. de l'Ouest : 5 Afr. Centrale : 1 Afr. de l'Est : 2
TNF	7	Kenya (Luo) Gabon Burkina Faso Inde Tanzanie République du Congo Nigeria	Aidoo et al., 2001 Meyer et al., 2002 Flori et al., 2005 Basu et al., 2010 Gichohi-Wainaina et al., 2015 Nguyen et al., 2017 Ojurongbe et al., 2018	Afr. de l'Ouest : 2 Afr. Centrale : 2 Afr. de l'Est : 2 Asie du Sud : 1
HP	4	Soudan Ghana Kenya (Giriama) Gambie	Elagib et al., 1998 Quaye et al., 2000 Atkinson et al., 2007 Cox et al., 2007	Afr. du Nord : 1 Afr. de l'Ouest : 2 Afr. de l'Est : 1
NOS2	4	Gabon Tanzanie Cameroun (semi-Bantu), Inde	Kun et al., 1998 Hobbs et al., 2002 Apinjoh et al., 2014 Kanchan et al., 2015	Afr. Centrale : 2 Afr. de l'Est : 1 Asie du Sud : 1
G6PD	3	Nigeria Kenya Mali (Fulani)	Bienzle et al., 1972 Ruwende et al., 1995 Maiga et al., 2014	Afr. de l'Ouest : 2 Afr. de l'Est : 1
HBB HbC	2	Burkina Faso (Mossi) Burkina Faso	Modiano et al., 2001 Rihet et al., 2004	Afr. de l'Ouest : 2
HBA1 *HBA2*	2	Vanuatu Tanzanie	Williams et al., 1996 Enevold et al., 2008	Afr. de l'Est : 1 Océanie : 1
NCR3	2	Burkina Faso République du Congo	Delahaye et al., 2007 Baaklini et al., 2017	Afr. de l'Ouest : 1 Afr. Centrale : 1
TIRAP	2	Iran Pakistan	Zakeri et al., 2011 Nawaz et al., 2015	Asie du Sud : 2
TLR9	2	Burundi Inde	Esposito et al., 2012 Sawian et al., 2013	Afr. de l'Est : 1 Asie du Sud : 1

Table 1.2 Gènes trouvés associés au moins deux fois dans les études d'association gènes candidats pour les accès palustres simples.

Pour *HBB*, nous avons distingué les associations concernant HbS de celles impliquant HbC. *N. assos, nombre de fois où l'association a été retrouvée ; Populations d'étude, pays où se sont déroulées les études, avec entre parenthèses le nom du groupe ethnique majoritaire (quand celui-ci est précisé).*

1.3.4 Les approches génome entier

Ces dix dernières années, plusieurs études d'association sur l'ensemble du génome (*Genome-Wide Association Study* ou GWAS) ont été réalisées pour les accès palustres graves. Cette approche, contrairement aux études gènes candidats ne fait pas d'hypothèse *a priori* sur la région du génome ou le gène potentiellement impliqué ; elle est basée sur une recherche systématique d'une association entre un phénotype et des centaines de milliers voire des millions de marqueurs répartis sur l'ensemble du génome. Elle utilise des panels denses de variants génétiques (jusqu'à 5 millions), génotypés par des puces à ADN (voire ces dernières années issus des résultats de séquençage haut-débit sur l'ensemble du génome). Les puces à ADN permettent de génotyper rapidement un sous-ensemble de variants du génome sur un grand nombre d'individus. Ces variants, en majorité des polymorphismes d'un seul nucléotide (SNPs pour S*ingle Nucleotide Polymorphisms*), sont sélectionnés pour être représentatifs de l'ensemble des variants communs du génome (avec une fréquence supérieure à 5% ou 1% suivant les puces). En effet, les variants proches dans une région ne sont pas indépendants et il existe le long des chromosomes, ce que l'on appelle des blocs de déséquilibre de liaison (DL) au sein desquels ils sont fortement corrélés.

Le principe général des GWAS est de tester l'association entre chaque variant pris individuellement et le phénotype, puis de corriger les résultats pour le grand nombre de tests effectués, afin de limiter la proportion de faux positifs. Dans le cas d'une GWAS utilisant une puce à ADN, et de manière générale dans les études d'association (que ce soit une étude gène candidat ou une étude sur l'ensemble du génome) un signal d'association au niveau d'un variant n'indique pas forcément que ce variant est impliqué dans le trait ou la maladie. Il peut être en DL avec un ou des variants

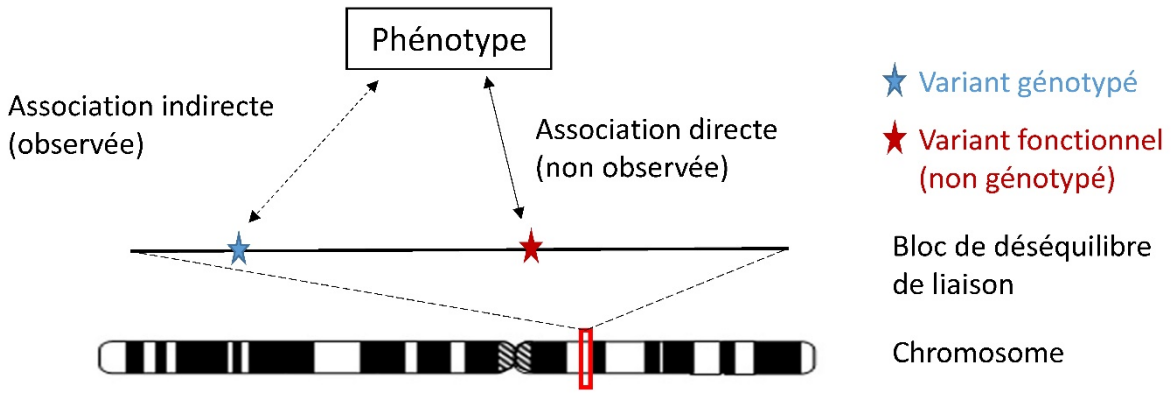


Figure 1.7 Association indirecte entre le variant génotypé et le phénotype

fonctionnels se trouvant dans le bloc de déséquilibre de liaison (association indirecte, Figure 1.7). Aussi des analyses fonctionnelles sont ensuite nécessaires pour identifier le variant responsable du signal d'association et son rôle dans la détermination du phénotype étudié.

La première GWAS réalisée sur les accès graves a été publiée en 2009 par Jallow *et al.* (Jallow et al., 2009), plusieurs années après la publication des premières GWAS (en 2004-2005), principalement réalisées sur des populations d'origine européenne. Cette étude, qui a porté sur environ 6000 enfants gambiens, a mis en avant surtout des problèmes méthodologiques pour l'application de cette approche dans les populations africaines. Elle pointe notamment le plus faible déséquilibre de liaison dans ces populations comparées aux populations européennes ou asiatiques, ainsi que la forte hétérogénéité génétique des populations africaines, même à l'échelle d'un seul pays. Les auteurs recommandent l'utilisation de puces spécifiques, plus adaptées à la diversité génétique des populations africaines, et notamment de puces plus denses permettant d'assurer une couverture plus large des variations communes du génome. Ils préconisent également l'utilisation de l'imputation statistique des génotypes manquants à partir de données de séquences d'un panel de référence mais soulignent l'importance d'utiliser des données de référence qui soient adaptées à la population d'étude. Les données de séquençage de la population des Yoruba du Nigéria, par exemple, se sont révélées inefficaces pour l'imputation statistique des génotypes manquants dans

l'échantillon gambien étudié, du fait de la divergence des profils de DL entre les deux populations à certains loci du génome.

Les études suivantes ont essayé de palier ces difficultés (Band et al., 2013; Malaria Genomic Epidemiology Network, 2019; Malaria Genomic Epidemiology Network et al., 2015; Ravenhall et al., 2018; Timmann et al., 2012), en utilisant des puces plus denses (> 1 million de SNPs), en améliorant la qualité de l'imputation et les méthodes pour prendre en compte la structure de population (facteur de confusion majeur dans les études d'association). L'étude de Timmann *et al.*, en 2012, a porté sur une seule région, la région de Kumasi au Ghana : les étapes de découverte et de réplication ont été réalisées dans deux échantillons de cette région, avant une seconde réplication dans l'échantillon gambien (Jallow et al., 2009). Les autres études, à l'exception de celle de Ravenhall *et al*., sont des études multicentriques réalisées dans le cadre du consortium Malaria Genomic Epidemiology Network (MalariaGEN), dont l'objectif était d'augmenter la taille des échantillons afin d'améliorer la capacité de détection des effets génétiques. L'étude la plus récente, parue en décembre dernier (Malaria Genomic Epidemiology Network, 2019), inclut des échantillons provenant de 11 pays (9 pays africains, le Vietnam et la Papouasie Nouvelle Guinée), soit environ 32 000 enfants (dont 17 000 dans l'étape de découverte). Enfin, la GWAS publiée par Ravenhall *et al.* (Ravenhall et al., 2018) s'est focalisée sur les données d'une étude du consortium MalariaGEN, réalisée en Tanzanie.

Toutes ces GWAS ont porté sur les accès graves et ont utilisé de grands échantillons de cas et de témoins, les cas étant recrutés dans les hôpitaux selon les critères définis par l'OMS. Comme nous l'avons vu précédemment, les accès graves constituent un ensemble hétérogène de syndromes. La plupart de ces études se sont intéressées aux cas graves dans leur ensemble sans distinction entre les sous-types. Seules deux études ont utilisé une définition restreinte des cas en se focalisant sur un sous-type ou en faisant la distinction entre plusieurs sous-types d'accès palustres graves (Malaria Genomic Epidemiology Network, 2019; Timmann et al., 2012). Timman *et al.* ont inclus dans la phase

initiale de leur étude (phase de découverte) uniquement les deux principaux sous-types (neuropaludisme et anémie grave) dont la gravité était avérée par la présence concomitante d'une acidose ou d'une détresse respiratoire. Dans sa dernière étude, le consortium MalariaGEN a considéré séparément trois groupes de cas graves : les cas de neuropaludisme, les cas d'anémie sévère et un groupe incluant tous les autres symptômes.

Les différentes GWAS, à l'exception de la première, confirment l'implication des variants déjà connus des gènes *HBB* et *ABO*. Elles ont permis également de mettre en évidence plusieurs nouveaux gènes : *ATP2B4*, *GYPA-B* et récemment un nouveau locus proche du gène *EPHA7*. *ATP2B4* a été identifié pour la première fois par Timman et al. au Ghana (Timmann et al., 2012) et son association a été répliquée depuis dans deux des quatre autres GWAS (Malaria Genomic Epidemiology Network, 2019; Malaria Genomic Epidemiology Network et al., 2015). *ATP2B4* code pour un transporteur de calcium de la membrane plasmique. Tout récemment, la dernière étude MalariaGEN a identifié le probable variant causal ainsi que le mécanisme par lequel il agirait : un des SNPs associés (rs10751451) perturbe un site de fixation du facteur de transcription GATA1, situé en amont d'un premier exon. Le SNP en question affecterait de manière spécifique les globules rouges au niveau d'un site d'initiation de la transcription qui n'est actif que dans ces cellules. L'allèle associé à une protection est corrélé à une diminution de l'expression d'*ATP2B4*; en revanche, le mécanisme par lequel *ATP2B4* affecte le parasite reste encore à élucider. Un second signal d'association a été identifié par la GWAS publiée par MalariaGEN en 2015 (Malaria Genomic Epidemiology Network et al., 2015) dans le cluster de gènes *GYP*. Ces gènes codent pour des glycophorines qui sont des récepteurs utilisés par *P. falciparum* pour l'invasion des globules rouges. Les travaux de Leffler *et al.* (Leffler et al., 2017) ont montré ensuite que ce signal peut s'expliquer par des variations du nombre de copies des gènes *GYPA* et *GYPB,* une configuration particulière (DUP4) responsable du groupe sanguin Dantu, réduisant le risque de formes sévères. Enfin, un nouveau signal d'association a été identifié sur le chromosome 6 par la dernière GWAS (Malaria Genomic Epidemiology Network, 2019), en plus des 4 gènes précédemment cités. Ce signal est situé à 700kb du gène codant le plus proche (*EPHA7)* ; les

analyses fonctionnelles n'ont pas permis pour le moment de faire des hypothèses sur le(s) variant(s) et le(s) gène(s) en cause.

Au final ces approches génome entier mettent en évidence les deux gènes majeurs déjà connus, *HBB* et *ABO*, ce qui semble valider la méthodologie utilisée, mais identifient par ailleurs peu de nouveaux gènes en comparaison des études gènes candidats. Même si l'on sait qu'une partie des associations identifiées par les approches gènes candidats sont des faux-positifs, en particulier quand elles n'ont pas fait l'objet de réplications, des gènes parmi les plus fréquemment répliqués ne sont pas retrouvés dans les GWAS : par exemple *NOS2 (nitric oxide synthase 2)* et *TNF* impliqués dans la réponse immunitaire ou les gènes *CD36* et *ICAM1,* des récepteurs de l'endothélium des vaisseaux sanguins impliqués dans le phénomène de cytoadhérence. Dans leur dernière étude, le consortium MalariaGEN estime que les gènes *HBB*, *ABO*, *GYPA-B* et *ATP2B4* agissent largement de manière indépendante et expliquent ensemble environ 2,5 % de la variabilité totale du phénotype (soit environ 10 % de l'héritabilité des formes graves), ce qui semble indiquer que les GWAS mettent en évidence uniquement la partie émergée de l'iceberg. Dans cette dernière étude, le consortium MalariaGEN met en avant plus particulièrement une hétérogénéité des effets des variants entre les populations pour un grand nombre de polymorphismes dans leur étude, dont ceux des gènes *HBB* et *ABO*. Par exemple pour l'allèle HbS, l'effet estimé varie d'un odds-ratio (OR) de 0,10 en Gambie à un OR de 0,47 au Cameroun, une différence qui ne semble pas expliquée, d'après les auteurs, par la présence en *sus* de l'allèle protecteur HbC dans les populations d'Afrique de l'Ouest. Cette hétérogénéité de manière générale pourrait être liée à la fois à des différences d'intensité de transmission du paludisme entre les pays et/ou à la variabilité génétique du parasite. Le succès relatif de ces méta-analyses pourrait également s'expliquer par la diversité génétique des populations qu'elles incluent et par l'hétérogénéité des effets entre les sous-types d'accès palustres graves. En effet, malgré la taille conséquente de l'échantillon (17 000 individus dans l'étape de découverte), l'identification de variants avec un effet modeste est rendue difficile car les variants conférant une protection ne sont pas forcément présents dans toutes les populations, leur distribution peut même

être limitée à quelques populations. C'est le cas par exemple de l'allèle HbC qui est présent uniquement en Afrique de l'Ouest ou du variant DUP4 pour le groupe sanguin Dantu qui est rare en dehors de l'Afrique de l'Est. L'hétérogénéité des effets des variants entre les sous-types de formes graves a pu réduire par le passé la capacité des précédentes GWAS à identifier les variants d'intérêt. En même temps, l'analyse par sous-type de phénotype réduit considérablement la taille des échantillons par pays et limite la puissance de l'analyse. Les auteurs recommandent aussi d'augmenter la taille des échantillons au sein des différents pays pour les études futures.

1.3.5 Prise en compte de la stratification de population dans les GWAS

La stratification de population, aussi communément appelée structure de population, désigne la présence de plusieurs populations ou sous-populations génétiquement distinctes dans un échantillon d'étude. Au cours de l'histoire de l'humanité, au fur et à mesure de l'expansion de l'Homme, en Afrique puis hors d'Afrique, des groupes de populations se sont formés. L'isolement géographique des différents groupes, combinée à des histoires évolutives différentes (adaptation à des environnements variés, migrations, variation aléatoire des fréquences alléliques aux cours des générations, appelée dérive génétique, effet fondateur) ont entraîné des différences de fréquences alléliques entre les populations. Une stratification de population peut être observée également à l'échelle locale quand plusieurs ensembles d'individus vivent dans une même zone mais avec des échanges génétiques limités du fait de différences socio-culturelles. C'est le cas de certaines ethnies par exemple en Afrique. Dans la suite du manuscrit, nous utiliserons également le terme de strates qui fait référence à des groupes distincts au sein d'une étude.

La stratification de population est le facteur de confusion majeur dans les études d'association et peut conduire à l'obtention de nombreux faux positifs. Si la prévalence (ou l'intensité) du phénotype d'intérêt varie entre les différentes strates de population, celui-ci va être associé à tous les variants génétiques dont la fréquence varie également entre les différentes strates dans le même sens. Une première solution a consisté à utiliser des approches basées sur des données familiales pour s'affranchir de ce problème, comme le TDT (pour Transmission Disequilibrium Test (Spielman et al.,

1993) ou plus récemment FBAT (pour Family Based Association Test (Rabinowitz & Laird, 2000). Ces méthodes, très efficaces, utilisent l'information de transmission des allèles des parents aux enfants atteints. Néanmoins, l'inconvénient de celles-ci est qu'elles nécessitent le recrutement et le génotypage des parents des individus et posent des problèmes de faisabilité lorsque les sujets de l'étude ont atteint un certain âge ou lorsque la taille des échantillons devient importante. Aussi, d'autres approches ont été développées avec l'avènement des GWAS: le *genomic control* (Devlin & Roeder, 1999) qui utilise la distribution empirique de la statistique sur l'ensemble du génome pour corriger pour l'inflation due à la stratification de population, et un ensemble de méthodes utilisant les données génomiques pour inférer la structure de population. Parmi les méthodes les plus utilisées figurent la méthode *Structured Association* proposée par Pritchard et *al.* (J. K. Pritchard et al., 2000), et l'analyse en composantes principales (ACP) qui est implémentée dans EIGENSTRAT (Price et al., 2006). La première procède en inférant les différentes strates de populations puis en testant l'association conditionnellement aux strates. L'ACP est une méthode de réduction linéaire de dimensions. Appliquée aux variants génétiques répartis sur l'ensemble du génome, elle permet d'identifier des axes orthogonaux de variation, appelés composantes principales (ou *principal component*, PCs) expliquant au mieux la variabilité existant parmi les individus de l'échantillon. Les premières PCs sont ensuite simplement incluses comme variables d'ajustement dans le modèle standard (régression linéaire ou régression logistique).

Le modèle linéaire mixte (MLM), qui incorpore la structure de variance-covariance des individus estimée à partir des données génomiques, a ensuite été proposé pour analyser les traits quantitatifs. Ce modèle peut être vu comme une généralisation de la méthode précédente où l'ensemble des PCs seraient incluses dans le modèle (Claire Dandine-Roulland, 2016). Il est particulièrement attractif dans la mesure où il permet de corriger à la fois pour la stratification de population et pour l'apparentement entre les individus (Sul et al., 2018), un second type de structure de population pouvant entrainer également une inflation de la statistique dans les études d'association. Son implémentation dans le cadre des GWAS a fait l'objet de nombreuses recherches (Sul et al., 2018),

pour réduire les temps de calcul à mesure que les échantillons et la densité des marqueurs ont augmenté. Une première méthode basée sur une approximation du modèle (GRAMMAR, (Aulchenko et al., 2007)) a été proposée avant qu'une méthode exacte puisse être appliquée (FaST LMM, (Lippert et al., 2011)). La méthode BOLT-LMM (Loh et al., 2015) a été développée plus récemment afin de pouvoir analyser des cohortes de centaines de milliers d'individus.

Pour les autres types de traits (binaires ou données répétées), le problème de temps de calcul pour les modèles mixtes non-linéaires étant encore plus prégnant, leur application n'était pas envisageable dans les GWAS jusqu'à récemment. Dans le cas des études cas/témoins par exemple, l'ajustement du modèle mixte logistique avec la PQL (pour *Penalized Quasi-likelihood* (Breslow & Clayton, 1993)) ne peut être réalisé dans une étude génome entier. Pour ces études, une alternative consistait alors à analyser le statut cas/témoins, codé respectivement 1 et 0, comme un trait quantitatif avec un MLM. Cependant Chen *et al.*, en 2016 (Chen et al., 2016), ont montré que lorsque la prévalence de la maladie était hétérogène entre les strates de population, cette approche présentait des biais au niveau de la distribution des *p*-valeurs, et ont proposé un test du score pour le modèle mixte logistique applicable à l'ensemble du génome. Dernièrement, les travaux de Chen et al., ont été repris afin d'adapter ce test pour les analyses de très grands échantillons tels que celui de la UKBiobank (Rusk, 2018; Zhou et al., 2018).

1.4 Le paludisme : pression de sélection au cours de l'histoire récente des populations humaines

L'Homme, tout comme les autres espèces animales ou végétales, est soumis à la sélection naturelle telle que définie par Charles Darwin : les individus les mieux adaptés à leur environnement, sont aussi les plus aptes à se reproduire et à transmettre leurs gènes. Aussi la sélection naturelle a-t-elle modelé au cours du temps le génome humain en sélectionnant les variants génétiques conférant un avantage (sélection positive) ou au contraire en éliminant les variants génétiques délétères (sélection négative ou purificatrice). Les études récentes de génétique des populations, utilisant les données du génome humain entier, mettent en évidence l'importance de l'adaptation locale des

populations à leur environnement, ainsi que le rôle prépondérant joué par les pathogènes dans la sélection naturelle (Fumagalli et al., 2011).

Aujourd'hui, le paludisme représente un fardeau pour les populations humaines, causant des centaines de milliers de décès par an chez les enfants. Mais son impact en termes de morbidité et de mortalité a été encore plus important avant le développement des traitements antipaludiques et la mise en place de moyens de lutte anti-vectorielle au XXe siècle. De ce fait, le paludisme est connu pour être l'une des plus fortes forces évolutives imposées par un pathogène dans l'histoire récente de l'humanité.

1.4.1 Les différentes formes de sélection naturelle

La sélection naturelle peut agir de différentes façons et on distingue trois grands types de sélection naturelle : la sélection purificatrice, la sélection positive et la sélection balancée.

La sélection purificatrice

Appelée aussi sélection négative**,** cette sélection diminue la fréquence des allèles délétères (qui confèrent un désavantage aux individus) dans la population. L'intensité de la sélection va dépendre du niveau de nuisance de la mutation. Si celle-ci est fortement délétère, entrainant par exemple une mortalité importante, l'allèle délétère sera rapidement éliminé de la population.

La sélection positive

Appelée aussi sélection directionnelle ou darwinienne, cette sélection, contrairement à la sélection purificatrice, va faire augmenter en fréquence les allèles qui confèrent un avantage aux individus. Les allèles favorables peuvent aller jusqu'à se fixer dans la population (balayage sélectif complet), les autres allèles à ce locus étant éliminés. Lorsque la fixation n'est pas atteinte, on parle de balayage sélectif incomplet ou partiel. Ce dernier cas correspond à une pression de sélection en cours, ou qui s'est affaiblie voire arrêtée avant que l'allèle n'ait eu le temps d'arriver à la fixation. Le temps nécessaire pour la fixation d'un allèle dépend de l'intensité de la sélection mais aussi d'autres

facteurs tels que l'effectif efficace de la population ou l'effet de dominance de l'allèle (s'il s'agit d'un allèle agissant sur le trait de manière dominante, récessive ou codominante).

Dans le modèle classique de sélection positive, appelé hard sweep, la sélection agit sur une nouvelle mutation bénéfique (parmi celles apparaissant aléatoirement) qui va se répandre rapidement dans la population. Deux scénarios légèrement différents de sélection positive ont ensuite été définis et regroupés sous le terme de soft sweep (Hermisson & Pennings, 2005). Dans le premier scénario la sélection positive agit sur un allèle déjà présent dans la population. Suite à une modification de la pression de sélection, due à un changement de l'environnement, du mode de vie ou du régime alimentaire, un allèle jusque-là neutre ou légèrement délétère devient bénéfique. Dans le second, la sélection positive agit sur plusieurs allèles à un même locus ayant un bénéfice équivalent. Dans ce cas, les allèles bénéfiques augmentent en fréquence sans qu'aucun d'entre eux n'atteigne la fixation.

La sélection balancée

Elle agit de manière à maintenir différents allèles dans la population (il n'y a pas de fixation ou de forte augmentation en fréquence d'un allèle). Il existe deux mécanismes principaux conduisant à ce type de sélection : l'avantage de l'hétérozygote et la sélection fréquence dépendante. Dans le premier cas, les individus porteurs de deux allèles différents, hétérozygotes, présentent un avantage sélectif par rapport aux individus homozygotes, ce qui maintient les différents allèles dans la population pendant une longue période de temps. Dans le second cas, la sélection va dépendre de la fréquence de l'allèle dans la population. Ainsi, un individu avec un phénotype considéré comme "rare" peut gagner en survie par rapport au reste du groupe, du fait de la rareté du trait dans la population : on parle alors de sélection fréquence-dépendante négative. La sélection fréquence dépendante intervient dans un contexte de compétition (par exemple dans le choix d'un partenaire dans de nombreuses espèces) ou dans un contexte d'interaction entre les espèces et notamment dans les relations hôte-pathogènes. Le haut niveau de polymorphisme du complexe majeur d'histocompatibilité (CMH) est en partie attribué à la sélection fréquence-dépendante négative (Borghans_2004). Il s'agit d'un phénomène de coévolution faisant intervenir également des pressions

de sélection également sur le pathogène. L'évolution va favoriser les agents pathogènes qui évitent la présentation par les molécules les plus fréquentes du CMH dans la population, entrainant au niveau de la population hôte une sélection des variants rares du CMH.

1.4.2 Détection des signatures de sélection récente à partir des données génétiques

La sélection naturelle laisse des signatures au niveau de l'ADN qui peuvent être détectées par un grand nombre de méthodes. Chaque type de sélection laisse des signatures distinctes. Dans le cas de la sélection positive, lorsqu'un allèle avantageux commence à augmenter en fréquence dans la population, il va entrainer avec lui et faire augmenter en fréquence les allèles proches par un phénomène appelé « auto-stop génétique ». Dans le modèle classique de *hard sweep*, la sélection agit sur un seul allèle nouvellement apparu sur un seul haplotype (combinaison d'allèles à différents sites polymorphiques sur un même chromosome et transmis ensemble lors de la méiose). Aussi la sélection va faire augmenter en fréquence l'ensemble de l'haplotype portant l'allèle avant que le mécanisme de recombinaison ne vienne casser le DL (au bout d'un grand nombre de générations). La région autour du locus va alors présenter une variation génétique réduite, ainsi qu'un spectre de fréquences spécifique avec un excès de fortes et de faibles fréquences alléliques (Figure 1.8a). La Figure 1.8b montre que les signatures sur le génome vont être différentes selon que la sélection naturelle cible une nouvelle mutation (mutation *de novo,* modèle *hard sweep*) ou un variant déjà présent dans la population (*standing variation* et modèle *soft sweep*) (Jonathan K. Pritchard et al., 2010; Vitti et al., 2013). En effet, dans le cas d'un *soft sweep*, la signature présente au niveau de l'ADN est plus subtile. La sélection agit sur un allèle déjà présent dans la population avant le début de la sélection ou sur plusieurs allèles à un même locus. Dans les deux cas, la sélection a pour conséquence de faire augmenter en fréquence plusieurs haplotypes et non un seul. En conséquence, l'augmentation en fréquence de cet allèle va entrainer une réduction moins importante de la diversité et la signature laissée par un *soft sweep* a alors tendance à être plus difficile à détecter en utilisant les tests standards de sélection. L'origine d'un *soft sweep* est le plus souvent liée à une

modification de la pression de sélection sur une population, comme dans le cas d'un changement d'environnement. Un variant jusqu'à présent neutre ou légèrement délétère devient avantageux.

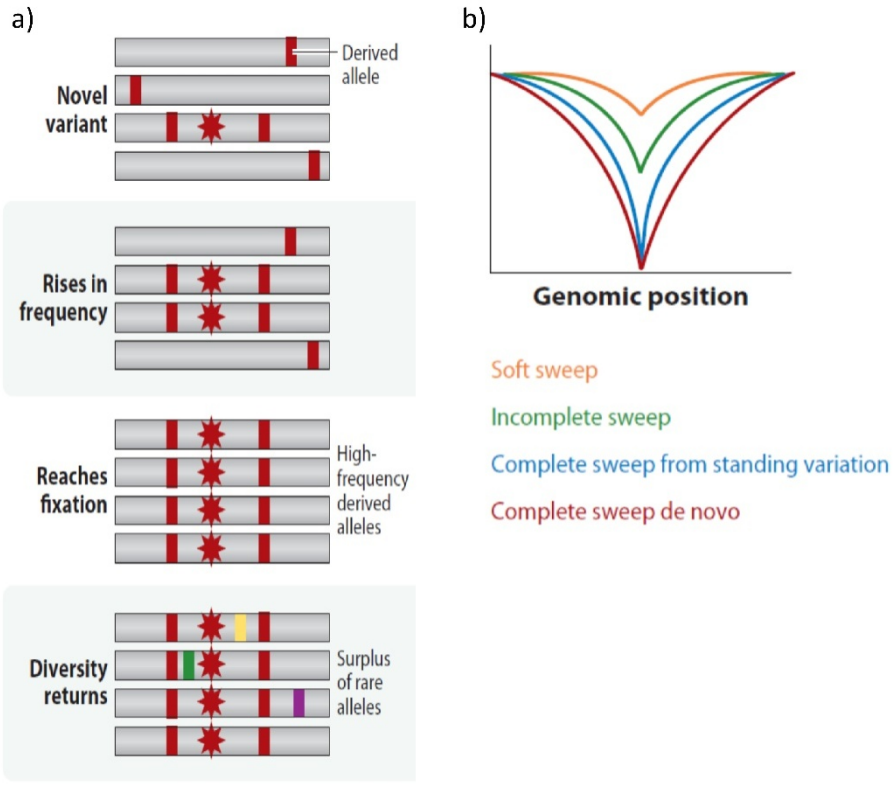


Figure 1.8 Réduction de diversité par le balayage sélectif: a) Processus associé au balayage sélectif complet b) Impact sur le niveau de diversité génétique du locus ciblé par la sélection positive en fonction du type de balayage sélectif. Le graphique représente la diversité nucléotidique en fonction de la distance au locus sélectionné. Source : Vitti et al., 2013 (à partir de la Figure 2)

De nombreuses méthodes sont utilisées pour détecter les signatures de sélection naturelle positive récente (Vitti et al., 2013).On distingue trois grandes familles de méthodes : les méthodes basées sur le spectre de fréquences alléliques (par exemple le test D de Tajima (Tajima, 1989) et le test H de Fay & Wu (Fay & Wu, 2000)) ; les méthodes basées sur le déséquilibre de liaison (DL) entre les loci et la longueur des haplotypes (test LRH pour *Long Range Haplotype* (Sabeti et al., 2002) et tous les tests dérivés) et les méthodes basées sur la différenciation génétique des populations. Ces dernières méthodes utilisent le fait que la sélection sur un allèle est particulière à un environnement donné, et que les populations évoluant dans des environnements distincts vont être soumises à différentes

pressions de sélection. Ainsi la comparaison des fréquences alléliques entre des populations peut permettre d'identifier des variants qui sont la cible de la sélection naturelle. La méthode la plus largement utilisée pour estimer la différenciation génétique de populations est l'indice Fst qui compare les variances des fréquences alléliques au sein et entre les populations (Wright, 1946).

Chaque famille de méthodes possède ses spécificités et permet de détecter avec plus ou moins de puissance les différentes formes de sélection naturelle. Par exemple, les méthodes basées sur le DL sont plus puissantes que celles basées sur le spectre de fréquences alléliques pour identifier les balayages sélectifs récents ou en cours (<30 000 ans), les balayages sélectifs partiels présentant des haplotypes longs mais une plus faible réduction de la diversité génétique. A l'inverse, les méthodes basées sur le spectre de fréquences alléliques sont particulièrement adaptées pour identifier des évènements plus anciens (30 000-50 000 ans), les distorsions de fréquences alléliques (tels que les fréquences élevées d'allèles dérivés) au locus persistant plus longtemps au cours des générations que les haplotypes longs (Grossman et al., 2013; Vitti et al., 2013).

Pour déterminer s'il existe des éléments en faveur de l'effet de la sélection à un locus, les différentes méthodes utilisent des tests de neutralité. Elles produisent des résumés statistiques qui sont comparés aux valeurs attendues sous l'hypothèse nulle de neutralité sélective. Dans le cas d'une recherche de signature de sélection naturelle dans un gène ou une région candidat(e), la distribution de la statistique sous l'hypothèse nulle est obtenue le plus souvent à partir de simulations basées sur la théorie de la coalescence (suivant différents scénarios démographiques). Elle peut être obtenue également de façon empirique en calculant la valeur de la statistique sur un ensemble de loci neutres indépendants génotypés sur les mêmes individus.

1.4.3 Des exemples de sélection balancée ou positive exercées par le paludisme

Parmi les gènes ou variants génétiques conférant une résistance accrue au paludisme lié à *P. falciparum,* certains sont considérés comme des cas d'école pour illustrer l'action de la sélection naturelle sur le génome humain : *HBB, G6PD, CD40LG* et *CD36*.

Hémoglobine S

Comme nous l'avons vu précédemment, A. C. Allison a montré que la sélection naturelle a induit une augmentation de la fréquence de l'allèle HbS responsable de la drépanocytose (Allison, 1954). L'allèle HbS est l'exemple le plus souvent cité pour illustrer la sélection balancée. A l'état homozygote, cet allèle est responsable de la drépanocytose, qui a des conséquences physiologiques très importantes, et qui était, jusqu'à il n'y a pas très longtemps encore, associée à une espérance de vie limitée (moins de 5 ans). Cet allèle aurait dû être purgé du fait de ses conséquences fortement délétères, mais l'avantage qu'il confère aux individus hétérozygotes face au paludisme l'a maintenu à des fréquences élevées, principalement en Afrique. La figure 4 montre l'avantage au niveau de la survie que confère le fait d'être hétérozygote par rapport aux homozygotes dans une cohorte de nouveau-nés au Kenya (Figure 4a) et la superposition de la distribution géographique de l'allèle HbS et du paludisme en Afrique (Figure 4b). Ce mode de sélection explique également la persistance d'autres anomalies du globule rouge qui confèrent une résistance au paludisme (Thomas N. Williams, 2006).

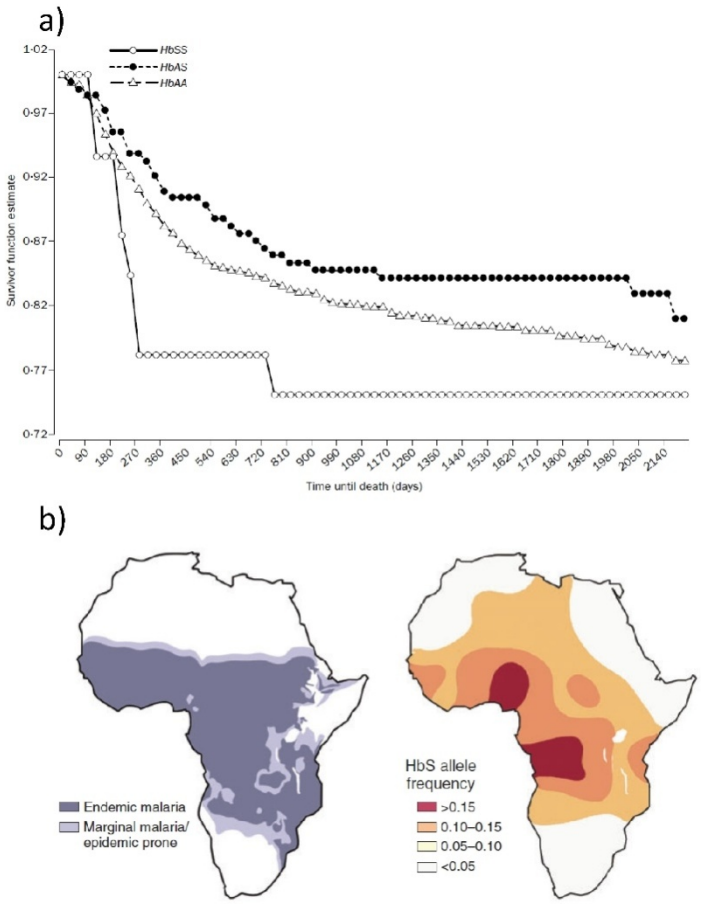


Figure 1.9 Illustration de la sélection balancée pour l'allèle HbS *: a) Courbe de survie pour 1022 enfants d'une cohorte de nouveaux-nés au Kenya – source : Aidoo et al., The Lancet, 2002*. Cette courbe illustre la meilleure survie des individus hétérozygotes HbAS comparés aux homozygotes HbAA et HbSS. *; b) Cartes montrant la superposition des distributions géographiques des zones d'endémie du paludisme (à gauche) et de la fréquence allélique de l'allèle HbS (à droite), suggérant un lien entre la pression de sélection exercée par le paludisme et la distribution de l'allèle HbS en Afrique. – source : Wellems and Fairhurst, 2005*

De nombreuses études ont porté également sur l'histoire évolutive de l'allèle HbS et nous renseignent sur l'origine du paludisme et l'histoire de l'exposition des populations humaines à cette infection (Laval et al., 2019). Du fait de la répartition géographique actuelle de HbS, une des hypothèses qui prévalait jusqu'à présent était que le paludisme s'était développé avec l'expansion de l'agriculture, il y a environ 10 000 ans (Kwiatkowski, 2005). Deux modèles évolutifs possibles étaient proposés pour l'origine de l'allèle HbS : une origine multiple, avec plusieurs événements de mutation survenant de façon concomitante à l'émergence de l'agriculture il y a environ 3000-5000 ans, ou une origine unique plus ancienne. Les études les plus récentes, basées sur des données de séquençage et

l'analyse des haplotypes portant l'allèle HbS, indiquent une origine unique de la mutation (rs334) (Laval et al., 2019; Shriner & Rotimi, 2018). Alors que Shriner *et al.* estiment son apparition il y a environ 7 300 ans en Afrique sous un modèle d'évolution neutre, l'étude de Laval *et al.* conclut à une origine beaucoup plus ancienne, il y a environ 22 000 ans, sous un modèle de sélection balancée, et à une pression de sélection conséquente antérieure à l'émergence de l'agriculture en Afrique. Ces résultats sont concordants avec la date estimée d'apparition du paludisme en Afrique subsaharienne à la fin du Pléistocène, par les études génétiques sur le parasite qui suggèrent une émergence *de P. falciparum* il y a environ 40 000 à 60 000 ans (Otto et al., 2018).

G6PD

G6PD est également un exemple classique de sélection naturelle récente liée au paludisme. De nombreux variants peuvent induire un déficit en G6PD, dont le variant *G6PD*-202A *(rs1050828)* commun en Afrique subsaharienne (prévalence d'environ 15 à 20%) et qui a été associé à une protection contre le paludisme. Sa prévalence dans les zones endémiques, comme pour le variant HbS, suggère une influence de la pression de sélection exercée par le paludisme. Sabeti *et al.* (Sabeti et al., 2002) montrent, à l'aide du test LRH basé sur la longueur des haplotypes transmis sans recombinaison, que l'haplotype portant la mutation présente un score EHH inhabituellement élevé. Le signal de sélection compatible avec une pression de sélection positive a été identifié comme significatif dans les trois populations africaines étudiées. Cependant le mode de sélection agissant sur le variant est toujours controversé (sélection positive ou sélection balancée), l'étude de la sélection naturelle à ce locus étant rendu plus complexe du fait de sa localisation sur le chromosome X.

CD40LG

CD40LG code pour une molécule présente à la surface des lymphocytes T, impliquée dans la réponse immunitaire. Il a été associé de manière répétée aux formes graves de paludisme. Comme le gène *G6PD*, il est localisé sur le chromosome X, et Sabeti et al. (Sabeti et al., 2002) ont montré que l'haplotype portant une mutation commune dans le promoteur, associée à la résistance, présentait

un score EHH particulièrement élevé, ce signal de sélection étant trouvé également significatif dans les trois populations africaines étudiées.

CD36

CD36 code pour un récepteur impliqué dans le phénomène de cytoadhérence des globules rouges infectés à l'endothélium vasculaire. Il a été associé de manière répétée aux formes graves de paludisme et la prévalence de la déficience de ce récepteur dans les populations d'Afrique et d'Asie de l'Est suggère que les variants responsables sont la cible de la sélection naturelle. Le variant le plus commun en Afrique est une mutation non-sens dans l'exon 10 (rs3211938). L'analyse des données HapMap[1] à ce locus dans la population Yoruba (Nigéria) a montré un EHH inhabituellement long pour l'haplotype portant la mutation (International HapMap Consortium, 2005). Fry et al. (Fry et al., 2009) ont ensuite évalué la distribution géographique globale de ce variant en utilisant les échantillons du panel HGDP-CEPH[2] et des échantillons de 15 groupes ethniques additionnels en Afrique (provenant de la Gambie, du Malawi, du Kenya et du Ghana), et ont testé la présence d'un signal de sélection naturelle avec la même approche basée sur l'EHH, dans deux populations en particulier, les Yoruba et les Gambiens. Ce polymorphisme est présent dans les populations d'Afrique subsahariennes où le paludisme est endémique, mais absent dans les populations du Sud de l'Afrique (Bantu d'Afrique du Sud et San de Namibie) ou dans les populations non-africaines pour lesquelles il n'y a pas eu de mélange récent avec des populations africaines. Cependant la fréquence élevée de l'allèle G associé à la résistance, observée dans la population Yoruba (26%) n'est pas retrouvée dans la plupart des autres populations d'Afrique subsaharienne (2% en Gambie et 9% près de la côte au Ghana, deux régions soumises également à une forte pression de sélection due au paludisme). Fry et al. confirment le signal de sélection positive au niveau du SNP rs3211938 dans la population Yoruba

[1] HapMap : projet international dont l'objectif était de développer une carte d'haplotypes du génome humain. Dans la phase initiale du projet (2002-2005), 270 individus de 4 populations ont été génotypés.

[2] HGDP-CEPH : Human Genome Diversity Project-Centre d'Etude du Polymorphisme Humain. Le panel HGDP-CEPH est une banque d'ADN contenant des échantillons de 1050 individus de 52 populations dans le monde, collectés par les différents laboratoires participant au projet.

mais ne mettent pas en évidence de signal significatif dans la population gambienne. Selon les auteurs, l'absence de résultat significatif dans cette dernière population est liée au fait que le test LRH présente une puissance limitée pour les allèles de faible fréquence car ces derniers sont souvent entourés par un haplotype long. Par ailleurs, plusieurs hypothèses sont avancées pour expliquer les plus faibles fréquences dans les populations en dehors du Nigéria, parmi lesquelles : une origine récente de l'allèle dans cette région, diffusé ultérieurement par migration aux autres populations; un balayage sélectif en cours dans les populations présentant une faible fréquence de l'allèle sélectionné mais à un stade moins avancé que dans la population nigériane; un événement local de sélection au Nigéria, qui pourrait être expliqué, par exemple, par l'effet d'une lignée spécifique de *P. falciparum*. Enfin, un fort signal de sélection est rapporté dans le gène *CD36* dans l'article décrivant le *1000 Genome Selection Browser* (Pybus et al., 2014), qui permet d'explorer et de visualiser les résultats d'un large panel de tests de neutralité sélective réalisés sur les données de séquençage du Projet 1000 Génomes [3](Phase 1 ; publication en avril 2012). Pour *CD36* le score maximum pour le test iHS (test basé également sur la statistique de l'EHH), est particulièrement élevé dans la population Yoruba (identique à celle du projet Hapmap) et fait partie des 1% des scores iHS les plus extrêmes observés dans la population.

1.4.4 Approches génome entier

Comme pour les études d'association, l'émergence des outils de génotypage et de séquençage à haut débit a conduit au développement de nouvelles méthodes statistiques et d'outils permettant la détection des signatures de sélection naturelle à l'échelle d'un génome. Chez l'Homme, la mise à disposition des données des Projet Hapmap puis 1000 Génomes a donné lieu à de nombreuses études génome entier qui ont permis d'identifier des centaines de gènes et de régions potentiellement soumis à la sélection naturelle. La plupart de ces études se sont concentrées sur

[3] Projet 1000 Génomes : projet international dont l'objectif initial était de séquencer au moins 1000 personnes dans le monde afin de produire un catalogue le plus détaillé possible des variations génétiques chez l'Homme. La phase 1, rendue publique en 2012, concerne les 1092 premiers échantillons de 14 populations. Dans la dernière phase du projet (phase 3), plusieurs populations (principalement d'Afrique et d'Asie du Sud) ont été ajoutées, portant le nombre total de populations à 26 et d'individus à 2504.

l'identification des signatures de sélection naturelle positive récente (depuis la migration hors d'Afrique il y a 50-60 000 ans). Bien que les différents types d'approches décrites à la section 1.4.2 ont été adaptées pour une analyse à l'échelle du génome entier, celles basées sur le DL ont été privilégiées pour l'étude du génome humain. Elles sont puissantes pour identifier les signatures de sélection les plus récentes et sont particulièrement utiles pour détecter les variants qui font l'objet d'un balayage sélectif partiel. En effet, depuis la sortie d'Afrique, on estime, en considérant un coefficient de sélection réaliste, que peu de nouvelles mutations ont eu le temps d'atteindre la fixation et que la plupart des balayages sélectifs depuis cette période sont des balayages partiels (Jonathan K. Pritchard et al., 2010). Ces méthodes présentent également l'avantage de permettre la détection des signatures de sélection balancée récente, qui sont comparables à celles laissées par un balayage partiel. Dans cette section sont décrites les principales méthodes basées sur le DL, les spécificités de l'approche génome entier pour la détection de signaux de sélection.

Les tests basés sur le déséquilibre de liaison

Les premières analyses génome entier ont porté sur la détection des signatures de sélection positive récente de type *hard sweep* qui sont les plus visibles et les plus faciles à détecter. Les méthodes les plus connues s'appuient sur la statistique EHH (pour *Extended Haplotype Homozygosity*) proposée par Sabeti *et al.* (Sabeti et al., 2002). L'EHH consiste à comparer le profil d'homozygotie de part et d'autre du variant d'intérêt pour les haplotypes portant ou non l'allèle soumis à la sélection. Pour un allèle donné, l'EHH à une distance *x* du SNP est définie comme la probabilité que deux haplotypes choisis au hasard parmi ceux qui portent cet allèle, soient homozygotes à tous les SNPs sur une distance *x* à partir du SNP d'intérêt. L'EHH permet ainsi de détecter la transmission d'haplotypes longs que la recombinaison n'a pas eu le temps de casser. Le test LRT (pour *Likelihood Ratio Test*), proposé initialement par Sabeti et al., 2002, dans le contexte des analyses gènes candidats, combine l'EHH relatif à la fréquence de l'allèle considéré, cherchant ainsi à détecter des haplotypes particulièrement longs pour leur fréquence allélique, suggérant une augmentation en fréquence rapide due à l'action de la sélection naturelle. Le test iHS (pour *integrated Haplotype Score*) proposé

par Voight *et al.* (Voight et al., 2006) est une extension de l'EHH permettant son utilisation pour des analyses génome entier. Il compare l'aire sous la courbe définie par l'EHH pour l'allèle ancestral et l'allèle dérivé.

$$iHS = ln\left(\frac{iHH_A}{iHH_D}\right)$$

avec :

- iHH : l'intégrale de la fonction EHH autour du SNP d'intérêt
- iHH_A et iHH_D: l'iHH pour les allèles ancestral et dérivé respectivement

Cette approche est illustrée, ci-dessous, par un exemple extrait de l'article de Voight *et al.* (Figure 1.9) pour un SNP ayant un score iHS particulièrement élevé dans les populations d'Asie de l'Est. Au centre sont représentés les haplotypes de l'échantillon, regroupés en fonction du statut ancestral de l'allèle (groupe d'haplotypes du haut, indiqué par le trait vertical bleu) ou dérivé (groupe d'haplotypes du bas, indiqué par le trait rouge). Les couleurs bleu et rouge des traits horizontaux indiquent les régions où les haplotypes sont homozygotes. A droite figure le graphique de l'EHH en fonction de la distance au SNP d'intérêt pour les deux allèles et à gauche les valeurs de la statistique iHS pour un ensemble de SNPs de cette région chromosomique.

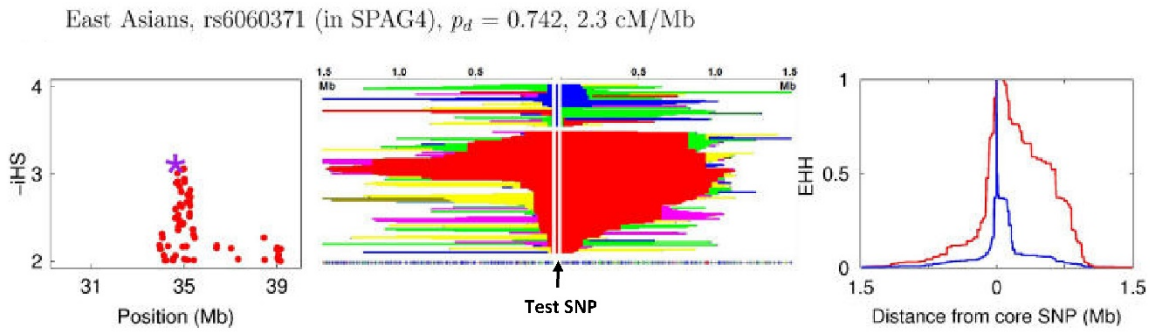


Figure 1.10 Exemple de signal de sélection détecté par le test iHS.
*Au centre figure la représentation des haplotypes, regroupés en fonction de l'allèle ancestral (en haut) et de l'allèle dérivé (en bas). En bleu et rouge sont représentées les portions d'haplotypes identiques par descendance pour les deux allèles. A droite l'*Extended Haplotype Homozygosity calculé en fonction de la distance au SNP d'intérêt (core SNP). A gauche, le score iHS calculé pour le SNP (étoile violette) et pour les autres SNPs de la région (points rouges). Source : Voight *et al.*, 2006

Ce test présente une bonne puissance pour détecter les balayages sélectifs incomplets, quand le variant sélectionné a atteint une fréquence intermédiaire dans la population (50-80%), mais est peu puissant lorsque celui-ci atteint une fréquence élevée proche de la fixation (>80%) (Pickrell et al., 2009). Une autre extension de l'EHH, l'XP-EHH (pour *Cross-Population Extended Haplotype Homozygosity*) proposée par Sabeti et al. (Sabeti et al., 2007) est complémentaire de cette première méthode car elle est particulièrement puissante pour identifier les balayages sélectifs où l'allèle sélectionné est proche de la fixation. Cette méthode utilise la statistique iHH pour comparer l'homozygotie des haplotypes entre deux populations : la population testée et une population prise comme référence. Le score XP-EHH correspond au logarithme du rapport de l'iHH entre les deux populations, l'iHH étant calculé sur l'ensemble des haplotypes sans distinction de l'allèle ancestral et de l'allèle dérivé. D'autres méthodes basées sur le DL ont été développées en parallèle de celles utilisant l'EHH. Ces approches utilisent un calcul de l'homozygotie directement à partir des données génotypiques à chaque SNP (LDD test , E. T. Wang et al., 2006), ce qui présente l'avantage de ne pas nécessiter d'étape de reconstruction des haplotypes, ou procèdent par identification de régions IBD (pour *Identity by descent*) (Albrechtsen et al., 2010). L'objectif de cette dernière approche est similaire à celui des méthodes basées sur l'EHH : rechercher des longues portions d'ADN partagées entre les individus, mais elle utilise une méthode de calcul différente.

Au début des années 2010, plusieurs méthodes composites ont été également développées (Vitti et al., 2013) comme le test CMS (Grossman et al., 2010). Ces méthodes combinent à un même SNP différents tests indépendants pouvant apporter des informations complémentaires (par exemple l'iHS, l'XP-EHH, le F_{ST} et deux autres tests basés sur le spectre de fréquences alléliques pour le CMS). Ces méthodes permettent de réduire le taux de faux positifs et/ou d'obtenir une meilleure résolution spatiale pour identifier le variant causal.

Plus récemment, de nouvelles méthodes, toujours basées sur l'EHH, ont été proposées pour permettre la détection des *soft sweeps* où la sélection agit sur un allèle déjà présent dans la

population ou sur plusieurs allèles à un même locus. Ces signatures de sélection sont difficilement détectables par les méthodes standards telles que l'iHS, car la sélection va entrainer une augmentation en fréquence de plusieurs haplotypes et non d'un seul. Le test nS_L (Ferrer-Admetlla et al., 2014) est comparable au test iHS dans la mesure où il utilise le rapport entre les mesures d'homozygotie des haplotypes dans les groupes d'haplotypes portant l'allèle ancestral et l'allèle dérivé. Cependant l'homozygotie et la longueur des haplotypes sont définies de manière différente. Les haplotypes sont comparés deux à deux et non globalement dans chaque groupe ; l'homozygotie est calculée en réalisant la moyenne de la longueur des haplotypes identiques par état (IBS pour *Identical by state*, c'est-à-dire homozygotes aux différents sites polymorphes). La longueur des haplotypes est mesurée par le nombre de sites polymorphes présents dans l'ensemble de l'échantillon et non par la distance génétique, ce qui le rend plus robuste aux variations du taux de recombinaison dans le génome que l'iHS. En 2016, Garud et al. (Garud et al., 2015) ont proposé une seconde approche qui regroupe les deux haplotypes les plus fréquents dans le calcul de l'EHH (test H12). Cette approche a été ensuite adaptée et intégrée dans le cadre de l'EHH par Torres et al. (iHH12, (Torres et al., 2019). Ces deux approches sont à la fois puissantes pour détecter les *soft sweep* et les *hard sweep*.

Enfin, Field et al. (Field et al., 2016) ont introduit la méthode SDS adaptée aux données de séquençage pour mettre en évidence des signatures de sélection positive très récente (<2000 ans). Cette méthode utilise la distribution des mutations singletons (existant en un seul exemplaire dans l'échantillon) autour des allèles ancestral et dérivé au locus d'intérêt pour identifier les signatures de sélection.

Principes des approches génome entier

Comme pour les études d'association, les tests de sélection positive sont réalisés à chaque SNP du génome. L'analyse est généralement limitée aux SNPs fréquents (fréquence de l'allèle mineur MAF>0.05 ou plus rarement MAF> 0.01). La plupart des méthodes basées sur le DL requièrent des données phasées, c'est-à-dire que les haplotypes pour chacun des deux chromosomes soient

reconstruits avant l'analyse. Ces tests nécessitent également une carte de distance génétique (basée sur le taux de recombinaison) ou de distance physique (dans de rares cas, comme par exemple pour le test nS_L).

Les principales difficultés pour la mise en évidence de signatures de sélection naturelle (que ce soit au niveau de gènes candidats ou dans les approches génome entier) sont dues au fait que des facteurs autres que la sélection peuvent être également responsables de déviations par rapport au modèle neutraliste. Des événements liés à l'histoire démographique des populations tels que les goulots d'étranglement et les expansions majeures peuvent laisser des signatures qui ressemblent à celles de la sélection positive. Afin de corriger ce facteur de confusion, les études génome entier utilisent, dans la majorité des cas, une approche empirique consistant à comparer la valeur de la statistique de test calculée pour le locus d'intérêt à la distribution de la statistique générée sur l'ensemble du génome. En effet, les événements liés à l'histoire démographique affectent l'ensemble du génome alors que la sélection agit de manière spécifique au niveau d'un locus. Ainsi, cette approche empirique peut être utilisée pour faire la différence entre les effets des facteurs démographiques et les événements de sélection naturelle ciblant des régions génomiques précises. Aussi, dans ces études, les résultats obtenus pour chaque SNP sont standardisés sur l'ensemble du génome (en soustrayant la moyenne des statistiques sur l'ensemble du génome et en divisant par l'écart-type), puis les SNPs présentant des valeurs extrêmes de la statistique (par exemple inférieures au 1er ou supérieure au 99ème centile) sont identifiés comme des cibles potentielles de la sélection. Pour les tests portant sur une seule population (iHS, nS_L, ect), les statistiques de test étant sensibles à la fréquence allélique, la standardisation se fait par bin de fréquence, c'est-à-dire par sous-groupe de SNPs défini par leur fréquence allélique. Une autre difficulté pour la détection des cibles de la sélection naturelle est liée à la variation du taux de recombinaison selon les régions du génome. Ferrer-Admetlla *et al.* (Ferrer-Admetlla et al., 2014) ont montré que les méthodes standards utilisées pour détecter les *hard sweeps* (dont les tests basés sur l'*EHH*) étaient sensibles au taux de recombinaison dans la région. Ainsi, en utilisant ces tests, les signatures de sélection fortes sont

détectées de manière plus fréquente dans les régions présentant un faible taux de recombinaison. Le test nS_L qu'ils proposent est supposé s'affranchir de ce second problème en utilisant le nombre de sites polymorphes pour mesurer les longueurs d'haplotypes.

Enfin, afin d'identifier les régions (et non plus les variants) présentant un signal de sélection significatif, certaines études ajoutent une étape où les statistiques sont examinées à l'intérieur de fenêtres non chevauchantes. La motivation première pour l'analyse par fenêtre non chevauchante était de faciliter la comparaison des signaux de sélection entre les populations (Pickrell et al., 2009; Voight et al., 2006). Elle permet également d'améliorer la puissance de détection des balayages sélectifs. Il est apparu, notamment pour l'iHs, qu'il était plus fiable de considérer les régions présentant plusieurs SNPs avec une valeur élevée plutôt que celles n'en présentant qu'un seul, une valeur élevée isolée pouvant en effet plus probablement résulter du hasard. Les fenêtres sont définies en fonction d'un nombre de SNPs ou d'un nombre de paires de bases (la taille la plus communément utilisée actuellement est 100Kb). En ce qui concerne l'iHS, il existe un certain consensus dans la littérature pour utiliser la proportion de SNPs présentant une valeur dépassant un certain seuil, par exemple le pourcentage de SNPs avec une valeur extrême. Cette approche est moins courante pour l'XP-EHH où la proportion de valeurs extrêmes de la statistique ou la valeur maximale de la statistique sont plutôt utilisées. Les approches plus récentes pour la détection des *soft sweeps* ont été encore peu utilisées avec une approche par fenêtre.

1.5 Problématique et objectifs

Comme nous l'avons vu précédemment, les efforts de recherche pour l'identification des gènes impliqués dans le paludisme se sont essentiellement focalisés sur les accès graves bien que les études sur les phénotypes non graves présentent un intérêt certain pour la compréhension de la physiopathologie et des mécanismes de résistance naturelle à la maladie. L'objectif de cette thèse était d'étendre l'approche d'association génome entier aux formes simples du paludisme, au travers de l'étude de deux cohortes de nouveau-nés au Sud Bénin (au total 800 enfants), suivis pendant 18-24 mois par l'UMR261 (MERIT IRD/Université de Paris).

Dans une première partie nous présentons la stratégie d'analyse et les résultats de la GWAS sur les formes simples de paludisme dans ces deux cohortes. Cette étude est assez éloignée méthodologiquement des GWAS classiques réalisées sur des grands échantillons de cas et de témoins: les études sur les formes simples de paludisme impliquent en effet un suivi longitudinal des individus en population pour évaluer la susceptibilité des individus aux infections palustres, et la diversité génétique au sein des populations africaines nécessite de prendre en compte la structure de population qui est un facteur de confusion potentiel. Dans cette étude, l'association a été testée avec deux phénotypes, la récurrence des accès palustres simples et la récurrence des infections dans leur ensemble (incluant les accès palustres et les infections asymptomatiques). La récurrence des évènements a été analysée avec un modèle mixte de Cox, permettant à la fois de tenir compte de l'ensemble des infections détectées et d'ajuster sur le risque environnemental qui varie au cours du temps. Ce modèle n'étant pas applicable sur l'ensemble du génome, nous avons opté pour une stratégie d'analyse en deux étapes, avec dans un premier temps l'utilisation d'un modèle mixte de Cox pour la définition d'un phénotype ajusté sur les covariables (facteurs environnementaux et facteurs non génétiques de l'hôte) et dans un second temps, le test de l'association avec l'ensemble des polymorphismes avec un modèle mixte linéaire. L'étude d'association a révélé plusieurs signaux d'association forts et a été poursuivie par une étude fonctionnelle *in silico* avec la plateforme FUMA,

afin d'identifier de potentiels variants fonctionnels au niveau des régions présentant les signaux d'association.

Nous nous sommes intéressés ensuite plus généralement aux modèles mixtes non-linéaires. Dans le cas d'un trait binaire, il serait particulièrement attractif d'utiliser un modèle mixte logistique pour prendre en compte la structure de population. Cependant, comme pour le modèle mixte de Cox, l'estimation de la vraisemblance de ces modèles, par un processus itératif, est complexe et couteuse en temps de calcul, ce qui limite leur utilisation. La solution largement utilisée jusqu'à présent pour les GWAS sur des échantillons de cas et de témoins était d'analyser le statut cas/témoin, codés respectivement 1 et 0, comme un trait quantitatif avec un modèle mixte linéaire. Cependant, Chen *et al.* (Chen et al., 2016) ont montré en 2016 que cette méthode était inappropriée dans certaines situations et ont proposé un test du score pour la régression logistique mixte, applicable à l'ensemble du génome. La contrepartie de ce test est qu'il ne permet pas d'obtenir les effets des SNPs. Dans cette deuxième partie, deux méthodes sont proposées pour estimer l'effet des polymorphismes avec le modèle mixte logistique. Ces méthodes sont évaluées à l'aide de deux jeux de simulations, l'un basé sur les données de la GWAS, l'autre sur des données génétiques simulées à l'aide d'un modèle de coalescence. Les simulations montrent que les effets des variants sont de manière générale bien estimés. Nous avons ensuite évalué la capacité de différentes méthodes (régression logistique, régression linaire mixte, régression logistique mixte utilisant le test du score et les deux nouvelles méthodes proposées) à corriger pour la structure de population dans le cas de la GWAS. Nous proposons également une extension du quantile-quantile plot (QQplot) stratifié de Chen *et al.* permettant de diagnostiquer une correction incomplète de la structure de population quand la structure de population n'est pas clairement identifiée au départ.

La dernière partie est consacrée à l'étude des signatures de sélection naturelle par une approche génome entier dans ces mêmes cohortes. Le paludisme ayant constitué une des plus fortes pressions de sélection que l'Homme ait connue dans son histoire récente, nous explorons ici la possibilité d'exploiter l'information de sélection naturelle pour augmenter la puissance de l'analyse et améliorer

la détection des signaux d'association. Une étude sur la résistance au paludisme grave (Ayodo et al., 2007), portant sur 10 gènes candidats dont *HBB,* a fourni la preuve de concept que combiner les résultats des tests de sélection naturelle avec ceux des tests d'association permettait d'augmenter la puissance de l'analyse d'association d'un ordre de magnitude de 1 à 2. Nous proposons d'appliquer cette approche dans le cadre d'une GWAS. Différentes méthodes basées sur l'EHH ont été utilisées pour identifier les signatures de sélection positive ou balancée récente dans la population d'étude : les tests standards iHS et XP-EHH et le test nS_L permettant d'identifier les *soft sweeps*. Les premières analyses croisant les informations de sélection naturelle et d'association mettent en évidence plusieurs régions chromosomiques d'intérêt potentiel où les signaux d'association et de sélection co-localisent ; cependant ces analyses montrent également la difficulté à mettre en évidence les signaux de sélection liés au paludisme avec les outils disponibles.

2 LES DONNEES

2.1 Les suivis de cohortes

Deux cohortes de nouveau-nés ont été mises en place successivement par l'UMR MERIT dans le sud du Bénin. Cette région est caractérisée par un climat subtropical, avec deux saisons des pluies (une saison des pluies longue d'avril à juillet et une seconde, plus courte, entre août et septembre). Dans le district de Tori-Bossito, une étude épidémiologique et entomologique a montré que les accès cliniques étaient essentiellement dus à *P. falciparum* (97%) et que l'incidence des accès palustres était de 1,5 (IC à 95% : [1,2-1,9]) par an et par enfant, pour les enfants de moins de 5 ans (Damien et al., 2010).

Ces deux cohortes ont été suivies dans le cadre de deux projets multidisciplinaires, dirigés par le Dr André Garcia et financés par l'ANR (2007 et 2010). Ils ont permis de recueillir, en dehors des données parasitologiques et cliniques, des informations sur de nombreux facteurs de risque potentiels : à la fois des facteurs individuels (par exemple le niveau socio-économique, l'utilisation de moustiquaires, la présence d'une infection placentaire) et des facteurs environnementaux.

2.1.1 La cohorte de Tori-Bossito

Les enfants de la cohorte de Tori-Bossito ont fait partie d'une étude sur les premières infections palustres (Le Port et al., 2012) qui s'est déroulée de juin 2007 à janvier 2010 dans 9 villages du district de Tori-bossito situés à 40 km au nord-est de Cotonou (Figure 2.1). Au total, 656 nouveau-nés ont été inclus à la naissance et suivis jusqu'à l'âge de 18 mois. Les enfants ont été suivis de manière active pour le paludisme. Une visite à domicile par semaine par un infirmier du programme était réalisée avec une prise de température systématique. En cas de fièvre (température axillaire ≥ 37°5) ou d'historique de fièvre dans les 24h, un diagnostic de paludisme était réalisé au centre de santé avec un test de diagnostic rapide (TDR) et une goutte épaisse (GE). Une fois par mois, une GE systématique (indépendamment de tout signe clinique) était effectuée afin de détecter les infections asymptomatiques. A tout moment, en cas de fièvre ou de tout autre signe clinique, lié ou non au

paludisme, les mères étaient invitées à se rendre avec leur enfant au centre de santé où le même protocole (température, GE et TDR) était appliqué. Les enfants présentant une infection symptomatique (associée à une fièvre) étaient traités par une combinaison thérapeutique à base d'artémisinine (CTA) comme recommandé par le Programme national de lutte contre le paludisme (PNLP).

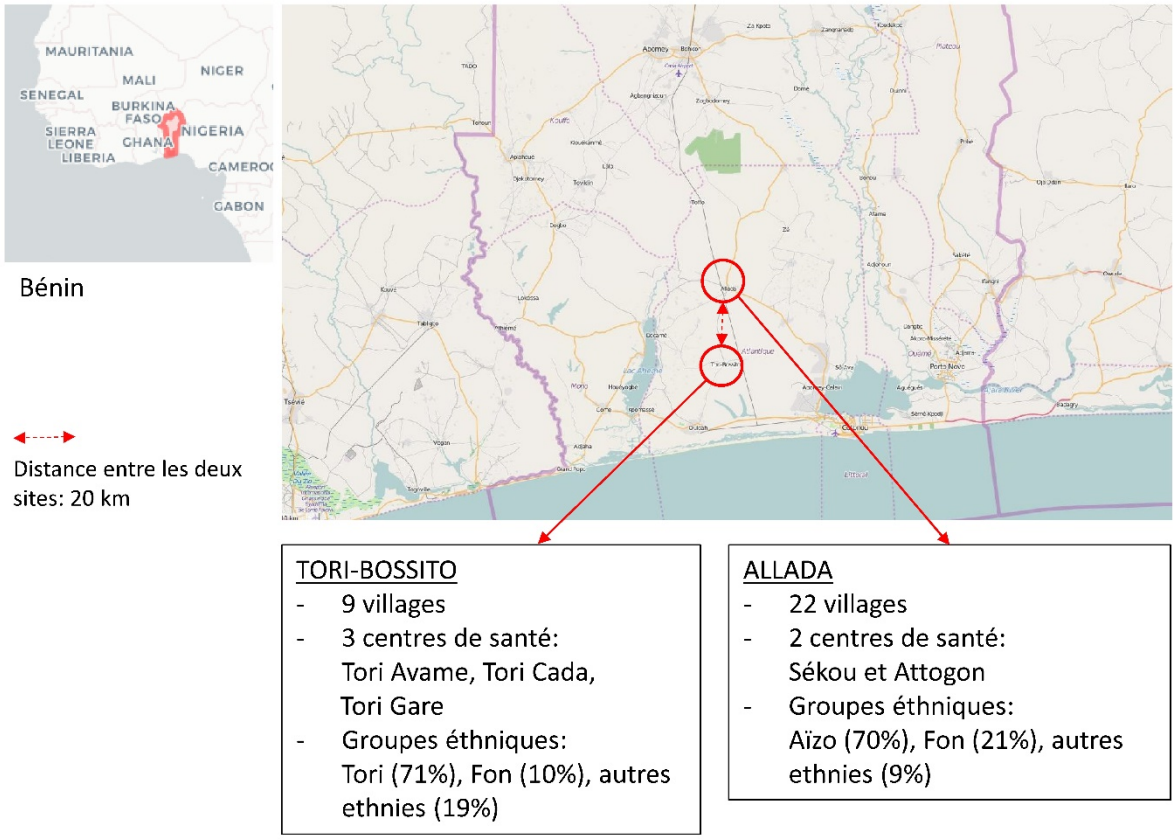


Figure 2.1 Les deux sites d'étude au sud du Bénin

2.1.2 La cohorte d'Allada

Les enfants de la cohorte d'Allada ont fait partie d'un suivi mère-enfant dans le district d'Allada, une zone semi-rurale située à 55 kilomètres au nord de Cotonou et à 20 km du premier site d'étude (d'Almeida et al., 2017). La cohorte est composée de 400 enfants qui ont été inclus à la naissance parmi les nouveau-nés de mères ayant participé à un essai clinique multicentrique pour la prévention du paludisme pendant la grossesse (essai MiPPAD, «Malaria in Pregnancy Preventive Alternative Drugs, NCT00811421, (González et al., 2014a)). La première année, ces enfants ont fait

l'objet d'un suivi passif pour le paludisme dans le cadre du projet MiPPAD (Accrombessi et al., 2015). Une visite systématique était programmée à 6, 9 et 12 mois. En dehors de ces visites, les mères étaient invitées à se rendre avec leur enfant au centre de santé quel que soit le problème de santé rencontré par l'enfant. Au centre de santé, en cas de fièvre (température axillaire ≥ 37°5) ou d'historique de fièvre dans les 24h, un diagnostic de paludisme était également réalisé avec un test de diagnostic rapide (TDR) et une goutte épaisse (GE). Parmi ces 400 enfants, 306 ont été suivis au-delà de la première année et ont fait l'objet d'un suivi parasitologique et clinique actif entre 12 et 24 mois, similaire à celui de la cohorte de Tori-Bossito.

Le risque spatio-temporel d'exposition au vecteur

Pour les deux cohortes, des données environnementales (informations sur les caractéristiques de la maison et sur son environnement immédiat), géographiques (images satellites, type de sol, cours d'eau à proximité, indice de végétation, précipitations, etc.) et entomologiques ont été recueillies durant le suivi. Ces données ont permis de modéliser, pour chaque enfant, un risque spatio-temporel d'exposition aux piqûres d'Anophèles (Cottrell et al., 2012). Un risque d'exposition a été estimé pour chaque enfant, une fois par mois (au moment de la visite mensuelle).

La même méthodologie statistique a été utilisée pour modéliser le risque d'exposition dans les deux cohortes mais à partir de mesures entomologiques différentes sur le terrain. Pour la cohorte de Tori-Bossito, des captures de moustiques sur Homme ont été effectuées toutes les six semaines, dans plusieurs points des villages, alors que pour la cohorte d'Allada, les captures ont été effectuées par des pièges lumineux (*CDC ligth trap*) dans la chambre des enfants, pour un sous-échantillon de la cohorte, deux nuits consécutives par mois.

2.2 Contrôle qualité des données

2.2.1 Les données du suivi palustre

Pour chaque cohorte, nous avons travaillé à partir d'un fichier regroupant l'ensemble des visites du suivi (les visites à domicile et les visites au centre de santé) et incluant pour chaque visite les

informations sur la date de la visite, les données cliniques (température, antécédent de fièvre au cours des dernières 24h), les résultats des tests diagnostiques (TDR et GE) ainsi que d'autres données relevées tout au long du suivi, telles que l'information concernant l'utilisation de la moustiquaire ou l'estimation du risque d'exposition.

Un protocole standard de contrôle qualité des données a été appliqué, avec notamment une vérification des valeurs extrêmes et de la cohérence entre les données (cohérence entre les dates, entre les deux tests de diagnostics (TDR et GE), entre les différentes lectures des lames de GE, etc.). Lorsque plusieurs visites avaient eu lieu le même jour pour un même enfant (par exemple une visite à domicile suivie, en cas de fièvre, d'une visite au centre de santé), ces données ont été agrégées afin d'obtenir une seule ligne de données par jour.

Les accès palustres et les infections asymptomatiques ont ensuite été codés de manière automatique avec un script R (R Core Team, 2017). Un accès palustre a été défini comme un test de diagnostic positif (TDR et/ou GE) associé à une fièvre (de température axillaire ≥ 37 ° 5) ou un antécédent de fièvre au cours des dernières 24h. Une infection asymptomatique a été définie comme une GE positive (GE systématique réalisée lors de la visite mensuelle) en absence de fièvre et d'antécédent de fièvre, et sans diagnostic d'accès palustre dans les trois jours suivants. Comme il s'agit ici d'une étude sur de très jeunes enfants (avant le développement d'une immunité protectrice), il n'a pas été appliqué de seuil de parasitémie pour définir un accès palustre. A la suite des accès palustres, tous les enfants ayant reçu un traitement antipaludique, nous avons considéré une période de 14 jours après le diagnostic pendant laquelle l'enfant n'est plus considéré comme à risque. Le diagramme ci-dessous (Figure 2.2) donne une illustration des données dont nous disposons. Il représente les données des accès palustres pour cinq individus. Tous les individus ont fait au moins un accès palustre (triangle rouge). L'individu 4, par exemple, a fait un accès palustre à 4, 9 et 11 mois. La discontinuité du trait après un accès palustre indique la période pendant laquelle l'individu n'est plus considéré comme à risque. Quatre des cinq enfants ont été suivis pendant la totalité des 18 mois

(indiqué par un cercle noir). L'individu 1 a quitté l'étude prématurément, entraînant une censure des données (rond gris).

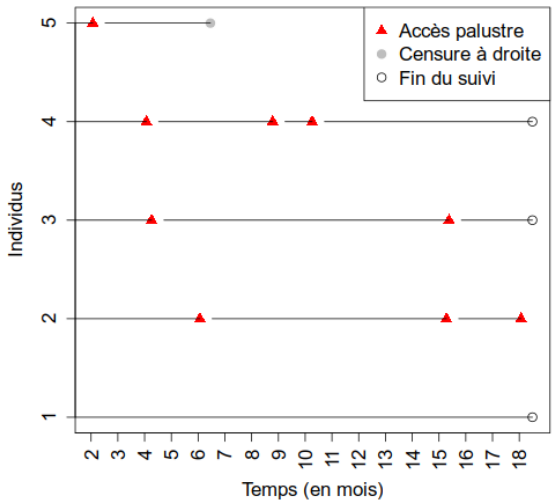


Figure 2.2 Diagramme des données de récurrence des accès palustres pour 5 individus de la cohorte de Tori-Bossito

2.2.2 Les données génétiques

Au total, 820 enfants (536 dans la cohorte de Tori-Bossito et 283 dans celle d'Allada) ont été génotypés avec la puce Illumina HumanOmni5-4v1 (4,2 millions de SNPs ou d'insertions/délétions de petite taille) par le Centre National de Recherche en Génomique Humaine (CNRGH, CEA, Evry, France). Nous avons effectué ensuite un contrôle qualité des données de génotypage en nous appuyant sur les étapes et critères définis pour les GWAS par Anderson et *al.* (Anderson et al., 2010). Les différentes étapes du contrôle qualité sont reprises dans le *flow chart* de la Figure 2.3. Le contrôle qualité des échantillons d'ADN a consisté :

- à comparer le sexe génotypique à celui reporté dans la base de données,

- à représenter l'hétérozygotie versus le taux de génotypage par individu (des valeurs basses par rapport à l'ensemble de l'échantillon pour ces deux variables étant révélatrices d'une mauvaise qualité de l'échantillon),

- et à examiner les relations de parenté entre les individus en estimant la matrice de corrélation génétique (ou *Genetic Relationship Matrix,* GRM) telle que définie dans l'encadré de la Figure 2.4.

Après exclusion d'un individu présentant des valeurs aberrantes d'hétérozygotie et de taux de génotypage, tous les autres échantillons avaient un taux de génotypage > 0,97 (moyenne = 0,998) et ont été conservés pour l'analyse. Une paire d'individus a été supprimée en raison d'un coefficient de

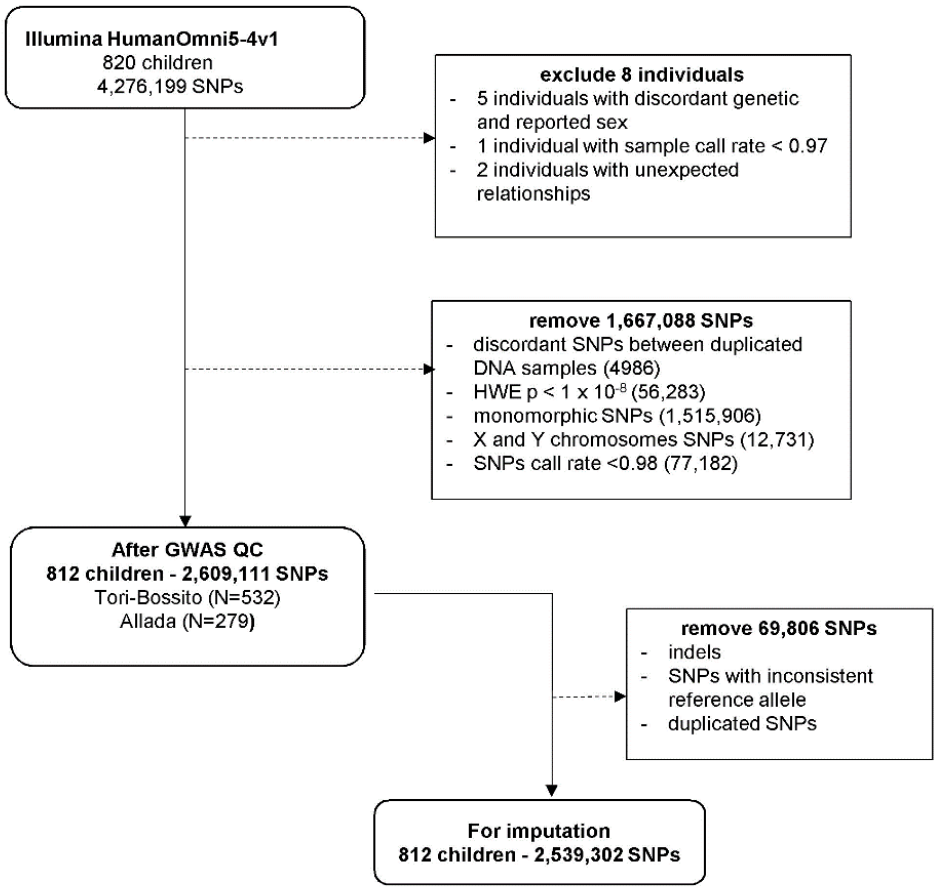


Figure 2.3 Flow chart du contrôle qualité des données génétiques

parenté étonnamment élevé (Φ = 0,27, correspondant à des frères et sœurs). La GRM a par ailleurs révélé un certain degré de parenté entre les enfants dans les deux cohortes. Ceci était attendu sachant que le recrutement des nouveau-nés a été réalisé dans des villages ruraux, c'est-à-dire dans des petites communautés endogames. Pour un certain nombre de paires, les coefficients de parenté Φ étaient compris entre 0,10 et 0,16, ce qui correspond à des niveaux de parenté pour des demi-

frères et sœurs, oncle-neveu voire des ¾ frères et sœurs (deux enfants ayant le même père et dont les mères sont sœurs par exemple). L'ensemble de ces individus ont été conservés pour les analyses d'association.

Le contrôle qualité des marqueurs génétiques a consisté à éliminer les variants présentant un écart très significatif à l'équilibre de Hardy Weinberg ($P < 10^{-8}$, test réalisé sur les individus non

Matrice de corrélation génétique (GRM)

Les données génotypiques sur l'ensemble du génome permettent d'estimer avec précision le degré de parenté entre les individus sans avoir recours à l'information des généalogies. La matrice GRM peut être estimée par la matrice de variance-covariance calculée à partir de l'information sur le nombre d'allèle mineur à chaque SNP (Yang et al.2011).

La GRM est une matrice de taille (n x n), *n* étant le nombre d'individus. Pour deux individus i et j, la covariance sur un ensemble de *m* SNPs, s'écrit :

$$GRM_{ij} = \frac{1}{m}\sum_{k=1}^{m}\frac{(G_{ik} - 2p_k)(G_{jk} - 2p_k)}{2p_k(1-p_k)}$$

avec *k* un SNP spécifique, G_{ik} = 0, 1 ou 2, le nombre d'allèles mineurs pour l'individu *i* au SNP *k* et p_k la fréquence de l'allèle mineur au même SNP. Alors $2p_k$ est l'espérance G_{ik} dans la population et $2p_k(1-p_k)$, sa variance.

La GRM est utilisée en génétique pour estimer le coefficient de parenté φ entre les individus ($GRM = 2\varphi$) et pour ajuster sur les relations de parenté dans les GWAS. Pour estimer φ, cette méthode nécessite que les SNPs soient indépendants. Une approche consiste à réduire le nombre initial de SNPs en supprimant ceux qui sont en DL afin d'obtenir un sous-ensemble de SNPs approximativement indépendants. Dans notre étude, la GRM a été calculée à partir des SNPs communs (MAF>0,05) et après filtre des SNPs en DL (r^2> 0,2)

Figure 2.4 Matrice de corrélation génétique (GRM)

apparentés) et ceux dont le taux de génotypage est inférieur à 0,98. A partir d'échantillons d'ADN dupliqués (21 paires d'échantillons), nous avons également identifié et éliminé des SNPs présentant une mauvaise reproductibilité (SNPs montrant au moins une discordance). Enfin, les variants monomorphes ainsi que les variants des chromosomes X et Y ont été exclus. En effet, pour ce travail de thèse, nous avons limité l'analyse aux chromosomes autosomaux, l'analyse des chromosomes X et Y demandant l'utilisation de méthodes spécifiques.

Après le contrôle qualité, les données de génotypage de 2 609 111 variants étaient disponibles pour 812 enfants. Une étape d'imputation a été réalisée par Pierre Luisi, post-doctorant au Centre National d'investigation scientifique et technique à Cordoba (Argentine) et qui a également initié l'analyse sur l'ensemble des infections (deuxième phénotype de la GWAS). Cette étape d'imputation permet de densifier la couverture du génome, en récupérant l'information pour un certain nombre de variants non présents sur la puce (et donc non génotypés) mais en déséquilibre de liaison avec des variants génotypés. Brièvement, l'imputation a été réalisée sur le serveur du Michigan (Michigan Imputation Server) avec l'algorithme minimac 3, après une étape de contrôle qualité supplémentaire (Figure 2.4). Avant l'imputation, les données génotypiques ont été préphasées avec le logiciel SHAPEIT v2, en utilisant comme panel de référence l'ensemble des haplotypes du projet 1000 Genomes v5. Les données imputées ont ensuite été filtrées pour ne conserver que les SNPs avec une très bonne qualité d'imputation ($R^2 > 0.8$).

Au final, le jeu de données complet (variants génotypés et imputés) incluait 15 566 900 variants avec une MAF supérieure à 0,01. Les analyses de ce manuscrit ont porté sur 800 enfants (525 dans la cohorte de Tori-Bossito et 275 dans la cohorte d'Allada), après exclusion *a posteriori* de 12 enfants avec une mauvaise qualité de suivi. L'ensemble du contrôle qualité a été réalisé avec le logiciel R (R Core Team, 2017). Le contrôle qualité des données génétiques a été réalisé avec le paquet gaston (Perdry & Dandine-Roulland, 2018).

2.3 Stratification de population

Nous avons examiné ensuite la stratification de population de nos deux échantillons, c'est-à-dire la présence éventuelle de sous-groupes génétiquement homogènes (et différenciés entre eux). Caractériser la stratification de population est important aussi bien pour les analyses d'association, car c'est un potentiel facteur de confusion pouvant conduire à l'obtention de faux positifs, que pour la détection des signaux de sélection naturelle, qui nécessite un échantillon provenant d'une population homogène. D'après l'origine ethnique déclarée par la mère, la composition en groupes

ethniques diffère significativement entre les deux sites d'étude. Les deux groupes ethniques majoritaires sont les Tori (74%) et les Fon (11%) dans la première cohorte et les Aïzo (68%) et les Fon (22%) dans la seconde. Les autres groupes ethniques représentent moins de 10% des individus dans chaque cohorte.

L'analyse en composantes principales (ACP) est la méthode standard utilisée pour estimer la stratification de population dans les études basées sur des données génomiques. Il s'agit d'une méthode classique de réduction de dimension utilisée pour l'exploration de données quantitatives complexes. Le principe de l'ACP est d'identifier des axes orthogonaux de variation, appelés composantes principales (PCs pour *principal component*) qui expliquent au mieux la variabilité existant parmi les individus de l'échantillon. Les individus sont ensuite projetés dans un sous-espace défini par les premières PCs, permettant ainsi d'identifier des profils de similarités/différences entre les individus. Dans le contexte des données génétiques, les SNPs sont considérés comme autant de variables quantitatives sur lesquelles l'ACP est appliquée. Les PCs correspondent alors à des axes orthogonaux expliquant au mieux la diversité génétique existant au sein de l'échantillon. Dans ce sous-espace, les individus issus d'une même population vont se regrouper sous forme d'amas, alors que des individus provenant de différentes origines vont former des groupes distincts. Suivant l'échelle de l'échantillonnage (au niveau d'un continent, d'un pays, d'une région) la représentation graphique des premières PCs peut être utilisée pour estimer une structure entre plusieurs populations ou une sous-structure au sein d'une même population.

2.3.1 Analyse en composantes principales sur les deux cohortes

Une première ACP a été réalisée sur les 624 individus non apparentés ($\Phi < 0.05$) des deux cohortes. L'ACP a été obtenue à partir de la matrice GRM, tel que définie à la section 2.2.2, calculée sur un sous-ensemble de SNPs communs (MAF > 0.05) et après filtre des SNPs en DL ($r^2 > 0.2$). Les 151 individus exclus initialement ont ensuite été projetés sur les PCs, c'est-à-dire représentés dans les différents plans comme des individus supplémentaires.

L'ACP ne révèle pas d'individus ou de groupe d'individus ayant une composition génétique distincte (Figure 2.5). La PC1 sépare les deux cohortes alors que la PC2 révèle une certaine hétérogénéité dans la cohorte de Tori-Bossito (Figure 2.5a). Etonnamment, la diversité génétique observée au sein de ces cohortes n'apparait pas liée au groupe ethnique déclaré par la mère (Figure 2.5c) mais au centre de santé où l'enfant consulte (Figure 2.5d). La structure génétique apparait donc liée à la localisation géographique dans cette région du Bénin, plutôt qu'à l'appartenance ethnique.

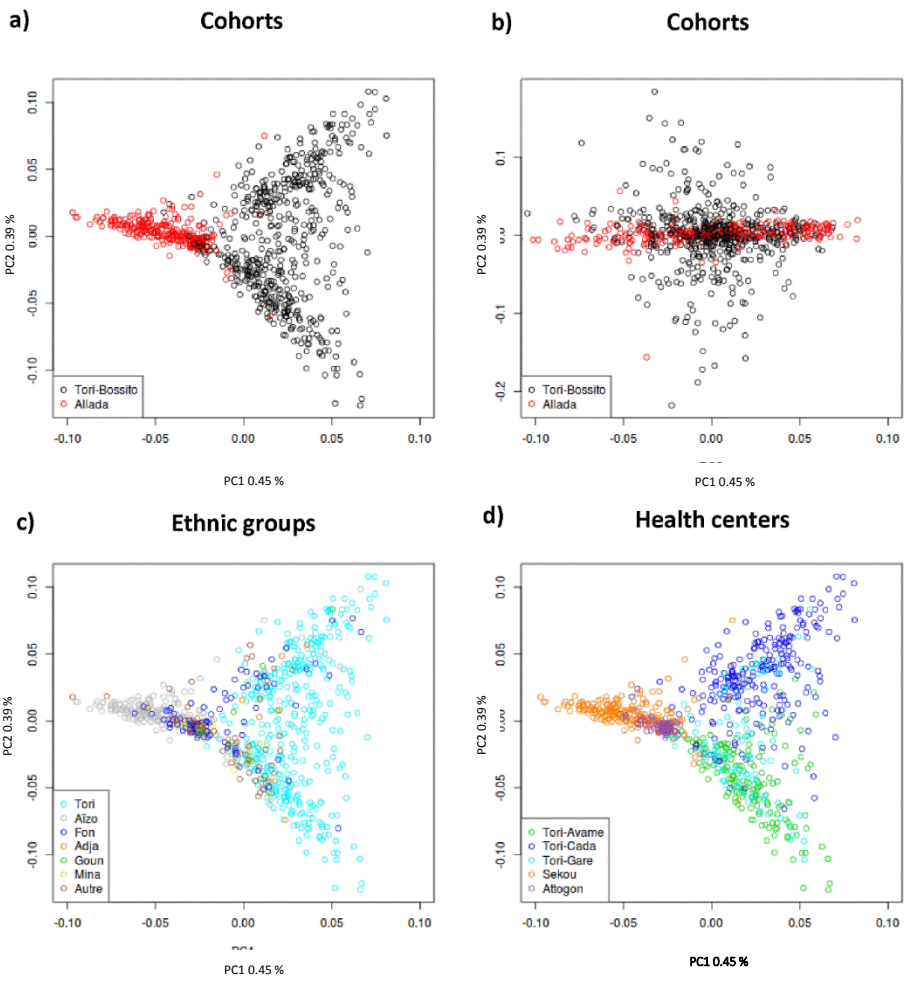


Figure 2.5 Analyse en composantes principales sur les individus des deux cohortes
a) et b) deux premiers plans principaux avec une coloration des individus en fonction de la cohorte, c) et d) premier plan principal avec une coloration des individus en fonction du groupe ethnique et du centre de santé

2.3.2 Analyse en composantes principales incluant des populations africaines du projet 1000 Génomes

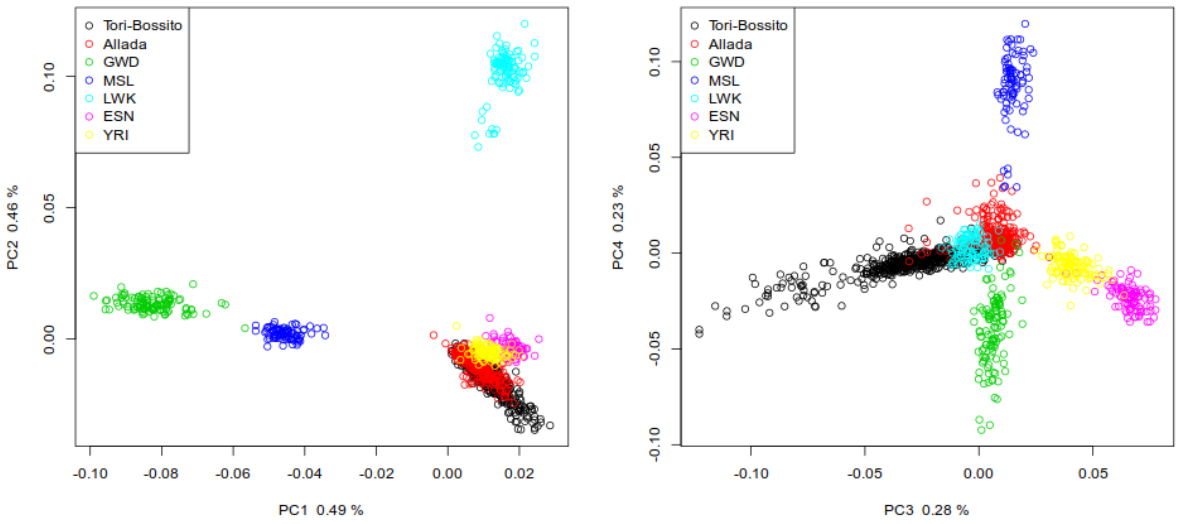


Figure 2.6 Analyse en composante principales incluant des populations d'Afrique subsaharienne
GWD = Gambien, MSL = Mende de Sierra Leone, LWK = Luhya du Kenya, ESN = Esan du Nigeria, YRI = Yoruba du Nigeria

Afin de comparer notre population aux autres populations d'Afrique subsaharienne, nous avons réalisé une seconde ACP en incluant plusieurs populations africaines du projet 1000 Génomes (les Gambiens, les Mende de la Sierra Leone, les Luhya du Kenya, et les Esan et les Yoruba du Nigéria). Les données génétiques des deux cohortes ont été fusionnées avec celles des données 1000 Génomes sur les positions communes. L'ACP a été appliquée comme dans la première analyse sauf que cette fois, seuls 100 individus non apparentés de chacune des deux cohortes ont été inclus dans l'analyse (taille d'échantillon similaire à celle des populations du projet 1000 Génomes) afin de donner un poids équivalent aux différentes populations dans l'analyse. Comme pour la première analyse, l'ensemble des individus des deux cohortes sont représentés dans les deux premiers plans principaux (Figure 2.6), les individus exclus de l'analyse ayant été projetés ensuite sur les PCs. On observe sur la PC1 un *cline* de variation d'ouest en est. La PC2 sépare la population Kenyane du reste des populations d'Afrique de l'ouest. Nos deux populations se superposent et apparaissent proches génétiquement des deux populations nigérianes (Yoruba et Esan). On n'observe pas d'hétérogénéité plus importante dans les deux cohortes du Sud Bénin que dans les échantillons de populations du projet 1000 Génomes.

En résumé, nos deux cohortes présentent une certaine homogénéité génétique, lorsque l'on se situe à l'échelle du continent africain. On observe cependant une sous-structure qui apparait liée aux sites d'études et, au sein des sites d'étude, au centre de santé.

3 ETUDE D'ASSOCIATION GENOME ENTIER SUR LES FORMES SIMPLES DE PALUDISME

Cette analyse génome entier constituant la première GWAS sur les formes simples de paludisme, l'objectif était de rechercher des variants génétiques avec des effets forts. Réalisée sur deux cohortes de jeunes enfants, suivis de la naissance à 18-24 mois, période au cours de laquelle l'immunité acquise n'est pas encore efficace, cette étude cible plus particulièrement les facteurs génétiques impliqués dans l'immunité innée ou dans les mécanismes de résistance innée.

Les études d'association ayant porté sur les formes simples de paludisme ont considéré des phénotypes divers : le fait de développer ou non un accès palustre pendant une saison de transmission, le délai de survenue de la première infection (ou du premier accès palustre), le nombre d'infections au cours du suivi ou encore les niveaux de densité parasitaire. Nous avons fait le choix pour cette première analyse de considérer l'ensemble des infections au cours du suivi afin de prendre en compte le maximum d'informations disponibles pour définir la sensibilité au paludisme. Une première analyse a porté sur les accès palustres simples, ciblant des variants génétiques pouvant être impliqués dans les différentes étapes, de l'infection au développement d'une forme symptomatique (transition A et B du schéma de la Figure 4). La seconde analyse a porté sur l'ensemble des infections (incluant les accès palustres et les infections asymptomatiques détectées une fois par mois lors des visites systématiques) et cible plus particulièrement les facteurs impliqués dans la première transition, c'est-à-dire de l'infection au développement de la parasitémie.

La pratique courante dans les études d'épidémiologie ou de génétique sur le paludisme qui considèrent l'ensemble des infections est d'analyser le nombre total d'infections par enfant en utilisant un modèle de poisson ou une régression binomiale négative. Dans cette étude, nous avons opté pour un modèle de Cox mixte car il présente l'avantage de permettre un ajustement sur des variables dépendantes du temps. En effet, dans ces cohortes, le risque d'exposition environnementale a été estimé à un niveau individuel de manière répétée tout au long du suivi; aussi

le modèle de Cox mixte permet de tenir compte des variations d'exposition tout au long de l'année alors que les deux autres modèles ne peuvent prendre en compte qu'une variable résumée de cette information (comme la moyenne sur l'ensemble du suivi par exemple). Ceci est d'autant plus important dans notre étude qu'il s'agit d'une cohorte de nouveau-nés, où les enfants ne sont pas tous inclus à la même période et donc n'expérimentent pas les saisons d'intense transmission au même âge.

La GWAS a été réalisée en suivant une stratégie découverte/réplication, en utilisant la cohorte de Tori-Bossito (n=525) comme cohorte de découverte, et celle d'Allada (n=250) comme cohorte de réplication. Dans la seconde cohorte, l'analyse a été limitée à la période allant de 12 à 24 mois, le protocole de suivi n'ayant pas été similaire à celui de la cohorte de Tori-Bossito la première année. En conséquence, 25 enfants supplémentaires ont été exclus par rapport au nombre d'enfants disponibles après le contrôle qualité, ces enfants n'ayant pas été suivis au cours de leur deuxième année.

Je présente tout d'abord dans une première partie les modèles mixtes, et en particulier le modèle mixte linéaire utilisé pour prendre en compte la structure de population dans les GWAS ainsi que le modèle de Cox mixte pour l'analyse des évènements récurrents. Les principaux résultats de la GWAS sont détaillés ensuite dans la seconde partie.

3.1 Les modèles mixtes

Un modèle mixte est un modèle incluant à la fois des effets fixes et des effets aléatoires. Les facteurs à effets fixes correspondent à des variables dont les effets sont des paramètres à estimer dans le modèle. Ces effets sont estimés au niveau de la population. Ils sont dits « fixes » car identiques d'un individu à l'autre pour un niveau donné de la variable. Ils interviennent dans l'estimation de la moyenne du modèle. Pour les facteurs à effets aléatoires, on ne s'intéresse pas, en général, aux estimations de ces effets (tirés aléatoirement dans une loi normale) mais on cherche à prendre en compte leur variance dans le modèle. Les effets aléatoires permettent notamment de modéliser un

grand nombre d'effets, par exemple dans le cas de données mesurées de manière répétée, des effets propres à chaque individu. Les différents paramètres de variance à estimer dans le modèle (variances des erreurs du modèle, variances des facteurs à effets aléatoires) sont regroupés sous le terme de composantes de la variance.

Les techniques standards de régression (régression linaire, régression logistique, etc.) supposent que les données soient indépendantes et identiquement distribuées. Lorsque les données présentent une structure de corrélation, l'introduction d'un effet aléatoire va permettre de prendre en compte cette corrélation dans le modèle. Dans le cas de l'exemple ci-dessus, la prise en compte de la corrélation entre les mesures répétées chez un même individu se fait *via* l'introduction d'effets aléatoires individuels. Cette corrélation peut avoir plusieurs autres origines : l'emboîtement des données (par exemple dans une étude avec des individus inclus dans plusieurs centres hospitaliers) ou encore dans les études génétiques, la présence d'individus apparentés dans l'échantillon. Nous présentons, ci-dessous, les deux modèles mixtes utilisés dans le cadre de cette étude.

3.1.1 Le modèle linaire mixte pour la prise en compte de la structure de population

Le modèle linéaire mixte est actuellement la méthode de référence pour corriger la structure de population dans les études d'association car elle permet de corriger à la fois la stratification de population et l'apparentement entre les individus, ces deux situations induisant une inflation de la statistique de test et l'obtention de faux positifs.

Si l'on considère *Y*, un phénotype quantitatif mesuré sur *n* individus, *X* la matrice des covariables sur lesquelles on souhaite ajuster le modèle, *G* le génotype du SNP testé, le modèle peut s'écrire :

$$Y_i = X_i\beta + G_i\gamma + u_i + \varepsilon$$

avec :

- βet *γ* des effets fixes. β est le vecteur des effets associés aux covariables et *γ* l'effet du SNP que l'on cherche à estimer

- X_i le vecteur des covariables pour l'individu *i*,
- u_i l'effet aléatoire associé à l'individu *i* ; le vecteur $u = (u_1 \cdots u_n)$ suit une loi $N(0, \tau \mathrm{K})$
- ε le terme d'erreur aléatoire, $\varepsilon \sim N(0, \sigma^2 I_n)$

Lorsque l'échantillon inclut des individus avec des relations de parenté proches, l'utilisation d'un modèle mixte est assez intuitive, les données génétiques n'étant plus indépendantes. Leur corrélation va dépendre du lien de parenté existant entre les individus, et plus précisément de la quantité de gènes qu'ils ont en commun. Le modèle mixte prend en compte la corrélation entre les données en mettant des effets aléatoires individuels qui ont une structure de variance-covariance qui dépend d'une matrice K, des relations de parenté entre les individus deux à deux. Les relations de parenté peuvent être estimées soit à partir de la reconstitution des généalogies dans les familles, alors $K_{ij} = 2\varphi_{ij}$ où φ_{ij} correspondant au coefficient de parenté estimé entre les individus *i* et *j*, soit lorsque les données sont disponibles sur l'ensemble du génome, par la matrice GRM définie dans la section 2.2.2.

La stratification de population, tout comme l'apparentement, fait intervenir des corrélations génétiques entre les individus. En effet, la stratification de population implique la présence de plusieurs populations ou sous-populations distinctes dans un échantillon, et les individus appartenant à une même sous-population peuvent être vus comme un ensemble d'individus partageant un même ancêtre commun et ayant des caractéristiques génétiques communes (Astle and Balding, 2009).

Les estimations obtenues par la GRM englobent l'ensemble des corrélations génétiques existant entre les individus. Aussi lorsque l'on introduit dans le modèle un effet aléatoire qui dépend de la GRM, celui-ci permet de prendre en compte les deux types de structure. Ce modèle parait particulièrement adapté à notre étude, l'échantillon présentant à la fois une sous-structure de population et un certain degré d'apparentement entre les individus.

3.1.2 Le modèle de Cox mixte

Le modèle de Cox mixte, encore appelé modèle de fragilité (*frailty model* en anglais) est une extension du modèle de Cox adaptée à l'analyse des données de survie corrélées telles que les données d'événements récurrents. Le terme de fragilité fait référence au fait qu'indépendamment des variables incluses dans le modèle, certains individus vont présenter un risque plus élevé et sont donc plus fragiles.

Le modèle de Cox permet d'exprimer le risque instantané de survenue de l'événement, qui s'écrit :

$$\lambda_i(t) = \lambda_0(t)e^{X_{i,t}\beta},$$

avec :

- β le vecteur des effets associés aux covariables
- $X_{i,t}$ le vecteur des covariables, pouvant inclure des variables dépendantes du temps, pour l'individu $_i$
- $\lambda_0(t)$ la fonction de risque de base, qui correspond au risque instantané de l'évènement lorsque l'ensemble des covariables sont nulles ou que $X\beta = 0$.

Dans le cas de l'analyse d'événements récurrents, la corrélation entre les événements est prise en compte par l'ajout d'un effet aléatoire individuel. Le risque instantané de survenue de l'événement estimé par le modèle de Cox mixte peut s'écrire (Therneau & Grambsch, 2000):

$$\lambda_i(t) = \lambda_0(t)e^{X_{i,t}\beta+u_i} \quad (1)$$

avec :

- β le vecteur des effets fixes associé aux covariables
- $X_{i,t}$ le vecteur des covariables, pouvant inclure des variables dépendantes du temps, pour l'individu $_i$
- $\lambda_0(t)$ la fonction de risque de base
- u_i l'effet aléatoire associé à l'individu *i* ; le vecteur $u = (u_1 \cdots u_n)$ suit une loi $N(0, \sigma^2 I_n)$

Dans ce modèle, les effets aléatoires individuels sont tirés dans une loi normale et sont indépendants les uns des autres.

Dans notre étude, afin de prendre en compte la structure de population, le modèle le plus adapté pour tester l'association avec la récurrence des infections palustres est un modèle de Cox mixte incluant des effets aléatoires individuels qui ne sont plus indépendants mais dont la structure dépend de la matrice GRM.

$$\lambda_i(t) = \lambda_0(t)e^{X_{i,t}\beta + G_i\gamma + u_i} \qquad (2)$$

avec :

- γ l'effet du SNP que l'on cherche à estimer
- u_i l'effet aléatoire associé à l'individu *i* ; le vecteur $u = (u_1 \cdots u_n)$ suit une loi $N(0, \tau K_n)$

3.1.3 Stratégie d'analyse en deux étapes de la GWAS

Le modèle de Cox mixte (2) ne pouvant pas être appliqué sur l'ensemble du génome du fait du temps de calcul, l'étude d'association a été réalisée en deux étapes, permettant d'obtenir rapidement une approximation du modèle. Dans une première étape, un modèle de Cox mixte tel que défini en (1) avec des effets aléatoires individuels indépendants a été utilisé pour estimer les effets des facteurs environnementaux et des facteurs individuels autres que génétiques, sur le risque de survenue des infections. Pour définir les variables à inclure dans le modèle, nous avons utilisé une procédure pas à pas descendante et conservé dans le modèle final uniquement les variables avec une $p < 0.05$. Ce modèle nous permet ensuite d'obtenir le *« Best Linear Unbiased Predictor »* (BLUP) $\hat{u}$ de u, qui correspond à différents risques individuels (ou fragilité individuelle) une fois les covariables prises en compte. Dans la seconde étape, l'association a été testée avec les SNPs avec un modèle mixte linéaire, pour corriger pour la structure de population, en considérant comme variable à expliquer la fragilité individuelle $\hat{u}$ obtenue dans le premier modèle :

$$\hat{u}_i = G_i\gamma + v_i + \varepsilon$$

avec v_i l'effet aléatoire associé à l'individu *i*, dont la structure de variance-covariance dépend de K, la matrice GRM.

La même stratégie d'analyse a été appliquée dans les analyses de découverte et de réplication. La seule différence est que pour l'analyse de réplication, seuls les SNPs trouvés associés avec une p-valeur < 10^{-5} dans la première cohorte ont été testés.

Ensuite, pour les régions présentant un signal d'association fort (*p*-valeurs aux alentours du seuil de signification dans la cohorte de découverte et/ou p-valeurs <0.05 dans la cohorte de réplication), les données ont été réanalysées en une seule étape avec le modèle de Cox mixte (2) afin de contrôler la qualité de l'approximation. Les mêmes facteurs non génétiques que dans la première étape d'ajustement sur les covariables ont été inclus dans le modèle.

L'ensemble des analyses statistiques ont été réalisées avec le logiciel R (R Core Team, 2017), avec le paquet gaston (Perdry & Dandine-Roulland, 2018) pour les tests d'association sur l'ensemble du génome et avec le paquet coxme (Terry M. Therneau, 2018) pour le modèle de Cox mixte.

3.2 Résultats

3.2.1 Le suivi palustre dans les deux cohortes

Pour les 525 nouveau-nés inclus dans l'échantillon de découverte (cohorte de Tori-Bossito), la durée moyenne de suivi (écart type, ET) était de 16,9 mois (2,83). Au cours du suivi, 342 d'entre eux (65,1%) ont fait au moins un accès palustre simple, et 359 (68,4%) au moins une infection (un accès palustre ou une infection asymptomatique). Le nombre total d'infections au cours du suivi varie de 0 à 10 pour les accès palustres et de 0 à 16 pour l'ensemble des infections.

Dans la cohorte de réplication (cohorte d'Allada), les 250 enfants ont été suivis en moyenne 11,9 mois (ET = 1,72) sur les 12 mois de suivi considérés. On observe dans cette cohorte une proportion plus élevée d'enfants ayant fait au moins accès palustre simple (83,2%) ou au moins une infection au cours du suivi (86,8%). L'étendue du nombre d'infections est similaire à celle de la première cohorte : de 0 à 9 accès palustres par enfant et de 0 à 14 infections au total. Les distributions du nombre

d'infections, par cohorte et par type d'infections sont présentées en annexe (Annexe 2 Figure 1). Le calcul de l'incidence des accès palustres par mois, tout au long du suivi (Annexe 2 Figure 2a) montre que l'incidence des accès palustres n'est pas plus élevée dans la cohorte de réplication que dans celle de découverte. Le pourcentage plus élevé d'infections observées semble plus vraisemblablement lié au fait que les enfants suivis dans la cohorte de réplication sont plus âgés (12-24 mois) que ceux de la cohorte de découverte (0-18 mois) et que le risque de survenue d'un événement est plus élevé la seconde année.

3.2.2 Ajustement sur les facteurs individuels et environnementaux

Un grand nombre de facteurs de risque potentiels ont été considérés. La plupart sont communs aux deux cohortes. Ils peuvent se regrouper en plusieurs grandes catégories :

- **des facteurs relatifs à l'enfant** : le sexe, le petit poids de naissance (<2500 g), l'anémie de l'enfant à la naissance (hémoglobine <11g/dl, mesurée dans le sang de cordon), le centre de santé où sont effectuées les consultations,
- **des facteurs relatifs à l'infection de la mère par le paludisme pendant la grossesse** : l'infection placentaire (apposition placentaire positive), l'infection palustre durant le suivi de grossesse (au moins une infection, pour la cohorte d'Allada uniquement), la prise d'un traitement intermittent préventif pendant la grossesse (cohorte de Tori-Bossito) ou le bras de l'essai clinique (cohorte d'Allada),
- **d'autres facteurs relatifs à la mère :** l'âge (en quatre classes : « ≤20 » ; « 20-25 » ; « 25-30 » ; « >30 »), le niveau d'éducation, l'anémie à l'accouchement (hémoglobine <11g/dl), la parité (primipare *vs* multipare), le statut marital,
- **des facteurs relatifs à l'exposition aux moustiques :** le risque individuel d'exposition aux piqûres d'anophèles, l'utilisation de la moustiquaire (codée en 4 catégories selon la fréquence de réponse positive de la mère à la question: « Est-ce que l'enfant a dormi sous moustiquaire la veille ? », posée lors des visites systématiques), la saison de transmission (une catégorie par année pour les saisons des pluies, par exemple « saison des pluies 2007 »,

« saison des pluies 2008 » et « saison des pluies 2009 » pour la cohorte de Tori-Bossito ; une catégorie « saison sèche » pour l'ensemble des mesures en dehors de la saison des pluies).

Plus de détails sur ces facteurs de risque non génétiques sont disponibles dans les tables supplémentaires du premier article (en Annexe).

Pour prendre en compte le risque environnemental dépendant du temps, le suivi a été divisé en intervalles d'un mois suivant les visites mensuelles. La fin d'un intervalle correspond soit à la survenue d'un événement codé 1 (accès palustre dans le cas de la première analyse, accès palustre ou infection asymptomatique dans le cas de la seconde) soit à la visite systématique mensuelle suivante (événement codé 0 si l'enfant ne présente pas d'infection). A chaque événement est associé le risque d'exposition mesuré à la visite antérieure.

Les résultats du modèle de Cox mixte dans cette première étape sont présentés ci-dessous pour les deux phénotypes considérés : la récurrence des accès palustres simples (APS) et la récurrence de l'ensemble des infections (EI), dans la cohorte de découverte (Table 3.1). Le modèle final contient peu de facteurs par rapport à l'ensemble des facteurs considérés dans cette étude et ce sont les mêmes pour les deux phénotypes : le centre de santé, le risque d'exposition (qui correspond au logarithme du risque d'exposition aux piqûres d'anophèles tel que défini par Cottrell *et al*., 2012).

	Accès palustres simples		**Ensemble des infections**	
Covariables	HR (95%CI)	*P*	HR (95%CI)	*P*
Centre de santé				
Tori Avamè	Référence		Référence	
Tori Cada	2.38 (1.90-2.97)	1.4e-14	2.55 (2.03-3.19)	1.1e-16
Tori Gare	1.52 (1.14-2.03)	4.3e-03	1.77 (1.32-2.35)	9.0e-05
Risque d'exposition	1.26 (1.16-1.37)	6.4e-09	1.24	1.4e-09
Saisons de transmission				
saison sèche	Référence		Référence	
saison des pluies 2007	1.76 (1.06-2.90)	2.8e-02	2.15 (1.38-3.36)	6.8e-04
saison des pluies 2008	1.46 (1.15-1.87)	1.7e-03	1.62 (1.31-2.00)	6.2e-06
saison des pluies 2009	3.27 (2.48-4.31)	1.1e-16	2.57 (2.00-3.28)	7.9e-14

Table 3.1 Modèles de Cox finaux d'ajustement sur les covariables pour les accès palustres simples et l'ensemble des infections dans la cohorte de découverte

HR, Hazard Ratio. Les modèles incluent les facteurs trouvés associés avec une p-valeur < 0.05. Pour les 550 enfants, un total de 1072 infections, dont 811 accès palustres simples, ont été observées.

.

Des résultats similaires ont été obtenus pour la seconde cohorte. Les facteurs trouvés associés sont également identiques pour les deux phénotypes. Les mêmes trois facteurs sont inclus dans le modèle final ainsi que deux autres facteurs relatifs à la mère (le niveau de scolarité de la mère et le statut marital). Les estimations des effets aléatoires individuels û dans ces modèles sont ensuite utilisées comme phénotype dans l'analyse génome entier.

3.2.3 Analyse d'association génome entier

Dans la cohorte de découverte, l'analyse d'association a été réalisée pour 15 566 900 variants génotypés ou imputés, disponibles après le contrôle qualité, avec une MAF supérieure à 0,01. L'association a été testée avec un modèle additif, en considérant le nombre d'allèles mineurs pour les SNPs génotypés et le dosage des allèles pour les SNPs imputés. Le dosage correspond au nombre d'allèles mineurs attendu au locus compte tenu des probabilités des génotypes estimés lors de l'imputation. Cette mesure tient compte de l'incertitude pouvant exister lors de l'imputation. Les résultats obtenus sur l'ensemble du génome sont représentés ci-dessous pour la récurrence des accès palustres simples (APS, Figure 3.1a) et pour la récurrence de l'ensemble des infections (EI, Figure 3.1b). A droite de la figure, les diagrammes quantile-quantile (Q-Q plot) montrent que la structure de population est bien prise en compte dans l'analyse. On n'observe pas d'inflation de la statistique : le facteur génomique d'inflation (λ) est très proche de 1 dans les deux analyses. Les Manhattan plots, à gauche de la figure, représentent les résultats des tests d'association (le - $\log_{10}$ de la *p*-valeur) en fonction de la position des SNPs sur le génome.

Dans les deux analyses, on observe plusieurs signaux d'association forts. Pour quatre d'entre eux, les *p*-valeurs sont proches du seuil de signification communément admis dans les GWAS ($p = 5 \times 10^{-8}$) : deux signaux pour APS, situés dans le gène *SYT16* (région 14q23.2, meilleur SNP rs375961263, $p = 3.7 \times 10^{-8}$) et dans le gène *PTPRM* (région 18p11.23, meilleur SNP rs113776891, $p = 3.77 \times 10^{-8}$); deux autres pour EI situés dans le gène *ACER3*, (région 11q13.5, meilleur SNP rs77147099, $p = 6{,}85 \times 10^{-8}$) et dans le gène *PTPRT* (région 20q12, meilleur SNP rs111968843, $p = 9{,}70 \times 10\text{-}8$). Les signaux les plus forts observés pour EI dans *ACER3* et *PTPRT*, apparaissent également dans l'analyse des APS avec une

p-valeur <10^{-5} (*p* = 1,78 x 10^{-7} et *p* = 1,60 x 10^{-6} respectivement pour les meilleurs SNPs dans *PTPRT* et *ACER3*).

Nous avons ensuite sélectionné les SNPs avec une *p*-valeur inférieure à 10^{-5} (un niveau pouvant être

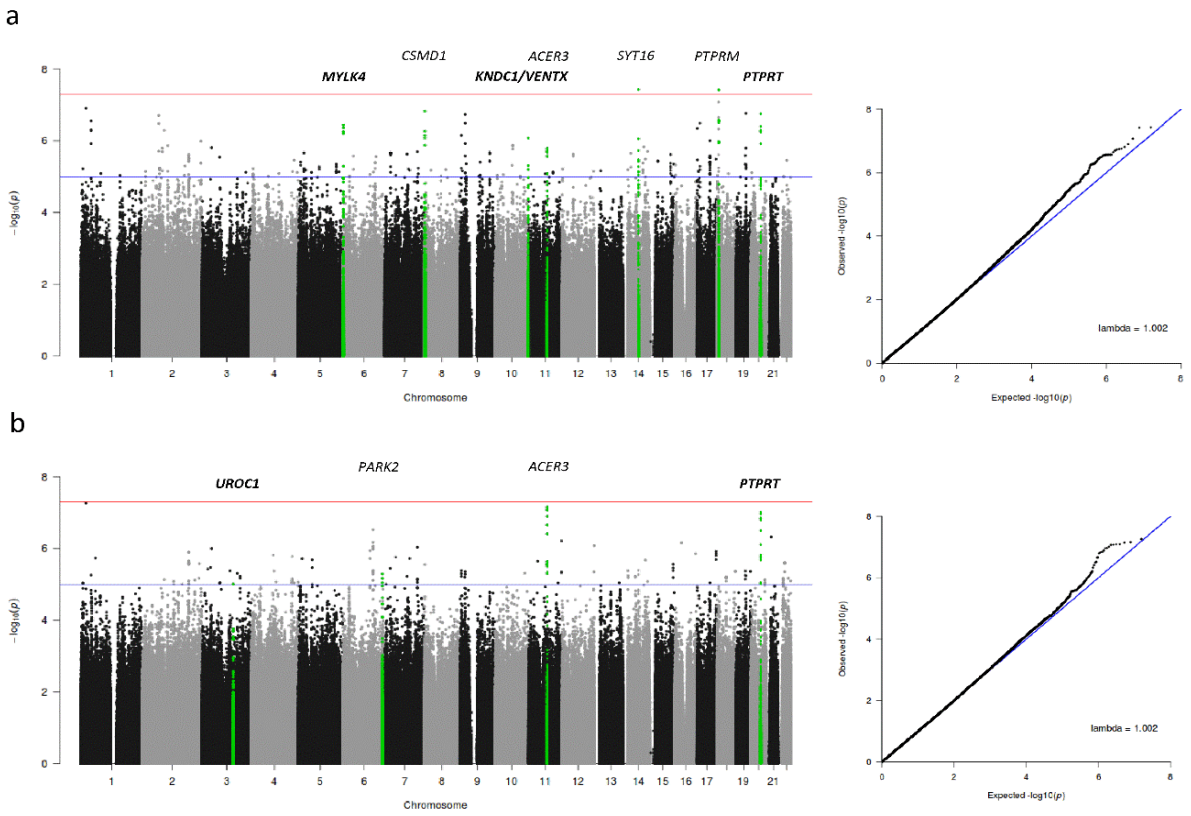


Figure 3.1 Manhattan plot et Q-Q plot de l'analyse génome entier dans la cohorte de découverte *: a) pour la récurrence des accès palustres, b) pour la récurrence de l'ensemble des infections. Les lignes horizontales dans le Manhattan plot indiquent le seuil de signification statistique (p = 5 x 10^{-8}, en rouge) et le seuil à partir duquel les SNPs ont été testés dans la cohorte de réplication (p = 10^{-5}, en bleu)*

interprété comme une association suggestive) pour l'étape de réplication. L'association a d'abord été testée de la même manière que dans la cohorte de découverte, en utilisant le phénotype intermédiaire ajusté sur les covariables et un modèle linéaire mixte (MLM). Puis, pour les signaux d'association dont la *p*-valeur était proche du seuil de signification dans la cohorte de découverte ou pour ceux dont l'association était répliquée (MLM, *p*<0,05), les données ont été réanalysées avec le modèle de Cox mixte (2) dont la structure de variance-covariance dépend de la GRM (MMCox), ce modèle constituant notre *gold standard*. Une méta-analyse des résultats obtenus dans les deux

cohortes avec le MMCox a ensuite été réalisée avec le logiciel METAL (Willer & Li, 2010) afin d'évaluer un effet et une p-valeur globale pour les variants sur les deux cohortes.

Au total 356 et 214 variants ont été testés dans la cohorte de réplication, pour les phénotypes APS et EI respectivement. Le signal pour lequel on observe les plus faibles *p*-valeurs dans l'étude de réplication est associé au phénotype APS et est situé dans la région chromosomique 6p25.2, au niveau du gène *MYLK4* (Table 3.2). Ce signal ne faisait pas partie des signaux les plus forts dans la cohorte de découverte mais le SNP avec la plus faible *p*-valeur dans la cohorte de réplication (rs72840075, $p = 0{,}0063$) est également celui pour lequel on a le plus d'évidence d'association globalement dans la méta-analyse sur les deux cohortes ($p = 5{,}29 \times 10^{-8}$).

Locus	SNP	MAF Tori-Bossito	*p* MMCox Tori-Bossito	*p* MMCox Allada[a]	Gène le plus proche	*p* Méta-analyse
6p25.2	rs76088706	0.025	2.68×10^{-6}	0.018	C6orf195(intergénique)	3.65×10^{-7}
	rs140858180	0.026	2.53×10^{-6}	0.030	C6orf195(intergénique)	6.02×10^{-7}
	rs547331171	0.026	2.51×10^{-6}	0.028	*MYLK4* (intergénique)	5.57×10^{-7}
	rs142480106	0.019	1.14×10^{-6}	0.0063	*MYLK4* (UTR3)	5.29×10^{-8}
	rs144194334	0.026	2.53×10^{-6}	0.026	*MYLK4* (UTR3)	5.21×10^{-7}
	rs72840075	0.030	1.05×10^{-5}	0.0028	*MYLK4* (intronique)	2.01×10^{-7}
10q26.3	rs182416945	0.12	5.56×10^{-6}	0.01	*KNDC1* (intronique)	7.94×10^{-7}
14q23.2	rs61743638	0.047	1.95×10^{-5}	0.04	*SYT16 (intronique)*	6.21×10^{-6}
	rs62639692	0.047	1.87×10^{-5}	0.04	*SYT16 (intronique)*	6.00×10^{-6}
18p11.32	rs113776891	0.076	1.75×10^{-7}	0.13	*PTPRM* (intronique)	1.15×10^{-6}
20q12	rs6124419	0.12	2.51×10^{-6}	0.02	*PTPRT* (intronique)	6.82×10^{-7}

Table 3.2 Résultats des tests d'association avec le modèle de Cox mixte pour les accès palustres simples.
Cette table inclut uniquement les SNPs dont l'association réplique ou pour le fort signal dans la cohorte de découverte qui ne réplique pas (dans PTPRM), le SNP avec la plus faible p-valeur.
MMCox : modèle de Cox mixte.
[a] p-valeurs du test unilatéral

L'association est répliquée pour cinq autres signaux à un seuil de 0,05 (Table 3.2 pour APS et Table 3.3 pour EI). L'association est retrouvée pour trois des quatre signaux principaux mentionnés dans l'analyse de la cohorte de découverte. Il s'agit des signaux situés dans le gène *SYT16* pour APS, dans *ACER3* pour EI et dans *PTPRT* pour les deux phénotypes. Seul le signal situé dans le gène *PTPRM* ne réplique pas (*p* = 0,13, test unilatéral). Dans le gène *PTPRT*, l'association est retrouvée pour un SNP, rs6124419, le même pour les deux phénotypes. L'analyse de réplication met en évidence également

deux autres régions chromosomiques : la région 10q26.3 pour APS avec un SNP situé dans le gène *KNDC1* et la région 3q21.3 pour EI, avec un SNP situé dans un exon du gène *UROC1*.

Locus	SNP	MAF Tori-Bossito	*p* MMCox Tori-Bossito	*p* MMCox Allada[a]	Gène le plus proche	*p* Méta-analyse
3q21.3	rs9871671	0.11	5.00×10^{-5}	0.02	*UROC1* (exonique)	8.25×10^{-6}
11q13.5	rs545152253	0.017	1.16×10^{-5}	0.045	*ACER3* (intronique)	1.62×10^{-5}
	rs115655584	0.014	5.31×10^{-5}	0.041	*ACER3* (intronique)	2.35×10^{-5}
20q12	rs6124419	0.12	1.35×10^{-6}	0.02	*PTPRT* (intronique)	4.10×10^{-7}

Table 3.3 Résultats des tests d'association avec le modèle de Cox mixte pour l'ensemble des infections.
Cette table inclut uniquement les SNPs dont l'association réplique.
MMCox : modèle de Cox mixte.
[a] p-valeurs du test unilatéral

Pour illustrer l'effet des SNPs observés dans les deux cohortes, nous avons calculé le taux d'incidence des accès palustres en fonction du génotype rs6124419 pour *PTPRT*, dans trois groupes basés sur le niveau d'exposition au vecteur (Figure3.2). Les effets sont très cohérents dans les deux cohortes en fonction des différentes catégories de risque d'exposition.

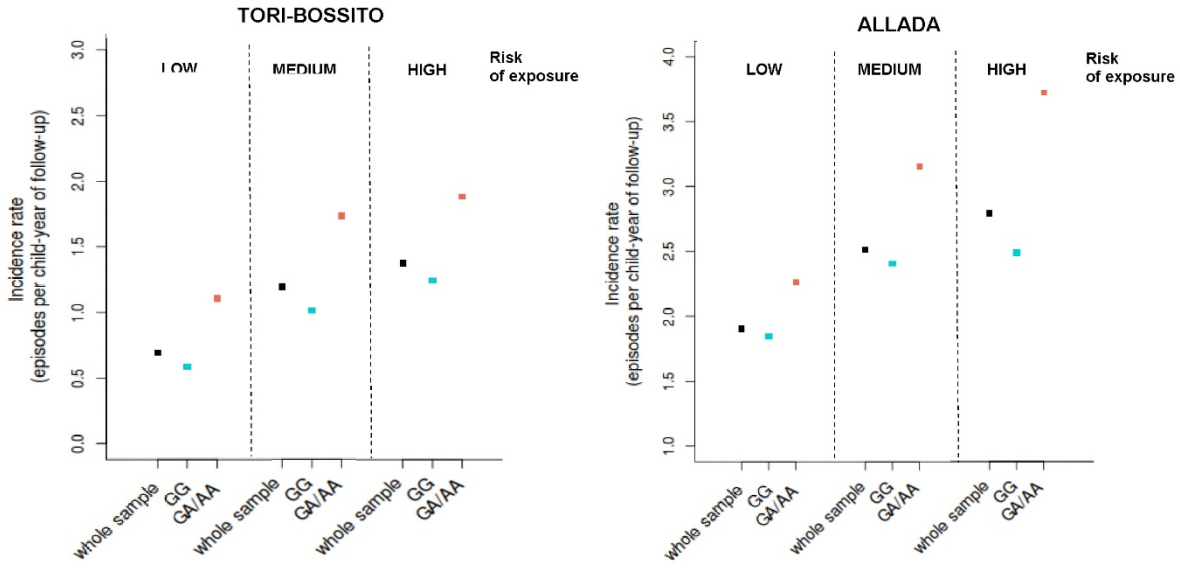


Figure 3.2 Taux d'incidence brut des accès palustres pour le SNP rs6124419 situé dans PTPRT
Les trois classes ont été définies à partir du risque d'exposition moyen calculé sur l'ensemble du suivi pour chaque enfant. Faible, moyen et élevé correspondent aux trois tertiles de la distribution

Les différents signaux mentionnés ont été examinés par un graphique réalisé avec le logiciel LocusZoom (Pruim et al., 2010) et représentant une région de 200 kb autour du SNP principal (Figure 3.3). Le SNP principal désigne ici (et dans la suite de l'étude) le SNP avec la plus faible *p*-valeur dans l'analyse de réplication ou la plus faible *p*-valeur dans l'analyse de découverte pour les signaux dont l'association ne réplique pas. Comme pour le Manhattan plot, les *p*-valeurs sont représentées en fonction de la position physique du SNP sur le chromosome. Ce graphique intègre en plus un ensemble d'informations, telles que le DL entre les SNPs de la région et le SNP principal, le type de SNP (génotypé ou imputé), les taux de recombinaison dans la région. Pour l'ensemble des pics d'association, les données de DL et de *p*-valeurs sont très cohérentes. Les pics contiennent à la fois des SNPs imputés et génotypés ce qui permet d'exclure un artéfact lié à l'imputation. On observe deux pics plus particuliers : un pic très étroit dans le gène *PTPRT* (similaire pour les deux phénotypes) dans une région d'environ 20 Kb incluant les exons 20 à 22 du gène, et un autre qui s'étend sur tout le gène *ACER3*, qui semble être inclus dans un large bloc de DL.

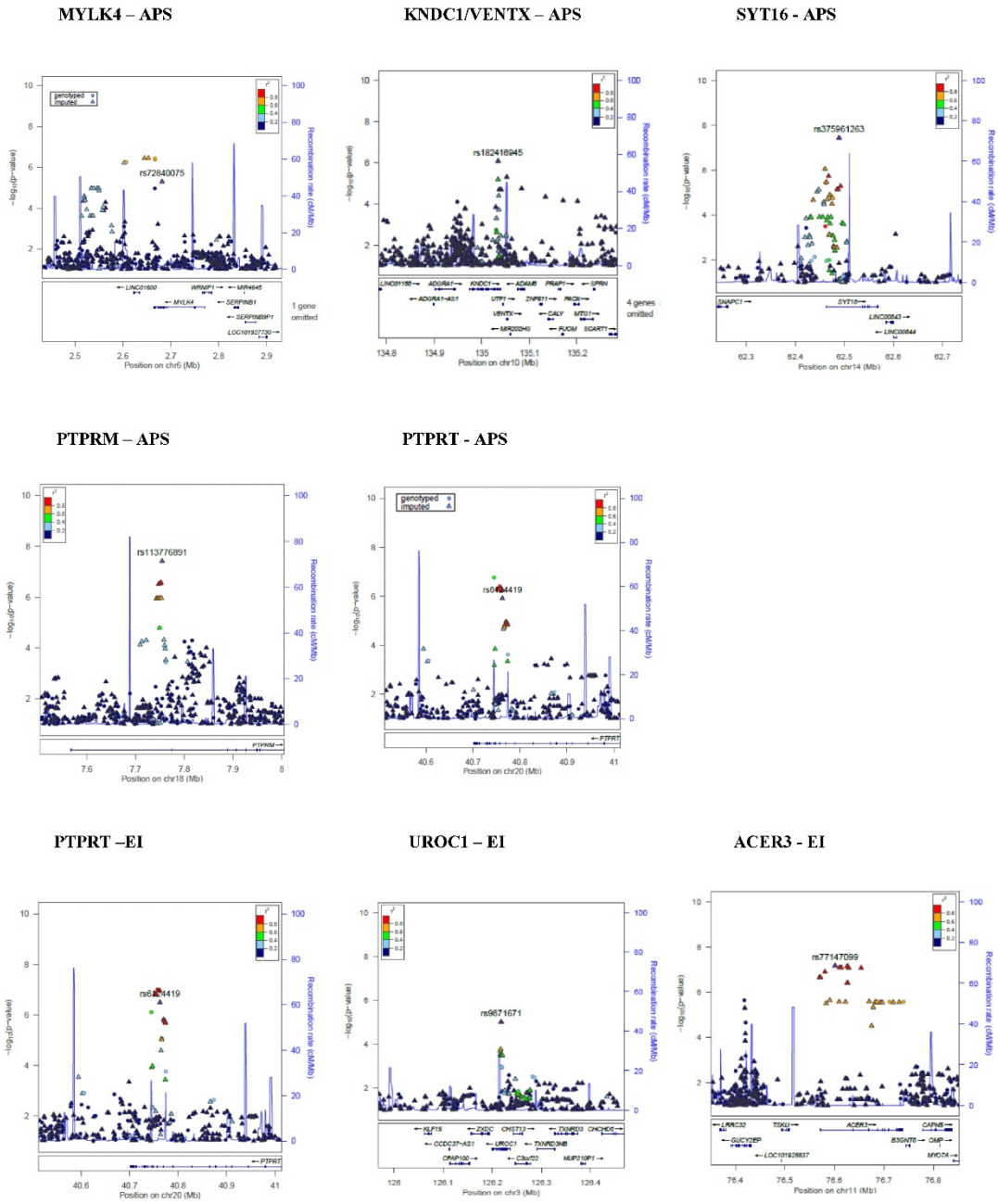


Figure 3.3 Graphique régional produit par LocusZoom pour les principaux signaux d'association, pour les accès palustres simple (APS) et pour l'ensemble des infections (EI)

Les SNPs imputés sont représentés par des triangles, les SNPs génotypés par des cercles. Le SNP principal figure en violet avec une annotation de son identifiant *rs*. La couleur des autres SNPs indique le niveau de DL entre le SNP et le SNP principal. Le DL a été calculé pour cette représentation, à partir des données des deux populations nigérianes du projet 1000 Génomes (YRI et ESN). Le panneau en dessous du graphique permet de visualiser les gènes présents dans la région. Ces graphiques sont visibles dans leur taille normale dans l'article 1, dans le document principal et dans les figures supplémentaires

A la suite, une analyse conditionnelle avec un MMCox incluant le SNP principal comme covariable a été réalisée dans chaque région, afin de déterminer si plus d'un SNP était à l'origine du signal observé. Aucun signal d'association résiduel n'a été trouvé dans les différentes régions.

3.2.4 Analyse fonctionnelle *in silico*

Dans les GWAS, l'association détectée est typiquement une association indirecte ; la plupart du temps le variant causal n'est pas présent dans les données. Des étapes dites *post-GWAS* ou de cartographie fine du signal (*fine-mapping* en anglais) et d'analyse fonctionnelle sont alors nécessaires pour identifier le variant causal et le(s) gène(s) susceptible(s) d'être impliqué(s) dans la physiopathologie de la maladie. Ces étapes étant complexes et coûteuses à mettre en œuvre, des outils bio-informatiques ont été développés ces dernières années, pour aider à sélectionner des variants d'intérêt d'après leurs conséquences fonctionnelles, et le(s) gène(s) « cibles » qui apparaissent les plus susceptibles d'être impliqués (Schaid et al., 2018). Ces outils utilisent des données de séquençage issues de grands projets internationaux afin de densifier la région chromosomique en variants génétiques ainsi que de nombreuses bases de données sur la fonctionnalité des variants maintenant accessibles à la communauté scientifique.

Nous avons poursuivi l'analyse des signaux d'association de cette étude à l'aide de la plateforme en ligne FUMA (Watanabe et al., 2017). Dans un premier temps, FUMA individualise les régions chromosomiques en DL autour des signaux d'association, puis réalise une cartographie fine en récupérant à partir des données de séquençage du projet 1000 Genomes (KGP) l'ensemble des variants en DL avec le meilleur SNP de la région (r^2>0,6). Ces variants sélectionnés ainsi que ceux de la GWA (également en DL avec un r^2>0,6) constituent un ensemble de SNPS candidats. La position de ces SNPs sur le génome délimite la région candidate. Les gènes susceptibles d'être impliqués sont identifiés ensuite par plusieurs approches : i) par cartographie positionnelle (ou *positional mapping*) d'après la position physique des SNP sur le génome; ii) par cartographie eQLT (ou *eQTL mapping*) en utilisant les associations entre les SNPs et les données d'expression de gènes dans différents tissus ; iii) par cartographie des interactions 3D de la chromatine (ou *chromatin*

interactions mapping) : un gène est identifié comme une cible potentielle si le segment de chromatine incluant le gène interagit avec le segment incluant le SNP. FUMA intègre des bases de données bioinformatiques récentes, telles que le *Combined Annotation Dependent Depletion score* (CADD score, (Kircher et al., 2014; Rentzsch et al., 2019), RegulomeDB (RDB score, (Boyle et al., 2012), les données d'expression de gènes provenant du projet GTEx ou encore les interactions 3D de la chromatine obtenues à partir de données HI-C (Schmitt et al., 2016).

Nous avons exploré avec cette méthode les signaux mentionnés ci-dessus, mais aussi plus largement l'ensemble des signaux d'association dans la cohorte de découverte incluant un minimum de trois SNPs avec une p-valeur inférieure à 10^{-5}. La cartographie fine a ainsi été réalisée dans 30 régions indépendantes pour APS et 15 régions pour EI à partir des données KGP des populations africaines (phase 3), et en prenant comme SNP de référence le SNP principal. Pour identifier les gènes d'intérêt à partir des candidats obtenus dans chaque région, les trois approches ont été appliquées. L'analyse par cartographie positionnelle a été limitée aux SNPs fonctionnels, c'est-à-dire aux SNPs codants ou non-codants ayant un score CADD> 12,37 (ce seuil ayant été défini par les auteurs de la méthode pour identifier les variants délétères), ou aux SNPs avec un RDB score ≤ 2, susceptibles d'affecter la liaison avec des éléments régulateurs dans les régions non codantes. Un gène est identifié comme un candidat probable si un SNP fonctionnel est localisé dans la partie codante du gène ou à une distance inférieure à 10 kb de celle-ci. Pour les cartographies eQTL et IC, les données d'expression de gènes et d'interactions de la chromatine étant spécifiques aux tissus, l'analyse a porté sur les tissus qui paraissaient pertinents pour les infections palustres non graves : la peau, les fibroblastes de la peau, le sang total, les lymphocytes B, le foie et la rate. Les données eQTL proviennent de la base de données GTex v7 et de trois bases de données d'associations eQTL identifiées dans le sang total (Blood eQTL, BIOS QTL et eQTLGen).

L'analyse par cartographie physique et eQTL identifie 20 gènes pour APS et 8 pour EI. Les résultats de l'analyse pour 18 régions candidates où un gène « cible» est trouvé sont résumés dans la Table 3.4.

Ces régions correspondent pour EI à l'ensemble des régions pour lesquelles au moins un gène a été identifié; pour APS, aux régions dont le signal d'association réplique ou, dans le cas d'une absence de réplication, aux régions les plus fortement associées.

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

Table 3.4 Gènes identifiés par FUMA d'après les conséquences fonctionnelles des SNPs dans les principaux signaux d'association de la cohorte de découverte

Les régions génomiques présentées sont celles avec au moins un gène identifié par cartographie positionnelle ou cartographie eQTL. Pour chaque région génomique, le SNP principal a été identifié comme le SNP dont l'association réplique ou le SNP avec la p-valeur la plus faible. La p-valeur de la GWA initiale correspond à celle obtenue avec un modèle mixte linéaire sur les phénotypes ajustés. Les cases rouges indiquent si le gène a été identifié par cartographie physique (pos.), cartographie eQTL (eQTL) ou cartographie des interactions de la chromatine (CI). Pour la cartographie physique, les identifiants rs des SNP délétères sont donnés (rsID) avec le score CADD et le score RDB observés pour ces SNPs. Pour la cartographie eQTL, le nombre d'associations eQTL significatives (nSNP) est indiqué, ainsi que la direction de l'effet de l'allèle sur l'expression du gène, pour l'allèle à risque dans la GWAS (eQTL dir.). Les résultats sur l'ensemble des régions candidates figurent dans l'article publié.

Pour tous les signaux dont l'association réplique à l'exception de celui dans la région 14q23.2, un gène a été identifié comme candidat potentiel. Ce gène correspond au gène dans lequel se situe le signal d'association pour 4 des signaux : *MYLK4 (Myosin Light Chain Kinase Family Member 4*) dans la région 6p25.2, *PTPRT (Protein Tyrosine Phosphatase Receptor Type T*) dans la région 20q12, *UROC1* (*Urocanate Hydratase 1*) dans la région 3q21.3, et *ACER3* (*Alkaline Ceramidase 3*) dans la région 11q35.5. Pour la région 10q26.3, FUMA identifie par cartographie eQTL le gène *VENTX* (*VENT Homeobox*) situé à 17Kb du SNP principal : deux SNPs, rs182416945 (celui dont l'association réplique)

et rs138609386 (en DL complet, r^2 =1, avec le premier) sont des eQTL significatifs répertoriés dans une base de données eQTLGen (eQTL dans le sang total). L'allèle à risque dans nos données est associé à une plus faible expression de *VENTX*, un gène codant pour la protéine VENT Homeobox. Pour les gènes *PTPRT* et *UROC1,* un SNP annoté comme délétère est identifié dans la région codante des gènes où se situe le signal. On peut noter ici que pour *UROC1*, celui-ci correspond au SNP dont l'association réplique ; il s'agit d'un SNP non-sens, dont le score CADD est de 26,5 indiquant que ce variant est prédit comme faisant partie des 1% des substitutions les plus délétères du génome. Aussi, ce variant a une probabilité extrêmement forte d'entrainer une modification au niveau de la protéine. Dans la région candidate 6p25.5, le gène *MYLK4* est identifié par une association eQTL significative (pour le SNP rs72840075 dont l'association réplique), qui est retrouvé dans deux bases de données sur le sang total (eQTLGen and BIOS eQTl) ; l'allèle a risque dans nos données est associé à une plus forte expression de *MYLK4*. Dans la région candidate 11q13.5, de multiples éléments ciblent le gène *ACER3* : un grand nombre d'eQTL significatifs (21 SNPs au total) et deux SNPs susceptibles d'affecter la liaison avec des éléments régulateurs (RDB = 2b). Les eQTL sont identifés dans des cultures de lignées cellulaires de fibroblastes (base de données GTex v7). Enfin, dans la région 18p11.23, FUMA met en évidence le gène *PTPRM*, qui code pour une seconde protéine tyrosine phosphatase de type récepteur : deux SNPs, rs138057394 et rs112288193 (r^2 = 0,74 avec le SNP principal, pour les deux SNP) sont annotés comme SNPs délétères (score CADD de 19,58 et 14,64 respectivement).

3.3 Discussion

Cette GWAS met en évidence plusieurs signaux d'association forts avec la sensibilité aux accès palustres simples et à l'ensemble des infections dans la cohorte de découverte. Même s'ils n'atteignent pas strictement le seuil de signification de 5×10^{-8} admis pour les GWAS, ces signaux d'association apparaissent particulièrement élevés, compte tenu de l'effectif de la cohorte. L'analyse de réplication dans la seconde cohorte met en avant 8 signaux, dont l'association est répliquée à un

seuil de 0,05; l'association est répliquée en particulier pour trois des quatre plus forts pics d'association observés dans la cohorte de découverte.

La région chromosomique 6p25.2 est la région pour laquelle nous avons le plus d'éléments en faveur d'une association d'un point de vue statistique. L'association est répliquée pour quatre SNPs ; la p-valeur pour le meilleur SNP dans l'analyse de réplication est inférieure à 0,005 et approche le seuil de signification dans la méta-analyse sur les deux cohortes (p= 5,29 x 10^{-8}). FUMA identifie *MYLK4,* comme le gène le plus probablement associé à ce signal d'après la fonctionnalité de SNPS. *MYLK4* appartient à la famille des kinases des chaînes légères de myosine (MYLK) qui catalysent l'interaction de la myosine avec les filaments d'actine. Il n'existe pas de relation évidente avec le paludisme, cependant les membres de la famille MYLK jouant un rôle essentiel dans l'organisation du cytosquelette d'actine/myosine et dans la motilité cellulaire (Tan & Leung, 2009) et *MYLK4* pourrait jouer un rôle dans la structure de la membrane des globules rouges.

Pour les autres signaux dont l'association réplique, FUMA met en évidence des gènes dont la fonction est pertinente dans la physiopathologie du paludisme simple.
PTPRT et *ACER3* présentent un intérêt tout particulier. Les signaux associés à ces gènes sont observés pour les deux phénotypes et sont proéminents pour EI. Cela semble indiquer que ces gènes jouent un rôle dans la phase précoce de la maladie (au moment de l'infection ou durant la phase hépatique). *PTPRT* code pour une protéine tyrosine phosphatase de type récepteur (PTPR), une famille de protéines transmembranaires qui ont un domaine extracellulaire impliqué dans l'agrégation des cellules et un domaine phosphatase qui joue un rôle dans la transmission des signaux à l'intérieur de la cellule. *PTPRT* est impliqué dans la voie métabolique STAT3 qui joue un rôle dans de nombreuses fonctions physiologiques dont la réponse immunitaire. Elle intervient par exemple dans la phase de réponse aigüe de l'inflammation, une réponse non-spécifique du système immunitaire inné aux infections par les pathogènes (Suarez et al., 2018). Le rôle de *PTPRT* dans la régulation de cette voie métabolique a été démontré, STAT3 étant identifié comme un substrat direct de PTPRT (Peyser et al.,

2016; X. Zhang et al., 2007). De façon intéressante, STAT3 est un transmetteur de signal et un activateur de transcription qui se comporte de manière similaire à NF-κB, dont la voie métabolique est connue pour moduler la réponse de l'hôte à *P. falciparum*. Les sites de fixation se chevauchent notamment dans les régions de régulation des gènes. De plus, STAT3 a été associé au neuropaludisme dans des études expérimentales sur le modèle murin. Le second gène *ACER3* est une céramidase alcaline, impliquée dans la dégradation des céramides en sphingosine-1-phosphatase (S1P). Les céramides ont une action anti-plasmodiale (Heung et al., 2006; Labaied et al., 2004), qui est inhibée par S1P. Dans nos données, l'allèle à risque est associé à une diminution de l'expression d'*ACER3*, indiquant ainsi une activité céramidase plus faible chez les enfants présentant un risque plus élevé d'infections palustres. *ACER3* a également été impliqué dans l'activation de l'immunité innée chez la souris (K. Wang et al., 2016).

Les deux autres gènes mis en évidence *VENTX* et *UROC1* paraissent également pertinents. *VENTX* est un facteur de transcription (une homéoprotéine) qui contrôle la prolifération et la différenciation des cellules hématopoïétiques et immunitaires (Gao et al., 2012; Wu et al., 2011, 2014). Il est associé aux APS dans notre étude, et donc serait impliqué dans le développement des symptômes cliniques de la maladie. Les travaux de Wu et al. ont montré que VENTX était un régulateur clé de la différenciation des macrophages et des cellules dendritiques. Ces deux types de cellules immunitaires, dérivées de monocytes, sont impliqués dans la réponse immunitaire innée et jouent un rôle essentiel dans la réponse aux infections palustres (Chua et al., 2013). Le dernier gène *UROC1* apparait pertinent du fait de son lieu d'expression dans le foie (données GTEX RNA-seq) où s'effectue la première phase du développement du parasite, mais aussi du fait que le SNP dont l'association réplique soit une mutation faux-sens annotée comme fortement délétère. *UROC1* code pour une enzyme, l'urocanate hydratase ou urocanase, qui catabolise l'acide urocanique en acide 4-imidazolone-5-propionique dans le foie. Elle pourrait avoir un rôle direct dans le foie, ou par l'intermédiaire de l'acide urocanique. L'acide urocanique se trouve dans la peau et la transpiration. Il protège la peau des rayons ultraviolets mais il a été démontré également qu'il est un agent chimio-attractant majeur

pour un parasite nématode qui infeste l'Homme par la peau (Safer et al., 2007). On peut émettre l'hypothèse que l'acide urocanique soit aussi un attractant pour les moustiques. La présence d'un allèle délétère implique une moindre production de l'urocanase et par conséquent pourrait entrainer une accumulation d'acide urocanique. Dans notre étude les individus porteurs de l'allèle délétère sont plus à risque de développer une infection palustre, ce qui est compatible avec l'hypothèse ci-dessus. *UROC1* est associé à l'ensemble des infections, son action potentielle au niveau de la peau ou du foie est cohérente avec un rôle dans la phase précoce de la maladie.

4 MODELE LOGISTIQUE MIXTE POUR LA CORRECTION DE LA STRUCTURE DE POPULATIONS DANS LES GWAS

La GWAS sur les formes simples de paludisme nous a amené à nous intéresser à l'utilisation des modèles non-linéaires mixtes dans les GWAS, dont la complexité numérique les rend difficilement applicables à l'échelle du génome avec les moyens de calcul actuels. Nous nous sommes posé la question, à la suite, de la stratégie à adopter dans le cas le plus simple, une analyse d'un trait binaire (par exemple le statut cas/témoins pour une maladie) dans des cohortes telles que celles du Sud Bénin, présentant à la fois une structure (ou sous-structure) de population et un niveau d'apparentement substantiel entre les individus. Le modèle logistique mixte (ou régression logistique mixte, MLR), comme le modèle de Cox mixte représente un fardeau trop important en terme de calcul pour être applicable dans une GWAS, même en utilisant une solution approchée avec la PQL (pour *Penalized Quasi-Likelihood*). Pour un trait binaire, la solution en deux étapes mise en œuvre dans la GWAS ne peut être appliquée, car elle nécessite des données répétées afin d'estimer la fragilité (ou effet aléatoire individuel) des individus.

Notre étude a été motivée également par les travaux de Chen *et al.* , 2016 (Chen et al., 2016). La solution alternative largement utilisée jusqu'à présent était d'analyser le statut cas/témoins codé 1 et 0 respectivement comme un trait quantitatif avec un modèle linéaire mixte (MLM). Chen *et al.* ont montré, dans le cadre d'une GWAS sur l'asthme incluant plusieurs populations d'origine caribéenne et latino-américaine, qu'en présence d'une hétérogénéité de prévalence de la maladie entre les strates de populations, le MLM était inapproprié pour corriger la structure de population. Cette GWAS incluait en particulier environ 15% d'individus de Porto Rico pour lesquels la prévalence de l'asthme était beaucoup plus élevée (25,6%) que dans les autres populations (de 3,9 à 9,6%). Ils ont observé que le MLM conduisait à l'obtention de p-valeurs conservatrices pour certaines catégories de SNPs et anti-conservatrices pour d'autres, en fonction des fréquences alléliques dans les deux strates. Ils ont montré également que ces biais avec le MLM n'étaient pas diagnostiqués par un Q-Q

plot standard, communément utilisé pour attester d'une absence d'inflation liée à la structure de population. En effet, les p-valeurs étant anti-conservatrices pour des catégories de SNPs et conservatrices pour d'autres, la distribution globale des p-valeurs apparait correcte. Chen *et al.* ont mis en évidence ces biais par un Q-Q plot stratifié, où les distributions des p-valeurs sont représentées par catégories de SNPs en fonction de leur fréquence dans les deux strates et ont proposé GMMAT, un test du score pour la MLR permettant d'obtenir les p-valeurs dans les GWAS de manière efficace. Cependant, ce test ne permet pas d'estimer les effets des SNPs (i. e. les odds ratio), ce qui est un inconvénient également du MLM. Les odds ratio peuvent être estimés *a posteriori* pour les SNPs ou régions d'intérêt d'une GWAS avec un MLR utilisant la PQL, mais les effets des SNPs sur l'ensemble du génome sont parfois nécessaires comme dans le cas d'une méta-analyse.

Dans une première étude de simulation nous avons évalué la capacité du MLM, de GMMAT, et également de la régression logistique incluant les premières PCs, à corriger la structure de population dans la GWAS du Sud Bénin, en se mettant dans la même situation que Chen et *al.*, avec une forte hétérogénéité de prévalence entre les cohortes. Nous avons ensuite étudié le comportement de deux méthodes approchées permettant d'estimer les odds ratio (OR) dans la MLR à partir des données de la GWAS et d'un jeu de données de taille plus conséquente obtenu par simulation à l'aide d'un modèle de coalescence. Enfin, nous proposons une extension du Q-Q plot stratifié proposé par Chen *et al.* pour permettre le diagnostic d'une correction incomplète de la structure de population lorsque les deux strates de populations ne sont pas clairement définies *a priori*.

4.1 Adéquation des méthodes existantes pour l'analyse des données du Sud Bénin

4.1.1 Methodes

Evaluation de l'erreur de type I

L'erreur de type I a été évaluée pour la régression logistique (LR), le MLM et GMMAT, en simulant un phénotype binaire sous l'hypothèse d'absence d'effet des génotypes. Nous avons considéré une hétérogénéité de prévalence équivalente à celles de l'étude de Chen *et al.* (Chen et al., 2016), avec

une prévalence de 0,30 dans la strate avec un risque plus élevé (cohorte d'Allada) et une prévalence de 0,05 dans la seconde (cohorte de Tori-Bossito). Le trait binaire a été simulé avec un modèle de régression logistique mixte. Pour un individu *i*, la probabilité d'être malade a été calculée par :

$$logit\ P(Y_i = 1) = \alpha_0 + \alpha_1 Z_i + u_i \qquad (1)$$

avec

- $Z_i = 1$ si l'individu appartient à la cohorte avec un risque plus élevé ou $Z_i = 0$ sinon
- u_i l'effet aléatoire associé à l'individuel *i ;* le vecteur $u = (u_1 \cdots u_n)$ suit une loi $N(0, \tau \mathrm{K})$, avec K la matrice GRM, correspondant à un effet polygénique

Les coefficients α_0 et $\alpha_1$ont été définis de manière à obtenir des prévalences de 0,30 et 0,05 dans les deux cohortes sans tenir compte des effets aléatoires ($\alpha_0 = logit\ (0.05)$ et $\alpha_1 = logit\ (0.30) - logit\ (0.05)$). Les effets aléatoires ont été simulés avec $\tau = 1$.

Une fois le phénotype obtenu, l'association a été testée avec les SNPs de la puce HumanOmni5 ayant une MAF supérieure à 5% (n=1 847 505 SNPs), avec les différentes méthodes. Les analyses ont été réalisées dans un premier temps sans inclure de PCs dans les modèles puis en incluant les 10 premières PCs en effets fixes.

Le Q-Q plot stratifié

Nous avons utilisé le Q-Q plot stratifié tel que défini par Chen *et al.* pour évaluer l'erreur de type I avec les différentes méthodes testées. Ce Q-Q plot implique de définir deux strates *a priori.* Dans leur étude, les deux strates indexées par $i = 0$ ou $i = 1$ ont été définies en fonction de la prévalence de la maladie, $i = 1$ correspondant à la population avec un risque plus élevé et $i = 0$ aux autres populations. Trois catégories de SNPs sont ensuite définies en fonction du rapport de la variance des génotypes entre les deux strates. Soit *G* le génotype au SNP considéré, sous l'hypothèse de panmixie à l'intérieur des strates, la variance du génotype dans la strate *i* est égale à $var_i(G) = 2p_i q_i$, avec p_i et q_i les fréquences des allèles du SNP. Les catégories sont définies de la façon suivante en fonction de $r(G) = var_1(G)/var_0(G)$ et d'un seuil *th* = 0.8 :

- SNPs avec $r(G) < th$ (catégorie 1)
- SNPs avec $th \leq r(G) < 1/th$ (catégorie 2)
- SNPs avec $1/th \leq r(G)$ (catégorie 3)

L'extension du Q-Q plot stratifié

Nous avons ensuite proposé une extension de cette méthode permettant de diagnostiquer une correction incomplète de la structure de population, non plus à partir de l'information dans les deux strates (une variable binaire codé 0 ou 1) mais à partir d'une variable quantitative quelle qu'elle soit, dont les valeurs sont comprises dans l'intervalle [0,1]. Soit $G \in \{0,1,2\}^n$ le vecteur des génotypes, *Z* un vecteur de taille *n* dont les éléments sont inclus dans [0,1] et $\mathbf{1}$ un vecteur de uns, on peut poser :

$$q_1 = \frac{1}{2}\frac{Z'G}{Z'\mathbf{1}} \; et \; q_0 = \frac{1}{2}\frac{(\mathbf{1}-Z)'G}{(\mathbf{1}-Z)'\mathbf{1}}$$

Ces quantités correspondent aux fréquences alléliques dans les deux strates lorsque *Z* est la variable indicatrice de l'appartenance aux strates. On pose ensuite $p_i = 1 - q_i$ et $r(G) = (2p_1q_1)/(2p_0q_0)$. Les catégories de SNPs sont ensuite définies de la même manière. Le point important de cette extension est qu'elle permet d'utiliser les coordonnées des individus sur les premières PCs de l'ACP en l'absence d'information sur les strates ou lorsque plus de deux strates sont observées dans l'échantillon. Un diagnostic peut ainsi être réalisé à partir de l'information de chaque PC, prise individuellement, une fois les valeurs rapportées dans un intervalle de [0,1].

4.1.2 Résultats

Evaluation de l'erreur de type I

La Figure 4.1 présente les Q-Q plots stratifiés pour l'analyse d'un phénotype simulé avec une différence importante de prévalence entre les strates (prévalences de 0.05 et 0.30 dans les cohortes de Tori-Bossito et d'Allada respectivement) avec les différents modèles : régression logistique (LR), régression logistique mixte (MLR, test du score de Chen *et al.*) et modèle linéaire mixte (MLM). Les catégories de SNPs correspondent à celles définies par Chen *et al.* : les SNPs de la catégorie 1 (11,4%) ont des MAFs sensiblement inférieures dans la strate à plus haut risque (cohorte de d'Allada), les

SNPs de la catégorie 2 (77,6%) des MAFs similaires dans les deux strates et les SNPs de la catégories 3 (11,0%) des MAFs sensiblement plus élevées dans la strate à plus haut risque.

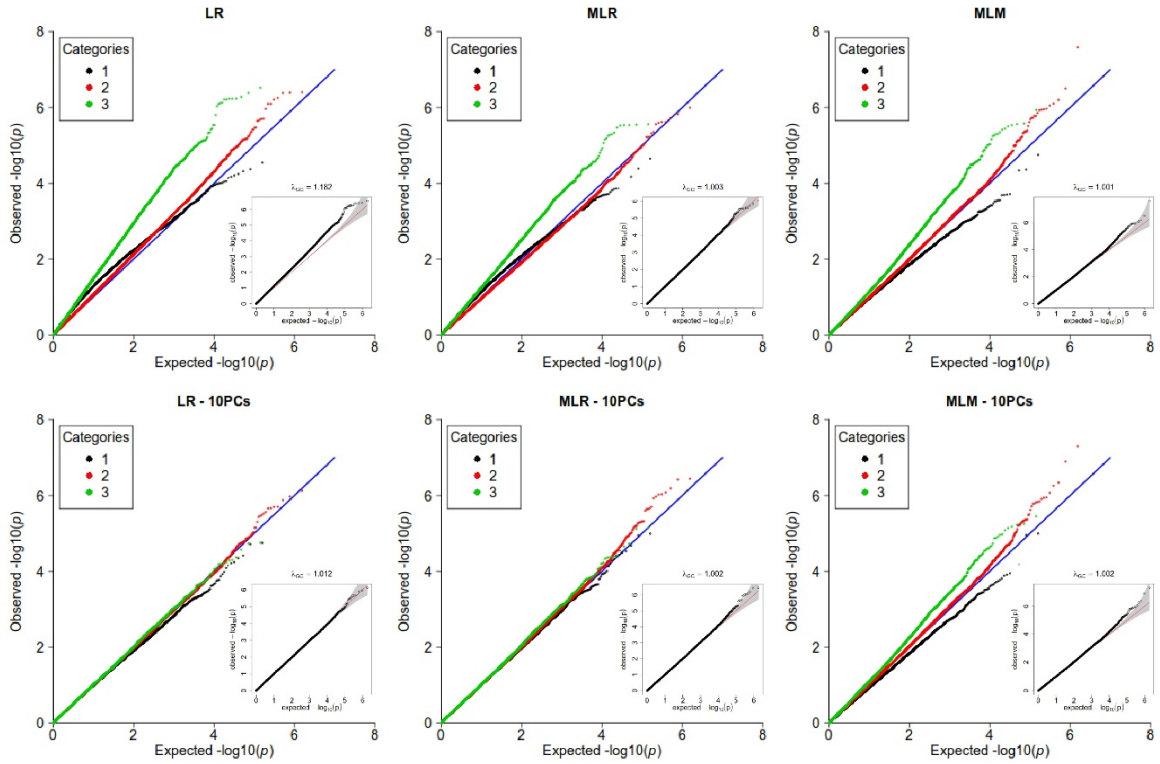


Figure 4.1 Q-Q plots stratifiés en absence d'effet des génotypes dans la cohorte du Sud Bénin
LR, régression logistique ; MLR, régression logistique mixte utilisant le test du score ; MLM, modèle linéaire mixte. Les catégories de SNPs correspondent à celles définies dans Chen et al. (2016). En vignette figure le Q-Q plot non stratifié correspondant

Lorsqu'aucune PC n'est incluse (première ligne de la figure), une inflation de la statistique est observée pour la LR (λ = 1,182). Sur la base du Q-Q plot non stratifié (en miniature), les deux méthodes MLM et MLR paraissent corriger la structure de la population; cependant, le Q-Q plot stratifié montre une inflation de la statistique pour les SNP dans les catégories 1 et 3 pour les deux modèles associé à une déflation de la statistique pour les SNPs de catégorie 1 dans le cas du MLM.

Lorsque 10 PCs sont incluses dans les modèles (deuxième ligne de la figure), cette différence de comportement entre les catégories de SNPs persiste pour le MLM mais la correction est adéquate pour toutes les catégories de SNPs pour la LR et la MLR.

Extension du Q-Q plot stratifié

Nous avons comparé les Q-Q plots stratifiés obtenus avec l'information sur les deux strates, tels que proposés par Chen *et al.,* avec ceux obtenus avec les coordonnées des individus sur les premières PC à partir de la méthode proposée. La Figure 4.2 montre les Q-Q plots pour les analyses avec un MLM et une MLR incluant 10 PCs pour les données de la GWAS (analyse sous l'hypothèse nulle de la section 4.2.1). Des différences entre les Q-Q plots peuvent être observées, cependant les deux Q-Q plots stratifiés conduisent à la même conclusion, c'est-à-dire une correction adéquate dans le cas de la MLR et une inflation/déflation des p-valeurs dans le cas du MLM.

Une comparaison similaire a été réalisés pour la cohorte simulée de 10 000 individus (Figure supplémentaire 2 de l'article 2). Pour cette cohorte qui présente une structure de population plus complexe, l'inadéquation du MLM peut être diagnostiquée à partir des données de la première PC mais également à partir des données de la deuxième PC.

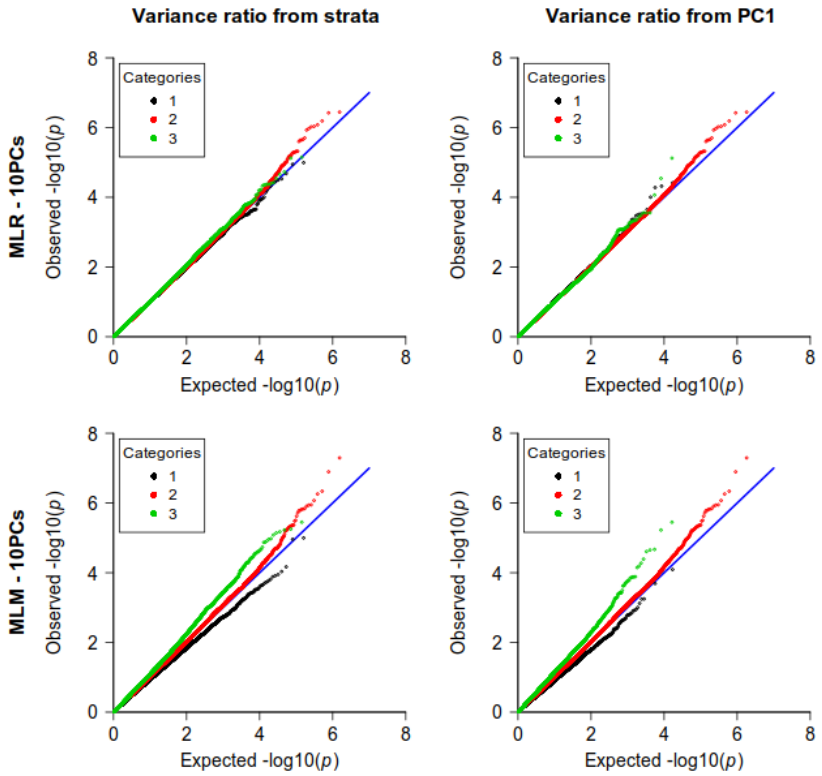


Figure 4.2 Q-Q plots stratifiés obtenus à partir de deux strates (à gauche) et à partir des coordonnées de la première PC (à droite) pour les simulations basées sur les données de la GWAS

4.2 Méthodes proposées pour l'estimation des effets des SNPs

4.2.1 Les méthodes AMLE et Offset

Pour tester l'association à chaque SNP avec une MRL, on considère le modèle suivant:

$$logit\ P(Y_i = 1) = X_i\beta + G_i\gamma + u_i \qquad (2)$$

avec

- β le vecteur des effets associés aux covariables et *γ* l'effet du SNP que l'on cherche à estimer
- X_i le vecteur des covariables pour l'individu *i*,
- u_i l'effet aléatoire associé à l'individu *i* ; le vecteur $u = (u_1 \cdots u_n)$ suit une loi $N(0, \tau \mathrm{K})$, avec K la GRM

Le calcul de la vraisemblance de ce modèle nécessite une intégration numérique et serait extrêmement lourd en temps de calcul et l'approche classique consiste à utiliser une solution approchée, donnée par l'algorithme de la PQL (pour *Penalized Quasi-Likelihood*) qui est une séquence d'approximations du MLR par un MLM. Cependant bien que beaucoup moins lourde, la PQL reste trop coûteuse en temps de calcul pour permettre l'estimation des effets des SNPs dans une GWAS.

Nous avons étudié le comportement de deux méthodes approchées pour obtenir les effets des SNPs de manière rapide. La première est basée sur une approximation du maximum de vraisemblance du modèle estimée par la PQL (AMLE pour *Approximate Maximum Likelihood Estimate*) et est analogue à celle implémentée dans SAIGE (Zhou et al., 2018). La seconde (Offset) consiste à estimer, une seule fois, les effets individuels (fixes et aléatoires) dans un modèle logistique mixte qui n'inclut pas les SNPs, puis à estimer les effets des SNPs dans un modèle logistique incluant les effets individuels sous forme d'un offset.

La méthode AMLE

On note $\kappa = (\beta, \tau)$ le vecteur des paramètres de nuisance et $\ell(\kappa, \gamma)$ la vraisemblance du modèle (2). Dans un premier temps, on calcul $\hat{\kappa}_0$ qui maximise $\ell(\kappa, 0)$ sous l'hypothèse nulle d'absence d'effet

du SNP par l'estimateur du maximum de vraisemblance (EMV). Cette estimation avec la PQL est réalisée une seule fois pour l'ensemble des tests de la GWAS.

Ensuite, on fixe $\kappa = \hat{\kappa}_0$; et pour chaque SNP, $\ell(\hat{\kappa}_0, \gamma)$ est approchée au second ordre par :

$$\ell(\hat{\kappa}_0, \gamma) \simeq \ell(\hat{\kappa}_0, 0) + \frac{\partial}{\partial \gamma}\ell(\hat{\kappa}_0, 0)\gamma + \frac{1}{2} \cdot \frac{\partial^2}{\partial \gamma^2} \ell(\hat{\kappa}_0, 0)\gamma^2$$

qui est maximale pour :

$$\hat{\gamma} = -\frac{\frac{\partial}{\partial \gamma}\ell(\hat{\kappa}_0, 0)}{\frac{\partial^2}{\partial \gamma^2}\ell(\hat{\kappa}_0, 0)}$$

Cet estimateur de γ ne correspond pas à l'EMV, mais il est proche de celui-ci lorsque les effets ne sont pas trop grands ou lorsque $\hat{\kappa}$ et $\hat{\gamma}$ sont indépendants.

La méthode offset

Comme pour la première méthode, la méthode offset est basée sur une seule estimation de la vraisemblance du MLR, sous l'hypothèse nulle d'absence d'effet du SNP. La première étape est identique ; les effets fixes et aléatoires sont estimés par $\hat{\beta}_0$ et $\hat{u}_0$ dans le modèle (2) où $\gamma = 0$.

Ensuite, pour chaque SNP, l'association est testée avec un modèle logistique incluant uniquement des effets fixes, dans lequel on intègre les vecteurs $\hat{\beta}_0$ et $\hat{u}_0$ comme des offsets. Cette deuxième étape consiste, pour chaque SNP :

- à calculer $\tilde{G}$ le résidu de la régression linaire de *G* par *X* (la matrice des covariables)
- à estimer γ avec un modèle logistique standard

$$logit\, P(Y_i = 1) = X_i\hat{\beta}_0 + \hat{u}_{0i} + \tilde{G}_i\gamma$$

Cette approche incluant une étape intermédiaire de régression du génotype par les covariables est motivée par le fait qu'une approche similaire appliquée avec un modèle linéaire standard permet d'obtenir une estimation $\hat{\gamma}$ identique à l'estimation de γ dans un modèle linéaire en une seule étape $Y_i = X_i\beta + G_i\gamma$.

L'avantage des deux méthodes que nous proposons est que de la même manière que pour le test du score, elles ne nécessitent qu'une seule estimation de la vraisemblance du modèle sous H_o pour l'ensemble de la GWAS, ce qui réduit considérablement le temps de calcul.

4.2.2 Evaluation du comportement des méthodes approchées

Les propriétés des méthodes (erreur de type I, puissance et biais d'estimation des effets) ont été évaluées à partir des données génétiques de la GWAS, puis à partir d'un jeu de données simulées afin de valider les résultats sur une seconde cohorte de taille plus importante.

Simulation d'une cohorte basée sur un modèle de coalescence

L'objectif de ces simulations était d'obtenir des données génétiques pour un large échantillon d'individus présentant à la fois une stratification de population et des individus apparentés. Ces simulations ont été réalisées avec le logiciel *ms* (Hudson, 2002) sur le modèle de celles réalisées par Chen *et al.* : la procédure décrite a été scrupuleusement suivie, les même paramètres ont été utilisés. Ces simulations utilisent une grille où des sous-populations sont générées par cellule et sont basées sur un modèle de *stepping stone* avec des migrations symétriques entre les cellules adjacentes de la grille. Cette procédure est utilisée couramment pour simuler les génotypes d'une population avec une structure spatiale continue (Bradburd et al., 2016; Mathieson & McVean, 2012). Les paramètres utilisés (une grille 20 x 20 et un taux de migration entre les cellules adjacentes de 10) permettent d'obtenir un indice de fixation de Wright (Fst) de l'ordre de 0.01 lorsque la grille est divisée en deux parties égales, un niveau équivalent à celui qui est observé en Europe (Mathieson & McVean, 2012). Dans un premier temps, un total de 10 millions de SNPs indépendants pour 8 000 individus ont été simulés. Ensuite, pour obtenir des individus apparentés, une cohorte d'enfants a été simulée à partir de ces individus « fondateurs ». Au sein de chaque cellule, des couples d'individus ont été formés de manière aléatoire (10 couples par cellule) et pour chaque couple deux enfants ont été simulés par *gene dropping*. Nous avons ainsi obtenus un échantillon de 16 000 individus (8 000 fondateurs et 8 000 enfants), à partir duquel 10 000 individus ont été extrait de manière aléatoire pour constituer la cohorte.

Deux strates ont été définies en fonction de la position des individus sur la grille, les individus situés dans le quart supérieur gauche de la grille appartenant à la population à plus haut risque (Figure 4.3).

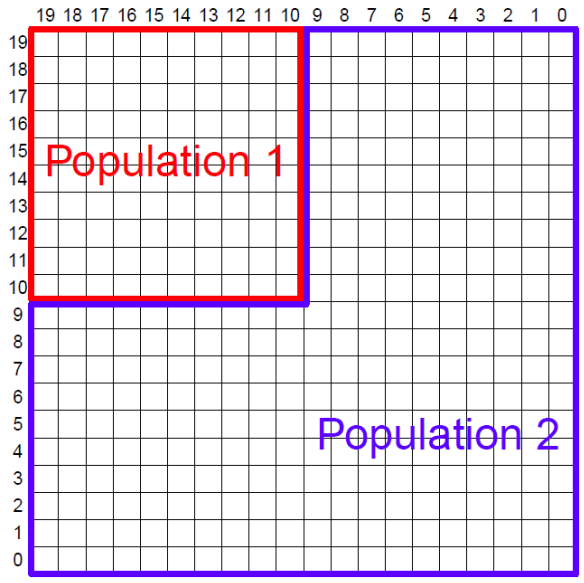


Figure 4.3 Grille utilisée pour simuler les génotypes avec une structure spatiale continue
Le carré rouge en haut à gauche indique la strate de population avec un risque plus élevé. Source : Chen et al., 2016

Evaluation de l'erreur de type I

L'erreur de type I pour les deux méthodes proposées a été évaluée de la même manière que pour les précédentes méthodes (section 4.1.1. *Evaluation de l'erreur de type I*) avec la simulation d'un phénotype dont la prévalence dépend de la strate.

Pour les simulations réalisées à partir des données de GWAS du Sud Bénin, le test de Wald pour la méthode AMLE étant identique à GMMAT, les résultats obtenus pour celui-ci à la section 4.1.2 s'applique également pour la méthode AMLE. Pour la méthode offset, nous avons réutilisé le phénotype simulé précédemment à la section 4.1.1.

Pour la cohorte obtenue à partir d'un modèle de coalescence, un phénotype a été simulé de manière similaire en utilisant l'équation (1) et $Z_i = 1$ si l'individu appartient à la partie supérieure gauche de la grille. La seule différence est que nous avons utilisé la matrice φ des coefficients de parenté pour obtenir un phénotype dont la structure de corrélation dépend des relations familiales ($K = 2\varphi$, où φ

est La matrice des corrélations génotypiques attendues entre les individus, i. e. 0.5 pour les apparentés du premier ordre, 1 sur la diagonale et 0 sinon). L'association a ensuite été testée avec les SNPs ayant une MAF >0.05 (2 840 903 SNPs), en utilisant comme matrice de variance-covariance, la GRM calculée à partir de 100 000 SNPs tirés aléatoirement.

Estimation des effets des SNPs

Une seconde série de simulations a été réalisée à partir des données génétiques de la GWAS pour évaluer les biais des estimations des SNPs. Le modèle pour simuler le phénotype inclut cette fois l'effet d'un SNP :

$$logit\ P(Y_i = 1) = \alpha_0 + \alpha_1 Z_i + G_i\gamma + u_i\ ,$$

avec G_i le génotype de l'individu *i* pour un SNP tiré aléatoirement dans les données, et Z_i et u_i tels que définit dans le modèle (1). L'effet du SNP utilisé pour simuler le phénotype a ensuite été testé avec les deux méthodes et avec un MLR utilisant la PQL, qui constitue le *gold standard* actuel.

Différents scenarios ont été explorés:

A. Un effet cohorte modéré ($p_0 = 0{,}10$ et $p_1 = 0{,}20$) et un effet polygénique modéré ($\tau = 0{,}30$)

B. Un effet cohorte modéré ($p_0 = 0{,}10$ et $p_1 = 0{,}20$) et un effet polygénique important ($\tau = 1$)

C. Un effet cohorte important ($p_0 = 0{,}05$ et $p_1 = 0{,}30$) et un effet polygénique modéré ($\tau = 0{,}30$)

Les coefficients α_0 et α_1 ont été calculés comme dans les premières simulations par respectivement $logit\ (p_0)$ et $logit\ (p_1) - logit\ (p_0)$; le vecteur de génotype *G* est centré pour obtenir dans chaque strate les prévalences attendues. Pour chaque scénario, les simulations ont été réalisées en tirant aléatoirement le SNP dans 3 intervalles de MAF différents (0,05 ;0,10], (0,20 ;0,25] et (0,45 ;0,50] et pour un effet du SNP $\gamma = \log(1{,}5)$ et $\gamma = \log(2)$ correspondant respectivement à $OR = 1{,}5$ et 2. Les simulations ont été répétées 100 fois dans chaque condition en redéfinissant à chaque fois un nouveau vecteur d'effets aléatoires.

Comparaison des puissances

Des simulations additionnelles ont été réalisées sur le modèle des précédentes (paragraphe ci-dessus *Estimation des effets des SNP*) pour comparer la puissance des méthodes présentant une erreur de type I correcte. Deux scénarios avec ou sans effet cohorte ont été envisagés avec à chaque fois un effet polygénique aléatoire important ($\tau = 1$). Pour les simulations basées sur les données de GWAS, nous avons considéré un $OR = 3$ avec un effet cohorte modéré ($p_0 = 0{,}10$ et $p_1 = 0{,}20$) et un $OR = 2{,}5$ sans effet cohorte. Les simulations basées sur la cohorte obtenue par un modèle de coalescence, ont été réalisées sur un sous-échantillon de 5 000 individus, avec un $OR = 1{,}5$ soit avec un effet cohorte important ($p_0 = 0{,}05$ et $p_1 = 0{,}30$) soit sans effet cohorte. Les puissances ont été estimées à partir de 1 000 réplicats.

4.2.3 Résultats

Erreur de type I

Des Q-Q plots similaires ont été réalisés pour les analyses avec les deux méthodes proposées. La Figure 4.4 présente les Q-Q plots stratifiés obtenus pour les analyses incluant 10 PCs pour les deux jeux de données, les données de la GWAS (à gauche) et celle de la cohorte simulée de 10 000 individus (à droite). La méthode AMLE (en haut) corrige de manière adéquate la structure de population dans les deux jeux de données. Les résultats obtenus pour les données de la GWAS sont strictement identiques à ceux de la MLR à la section 4.2.1, le test de Wald pour la méthode AMLE étant identique au test du score pour la MLR. La méthode offset corrige de manière adéquate la structure de population dans le cas des données de la GWAS mais est légèrement trop conservatrice dans le cas de la cohorte simulée, quelles que soient les catégories de SNPs.

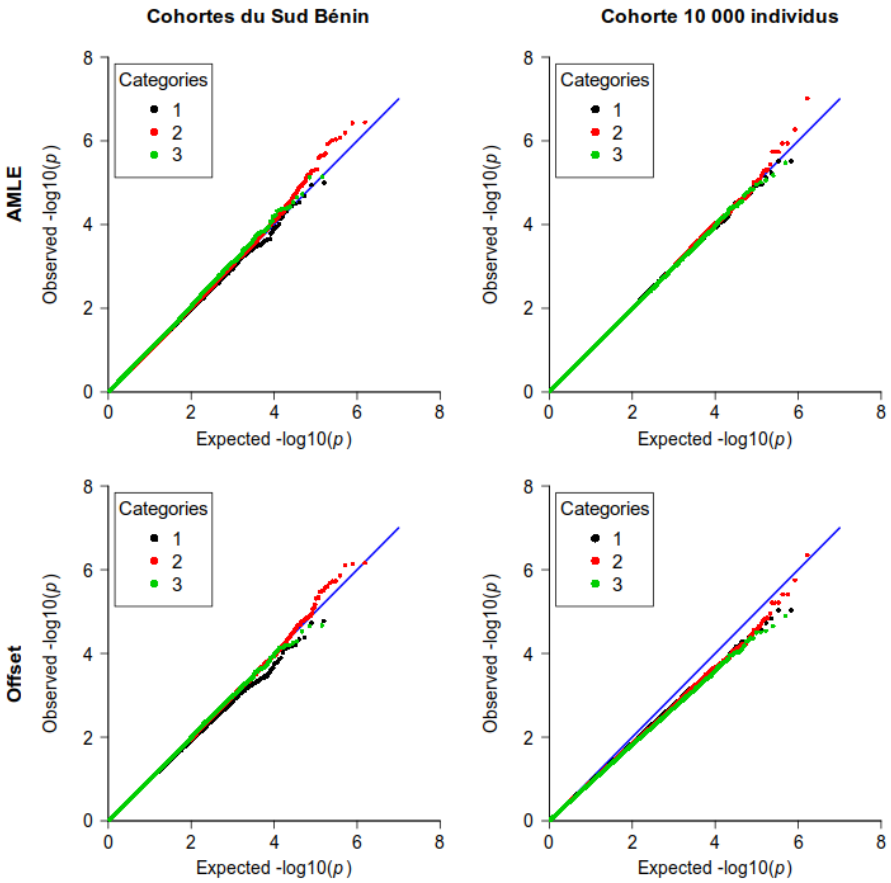


Figure 4.4 Q-Q plots stratifiés pour les méthodes AMLE et Offset, pour les simulations basées sur les données du Sud-Bénin (à gauche) et celle obtenues avec un modèle de coalescence (à droite).
Les catégories de SNPs correspondent à celles définies dans Chen et al. (2016)

Estimation de l'effet des SNPs

La figure 4.5 montre les biais d'estimation des effets des SNPs ($\hat{\gamma} - \gamma$) des méthodes AMLE et Offset comparés à ceux de la PQL. Ils ont été évalués pour deux valeurs de γ simulé (équivalant à des ORs de 1,5 et 2) dans trois scénarios (A, B, C) correspondant à des niveaux différents de l'effet cohorte et de l'effet polygénique aléatoire. Dans ces simulations réalisées à partir des données de la GWAS, la PQL ne montre aucun biais dans pratiquement tous les scénarios quelle que soit la MAF du SNPs. On peut noter un très léger biais négatif dans le scenario B où l'effet polygénique est important.

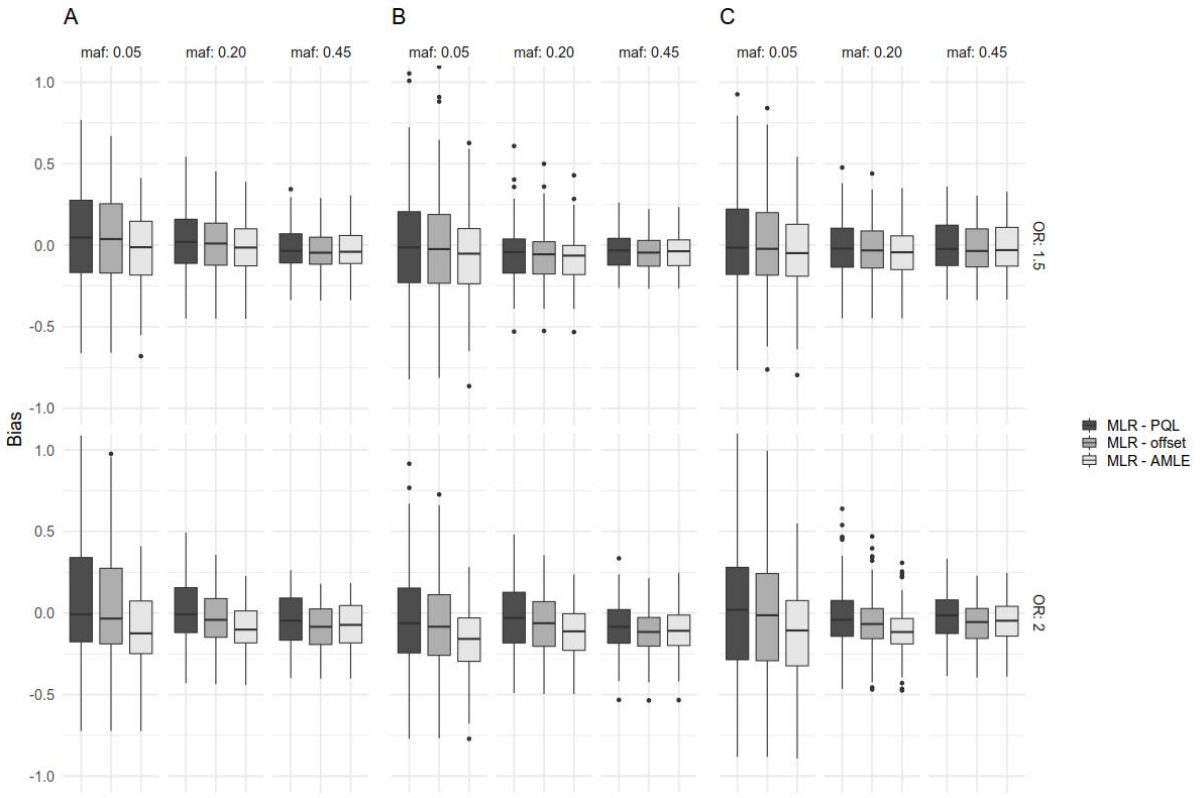


Figure 4.5 Biais d'estimation de l'effet des SNPs dans trois scenarios: *A) effet cohorte et effet polygénique modéré; B) effet cohorte modérée et effet polygénique important; C) effet cohorte important, effet polygénique modéré. Les biais pour $\hat{\gamma}$ ont été estimés pour deux valeurs de γ (équivalent à un OR=1.5 et un OR=2) et trois intervalles de MAF (0.05-0.10, 0.20-0.25, 0.45-0.50)*

Les méthodes AMLE et offset ont tendance à présenter un biais négatif comparé à la PQL. Ce biais est minime et est du même ordre de grandeur pour les deux méthodes lorsque l'effet du SNP est modéré ($OR = 1{,}5$). Dans les différents scénarios, on observe un biais moyen maximum de -0.08, ce qui correspond à un OR estimé de 1.4.

Lorsque l'effet du SNPs augmente (OR = 2), des biais plus importants sont observés quel que soit le scénario. Ces biais ont tendance à être plus élevés pour la méthode AMLE. Ils peuvent atteindre dans certaine situation -0.1, correspondant à un OR estimé de 1.8. Les biais apparaissent légèrement plus importants pour les MAF faibles (comprises entre 0.05 et 0.10).

Comparaison de puissance des méthodes

Les puissances ont été comparées pour les méthodes LR, AMLE et Offset dans les données de la GWAS (Table 4.1). Ces différentes méthodes ont montré une erreur de type I correcte quelles que

soient les catégories de SNPs lorsque 10 PCs étaient incluses en effets fixes. La méthode AMLE (équivalent au test du score) est légèrement plus puissante que les autres dans les différents scenarios évalués. La LR en particulier apparait moins puissante en présence d'une hétérogénéité de prévalence entre les strates.

Scenario	p_0	p_1	OR	MAF bin	LR	AMLE	Offset
Effet cohorte modéré	0.10	0.20	3	(0.20 ; 0.25]	0.179	0.261	0.268
Effet cohorte modéré	0.10	0.20	3	(0.45 ; 0.50]	0.899	0.920	0.906
Pas d'effet cohorte	0.30	0.30	2.5	(0.20 ; 0.25]	0.473	0.496	0.458
Pas d'effet cohorte	0.30	0.30	2.5	(0.45 ; 0.50]	0.926	0.928	0.913

Table 4.1 Puissances des méthodes dans les simulations basées sur les données du Sud Bénin
p0 et p1, les prévalences dans les deux strates ; OR, odds ratio correspondant à l'effet du SNP simulé ; MAF bin, intervalle de fréquence du SNP.
Toutes les analyses incluent 10 PCs. Les puissances ont été calculées au seuil de signification des GWAS (5 x 10-8) pour la régression logistique (LR),et les méthodes AMLE (équivalent au test du score) et Offset

Les résultats pour la cohorte simulée sont présentés dans la table ci-dessous (Table 4.2). La LR n'a pas été incluse dans la comparaison de puissance car une forte inflation de l'erreur de type I a été observée (simulations non présentées). La méthode AMLE est également légèrement plus puissante que la méthode Offset, ce qui est attendu vu que la méthode offset est conservatrice dans cet échantillon.

Scenario	p_0	p_1	OR	MAF bin	AMLE	Offset
Effet cohorte important	0.05	0.30	1.5	(0.20 ; 0.25]	0.251	0.202
Effet cohorte important	0.05	0.30	1.5	(0.45 ; 0.50]	0.544	0.464
Pas d'effet cohorte	0.30	0.30	1.5	(0.20 ; 0.25]	0.859	0.834
Pas d'effet cohorte	0.30	0.30	1.5	(0.45 ; 0.50]	0.989	0.987

Table 4.2 Puissances des méthodes dans les simulations basées sur données issues d'un modèle de coalescence
p_0 et p_1, les prévalences dans les deux strates ; OR, odds ratio correspondant à l'effet du SNP simulé ; MAF bin, intervalle de fréquence du SNP.
Toutes les analyses incluent 10 PCs. Les puissances ont été calculées au seuil de signification des GWAS (5 x 10^{-8}) pour les méthodes AMLE (équivalent au test du score) et Offset

4.3 Discussion

Un premier objectif de ce travail était de regarder si les résultats observés par Chen *et al.* avec le MLM pour l'analyse d'un trait binaire pouvaient être retrouvés dans une étude sur des populations africaines comme celle du Sud Bénin. Alors que leur GWAS incluait des individus de différentes populations d'origine caribéenne et latino-américaine, les deux cohortes du Sud Bénin sont géographiquement proches, les deux sites d'étude étant distants de 20 km et présentant une sous-structure génétique liée aux sites d'étude. En simulant un phénotype avec une différence de prévalence identique à celle observée dans leur étude, nous avons retrouvé des résultats similaires, c'est à dire des p-valeurs conservatrices pour les SNPs avec des MAFs sensiblement inférieures dans la strate à plus haut risque et des p-valeurs anti-conservatrices pour les SNPs avec des MAFs sensiblement plus élevées dans cette même strate. Une hétérogénéité de prévalence peut résulter de facteurs environnementaux (facteurs nutritionnels, comportementaux, d'exposition à la maladie, *ect.*) et peut ne pas être si rare, en pratique, dans les GWAS. Dans le cas d'une diversité génétique substantielle, si ces facteurs ne sont pas pris en compte dans l'analyse, le MLM conduit à une analyse incorrecte de l'association. Les analyses sur les données de ces deux cohortes montrent qu'en Afrique les différences de fond génétique observées à une échelle locale peuvent être une source de biais dans les analyses d'association.

Les simulations réalisées sur les données de la GWAS montrent que lorsque les premières PCs sont incluses dans les analyses, le MLR et la régression logistique standard corrigent de manière adéquate la structure de population. Aussi, une analyse logistique standard incluant des PCs peut être envisagée pour l'analyse d'un trait binaire dans l'étude du Sud Bénin. Cette méthode, plus simple, est suffisante pour corriger la structure de population bien qu'un niveau d'apparentement substantiel soit observé dans l'échantillon. Cependant l'étude de puissance montre qu'elle tend à être moins puissante que le MLR, en particulier lorsqu'il existe une différence de prévalence entre les strates. Des simulations similaires réalisées sur les données de la cohorte de 10 000 individus (que nous

n'avons pas présentées ici mais qui figurent dans l'article 2), montrent par ailleurs que dans le cas d'un niveau d'apparentement plus important (avec des apparentés au premier degré) le MLR est la seule méthode adéquate pour corriger la structure de population. On peut noter également qu'il est nécessaire d'inclure les premières PCs comme des effets fixes dans le MLR, en plus des effets aléatoires, comme préconisé par d'autres auteurs auparavant (Price et al., 2010; Y. Zhang & Pan, 2015).

Concernant le diagnostic d'une erreur de type I incorrecte, Chen *et al.* ont montré qu'un Q-Q plot standard n'était pas suffisant en présence d'une hétérogénéité de prévalence entre les strates. Le Q-Q plot stratifié qu'ils proposent est basé sur la définition de deux strates de population. L'extension proposée ici permet de réaliser un Q-Q plot stratifié à partir de l'information des premières PCs. Les Q-Q plot réalisés dans nos simulations conduisent à un diagnostic similaire au Q-Qplot réalisé avec l'information des deux strates, et devrait permettre de faciliter et d'améliorer le diagnostic d'une correction incomplète de la structure de population.

Les études de simulations montrent que les deux méthodes, AMLE et Offset proposées pour estimer les effets des SNPs sur l'ensemble du génome présentent de bonnes propriétés. Pour la méthode AMLE, les mêmes conclusions que pour le MLR peuvent être tirées concernant l'erreur de type I, le test de Wald pour la méthode AMLE étant équivalent au test du score de Chen (Chen et al., 2016). Cette méthode présente donc une erreur de type I correcte pour les deux jeux de données. La seconde méthode montre des performances similaires dans les données de la GWAS du Sud Bénin, mais est légèrement trop conservatrices dans le jeu de données simulées à partir d'un modèle de coalescence, comportant un niveau d'apparentement important. Concernant l'estimation des effets des SNPs, un léger biais négatif est observé pour les deux méthodes dans les données du Sud Bénin. Celui-ci apparait plus important pour la méthode AMLE que pour la méthode Offset, alors que la PQL est non-biaisée dans ces données. Il est connu qu'une hétérogénéité dans les données non prise en compte induit des biais négatifs dans la régression logistique (Cramer, 2007; Gail et al., 1984), aussi

ces résultats laissent supposer que l'hétérogénéité entre les strates de population n'est pas complétement prise en compte par les deux méthodes approchées. Cependant, ces biais ne sont notables que pour des ORs relativement élevés pour les GWAS (OR=2), ce qui devrait avoir un impact négligeable. L'étude de puissance montre que la méthode AMLE (équivalente au MLR) est légèrement plus puissante dans tous les scénarios et en particulier pour le jeu de données simulées où une déflation de la statistique est observée pour la méthode Offset.

Comme nous l'avons mentionné précédemment (section 4.2.1), la méthode AMLE est analogue à celle implémentée dans SAIGE (Zhou et al., 2018). Zhou *et al.*, ont présenté cette méthode sans justification de la formule, ni simulations pour évaluer les biais des estimations, dans le cadre des analyses sur de grand échantillons tels que celui de la UKBiobank, où le rapport cas/témoins est très déséquilibré. Il faut noter cependant que si la méthode AMLE est analogue, elle n'est pas adaptée pour l'analyse d'échantillons très déséquilibrés. En effet SAIGE, inclut d'autres développements méthodologiques spécifiques à ce type d'échantillon.

L'ensemble de ces méthodes ont été implémentées par Hervé Perdry dans un paquet R milorGWAS disponible sur le CRAN.

Les méthodes proposées ici pour estimer les effets des SNPs s'appuient sur des concepts simples qui peuvent être appliqués à d'autres modèles. L'utilisation de méthodes équivalentes pour le modèle de Cox Mixte permettrait par exemple de réaliser la GWAS sur la récurrence des infections en une seule étape sans estimer un phénotype intermédiaire de fragilité.

.

5 IDENTIFICATION DES SIGNAUX DE SELECTION NATURELLE RECENTE

Une des raisons principales pour lesquelles l'approche d'association génome entier n'avait pas encore été appliquée sur les phénotypes liés au paludisme simple est la difficulté de constituer un échantillon de taille suffisante. En effet, les études sur les formes simples de paludisme impliquent un suivi longitudinal en population, long, difficile à mettre en œuvre et coûteux, ce qui limite la taille des échantillons. Aussi, le paludisme ayant constitué une des plus fortes pressions de sélection que l'Homme ait connue dans son histoire récente, un des objectifs initiaux de l'étude génétique sur les cohortes de Tori-Bossito et d'Allada était d'explorer la possibilité d'intégrer l'information sur les signaux de sélection naturelle pour permettre d'augmenter la puissance des analyses d'association.

Ayodo *et al.* (Ayodo et al., 2007) ont fourni la preuve de concept que combiner l'information de la sélection naturelle avec celle des tests d'association pouvait permettre d'augmenter la puissance des analyses d'association. Dans leur étude sur le paludisme grave, incluant environ 500 cas et 500 témoins appartenant au groupe ethnique Luo du Kenya, ils ont étudié 10 variants génétiques précédemment associés à la résistance au paludisme grave. L'étude sur la sélection naturelle a été réalisée en calculant pour chaque variant le niveau de différenciation génétique entre populations exposées au paludisme (leur échantillon Luo ainsi que les Yoruba du projet international HapMap) et non exposées (des échantillons de témoins des ethnies Maasaï et Kikuyu qu'ils ont collectés au Kenya). Alors que l'analyse d'association seule met en évidence l'effet d'un seul gène après correction pour les tests multiples (*HBB* rs334, $p=8,0 \times 10^{-4}$) et d'un autre au seuil nominal de 0,05 (*CD36* rs3211938, p=0,03), la combinaison de l'information de sélection naturelle avec celle du test d'association a permis de diminuer les p-valeurs d'un ordre de magnitude de 1 à 2 ($p= 1,8 \times 10^{-5}$ et $p=4,3 \times 10^{-4}$ pour les variants dans *HBB* et *CD36* respectivement). Dans le cas des GWAS en particulier, le gain de puissance lié à la prise en compte de l'information de sélection naturelle peut permettre de mettre en évidence des associations qui n'auraient pas atteint sinon le seuil de signification très exigeant des GWAS.

Pour étendre cette approche à l'ensemble du génome et détecter les signaux de sélection naturelle dans notre population, nous nous sommes focalisés sur les méthodes basées sur les haplotypes longs. Ces méthodes sont efficaces pour mettre en évidence les signaux de sélection positive ou de sélection balancée récents, et l'on suppose que la pression exercée par le paludisme a été particulièrement intense au cours des 10 000 dernières années. Ces méthodes apparaissent également adaptées pour mettre en évidence les signaux de sélection liés au paludisme, celles-ci ayant permis de détecter des signaux de sélection naturelle dans les principaux gènes associés au paludisme : *HBB* (Hanchard et al., 2006), *G6PD* (Sabeti et al., 2002), *CD40LG* (Sabeti et al., 2002), *CD36* (Fry et al., 2009; International HapMap Consortium, 2005; Pybus et al., 2014).

Pour identifier les loci soumis à la sélection, nous avons utilisé deux méthodes complémentaires : l'iHS (Voight et al., 2006) qui a une bonne puissance pour détecter les balayages sélectifs incomplets (fréquence de l'allèle favorable entre 0,50 et 0,80) et l'XP-EHH (Sabeti et al., 2007) qui est plus puissant pour détecter les balayages sélectifs proches de la fixation (fréquence de l'allèle favorable > 0,80). Ces deux méthodes ont été développées avec l'hypothèse d'un balayage de type *hard sweep* où la sélection agit sur une seule mutation nouvellement apparue dans la population. Dans le cas du paludisme, la pression de sélection est présente depuis suffisamment longtemps pour s'attendre à de tels balayages sélectifs. Cependant on ne peut pas exclure que des modifications du climat au cours du temps aient pu entrainer des discontinuités dans l'exposition des populations au paludisme, pouvant avoir donné lieu à une sélection de variants déjà présents à une certaine fréquence dans la population ; de plus la sélection naturelle a pu agir sur plusieurs variants bénéfiques dans un même locus entrainant des balayages sélectifs de type *soft sweep*. Aussi nous avons utilisé également le test nS_L (Ferrer-Admetlla et al., 2014) développé plus récemment qui permet de mieux détecter les signaux de type *soft sweep*.

Les résultats de l'étude de sélection naturelle sont présentés ici, avec une liste des régions candidates montrant les plus forts niveaux de preuves de sélection naturelle dans les deux cohortes

du Sud Bénin. Nous avons ensuite croisé les informations de sélection naturelle et d'association afin d'identifier les régions où les signaux co-localisent. Dans cette étude nous n'avons pas observé d'enrichissement en signaux de sélection naturelle dans les régions trouvées associées au paludisme dans la GWAS. Nous avons poursuivi l'analyse en explorant de manière plus approfondie les régions présentant les plus hauts niveaux de preuve que ce soit pour l'association ou la sélection naturelle.

5.1 Méthodes

5.1.1 Phasage des données génotypiques et filtre des données

Les analyses de sélection naturelle ont été réalisées à partir des données de la puce HumanOmni5-4v1 et plus précisément à partir des fichiers VCF préparés pour l'imputation, incluant 2 539 532 SNPs (après le contrôle qualité et l'alignement des SNPs sur le génome de référence, Figure 2.4). Les données ont été phasées en utilisant l'ensemble des individus de l'échantillon (n=800) avec SHAPEITv2 (Delaneau et al., 2013) sans utiliser de panel de référence. Les différentes étapes de la préparation des fichiers, avec les outils et logiciels utilisés sont résumées dans les Figure 5.1 et 5.2 pour les tests intra-populationnels et les tests XP-EHH, respectivement.

Les analyses de sélection naturelle ont été réalisées sur un sous-échantillon d'individus non apparentés (n=638) identifiés avec KING (Manichaikul et al., 2010), en considérant les deux cohortes ensemble (Figure 5.1). En effet, l'analyse de la structure de la population a montré que les deux cohortes sont proches d'un point de vue génétique lorsque l'on se positionne à l'échelle des populations d'Afrique de l'Ouest (les deux populations se superposent sur le premier plan de l'ACP lorsque l'on inclut les populations KGP, section 2.3.2 Figure 2.6). Aussi, il semble raisonnable de les considérer comme un groupe homogène et de rechercher des signaux de sélection communs. KING identifie trente-trois paires d'apparentés au second degré et 172 paires d'apparentés au troisième degré. Les individus présentant un niveau d'apparentement jusqu'au troisième degré avec un autre individu de l'échantillon ont été exclus, la taille de l'échantillon étant suffisamment grande pour permettre un seuil strict.

Pour l'ensemble des tests de sélection, nous avons également filtré les SNPs sur la MAF pour ne garder que ceux avec une MAF>0,05, comme préconisé par Grossman *et al.*(Grossman et al., 2013) pour éviter que des variants rares ne viennent briser prématurément un haplotype long dans les balayages sélectifs proches de la fixation.

5.1.2 Tests de sélection naturelle positive ou balancée récente

Nous avons utilisé le logiciel selscan v1.2.0 (Szpiech & Hernandez, 2014) pour réaliser les tests de sélection sur l'ensemble du génome.

Tests intra-populationnels

Les tests iHS et nS_L ont été calculés pour les variants avec une MAF>0,05 (1 760 953 SNPs) en utilisant les paramètres par défaut dans selscan. Pour le test iHS, les distances génétiques ont été estimées avec la carte génétique HapMap phase II b37. Les deux statistiques ont été standardisées à l'intérieur de 100 *bins* de fréquence allélique avec le programme *norm* associé à selscan. Pour identifier les signaux de sélection significatifs, nous avons utilisé la distribution empirique des scores sur l'ensemble du génome. Nous avons considéré comme valeurs extrêmes les scores appartenant aux premier et 99^{eme} centiles de la distribution (correspondant à une valeur absolue du score supérieure à 2,60 pour iHS et supérieure à 2,54 pour nS_L). Le génome a ensuite été découpé en fenêtres de 100 Kb non chevauchantes et la proportion de valeurs extrêmes à l'intérieur de chaque fenêtre a été calculée, en excluant les fenêtres avec moins de 20 SNPs. Nous avons considéré comme signal significatif les 1% (puis 5%) des fenêtres avec les plus fortes proportions de scores extrêmes. Cette procédure, utilisant la proportion de scores élevés au sein de fenêtres, a été montrée comme étant plus puissante pour l'iHS que celle utilisant directement les valeurs des scores (Pickrell et al., 2009; Triska et al., 2015).

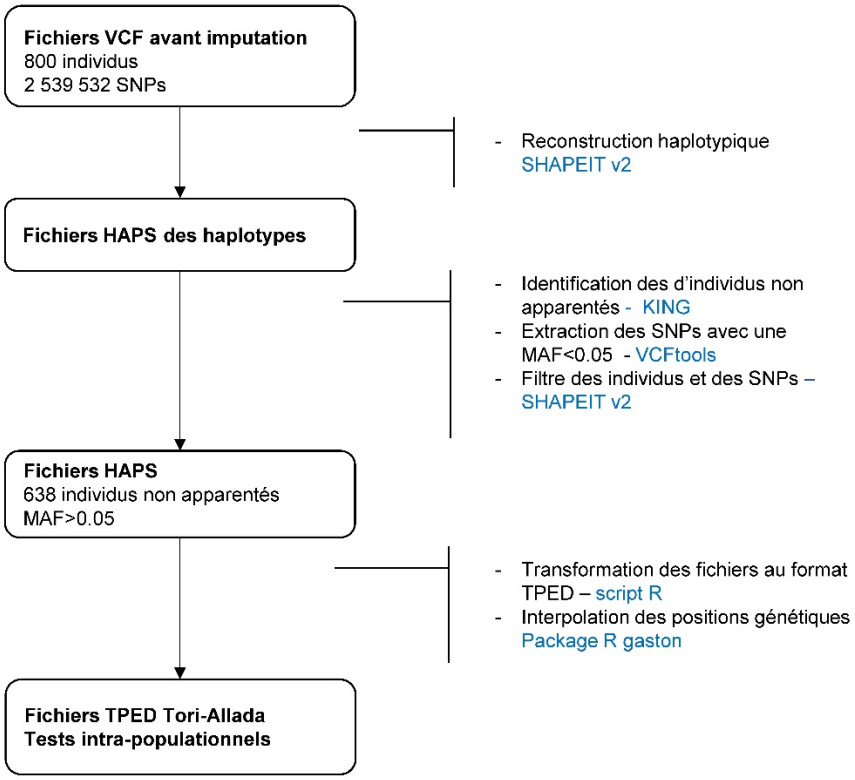


Figure 5.1 Préparation des fichiers d'analyse pour les tests intra-populationnels

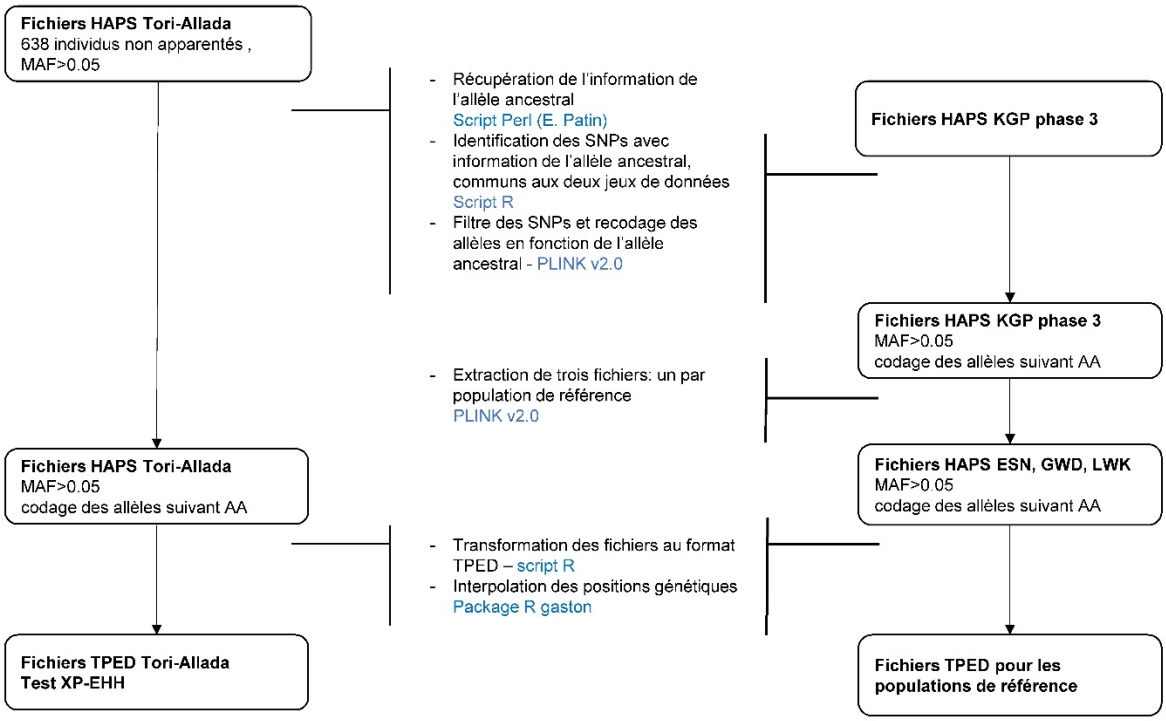


Figure 5.2 Préparation des fichiers d'analyse pour le test XP-EHH

Test XP-EHH

Nous avons comparé notre population à trois populations d'Afrique sub-saharienne des données 1000 Genomes, présentant un niveau croissant de divergence génétique avec la nôtre :

- une population géographiquement et génétiquement proche : les Esans du Nigéria (ESN),
- une population à l'extrême ouest de l'Afrique : les gambiens (GWD),
- une population d'Afrique de l'est : les Luhyas du Kenya (LWK).

Nous avons utilisé des populations africaines comme populations de référence afin d'éviter les artéfacts liés à une histoire démographique trop différente. Les effets fondateurs et les fortes expansions de population qu'ont connus les populations non africaines après la sortie d'Afrique peuvent affecter de manière importante la statistique XP-EHH. Les trois populations choisies sont toutes exposées au paludisme (il aurait été pertinent de réaliser une comparaison avec une population non exposée mais nous ne disposions pas de données génome entier pour une population africaine n'ayant jamais été exposée au paludisme) ; cependant les populations d'Afrique de l'ouest et d'Afrique de l'est présentent des ascendances génétiques distinctes (Gouveia et al., 2019; Triska et al., 2015) qui ont pu donner lieu à des adaptations différentes au même agent infectieux ; nous avons également considéré une population plus proche de la nôtre, les mécanismes d'adaptation pouvant survenir à un niveau local, comme montré dans le cas du gène CD36 par *Fry et al.* (Fry et al., 2009).

L'XP-EHH compare l'iHH (l'intégrale de la fonction EHH autour du SNP d'intérêt) aux mêmes SNPs entre la population d'intérêt et une population de référence. L'information sur l'état ancestral ou dérivé des allèles est importante dans le cas de ce test car elle permet de déterminer dans quelle population est observé le signal de sélection. Aussi, par rapport aux tests iHS et nS_L, la mise en œuvre du test XP-EHH nécessite une étape supplémentaire dans la préparation des fichiers afin de récupérer l'information sur l'allèle ancestral et de recoder les allèles en fonction de leur état ancestral ou dérivé (en haut de la Figure 5.2). Le test XP-EHH a été mis en œuvre pour les SNPs avec une MAF>0,05 dans la population du Sud Bénin, dont l'information sur l'allèle ancestral était

disponible (1 727 300 ± 200 SNPs, suivant la population de référence). L'information sur l'état ancestral ou dérivé a été obtenue à partir de la séquence ancestrale GRCh37 e71 d'Ensembl (Flicek et al., 2013) qui utilise un alignement multiple de six primates.

Les scores d'XP-EHH ont été standardisés par le programme *norm* en utilisant la moyenne et l'écart-type des valeurs sur l'ensemble du génome. La même approche que pour les tests intra-populationnels a ensuite été appliquée pour identifier les signaux significatifs, en utilisant la proportion de valeurs extrêmes dans des fenêtres non-chevauchantes de 100 kb.

5.1.3 Co-localisation des signaux d'association et de sélection

Enrichissement en signaux de sélection naturelle parmi les loci trouvés associés

Nous avons d'abord recherché s'il existait un enrichissement en signaux de sélection naturelle parmi les signaux d'association détectés pour le paludisme simple sur ces mêmes cohortes. Pour cela nous avons croisé l'information des signaux d'association et de sélection au sein des fenêtres de 100 Kb définies pour identifier les signaux de sélection naturelle.

Pour les signaux de sélection, les trois tests effectués ciblant des balayages sélectifs sensiblement différents (de type hard sweep pour iHS, de type hard sweep et soft sweep pour nS_L) ou étant plus puissants dans certaines conditions (balayages sélectifs proches de la fixation pour XP-EHH), nous avons considéré qu'une fenêtre présentait un signal de sélection significatif si au moins un des trois tests était significatif. Les analyses ont été réalisées avec un seuil de signification de 0,01 communément utilisé dans les scans de sélection naturelle, puis avec un seuil moins stringent de 0,05. Pour les signaux d'association, afin d'obtenir l'information d'association globale sur les deux cohortes, nous avons réalisé l'analyse génome entier dans la cohorte d'Allada avec le phénotype estimé dans l'étape de réplication de la GWAS et utilisé les résultats de la méta-analyse des deux cohortes obtenus avec le logiciel METAL (Willer & Li, 2010). A chaque fenêtre définie pour l'analyse des signaux de sélection naturelle, nous avons attribué la p-valeur minimum dans la fenêtre. Les

analyses ont été réalisées pour les deux phénotypes considérés dans la GWAS, la récurrence des accès palustres simples et la récurrence de l'ensemble des infections.

Afin d'évaluer l'enrichissement en signaux de sélection naturelle parmi les loci associés au paludisme simple, les fenêtres ont été divisées en quatre catégories en fonction de la p-valeur de l'analyse d'association (<10-5,]10-4 – 10-5],]10-3-10-4] et >10-3) dans lesquelles nous avons calculé le pourcentage de signaux de sélection naturelle significatifs. On émet l'hypothèse que dans les catégories de p-valeurs les plus significatives pour l'association, qui incluent une proportion moindre de faux positifs, on devrait observer plus de signaux de sélection naturelle si les loci associés au paludisme simple sont la cible de la sélection naturelle. Ainsi on devrait observer un enrichissement en signaux de sélection naturelle dans les catégories de p-valeurs les plus faibles.

Co-localisation parmi les plus forts signaux de sélection et d'association

Nous avons ensuite identifié parmi les plus forts signaux de sélection ceux associés au paludisme simple et inversement. Dans le premier cas nous avons considéré les 25 fenêtres avec les plus forts signaux de sélection pour chacun des tests de sélection et identifié celles dont la p-valeur minimum pour l'association était inférieure à 10^{-3}. Pour le test XP-EHH, les 25 premières fenêtres ont été sélectionnées en fonction du score maximum dans la fenêtre, étant donné que plus de 25 fenêtres ont un pourcentage de scores extrêmes égal à 1 pour chacune des populations de référence. Dans le second cas, nous avons sélectionné les fenêtres dont la p-valeur minimum de l'association était inférieure à 10^{-5} et identifié celles présentant une évidence de sélection au seuil de 0,05.

Ces fenêtres ont ensuite été explorées pour rechercher de potentiels gènes candidats impliqués dans le paludisme simple en utilisant la plateforme FUMA (Watanabe et al., 2017) (utilisée également pour l'analyse fonctionnelle *in silico* dans la GWAS et décrite à la section 3.2.4). Dans chaque fenêtre, le SNP avec la plus faible p-valeur a été soumis à FUMA comme SNP principal, une cartographie fine et une recherche des gènes candidats a été réalisée par cartographie positionnelle et cartographie eQTL. Les mêmes critères que dans l'analyse *post*-GWAS ont été appliqués (section 3.2.4). La

cartographie fine a été réalisée en identifiant les SNPs en déséquilibre de liaison (DL, r^2>0,6) avec les SNPs soumis, dans les données de la GWAS et dans les données KGP des populations africaines (phase 3). L'analyse par cartographie positionnelle a été réalisée pour les SNPs annotés comme fonctionnels d'après les bases de données CADD (score CADD> 12,37) et RegulomDB (score RDB score ≤ 2). Un gène a été identifié comme candidat si le SNP est situé à une distance <10 Kb de celui-ci. Pour la cartographie eQTL, nous avons ciblé les mêmes tissus que dans les analyses *post*-GWAS (i.e. : la peau, les fibroblastes de la peau, le sang total, les lymphocytes B, le foie et la rate) et utilisé les même bases de données, à l'exception de la base de données GTex pour laquelle nous avons utilisé la dernière version (v8 au lieu de v7).

Enfin, les gènes candidats identifiés dans ces fenêtres ainsi que les gènes présents dans les fenêtres avec les plus fortes évidences de sélection naturelle (les 10 premières de chaque test) ont été comparés à deux autres listes de gènes : la liste des gènes trouvés associés au moins une fois à une des trois formes de paludisme (infections asymptomatiques, accès palustres simples, accès palustre graves) que nous avons établie à partir de la littérature (section 1.3.3, Figure 1.11), et une liste de gènes candidats influençant la formation et la fonction du globule rouge identifiée dans une méta-analyse de données de GWAS portant sur ~ 135 000 individus (van der Harst et al., 2012).

5.2 Résultats

5.2.1 Vue d'ensemble des résultats sur le génome entier

Les Figures 5.3 et 5.4 présentent les résultats des tests de sélection naturelle par SNP, après standardisation des données sur l'ensemble du génome. Pour les tests intra-populationnels (iHS et nS_L, Figure 5.3), la valeur absolue du score a été représentée. En effet le sens positif ou négatif des scores n'apporte pas d'information dans cette analyse, les allèles n'ayant pas été codés suivant leur état ancestral ou dérivé. Pour l'XP-EHH (Figure 5.4) en revanche, nous avons représenté uniquement les scores positifs qui indiquent un signal de sélection dans notre population comparée à la

population de référence. La ligne bleue, dans ces figures, indique la valeur correspondant aux 1% des scores les plus élevés, qui a été utilisée dans l'analyse par fenêtre pour définir les scores extrêmes.

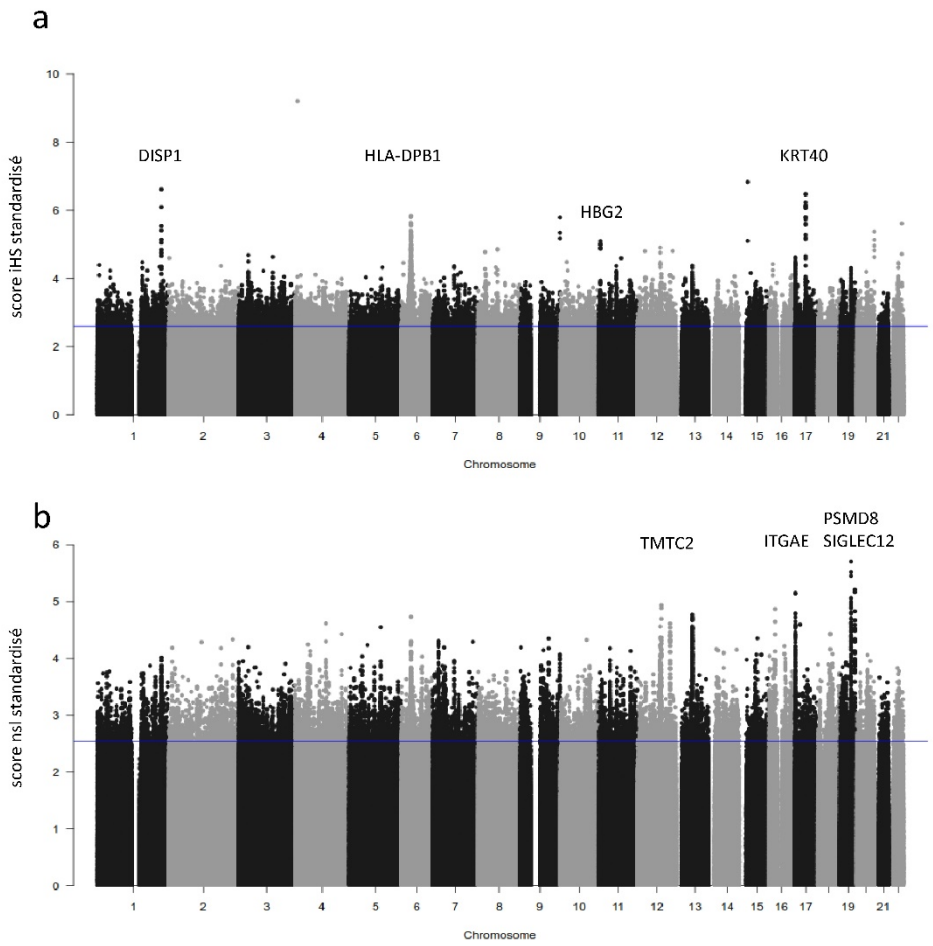


Figure 5.3 Manhattan plots des tests de sélection intra-populationnels: a) Test IHS, b)Test nS_L.
La ligne bleue indique la valeur du score correspondant aux 1% des scores les plus élevés sur l'ensemble du génome.Pour les signaux les plus proéminents, les gènes reportés correspondent au gène (±10Kb) dans lequel se situe le SNP avec le score le plus élevé. Deux gènes sont reportés dans une même région s'il existe deux signaux distincts forts dans cette région. Les régions des gènes HLA-DPB1 et HBG2 incluent plusieurs gènes ; nous avons représenté uniquement celui qui était exprimé dans les données GTEx afin de ne pas alourdir la figure.

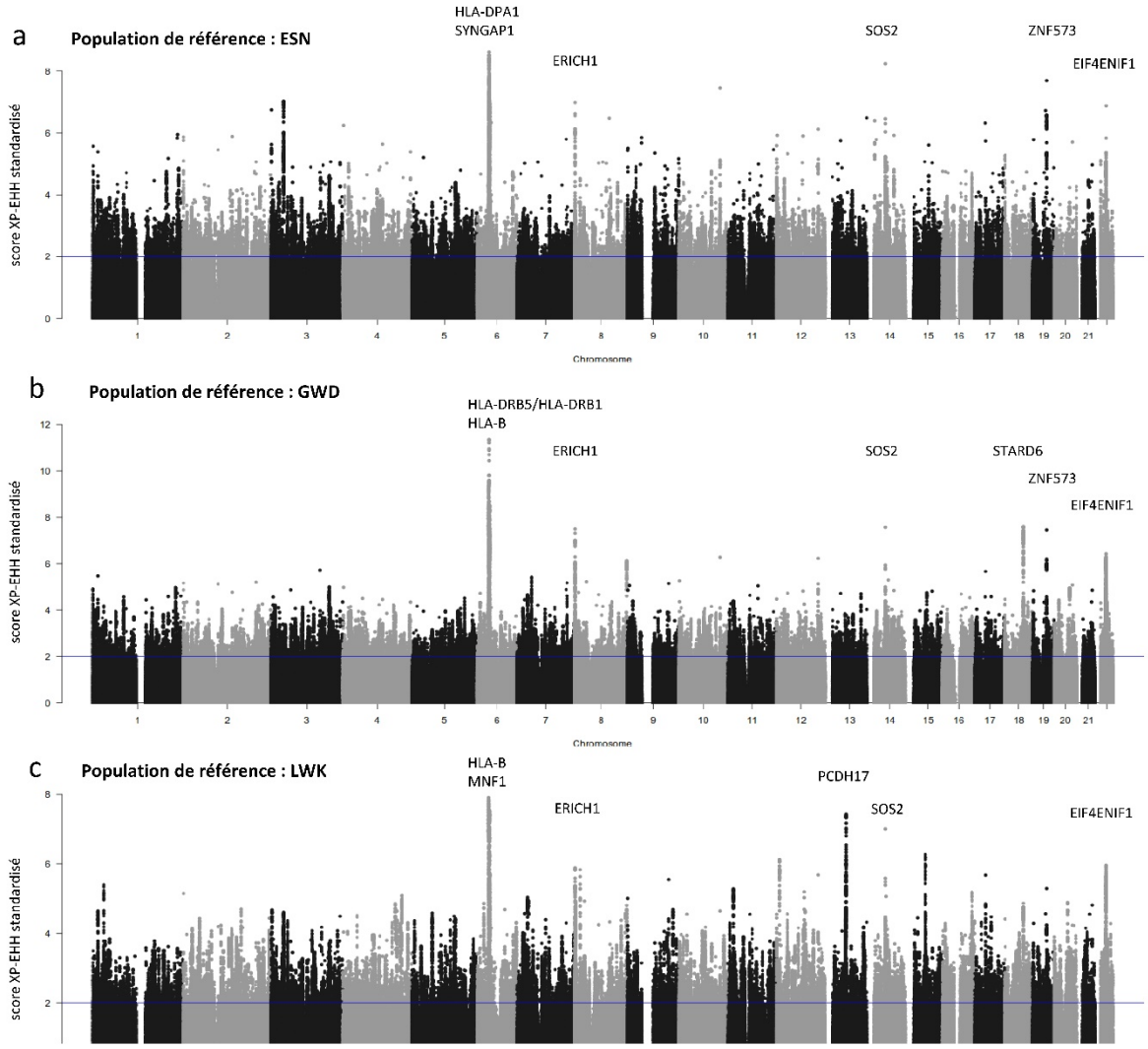


Figure 5.4 Manhattan plots des analyses XP-EHH avec trois populations de référence: a) ESN (Esans du Nigéria), b) GWD (Gambiens), c) LWK (Luhyas du Kenya).
La ligne bleue indique la valeur du score correspondant aux 1% des scores les plus élevés sur l'ensemble du génome. Pour les signaux les plus proéminents, les gènes reportés correspondent au gène (±10Kb) dans lequel se situe le SNP avec le score le plus élevé. Deux gènes sont reportés dans une même région s'il existe deux signaux distincts forts dans cette région.

Concernant les tests intra-populationnels (Figure 5.3), on observe un signal fort dans la région HLA pour le test iHS (SNP avec le score le plus élevé : rs5694772, max|iHS|=5,80, p-valeur empirique (p_{emp})=7,38 x 10^{-6}). Le SNP rs5694772 est situé dans les gènes *HLA-DPA1* et *HLA-DPB2* et constitue un eQTL du gène *HLA-DPB2.* Les autres principaux signaux correspondent à des signaux de sélection naturelle déjà observés dans de précédentes études de sélection naturelle récente dans les populations africaines. Les signaux au niveau des gènes *ITGAE* (rs4602070, max |nsl|=5,14, p_{emp}=2,82 x 10^{-6}), *KRT40* (rs372696652, max|iHs|=6,48 , p_{emp}=1,70 x 10^{-6}) et *DISP1* (rs17163519, max |iHS|=6,61

p_{emp}=1,13 x 10-6) ont été détectés de manière répétée dans des populations d'Afrique de l'ouest (Grossman et al., 2013; Triska et al., 2015), et dans une analyse englobant l'ensemble des populations d'Afrique sub-saharienne du projet African Genome Variation (AGVP, (Gurdasani et al., 2015). *ITGAE* (également appelé CD103) code pour un récepteur de la cadhérine E, et est exprimé de manière importante dans les lymphocytes T (lymphocytes T intra-épithéliaux de la muqueuse intestinale et lymphocyte Treg), *KRT40* code pour une protéine de la famille des kératines épithéliales présente dans l'épiderme, et *DISP1* (Dispatched RND Transporteur Family Member 1) apparait impliqué dans le développement précoce de l'embryon. Les gènes *SIGLEC12 et TMTC2* ont également été identifiés comme des cibles potentielles de sélection naturelle dans les populations AGVP (Gurdasani et al., 2015), ainsi que dans l'étude de Triska *et al.* (Triska et al., 2015) pour *TMTC2*. La fonction de *SIGLEC12* est particulièrement intéressante. Il intervient dans la réponse immunitaire innée et dans le traitement et la présentation des antigènes par les molécules du complexe majeur d'histocompatibilité de classe I. Ce gène a été montré comme potentiellement sélectionné par la trypanosomose humaine africaine dans l'étude de Gurdasani *et al.* (Gurdasani et al., 2015). On peut noter également au début du chromosome 11, un signal iHS fort (rs201924885, max |iHS|=5,6, p_{emp}=2,55 x 10^{-5}) proche du gène *HBB* (250Kb). Les SNPs avec des valeurs de l'iHS proches de 5 (n=10) sont situés dans une région restreinte de 2,5 Kb incluant les gènes *HBG2*, *HBE1* et *OR51B5,* à 250 Kb du gène *HBB.* Les données GTEx dans cette région montrent une forte expression de *HBG2* dans le sang et une absence d'expression des deux autres gènes quel que soit le tissu.

Les analyses XP-EHH avec trois populations de référence différentes révèlent également des signaux particulièrement élevés au niveau de la région HLA (Figure 5.4). Ces signaux sont situés dans des gènes distincts suivant la population de référence considérée, à l'exception d'un signal dans *HLA-B* commun aux analyses avec les deux populations les plus éloignées de notre population d'étude (GWD et LWK). La majorité des autres signaux forts sont observés dans au moins deux des trois analyses réalisées. Les signaux dans les chromosomes 8, 14 et 22, en particulier, sont observés avec les trois populations de référence ; ils ciblent des gènes identiques, respectivement *ERICH1*

(Glutamate rich 1), *SOS2* (Ras/Rho Guanine Nucleotide Exchange Factor 2) et *EIF4ENIF1 (*Eukaryotic Translation Initiation Factor 4E Nuclear Import Factor 1). La fonction de *ERICH1* est inconnue *;* SOS2 code pour une protéine impliquée dans la régulation positive des protéines *ras* et *EIF4ENIF1* code pour une protéine de transport nucléoplasmique pour le facteur d'initiation de la traduction eIF4E.

5.2.2 Analyse par fenêtres de 100 kb

Pour identifier sur l'ensemble du génome les régions potentiellement ciblées par la sélection naturelle, nous avons utilisé une approche par fenêtres non chevauchantes de 100 Kb.

La Figure 5.5 présente les dix régions avec les plus forts signaux de sélection naturelle pour chacun des cinq tests de sélection réalisés. Ces régions correspondent à une fenêtre ou à plusieurs fenêtres consécutives. En effet, lorsque plusieurs fenêtres adjacentes étaient présentes dans les dix premières fenêtres, elles ont été agrégées et nous avons considéré les fenêtres suivantes au-delà des dix premières (la première région dans le chromosome 13, par exemple, a une taille de 500 Kb car elle inclut cinq fenêtres). Nous pouvons remarquer que les régions avec les plus forts signaux de sélection pour l'iHS sont également toutes identifiées au seuil de 0,01 par le test $nS_{L,}$ ce qui est concordant avec le fait que le test nS_L est au moins aussi puissant que l'iHS pour détecter les signaux de type *hard sweep* dans différents scénarios de sélection comme le décrivent les auteurs dans l'article initial présentant le nS_L (Ferrer-Admetlla et al., 2014). En revanche, le test nS_L met en évidence deux régions qui ne sont pas détectées par le test iHS, qui pourraient correspondre à des signaux de type soft sweep pour lequel ce test a une plus grande puissance de détection. Des régions sont identifiées à la fois par les tests intra-populationnels et le test XP-EHH, ce qui renforce le fait que ces régions soient des cibles réelles de la sélection naturelle. Pour trois régions en particulier, nous observons un signal significatif au seuil de 0,01 pour l'ensemble des cinq tests. Ces trois régions font partie des signaux proéminents observés dans les Manhattan plots (Figures 5.3 et 5.4). Il s'agit de la région HLA incluant le gène *HLA-DPB1*, d'une région proche de la région HLA incluant le gène *MNF1* et de la région sur le chromosome 22 incluant le gène *EIF4ENIF1*. Le gène *MNF1* (appelée depuis *UQCC2* dans la nomenclature HGNC pour *Ubiquinol-Cytochrome C Reductase Complex Assembly Factor 2)* code

pour une protéine située dans la membrane interne des mitochondries, et joue un rôle dans la régulation de la sécrétion d'insuline, dans la production d'ATP mitochondrial et dans la formation des tissus musculaires. Ces analyses mettent en avant également une large région de 500 kb sur le chromosome 13 (région chromosomique 13q21.1) pour laquelle aucun gène n'est répertorié. Cette région apparait dans les dix premiers signaux pour trois tests (iHS, $nS_{L,}$ et XP-EHH avec la population LWK comme population de référence). De façon intéressante, elle a aussi été identifiée comme faisant partie des dix premiers signaux de sélection, avec le test iHS, dans plusieurs populations d'Afrique de l'ouest (Gambien, Mende, Esan et Burkinabé) (Triska et al., 2015). Elle n'a pas été identifiée par contre avec le second test mis en œuvre dans cette même étude, un test XP-EHH utilisant comme population de référence la population italienne de 1000 Genome.

En ce qui concerne les gènes impliqués dans le paludisme et connus pour être la cible de la sélection naturelle, nous observons un signal significatif à 0,05 pour les gènes *HBB* et *CD36* pour les deux tests intra-populationnels (résultats non montrés). Les autres gènes étudiés (*HBA* et *DARC*) ne présentent pas d'éléments en faveur d'une action de la sélection naturelle avec les tests utilisés. Les gènes *G6PD* et *CD40LG* sont situés sur le chromosome X que nous n'avons pas analysé dans cette étude.

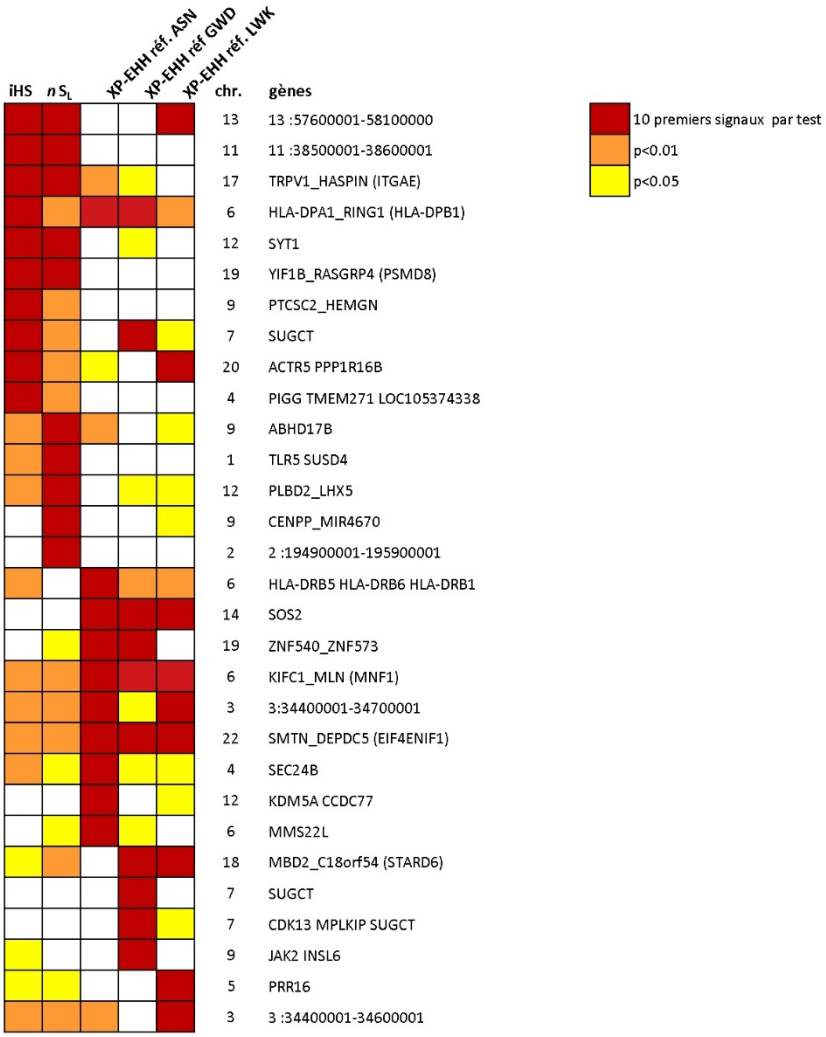


Figure 5.5 Dix premières régions identifiées pour chaque test de sélection naturelle par l'analyse par fenêtres de 100 Kb. Les 10 premières régions sont indiquées en rouge. Les couleurs orange et jaune renseignent si ces mêmes régions sont détectées au seuil de 0,01 et 0,05 respectivement par les autres tests. Ces régions correspondent à une ou à plusieurs fenêtres consécutives ; les fenêtres adjacentes dans les premiers signaux ont été agrégées. Les gènes présents dans la région sont reportés, avec l'ensemble des gènes présents lorsque la région inclut au plus trois gènes; sinon les premier et dernier sont séparés par un tiret bas. Les gènes reportés dans la description des résultats par SNP (section 5.2.1) sont mentionnés entre parenthèses. Lorsqu'aucun gène n'est présent dans la région, la position de la fenêtre est reportée.

5.2.3 Co-localisation des signaux d'association et de sélection

Nous avons ensuite croisé les informations de sélection naturelle avec celles de l'association avec le paludisme simple provenant de la GWAS. Ce croisement a été réalisé en considérant la p-valeur minimum de l'association dans chaque fenêtre.

Enrichissement en signaux de sélection naturelle parmi les loci trouvés associés au paludisme

Dans un premier temps, nous avons regardé s'il existait un enrichissement en signaux de sélection naturelle dans les fenêtres présentant des niveaux de preuve croissants d'association avec les phénotypes étudiés (accès palustres simples –APS- et ensemble des infections -EI). Les graphiques en barre de la Figure 5.6 représentent les pourcentages de fenêtres du génome présentant une évidence de sélection naturelle aux seuils de 0,01 (orange) et 0,05 (jaune) pour différentes catégories de SNPs en fonction de la p-valeurs de l'association. La partie haute de la Figure 5.6 présente les résultats obtenus avec les APS et la partie basse ceux obtenus avec l'EI. Les graphiques à gauche montrent les résultats obtenus lorsque l'on considère qu'une fenêtre est significative si au moins un des cinq tests de sélection est significatif. Les autres graphiques ont été obtenus en considérant les résultats d'un seul test, l'iHS au centre et le nS_L à droite. Dans aucun des cas considérés nous n'observons d'enrichissement en signaux de sélection naturelle dans les catégories de p-valeurs décroissantes. La plus forte différence de proportion a été observée pour le phénotype APS en considérant l'ensemble des tests de sélection, mais celle-ci n'est pas significative (test du χ^2 global, p=0.89) et les proportions de signaux de sélection naturelle observées dans les régions les plus fortement associées ($p<10^{-5}$) ne vont pas dans le sens attendu.

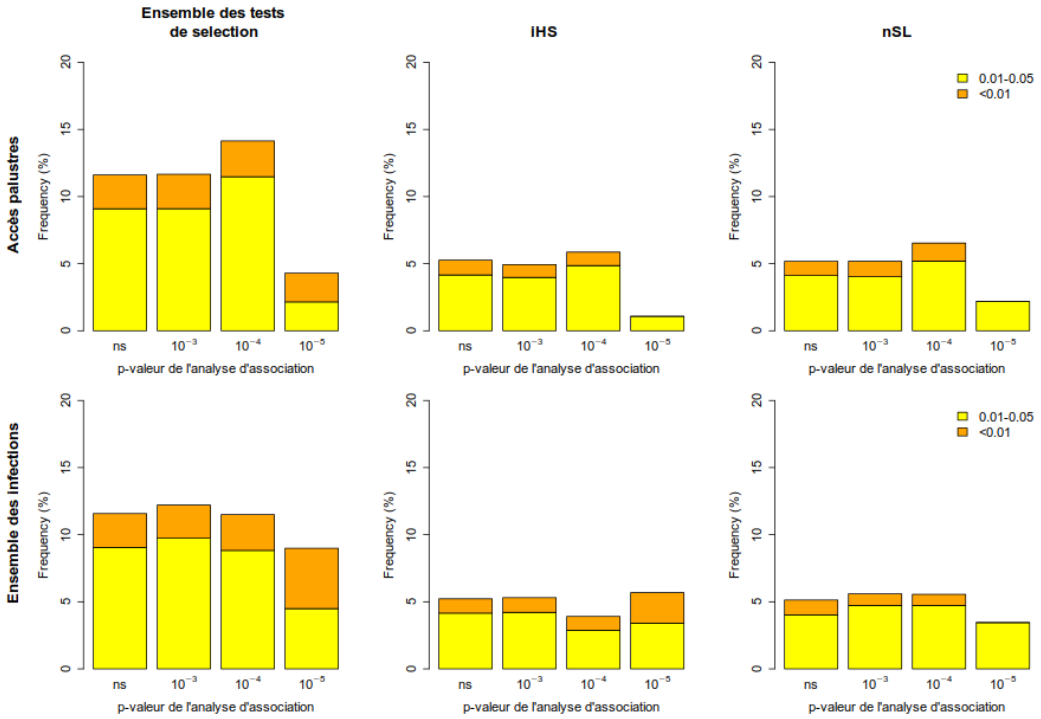


Figure 5.6 Pourcentage de fenêtres présentant un signal de sélection naturelle en fonction de la p-valeur de l'association avec les phénotypes palustres. Les graphiques du haut représentent les résultats obtenus pour les accès palustres simples et ceux du bas, pour l'ensemble des infections palustres. Nous avons considéré qu'une fenêtre présentait un signal de sélection (i) si au moins un des cinq tests de sélection réalisés était significatif (graphes de gauche), (ii) si le test iHS était significatif (graphes du centre), ou (iii) si le test nSL était significatif (graphes de droite). Deux seuils de signification statistique ont été considérés pour les tests de sélection : 0,05 (en jaune) et 0,01 (en orange).

Co-localisation parmi les plus forts signaux de sélection et d'association

Parmi les fenêtres présentant les signaux de sélection les plus forts (les 25 premières fenêtres pour chaque test), 16 incluent un signal d'association à $p<10^{-3}$ et sont présentées dans la Table 5.1. La plus faible p-valeur observée pour l'association se situe dans le gène *SOS2* ($p=6,2 \times 10^{-6}$ pour le phénotype EI). C'est également dans ce gène qu'est identifié un des plus fort signaux de sélection par l'XP-EHH et ce, pour les trois populations de référence (section 5.2.1, Figure 5.4 et section 5.5, Figure 5.5). La cartographie eQTL à partir du signal d'association identifie un grand nombre de gènes candidats dont *SOS2*.

Fenêtre	iHs	nS_L	XP-EHH			P association		SNP	Gène le plus proche[a]	Gène FUMA	
			ESN	GWD	LWK	APS	EI			position	eQTL
2 :98800001-98900001						8,7 x 10-4	3,3 x 10-4	2 :98839367	*VWA3B*	-	*INPP4A, ACTR1B, ZAP70, UNC50*
3 :34500001-34600001						1,9 x 10-4	6,5 x 10-4	3 :34522789	-	*AC018359.1* (lncRNA)	-
3 :34600001-34700001						6,0 x 10-4	-	3 :34699157	-		
4 :140300001-140400001						4,3 x 10-5	1,8 x 10-4	4 :140324714	*NAA15*	*NAA15, NDUFC1*	*NDUFC1*
5 :119700001-119800001						4,6 x 10-4	-	5 :119737830	*PRR16 (50kb)*	nr	nr
9 :5100001-5200001						-	8.7 x 10-5	9 :5167245	*INSL6*	-	-
9 :95200001-95300001						5,2 x 10-4	5,9 x 10-5	9 :95209094	*CENPP*	-	-
10 :68800001-68900001						4,0 x 10-4	-	10 :68876251	*CTNNA3*	*CTNNA3*	-
10 :74200001-74300001						7,7 x 10-4	-	10 :74229143	*MICU1*	*MCU, HMGN2P34, Y_RNA*	*SPOCK2, MICU1, FAM149B1, ASCC1, DNAJB12, ANAPC16, NUDT13, DDIT4*
14 :50600001-50700001						4,7 x 10-4	6,2 x 10-6	14 :50667641	*SOS2*	-	*MAP4K5, L2HGDH, VCPKMT, SOS2, CDKL1, ATP5S, C14orf183, ATL1, C14orf182*
14 :50700001-50800001						8,2 x 10-4	2,5 x 10-4	14 :50733863 * 14 :50718816	*L2HGDH* *L2HGDH*		
17 :39100001 39200001						4.1 x 10-4	-	17 :39140988	*KRT40*	nr	nr
22 :31000001-31100001						-	4,0 x 10-4	22 :31049964	*SLC35E4 DUSP18*	*OSBP2*	*SF3A1, SLC35E4, PIK3IP1, DUSP18, OSBP2, TCN2, SELM, PISD*
22 :31100001-31200001						-	4,0 x 10-4	22 :31193188	*OSBP2*		
22 :32000001-32100001						-	5,7 x 10-5	22 :32076316	*PRR14L*	*PRR14L*	-
22 :32100001-32200001						-	4,5 x 10-5	22 :32108997	*PRR14L* (intron)		

Table 5.1 Fenêtres incluant les plus forts signaux de sélection et montrant une association à $p<10^{-3}$ avec les phénotypes palustres dans la GWAS (méta-analyse)

Le quadrillage de couleur à gauche indique le(s) test(s) avec le(s)quel(s) la région a été identifiée comme une cible potentielle de la sélection naturelle : en rouge, les régions figurant parmi les 25 fenêtres présentant les scores les plus significatifs ; en orange, les régions avec une évidence de sélection à p<0,01 ; en jaune, les régions avec une évidence de sélection à p<0,05. Les trois colonnes suivantes rapportent la p-valeur minimum de l'association dans la fenêtre pour la méta-analyse sur les accès palustres simples (APS) et pour celle sur l'ensemble des infections (EI). Dans les deux dernières colonnes sont renseignés les gènes candidats identifiés par cartographie positionnelle (position) et cartographie eQTL (eQTL) avec la plateforme FUMA.

[a] la distance au gène est indiquée entre parenthèse, lorsque celle-ci n'est pas indiquée le SNP est situé dans le gène

**les p-valeurs minimums pour les phénotypes APS et EI sont obtenues pour deux SNPs différents ; lncRNA ; long non-coding RNA ; nr, cartographie non réalisée par FUMA.*

Parmi les autres gènes cibles de la sélection au sein desquels on observe également une association, on peut noter *KRT40* (chromosome 17, p=4.1 x 10^{-4} pour le phénotype APS) et *PRR14L* (chromosome 22, p≤5,7 x 10^{-5} pour le phénotype EI). *KRT40* présente un des trois signaux iHS les plus élevés dans notre étude (section 5.2.1, Figure 5.4) et il a été identifié de manière répétée comme cible potentielle de la sélection dans de nombreuses populations africaines, notamment d'Afrique de l'Ouest (Grossman et al., 2013; Gurdasani et al., 2015; Triska et al., 2015). La cartographie fine avec FUMA n'a pas pu être réalisée pour ce signal d'association (noté *nr* dans la table) ; il fait partie des trois signaux pour lesquels FUMA n'a pu identifier de région candidate, probablement du fait que les SNPs avec les plus faibles p-valeur dans ces régions (soumis à FUMA comme SNPs principaux) ne sont pas présents dans la base de données que la plateforme utilise. Le SNP 17 :39140988 (*KRT40*) est un variant fréquent dans toutes les populations KGP, mais étonnamment n'est pas présent dans la dernière version de dbSNP (dbSNP153). Les autres SNPs sont des variants rares (MAF<0,005) ou absents dans les populations d'origine européenne. *PRR14L* (Proline Rich 14 like) est situé dans une fenêtre identifiée comme une cible de sélection par les trois tests XP-EHH, comme pour *SOS2*. Pour ce signal d'association, deux variants introniques délétères sont identifiés au sein du gène, dont un faisant partie des 1% des substitutions les plus délétères du génome (score CADD de 25,4). La fonction de ce gène est encore inconnue.

La Table 5.2 présente les fenêtres comportant un signal d'association avec une p-valeur inférieure à 10^{-5} avec l'un des deux phénotypes palustres et un signal de sélection significatif au seuil de 0,05. En dehors de la fenêtre incluant *SOS2* (présente également dans la première table), un signal d'association est observé proche du gène *HLA-DQ2* (avec le phénotype EI p=9,5 x 10^{-6}) dans une région présentant un signal de sélection avec l'ensemble des tests à l'exception de nS_L. Un double signal d'association (avec le phénotype EI : p<3,4 x 10^{-6}) et de sélection (avec les tests iHS et nS_L) est également détecté dans le gène *LARGE*, un gène déjà décrit pour être une cible potentielle de la sélection naturelle, probablement en lien avec la fièvre de Lassa (Grossman et al., 2013; Gurdasani et

al., 2015). Dans ces régions, la cartographie fine avec FUMA identifie très peu de gènes candidats, que ce soit en postionnel ou en eQTL.

	P association				XP-EHH				Gène FUMA	
SNP	APS	EI	iHS	*n*SL	ESN	GWD	LWK	Gène le plus proche	Position	eQTL
2:105486737	6,6 x 10-6	-						LINC01159	-	-
4:56259600	-	5,2 x 10-6						TMEM165 (5kb)	-	-
4:100147065	4,4 x 10-6	5,7 x 10-6						LOC100507053	-	-
6:32676392	-	9,5 x 10-6						HLA-DQA2 (20Kb)	nr	nr
7:21360483	-	3,4 x 10-6						-	-	SP4
12:47064593	4,3 x 10-6	-						LOC100288798	-	-
12:82216479	-	6,4 x 10-6							-	-
18:7013445	4,4 x 10-6	-						LAMA1	-	-
14 :50667641	4,7 x 10-4	6,2 x 10-6						SOS2	-	mlt
22:34150327	-	3,4 x 10-6						LARGE	-	-

Table 5.2 Fenêtres présentant un fort signal d'association avec les phénotypes palustres ($p < 10^{-5}$) et un signal de sélection avec au moins l'un des cinq tests réalisés

Pour chaque fenêtre le SNP avec la p-valeur minimum est renseignée. Le quadrillage de couleur indique le(s) test(s) avec le(s)quel(s) la région a été identifiée comme une cible de la sélection naturelle : en rouge, les régions figurant parmi les 25 fenêtres les plus significatives ; en orange, les régions avec une évidence de sélection à p<0.01 ; en jaune, les régions avec une évidence de sélection à p<0.05. Dans les deux dernières colonnes sont renseignés les gènes candidats identifiés par cartographie positionnelle (position) et cartographie eQTL (eQTL) avec la plateforme FUMA.

[a] la distance au gène est indiquée entre parenthèse, lorsque celle-ci n'est pas indiquée le SNP est situé dans le gène

nr, cartographie non réalisée par FUMA, mlt, de multiple SNPs sont identifiés par cartographie eQTL pour SOS (cf. Table5.1).

Nous avons ensuite recherché parmi les gènes identifiés (gènes les plus proches et gènes identifiés d'après la fonctionnalité des SNPs par FUMA décrits dans les Tables 5.1 et 5.2), si certains avaient été trouvés associés à l'une des trois formes de paludisme dans la littérature, ou à la fonction du globule rouge dans une méta-analyse de données de GWAS (van der Harst et al., 2012). Aucun de ces gènes n'est présent dans la base de données des associations avec le paludisme que nous avons établie (section 1.3.3, Figure 1.11), et seul *HLA-DQA2* a été associé à la fonction des globules rouges. La même recherche pour les gènes présents dans les dix régions les plus significatives identifiées par chaque test de sélection met en évidence deux gènes associés auparavant au paludisme : *TLR5* qui a été associé aux formes asymptomatiques à *P. vivax* dans une étude au Brésil (Costa et al., 2017) et *HLA-DRB1* qui a été associé aux formes graves de paludisme dans trois populations indépendantes (Hananantachai et al., 2005; Hill et al., 1991; Osafo-Addo et al., 2008).

5.3 Discussion

Cette étude constitue la première analyse des signatures de sélection naturelle récente à l'échelle du génome dans une population du Bénin. Parmi les plus forts signaux de sélection identifiés par les tests intra-populationnels (iHS et nS_L), nous retrouvons des signaux déjà détectés à plusieurs reprises dans des populations d'Afrique de l'ouest par des approches génome entier : *ITGAE, KRT40, DISP1* (Grossman et al., 2013; Gurdasani et al., 2015; Triska et al., 2015). D'autres régions parmi les dix premières pour chacun des cinq tests de sélection réalisés ont été également identifiées auparavant par ces mêmes études : la région incluant *SYT1* sur le chromosome 12 (Grossman et al., 2013; Gurdasani et al., 2015; Triska et al., 2015), la région HLA incluant *HLA-DPB1* (Grossman et al., 2013), la région incluant *SUGCT* sur le chromosome 7 (appelée auparavant C7orf10) (Grossman et al., 2013; Triska et al., 2015), une large région sur le chromosome 13 n'ayant pas de gène connu (Triska et al., 2015) et enfin la région incluant le gène *TLR5* sur le chromosome 1 (Grossman et al., 2013). Ces régions ont été identifiées dans Triska *et al.* comme faisant partie des 10 premières régions identifiées par le test iHS dans plusieurs populations d'Afrique de l'Ouest (dans quatre populations sur les cinq suivantes à chaque fois : Gambiens, Mende, Yoruba, Esan et Burkinabés); dans les 0,1% des signaux avec le plus fort score iHS dans l'ensemble des populations africaines du panel AGVP (Gurdasani et al., 2015); dans les régions identifiées par Grossman *et al.*, en utilisant le test *Composite of Multiple Signals* (CMS) dans la population Yoruba. Dans ces régions, nous ne mettons pas en évidence de gènes associés précédemment au paludisme à l'exception de *HLA-DRB1* et *TLR5. HLA-DRB1* a été associé à plusieurs reprises avec les formes graves de paludisme. *TLR5* est principalement connu pour modifier la voix métabolique NF-κB en réponse aux bactéries flagellées et a été associé une seule fois aux formes asymptomatiques dues à *P. vivax*.

Les gènes associés aux formes graves de paludisme et connus pour être la cible de la sélection naturelle n'apparaissent pas parmi les signaux de sélection les plus forts à l'échelle du génome dans notre étude. En utilisant l'approche par fenêtres non chevauchantes, *HBB* et *CD36* présentent un

signal de sélection significatif à 5% avec les tests intra-populationnels. Les signatures de sélection naturelle dans ces deux gènes ont été mises en évidence auparavant par des méthodes basées sur les haplotypes longs, mais avec des approches gènes candidats (Fry et al., 2009; Hanchard et al., 2006) ou des approches génome entier considérant les scores du test de neutralité à l'échelle du SNP et non à l'échelle d'une fenêtre (International HapMap Consortium, 2005; Pybus et al., 2014). C'est le cas, par exemple, dans l'étude du consortium Hapmap (International HapMap Consortium, 2005), qui identifie dans la population YRI, les gènes *HBB* et *CD36* comme étant la cible de la sélection en se basant sur la distribution de la statistique du test *long range haplotype* (LHR) en fonction de la fréquence du SNP. Ces dernières années, plusieurs études ont tenté de mettre en évidence des signatures de sélection naturelle récente en lien avec le paludisme, par des approches génome entier avec un succès mitigé. Triska *et al*. (Triska et al., 2015) ont étudié les signatures récentes de sélection dans un ensemble de populations africaines (plusieurs populations du Sahel et les population subsaharienne du projet 1000 *Genomes*) en utilisant le test iHS et le test XP-EHH (en prenant la population italienne de 1000 *Genomes* comme population de référence). Ils ont recherché la présence de signaux de sélection naturelle dans les gènes inclus dans la voie métabolique du paludisme (*malaria pathway* dans la base de données KEGG). Pour les différentes populations incluses dans leur étude, ils rapportent les gènes qui sont localisés dans l'une des 300 premières fenêtres identifiées par le test iHS ou par le test XP-EHH, dans au moins une des populations étudiées. Ils identifient uniquement cinq gènes comme étant la cible de la sélection naturelle dans au moins une population : *DARC* et *CD36* ainsi que trois autres gènes (*MET, KLRK1 et THB53*) qui ne sont pas inclus dans la liste des gènes trouvés associés au paludisme que nous avons établie à partir de la littérature (description de la liste dans la section 1.3.3 de ce mémoire). Par une approche différente évaluant le niveau de différenciation génétique entre des populations résidant en zones endémiques et d'autres résidant en zones non endémiques, Gurdasani *et al.* (Gurdasani et al., 2015) ont mis en évidence une liste de gènes dont le lien avec le paludisme n'a pas été clairement démontré à part pour deux gènes, *HBB* et *IL10*. Enfin, une étude récente (Gouveia et al., 2019) s'est intéressée aux

signatures de sélection naturelle dans les populations de la ceinture transafricaine du lymphome de Burkitt (une région avec une transmission intense du paludisme). Ils ont appliqué deux méthodes basées sur le niveau de différenciation génétique (PBS, pour *population branch statistic et XP-EHH)* dans deux populations situées de part et d'autre de la ceinture, au Ghana et en Ouganda, avec comme populations de référence des populations non-exposées au paludisme. Des valeurs de PBS extrêmes ont été identifiées dans les populations du nord de l'Ouganda, dans la région du gène *ATP2B4*, qui a été associé à plusieurs reprises avec les formes graves du paludisme. Les autres gènes identifiés n'ont pas de relation évidente avec le paludisme. Ces différentes études montrent une certaines difficulté à mettre en évidence les signaux de sélection en lien avec le paludisme par une approche génome entier. Dans le cas des approches basées sur les haplotypes longs, il est possible que les modèles de type *hard sweep* ou *soft sweep* ne correspondent pas à la réalité, peut-être beaucoup plus complexe, des signatures de sélection laissées par paludisme, notamment du fait de l'histoire longue du paludisme et paradoxalement de la pression de sélection particulièrement intense exercée (Karlsson et al., 2014). La pression de sélection a pu agir par exemple sur de multiples variants à un même locus (comme au locus *HLA* ou *HBA*) ou favoriser des variants dans un grand nombre de gènes, correspondant à une adaptation polygénique et dont les signaux sont difficilement détectables avec les méthodes actuelles.

En croisant de manière simple les informations de sélection naturelle et d'association au sein des fenêtres, nous ne mettons pas en évidence d'enrichissement en signaux de sélection naturelle dans les régions présentant des niveaux de preuves croissants d'association avec les formes simples de paludisme. Plusieurs hypothèses, peuvent expliquer cette observation. La première est que les gènes associés à la protection contre les formes non compliquées du paludisme, contrairement à ceux associés au paludisme grave, ne seraient pas la cible de la sélection naturelle récente, ou de manière pas assez forte pour être détectée par les outils actuels. Nous pouvons émettre l'hypothèse que les facteurs génétiques qui protègent contre l'infection par *P.falciparum* ou contre le développement d'une forme clinique du paludisme, s'ils ont un effet suffisamment fort, réduisent *de facto* de façon

importante le risque de faire une forme grave (et donc de décès) et soient la cible de la sélection naturelle. Cependant si nous considérons le cas de variants n'apportant qu'une protection partielle contre les infections et qui ne sont pas associés à la gravité de la maladie, nous pouvons nous attendre à ce qu'ils soient l'objet d'une sélection moins forte, d'autres facteurs (génétiques et non génétiques) jouant un rôle sur la sévérité et la mortalité.

La proportion de faux-positifs inhérente aux approches génome entiers (que ce soit pour les tests d'association ou les tests de sélection) pourrait masquer également la relation entre les deux types de signaux. Dans la GWAS, le seuil considéré de 1×10^{-5} pour définir les régions avec les plus forts niveaux d'association implique la présence d'une proportion importante de faux positifs, même dans cette catégorie incluant les signaux d'association les plus forts (seul le seuil de 5×10^{-8} permettant de contrôler le niveau de faux positifs à 5%). De même, pour les tests de sélection naturelle, l'approche empirique basée sur la distribution des scores statistiques sur l'ensemble du génome permet de déterminer à quel point le score observé à un SNP est extrême par rapport à ce qui est observé dans l'ensemble du génome mais ne permet pas d'évaluer le taux de faux-positifs parmi les régions présentant des scores extrêmes (François et al., 2016). Cette approche, si elle est puissante pour détecter des signaux de sélection dans des gènes bien connus pour leur rôle adaptatif, présente l'inconvénient de ne pas contrôler le taux de faux positifs. Enfin, comme mentionné ci-dessus on peut envisager également que les méthodes mises en œuvre ne soient pas adaptées pour détecter les signaux en lien avec le paludisme.

Nous nous sommes concentrés ensuite sur les régions où les signaux d'association et de sélection colocalisent dans les régions présentant soit un signal extrême de sélection (parmi les 25 premières fenêtres de chaque test) soit un signal fort d'association ($p<10^{-5}$).Nous n'avons pas encore examiné de façon approfondie (au moment où est rédigé ce manuscrit), par manque de temps, la robustesse des signaux d'association dans ces régions (par un examen des signaux à l'échelle régionale par LocusZoom), ni la fonction de l'ensemble des gènes présents dans ces régions ou identifiés par

FUMA. Ces premières analyses mettent en avant certains gènes ou régions avec un intérêt potentiel : *PRRL14* dont le signal d'association ($p=4,5 \times 10^{-5}$ avec le phénotype IE) est en DL avec une mutation fortement délétère, et qui est situé dans une région présentant des valeurs de scores extrêmes avec les trois tests XP-EHH ; la région incluant *SOS2* ($p = 6,2 \times 10^{-6}$ avec le phénotype EI) qui est la seule région parmi celles avec un signal de sélection extrême à présenter un signal d'association à $p<10^{-5}$. Nous pouvons noter également parmi les régions avec un signal fort d'association ($p<10^{-5}$) les gènes *HLA-DQA2* et *LARGE. HLA-DQA2* a été associé au nombre de globules rouges (van der Harst et al., 2012) et la même région entre *HLA-DQB1* et *HLA-DQA2* a été associée à la tuberculose dans une GWAS sur 200 000 individus d'origine européenne (Tian et al., 2017) et au lupus érythémateux systémique (Alarcón-Riquelme et al., 2016), une maladie qui partage avec le paludisme un certain nombre de gènes de susceptibilité/résistance. *LARGE* est connu pour être une cible de sélection naturelle dans les populations ouest africaines (Andersen et al., 2012; Grossman et al., 2013; Gurdasani et al., 2015) et sa relation avec le virus de Lassa a été démontrée (Andersen et al., 2012). Il est curieux d'observer un signal d'association avec le paludisme dans ce gène. Il code pour une glycosyltransférase qui modifie le récepteur α-distroglycan nécessaire à l'infection des cellules par le virus de Lassa. Ce signal peut évidemment être un faux-positif dans le cas de la GWAS mais mérite d'être investigué. Par ailleurs, ces analyses ne mettent pas en évidence de signaux de sélection naturelle dans les principaux gènes identifiés par la GWAS *(MYLK4, PTPRT, ACER3, VENTX ou UROC1)*.Cependant les SNPs trouvés associés dans la GWAS ont des MAF comprises entre 0,02 et 0,12, et il est connu que les méthodes basées sur les haplotypes longs présente de manière générale une puissance limitée pour détecter la sélection naturelle lorsque les allèles ont une fréquence faible.

Il est difficile à ce stade d'évaluer si l'information de sélection naturelle récente peut aider à identifier des gènes d'intérêt impliqués dans le paludisme simple dans les analyses génome entier.

Ces analyses doivent être poursuivies. Nous envisageons en premier lieu d'utiliser une approche différente pour identifier les signaux de sélection à partir des tests de sélection déjà réalisés. L'approche par fenêtres non chevauchantes est celle implémentée dans le logiciel Selscan mais elle n'est pas forcément la plus adaptée. Le découpage par fenêtre s'effectue de manière arbitraire et peut séparer des signaux de sélection et d'association. Une analyse par fenêtre glissante permettrait de localiser de manière plus précise les signaux de sélection. Elle permettrait également d'obtenir une information pour chaque SNP et de croiser les deux informations à l'échelle du SNP et non plus à celle de la fenêtre. Utiliser directement l'information du test de neutralité parait également pertinent lorsque l'on examine les signaux de sélection dans les gènes *HBB* et *CD36*. Les scores maximums observés pour ces deux gènes sont respectivement de 3,87 et 3,20 pour l'iHS, ce qui les classe parmi les 0,001% et 0,01% des scores les plus élevés sur l'ensemble du génome. Cependant le pourcentage de scores extrêmes dans les fenêtres est respectivement de 7,6% et 8,8% pour la fenêtre, les classant uniquement parmi les 5% des fenêtres avec un pourcentage de score de sélection. Pybus *et al* (Pybus et al., 2014), présentent un pourcentage de scores extrêmes similaires pour le test iHS pour le gène *CD36* dans la population YRI (7,1%) mais concluent à la présence d'un signal de sélection sur la base du score maximum (4,63). Puis si les données génome entier d'une telle population venaient à être disponibles dans le futur, nous envisageons de compléter nos analyses par un test XP-EHH utilisant une population de référence non-exposée au paludisme et présentant une histoire démographique proche des populations du sud Bénin.

6 DISCUSSION ET PERSPECTIVES

L'étude des déterminants génétiques constituait un des principaux objectifs lors de la constitution des deux cohortes de nouveau-nés du Sud Bénin. Avant cette thèse, les premières analyses sur ces cohortes se sont focalisées sur l'impact du paludisme pendant la grossesse sur la sensibilité des enfants aux premières infections palustres et sur le phénomène de tolérance immunitaire chez le jeune enfant, avec notamment l'étude du rôle du gène *HLA-G*. Cependant l'idée d'une analyse d'association génome entier sur ces deux cohortes de très jeunes enfants (c'est-à-dire dans des conditions facilitant la détection des facteurs de résistance génétique pour le paludisme avant le développement de l'immunité acquise) était bien présente depuis le début. Ce projet s'est concrétisé fin 2014 (après la fin du suivi de la deuxième cohorte), par la mise en place d'un partenariat entre l'équipe d'épidémiologie génétique de l'UMR216 MERIT et le Centre National de Recherche en Génomique Humaine (CNRGH, CEA, Evry, France) qui a accepté de financer le génotypage haute densité de ces deux cohortes. Les travaux de cette thèse ont permis la réalisation d'une première GWAS sur ces données, en utilisant comme phénotype la récurrence des infections, et d'étudier également à l'échelle du génome entier les signaux de sélection naturelle positive ou balancée récente.

Au cours des dix dernières années, les efforts de recherche pour l'identification de gènes impliqués dans la sensibilité au paludisme à *P. falciparum* se sont concentrés sur les formes graves de paludisme, avec plusieurs GWAS et analyses multicentriques publiées portant sur de grands échantillons provenant de différents pays. L'analyse d'association présentée dans ce manuscrit constitue la première GWAS sur les formes simples de paludisme, sur deux cohortes du Sud Bénin ayant fait l'objet d'un suivi rapproché pour le paludisme, privilégiant ainsi une définition précise du phénotype. Cette étude génome entier sur la récurrence des accès palustres et de l'ensemble des infections (incluant les infections asymptomatiques) entre la naissance et 18-24 mois montre des signaux d'association forts, à la limite du seuil de significativité communément admis dans les GWAS.

La plus forte évidence d'association est trouvée pour le gène *MYLK4* avec le risque d'accès palustres simples : l'association est répliquée pour plusieurs SNPs dans la seconde cohorte (dont deux avec une p-valeur<0,005); et il s'agit du signal pour lequel on observe la plus faible p-valeur pour la méta-analyse ($p = 5,29 \times 10^{-8}$). MYLK4 fait partie de la famille des kinases des chaînes légères de myosine impliquées dans l'organisation du cytosquelette d'actine/myosine et dans la motilité cellulaire (Tan et Leung_2009). Les autres signaux identifiés par l'étude de réplication mettent en évidence des gènes dont le rôle apparait très pertinent dans les premières étapes de développement des infections palustres (*PTPRT*, *ACER3* et *UROC1*) ou qui ont été impliqués dans la réponse immunitaire innée (*VENTX* et *ACER3* impliqués également dans l'activation de la réponse immunitaire innée dans le modèle murin), qui sont les facteurs génétiques plus particulièrement recherchés par notre étude.

Les niveaux de réplication peuvent paraitre faibles ; les p-valeurs pour les SNPs dont l'association réplique sont comprises entre 0,01 et 0,05 pour l'ensemble des gènes à l'exception de *MYLK4*. Ce faible niveau de réplication peut s'expliquer par le fait que la taille de l'échantillon de la seconde cohorte soit réduite (n=250) et la durée de suivi également plus courte que dans la première cohorte (12 mois versus 18 mois). Nous ne disposons pas de contrôle positif, tel que l'allèle *HbS* par exemple dans le cas des études sur le paludisme grave*,* pour conforter la pertinence de l'analyse. L'allèle HbS a aussi été associé aux formes simples de paludisme; cependant plusieurs études ont montrés que l'effet protecteur de l'allèle était lié à l'âge et n'était pas détecté avant l'âge de deux ans (Gong et al., 2012; Thomas N. Williams et al., 2005). Nous ne détectons pas d'effet de cet allèle dans cette étude (p=0,24 et p=0,08 pour le SNP rs334 pour la récurrence des accès palustres simples respectivement dans les cohortes de découverte et de réplication). Cependant, lorsque nous examinons les effets des SNPs dans les deux cohortes, en représentant le taux d'incidence brut (pour le SNP situé dans PTPRT, section 3.2.3 Figure 3.2) ou la fonction cumulative moyenne des événements (graphiques présentés pour l'ensemble des SNPs dans les données supplémentaires de l'article 1), les effets des SNPs sont très concordants dans les deux cohortes et convaincants. Le fait que les analyses fonctionnelles *in vitro* identifient des éléments fonctionnels dans l'ensemble des signaux qui répliquent (à l'exception

du signal dans *SYT16*) est un argument également en faveur d'une réelle association. Ces résultats doivent bien sûr être confirmés par d'autres études, mais pourraient ouvrir des pistes de recherche, pour le développement de nouveaux traitements ou d'outils de lutte.

Nous avons ensuite évalué la capacité de différentes méthodes existantes pour corriger la structure de population dans le cas d'un trait binaire, dans les données de ces deux cohortes. L'étude, réalisée à partir d'un phénotype simulé, montre qu'une des stratégies les plus simples qui consiste à inclure les premières PCs dans un modèle standard de régression peut être envisagée pour l'analyse des données du Sud Bénin. Même dans le cas d'une forte hétérogénéité entre les strates, la correction de la structure de population apparait adéquate. Nous montrons néanmoins que le modèle logistique mixte est légèrement plus puissant pour mettre en évidence les effets des SNPs. Les modèles mixtes ont été préconisés dans le cas des GWAS sur de grands échantillons, afin de prendre en compte à la fois la stratification de population et le fait qu'un certain nombre d'apparentés sont inclus de manière inévitable dans ce type d'échantillon (Sul et al., 2018). Dans notre étude, le niveau d'apparentement et de structure de population pourtant non négligeable n'apparait pas suffisant pour entraîner une inflation de l'erreur de type I lorsque l'on inclut simplement les premières PCs dans la régression logistique. Cette étude montre également que les différences de fonds génétique, à une échelle locale en Afrique, peuvent entrainer des biais dans les études d'association. L'analyse avec un modèle linéaire mixte (MLM) au lieu d'une régression logistique mixte (MLR) montre des biais similaires à ce qui est observé entre des populations plus éloignées géographiquement en Amérique du Sud et dans les Caraïbes. Une attention particulière doit donc être portée dans les études d'association en population (par opposition aux études sur des données familiales) en Afrique à la méthode utilisée pour prendre en compte la structure de population. Les simulations réalisées sur un échantillon d'individus plus large et présentant des apparentés au premier degré montrent que dans certaines situations seule la régression logistique mixte (MLR) corrige de manière adéquate la structure de population. La méthode GMMAT (un modèle logistique mixte avec un test du score (Chen et al., 2016)) permet d'estimer de manière efficace les p-valeurs dans une analyse génome

entier mais ne permet pas d'en obtenir les effets. Aussi nous avons proposé et évalué les propriétés de deux méthodes pour estimer les effets des SNPs avec une MLR dans les GWAS. Les deux méthodes présentent un léger biais négatif (pour un OR simulé de 2, le biais de -0,1 sur l'échelle log correspond à un OR de 1,8). La présence d'un tel biais dans une régression logistique est attendue en présence d'une hétérogénéité non prise en compte dans l'analyse. La méthode AMLE montre de bonnes propriétés (erreur de type I et puissance) dans les deux jeux de données étudiés. La méthode Offset présente également de bonnes propriétés dans les données de la GWAS, avec des biais négatifs légèrement moins importants que la méthode AMLE pour l'estimation de l'effet des SNPs, mais apparait trop conservatrice dans la cohorte simulée où le niveau d'apparentement est plus élevé et la structure de population plus complexe. Les concepts sur lesquels reposent ces méthodes pourraient être appliqués également pour le modèle de Cox mixte. Bien que la stratégie d'analyse en deux étapes utilisée pour l'analyse de la récurrence des événements dans la GWAS apparaisse la plus pertinente au premier abord, il pourrait être intéressant de comparer les résultats obtenus avec ceux des méthodes Offset et AMLE déclinées pour le modèle de Cox mixte.

Une analyse des signaux de sélection naturelle a été réalisée dans ces mêmes cohortes en utilisant une approche basée sur les haplotypes longs qui permet de détecter des signaux de sélection récente dans les populations humaines. Les premières analyses croisant les informations de sélection naturelle et d'association au sein de fenêtres non chevauchantes de 100 kb ne montrent pas d'enrichissement en signaux de sélection naturelle dans les régions les plus associées dans la GWAS. La présence d'un tel enrichissement aurait pu nous indiquer que les gènes qui jouent un rôle dans le paludisme non grave sont la cible également de la sélection naturelle. L'analyse plus approfondie des régions présentant les plus forts signaux de sélection ou les plus forts signaux d'association identifie plusieurs régions ou gènes d'intérêt potentiel, mais ces résultats doivent encore être confirmés par un examen plus approfondi des signaux d'association. Nous envisageons de poursuivre cette analyse des signaux de sélection naturelle en lien avec le paludisme simple en considérant l'information de sélection naturelle au niveau du SNP et non plus au niveau d'une fenêtre génomique. Cette approche

apparait plus adaptée pour mettre en évidence les signaux d'association en lien avec le paludisme et plus pertinente également pour étudier la co-localisation des signaux, qui est évaluée alors en combinant simplement les p-valeurs des tests d'association et de sélection. Une autre piste pour poursuivre ces analyses est de compléter les analyses déjà réalisées par un test XP-EHH en utilisant une population de référence non-exposée au paludisme. Dans les bases de données génomiques publiques actuellement disponibles, nous n'avons pas identifié de population presentant une histoire démographique proche des populations du Sud Bénin et qui n'aurait pas été exposée au paludisme. Une solution pourrait être de prendre une population africaine plus éloignée telles que les Maasaï et Kikuyu utilisées comme populations de référence dans l'étude d'Ayodo et *al.*(Ayodo et al., 2007), non exposées au paludisme car vivant en altitude au Kenya, ou une population d'origine Bantu ayant migré de façon précoce dans une région plus au sud de l'Afrique, où le paludisme n'est pas endémique (Patin et al., 2017).

Des développements méthodologiques sont encore nécessaires à la fois au niveau de la co-localisation des signaux de sélection et d'association et au niveau du modèle de Cox mixte. Ils pourront servir à d'autres études génétiques portant sur des suivis longitudinaux, ou sur des populations fortement stratifiées, telles que celles mises en place au Bénin par l'UMR216.

RÉFÉRENCES

Abel, L., Cot, M., Mulder, L., Carnevale, P., & Feingold, J. (1992). Segregation analysis detects a major gene controlling blood infection levels in human malaria. *American Journal of Human Genetics*, *50*(6), 1308‑1317.

Accrombessi, M., Ouédraogo, S., Agbota, G. C., Gonzalez, R., Massougbodji, A., Menéndez, C., & Cot, M. (2015). Malaria in Pregnancy Is a Predictor of Infant Haemoglobin Concentrations during the First Year of Life in Benin, West Africa. *PloS One*, *10*(6), e0129510. https://doi.org/10.1371/journal.pone.0129510

Ackerman, H., Usen, S., Jallow, M., Sisay-Joof, F., Pinder, M., & Kwiatkowski, D. P. (2005). A comparison of case-control and family-based association methods : The example of sickle-cell and malaria. *Annals of Human Genetics*, *69*(Pt 5), 559‑565. https://doi.org/10.1111/j.1529-8817.2005.00180.x

Aidoo, M., McElroy, P. D., Kolczak, M. S., Terlouw, D. J., ter Kuile, F. O., Nahlen, B., Lal, A. A., & Udhayakumar, V. (2001). Tumor necrosis factor-alpha promoter variant 2 (TNF2) is associated with pre-term delivery, infant mortality, and malaria morbidity in western Kenya : Asembo Bay Cohort Project IX. *Genetic Epidemiology*, *21*(3), 201‑211. https://doi.org/10.1002/gepi.1029

Aidoo, Michael, Terlouw, D. J., Kolczak, M. S., McElroy, P. D., ter Kuile, F. O., Kariuki, S., Nahlen, B. L., Lal, A. A., & Udhayakumar, V. (2002). Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet (London, England)*, *359*(9314), 1311‑1312. https://doi.org/10.1016/S0140-6736(02)08273-9

Alarcón-Riquelme, M. E., Ziegler, J. T., Molineros, J., Howard, T. D., Moreno-Estrada, A., Sánchez-Rodríguez, E., Ainsworth, H. C., Ortiz-Tello, P., Comeau, M. E., Rasmussen, A., Kelly, J. A., Adler, A., Acevedo-Vázquez, E. M., Cucho-Venegas, J. M., García-De la Torre, I., Cardiel, M. H., Miranda, P., Catoggio, L. J., Maradiaga-Ceceña, M., … Jacob, C. O. (2016). Genome-Wide Association Study in an Amerindian Ancestry Population Reveals Novel Systemic Lupus Erythematosus Risk Loci and the Role of European Admixture. *Arthritis & Rheumatology (Hoboken, N.J.)*, *68*(4), 932‑943. https://doi.org/10.1002/art.39504

Albrechtsen, A., Moltke, I., & Nielsen, R. (2010). Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. *Genetics*, *186*(1), 295‑308. https://doi.org/10.1534/genetics.110.113977

Allen, S. J., Bennett, S., Riley, E. M., Rowe, P. A., Jakobsen, P. H., O'Donnell, A., & Greenwood, B. M. (1992). Morbidity from malaria and immune responses to defined Plasmodium falciparum antigens in children with sickle cell trait in The Gambia. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *86*(5), 494‑498.

Allison, A. C. (1954). Protection afforded by sickle-cell trait against subtertian malareal infection. *British Medical Journal*, *1*(4857), 290‑294.

Allison, A. C., & Clyde, D. F. (1961). Malaria in African children with deficient erythrocyte glucose-6-phosphate dehydrogenase. *British Medical Journal*, *1*(5236), 1346‑1349.

Amato, D., & Booth, P. B. (1977). Hereditary ovalocytosis in Melanesians. *Papua and New Guinea Medical Journal*, *20*(1), 26‑32.

Andersen, K. G., Shylakhter, I., Tabrizi, S., Grossman, S. R., Happi, C. T., & Sabeti, P. C. (2012). Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1590), 868‑877. https://doi.org/10.1098/rstb.2011.0299

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, *5*(9), 1564‑1573. https://doi.org/10.1038/nprot.2010.116

Apinjoh, T. O., Anchang-Kimbi, J. K., Njua-Yafi, C., Ngwai, A. N., Mugri, R. N., Clark, T. G., Rockett, K. A., Kwiatkowski, D. P., Achidi, E. A., & MalariaGEN Consortium. (2014). Association of candidate gene polymorphisms and TGF-beta/IL-10 levels with malaria in three regions of Cameroon : A case-control study. *Malaria Journal*, *13*, 236. https://doi.org/10.1186/1475-2875-13-236

Ashley, E. A., Dhorda, M., Fairhurst, R. M., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Anderson, J. M., Mao, S., Sam, B., Sopha, C., Chuor, C. M., Nguon, C., Sovannaroth, S., Pukrittayakamee, S., Jittamala, P., Chotivanich, K., Chutasmit, K., Suchatsoonthorn, C., … Tracking Resistance to Artemisinin Collaboration (TRAC). (2014). Spread of artemisinin resistance in Plasmodium falciparum malaria. *The New England Journal of Medicine*, *371*(5), 411‑423. https://doi.org/10.1056/NEJMoa1314981

Atkinson, S. H., Mwangi, T. W., Uyoga, S. M., Ogada, E., Macharia, A. W., Marsh, K., Prentice, A. M., & Williams, T. N. (2007). The haptoglobin 2-2 genotype is associated with a reduced incidence of Plasmodium falciparum malaria in children on the coast of Kenya. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *44*(6), 802‑809. https://doi.org/10.1086/511868

Aulchenko, Y. S., de Koning, D.-J., & Haley, C. (2007). Genomewide rapid association using mixed model and regression : A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, *177*(1), 577‑585. https://doi.org/10.1534/genetics.107.075614

Ayodo, G., Price, A. L., Keinan, A., Ajwang, A., Otieno, M. F., Orago, A. S. S., Patterson, N., & Reich, D. (2007). Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *American Journal of Human Genetics*, *81*(2), 234‑242. https://doi.org/10.1086/519221

Baaklini, S., Afridi, S., Nguyen, T. N., Koukouikila-Koussounda, F., Ndounga, M., Imbert, J., Torres, M., Pradel, L., Ntoumi, F., & Rihet, P. (2017). 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*, *12*(11), e0187818. https://doi.org/10.1371/journal.pone.0187818

Band, G., Le, Q. S., Jostins, L., Pirinen, M., Kivinen, K., Jallow, M., Sisay-Joof, F., Bojang, K., Pinder, M., Sirugo, G., Conway, D. J., Nyirongo, V., Kachala, D., Molyneux, M., Taylor, T., Ndila, C., Peshu, N., Marsh, K., Williams, T. N., … Malaria Genomic Epidemiological Network. (2013). Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genetics*, *9*(5), e1003509. https://doi.org/10.1371/journal.pgen.1003509

Barbier, M., Delahaye, N. F., Fumoux, F., & Rihet, P. (2008). Family-based association of a low producing lymphotoxin-alpha allele with reduced Plasmodium falciparum parasitemia. *Microbes and Infection*, *10*(6), 673‑679. https://doi.org/10.1016/j.micinf.2008.03.001

Basu, M., Maji, A. K., Chakraborty, A., Banerjee, R., Mullick, S., Saha, P., Das, S., Kanjilal, S. D., & Sengupta, S. (2010). Genetic association of Toll-like-receptor 4 and tumor necrosis factor-alpha polymorphisms with Plasmodium falciparum blood infection levels. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *10*(5), 686‑696. https://doi.org/10.1016/j.meegid.2010.03.008

Beet, E. A. (1949). The genetics of the sickle-cell trait in a Bantu tribe. *Annals of Eugenics*, *14*(4), 279‑284. https://doi.org/10.1111/j.1469-1809.1947.tb02402.x

Berger, S. S., Turner, L., Wang, C. W., Petersen, J. E. V., Kraft, M., Lusingu, J. P. A., Mmbando, B., Marquard, A. M., Bengtsson, D. B. A. C., Hviid, L., Nielsen, M. A., Theander, T. G., & Lavstsen, T. (2013). Plasmodium falciparum expressing domain cassette 5 type PfEMP1 (DC5-PfEMP1) bind PECAM1. *PloS One*, *8*(7), e69117. https://doi.org/10.1371/journal.pone.0069117

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briët, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., … Gething, P. W. (2015). The effect of

malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, *526*(7572), 207‑211. https://doi.org/10.1038/nature15535

Bienzle, U., Ayeni, O., Lucas, A. O., & Luzzatto, L. (1972). Glucose-6-phosphate dehydrogenase and malaria. Greater resistance of females heterozygous for enzyme deficiency and of males with non-deficient variant. *Lancet (London, England)*, *1*(7742), 107‑110.

Boldt, A. B. W., Messias-Reason, I. J., Lell, B., Issifou, S., Pedroso, M. L. A., Kremsner, P. G., & Kun, J. F. J. (2009). Haplotype specific-sequencing reveals MBL2 association with asymptomatic Plasmodium falciparum infection. *Malaria Journal*, *8*, 97. https://doi.org/10.1186/1475-2875-8-97

Bouaziz, O., Courtin, D., Cottrell, G., Milet, J., Nuel, G., & Garcia, A. (2018). Is Placental Malaria a Long-term Risk Factor for Mild Malaria Attack in Infancy? Revisiting a Paradigm. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *66*(6), 930‑935. https://doi.org/10.1093/cid/cix899

Bousema, T., Drakeley, C., Gesase, S., Hashim, R., Magesa, S., Mosha, F., Otieno, S., Carneiro, I., Cox, J., Msuya, E., Kleinschmidt, I., Maxwell, C., Greenwood, B., Riley, E., Sauerwein, R., Chandramohan, D., & Gosling, R. (2010). Identification of hot spots of malaria transmission for targeted malaria control. *The Journal of Infectious Diseases*, *201*(11), 1764‑1774. https://doi.org/10.1086/652456

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., & Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790‑1797. https://doi.org/10.1101/gr.137323.112

Bradburd, G. S., Ralph, P. L., & Coop, G. M. (2016). A Spatial Framework for Understanding Population Structure and Admixture. *PLoS Genetics*, *12*(1), e1005703. https://doi.org/10.1371/journal.pgen.1005703

Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *88*(421), 9‑25. JSTOR. https://doi.org/10.2307/2290687

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedón, J. C., Redline, S., Papanicolaou, G. J., Thornton, T. A., Laurie, C. C., Rice, K., & Lin, X. (2016). Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *American Journal of Human Genetics*, *98*(4), 653‑666. https://doi.org/10.1016/j.ajhg.2016.02.012

Choi, H. W., Breman, J. G., Teutsch, S. M., Liu, S., Hightower, A. W., & Sexton, J. D. (1995). The effectiveness of insecticide-impregnated bed nets in reducing cases of malaria infection : A meta-analysis of published results. *The American Journal of Tropical Medicine and Hygiene*, *52*(5), 377‑382. https://doi.org/10.4269/ajtmh.1995.52.377

Chua, C. L. L., Brown, G., Hamilton, J. A., Rogerson, S., & Boeuf, P. (2013). Monocytes and macrophages in malaria : Protection or pathology? *Trends in Parasitology*, *29*(1), 26‑34. https://doi.org/10.1016/j.pt.2012.10.002

Claire Dandine-Roulland. (2016). *Modélisation de la composante génétique des maladies humaines : Données familiales et Modèles Mixtes.* Génétique humaine. Université Paris-Saclay.

Costa, A. G., Ramasawmy, R., Ibiapina, H. N. S., Sampaio, V. S., Xábregas, L. A., Brasil, L. W., Tarragô, A. M., Almeida, A. C. G., Kuehn, A., Vitor-Silva, S., Melo, G. C., Siqueira, A. M., Monteiro, W. M., Lacerda, M. V. G., & Malheiro, A. (2017). Association of TLR variants with susceptibility to Plasmodium vivax malaria and parasitemia in the Amazon region of Brazil. *PloS One*, *12*(8), e0183840. https://doi.org/10.1371/journal.pone.0183840

Cottrell, G., Kouwaye, B., Pierrat, C., le Port, A., Bouraïma, A., Fonton, N., Hounkonnou, M. N., Massougbodji, A., Corbel, V., & Garcia, A. (2012). Modeling the influence of local environmental factors on malaria transmission in Benin and its implications for cohort study. *PloS One*, *7*(1), e28812. https://doi.org/10.1371/journal.pone.0028812

Cox, S. E., Doherty, C., Atkinson, S. H., Nweneka, C. V., Fulford, A. J. C., Ghattas, H., Rockett, K. A., Kwiatkowski, D. P., & Prentice, A. M. (2007). Haplotype association between haptoglobin (Hp2) and Hp promoter SNP (A-61C) may explain previous controversy of haptoglobin and malaria protection. *PloS One*, *2*(4), e362. https://doi.org/10.1371/journal.pone.0000362

Cramer, J. S. (2007). Robustness of logit analysis : Unobserved heterogeneity and mis-specified disturbances. *Oxford Bulletin of Economics and Statistics*, *69*(4), 545‑555.

Crompton, P. D., Moebius, J., Portugal, S., Waisberg, M., Hart, G., Garver, L. S., Miller, L. H., Barillas, C., & Pierce, S. K. (2014). Malaria immunity in man and mosquito : Insights into unsolved mysteries of a deadly infectious disease. *Annual review of immunology*, *32*, 157‑187. https://doi.org/10.1146/annurev-immunol-032713-120220

Crompton, P. D., Traore, B., Kayentao, K., Doumbo, S., Ongoiba, A., Diakite, S. A. S., Krause, M. A., Doumtabe, D., Kone, Y., Weiss, G., Huang, C.-Y., Doumbia, S., Guindo, A., Fairhurst, R. M., Miller, L. H., Pierce, S. K., & Doumbo, O. K. (2008). Sickle cell trait is associated with a delayed onset of malaria : Implications for time-to-event analysis in clinical studies of malaria. *The Journal of Infectious Diseases*, *198*(9), 1265‑1275. https://doi.org/10.1086/592224

d'Almeida, T. C., Sadissou, I., Milet, J., Cottrell, G., Mondière, A., Avokpaho, E., Gineau, L., Sabbagh, A., Massougbodji, A., Moutairou, K., Donadi, E. A., Favier, B., Carosella, E., Moreau, P., Rouas-Freiss, N., Courtin, D., & Garcia, A. (2017). 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*, *12*(2), e0171117. https://doi.org/10.1371/journal.pone.0171117

Damien, G. B., Djènontin, A., Rogier, C., Corbel, V., Bangana, S. B., Chandre, F., Akogbéto, M., Kindé-Gazard, D., Massougbodji, A., & Henry, M.-C. (2010). Malaria infection and disease in an area with pyrethroid-resistant vectors in southern Benin. *Malaria Journal*, *9*, 380. https://doi.org/10.1186/1475-2875-9-380

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2017). Next-generation genotype imputation service and methods. *Nature genetics*, *48*(10), 1284‑1287. https://doi.org/10.1038/ng.3656.Next-generation

Delahaye, N. F., Barbier, M., Fumoux, F., & Rihet, P. (2007). Association analyses of NCR3 polymorphisms with P. falciparum mild malaria. *Microbes and Infection*, *9*(2), 160‑166. https://doi.org/10.1016/j.micinf.2006.11.002

Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5‑6. https://doi.org/10.1038/nmeth.2307

Demenais, F., Martinez, M., & Lathrop, M. (1996). Méthodes statistiques pour identifier les gènes dans les maladies multifactorielles. *Annales de l'Institut Pasteur / Actualités*, *7*(1), 3‑12. https://doi.org/10.1016/0924-4204(96)82110-X

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997‑1004. https://doi.org/10.1111/j.0006-341x.1999.00997.x

do Sambo, M. R., Penha-Gonçalves, C., Trovoada, M. J., Costa, J., Lardoeyt, R., & Coutinho, A. (2015). Quantitative trait locus analysis of parasite density reveals that HbS gene carriage protects severe malaria patients against Plasmodium falciparum hyperparasitaemia. *Malaria Journal*, *14*, 393. https://doi.org/10.1186/s12936-015-0920-z

Doolan, D. L., Dobaño, C., & Baird, J. K. (2009). Acquired Immunity to Malaria. *Clinical Microbiology Reviews*, *22*(1), 13‑36. https://doi.org/10.1128/CMR.00025-08

Driss, A., Hibbert, J. M., Wilson, N. O., Iqbal, S. A., Adamkiewicz, T. V., & Stiles, J. K. (2011). Genetic polymorphisms linked to susceptibility to malaria. *Malaria Journal*, *10*, 271. https://doi.org/10.1186/1475-2875-10-271

Elagib, A. A., Kider, A. O., Akerström, B., & Elbashir, M. I. (1998). Association of the haptoglobin phenotype (1-1) with falciparum malaria in Sudan. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *92*(3), 309‑311.

Enevold, A., Lusingu, J. P., Mmbando, B., Alifrangis, M., Lemnge, M. M., Bygbjerg, I. C., Theander, T. G., & Vestergaard, L. S. (2008). Reduced risk of uncomplicated malaria episodes in children with alpha+-thalassemia in northeastern Tanzania. *The American Journal of Tropical Medicine and Hygiene*, *78*(5), 714‑720.

Esposito, S., Molteni, C. G., Zampiero, A., Baggi, E., Lavizzari, A., Semino, M., Daleno, C., Groppo, M., Scala, A., Terranova, L., Miozzo, M., Pelucchi, C., & Principi, N. (2012). Role of polymorphisms of toll-like receptor (TLR) 4, TLR9, toll-interleukin 1 receptor domain containing adaptor protein (TIRAP) and FCGR2A genes in malaria susceptibility and severity in Burundian children. *Malaria Journal*, *11*, 196. https://doi.org/10.1186/1475-2875-11-196

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405‑1413.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, *31*(5), 1275‑1291. https://doi.org/10.1093/molbev/msu077

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I., & Pritchard, J. K. (2016). Detection of human adaptation during the past 2000 years. *Science (New York, N.Y.)*, *354*(6313), 760‑764. https://doi.org/10.1126/science.aag0776

Flatz, G. (1967). Hemoglobin E : Distribution and population dynamics. *Humangenetik*, *3*(3), 189‑234. https://doi.org/10.1007/bf00273124

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A. K., Keenan, S., … Searle, S. M. J. (2013). Ensembl 2013. *Nucleic Acids Research*, *41*(Database issue), D48-55. https://doi.org/10.1093/nar/gks1236

Flori, L., Delahaye, N. F., Iraqi, F. A., Hernandez-Valladares, M., Fumoux, F., & Rihet, P. (2005). TNF as a malaria candidate gene : Polymorphism-screening and family-based association analysis of mild malaria attack and parasitemia in Burkina Faso. *Genes and Immunity*, *6*(6), 472‑480. https://doi.org/10.1038/sj.gene.6364231

François, O., Martins, H., Caye, K., & Schoville, S. D. (2016). Controlling false discoveries in genome scans for selection. *Molecular Ecology*, *25*(2), 454‑469. https://doi.org/10.1111/mec.13513

Fry, A. E., Ghansa, A., Small, K. S., Palma, A., Auburn, S., Diakite, M., Green, A., Campino, S., Teo, Y. Y., Clark, T. G., Jeffreys, A. E., Wilson, J., Jallow, M., Sisay-Joof, F., Pinder, M., Griffiths, M. J., Peshu, N., Williams, T. N., Newton, C. R., … Kwiatkowski, D. P. (2009). Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. *Human Molecular Genetics*, *18*(14), 2683‑2692. https://doi.org/10.1093/hmg/ddp192

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics*, *7*(11), e1002355. https://doi.org/10.1371/journal.pgen.1002355

Gail, M. H., Wieand, S., & PIANTADOSI, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, *71*(3), 431‑444. https://doi.org/10.1093/biomet/71.3.431

Gao, H., Wu, X., Sun, Y., Zhou, S., Silberstein, L. E., & Zhu, Z. (2012). Suppression of homeobox transcription factor VentX promotes expansion of human hematopoietic stem/multipotent progenitor cells. *The Journal of Biological Chemistry*, *287*(35), 29979‑29987. https://doi.org/10.1074/jbc.M112.383018

Garcia, A., Cot, M., Chippaux, J. P., Ranque, S., Feingold, J., Demenais, F., & Abel, L. (1998). Genetic control of blood infection levels in human malaria : Evidence for a complex genetic model. *The American Journal of Tropical Medicine and Hygiene*, *58*(4), 480‑488.

Garcia, A., Marquet, S., Bucheton, B., Hillaire, D., Cot, M., Fievet, N., Dessein, A. J., & Abel, L. (1998). Linkage analysis of blood Plasmodium falciparum levels : Interest of the 5q31-q33 chromosome region. *The American Journal of Tropical Medicine and Hygiene*, *58*(6), 705‑709.

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genetics*, *11*(2), e1005004. https://doi.org/10.1371/journal.pgen.1005004

Ghanchi, N. K., Hasan, Z., Islam, M., & Beg, M. A. (2015). MAD 20 alleles of merozoite surface protein-1 (msp-1) are associated with severe Plasmodium falciparum malaria in Pakistan. *Journal of Microbiology, Immunology, and Infection = Wei Mian Yu Gan Ran Za Zhi*, *48*(2), 213‑218. https://doi.org/10.1016/j.jmii.2014.01.004

Gichohi-Wainaina, W. N., Melse-Boonstra, A., Feskens, E. J., Demir, A. Y., Veenemans, J., & Verhoef, H. (2015). Tumour necrosis factor allele variants and their association with the occurrence and severity of malaria in African children : A longitudinal study. *Malaria Journal*, *14*, 249. https://doi.org/10.1186/s12936-015-0767-3

Gomes, P. S., Bhardwaj, J., Rivera-Correa, J., Freire-De-Lima, C. G., & Morrot, A. (2016). Immune Escape Strategies of Malaria Parasites. *Frontiers in Microbiology*, *7*. https://doi.org/10.3389/fmicb.2016.01617

Gong, L., Maiteki-Sebuguzi, C., Rosenthal, P. J., Hubbard, A. E., Drakeley, C. J., Dorsey, G., & Greenhouse, B. (2012). Evidence for both innate and acquired mechanisms of protection from Plasmodium falciparum in children with sickle cell trait. *Blood*, *119*(16), 3808‑3814. https://doi.org/10.1182/blood-2011-08-371062

González, R., Mombo-Ngoma, G., Ouédraogo, S., Kakolwa, M. A., Abdulla, S., Accrombessi, M., Aponte, J. J., Akerey-Diop, D., Basra, A., Briand, V., Capan, M., Cot, M., Kabanywanyi, A. M., Kleine, C., Kremsner, P. G., Macete, E., Mackanga, J.-R., Massougbodgi, A., Mayor, A., … Menéndez, C. (2014a). Intermittent preventive treatment of malaria in pregnancy with mefloquine in HIV-negative women : A multicentre randomized controlled trial. *PLoS Medicine*, *11*(9), e1001733. https://doi.org/10.1371/journal.pmed.1001733

González, R., Mombo-Ngoma, G., Ouédraogo, S., Kakolwa, M. A., Abdulla, S., Accrombessi, M., Aponte, J. J., Akerey-Diop, D., Basra, A., Briand, V., Capan, M., Cot, M., Kabanywanyi, A. M., Kleine, C., Kremsner, P. G., Macete, E., Mackanga, J.-R., Massougbodgi, A., Mayor, A., … Menéndez, C. (2014b). Intermittent preventive treatment of malaria in pregnancy with mefloquine in HIV-negative women : A multicentre randomized controlled trial. *PLoS Medicine*, *11*(9), e1001733. https://doi.org/10.1371/journal.pmed.1001733

Gouveia, M. H., Bergen, A. W., Borda, V., Nunes, K., Leal, T. P., Ogwang, M. D., Yeboah, E. D., Mensah, J. E., Kinyera, T., Otim, I., Nabalende, H., Legason, I. D., Mpoloka, S. W., Mokone, G. G., Kerchan, P., Bhatia, K., Reynolds, S. J., Birtwum, R. B., Adjei, A. A., … Mbulaiteye, S. M. (2019). Genetic signatures of gene flow and malaria-driven natural selection in sub-Saharan populations of the « endemic Burkitt Lymphoma belt ». *PLoS Genetics*, *15*(3), Article 3. https://doi.org/10.1371/journal.pgen.1008027

Greenwood, B. M. (1989). The microepidemiology of malaria and its importance to malaria control. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *83 Suppl*, 25‑29.

Grossman, S. R., Andersen, K. G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D. J., Griesemer, D., Karlsson, E. K., Wong, S. H., Cabili, M., Adegbola, R. A., Bamezai, R. N. K., Hill, A. V. S., Vannberg, F. O., Rinn, J. L., 1000 Genomes Project, Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2013). Identifying recent adaptations in large-scale genomic data. *Cell*, *152*(4), 703‑713. https://doi.org/10.1016/j.cell.2013.01.035

Grossman, S. R., Shlyakhter, I., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science (New York, N.Y.)*, *327*(5967), 883‑886. https://doi.org/10.1126/science.1183863

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R. S., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., … Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, *517*(7534), 327‑332. https://doi.org/10.1038/nature13997

Haldane, J. B. S. (1949). The Rate of Mutation of Human Genes. *Hereditas*, *35*(S1), 267‑273. https://doi.org/10.1111/j.1601-5223.1949.tb03339.x

Hananantachai, H., Patarapotikul, J., Ohashi, J., Naka, I., Looareesuwan, S., & Tokunaga, K. (2005). Polymorphisms of the HLA-B and HLA-DRB1 genes in Thai malaria patients. *Japanese Journal of Infectious Diseases*, *58*(1), 25‑28.

Hanchard, N. A., Rockett, K. A., Spencer, C., Coop, G., Pinder, M., Jallow, M., Kimber, M., McVean, G., Mott, R., & Kwiatkowski, D. P. (2006). Screening for recently selected alleles by analysis of human haplotype similarity. *American Journal of Human Genetics*, *78*(1), 153‑159. https://doi.org/10.1086/499252

Hermisson, J., & Pennings, P. S. (2005). Soft sweeps : Molecular population genetics of adaptation from standing genetic variation. *Genetics*, *169*(4), 2335‑2352. https://doi.org/10.1534/genetics.104.036947

Heung, L. J., Luberto, C., & Del Poeta, M. (2006). Role of Sphingolipids in Microbial Pathogenesis. *Infection and Immunity*, *74*(1), 28‑39. https://doi.org/10.1128/IAI.74.1.28-39.2006

Hill, A. V., Allsopp, C. E., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J., & Greenwood, B. M. (1991). Common west African HLA antigens are associated with protection from severe malaria. *Nature*, *352*(6336), 595‑600. https://doi.org/10.1038/352595a0

Hobbs, M. R., Udhayakumar, V., Levesque, M. C., Booth, J., Roberts, J. M., Tkachuk, A. N., Pole, A., Coon, H., Kariuki, S., Nahlen, B. L., Mwaikambo, E. D., Lal, A. L., Granger, D. L., Anstey, N. M., & Weinberg, J. B. (2002). A new NOS2 promoter polymorphism associated with increased nitric oxide production and protection from severe malaria in Tanzanian and Kenyan children. *Lancet (London, England)*, *360*(9344), 1468‑1475. https://doi.org/10.1016/S0140-6736(02)11474-7

Holmberg, V., Schuster, F., Dietz, E., Sagarriga Visconti, J. C., Anemana, S. D., Bienzle, U., & Mockenhaupt, F. P. (2008). Mannose-binding lectin variant associated with severe malaria in young African children. *Microbes and Infection*, *10*(4), 342‑348. https://doi.org/10.1016/j.micinf.2007.12.008

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, *18*(2), 337‑338. https://doi.org/10.1093/bioinformatics/18.2.337

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299‑1320. https://doi.org/10.1038/nature04226

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., Kivinen, K., Bojang, K. A., Conway, D. J., Pinder, M., Sirugo, G., Sisay-Joof, F., Usen, S., Auburn, S., Bumpstead, S. J., Campino, S., Coffey, A., Dunham, A., Fry, A. E., … Malaria Genomic Epidemiology Network. (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genetics*, *41*(6), 657‑665. https://doi.org/10.1038/ng.388

Jepson, A. P., Banya, W. A., Sisay-Joof, F., Hassan-King, M., Bennett, S., & Whittle, H. C. (1995). Genetic regulation of fever in Plasmodium falciparum malaria in Gambian twin children. *The Journal of Infectious Diseases*, *172*(1), 316‑319. https://doi.org/10.1093/infdis/172.1.316

Jepson, A., Sisay-Joof, F., Banya, W., Hassan-King, M., Frodsham, A., Bennett, S., Hill, A. V., & Whittle, H. (1997). Genetic linkage of mild malaria to the major histocompatibility complex in Gambian children : Study of affected sibling pairs. *BMJ (Clinical Research Ed.)*, *315*(7100), 96‑97.

Kakuru, A., Staedke, S. G., Dorsey, G., Rogerson, S., & Chandramohan, D. (2019). Impact of Plasmodium falciparum malaria and intermittent preventive treatment of malaria in pregnancy on the risk of malaria in infants : A systematic review. *Malaria Journal*, *18*(1), 304. https://doi.org/10.1186/s12936-019-2943-3

Kanchan, K., Pati, S. S., Mohanty, S., Mishra, S. K., Sharma, S. K., Awasthi, S., Indian Genome Variation Consortium, Venkatesh, V., & Habib, S. (2015). Polymorphisms in host genes encoding NOSII, C-reactive protein, and adhesion molecules thrombospondin and E-selectin are risk factors for Plasmodium falciparum malaria in India. *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology*, *34*(10), 2029‑2039. https://doi.org/10.1007/s10096-015-2448-0

Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nature Reviews. Genetics*, *15*(6), 379‑393. https://doi.org/10.1038/nrg3734

Kim, D., Fedak, K., & Kramer, R. (2012). Reduction of malaria prevalence by indoor residual spraying : A meta-regression analysis. *The American Journal of Tropical Medicine and Hygiene*, *87*(1), 117‑124. https://doi.org/10.4269/ajtmh.2012.11-0620

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310‑315. https://doi.org/10.1038/ng.2892

Knox, T. B., Juma, E. O., Ochomo, E. O., Pates Jamet, H., Ndungo, L., Chege, P., Bayoh, N. M., N'Guessan, R., Christian, R. N., Hunt, R. H., & Coetzee, M. (2014). An online tool for mapping insecticide resistance in major Anopheles vectors of human malaria parasites and review of resistance status for the Afrotropical region. *Parasites & Vectors*, *7*, 76. https://doi.org/10.1186/1756-3305-7-76

Kreuels, B., Ehrhardt, S., Kreuzberg, C., Adjei, S., Kobbe, R., Burchard, G. D., Ehmen, C., Ayim, M., Adjei, O., & May, J. (2009). Sickle cell trait (HbAS) and stunting in children below two years of age in an area of high malaria transmission. *Malaria Journal*, *8*, 16. https://doi.org/10.1186/1475-2875-8-16

Kumaratilake, L. M., Ferrante, A., & Rzepczyk, C. M. (1990). Tumor necrosis factor enhances neutrophil-mediated killing of Plasmodium falciparum. *Infection and Immunity*, *58*(3), 788‑793.

Kun, J. F., Mordmüller, B., Lell, B., Lehman, L. G., Luckner, D., & Kremsner, P. G. (1998). Polymorphism in promoter region of inducible nitric oxide synthase gene and protection against malaria. *The Lancet*, *351*(9098), 265‑266. https://doi.org/10.1016/S0140-6736(05)78273-8

Kwiatkowski, D. P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *American Journal of Human Genetics*, *77*(2), 171‑192. https://doi.org/10.1086/432519

Labaied, M., Dagan, A., Dellinger, M., Gèze, M., Egée, S., Thomas, S. L., Wang, C., Gatt, S., & Grellier, P. (2004). Anti-Plasmodium activity of ceramide analogs. *Malaria Journal*, *3*, 49. https://doi.org/10.1186/1475-2875-3-49

Labie, D., Richin, C., Pagnier, J., Gentilini, M., & Nagel, R. L. (1984). Hemoglobins S and C in Upper Volta. *Human Genetics*, *65*(3), 300‑302. https://doi.org/10.1007/bf00286522

Laval, G., Peyrégne, S., Zidane, N., Harmant, C., Renaud, F., Patin, E., Prugnolle, F., & Quintana-Murci, L. (2019). Recent Adaptive Acquisition by African Rainforest Hunter-Gatherers of the Late Pleistocene Sickle-Cell Mutation Suggests Past Differences in Malaria Exposure. *American Journal of Human Genetics*, *104*(3), 553‑561. https://doi.org/10.1016/j.ajhg.2019.02.007

Lavstsen, T., Turner, L., Saguti, F., Magistrado, P., Rask, T. S., Jespersen, J. S., Wang, C. W., Berger, S. S., Baraka, V., Marquard, A. M., Seguin-Orlando, A., Willerslev, E., Gilbert, M. T. P., Lusingu, J., & Theander, T. G. (2012). Plasmodium falciparum erythrocyte membrane protein 1 domain cassettes 8 and 13 are associated with severe malaria in children. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(26), E1791-1800. https://doi.org/10.1073/pnas.1120455109

Le Hesran, J. Y., Personne, I., Personne, P., Fievet, N., Dubois, B., Beyemé, M., Boudin, C., Cot, M., & Deloron, P. (1999). Longitudinal study of Plasmodium falciparum infection and immune responses in infants with or without the sickle cell trait. *International Journal of Epidemiology*, *28*(4), 793‑798.

Le Port, A., Cottrell, G., Martin-Prevel, Y., Migot-Nabias, F., Cot, M., & Garcia, A. (2012). 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*, *2*(2), e000342. https://doi.org/10.1136/bmjopen-2011-000342

Le Port, A., Watier, L., Cottrell, G., Ouédraogo, S., Dechavanne, C., Pierrat, C., Rachas, A., Bouscaillou, J., Bouraima, A., Massougbodji, A., Fayomi, B., Thiébaut, A., Chandre, F., Migot-Nabias, F., Martin-Prevel, Y., Garcia, A., & Cot, M. (2011). Infections in infants during the first 12 months of life : Role of placental malaria and environmental factors. *PloS One*, *6*(11), e27516. https://doi.org/10.1371/journal.pone.0027516

Leffler, E. M., Band, G., Busby, G. B. J., Kivinen, K., Le, Q. S., Clarke, G. M., Bojang, K. A., Conway, D. J., Jallow, M., Sisay-Joof, F., Bougouma, E. C., Mangano, V. D., Modiano, D., Sirima, S. B., Achidi, E., Apinjoh, T. O., Marsh, K., Ndila, C. M., Peshu, N., … Malaria Genomic Epidemiology Network. (2017). Resistance to malaria through structural variation of red blood cell invasion receptors. *Science (New York, N.Y.)*, *356*(6343), Article 6343. https://doi.org/10.1126/science.aam6393

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833‑835. https://doi.org/10.1038/nmeth.1681

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., & Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*(3), 284‑290. https://doi.org/10.1038/ng.3190

Lyke, K. E. (2017). Steady progress toward a malaria vaccine. *Current Opinion in Infectious Diseases*, *30*(5), 463‑470. https://doi.org/10.1097/QCO.0000000000000393

Mackinnon, M. J., Mwangi, T. W., Snow, R. W., Marsh, K., & Williams, T. N. (2005). Heritability of malaria in Africa. *PLoS Medicine*, *2*(12), e340. https://doi.org/10.1371/journal.pmed.0020340

Maiga, B., Dolo, A., Campino, S., Sepulveda, N., Corran, P., Rockett, K. A., Troye-Blomberg, M., Doumbo, O. K., & Clark, T. G. (2014). Glucose-6-phosphate dehydrogenase polymorphisms and susceptibility to mild malaria in Dogon and Fulani, Mali. *Malaria Journal*, *13*, 270. https://doi.org/10.1186/1475-2875-13-270

Malaria Genomic Epidemiology Network. (2019). Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nature Communications*, *10*(1), 5732. https://doi.org/10.1038/s41467-019-13480-z

Malaria Genomic Epidemiology Network, Band, G., Rockett, K. A., Spencer, C. C. A., & Kwiatkowski, D. P. (2015). A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature*, *526*(7572), 253‑257. https://doi.org/10.1038/nature15390

malERA. (2017). malERA : An updated research agenda for insecticide and drug resistance in malaria elimination and eradication. *PLoS Medicine*, *14*(11), e1002450. https://doi.org/10.1371/journal.pmed.1002450

Mangano, V. D., Kabore, Y., Bougouma, E. C., Verra, F., Sepulveda, N., Bisseye, C., Santolamazza, F., Avellino, P., Tiono, A. B., Diarra, A., Nebie, I., Rockett, K. A., Sirima, S. B., Modiano, D., &

MalariaGEN Consortium. (2015). Novel Insights Into the Protective Role of Hemoglobin S and C Against Plasmodium falciparum Parasitemia. *The Journal of Infectious Diseases*, *212*(4), 626‑634. https://doi.org/10.1093/infdis/jiv098

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*, *26*(22), 2867‑2873. https://doi.org/10.1093/bioinformatics/btq559

Marquet, S. (2017). Overview of human genetic susceptibility to malaria : From parasitemia control to severe disease. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*. https://doi.org/10.1016/j.meegid.2017.06.001

Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, *44*(3), 243‑246. https://doi.org/10.1038/ng.1074

Meyer, C. G., May, J., Luty, A. J., Lell, B., & Kremsner, P. G. (2002). TNFalpha-308A associated with shorter intervals of Plasmodium falciparum reinfections. *Tissue Antigens*, *59*(4), 287‑292.

Migot-Nabias, F., Pelleau, S., Watier, L., Guitard, J., Toly, C., De Araujo, C., Ngom, M. I., Chevillard, C., Gaye, O., & Garcia, A. (2006). Red blood cell polymorphisms in relation to Plasmodium falciparum asymptomatic parasite densities and morbidity in Senegal. *Microbes and Infection*, *8*(9‑10), 2352‑2358. https://doi.org/10.1016/j.micinf.2006.03.021

Modiano, D., Luoni, G., Sirima, B. S., Simporé, J., Verra, F., Konaté, A., Rastrelli, E., Olivieri, A., Calissano, C., Paganotti, G. M., D'Urbano, L., Sanou, I., Sawadogo, A., Modiano, G., & Coluzzi, M. (2001). Haemoglobin C protects against clinical Plasmodium falciparum malaria. *Nature*, *414*(6861), 305‑308. https://doi.org/10.1038/35104556

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics*, *7*(3), 277‑318.

Nagel, R. L., & Roth, E. F. (1989). Malaria and red cell genetic defects. *Blood*, *74*(4), 1213‑1221.

Nawaz, S. K., Ahmed, B., Arshad, N., Rani, A., Rasool, H., & Arshad, M. (2015). Role of S180L polymorphism in etiology of malaria caused by Plasmodium falciparum in a small group of Pakistani population. *Bosnian Journal of Basic Medical Sciences*, *15*(4), 20‑23. https://doi.org/10.17305/bjbms.2015.413

Neel, J. V. (1949). The Inheritance of Sickle Cell Anemia. *Science (New York, N.Y.)*, *110*(2846), 64‑66. https://doi.org/10.1126/science.110.2846.64

Nguyen, T. N., Baaklini, S., Koukouikila-Koussounda, F., Ndounga, M., Torres, M., Pradel, L., Ntoumi, F., & Rihet, P. (2017). Association of a functional TNF variant with Plasmodium falciparum parasitaemia in a congolese population. *Genes and Immunity*, *18*(3), 152‑157. https://doi.org/10.1038/gene.2017.13

Ojurongbe, O., Funwei, R. I., Snyder, T. J., Farid, I., Aziz, N., Li, Y., Falade, C. O., & Thomas, B. N. (2018). Genetic variants of tumor necrosis factor-α -308G/A (rs1800629) but not Toll-interacting proteins or vitamin D receptor genes enhances susceptibility and severity of malaria infection. *Immunogenetics*, *70*(2), 135‑140. https://doi.org/10.1007/s00251-017-1032-4

Okebe, J., Amambua-Ngwa, A., Parr, J., Nishimura, S., Daswani, M., Takem, E. N., Affara, M., Ceesay, S. J., Nwakanma, D., & D'Alessandro, U. (2014). The prevalence of glucose-6-phosphate dehydrogenase deficiency in Gambian school children. *Malaria Journal*, *13*, 148. https://doi.org/10.1186/1475-2875-13-148

Okiro, E. A., Al-Taiar, A., Reyburn, H., Idro, R., Berkley, J. A., & Snow, R. W. (2009). Age patterns of severe paediatric malaria and their relationship to Plasmodium falciparum transmission intensity. *Malaria Journal*, *8*, 4. https://doi.org/10.1186/1475-2875-8-4

Osafo-Addo, A. D., Koram, K. A., Oduro, A. R., Wilson, M., Hodgson, A., & Rogers, W. O. (2008). HLA-DRB1*04 allele is associated with severe malaria in northern Ghana. *The American Journal of Tropical Medicine and Hygiene*, *78*(2), 251‑255.

Otto, T. D., Gilabert, A., Crellen, T., Böhme, U., Arnathau, C., Sanders, M., Oyola, S. O., Okouga, A. P., Boundenga, L., Willaume, E., Ngoubangoye, B., Moukodoum, N. D., Paupy, C., Durand, P., Rougeron, V., Ollomo, B., Renaud, F., Newbold, C., Berriman, M., & Prugnolle, F. (2018). Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria. *Nature microbiology*, *3*(6), 687‑697. https://doi.org/10.1038/s41564-018-0162-2

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J.-M., Pereira, J. B., Fernandes, V., Pereira, L., … Quintana-Murci, L. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science (New York, N.Y.)*, *356*(6337), 543‑546. https://doi.org/10.1126/science.aal1988

Perdry, H., & Dandine-Roulland, C. (2018). *Gaston : Genetic Data Handling (QC, GRM, LD, PCA) and Linear Mixed Models*.

Peyser, N. D., Freilino, M., Wang, L., Zeng, Y., Li, H., Johnson, D. E., & Grandis, J. R. (2016). Frequent promoter hypermethylation of PTPRT increases STAT3 activation and sensitivity to STAT3 inhibition in head and neck cancer. *Oncogene*, *35*(9), 1163‑1169. https://doi.org/10.1038/onc.2015.171

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., & Pritchard, J. K. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, *19*(5), 826‑837. https://doi.org/10.1101/gr.087577.108

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904‑909. https://doi.org/10.1038/ng1847

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews. Genetics*, *11*(7), 459‑463. https://doi.org/10.1038/nrg2813

Pritchard, J. K., Stephens, M., Rosenberg, N. A., & Donnelly, P. (2000). Association mapping in structured populations. *American Journal of Human Genetics*, *67*(1), 170‑181. https://doi.org/10.1086/302959

Pritchard, Jonathan K., Pickrell, J. K., & Coop, G. (2010). The genetics of human adaptation : Hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology: CB*, *20*(4), R208-215. https://doi.org/10.1016/j.cub.2009.11.055

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R., & Willer, C. J. (2010). LocusZoom : Regional visualization of genome-wide association scan results. *Bioinformatics*, *26*(18), 2336‑2337. https://doi.org/10.1093/bioinformatics/btq419

Pybus, M., Dall'Olio, G. M., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., Laayouni, H., Bertranpetit, J., & Engelken, J. (2014). 1000 Genomes Selection Browser 1.0 : A genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Research*, *42*(Database issue), D903-909. https://doi.org/10.1093/nar/gkt1188

Quaye, I. K., Ekuban, F. A., Goka, B. Q., Adabayeri, V., Kurtzhals, J. A., Gyan, B., Ankrah, N. A., Hviid, L., & Akanmori, B. D. (2000). Haptoglobin 1-1 is associated with susceptibility to severe Plasmodium falciparum malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *94*(2), 216‑219.

R Core Team. (2017). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org

Rabinowitz, D., & Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity*, *50*(4), 211‑223. https://doi.org/10.1159/000022918

Ravenhall, M., Campino, S., Sepúlveda, N., Manjurano, A., Nadjm, B., Mtove, G., Wangai, H., Maxwell, C., Olomi, R., Reyburn, H., Drakeley, C. J., Riley, E. M., Clark, T. G., & in collaboration with MalariaGEN. (2018). Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. *PLoS Genetics*, *14*(1), e1007172. https://doi.org/10.1371/journal.pgen.1007172

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD : Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886‑D894. https://doi.org/10.1093/nar/gky1016

Rihet, P., Abel, L., Traoré, Y., Traoré-Leroux, T., Aucan, C., & Fumoux, F. (1998). Human malaria : Segregation analysis of blood infection levels in a suburban area and a rural area in Burkina Faso. *Genetic Epidemiology*, *15*(5), 435‑450. https://doi.org/10.1002/(SICI)1098-2272(1998)15:5<435::AID-GEPI1>3.0.CO;2-#

Rihet, P., Traoré, Y., Abel, L., Aucan, C., Traoré-Leroux, T., & Fumoux, F. (1998). Malaria in humans : Plasmodium falciparum blood infection levels are linked to chromosome 5q31-q33. *American Journal of Human Genetics*, *63*(2), 498‑505. https://doi.org/10.1086/301967

Rihet, Pascal, Flori, L., Tall, F., Traore, A. S., & Fumoux, F. (2004). Hemoglobin C is associated with reduced Plasmodium falciparum parasitemia and low risk of mild malaria attack. *Human Molecular Genetics*, *13*(1), 1‑6. https://doi.org/10.1093/hmg/ddh002

Ringelhann, B., Hathorn, M. K., Jilly, P., Grant, F., & Parniczky, G. (1976). A new look at the protection of hemoglobin AS and AC genotypes against plasmodium falciparum infection : A census tract approach. *American Journal of Human Genetics*, *28*(3), 270‑279.

RTS,S Clinical Trials Partnership. (2015). Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa : Final results of a phase 3, individually randomised, controlled trial. *Lancet (London, England)*, *386*(9988), 31‑45. https://doi.org/10.1016/S0140-6736(15)60721-8

Rusk, N. (2018). The UK Biobank. *Nature Methods*, *15*(12), 1001. https://doi.org/10.1038/s41592-018-0245-2

Ruwende, C., Khoo, S. C., Snow, R. W., Yates, S. N., Kwiatkowski, D., Gupta, S., Warn, P., Allsopp, C. E., Gilbert, S. C., & Peschu, N. (1995). Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature*, *376*(6537), 246‑249. https://doi.org/10.1038/376246a0

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832‑837. https://doi.org/10.1038/nature01140

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., … Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913‑918. https://doi.org/10.1038/nature06250

Safer, D., Brenes, M., Dunipace, S., & Schad, G. (2007). Urocanic acid is a major chemoattractant for the skin-penetrating parasitic nematode Strongyloides stercoralis. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(5), 1627‑1630. https://doi.org/10.1073/pnas.0610193104

Sahu, P. K., Pati, S. S., & Satpathy, R. (2008). Association of msp-1, msp-2 and pfcrt genes with the severe complications of Plasmodium falciparum malaria in children. *Annals of Tropical Medicine and Parasitology*, *102*(5), 377‑382. https://doi.org/10.1179/136485908X300814

Salih, N. A., Hussain, A. A., Almugtaba, I. A., Elzein, A. M., Elhassan, I. M., Khalil, E. A. G., Ishag, H. B., Mohammed, H. S., Kwiatkowski, D., & Ibrahim, M. E. (2010). Loss of balancing selection in the betaS globin locus. *BMC Medical Genetics*, *11*, 21. https://doi.org/10.1186/1471-2350-11-21

Sawian, C. E., Lourembam, S. D., Banerjee, A., & Baruah, S. (2013). Polymorphisms and expression of TLR4 and 9 in malaria in two ethnic groups of Assam, northeast India. *Innate Immunity*, *19*(2), 174‑183. https://doi.org/10.1177/1753425912455675

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature reviews. Genetics*, *19*(8), 491‑504. https://doi.org/10.1038/s41576-018-0016-z

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L., & Ren, B. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports*, *17*(8), 2042‑2059. https://doi.org/10.1016/j.celrep.2016.10.061

Shriner, D., & Rotimi, C. N. (2018). Whole-Genome-Sequence-Based Haplotypes Reveal Single Origin of the Sickle Allele during the Holocene Wet Phase. *American Journal of Human Genetics*, *102*(4), 547‑556. https://doi.org/10.1016/j.ajhg.2018.02.003

Smith, J. D., Chitnis, C. E., Craig, A. G., Roberts, D. J., Hudson-Taylor, D. E., Peterson, D. S., Pinches, R., Newbold, C. I., & Miller, L. H. (1995). Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell*, *82*(1), 101‑110. https://doi.org/10.1016/0092-8674(95)90056-x

Snow, R. W., Omumbo, J. A., Lowe, B., Molyneux, C. S., Obiero, J. O., Palmer, A., Weber, M. W., Pinder, M., Nahlen, B., Obonyo, C., Newbold, C., Gupta, S., & Marsh, K. (1997). Relation between severe malaria morbidity in children and level of Plasmodium falciparum transmission in Africa. *Lancet (London, England)*, *349*(9066), 1650‑1654. https://doi.org/10.1016/S0140-6736(97)02038-2

Sokhna, C. S., Rogier, C., Dieye, A., & Trape, J. F. (2000). Host factors affecting the delay of reappearance of Plasmodium falciparum after radical treatment among a semi-immune population exposed to intense perennial transmission. *The American Journal of Tropical Medicine and Hygiene*, *62*(2), 266‑270.

Spielman, R. S., McGinnis, R. E., & Ewens, W. J. (1993). Transmission test for linkage disequilibrium : The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, *52*(3), 506‑516.

Stirnadel, H. A., Stöckle, M., Felger, I., Smith, T., Tanner, M., & Beck, H. P. (1999). Malaria infection and morbidity in infants in relation to genetic polymorphisms in Tanzania. *Tropical Medicine & International Health: TM & IH*, *4*(3), 187‑193.

Suarez, A. A. R., Renne, N. V., Baumert, T. F., & Lupberger, J. (2018). Viral manipulation of STAT3 : Evade, exploit, and injure. *PLOS Pathogens*, *14*(3), e1006839. https://doi.org/10.1371/journal.ppat.1006839

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., & Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75‑81. https://doi.org/10.1038/nature15394.An

Sul, J. H., Martin, L. S., & Eskin, E. (2018). Population structure in genetic studies : Confounding factors and mixed models. *PLoS Genetics*, *14*(12), e1007309. https://doi.org/10.1371/journal.pgen.1007309

Szpiech, Z. A., & Hernandez, R. D. (2014). selscan : An efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, *31*(10), 2824‑2827. https://doi.org/10.1093/molbev/msu211

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585‑595.

Tan, I., & Leung, T. (2009). Myosin light chain kinases : Division of work in cell migration. *Cell Adhesion & Migration*, *3*(3), 256‑258.

Terry M. Therneau. (2018). *coxme : Mixed Effects Cox Models.* https://CRAN.R-project.org/package=coxme

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data : Extending the Cox Model*. Springer-Verlag. //www.springer.com/us/book/9780387987842

Tian, C., Hromatka, B. S., Kiefer, A. K., Eriksson, N., Noble, S. M., Tung, J. Y., & Hinds, D. A. (2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nature Communications*, *8*(1), 599. https://doi.org/10.1038/s41467-017-00257-5

Timmann, C., Thye, T., Vens, M., Evans, J., May, J., Ehmen, C., Sievertsen, J., Muntau, B., Ruge, G., Loag, W., Ansong, D., Antwi, S., Asafo-Adjei, E., Nguah, S. B., Kwakye, K. O., Akoto, A. O. Y., Sylverken, J., Brendel, M., Schuldt, K., … Horstmann, R. D. (2012). Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*, *489*(7416), 443‑446. https://doi.org/10.1038/nature11334

Torres, R., Szpiech, Z. A., & Hernandez, R. D. (2019). Correction : Human demographic history has amplified the effects of background selection across the genome. *PLoS Genetics*, *15*(1), e1007898. https://doi.org/10.1371/journal.pgen.1007898

Triska, P., Soares, P., Patin, E., Fernandes, V., Cerny, V., & Pereira, L. (2015). Extensive Admixture and Selective Pressure Across the Sahel Belt. *Genome Biology and Evolution*, *7*(12), 3484‑3495. https://doi.org/10.1093/gbe/evv236

van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D. S., Elling, U., Allayee, H., Li, X., Radhakrishnan, A., Tan, S.-T., Voss, K., Weichenberger, C. X., Albers, C. A., Al-Hussani, A., Asselbergs, F. W., Ciullo, M., Danjou, F., … Chambers, J. C. (2012). Seventy-five genetic loci influencing the human red blood cell. *Nature*, *492*(7429), 369‑375. https://doi.org/10.1038/nature11677

Verra, F., Mangano, V. D., & Modiano, D. (2009). Genetics of susceptibility to Plasmodium falciparum : From classical malaria resistance genes towards genome-wide association studies. *Parasite Immunology*, *31*(5), 234‑253. https://doi.org/10.1111/j.1365-3024.2009.01106.x

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, *47*, 97‑120. https://doi.org/10.1146/annurev-genet-111212-133526

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, *4*(3), e72. https://doi.org/10.1371/journal.pbio.0040072

Wang, E. T., Kodama, G., Baldi, P., & Moyzis, R. K. (2006). Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(1), 135‑140. https://doi.org/10.1073/pnas.0509691102

Wang, K., Xu, R., Snider, A. J., Schrandt, J., Li, Y., Bialkowska, A. B., Li, M., Zhou, J., Hannun, Y. A., Obeid, L. M., Yang, V. W., & Mao, C. (2016). Alkaline ceramidase 3 deficiency aggravates colitis and colitis-associated tumorigenesis in mice by hyperactivating the innate immune system. *Cell Death & Disease*, *7*, e2124. https://doi.org/10.1038/cddis.2016.36

Wassmer, S. C., & Grau, G. E. R. (2017). Severe malaria : What's new on the pathogenesis front? *International Journal for Parasitology*, *47*(2‑3), 145‑152. https://doi.org/10.1016/j.ijpara.2016.08.002

Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, *8*(1), 1826. https://doi.org/10.1038/s41467-017-01261-5

Weatherall, J. D. (2004). J. B. S. Haldane and the Malaria Hypothesis. In K. R. Dronamraju (Éd.), *Infectious Disease and Host-Pathogen Evolution* (1re éd., p. 18‑36). Cambridge University Press. https://doi.org/10.1017/CBO9780511546259.003

White, N. J. (2008). Plasmodium knowlesi : The fifth human malaria parasite. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *46*(2), 172‑173. https://doi.org/10.1086/524889

WHO | Severe malaria. (2014). *Tropical Medicine & International Health: TM & IH*, *19 Suppl 1*, 7‑131. https://doi.org/10.1111/tmi.12313_2

WHO | World malaria report. (2019). WHO. http://www.who.int/malaria/publications/world-malaria-report-2019/en/

Willer, C. J., & Li, Y. (2010). METAL : Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)*, *26*(17), 2190‑2191. https://doi.org/10.1093/bioinformatics/btq340

Williams, T. N., Maitland, K., Bennett, S., Ganczakowski, M., Peto, T. E., Newbold, C. I., Bowden, D. K., Weatherall, D. J., & Clegg, J. B. (1996). High incidence of malaria in alpha-thalassaemic children. *Nature*, *383*(6600), 522‑525. https://doi.org/10.1038/383522a0

Williams, Thomas N. (2006). Human red blood cell polymorphisms and malaria. *Current Opinion in Microbiology*, *9*(4), 388‑394. https://doi.org/10.1016/j.mib.2006.06.009

Williams, Thomas N., Mwangi, T. W., Roberts, D. J., Alexander, N. D., Weatherall, D. J., Wambua, S., Kortok, M., Snow, R. W., & Marsh, K. (2005). An immune basis for malaria protection by the sickle cell trait. *PLoS Medicine*, *2*(5), e128. https://doi.org/10.1371/journal.pmed.0020128

Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics*, *31*, 39‑59.

Wu, X., Gao, H., Bleday, R., & Zhu, Z. (2014). Homeobox transcription factor VentX regulates differentiation and maturation of human dendritic cells. *The Journal of Biological Chemistry*, *289*(21), 14633‑14643. https://doi.org/10.1074/jbc.M113.509158

Wu, X., Gao, H., Ke, W., Giese, R. W., & Zhu, Z. (2011). The homeobox transcription factor VentX controls human macrophage terminal differentiation and proinflammatory activation. *The Journal of Clinical Investigation*, *121*(7), 2599‑2613. https://doi.org/10.1172/JCI45556

Zakeri, S., Pirahmadi, S., Mehrizi, A. A., & Djadid, N. D. (2011). Genetic variation of TLR-4, TLR-9 and TIRAP genes in Iranian malaria patients. *Malaria Journal*, *10*, 77. https://doi.org/10.1186/1475-2875-10-77

Zhang, X., Guo, A., Yu, J., Possemato, A., Chen, Y., Zheng, W., Polakiewicz, R. D., Kinzler, K. W., Vogelstein, B., Velculescu, V. E., & Wang, Z. J. (2007). Identification of STAT3 as a substrate of receptor protein tyrosine phosphatase T. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(10), 4060‑4064. https://doi.org/10.1073/pnas.0611665104

Zhang, Y., & Pan, W. (2015). Principal component regression and linear mixed model in association analysis of structured samples : Competitors or complements? *Genetic Epidemiology*, *39*(3), 149‑155. https://doi.org/10.1002/gepi.21879

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W.-Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, *50*(9), 1335‑1341. https://doi.org/10.1038/s41588-018-0184-y

ANNEXES

Annexe 1 : Les études d'association gènes candidats

Définition des cinq régions en Afrique

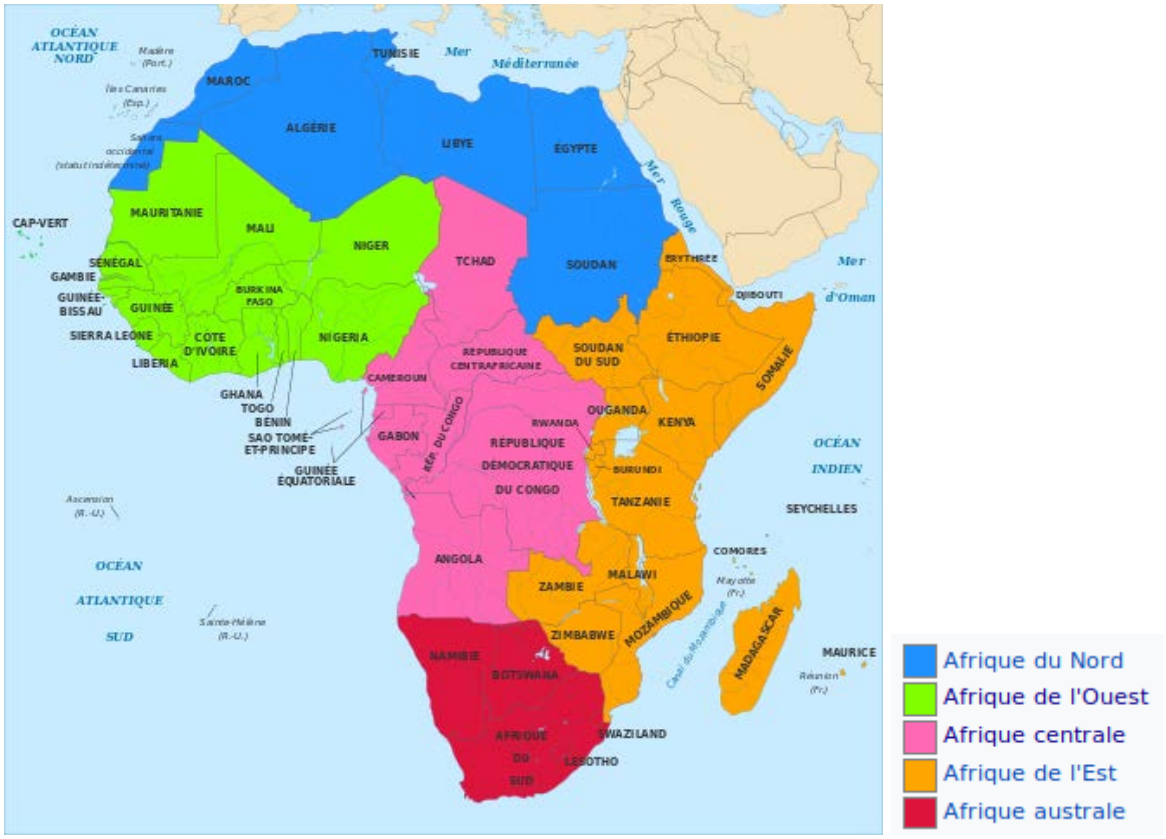


Figure 7 Régions d'Afrique selon l'ONU

Gènes associés avec les formes graves de paludisme

Gene	Nombre d'associations	Pays (groupe ethnique majoritaire quand précisé)	Nombre d'associations par région
HBB* (HbS)	20	Burkina Faso : 1 Mali : 1 (Bambara) Gambie : 4 Sénégal : 1 Ghana : 4 (Akan et Northerner, Dagomba) Nigeria : 2 Cameroun : 1 (Bantu et semi-Bantu) Gabon : 1 Angola : 1 Kenya : 2 (Luo) Tanzanie : 1 Malawi : 1	Afr. de l'Ouest : 13 Afr. Centrale: 3 Afr. de l'Est : 4
ABO*	20	Burkina Faso : 1 Mali: 2 (Dogon, Bambara) Gambie: 3 Ghana: 1 Nigeria: 2 Cameroun: 1 Gabon: 1 Kenya: 3 (Giriama) Malawi : 2 Tanzanie : 1 Zimbabwe : 1 Inde : 1 Vietnam : 1	Afr. de l'Ouest : 9 Afr. Centrale: 2 Afr. de l'Est : 7 Asie du Sud-Est : 1
G6PD*	13	Burkina Faso : 1 Mali : 2 (Dogon et Malinke) Gambie : 2 Ghana :2 Cameroun: 1 (Bantu et semi-Bantu) Nigéria: 1 Kenya: 1 (Giriama) Tanzanie: 1 Malawi: 1 Vietnam: 1	Afr. de l'Ouest : 8 Afr. Centrale: 1 Afr. de l'Est : 3 Asie du Sud-Est : 1
*NOS2**	10	Gambie : 2 Cameroon : 1 (Bantu et semi-Bantu) Gabon : 1 Kenya : 1 Tanzanie : 1 Inde: 3 (origine Austro-Asiatique, origine Indo-Européenne) Thaïlande: 1	Afr. de l'Ouest : 2 Afr. Centrale: 2 Afr. de l'Est : 2 Asie du Sud : 3 Asie du Sud-Est : 1
*TNF**	9	Gambie : 3 Nigéria : 1 Kenya : 1 Inde : 1 Sri Lanka : 1 Thaïlande : 1 Vietnam : 1	Afr. de l'Ouest : 4 Afr. de l'Est : 1 Asie du Sud : 2 Asie du Sud-Est : 2
CD40LG*	6	Mali : 1 Gambie : 2 Kenya : 1	Afr. de l'Ouest : 3 Afr. de l'Est : 2 Asie du Sud : 1

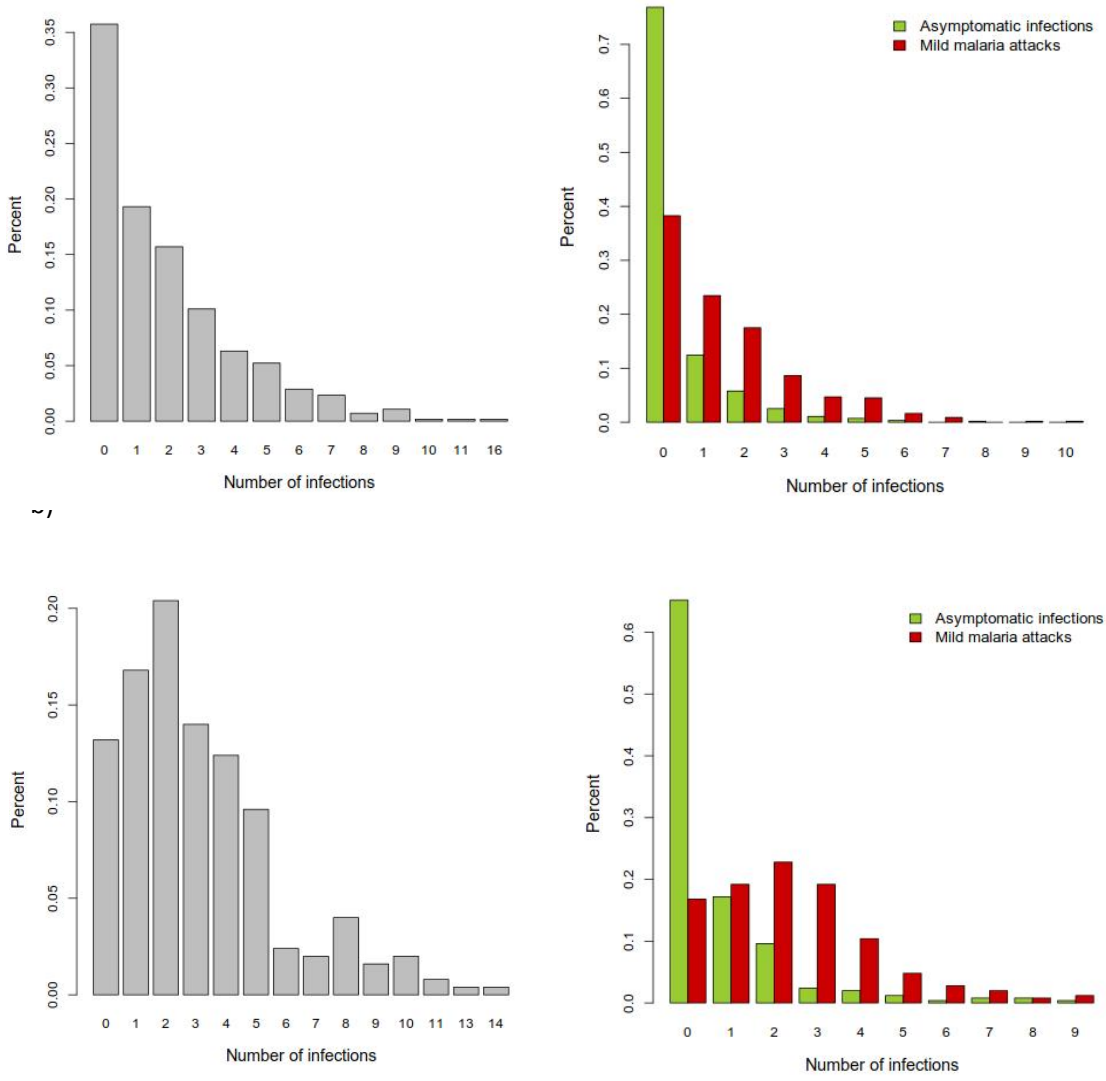
		Tanzanie : 1 Inde : 1	
*CR1**	6	Kenya : 2 (Luo, Girimia) Inde : 3 (origine Indo-Européenne) Thaïlande : 1	Afr. de l'Est : 2 Asie du Sud : 3 Asie du Sud-Est : 1
*CD36**	5	Gambie : 1 Kenya : 2 Inde : 1 Thaïlande : 1	Afr. de l'Ouest : 1 Afr. de l'Est : 2 Asie du Sud : 1 Asie du Sud-Est : 1
FCGR2A	5	Soudan : 1 Ghana : 1 Kenya : 1 Inde : 1 Thaïlande : 1	Afr. du Nord : 1 Afr. de l'Ouest : 1 Afr. de l'Est : 1 Asie du Sud : 1 Asie du Sud-Est : 1
HBA1 HBA2	5	Ghana : 2 (Dagomba, Akan et Northerner) Kenya : 1 (Mijikenda) Tanzanie : 1 Papouasie-Nouvelle-Guinée : 1	Afr. de l'Ouest : 2 Afr. de l'Est : 2 Océanie : 1
*ICAM1**	5	Gabon : 1 Kenya : 1 Tanzanie : 1 India : 1 Vietnam : 1	Afr. Centrale: 1 Afr. de l'Est : 2 Asie du Sud : 1 Asie du Sud-Est : 1
IL12B	5	Mali : 2 Kenya : 1 Thaïlande : 2	Afr. de l'Ouest : 2 Afr. de l'Est : 1 Asie du Sud-Est : 2
*TLR9**	5	Kenya : 1 Burundi : 1 Inde : 3 (Mundari, Bodo-Kachari et Nepalis)	Afr. de l'Est : 2 Asie du Sud : 3
ATP2B4*	4	Gambie : 1 Ghana : 1 Kenya : 1 Malawi : 1	Afr. de l'Ouest : 2 Afr. de l'Est : 2
HBB (HbC)	4	Mali : 1 (Dogon) Sénégal : 1 Ghana : 2 (Dagomba, Akan et Northerner)	Afr. de l'Ouest : 4
IFNG	4	Mali : 1 Gambie : 1 Inde : 2 (origine Austro-Asiatique, origine Indo-Européenne)	Afr. de l'Ouest : 2 Asie du Sud : 2
*IL13**	4	Kenya : 1 Tanzanie : 1 Thaïlande : 1 Vietnam : 1	Afr. de l'Est : 2 Asie du Sud-Est : 2
PECAM1	4	Inde : 2 (origine Austro-Asiatique, origine Indo-Européenne) Thaïlande : 2	Asie du Sud : 2 Asie du Sud-Est : 2
ADCY9	3	Cameroun : 2 (Bantu, semi-Bantu) Tanzanie : 1	Afr. Centrale : 2 Afr. de l'Est : 1
ADORA2A	3	Gambie : 1 Malawi : 1 Inde : 1	Afr. de l'Ouest : 1 Afr. de l'Est : 1 Asie du Sud : 1
HLA-DRB1	3	Gambie : 1 Ghana : 1 Thaïlande : 1	Afr. de l'Ouest: 2 Asie du Sud-Est : 1

IFNAR1	3	Gambie : 1 Angola : 1 Inde : 1 (origine Indo-Européenne)	Afr. de l'Ouest : 1 Afr. Centrale : 1 Asie du Sud : 1
*IL10**	3	Gambie : 1 Cameroun : 1 Kenya : 1	Afr. de l'Ouest : 1 Afr. Centrale : 1 Afr. de l'Est : 1
*IL1A**	3	Gambie : 1 Tanzanie : 1 Vietnam : 1	Afr. de l'Ouest : 1 Afr. de l'Est : 1 Asie du Sud-Est : 1
*IL4**	3	Mali : 1 Ghana : 1 Inde : 1	Afr. de l'Ouest: 2 Asie du Sud : 1
MBL2	3	Ghana : 1 (Dagomba) Gabon : 1 Inde : 1	Afr. de l'Ouest : 1 Afr. Centrale : 1 Asie du Sud : 1
SLC4A1	3	Ghana : 1 Papouasie-Nouvelle-Guinée : 2	Afr. de l'Ouest : 1 Océanie : 2
APOBEC3B	2	Inde : 2 (origine Austro-Asiatique, origine Indo-Européenne)	Asie du Sud : 2
*GNAS**	2	Gambie : 1 Ghana : 1	Afr. de l'Ouest : 2
HLA-B	2	Gambie : 1 Thaïlande : 1	Afr. de l'Ouest: 1 Asie du Sud-Est : 1
HMOX1	2	Gambie : 1 Birmanie : 1 (Myanmarese)	Afr. de l'Ouest: 1 Asie du Sud-Est : 1
HP	2	Ghana : 2 (Dagomba)	Afr. de l'Ouest: 2
IFNGR1	2	Gambie : 1 Inde : 1 (origine Indo-Européenne)	Afr. de l'Ouest: 1 Asie du Sud : 1
KIR2DL3	2	Inde : 1 Thaïlande : 1	Asie du Sud : 1 Asie du Sud-Est : 1
MIF	2	Kenya : 1 (Luo) Inde : 1	Afr. de l'Est : 1 Asie du Sud : 1
RNASE3	2	Senegal : 1 Ghana : 1	Afr. de l'Ouest: 2
RTN3	2	Cameron : 1 (Bantu et semi-Bantu) Tanzanie : 1	Afr. Centrale : 1 Afr. de l'Est : 1
SELE	2	Inde : 2 (origine Austro-Asiatique, origine Indo-Européenne)	Asie du Sud : 2
*TLR1**	2	Inde : 1 Papouasie-Nouvelle-Guinée : 1	Aie du Sud : 1 Océanie : 2
TLR2	2	Ouganda : 1 Inde : 1	Afr. de l'Est : 1 Asie du Sud : 1
*TLR4**	2	Ghana : 1 Nigeria : 1	Afr. de l'Ouest: 2

Table 3 Gènes trouvés associés au moins deux fois avec les formes graves de paludisme dans les études gènes candidats
** Gènes ayant été inclus dans l'étude multicentrique portant sur 11 890 cas de paludisme grave et 17 441 contrôles dans 11 pays d'Afrique, d'Asie et d'Océanie (Malaria Genomic Epidemiology Network,2014) ; en caractères gras, les gènes ayant montré une évidence forte d'association dans la méta-analyse de cette étude.*

Annexe 2 : Données sur les infections palustres dans les deux cohortes

a)



b)

Figure 18 Distribution du nombre d'infections palustres dans a) la cohorte de Tori-Bossito et b) la cohorte d'Allada. A gauche en gris est représenté le nombre total d'infections au cours du suivi par enfant (infections asymptomatiques et accès palustres) ; à droite le nombre d'infections par enfant par type d'infections.

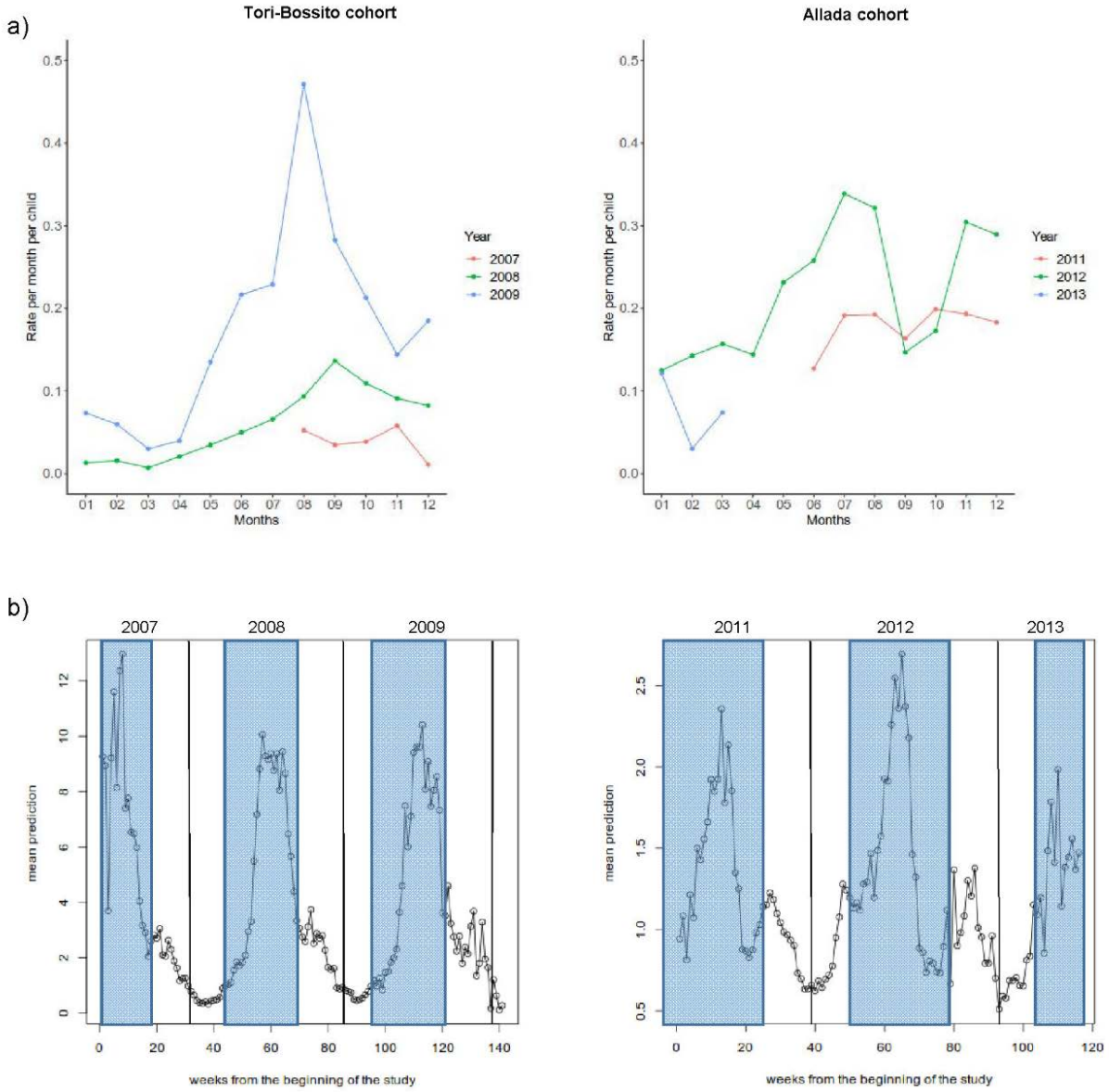


Figure 19 Taux d'incidence des accès palustres et moyenne de prédiction environnementale pour les deux cohortes. a) Le taux d'incidence (en nombre d'accès palustres par mois et par enfant) est représenté pour les mois où plus de 20 enfants sont inclus dans le suivi. Des couleurs différentes ont été attribuées pour chaque année dans les deux suivis. Le pic de transmission est observé entre juin et septembre, en léger décalage avec la grande saison des pluies (avril à juillet). L'incidence est faible à la fin de la saison des pluies mais non nulle. b) Moyenne de la prédiction environnementale par semaine de suivi. Les rectangles bleus indiquent la période avec les deux saisons des pluies (une grande saison des pluies puis une petite séparées par environ un mois sans précipitation).

Annexe 3 : Articles

Article 1

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

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

[1]MERIT, IRD, Université Paris 5, Sorbonne Paris Cité, Paris, 75006, France

[2] Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, F-91057, Evry, France

[3]Centro de Investigación y Desarrollo en Inmunología y Enfermedades Infecciosas, Consejo Nacional de Investigaciones Científicas y Técnicas, Cordoba, Argentina

[4]Facultad de Filosofía y Humanidades, Universidad Nacional de Córdoba, Argentina

[5]Faculty of Medicine of Ribeirão Preto, University of São Paulo, Brazil

[6]Centre d'Etude et de Recherche sur le Paludisme Associé à la Grossesse et l'Enfance, Faculté des Sciences de la Santé, Cotonou, Bénin

[7]Université Paris-Saclay, Centre de recherche en Epidémiologie et Santé des Populations, Institut National de la Santé et de la Recherche Médicale, Villejuif, France

* Joint last authors. The authors have no competing financial interests.

Abstract

Recent research efforts to identify genes involved in malaria susceptibility using genome-wide approaches have focused on severe malaria. Here we present the first GWAS on non-severe malaria designed to identify genetic variants involved in innate immunity or innate resistance mechanisms. Our study was performed on two cohorts of infants from southern Benin (525 and 250 individuals used as discovery and replication cohorts respectively) closely followed from birth to 18-24 months of age, with an assessment of a space- and time-dependent environmental risk of exposure. Both the recurrence of mild malaria attacks and the recurrence of malaria infections as a whole (symptomatic and asymptomatic) were considered. Post-GWAS functional analyses were performed using positional, eQTL and chromatin interaction mapping to identify the genes underlying association signals. Our study highlights a role of *PTPRT*, a tyrosine phosphatase receptor involved in STAT3 pathway, in the protection against both mild malaria attacks and malaria infections (p=9.70x10^{-8} and p=1.78x10^{-7} respectively in the discovery cohort). Strong statistical support was also found for a role of *MYLK4* (meta-analysis, p=5.29 x 10^{-8} with malaria attacks), and for several other genes whose biological functions are relevant in malaria infection. Results shows that GWAS on non-severe malaria can successfully identify new candidate genes and inform physiological mechanisms underlying natural protection against malaria.

Introduction

In spite of numerous prevention and control efforts in recent years, malaria remains a major global public health problem with 219 million cases and ~ 435,000 deaths in 2017 (World Health Organization 2018). In Africa, Bhatt et *al.* (Bhatt et al. 2015) have estimated that *Plasmodium falciparum* infection prevalence halved between 2000 and 2015, and that the incidence of clinical disease fell by 40%, owing mainly to the large distribution of insecticide-treated nets, the most widespread intervention. However the fight against malaria is currently facing numerous challenges. A decrease in the global reduction of malaria cases and deaths is observed at present, with no significant progress made over the 2015–2017 period. The WHO African region which represents the region with the highest malaria burden (92% of malaria cases and 93% of malaria-related deaths in 2017) is also the area where medical advances have been hardest to achieve recently. Moreover control efforts are threatened by insecticide resistance and the possible spread of resistance to artemisinin (an essential component of the most efficient anti-malaria drugs, the artemisinin-based combination therapy -ACT) from Southeast Asia to Africa. In absence of alternative treatments with the same efficacy and tolerability as ACT and of an efficient vaccine, fundamental malaria research continues to be essential in order to better understand the physiopathology of the disease.

P. falciparum is the most prevalent malaria parasite in sub-Saharan Africa (99.7% of malaria cases in African region). Different clinical presentations of *P. falciparum* malaria exist, from asymptomatic (parasite carriage without any clinical sign), or mild forms (parasitemia with fever), to severe forms which may ultimately lead to death. This variability of clinical presentation is thought to be attributable to environmental factors, as well as parasite and human host factors, among which genetic variation could play a major role (Verra, Mangano, and Modiano 2009; Kwiatkowski 2005). In 2005, Mackinnon et *al.*(Mackinnon et al. 2005) estimated that 24% and 25% of the variability in the incidence of mild malaria and severe malaria, respectively, could be explained by genetic factors. Numerous genetic epidemiological studies have attempted to identify gene polymorphisms associated with susceptibility or resistance to different malaria phenotypes (Kwiatkowski 2005; Driss et al. 2011; Marquet 2017).

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

associated with a protection against parasite invasion and clinical malaria attacks (Marquet 2017).

Other association studies have focused on chromosomal regions linked with parasite infection levels and mild malaria susceptibility, in particular 5q31-33 and 6p21-23 (Garcia et al. 1998; Rihet et al. 1998; Flori et al. 2003; Sakuntabhai et al. 2008; Jepson et al. 1997; Brisebarre et al. 2014; Timmann et al. 2007; Milet et al. 2010). In this last region, *TNF* has been the most studied candidate; *NCR3,* encoding a cell membrane receptor of natural killer, has been also repeatedly associated with mild malaria (Flori et al. 2003; Baaklini et al. 2017).Since 2010, several genome-wide association studies (GWAS) and meta-analysis have been published on severe malaria (Timmann et al. 2012; Band et al. 2013; Malaria Genomic Epidemiology Network et al. 2015; Ravenhall et al. 2018; Malaria Genomic Epidemiology Network and Malaria Genomic Epidemiology Network 2014). These studies confirmed the involvement of previously known susceptibility genes (*HBB* and *ABO*) and revealed new genes (*ATP2B4*, the cluster of genes *GYP/FREM3,* among which *GYPA* and *GYPB* appear to play a central role (Leffler et al. 2017).

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

In our study, infants were closely followed with symptomatic and asymptomatic malaria infections recorded. Furthermore geographical, climatic, behavioral and entomological data were collected in concomitant environmental risk surveys throughout the follow-up, to estimate a spatial- and time-dependent risk of exposure. Association studies were performed with the recurrence of mild malaria attacks (RMM) and the recurrence of malaria infections (RMI) including mild malaria attacks and asymptomatic infections, taking into account a time-dependent risk of exposure. We find convincing evidence supporting the involvement of *PTPRT, MYLK4, VENTX* and *UROC1* as well as strong association signals in *PTPRM, ACER3* and *CSMD1* in discovery cohort, whose biological

functions likely pertain to susceptibility to malaria infection.

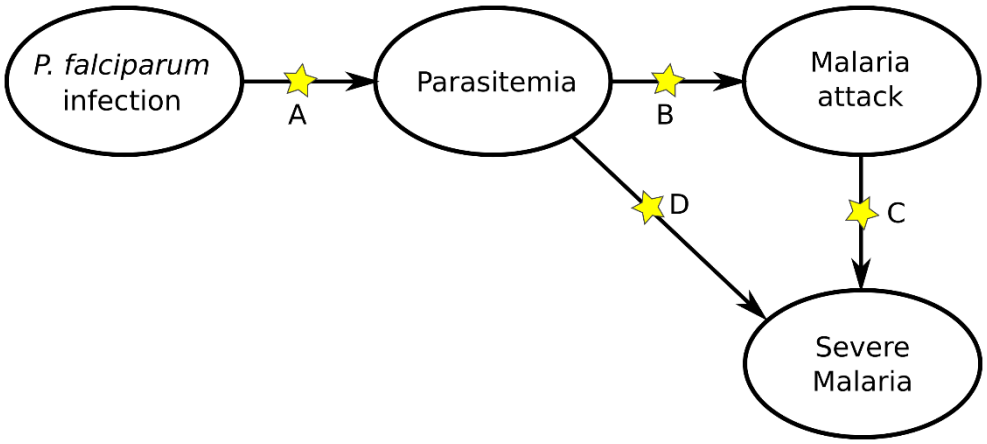


Fig. 1 The path from *P. falciparum* infection to severe malaria, through parasitemia and malaria attacks (parasitemia with fever)
The process may stop at each transition A, B, C and D, depending on several factors including the genetics of the host. Studies on severe malaria mainly target factors involved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

Results

Phenotypic and genotypic data

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

Principal Component Analysis (PCA) performed on the two samples did not reveal any ethnic outliers and showed a low degree of stratification (Online Resource Fig. S1). PC1 separated the two cohorts and PC2 likely reflected some heterogeneity in the Tori-Bossito cohort. We observed that infants did not cluster by reported ethnicity, but rather by cohort and health center, indicating that the self-reported ethnicity of the mother is not a major factor of genetic heterogeneity in our samples. The PCA including 1000 Genomes Project (KGP) African populations (as described in methods) confirmed the absence of outlier in our population (Online Resource Fig. S2). As expected, our two populations overlapped and clustered with Nigerian populations (Yoruba and Esan).

After quality control, 775 infants with more than three months follow-up were available for GWAS (525 in the discovery cohort and 250 in the replication one). Main characteristics of infants are presented in Online Resource Tables S1-S2 and were compared with those of infants who were excluded. For the discovery cohort, no significant differences were observed except for ethnic group composition which appeared to be enriched in Tori at the expense of less prevalent ethnic groups ($p<0.04$). For the replication cohort, only infants followed between 12 and 24 months of age were included in the GWAS. Thus a higher proportion of infants were excluded from analyses (infants lost to follow-up during the first year). We observed in our sample a lower proportion of infants with low birth weight and born to mothers who reported having never attended school ($p<0.01$ and $p=0.06$ respectively).

The protocol of follow-up allowed to detect mild malaria attacks, defined as a positive thick blood smear (TBS) or rapid diagnostic test (TBS) associated with fever (axillary temperature $\geq$ 37.5 C) or a history of fever, and asymptomatic infections (by a TBS at scheduled visits every month). For the 525 infants included in the discovery sample, mean (SD) length of follow-up was 16.9 months (2.83). A total of 342 infants (65.1%) experienced at least one mild malaria attack (from one to 10

episodes) and 359 infants (68.4%) were observed with at least one malaria infection (either a malaria attack or an asymptomatic infection, range from one to 16 malaria infections by infant). For the replication cohort of 250 infants, mean (SD) length of follow-up was 11.9 months (1.72). A proportion of 83.2% and 86.8% children experienced at least one malaria attack (range 1 to 9), or at least one malaria infection (range 1 to 14) during the follow-up, respectively. The distributions of the numbers of episodes by infant are represented in Online Resource Fig. S6.

Adjusting on main epidemiological and environmental determinants

Genome-wide association analyses were performed in two steps because computational burden inherent to the random effect Cox model (Therneau and Grambsch 2000) makes it inappropriate for the large number of tests in GWAS.

In the first step, the recurrence rate of malaria episodes was modeled using a random effect Cox model for recurrent events, taking into account the epidemiological and environmental factors described in Online Resource Tables S1-S2 that might influence malaria infections in infants. A few covariates among all those considered were retained in the final model (Online Resource Tables S3-S4). In each cohort, the same covariates were found associated for the recurrence of mild malaria attacks (RMM) and the recurrence of malaria infections (RMI). The levels of exposure to vector bites, health centers and transmission seasons were associated with the risk of infection in both cohorts. In addition, for the Allada cohort, significant effects were observed for marital status and for maternal education level. No effects of placental infection (detected by a thick and thin placental smears), low birth weight (<2500 g) or HbS allele carriage were detected.

Genome-wide association analyses

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

The analysis revealed four association signals with a *p*-value very close to the genome-wide association threshold of $5x10^{-8}$: two signals for RMM (Fig. 2a), located in *SYT16* (14q23.2 region, lead SNP rs375961263, $p=3.7x10^{-8}$) and in *PTPRM*, a gene encoding a receptor-type protein tyrosine phosphatase (18p11.23 region, lead SNP rs113776891, $p=3.77x10^{-8}$) ; and two signals for RMI (Fig. 2b): one located in *ACER3*, a gene encoding an alkaline ceramidase (11q13.5 region,

lead SNP rs77147099, p=6.85x10^{-8}) and another one located in *PTPRT*, a gene encoding a second receptor-type protein tyrosine phosphatase (20q12 region, lead SNP rs111968843, p=9.70 x10^{-8}).

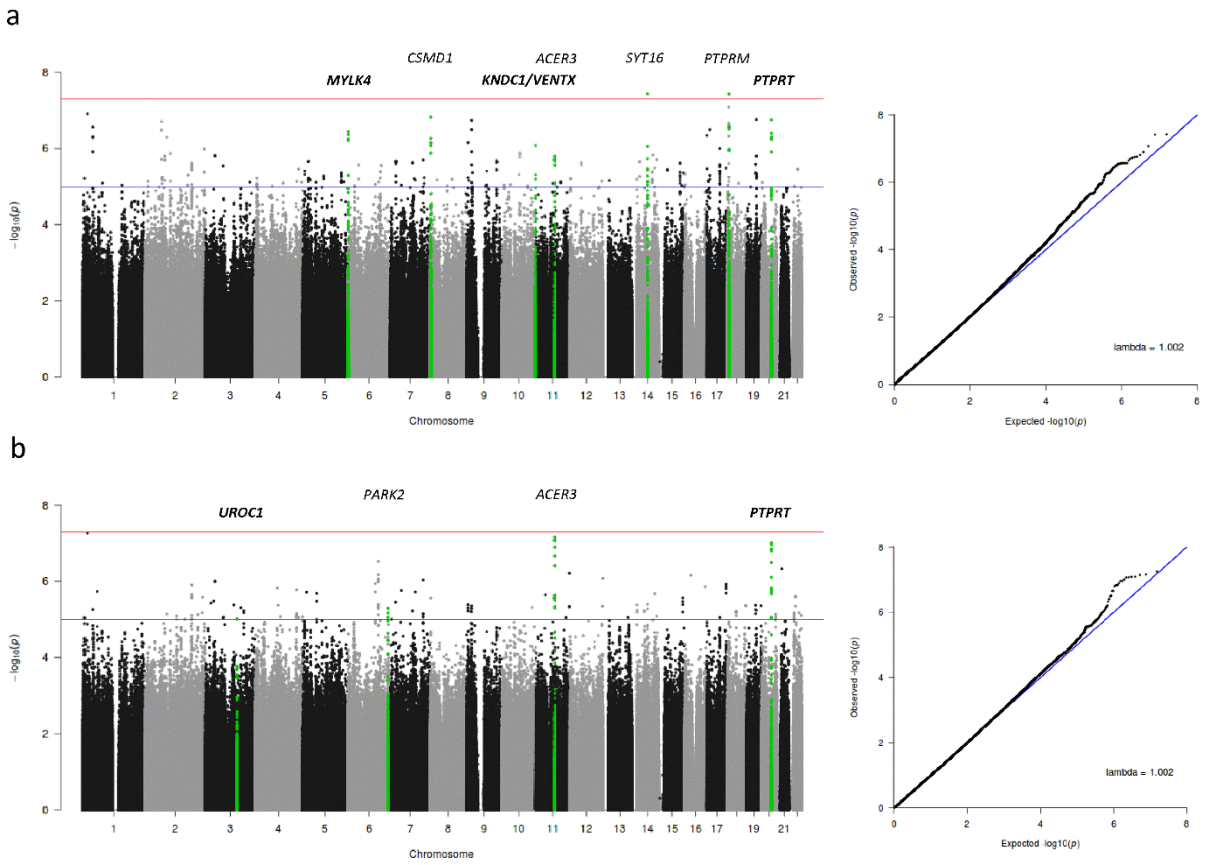


Fig. 2 Manhattan plots and QQ plots for the discovery association study
a) GWAS results for the recurrence of mild malaria attacks, b) GWAS results for the recurrence of malaria infections. The red line in the Manhattan plots indicates the threshold of significance (5 x 10^{-8}), the blue line the threshold of suggestive association used for replication (1 x 10^{-5}). The blue line in the QQ plots shows the 1:1 regression line (the expected distribution of p-value under the null hypothesis). Lambda is the genomic inflation factor calculated as the ratio of the median of the empirically observed distribution of the test statistic to the expected median.

A total of 356 and 214 variants were associated at p<1x10^{-5} with RMM and RMI respectively (Online Resource Tables S5-S6), and were tested for replication in the second cohort. For the highest association signals in discovery cohort aforementioned and for three others that showed evidence of replication (p<0.05), data were re-analyzed in a single step with a mixed Cox model for recurrent events accounting for cryptic relatedness and relevant covariates, without resorting to intermediate confounder-adjusted phenotypes. Results are presented in Tables 1 and 2. Highest statistical support was observed for SNPs located in the 6p25.2 locus with RMM (6 SNPs, min-p= 0.0028). The most strongly associated SNP in the replication cohort, rs72840075, is located in an intron of *MYLK4* which encodes a myosin light chain kinase (Fig. 3). A meta-analysis of these SNPs showed a borderline significant association for SNP rs142480106 (p=5.29 x 10^{-8}) (Table 1).

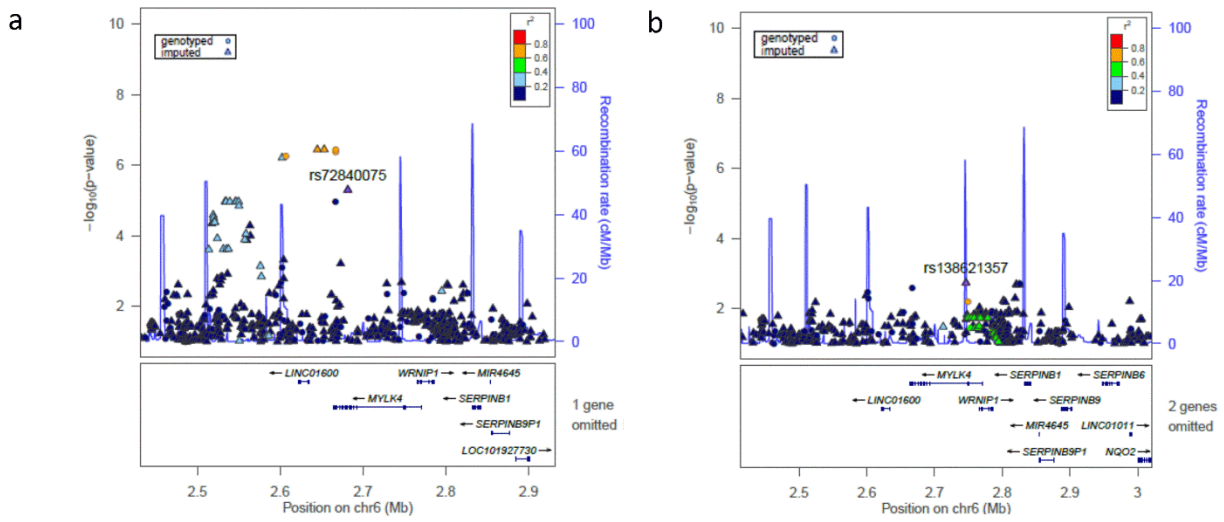


Fig. 3 Annotated regional association plot for the 6p25.2 locus
a) Association results with the recurrence of mild malaria attacks, pairwise LD are shown with rs72840075, the best associated SNP in replication analysis b) Association results in conditional analysis on rs72840075. Plots show LD calculated on Nigerian Population of KGP dataset (ESN and YRI populations) in the 250 kb region.

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

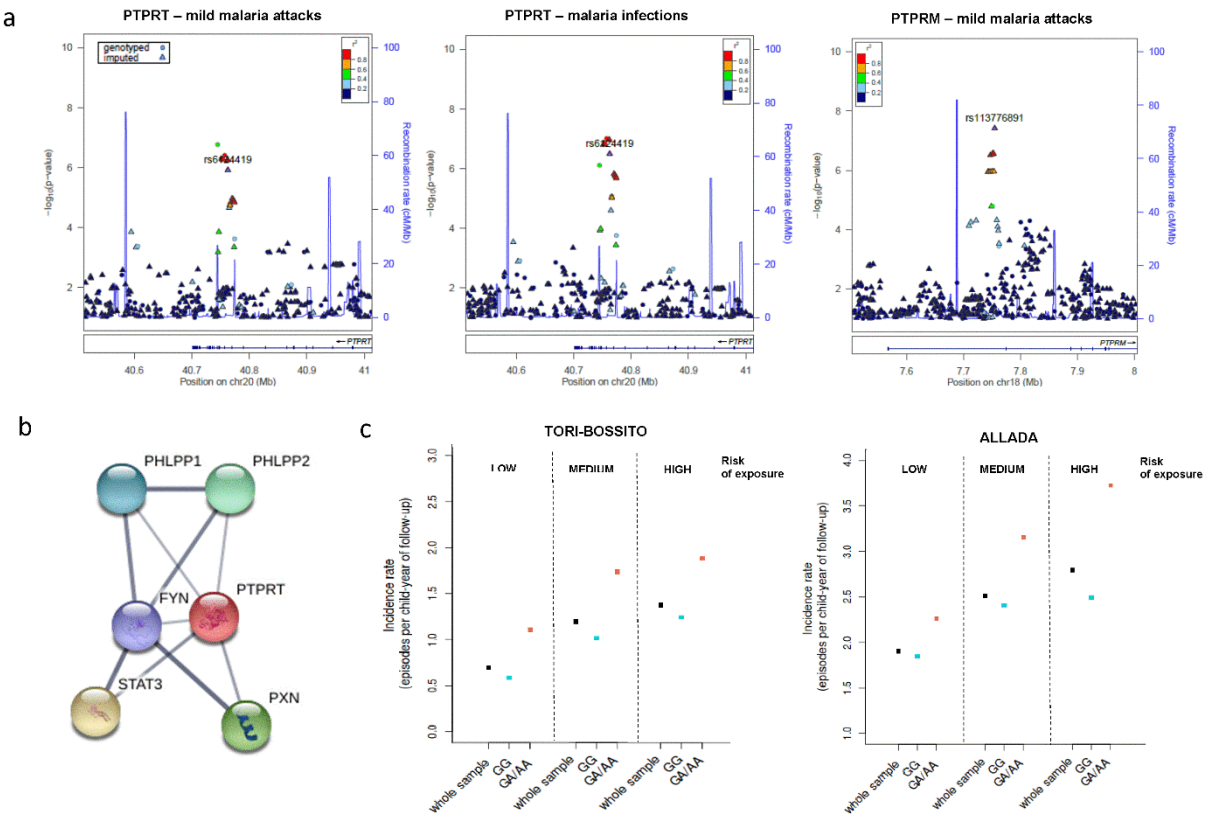


Fig. 4 a) Annotated regional association plots for *PTPRT* and *PTPRM* loci. b) Interaction network for PTPRT c) Crude incidence rate of mild malaria attacks for rs6124419 genotypes calculated for three classes of environmental exposure**.** Plots show LD calculated on Nigerian Population of KGP dataset (ESN and YRI populations). Replicating SNP are highlighted for PTPRT. The three classes were defined from the mean environmental exposure risk across all the follow-up calculated for each infant. Low, medium and high correspond to the three tertiles of the distribution.

Other evidences of replication in the Allada cohort were found in *SYT16* (rs61743638 and rs62639692) and in *KNDC1* (rs182416945) for RMM (table 1), as well as in *ACER3* (a series of 18 SNPs in high LD) and in *UROC1* (rs9871671) for RMI (Table 2). This last one, a missense SNP, is predicted to be probably damaging with a score of 0.99 in PolyPhen-2 (Adzhubei et al. 2010) and has a Combined Annotation Dependent Depletion **(**CADD) (Kircher et al. 2014; Rentzsch et al. 2019) score of 26.5, indicating that this variant is among the 1% most deleteriousness substitutions of the human genome.

For each signal aforementioned, a regional LocusZoom LD plot *(Pruim et al. 2010)* for the nearby region (± 200 kb) was realized and a conditional analysis was run (Fig. 4a and Online Resource Fig. S8 and S9). For none of them, a significant residual signal was found. To illustrate the effects of loci identified in both cohorts, we represented the Mean Cumulative functions of the episodes (Online Resource Fig. S10).

Functional annotation and gene mapping of GWAS results

To identify the genes and variants most likely associated with these four replicating association signals, and to select potential genes of interest pointed out by non-replicating signals, we used the FUMA web platform (Watanabe et al. 2017). FUMA identifies independent associated genomic regions and prioritizes genes based on functional consequences of SNPs in each of them.
Thirty independent associated regions were identified for RMM and 15 for RMI, based on the presence of at least 3 SNPs below the *p*-value threshold of 10^{-5} in a region (LD threshold to aggregate SNPs in one signal, $r^2 = 0.1$). Candidate SNPs (among SNPs present either in our data or in the AFR population of the KGP reference panel) were then defined based on LD ($r^2>0.6$) with the lead SNP in each of these regions, or with one of the replicating SNPs.
These candidate SNPs were mapped to genes, by positional mapping (deleterious coding SNPs with a CADD score > 12.37 or regulatory elements likely to affect binding in non-coding regions - Regulome database (Boyle et al. 2012) (RDB, score ≤ 2), and by eQTL mapping based on expression data from relevant tissue types for malaria (whole blood, cell-EBV transformed lymphocytes, liver, skin, cells transformed fibroblasts and spleen). These mappings identified 20 genes for RMM and 8 for RMI (Online Resource Tables S7-S8). Results for the strongest signals of association with both phenotypes are summarized in Fig. 5. We additionally performed chromatin interaction mapping with high chromosome contact map (Hi-C) data (Schmitt et al. 2016) with FUMA, allowing to identifying genes located in regions with significant chromatin interaction with the candidate SNPs (Fig. 5).

Phenotype	Locus	Location (start - end)	Lead SNP	MAF	P intial 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 x 10-7	**MYLK4**							1	+
	10q26.3	135034267-135052777	rs182416945	0.12	8.31 x 10-7	**VENTX**							2	-
	20q12	40745108-40774357	rs111968843	0.11	1.21 x 10-6	**PTPRT**				rs144104706	14.39			
	18p11.23	7742657-7754654	rs113776891	0.08	3.78 x 10-8	PTPRM				rs138057394 rs112288193	19.58	2b		
	18p11.31	7012462-7034932	rs143310084	0.02	8.32 x 10-8	LAMA1				rs566655	14.64			
	1p34.2	41902797-41932682	rs75470436	0.01	2.76 x 10-7	CTPS1							1	+
	2p12	80290041-80298707	rs6708548	0.08	1.58 x 10-6	CTNNA2				rs1484465 rs1484464	19.08			
	11q13.5	76570145-76654016	rs77147099	0.02	1.60 x 10-6	ACER3							7	-
	5p14.2	24307924-24461650	rs12519312	0.01	2.15 x 10-6	C5orf17							11	-
	5p14.3	22318243-22620207	rs141690513	0.03	2.16 x 10-6	CDH12				rs141690513	14.03			
Malaria infections RMI	20q12	40745108-40774357	rs111968843	0.11	9.70 x 10-8	**PTPRT**				rs144104706	14.39			
	3q21.3	126218211-126218211	rs9871671	0.11	9.79 x 10-6	**UROC1**				rs9871671	26.50			
	11q13.5	76418359-76739604	rs77147099	0.02	6.85 x 10-8	ACER3				rs141181984 rs11608202	13.95	2b	21	-
	6q22.31	122845480-123163778	rs146900853	0.11	3.00 x 10-7	SMPDL3A							1	+
	11q25	133411892-133432324	rs7929384	0.05	6.14 x 10-7	OPCML				rs7126528 rs7126348	17.04			
	2q32.1	188435603-188473204	rs7583251	0.16	1.25 x 10-6	TFPI				rs115976332		2b		
	6q26	161851951-162045282	rs189683911	0.08	5.16 x 10-6	PRKN				rs80324971		2a		
	20q13.32	58143026-58143026	rs144135811	0.15	6.83 x 10-6	MED15				rs114819925 rs5758468		1a	1	-

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

For all the replicating loci but *SYT16* FUMA identified a single gene. Four of these highlighted genes correspond to the gene in which the replicating SNP is located: *MYLK4* in 6p25.2 region, *PTPRT* in 20q12 region, *UROC1* in 3q21.3 region, *ACER3* in 11q35.5 region. The last highlighted gene, *VENTX* in 10q26.3 region, is located 17kb from the lead SNP. In *MYLK4* locus, the replicating SNP, rs72840075, is an eQTL in whole blood (eQTLGen and BIOS eQTl database, FDR<0.05). Allele at risk in our data is associated with a higher expression of this gene. One deleterious SNP is observed in *PTPRT* (r^2=0.78 with the replicating SNP, CADD = 14.39), and in *UROC1* (the aforementioned deleterious SNP). In *VENTX* locus, two SNPs, rs182416945 (the replicating one) and rs138609386, in complete DL (r^2 =1) with the first one, were identified as a significant eQTL in whole blood (eQTLGen database). The allele at risk in our data is associated with a lower expression of *VENTX*, a gene coding for the VENT Homeobox. Furthermore, *ACER3*

in 11q35.5 region is identified by the three mapping methods performed. Numerous significant eQTL associations were observed in the cells transformed fibroblast tissue (GTEx v7 database). For these eQTL, alleles at risk in our data were associated with a lower expression of the gene. This association signal is the strongest signal for RMI, with a *p*-value approaching the genome-wide significance threshold, ($p = 6.85 \times 10^{-8}$, Fig. 2) and it was also found associated with RMM ($p = 1.60 \times 10^{-6}$).

FUMA highlighted also *PTPRM* as the most probable gene associated with the prominent signal for RMM in 18p11.23 region. Two candidate SNPs, rs138057394 and rs112288193 (r^2=0.74 with the lead SNP, for both SNPs) are annotated as deleterious coding SNPs (CADD score of 19.58 and 14.64 respectively).

Candidate associations

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

Discussion

Genetic factors of resistance to malaria may be involved at different stages during the course of the disease: inoculation of parasite, development of asymptomatic parasitemia, appearance of clinical symptoms including fever, rapid progress to complications and manifestation of severe malaria (Kwiatkowski 2005). However research efforts to identify host genetic factors have so far mainly focused on severe malaria, the life threatening form of the disease. While some genetic factors may be in common between severe and mild forms, studies on severe malaria are not designed to discover genes affecting specifically the risk of malaria infection and the development of mild malaria. For this reason, studies based on non-severe malaria forms may help to complete our understanding of the disease.

Our study represents the first screening of genetic variations associated with non-severe forms of malaria at a genome-wide level. It was conducted on two birth cohorts closely followed with a protocol of surveillance which allowed to ascertain malaria infections as a whole. Infants were visited at home weekly or bi-monthly during all the follow-up for the systematic detection of fever and symptomatic malaria. Moreover, a TBS was performed monthly to detect asymptomatic infections (Le Port et al. 2012; d'Almeida et al. 2017). As the protocol of follow-up allowed to treat children early in the development of malaria infection, our sample can include a few cases that may later have come out as severe malaria. However, as all cases are included, our association study focuses on transitions A and B of Fig. 1. Another strength of this study is the assessment of environmental risk of exposure at an individual level all along the follow-up. Although there is consistent evidence for local variation in exposure to *P. falciparum,* which may partly explain the heterogeneity in malaria infection incidence observed in children (Greenwood 1989; Bousema et al. 2010), this factor has been seldom considered in genetic epidemiology studies. Here, entomological, climatic, environmental (information on house characteristics and on its immediate surrounding) as well as geographical data (soil type, watercourse nearby, vegetation index, rainfall, etc) were collected all along the follow-up. Altogether these data allowed modeling, for each child, an individual risk by means of a time-dependent variable (Cottrell et al. 2012). The benefit of using this estimated risk has been demonstrated in two studies on the impact of placental malaria on infant susceptibility (Le Port et al. 2013; Bouaziz et al. 2018). This risk turned out to have a highly significant effect on the recurrence of events in the present study ($p<1x10^{-8}$ and $p<1x10^{-6}$ for discovery and replication cohort respectively).

We find strong support for the involvement of protein tyrosine phosphatase (PTP) receptors. Highly suggestive association peaks were observed for *PTPRM* with RMM and association peaks observed

for *PTPRT* replicated for both phenotypes. Interestingly, both genes are paralogues and in each signal a deleterious coding SNP which could be the causal SNP is identified. Although the replication of *PTPRT* was weak in the second cohort (p=0.01), this can be explained by the fact that our replication cohort is relatively small in size, that infants in this cohort were not actively followed during the first year of life, which limited our analysis to the 12-24 months period. However, assessment of incidence rate of mild malaria attacks in function of groups of different level of exposure, showed very consistent results among exposure groups and cohorts which is an additional argument for a true association. It has to be noted that the higher incidence rate observed in replication cohort compared to discovery cohort (Fig. 4c) didn't appear to be linked to a higher malaria transmission in the Allada district study site. Incidence rate of malaria attacks calculated each month of the follow-up were not higher in the second cohort (Online Resource Fig. S7)4c. This is most probably due to the fact that children followed in the replication cohort are older (12-24 months) than those of the discovery one (0-18 months) and that the incidence rate in the first months of life are much lower than in the 12-24 months period.

PTPRT and PTPRM are proteins belonging to the PTP superfamily, which regulates diverse signaling pathways by catalyzing the removal of a phosphate group from specific signaling proteins. PTPRT and PTPRM encodes two PTP receptors of the same sub-family type IIb (Nikolaienko, Agyekum, and Bouyain 2012). These transmembrane proteins have an extracellular domain involved in cell-cell aggregation and a phosphatase domain which allows intra-cellular signaling. Interestingly PTPRT protein directly interacts with STAT3 as shown by the STRING interaction Network (Fig. 4b) (Szklarczyk et al. 2017) and is the only interactor for which a high confidence (>0.70) is reported in the STRING database. The role of PTPRT in regulating STAT3 pathways has been demonstrated, with STAT3 identified as a direct substrate of PTPRT (Zhang et al. 2007; Peyser et al. 2016). STAT3 is a signal transducer and activator of transcription which behaves similarly to NF-κB, the role of which in malaria infection is well-established. Moreover, STAT3 has been reported as associated with cerebral malaria severity in experimental studies (Liu et al. 2012; 2018). Liu et al. have demonstrated using murine models that *STAT3* is activated by *P. berghei* infection. The STAT3 pathway is thus very likely to be involved in the first stages of malaria infection and represents a potential drug target.

Other genes replicated with $p < 0.05$ in the validation cohort. *MYLK4* belongs to the family of myosin ligth chain kinases (MYLKs) that catalyse myosin interaction with actin filaments. Since members of MYLKs family play an essential role in the organization of the actin/myosin cytoskeleton, and in cell motility (Tan and Leung 2009), *MYLK4* may play a role in RBC membrane structure. At 10q26.3 locus, replication is observed for a SNP located in *KNDC1*, but

two eQTL associations identified *VENTX* as the most probable effector for this association signal. *VENTX* is a homeobox transcriptional factor that controls proliferation and differentiation of hematopoietic and immune cells (Gao et al. 2012; Wu et al. 2011; 2014). Wu et al. works have shown that *VENTX* is a key regulator of macrophage and dendritic cells differentiation. Both macrophages and dendritic cells are innate immune cells derived from monocytes that play an essential role in the response to malaria infections (Chua et al., 2013). Allele at risk of mild malaria attacks in our data was associated with a lower expression of *VENTX*, which would negatively impact the macrophages and/or dendritic cells response to malaria infection in the most susceptible children. Finally, a high deleterious missense SNP in *UROC1* (Urocanate Hydratase 1), a gene expressed in the liver (GTEX RNA-seq), is also associated. Urocanate Hydratase or urocanase is an enzyme that catabolizes urocanic acid to 4-imidazolone-5-propionic acid in liver. Urocanic acid is found in the skin and the sweat of humans, and has been shown to protect the skin from ultra violet radiation. It has also been demonstrated to be a major chemo-attractant for a skin-penetrating parasitic nematode (Safer et al. 2007). Thus, urocanic acid and urocanase could be hypothesized to play a role in the attractiveness of humans to mosquitoes. *ACER3* identified at 11q13.5 locus appears of particular relevance. It is associated both with a protection against asymptomatic and symptomatic malaria infections and multiple significant e-QTL associations were observed in the cells transformed fibroblast tissue. *ACER3* is an alkaline ceramidase, involved in the degradation of ceramides into sphingosine-1-phosphatase (S1P). Ceramides have an anti-plasmodium action (Labaied et al. 2004; Heung, Luberto, and Del Poeta 2006), which is inhibited by S1P; and it has been hypothesized that antimalarial drug artemisinin and mefloquine have an antiparasitic action through the activation of sphingomyelinase producing ceramide (Pankova-Kholmyansky et al. 2003). In our data allele at risk is associated with a decreased *ACER3* expression, thus indicating a lower ceramidase activity in the children with higher risk of malaria infections. *ACER3* has been also involved in the activation of innate immune in mice (K. Wang et al. 2016).

For only one of the replicating signals, *SYT16,* the functional annotation failed to identify a potential effector. The function of *SYT16* being yet unclear, it is hard to estimate the potential relevance of this association signal to malaria.

Finally, a strong association signal in *CSMD1* gene on chromosome 8 was evidenced in mild malaria attacks analysis. No functional evidence was found by FUMA for this association signal; however, an association signal in *CSMD1* was recently found in a GWAS of severe malaria in Tanzania (Ravenhall et al. 2018).

We also examined association results for the most associated candidate variants with malaria-

related phenotypes in the literature (Table 3). There was no evidence of replication for any variant, except rs40401 located in *IL3* ($p = 0.01$ and $p = 0.02$ for malaria attacks and malaria infections, respectively). The fact that our study target population differs substantially from those used in most published studies on mild malaria may partly explain the limited evidence of replication observed. *HBB* was not significantly associated with non-severe malaria in our sample. There are accumulating evidence for an age-related protective effect of HbS (Williams et al. 2005; Gong et al. 2012; Lopera-Mesa et al. 2015) with an acquired mechanism of protection from both asymptomatic and symptomatic *P. falciparum* infections in children. Our results are in line with two of these studies (Williams et al. 2005; Gong et al. 2012) which did not detect a protective effect before the age of two.

Our study has identified several genes whose biological function is relevant to malaria physiopathology and which could play a role in the control of malaria infection. Naturally, future studies are required to further validate these findings. However our results show that GWAS on non-severe malaria can successfully identify new candidate genes and inform physiological mechanisms underlying natural protection against malaria. Improving our understanding of the disease course is crucial for the development of effective control measures.

Materials and Methods

Study population – discovery cohort

The discovery cohort was composed of 525 infants followed-up from birth until 18 months (Le Port et al. 2012). This study took place from June 2007 to January 2010 in 9 villages of Tori-bossito district located 40 km North-East of Cotonou. Southern Benin is characterised by a subtropical climate, with 2 rainy seasons (a long rainy season from April to July and a short one in August and September). Clinical incidence of malaria mainly due to *P. falciparum* (97%) was estimated to be 1.5 (95%CI 1.2-1.9) malaria episodes per child (0-5 years) per year in this area (Damien et al. 2010) and entomological inoculation rate on average 15.5 infective bites per human per year in studied villages (Cottrell et al. 2012).

The design and the flow-chart of the study have been published elsewhere (Le Port et al. 2012). Briefly, 656 infants were included at birth and followed with a close parasitological and clinical survey until the age of 18 months. During the entire follow-up, infants were weekly visited at home by a nurse of the program and temperature was systematically controlled. In case of axillary temperature higher or equal to 37°5 (or a history of fever in the preceding 24 hours), child was referred to health center for medical screening where both a TBS and a rapid diagnostic test (RDT) were performed. Once a month a systematic TBS was performed to detect asymptomatic parasitemia. At any time in case of suspicious fever or clinical signs, related or not to malaria, mothers were invited to bring their infants to the health center where the same protocol (temperature, TBS and RDT) was applied. Symptomatic malaria infection was treated with artemether-lumefantrine combination therapy as recommended by the National Malaria Control Program. A total of 10589 TBS were performed along the follow-up which were read by two independent technicians.

To assess the risk of exposure to *Anopheles* bites, environmental (information on house characteristics and on its immediate surrounding) and geographical data (satellite images, soil type, watercourse nearby, vegetation index, rainfall, etc) were recorded. Throughout the study, every six weeks, human landing catches were performed in several points of the villages to evaluate spatial and temporal variations of *Anopheles* density. Altogether these data allowed modeling, for each child included in the follow-up, an individual risk of exposure by means of a space- and time-dependent variable(Cottrell et al. 2012).

Study population – replication cohort

The replication cohort was composed of 250 infants who were part of a mother-child study conducted from 2009 to 2013 in the district of Allada, a southern semi-rural area located 55

kilometres north of Cotonou and 20 km from the first study site (d'Almeida et al. 2017). Infants were born to mothers who participated in a multi-country clinical trial for the prevention of malaria in pregnancy (MiPPAD trial, "Malaria in Pregnancy Preventive Alternative Drugs, NCT00811421,(González et al. 2014)). The first 400 newborns from MiPPAD participants (Accrombessi et al. 2015) were enrolled in the present study. In brief, 400 offsprings were included at birth between January 2010 and June 2012 and among them 306 integrated the second year follow-up. During the first year, detection of malaria cases was passive. Children were seen at 6, 9 and 12 months of age for scheduled visit. Between 12 and 24 months the same protocol of follow-up as in the first study was set up. Infants were visited at home every two weeks by a nurse and temperature was taken. In case of axillary temperature greater than or equal to 37.5°C, child was referred to health center where a RDT and a TBS were performed. Each month a systematic TBS was made to detect asymptomatic malaria. During the whole 24 months follow-up period, in case of suspicious fever or any health problems, mothers were invited to visit health centers where the same protocol for malaria screening was applied (temperature, TBS and RDT). Clinical malaria infections were treated as recommended by the National Malaria Control Program. Environmental and geographical data were collected as well, to estimate risk of exposure for each infant. The methodology detailed in Cottrell et al., was used again except that this time *Anopheles* density was measured in the children's room, using a CDC light trap.

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

Ethics Approval

For the two cohort studies, both oral and written communications were provided to parent's children interested in participating. An informed consent, written in French and in Fon, was presented to, and signed by parents who agreed to participate. The protocols of these studies were approved by both the Beninese Ethical Committee of the Faculté des Sciences de la Santé (FSS) and the IRD Consultative Committee on Professional Conduct and Ethics (CCDE).

Genotyping and data quality control

Genotyping was conducted at the Centre National de Recherche en Génomique Humaine (CNRGH,

CEA, Evry, France). Before genotyping, a quality control was systematically performed on each DNA sample. All samples were quantified by fluorescence, in duplicate, using the Quant-It kits (Thermo Fischer Scientific). The lowest values systematically underwent a second measurement before any sample was excluded. The quality of material was estimated using about 10% of the total samples received (selected randomly throughout the collection) by performing: i) a quality check by migration on a 1% agarose gel to ensure the samples were not degraded, ii) a standard PCR amplification reaction on the samples to ensure that the genomic DNA was free of PCR inhibitors, iii) a PCR test to verify the gender of the individual (Wilson and Erlandsson 1998). All samples with concentrations below 20 ng/µL, or a major quality problem (degradation and/or amplification problems) were systematically excluded from the study.

After quality control, DNA samples were aliquoted in 96-well plates (JANUS liquid handling robot, Perkin Elmer) for genotyping; sample tracking was ensured by a systematic barcode scanning for each sample. Two DNA positive controls were systematically inserted in a random fashion into the plates. Genotyping was performed on Illumina HumanOmni5-4v1 chips, on a high throughput Illumina automated platform, in accordance with the standard automated protocol of Illumina ® 'Infinium HD Assay (Illumina ®, San Diego, USA).The PCR amplification of the genomic DNA was performed in a dedicated (« pre-PCR ») laboratory. Fragmentation and hybridization of the DNA on the chips were performed in a dedicated (« post-PCR ») laboratory. Several quality controls were systematically included during the process, such as visual inspection of the DNA pellets after precipitation, visual inspection of the deposited cocktail of reagents for hybridization, systematic verification of the temperature of the heating block during the extension and imaging steps. Reading of the chips was performed on iScan+ scanners (Illumina®, San Diego, USA).Primary analysis of the genotyping results was done using the GenomeStudio software (Illumina®, San Diego, USA). The analysis of the internal controls provided by Illumina and the randomly distributed positive controls allowed the validation of the technological process.

Standard control steps and criteria (Anderson et al. 2010) in GWAS were then applied to data. For DNA samples, individuals with discordant genotypic and reported sex were removed (n=5); heterozygosity versus sample call rate were examined, leading to removal of one clear outlier; all remaining samples had a call rate >0.97 (mean = 0.998) and were kept for analysis. A genetic relationship matrix (GRM) (Yang et al. 2011) was calculated on common variants (MAF > 0.05) and thinned data (r^2<0.2) to identify duplicates and examine relationships between individuals. A relatively high relatedness was observed in both cohorts which was expected as most of the recruitment was made in rural villages. One pair of individuals was removed because of an unexpectedly high kinship coefficient ($\Phi = 0.27$, corresponding to siblings). All others were kept

for association analysis (maximum kinship coefficient $\Phi = 0.16$). Markers with call rate <0.98, monomorphic SNPs, Hardy-Weinberg equilibrium (HWE) test $P<10^{-8}$ (calculated on unrelated individuals) were filtered out, as well as all non-autosomal SNPs. Furthermore, a set of 21 pairs of samples (duplicated DNA samples) were used to identify and remove SNPs with poor reproducibility (4986 SNPs showing at least one discordance). After these quality control steps, 2,609,111 markers were available for imputation, and 2,363,703 genotyped markers with MAF>0.01 in the discovery cohort for GWAS analysis (see flow chart for details, Online Resource Fig. S3-S4).

Population stratification

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

In order to compare our two Beninese populations with other African populations, a second PCA was performed including African populations of the KGP. Genetic data were merged on common positions, filtered and thinned as for the first PCA. Only 100 unrelated individuals from each of the two study cohorts were selected, to give them the same weight as KGP samples in PCA. The remaining individuals were projected on the factorial plan thus obtained.

Imputation

Imputation of the two cohorts together was performed on the Michigan Imputation Server (Das et al. 2016) (Das et al., 2017). The reference allele was homogenized to the reference genome by converting the genotypes to the positive strand when necessary; duplicated SNPs, indels, and SNPs with inconsistent reference allele with the reference genome were removed. The resulting genotype data containing 2,539,302 SNPs were uploaded to the Michigan Imputation Server(Das et al., 2017). A last QC step was performed on the server by comparing allele frequencies observed in the uploaded data and in the African reference panel from 1000 Genomes v5 (Sudmant et al. 2015) (Sudmant et al., 2015), and flipping strand for variants showing significant difference in allele frequency (χ2 statistics greater than 300). After this step, there was a high matched between genotype data and the reference panel (Online Resource Fig. S5).
Genotype data were finally imputed using the minimac3 algorithm provided by the Michigan Imputation Server. We selected SHAPEIT v2 for prephasing (Delaneau, Marchini, and Zagury 2011; Delaneau, Zagury, and Marchini 2013) and all haplotypes from 1000 Genomes v5 (Sudmant et al. 2015) (Sudmant et al., 2015)as the reference panel. We filtered out imputed SNPs with a

squared correlation (R^2) between input genotypes and expected continuous dosages below 0.8, or with a MAF below 0.01, which yielded an imputed genotype data set of 15,566,900 variants.

Malaria phenotypes definition

A mild malaria attack was defined as a positive RDT or TBS along with fever (axillary temperature ≥ 37.5°C) or a history of fever in the preceding 24 hours. Each mild malaria attack was recorded at the date of diagnosis at health center and was treated as recommended by the National Malaria Control Program. Children were then considered not to be at risk of malaria within the 14 days after receiving the anti-malarial treatment.

A malaria asymptomatic infection was defined as a positive TBS at monthly scheduled visit, without any fever or history of fever in the preceding 24 hours and without a diagnosis of malaria attacks within three days.

In the following paragraphs we describe the methodology used to perform association study to identify genetic factors affecting the individual rate of malaria attacks, then of malaria infections as a whole (ie both clinical and asymptomatic infections), while adjusting on epidemiological and environmental factors.

Adjusting on relevant epidemiological and environmental factors

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

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

where X_t is a design matric for (possibly time dependent) covariates with a fixed effect, Z is a matrix of indicator variables designed for including a vector of random individual effects b with variance proportional to the GRM 2Φ, allowing to account for cryptic relatedness and population structure. The matrix X_t includes covariates (sex, birthweights, risk of exposure at time t, etc). This model has the advantage of both taking into account incomplete follow-up and all the malaria infections (not only the first one), while incorporating time-dependent susceptibility variables.

To fit this model, the follow-up was divided in month intervals following scheduled visit. For each event we considered the risk of exposure estimated at the previous scheduled visit. A stepwise backward strategy was used to select relevant covariates. Factors considered were individual factors (sex, birth weight), health center, factors related to malaria during pregnancy (placental malaria,

infection during pregnancy - replication cohort only-, intake of intermittent preventive treatment – discovery cohort only- , arm of clinical trials - replication cohort only-), parity (primigravida vs multigravida), maternal anemia at delivery, use of bednet during the follow-up (categorical variable based on mother declaration during the follow-up), risk of exposure and transmission season (a time-dependent variable in four categories, both follow-up spanning over 3 years: dry season, rainy season 2007, rainy season 2009 and rainy season 2010 for example for the discovery cohort) and socio-economic factors (mother educational level, marital status). The final model included only covariates significant at $p \leq 0.05$).

Genome-wide association analyses

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

We thus used the Best Linear Unbiased Predictor (BLUP) $\hat{b}$ of b obtained from the fitted model (1), including all covariates selected by the stepwise backward procedure. This value $\hat{b}$ is an "individual frailty", corresponding to different values of individual log Hazard Ratios of malaria episodes, adjusted on epidemiological covariates. They 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 genotypes at the SNP to be tested, coded as 0, 1 or 2 according the number of alternate alleles, which corresponds to an additive model on the log hazard-ratio scale.

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

SNPs that showed an association at $p \leq 10^{-5}$ in the discovery cohort were selected for replication where the same two steps approach was applied. Gene-based annotations given by ANNOVAR (Kai Wang, Li, and Hakonarson 2010) and CADD score (Kircher et al. 2014) were added for these variants in Online Resource Tables S5-S6. All highly suggestive signals (around the significant threshold of $p \leq 5x10^{-8}$ in the discovery cohort, or which replicated at $p \leq 0.05$) were re-analysed with the mixed Cox model of equation (1) above. For these signals we additionally

performed a meta-analysis combining the results obtained by mixed Cox models, using the method implemented in METAL software (Willer and Li 2010).

All statistical analyses were done using R software (R Core Team 2017), with the package gaston (Hervé Perdry & Claire Dandine-Roulland 2018) for data quality control and genome wide association tests for both genotyped and imputed SNPs, and with the package coxme *(Terry M. Therneau 2018)* for the random effects Cox model. Manhattan plots were obtained with the package qqman (Stephen Turner 2014), and regional linkage disequilibrium plot for association signals with Locuszoom Standalone v1.4 (Pruim et al. 2010).

Gene mapping and biological prioritization

Loci that showed evidence of association at $p \leq 10^{-5}$ with one of the two phenotypes in the discovery cohort, were mapped to gene using the FUMA web plateform (FUMAGWAS v1.3.3; http://fuma.ctglab.nl/(Watanabe et al. 2017), designed for post-processing of GWAS results and prioritizing of genes. FUMA annotates candidate SNPs in genomic risk loci and subsequently maps them to prioritized genes based on (i) physical position mapping on the genome, (ii) expression quantitative trait loci (eQTL) mapping and (iii) 3D chromatin interactions (chromatin interaction mapping). It incorporates the most recent bioinformatics databases, such as Combined Annotation Dependent Depletion score (CADD score) (Kircher et al. 2014; Rentzsch et al. 2019), regulatory elements in the intergenic region of the genome (RegulomeDB) (Boyle et al. 2012), Genotype-Tissue Expression (GTEx) and 3D chromatin interactions from HI-C experiments (Schmitt et al. 2016).

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

Positional mapping was performed using maximum distance from SNPs to gene of 10 kb. Criteria to define deleterious SNPs was a CADD score >12.37 which indicates potential pathogenicity or a RDB score ≤ 2 which indicates that the SNP likely lies in a functional location (categories 1 and 2 of Regulome classification identified as “likely to affect binding”). EQTL mapping was performed using eQTL data from tissue types relevant for non-severe malaria. Search for eQTL associations was carried out in Blood eQTL (Westra et al. 2013), BIOS QTL (Zhernakova et al. 2017), eQTLGen, three databases of e-QTL associations identified in blood, and in the following tissue types of GTEx v7 database: whole blood, cells-transformed fibroblast, cells-EBV-transformed

lymphocytes, liver, skin and spleen. Chromatin interaction mapping was performed from Hi-C experiments data of five tissues/cell types identified as relevant for non-severe malaria: liver, spleen GM12878, IMR90 and trophoblast-like-cell. Hi-C experiments identified chromatin interaction between small chromosomal regions two by two. In this analysis a candidate SNP is mapped to a gene if a significant interaction is observed between a first region containing the candidate SNP and a second one where the gene is located. FDR was used to correct for multiple testing (FDR <0.05 for eQTL association and FDR <1 x 10^{-6} for chromatin interaction, the default values in FUMA).

Candidate associations

We report associations observed in the discovery cohort for genes (± 200 kb 3' and 5') showing the greatest evidence of association with malaria in the literature. Genes previously associated with mild malaria phenotype were selected from a recent review (Marquet 2017): *HbS* and *HbC* alleles of *HBB*, *IL3* from the 5q31-q33 region, *TNF*, *NCR3* from 6p21-p23 region, and *IL10*. Association was not assessed for the *HP* gene and apha-thalasemia condition, because the structural variants involved are not directly accessible through high density genotyping array (Higgs 2013). We also checked significance for genes previously associated with severe malaria that were confirmed by recent GWA and multi-center studies: *ABO, ATP2B4*, and *FREM3/GYP* locus (Timmann et al. 2012; Band et al. 2013; Malaria Genomic Epidemiology Network et al. 2015; Leffler et al. 2017)

References

Accrombessi, Manfred, Smaïla Ouédraogo, Gino Cédric Agbota, Raquel Gonzalez, Achille Massougbodji, Clara Menéndez, and Michel Cot. 2015. 'Malaria in Pregnancy Is a Predictor of Infant Haemoglobin Concentrations during the First Year of Life in Benin, West Africa'. *PloS One* 10 (6): e0129510. https://doi.org/10.1371/journal.pone.0129510.

Adzhubei, Ivan A., Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. 2010. 'A Method and Server for Predicting Damaging Missense Mutations'. *Nature Methods* 7 (4): 248–49. https://doi.org/10.1038/nmeth0410-248.

Almeida, Tania C. d', Ibrahim Sadissou, Jacqueline Milet, Gilles Cottrell, Amandine Mondière, Euripide Avokpaho, Laure Gineau, et al. 2017. '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* 12 (2): e0171117. https://doi.org/10.1371/journal.pone.0171117.

Anderson, Carl A., Fredrik H. Pettersson, Geraldine M. Clarke, Lon R. Cardon, Andrew P. Morris, and Krina T. Zondervan. 2010. 'Data Quality Control in Genetic Case-Control Association Studies'. *Nature Protocols* 5 (9): 1564–73. https://doi.org/10.1038/nprot.2010.116.

Baaklini, Sabrina, Sarwat Afridi, Thy Ngoc Nguyen, Felix Koukouikila-Koussounda, Mathieu Ndounga, Jean Imbert, Magali Torres, Lydie Pradel, Francine Ntoumi, and Pascal Rihet. 2017. '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* 12 (11): e0187818. https://doi.org/10.1371/journal.pone.0187818.

Band, Gavin, Quang Si Le, Luke Jostins, Matti Pirinen, Katja Kivinen, Muminatou Jallow, Fatoumatta Sisay-Joof, et al. 2013. 'Imputation-Based Meta-Analysis of Severe Malaria in Three African Populations'. *PLoS Genetics* 9 (5): e1003509. https://doi.org/10.1371/journal.pgen.1003509.

Bhatt, S., D. J. Weiss, E. Cameron, D. Bisanzio, B. Mappin, U. Dalrymple, K. Battle, et al. 2015. 'The Effect of Malaria Control on Plasmodium Falciparum in Africa between 2000 and 2015'. *Nature* 526 (7572): 207–11. https://doi.org/10.1038/nature15535.

Bouaziz, Olivier, David Courtin, Gilles Cottrell, Jacqueline Milet, Gregory Nuel, and André Garcia. 2018. 'Is Placental Malaria a Long-Term Risk Factor for Mild Malaria Attack in Infancy? Revisiting a Paradigm'. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 66 (6): 930–35. https://doi.org/10.1093/cid/cix899.

Bousema, Teun, Chris Drakeley, Samwel Gesase, Ramadhan Hashim, Stephen Magesa, Frank Mosha, Silas Otieno, et al. 2010. 'Identification of Hot Spots of Malaria Transmission for Targeted Malaria Control'. *The Journal of Infectious Diseases* 201 (11): 1764–74. https://doi.org/10.1086/652456.

Boyle, Alan P., Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, et al. 2012. 'Annotation of Functional Variation in Personal Genomes Using RegulomeDB'. *Genome Research* 22 (9): 1790–97. https://doi.org/10.1101/gr.137323.112.

Brisebarre, Audrey, Brice Kumulungui, Serge Sawadogo, Alexandre Atkinson, Séverine Garnier, Francis Fumoux, and Pascal Rihet. 2014. 'A Genome Scan for Plasmodium Falciparum Malaria Identifies Quantitative Trait Loci on Chromosomes 5q31, 6p21.3, 17p12, and 19p13'. *Malaria Journal* 13 (May): 198. https://doi.org/10.1186/1475-2875-13-198.

Cottrell, Gilles, Bienvenue Kouwaye, Charlotte Pierrat, Agnès le Port, Aziz Bouraïma, Noël Fonton, Mahouton Norbert Hounkonnou, Achille Massougbodji, Vincent Corbel, and André Garcia. 2012. 'Modeling the Influence of Local Environmental Factors on Malaria Transmission in Benin and Its Implications for Cohort Study'. *PloS One* 7 (1): e28812. https://doi.org/10.1371/journal.pone.0028812.

Damien, Georgia B., Armel Djènontin, Christophe Rogier, Vincent Corbel, Sahabi B. Bangana,

Fabrice Chandre, Martin Akogbéto, Dorothée Kindé-Gazard, Achille Massougbodji, and Marie-Claire Henry. 2010. 'Malaria Infection and Disease in an Area with Pyrethroid-Resistant Vectors in Southern Benin'. *Malaria Journal* 9 (December): 380. https://doi.org/10.1186/1475-2875-9-380.

Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, et al. 2016. 'Next-Generation Genotype Imputation Service and Methods'. *Nature Genetics* 48 (10): 1284–87. https://doi.org/10.1038/ng.3656.

Delaneau, Olivier, Jonathan Marchini, and Jean-François Zagury. 2011. 'A Linear Complexity Phasing Method for Thousands of Genomes'. *Nature Methods* 9 (2): 179–81. https://doi.org/10.1038/nmeth.1785.

Delaneau, Olivier, Jean-Francois Zagury, and Jonathan Marchini. 2013. 'Improved Whole-Chromosome Phasing for Disease and Population Genetic Studies'. *Nature Methods* 10 (1): 5–6. https://doi.org/10.1038/nmeth.2307.

Driss, Adel, Jacqueline M. Hibbert, Nana O. Wilson, Shareen A. Iqbal, Thomas V. Adamkiewicz, and Jonathan K. Stiles. 2011. 'Genetic Polymorphisms Linked to Susceptibility to Malaria'. *Malaria Journal* 10 (September): 271. https://doi.org/10.1186/1475-2875-10-271.

Flori, L., B. Kumulungui, C. Aucan, C. Esnault, A. S. Traoré, F. Fumoux, and P. Rihet. 2003. 'Linkage and Association between Plasmodium Falciparum Blood Infection Levels and Chromosome 5q31-Q33'. *Genes and Immunity* 4 (4): 265–68. https://doi.org/10.1038/sj.gene.6363960.

Gao, Hong, Xiaoming Wu, Yan Sun, Shuanhu Zhou, Leslie E. Silberstein, and Zhenglun Zhu. 2012. 'Suppression of Homeobox Transcription Factor VentX Promotes Expansion of Human Hematopoietic Stem/Multipotent Progenitor Cells'. *The Journal of Biological Chemistry* 287 (35): 29979–87. https://doi.org/10.1074/jbc.M112.383018.

Garcia, A., S. Marquet, B. Bucheton, D. Hillaire, M. Cot, N. Fievet, A. J. Dessein, and L. Abel. 1998. 'Linkage Analysis of Blood Plasmodium Falciparum Levels: Interest of the 5q31-Q33 Chromosome Region'. *The American Journal of Tropical Medicine and Hygiene* 58 (6): 705–9.

Gong, Lauren, Catherine Maiteki-Sebuguzi, Philip J. Rosenthal, Alan E. Hubbard, Chris J. Drakeley, Grant Dorsey, and Bryan Greenhouse. 2012. 'Evidence for Both Innate and Acquired Mechanisms of Protection from Plasmodium Falciparum in Children with Sickle Cell Trait'. *Blood* 119 (16): 3808–14. https://doi.org/10.1182/blood-2011-08-371062.

González, Raquel, Ghyslain Mombo-Ngoma, Smaïla Ouédraogo, Mwaka A. Kakolwa, Salim Abdulla, Manfred Accrombessi, John J. Aponte, et al. 2014. 'Intermittent Preventive Treatment of Malaria in Pregnancy with Mefloquine in HIV-Negative Women: A Multicentre Randomized Controlled Trial'. *PLoS Medicine* 11 (9): e1001733. https://doi.org/10.1371/journal.pmed.1001733.

Greenwood, B. M. 1989. 'The Microepidemiology of Malaria and Its Importance to Malaria Control'. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 83 Suppl: 25–29.

Hervé Perdry & Claire Dandine-Roulland. 2018. 'Gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models'.

Heung, Lena J., Chiara Luberto, and Maurizio Del Poeta. 2006. 'Role of Sphingolipids in Microbial Pathogenesis'. *Infection and Immunity* 74 (1): 28–39. https://doi.org/10.1128/IAI.74.1.28-39.2006.

Higgs, Douglas R. 2013. 'The Molecular Basis of α-Thalassemia'. *Cold Spring Harbor Perspectives in Medicine* 3 (1).

Jepson, A., F. Sisay-Joof, W. Banya, M. Hassan-King, A. Frodsham, S. Bennett, A. V. Hill, and H. Whittle. 1997. 'Genetic Linkage of Mild Malaria to the Major Histocompatibility Complex in Gambian Children: Study of Affected Sibling Pairs'. *BMJ (Clinical Research Ed.)* 315 (7100): 96–97.

Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay

Shendure. 2014. ‘A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants’. *Nature Genetics* 46 (3): 310–15. https://doi.org/10.1038/ng.2892.
Kwiatkowski, Dominic P. 2005. ‘How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria’. *American Journal of Human Genetics* 77 (2): 171–92. https://doi.org/10.1086/432519.
Labaied, Mehdi, Arie Dagan, Marc Dellinger, Marc Gèze, Stéphane Egée, Serge L. Thomas, Chunbo Wang, Shimon Gatt, and Philippe Grellier. 2004. ‘Anti-Plasmodium Activity of Ceramide Analogs’. *Malaria Journal* 3 (December): 49. https://doi.org/10.1186/1475-2875-3-49.
Le Port, Agnès, Gilles Cottrell, Fabrice Chandre, Michel Cot, Achille Massougbodji, and André Garcia. 2013. ‘Importance of Adequate Local Spatiotemporal Transmission Measures in Malaria Cohort Studies: Application to the Relation between Placental Malaria and First Malaria Infection in Infants’. *American Journal of Epidemiology* 178 (1): 136–43. https://doi.org/10.1093/aje/kws452.
Le Port, Agnès, Gilles Cottrell, Yves Martin-Prevel, Florence Migot-Nabias, Michel Cot, and André Garcia. 2012. ‘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* 2 (2): e000342. https://doi.org/10.1136/bmjopen-2011-000342.
Leffler, Ellen M., Gavin Band, George B. J. Busby, Katja Kivinen, Quang Si Le, Geraldine M. Clarke, Kalifa A. Bojang, et al. 2017. ‘Resistance to Malaria through Structural Variation of Red Blood Cell Invasion Receptors’. *Science (New York, N.Y.)* 356 (6343). https://doi.org/10.1126/science.aam6393.
Liu, Mingli, Audu S. Amodu, Sidney Pitts, John Patrickson, Jacqueline M. Hibbert, Monica Battle, Solomon F. Ofori-Acquah, and Jonathan K. Stiles. 2012. ‘Heme Mediated STAT3 Activation in Severe Malaria’. *PloS One* 7 (3): e34280. https://doi.org/10.1371/journal.pone.0034280.
Liu, Mingli, Wesley Solomon, Juan Carlos Cespedes, Nana O. Wilson, Byron Ford, and Jonathan K. Stiles. 2018. ‘Neuregulin-1 Attenuates Experimental Cerebral Malaria (ECM) Pathogenesis by Regulating ErbB4/AKT/STAT3 Signaling’. *Journal of Neuroinflammation* 15 (1): 104. https://doi.org/10.1186/s12974-018-1147-z.
Lopera-Mesa, Tatiana M., Saibou Doumbia, Drissa Konaté, Jennifer M. Anderson, Mory Doumbouya, Abdoul S. Keita, Seidina A. S. Diakité, et al. 2015. ‘Effect of Red Blood Cell Variants on Childhood Malaria in Mali: A Prospective Cohort Study’. *The Lancet. Haematology* 2 (4): e140-149. https://doi.org/10.1016/S2352-3026(15)00043-5.
Mackinnon, Margaret J., Tabitha W. Mwangi, Robert W. Snow, Kevin Marsh, and Thomas N. Williams. 2005. ‘Heritability of Malaria in Africa’. *PLoS Medicine* 2 (12): e340. https://doi.org/10.1371/journal.pmed.0020340.
Malaria Genomic Epidemiology Network, Gavin Band, Kirk A. Rockett, Chris C. A. Spencer, and Dominic P. Kwiatkowski. 2015. ‘A Novel Locus of Resistance to Severe Malaria in a Region of Ancient Balancing Selection’. *Nature* 526 (7572): 253–57. https://doi.org/10.1038/nature15390.
Malaria Genomic Epidemiology Network, and Malaria Genomic Epidemiology Network. 2014. ‘Reappraisal of Known Malaria Resistance Loci in a Large Multicenter Study’. *Nature Genetics* 46 (11): 1197–1204. https://doi.org/10.1038/ng.3107.
Marquet, Sandrine. 2017. ‘Overview of Human Genetic Susceptibility to Malaria: From Parasitemia Control to Severe Disease’. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, June. https://doi.org/10.1016/j.meegid.2017.06.001.
Milet, Jacqueline, Gregory Nuel, Laurence Watier, David Courtin, Yousri Slaoui, Paul Senghor, Florence Migot-Nabias, Oumar Gaye, and André Garcia. 2010. ‘Genome Wide Linkage Study, Using a 250K SNP Map, of Plasmodium Falciparum Infection and Mild Malaria

Attack in a Senegalese Population'. *PloS One* 5 (7): e11616. https://doi.org/10.1371/journal.pone.0011616.
Nikolaienko, Roman M., Boadi Agyekum, and Samuel Bouyain. 2012. 'Receptor Protein Tyrosine Phosphatases and Cancer'. *Cell Adhesion & Migration* 6 (4): 356–64. https://doi.org/10.4161/cam.21242.
Pankova-Kholmyansky, I., A. Dagan, D. Gold, Z. Zaslavsky, E. Skutelsky, S. Gatt, and E. Flescher. 2003. 'Ceramide Mediates Growth Inhibition of the Plasmodium Falciparum Parasite'. *Cellular and Molecular Life Sciences: CMLS* 60 (3): 577–87.
Peyser, N. D., M. Freilino, L. Wang, Y. Zeng, H. Li, D. E. Johnson, and J. R. Grandis. 2016. 'Frequent Promoter Hypermethylation of PTPRT Increases STAT3 Activation and Sensitivity to STAT3 Inhibition in Head and Neck Cancer'. *Oncogene* 35 (9): 1163–69. https://doi.org/10.1038/onc.2015.171.
Pruim, Randall J., Ryan P. Welch, Serena Sanna, Tanya M. Teslovich, Peter S. Chines, Terry P. Gliedt, Michael Boehnke, Gonçalo R. Abecasis, and Cristen J. Willer. 2010. 'LocusZoom: Regional Visualization of Genome-Wide Association Scan Results'. *Bioinformatics* 26 (18): 2336–37. https://doi.org/10.1093/bioinformatics/btq419.
R Core Team. 2017. 'R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing'. R Foundation for Statistical Computing. https://www.R-project.org.
Ravenhall, Matt, Susana Campino, Nuno Sepúlveda, Alphaxard Manjurano, Behzad Nadjm, George Mtove, Hannah Wangai, et al. 2018. 'Novel Genetic Polymorphisms Associated with Severe Malaria and under Selective Pressure in North-Eastern Tanzania'. *PLoS Genetics* 14 (1): e1007172. https://doi.org/10.1371/journal.pgen.1007172.
Rentzsch, Philipp, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. 2019. 'CADD: Predicting the Deleteriousness of Variants throughout the Human Genome'. *Nucleic Acids Research* 47 (D1): D886–94. https://doi.org/10.1093/nar/gky1016.
Rihet, P., Y. Traoré, L. Abel, C. Aucan, T. Traoré-Leroux, and F. Fumoux. 1998. 'Malaria in Humans: Plasmodium Falciparum Blood Infection Levels Are Linked to Chromosome 5q31-Q33'. *American Journal of Human Genetics* 63 (2): 498–505. https://doi.org/10.1086/301967.
Safer, Daniel, Mario Brenes, Seth Dunipace, and Gerhard Schad. 2007. 'Urocanic Acid Is a Major Chemoattractant for the Skin-Penetrating Parasitic Nematode Strongyloides Stercoralis'. *Proceedings of the National Academy of Sciences of the United States of America* 104 (5): 1627–30. https://doi.org/10.1073/pnas.0610193104.
Sakuntabhai, Anavaj, Rokhaya Ndiaye, Isabelle Casadémont, Chayanon Peerapittayamongkol, Chayanon Peerapittayamonkol, Christophe Rogier, Patricia Tortevoye, et al. 2008. 'Genetic Determination and Linkage Mapping of Plasmodium Falciparum Malaria Related Traits in Senegal'. *PloS One* 3 (4): e2000. https://doi.org/10.1371/journal.pone.0002000.
Schmitt, Anthony D., Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L. Tan, Yun Li, et al. 2016. 'A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome'. *Cell Reports* 17 (8): 2042–59. https://doi.org/10.1016/j.celrep.2016.10.061.
Stephen Turner. 2014. 'Qqman: Q-Q and Manhattan Plots for GWAS Data.' https://CRAN.R-project.org/package=qqman.
Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. 'An Integrated Map of Structural Variation in 2,504 Human Genomes'. *Nature* 526 (7571): 75–81. https://doi.org/10.1038/nature15394.
Szklarczyk, Damian, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, et al. 2017. 'The STRING Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible'. *Nucleic Acids Research* 45 (D1): D362–68. https://doi.org/10.1093/nar/gkw937.
Tan, Ivan, and Thomas Leung. 2009. 'Myosin Light Chain Kinases: Division of Work in Cell

Migration'. *Cell Adhesion & Migration* 3 (3): 256–58.
Terry M. Therneau. 2018. 'Coxme: Mixed Effects Cox Models.' https://CRAN.R-project.org/package=coxme.
Therneau, Terry M., and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. New York: Springer-Verlag. //www.springer.com/us/book/9780387987842.
Timmann, Christian, Jennifer A. Evans, Inke R. König, André Kleensang, Franz Rüschendorf, Julia Lenzen, Jürgen Sievertsen, et al. 2007. 'Genome-Wide Linkage Analysis of Malaria Infection Intensity and Mild Disease'. *PLoS Genetics* 3 (3): e48. https://doi.org/10.1371/journal.pgen.0030048.
Timmann, Christian, Thorsten Thye, Maren Vens, Jennifer Evans, Jürgen May, Christa Ehmen, Jürgen Sievertsen, et al. 2012. 'Genome-Wide Association Study Indicates Two Novel Resistance Loci for Severe Malaria'. *Nature* 489 (7416): 443–46. https://doi.org/10.1038/nature11334.
Verra, F., V. D. Mangano, and D. Modiano. 2009. 'Genetics of Susceptibility to Plasmodium Falciparum: From Classical Malaria Resistance Genes towards Genome-Wide Association Studies'. *Parasite Immunology* 31 (5): 234–53. https://doi.org/10.1111/j.1365-3024.2009.01106.x.
Wang, K., R. Xu, A. J. Snider, J. Schrandt, Y. Li, A. B. Bialkowska, M. Li, et al. 2016. 'Alkaline Ceramidase 3 Deficiency Aggravates Colitis and Colitis-Associated Tumorigenesis in Mice by Hyperactivating the Innate Immune System'. *Cell Death & Disease* 7 (March): e2124. https://doi.org/10.1038/cddis.2016.36.
Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. 'ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data'. *Nucleic Acids Research* 38 (16): e164. https://doi.org/10.1093/nar/gkq603.
Watanabe, Kyoko, Erdogan Taskesen, Arjen van Bochoven, and Danielle Posthuma. 2017. 'Functional Mapping and Annotation of Genetic Associations with FUMA'. *Nature Communications* 8 (1): 1826. https://doi.org/10.1038/s41467-017-01261-5.
Westra, Harm-Jan, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, et al. 2013. 'Systematic Identification of Trans EQTLs as Putative Drivers of Known Disease Associations'. *Nature Genetics* 45 (10): 1238–43. https://doi.org/10.1038/ng.2756.
Willer, Cristen J., and Yun Li. 2010. 'METAL: Fast and Efficient Meta-Analysis of Genomewide Association Scans'. *Bioinformatics (Oxford, England)* 26 (17): 2190–91. https://doi.org/10.1093/bioinformatics/btq340.
Williams, Thomas N., Tabitha W. Mwangi, David J. Roberts, Neal D. Alexander, David J. Weatherall, Sammy Wambua, Moses Kortok, Robert W. Snow, and Kevin Marsh. 2005. 'An Immune Basis for Malaria Protection by the Sickle Cell Trait'. *PLoS Medicine* 2 (5): e128. https://doi.org/10.1371/journal.pmed.0020128.
Wilson, J. F., and R. Erlandsson. 1998. 'Sexing of Human and Other Primate DNA'. *Biological Chemistry* 379 (10): 1287–88.
World Health Organization. 2018. 'WHO | World Malaria Report 2018'.
Wu, Xiaoming, Hong Gao, Ronald Bleday, and Zhenglun Zhu. 2014. 'Homeobox Transcription Factor VentX Regulates Differentiation and Maturation of Human Dendritic Cells'. *The Journal of Biological Chemistry* 289 (21): 14633–43. https://doi.org/10.1074/jbc.M113.509158.
Wu, Xiaoming, Hong Gao, Weixiong Ke, Roger W. Giese, and Zhenglun Zhu. 2011. 'The Homeobox Transcription Factor VentX Controls Human Macrophage Terminal Differentiation and Proinflammatory Activation'. *The Journal of Clinical Investigation* 121 (7): 2599–2613. https://doi.org/10.1172/JCI45556.
Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. 'GCTA: A Tool for Genome-Wide Complex Trait Analysis'. *American Journal of Human Genetics* 88 (1): 76–

82. https://doi.org/10.1016/j.ajhg.2010.11.011.
Zhang, Xiaodong, Ailan Guo, Jianshi Yu, Anthony Possemato, Yueting Chen, Weiping Zheng, Roberto D. Polakiewicz, et al. 2007. 'Identification of STAT3 as a Substrate of Receptor Protein Tyrosine Phosphatase T'. *Proceedings of the National Academy of Sciences of the United States of America* 104 (10): 4060–64. https://doi.org/10.1073/pnas.0611665104.
Zhernakova, Daria V., Patrick Deelen, Martijn Vermaat, Maarten van Iterson, Michiel van Galen, Wibowo Arindrarto, Peter van 't Hof, et al. 2017. 'Identification of Context-Dependent Expression Quantitative Trait Loci in Whole Blood'. *Nature Genetics* 49 (1): 139–45. https://doi.org/10.1038/ng.3737.

Acknowledgements

This research is a collaboration between the CEA/ Jacob/CNRGH and the IRD/UMR216. We wish to thanks the collaborators of the CERPAGE who participated actively in the longitudinal follow-ups. Longitudinal follow-ups were funded by the French National Research Agency (ANR-SEST 2006 040-01 and ANR-PRSP 2010 012-001); the French ministry of Research and Technology (REFS Nu2006-22) and the Institut de Recherche pour le Développement (IRD). We made use of data previously generated in the MiPPAD study (EDCTP-IP.07.31080.002). The genome-wide scan was supported by the Centre National de Recherche en Génomique Humaine (CNRGH). We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources.

Data Availability Statement

The datasets generated and analysed during the current study are available from the DataSuds repository (https://doi.org/10.23708/EXSQTM). The dataset are not publicly available due to patient confidentiality but are available for researchers who meet the criteria for access to confidential data. All other relevant data are available within the manuscript and its Online Resource files.

Tables

Table 1 Loci with the highest association signals in discovery cohort or which replicated at $p < 0.05$ for the recurrence of mild malaria attacks

Locus	SNP	Effect allele	*p* initial GWA[a]	Allele frequency Tori-Bossito	*p* Cox model Tori-Bossito[b]	HR (CI 95%) Tori-Bossito	Allele frequency Allada	*p* Cox model Allada[b]	HR (CI 95%) Allada	Nearest Genes	*p* Meta-analysis
6p25.2	rs76088706	T	5.56×10^{-7}	0.025	2.68×10^{-6}	2.22 (1.59-3.11)	0.016	0.018	1.75 (1.03-2.96)	C6orf195(intergenic)	3.65×10^{-7}
	rs140858180	T	3.64×10^{-7}	0.026	2.53×10^{-6}	2.18 (1.57-3.03)	0.017	0.030	1.65 (0.97-2.78)	C6orf195(intergenic)	6.02×10^{-7}
	rs547331171	T	3.61×10^{-7}	0.026	2.51×10^{-6}	2.18 (1.57-3.03)	0.016	0.028	1.66 (0.98-2.80)	*MYLK4* (intergenic)	5.57×10^{-7}
	rs142480106	T	4.35×10^{-7}	0.019	1.14×10^{-6}	2.52 (1.73-3.66)	0.012	0.0063	2.08 (1.17-3.70)	*MYLK4* (UTR3)	5.29×10^{-8}
	rs144194334	T	3.62×10^{-7}	0.026	2.53×10^{-6}	2.18 (1.57-3.03)	0.016	0.026	1.67 (0.99-2.82)	*MYLK4* (UTR3)	5.21×10^{-7}
	rs72840075	A	5.04×10^{-6}	0.030	1.05×10^{-5}	2.03 (1.48-2.79)	0.022	0.0028	1.88 (1.20-2.95)	*MYLK4* (intronic)	2.01×10^{-7}
10q26.3	rs182416945	A	8.31×10^{-7}	0.12	5.56×10^{-6}	1.55 (1.28-1.87)	0.19	0.01	1.28 (1.04-1.58)	*KNDC1* (intronic)	7.94×10^{-7}
14q23.2	rs61743638	A	3.72×10^{-6}	0.047	1.95×10^{-5}	1.69 (1.32-2.16)	0.038	0.04	1.37 (0.96-1.96)	*SYT16 (intronic)*	6.21×10^{-6}
	rs62639692	G	3.38×10^{-6}	0.047	1.87×10^{-5}	1.69 (1.33-2.16)	0.038	0.04	1.37 (0.96-1.96)	*SYT16 (intronic)*	6.00×10^{-6}
18p11.32	rs113776891	C	3.77×10^{-8}	0.076	1.75×10^{-7}	0.45 (0.33-0.62)	0.039	0.13	0.80 (0.53-1.18)	*PTPRM* (intronic)	1.15×10^{-6}
20q12	rs6124419	A	1.21×10^{-6}	0.12	2.51×10^{-6}	1.53 (1.28-1.83)	0.096	0.02	1.26 (1.01-1.58)	*PTPRT* (intronic)	6.82×10^{-7}

HR, hazard ratio; MAF, minor allele frequency

[a]p-value in GWA analysis performed in two steps using a confounder-adjusted phenotype.

[b]one-sided 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).

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

Locus	SNP	Effect allele	*p* initial GWA[a]	Allele frequency Tori-Bossito	*p* Cox model Tori-bossito[b]	HR (CI 95%) Tori-Bossito	Allele frequency Allada	*p* Cox model Allada[b]	HR (95%) Allada	Nearest Genes	*p* Meta-analysis
3q21.3	rs9871671	A	9.79×10^{-6}	0.11	5.00×10^{-5}	1.49 (1.22-1.81)	0.097	0.02	1.30 (1.01-1.66)	*UROC1* (exonic)	8.25×10^{-6}
11q13.5	rs545152253	"-"	3.90×10^{-7}	0.017	1.16×10^{-5}	2.64 (1.71-4.08)	0.018	0.045	1.61 (0.92-2.82)	*ACER3* (intronic)	1.62×10^{-5}
	rs115655584	G	2.80×10^{-6}	0.014	5.31×10^{-5}	2.58 (1.62-4.08)	0.016	0.041	1.65 (0.93-2.94)	*ACER3* (intronic)	2.35×10^{-5}
20q12	rs6124419	A	3.19×10^{-7}	0.12	1.35×10^{-6}	1.58 (1.31-1.91)	0.096	0.02	1.28 (1.00-1.64)	*PTPRT* (intronic)	4.10×10^{-7}

HR, hazard ratio; MAF, minor allele frequency

For 11q13.5 locus, 16 other SNPs which replicated (one in complete LD with rs545152253 and 15 in nearly complete LD with rs115655584) were not added to the table.

[a]p-value in GWA analysis performed in two steps using a the confounder-adjusted phenotype

[b]one-sided 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)

Table 3 Candidate SNP associations

Gene	SNP	Variant	Location (GRCh37)	MAF	Change	Malaria attacks		Malaria infections	
						β(CI 95%)	p	β(CI 95%)	p
HBB	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
TNF	rs1799964	3'UTR, *TNF* -857	6:31574531	0.19	T > C	0.002	0.92	0.02	0.59
	rs1800629	3'UTR, *TNF* -308	6:31543031	0.11	G > A	0.018	0.11	-0.011	0.80
IL10	rs1800871	3'UTR, *IL10* -819	1:206946634	0.38	C > T	0.023	0.23	0.02	0.51
IL3	rs40401	Exon 1, missense	5:131396478	0.71	C > T	0.050	0.01	-0.08	0.02
ABO	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
ATP2B4	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

MAF, minor allele frequency;

Results of association tests performed in the discovery cohort using the confounder-adjusted phenotype. β is the estimate of the SNP effect under an additive model. A negative value indicates a protective effect of the alternative allele and conversely.

Supplementary Tables

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

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

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

Table S1

Description of main characteristics of Tori-Bossito cohort (572 singletons followed > 28 days) for infants included or not in the GWAS

Characteristics	GWA sample (n=525)	Infants excluded from analysis (n=47)	P-value
Health center n(%)			
Tori Avamè	175(33.3)	13(27.7)	
Tori Cada	256(48.8)	20(42.6)	
Tori Gare	94(17.9)	14(29.8)	0.14
Ethnic group[a] n(%)			
Tori	380(73.6)	28(59.6)	
Fon	55(10.7)	5(10.6)	
Others	81(15.7)	14(29.8)	0.04
Maternal characteristics			
Age mean(SD)	27.4(5.5)	27.7(5.9)	0.73
Primigravidae n(%)	81(15.4)	5(10.6)	0.50
Placental malaria n(%)	58(11.1)	4(8.9)	0.83
Anaemia at delivery (Hb < 100g/L) n(%)	83(16.0)	13(28.3)	0.06
IPTp[b] use n(%)	438(83.4)	41(87.2)	0.68
Education of woman n(%)			
No education	445(84.8)	37(78.7)	
Partial primary	56(10.7)	6(12.8)	0.37
Complete primary or more	24(0.05)	4(8.6)	
Infant characteristics			
Gender male n(%)	261(49.7)	28(59.6)	0.25
Low birth weight (< 2500 g) n(%)	50(9.5)	4(8.5)	1
Bed net use n(%)			
seldom	36(7.8)	2(7.4)	
freq -	84(18.2)	6(22.2)	
freq +	129(28.0)	7(25.9)	
always	212(46.0)	12(44.4)	
non available	64	20	

[a]Ethnic group declared by the mother

[b]IPTp Intermittent preventive treatment during pregnancy with sulfadoxine pyrimethamine

Table S2

Description of main characteristics of Allada cohort (400 singletons) for infants included or not in the GWAS

Characteristics	GWA sample (n=250)	Infants excluded from analysis (n=150)	P-value
Health center n(%)			
Attogon	61(24.4)	35(23.3)	0.90
Sekou	189(75.6)	115(76.6)	
Ethnic group[a] n(%)			0.68
Aïzo	170(68.0)	108(72.0)	
Fon	55(22.0)	28(18.7)	
Others	25(10.0)	14(9.3)	
Maternal characteristics			
Age mean(SD)	26.2(5.5)	25.4(5.4)	0.13
Primigravidae n(%)	36(14.4)	27(18.0)	0.42
Placental malaria n(%)	25(10.1)	18(12.1)	0.66
Anaemia at delivery (Hb < 100g/L) n(%)	40(16.0)	31(20.7)	0.29
IPTp[b]group(%)			0.13
SP	77(30.8)	61(40.7)	
MQ	173(70.2)	89(60.3)	
Education of woman n(%)			0.06
No education	168(67.2)	115(76.7)	
Primary or more	82(32.8)	35(23.3)	
Marital status			0.73
married	245(98.0)	146(97.3)	
celibate	5(2.0)	4(2.7)	
Indicators of socioeconomic status:			
having electricity n(%)	87(34.8)	51(34.0)	0.96
having a television n(%)	80(32.0)	53(35.3)	0.57
having a refrigerator n(%)	13(5.2)	3(0.2)	0.19
rewarding activity of woman n(%)	113(45.2)	64(42.7)	0.70
Infant characteristics			
Gender male n(%)	121(48.4)	68(45.6)	0.67
Low birth weight (< 2500 g) n(%)	15(6.0)	21(14.0)	0.01
Bed net use n(%)			
seldom	11(4.4)	3(5.4)	
freq -	29(11.6)	3(5.4)	
freq +	29(11.6)	10(17.9)	
always	181(72.4)	40(71.4)	
non available	0	39	

[a] Ethnic group declared by the mother

[b] IPTp Intermittent preventive treatment during pregnancy in MiPPAD trials. Women were randomised to receive two doses of IPTp: SP, sulfadoxine-pyrimethamine; MQ, mefloquine.

Table S3

Random effects Cox models used to adjust for covariates, discovery cohort

	Mild malaria attacks		**Malaria infections**	
Covariates	HR (95%CI)	*P*	HR (95%CI)	*P*
Health center				
Tori Avamè	Reference		Reference	
Tori Cada	2.38 (1.90-2.97)	1.4e-14	2.55 (2.03-3.19)	1.1e-16
Tori Gare	1.52 (1.14-2.03)	4.3e-03	1.77 (1.32-2.35)	9.0e-05
Risk of exposure	1.26 (1.16-1.37)	6.4e-09	1.24	1.4e-09
Transmission season				
dry season	Reference		Reference	
rainy season 2007	1.76 (1.06-2.90)	2.8e-02	2.15 (1.38-3.36)	6.8e-04
rainy season 2008	1.46 (1.15-1.87)	1.7e-03	1.62 (1.31-2.00)	6.2e-06
rainy season 2009	3.27 (2.48-4.31)	1.1e-16	2.57 (2.00-3.28)	7.9e-14

HR, Hazard Ratio. Final model included covariates associated at P<0.05. Covariates selection lead to the same covariates set for both traits. Models were performed on 554 children followed more than 3 months, for whom 1072 malaria infections, among them 811 mild malaria attacks were observed.

Table S4

Random effects Cox models used to adjust for covariates, replication cohort

	Mild malaria attacks		**Malaria infections**	
Covariates	HR (95%CI)	*P*	HR (95%CI)	*P*
Health center				
Attogon	Reference		Reference	
Sekou	1.51 (1.16-1.96)	1.6e-03	1.33 (1.02-1.72)	0.02
Risk of exposure	1.50 (1.28-1.75)	2.9e-07	1.50 (1.29-1.74)	3.9e-08
Transmission season				
dry season	Reference		Reference	
rainy season 2011	1.11 (0.85-1.43)	0.44	1.19 (0.94-1.50)	0.13
rainy season 2012	1.53 (1.23-1.91)	1.1e-04	1.41 (1.16-1.72)	5.5e-04
Marital status				
Celibate	Reference		Reference	
married	0.47 (0.26-0.86)	0.01	0.49 (0.25-0.96)	0.03
Education of woman				
No education	Reference		Reference	
Primary or more	0.74 (0.59-0.94)	0.01	0.78 (0.62-0.99)	0.04

HR, Hazard Ratio. Final model included covariates associated at P<0.05. Covariates selection lead to the same covariates set for both traits. As few events were recorded in rainy season 2013 (end of the follow-up at the beginning of this rainy season) they were grouped with rainy season 2012. Models were performed on 295 children followed more than 3 months, for whom 901 malaria infections among them 687 mild malaria attacks were observed.

Supplementary Figures

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

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

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

Fig. S1

Principal component analysis on study data.

PCA were performed using 624 unrelated individuals for calculation of principal components (PC) with projection onto of remaining individuals: a) and b) plots of the first four PC by cohort; c) and d) plots of the first two PC by ethnic groups and health centers. The substructure observed appeared to be linked to the cohort and inside cohort to health centers rather than to ethnic groups.

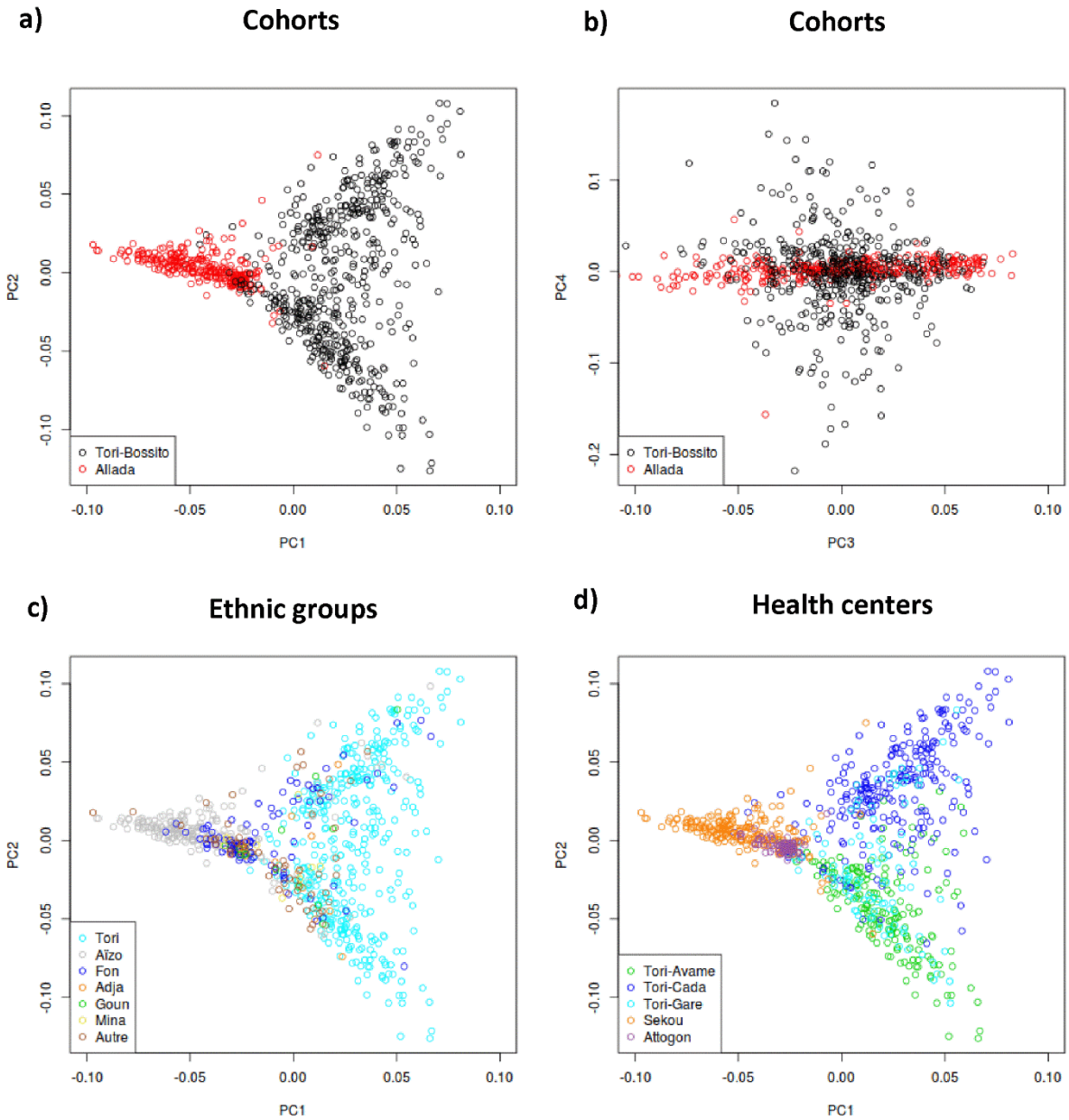
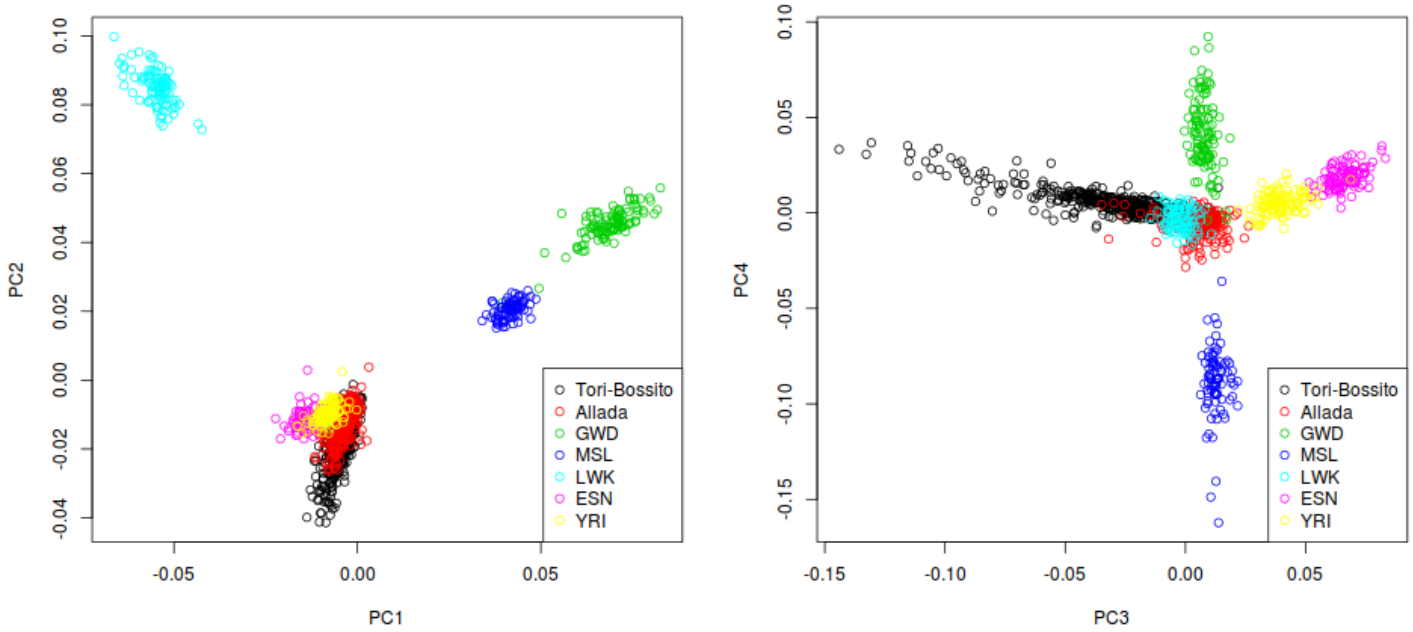


Fig. S2 Principal Component Analysis including KGP African population. PCA were performed including 100 unrelated individuals selected at random in each of the two study cohorts and African KGP populations (West African and Kenyan populations). The remaining individuals of our cohort were then projected on axes. PC1 to PC4 were plotted.



GWD = Gambian, MSL = Mende in Sierra Leone, LWK = Luhya in Kenya, ESN = Esan in Nigeria, YRI = Yoruba in Nigeria.

Fig. S3 QC flow chart for both cohorts prior to imputation. The upper box contain the initial number of children and SNPs genotyped before QC. The middle square boxes contain the number of children excluded and SNPs removed at the different steps of the QC. The bottom boxes contain the final number of children and SNPs available for GWAS and imputation.

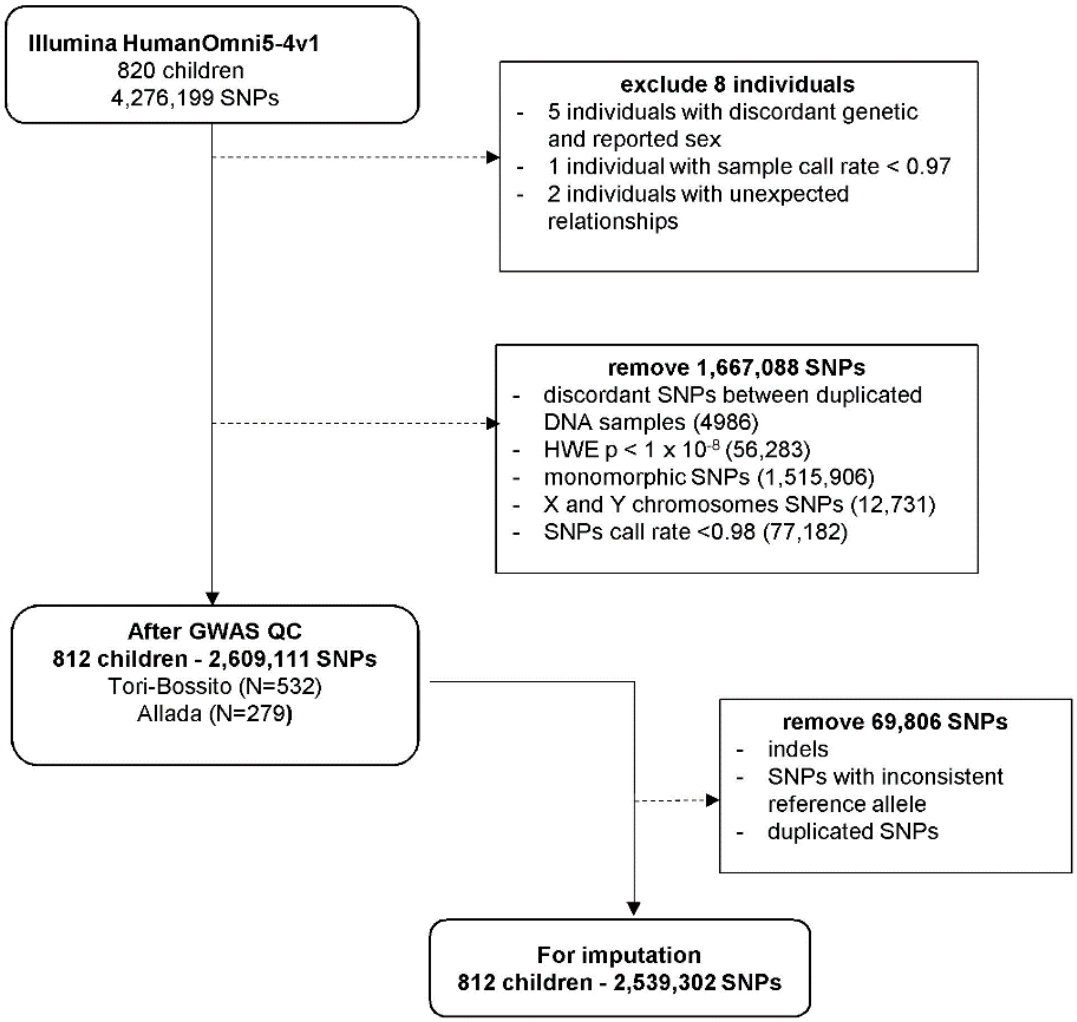


Fig. S4 Flow chart for both the discovery and replication cohorts prior to GWAS. The left boxes contain the numbers of children available after genotyping QC. Some children were additionally excluded due to a low quality follow-up or a too short follow-up. Right boxes contain the final numbers of children and SNPs available for GWAS.

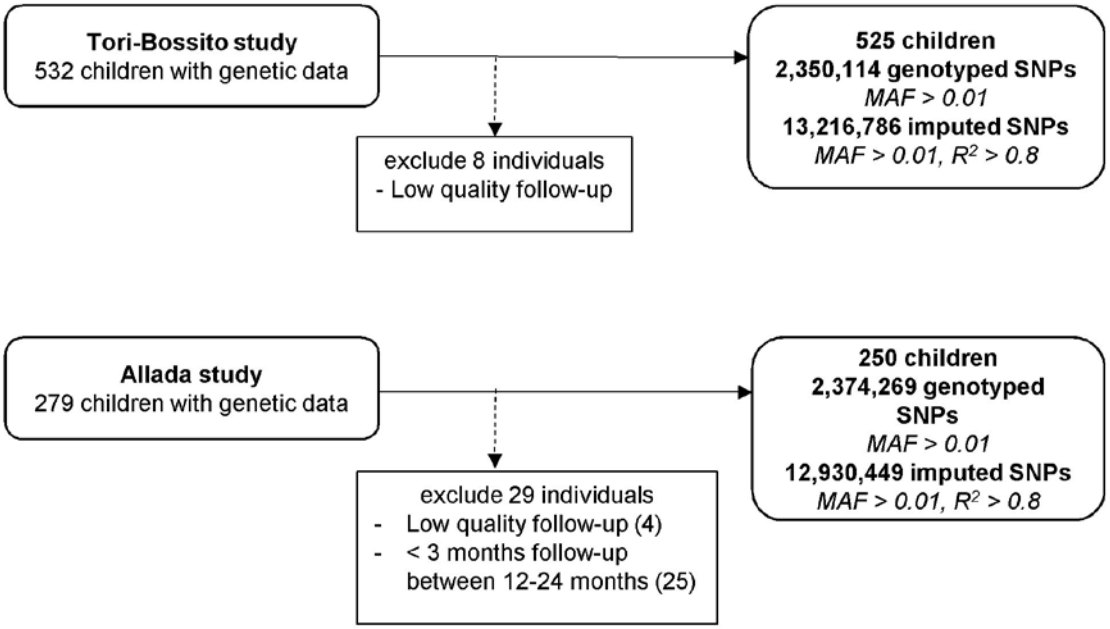


Fig. S5 Two cohorts (x-axis) versus African reference panel (y-axis) reference allele frequencies after pre-imputation QC steps.

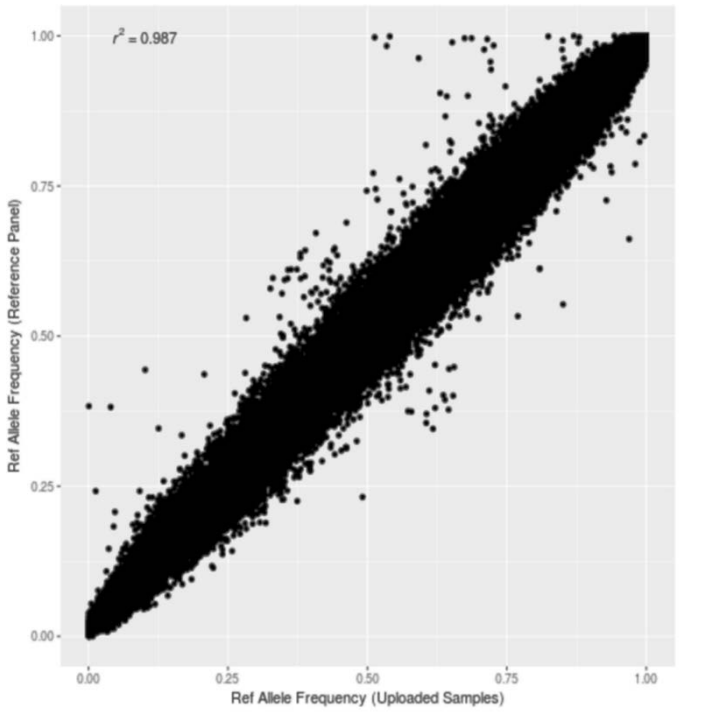
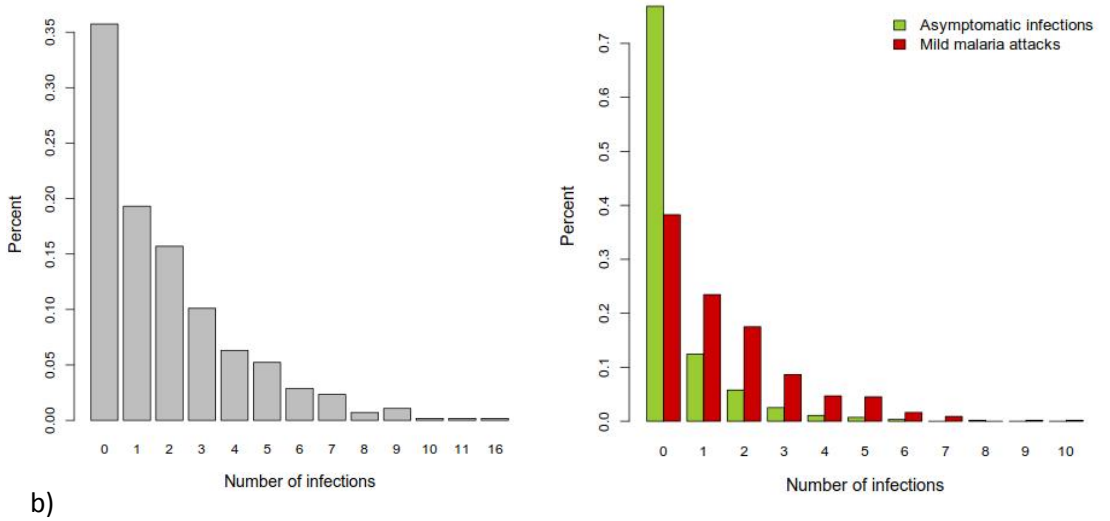


Fig S6 Distribution of malaria infections in a) discovery and b) replication cohorts. On the left, in grey, the number of malaria infections as a whole by infant (asymptomatic infections and mild malaria attacks), on the right the number of malaria infections by type.

a)



b)

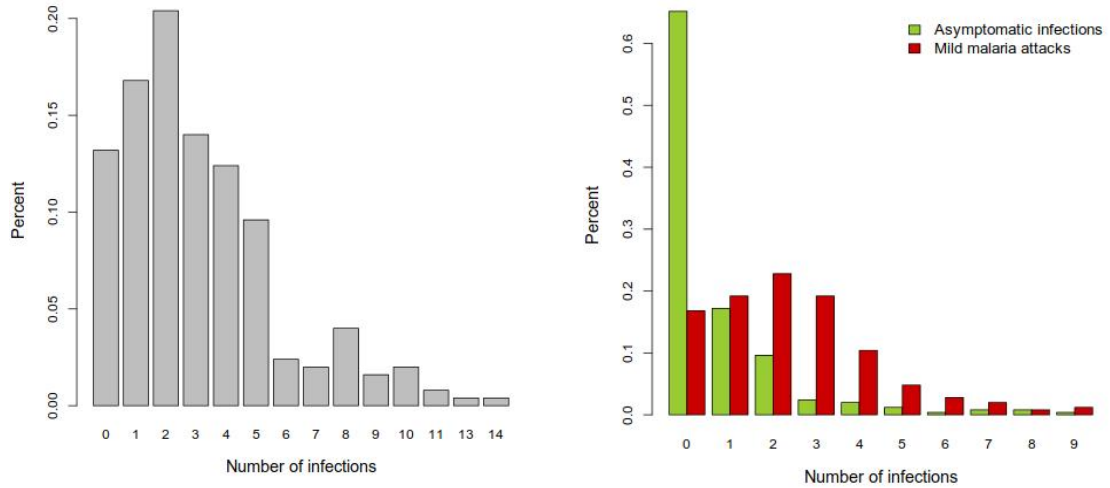


Fig S7. Incidence rate of malaria attacks and mean environmental prediction for the two cohorts. a) Incidence rate of malaria attacks per month per child during the follow-up. Incidence is shown for months for which more than 20 children were followed. Colors indicate the year of the follow-up. The pic of transmission is observed between June and September, slightly offset from the long rainy season (April-July). Malaria transmission is low at the end of dry season (January-March) but is non-null. In Tori-bossito, incidence rates are the lowest the first year and the highest the last one because infants were included at birth and the risk of malaria infections was increasing with age. b) Mean of the environmental prediction by week of follow-up. In both study, the environmental prediction estimate an *Anopheles density* but with two different methodologies for mosquitoes capture (human landing catches in Tori-Bossito and CDC ligth traps in child room in Allada). Ligth blue rectangles inform on the rainy season period (including a long one and a short one separated by one month without rain).

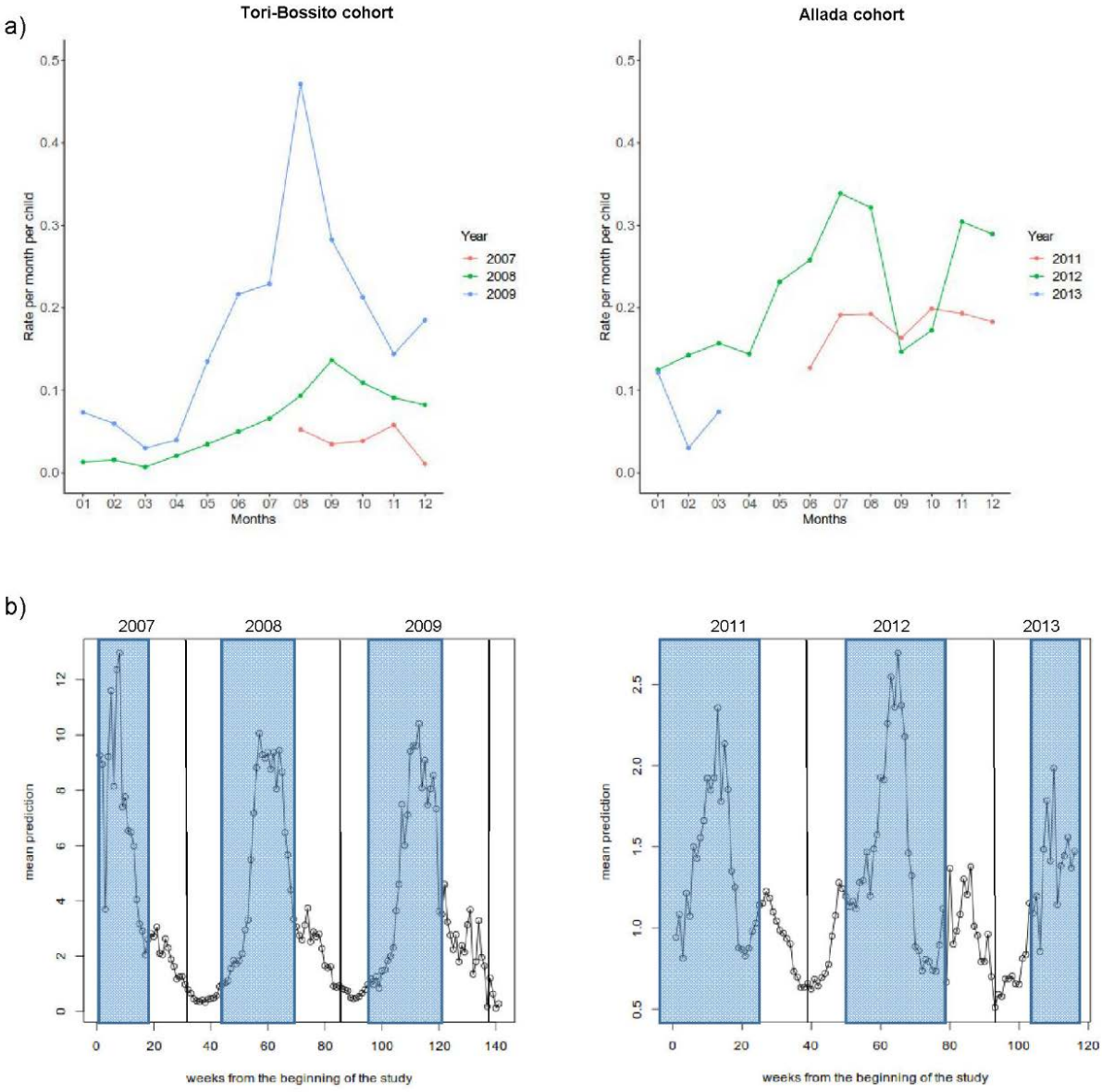


Fig. S8 Annotated regional association plots for main association signals. Regional association plots showing the results of association in the discovery cohort in a 200kb region around the lead SNPs together with the recombination rate (blue line) and the location of genes in the region. Genotyped SNPs are represented by circle, imputed ones by triangle. Color inform on the LD between the SNP and the lead SNPs (from dark blue - $r^2 < 0.2$ - to red - $r^2 > 0.8$), which was calculated on Nigerian Population of KGP dataset (ESN and YRI populations). To improve the graphic visual only SNPs with $p < 0.1$ were plotted.

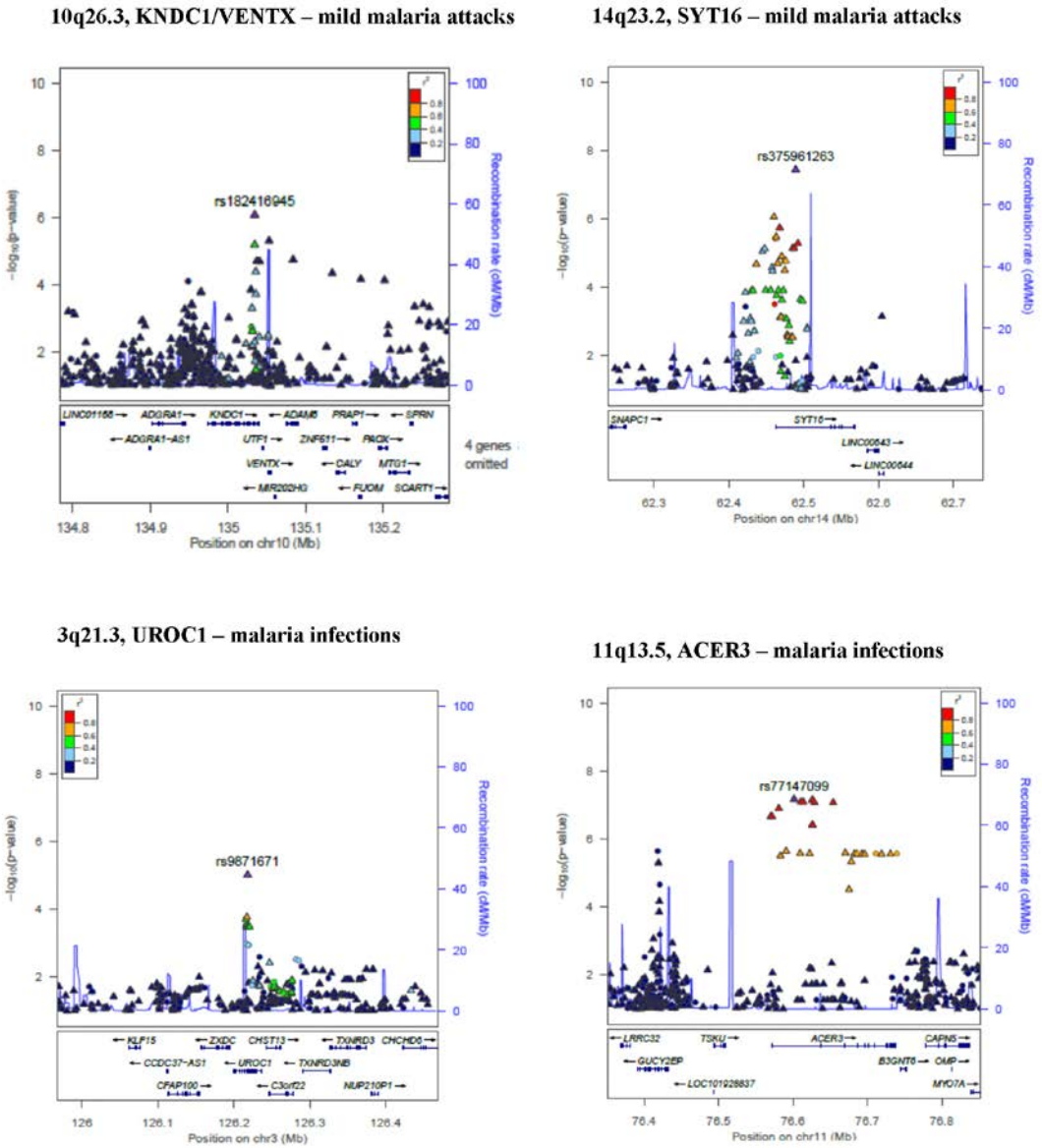
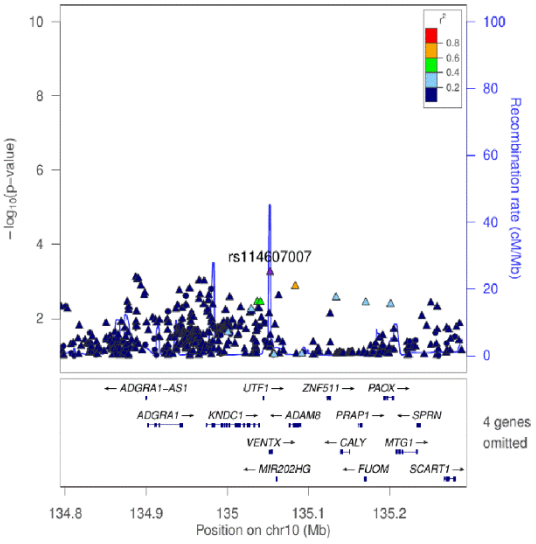
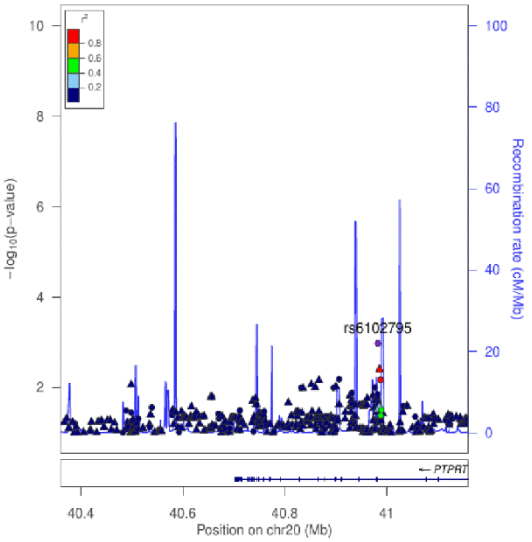


Fig. S9 Results of association analysis after conditioning on the lead SNP, for main association signals. Regional association plots showing the results of association tests conditioning on the lead SNP, in the discovery cohort, in a 200kb region around the SNPs. Genotyped SNPs are represented by circle, imputed ones by triangle. Color inform on the LD between the SNP and the lead SNPs (from dark blue - $r^2 < 0.2$ - to red - $r^2 > 0.8$), which was calculated on Nigerian Population of KGP dataset (ESN and YRI populations). To improve the graphic visual only SNPs with $p < 0.1$ were plotted.

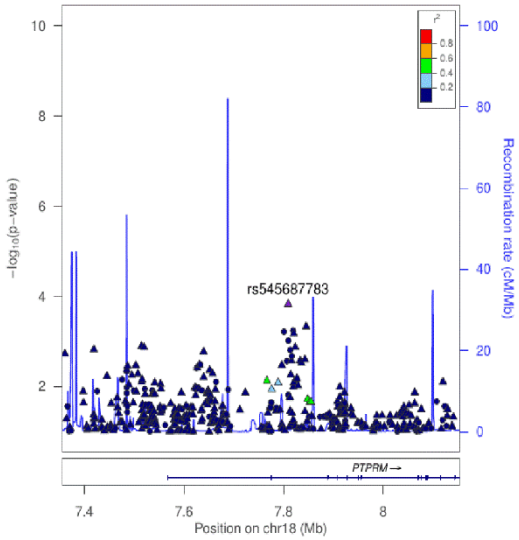
10q26.3, KNDC1/VENTX – mild malaria attacks



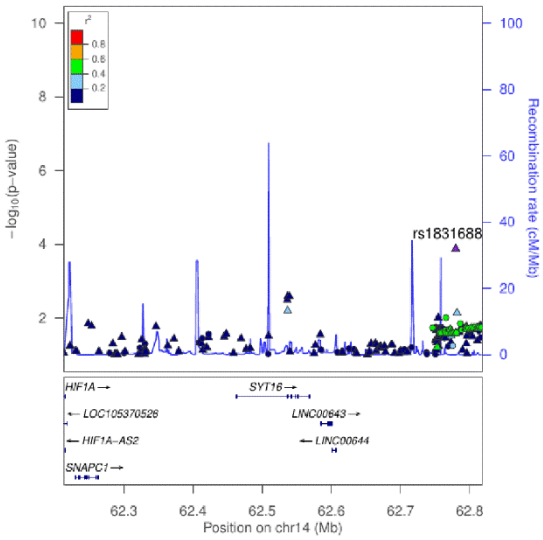
20q12, PTPRT – mild malaria attacks



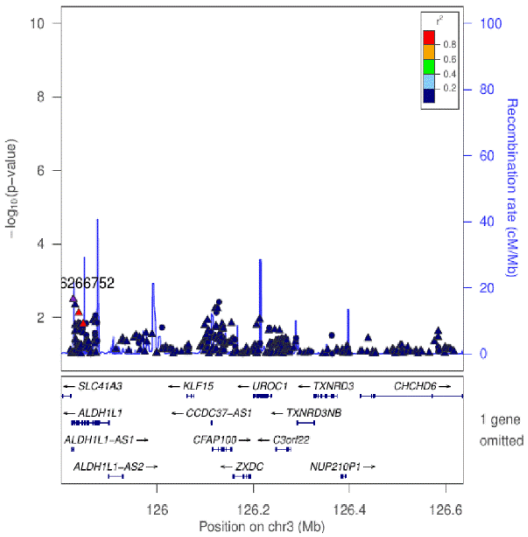
18p11.23, PTPRM – mild malaria attacks



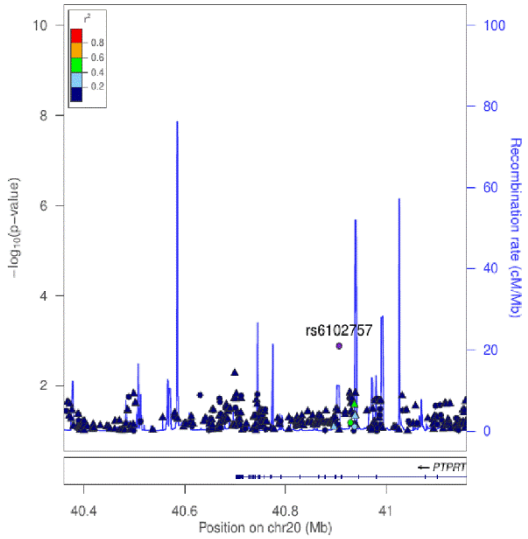
14q23.2, SYT16 – mild malaria attacks



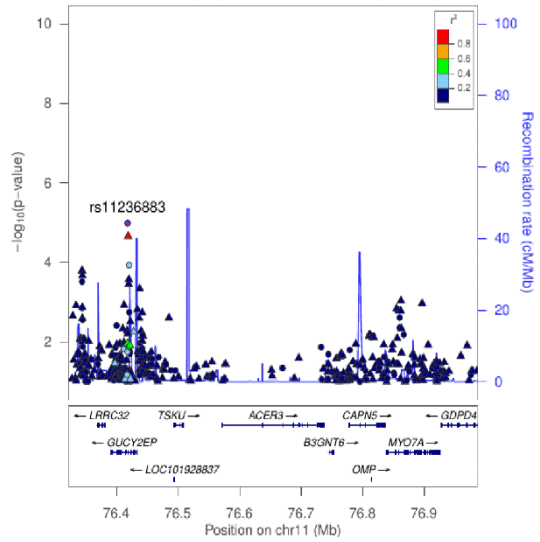
3q21.3, UROC1 –malaria infections



20q12, PTPRT –malaria infections

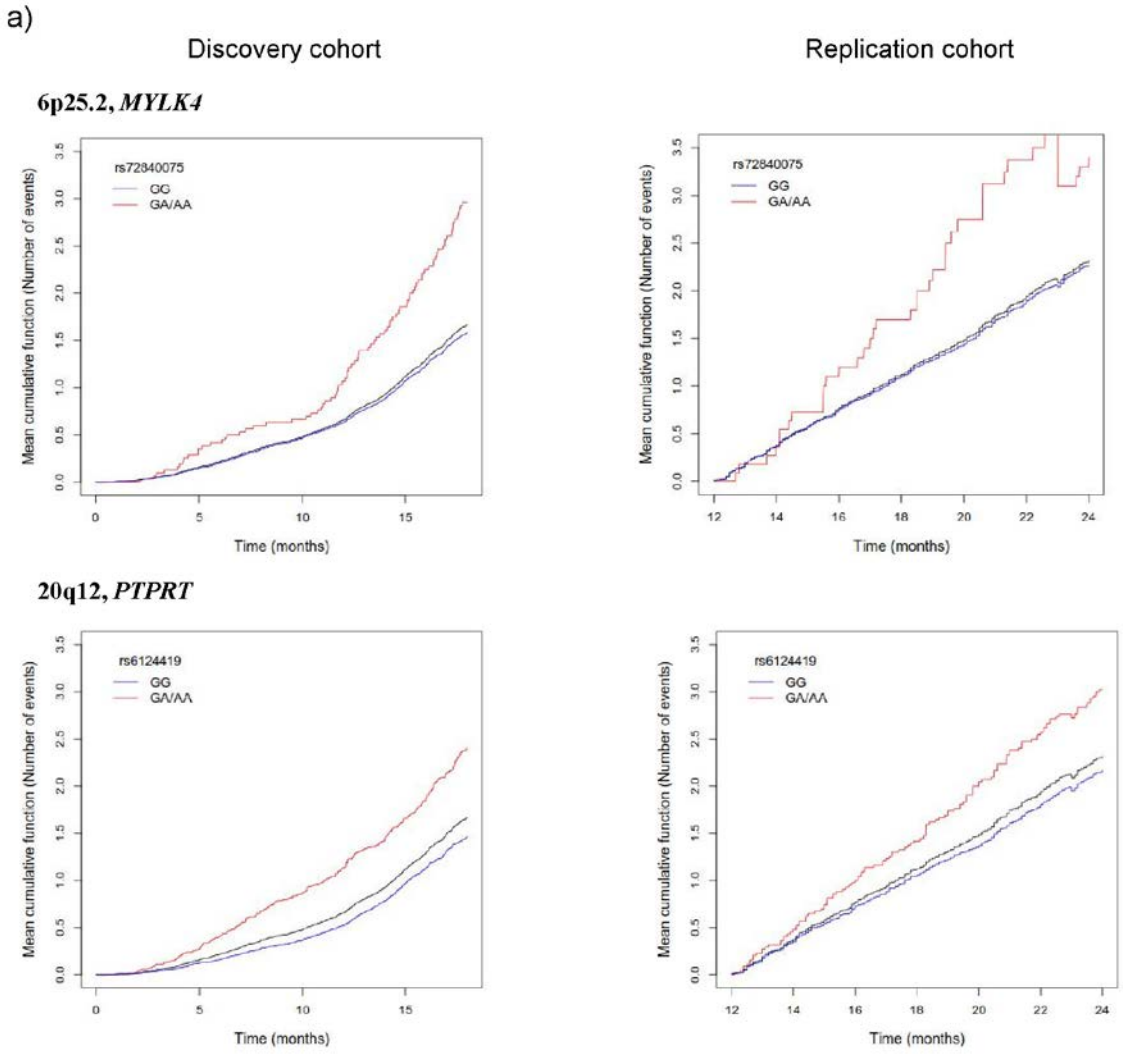


11q13.5, ACER3 – malaria infections



FigS10. Mean cumulative function by genotypes categories in discovery and replication cohorts for the main loci identified. a) Mean cumulative function (MCF) of mild malaria attacks. b) MCF of whole malaria infections.

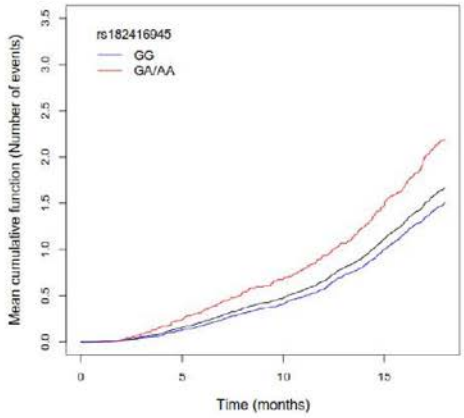
MCF is the average number of events experienced by an individual at each point in time since the start of the follow-up. The black line represent the MCF for whole cohort, the red line, MCF for individuals heterozygous or homozygous for the minor allele, the blue one, MCF for individuals homozygous for the major allele. These representations are not adjusted on covariates. Note that in the discovery cohort (follow-up from birth to 18 months) the rate of infection increases around circa 12 months of age; in the replication cohort (follow-up from 12 to 24 months), it is constant over time.

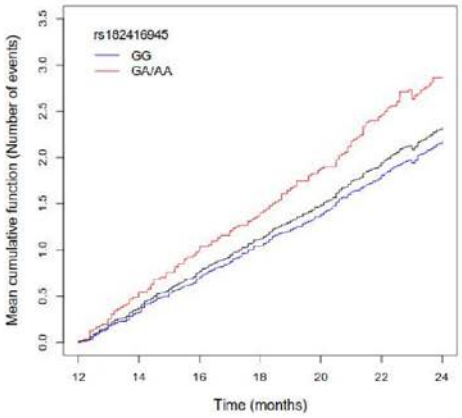


Discovery cohort

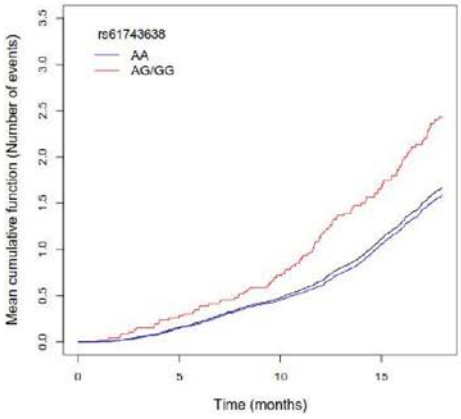
Replication cohort

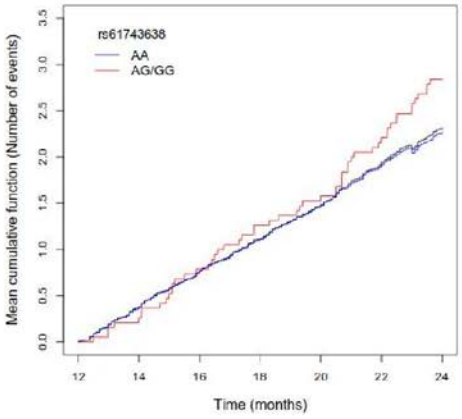
10q26.3, *KNDC1/VENTX*



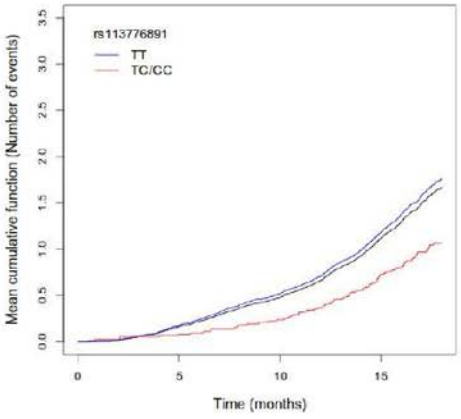


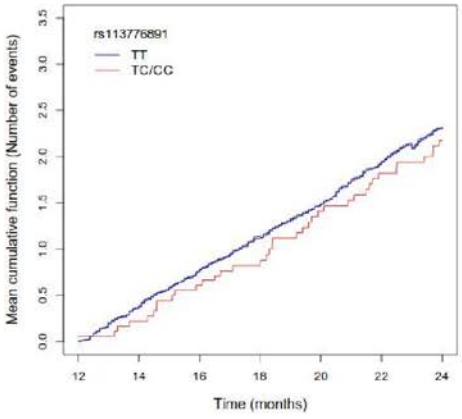
14q23.2, *SYT16*





18p11.32, *PTPRM*



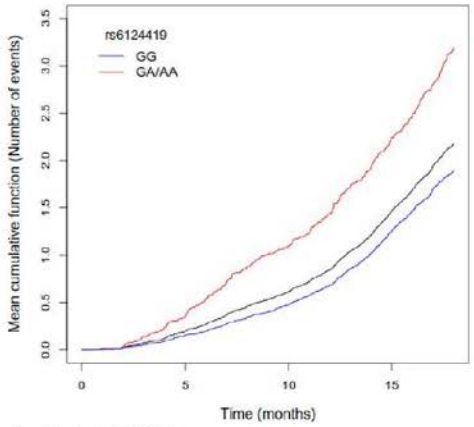


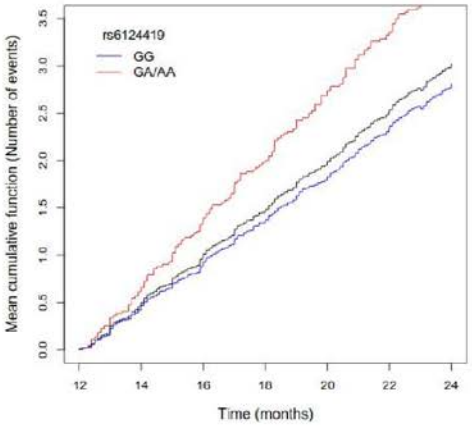
b)

Discovery cohort

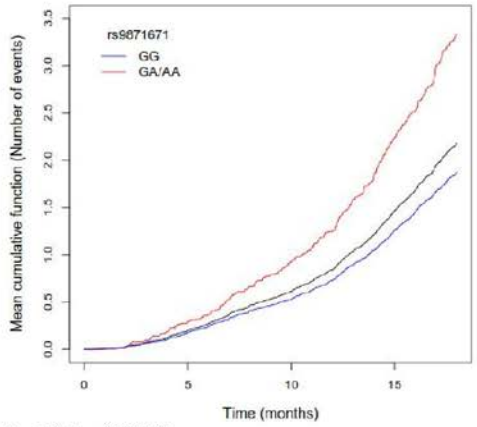
Replication cohort

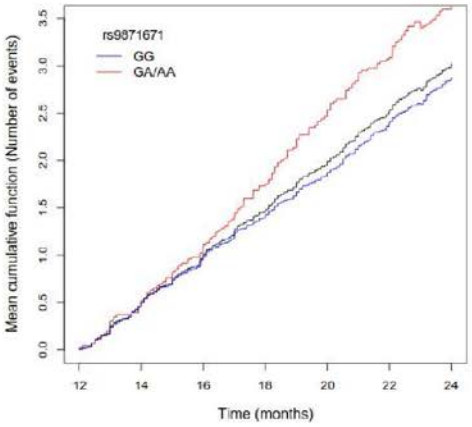
20q12, *PTPRT*



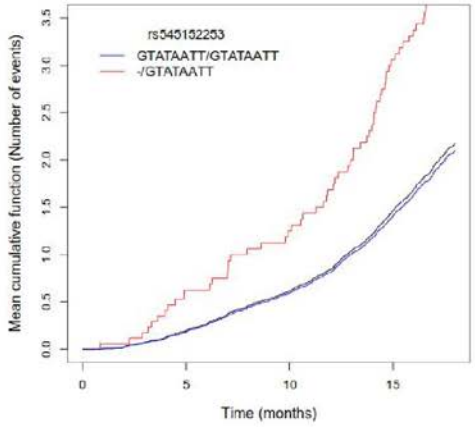


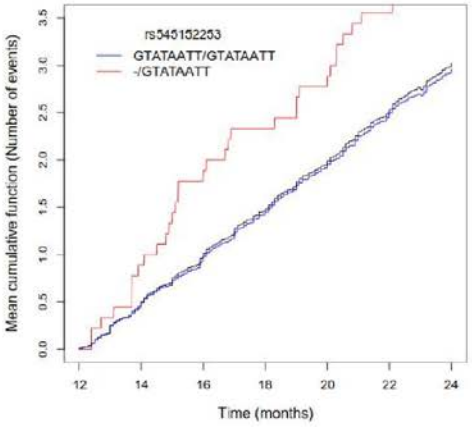
3q21.3, *UROC1*





11q13.5, *ACER3*





Article 2

BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

Mixed logistic regression in genome-wide association studies

Jacqueline Milet[1], David Courtin[1], André Garcia[1] and Hervé Perdry[2*]

*Correspondence:
herve.perdry@inserm.fr
[2] Université Paris-Saclay, UVSQ, Inserm, CESP, 94807 Villejuif, France
Full list of author information is available at the end of the article

Abstract

Background: Mixed linear models (MLM) have been widely used to account for population structure in case-control genome-wide association studies, the status being analyzed as a quantitative phenotype. Chen et al. proved in 2016 that this method is inappropriate in some situations and proposed GMMAT, a score test for the mixed logistic regression (MLR). However, this test does not produces an estimation of the variants' effects. We propose two computationally efficient methods to estimate the variants' effects. Their properties and those of other methods (MLM, logistic regression) are evaluated using both simulated and real genomic data from a recent GWAS in two geographically close population in West Africa.

Results: We show that, when the disease prevalence differs between population strata, MLM is inappropriate to analyze binary traits. MLR performs the best in all circumstances. The variants' effects are well evaluated by our methods, with a moderate bias when the effect sizes are large. Additionally, we propose a stratified QQ-plot, enhancing the diagnosis of *p* values inflation or deflation when population strata are not clearly identified in the sample.

Conclusion: The two proposed methods are implemented in the R package **milorGWAS** available on the CRAN. Both methods scale up to at least 10,000 individuals. The same computational strategies could be applied to other models (e.g. mixed Cox model for survival analysis).

Keywords: GWAS, Mixed-models, Logistic regression

Background

Population stratification has long been known to be at the origin of spurious associations in genetic association studies [1]: if the frequency of the phenotype of interest varies across the population strata, it will be associated to any allele the frequency of which varies accordingly. An early and elegant solution to this issue has been the use of family data, notably in the Transmission Disequilibrium Test (TDT) [2] and in the Family Based Association Test (FBAT) [3]. However, these methods imposed the ascertainment and genotyping of affected individuals' relatives, impairing their practical feasibility. The advent of Genome-Wide Association Studies (GWAS), demanding increasingly large samples to detect weaker and weaker effects, made the problem even more accurate.

Methods adapted to large scale population studies have thus been proposed, among which the Genomic Control [4] and Structured Association methods [5, 6]. The Genomic Control uses the empirical distribution of the genome-wide chi-square statistics to correct the statistic inflation attributable to population structure whereas Structured Association methods infer population strata from genome-wide data before testing for association conditional to the strata.

A major breakthrough was achieved in 2006 with EIGENSTRAT [7], also known as Principal Component Regression (PCR). This conceptually simple but extremely efficient method consists in incorporating the top Principal Components (PC) of the genotype data in a linear model. The next major advance was the introduction of mixed models, which can be interpreted as a generalization of PCR which incorporates all PCs in the model with random effects [8, 9]. Incorporating a few PCs with fixed effects as well in the mixed model might still prove useful to correct statistic inflation at SNPs with large allelic frequency variations across strata [8, 10]. Fast approximate [11] or exact [12] methods for genome-wide analysis of quantitative traits with mixed linear models (MLM) were soon made available.

The analogue of MLM for case-control studies is the mixed logistic regression (MLR). However, fitting the MLR with the Penalized Quasi-Likelihood (PQL) [13] is computationally heavy. Thus, in case-control studies the status was often coded as quantitative trait (0 for control subjects and 1 for disease subjects), and analyzed as such. Nevertheless, Chen et al. [14] proved that when disease prevalence was heterogeneous between populations strata, while the overall distribution of p values is well corrected by this method, it leads to conservatives p values for some SNPs and to anti-conservative p values for others. This behavior was made evident by the mean of quantile-quantile plots (QQ-plots) in which SNPs were categorized according to their allele frequencies in the different population strata. In order to address this issue, Chen et al. proposed GMMAT, a score test for the MLR, which is feasible in GWAS. The p values obtained with this test were shown to be well distributed in all SNP categories.

While the score test has a reduced computational burden, its drawback is the absence of an estimation of the variants' effects. When only the effects of genome-wide significant SNPs are needed, an obvious solution is to fit MLR models including each of these SNPs. When the SNPs' effects are needed for the whole genome, e.g. for meta-analysis purposes, it is desirable to have a computationally efficient method to estimate these effects. We propose two such methods in this paper.

The strategy to reduce the computational burden is to fit the MLR only once for a "null model". This null model may incorporate relevant covariates, but no SNP. The SNPs are then tested one by one. Roughly speaking, this is done by using an appropriate method to confront a vector of genotypes with the residues of the null model. This stragegy is shared by the score test (GMMAT) and by the two proposed methods; but while the score test provides association p values with no estimate of the variant's effects, the methods proposed here give such estimates.

One of the proposed method, named hereafter Approximate Maximum Likelihood Estimate (AMLE), is based on a first-order approximation of the MLR, which leads to an approximation of the SNPs effect. The association is tested by a Wald test, which is identical to the score test of [14]. A similar approach has been previously used for mixed

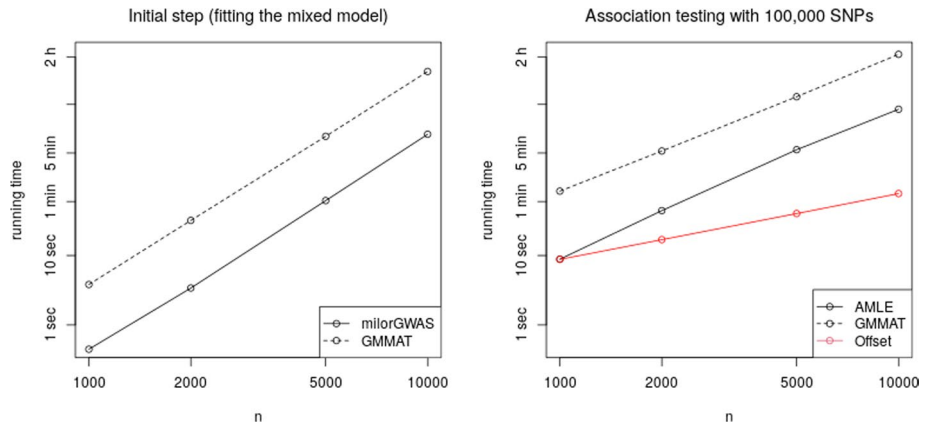


Fig. 1 Running times for sample sizes varying from $n = 1000$ to $n = 10{,}000$. All computations were run on a Intel i7 at 3.60 GHz with 8 MB of cache memory. Due to the log–log scale, the timing of an algorithm with complexity $O(n^{\alpha})$ appears as a straightline with slope α: slope 3 for the initial step, 2 for AMLE and GMMAT, and 1 for the Offset method

linear models [15] and has been recently adapted in SAIGE [15] for MLR, but without an evaluation of its capacity to properly estimate SNP effects.

The other method, named the Offset method, which bears similarities with the methods of [11], consists of first estimating individual effects in a mixed logistic regression model, and then incorporating these effects as an offset in a (non-mixed) logistic regression model.

In the sequel, we evaluate the capacity of logistic regression (LR), MLM, and MLR (using either the score test or the methods mentioned above) to properly take into account population structure associated with heterogeneous prevalence. While Chen et al. were interested in geographically distant populations, spanning several countries in South America, part of our work focuses two geographically very close populations in West Africa, using real genotype data from a recent GWAS [16]. We also use data simulated with a coalescent model [17], reproducing the simulations presented in [14]. We use similar simulations to evaluate the ability of the PQL and of our two methods to properly evaluate SNPs' effects.

Additionally, we propose to generalize the categorization of SNPs in QQ-plots presented in [14] to variables other than the population strata, including continuous variables as for example the first PC. The interest of this generalization is demonstrated using the same simulations.

All methods are implemented in the R package milorGWAS (for mixed logistic regression in GWAS), freely available on the Comprehensive R Archive Network (CRAN).

Results

Computational efficiency of the methods

We measured the running times of GMMAT and of the methods AMLE and Offset, implemented in our package milorGWAS, for a sample size n varying from 1000 to 10, 000. The results are reported in Fig. 1. The figure is on log–log scale, showing that the results are in good accordance with the asymptotic complexity of the methods, given in

the Methods section: the time taken by the initial step (fitting the mixed model) growing as n^3, it appears linear on log–log scale with a slope close to 3. Similarly, the complexity $O(n^2)$ of GMMAT and AMLE result in straight lines with a slope of 2. GMMAT is consistently slower than milorGWAS: milorGWAS runs roughly 8 times faster for the initial step, and 6 times faster for the testing step. Part of the difference in the testing step is due to the fact that GMMAT reads data and write results on disk, whereas milorGWAS uses the RAM.

The Offset method is in $O(n)$ while the AMLE and GMMAT are in $O(n^2)$, which makes this method appealing in terms of running times. However, in practice the initial step in $O(n^3)$ is the limiting factor: it would take more than 24 hours to fit the MLR model for $n = 50{,}000$. To manipulate larger samples such as those of biobanks, more complex workarounds and other approximate methods are needed [15].

Type I error in the presence of population structure

We analyzed two data sets with a simulated binary phenotype. Simulations were based either on real genotype data from a recent GWAS [16], or on large cohort simulated with a coalescent model. Both data sets present cryptic relatedness and population stratification, with two strata (or two cohorts) with different disease prevalence. The simulations are fully described in the Methods section.

The first simulated set uses genotypes of 800 individuals from South Benin, ascertained in two sites distant of 20 km from each other, forming the two population strata in which we set different prevalences for simulating the phenotypes. These data were analyzed using the logistic regression (LR), the mixed logistic regression (MLR) (with Chen's score test GMMAT or equivalently Wald test with AMLE) and the mixed linear model (MLM) (the status being analyzed as a quantitative trait with values 0 or 1), with or without the top 10 principal components (PCs) in the model.

The Fig. 2 displays the stratified QQ-plots of the corresponding p values, the 1, 847, 505 SNPs being split in three categories according to their allelic frequencies as described in the methods (we used a threshold $th = 0.8$). Category 1 contained 11.4% of the SNPs, categories 2 and 3, respectively 77.6% and 11.0% of the SNPs.

When no PCs are included (first row of the Fig 1), statistic inflation is observed for LR ($\lambda = 1.182$). Based on the non-stratified QQ-plot, both MLM and MLR appear to adequately correct for population structure; however the stratified QQ-plot shows that this is not the case for SNPs in categories 1 and 3. In particular, for MLM, there is a statistic inflation for SNPs in category 3, and a deflation for SNPs in category 1.

When 10 PCs are included in the models (second row of the figure), this difference of behavior between SNPs categories persists for MLM. However, both logistic regression and MLR show an adequate correction for all categories of SNPs.

The second simulation set consists in 10, 000 individuals simulated on a 20×20 grid, whose genotypes were obtained from a coalescent model. These data include some first order relatives. Ten millions independent SNPs were simulated, among which 2,840,903 had a minor allele frequency above 5%. Two strata were defined according to the individuals position on the grid (a "high risk strata" was defined as being the top left quarter of the grid), and a binary phenotype was simulated with different prevalences on these strata. Similar analyzes were performed on these data (Additional file 1: Figure 1). The

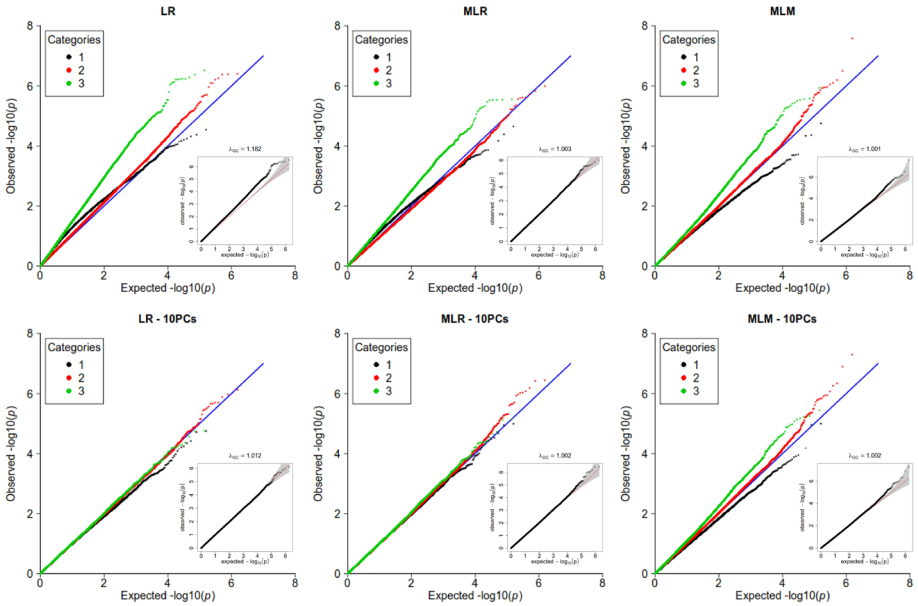


Fig. 2 Simulated data based on South Benin genotypes (1). Stratified QQ-plots for logistic regression (LR), mixed logistic regression (MLR) using Chen's score test (or AMLE), and mixed linear model (MLM) on the data simulated based on South Benin data. In each panel, a non-stratified QQ-plot is embedded. On the second row, 10 PCs were included as covariates. SNP categories are determined as in [14], based on the allele frequencies in the strata

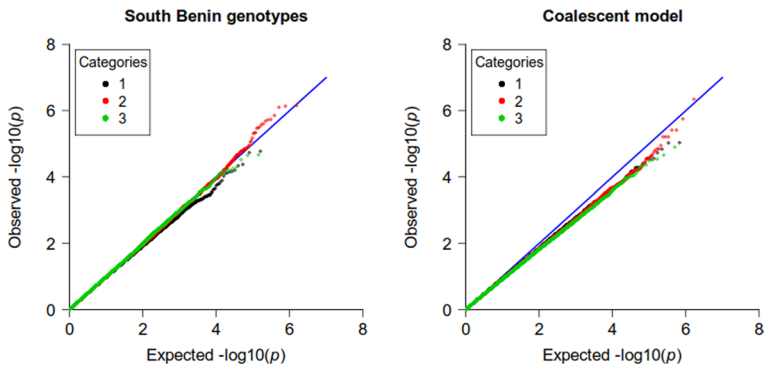


Fig. 3 Simulated data based on South Benin genotypes (2). Stratified QQ-plots for the Offset method, for simulations based on South Benin data (left) and on the coalescent model (right). SNP categories are determined as in [14], based on the allele frequencies in the strata

2,840,903 independent SNPs are at 23, 7% in category 1, 58.8% in category 2 and 17.5% in category 3. The same patterns of inflation and deflation of the test statistic are retrieved for most analyzes; the only notable difference is that, in this case, the logistic regression with 10 PCs does not correct the statistics inflation completely.

Both data sets were also analyzed using the Offset method for MLR. Figure 3 shows the QQ-plots for the Offset method with the top 10 PCs included in the model, on the two simulations sets. While it adequately corrects for population structure in the data from the South Benin, it is too conservative in the case of the simulated cohort.

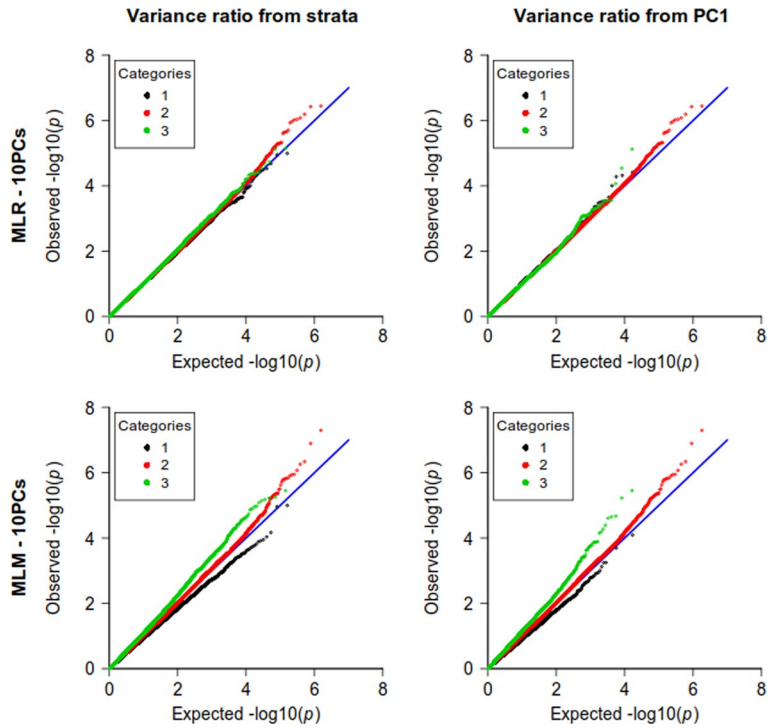


Fig. 4 Stratified quantile–quantile plots based on PCs. Stratified QQ-plots obtained from the allele frequencies in the two strata (left) and from the first PC coordinates (right) for simulations based on South Benin data

Extension of the stratified QQ-plot

We extended the stratified QQ-plot proposed in [14] to the case in which strata are not clearly defined, or the strata information is missing. Our extension relies on the use of the first PCs instead of the strata.

Figure 4 compares the stratified QQ-plots obtained using either strata information (as in Figs. 2 and 3) or the first PC, for two of the analyses already considered in Fig. 2, the MLM and the MLR, both including 10 PCs as fixed effects. The same comparisons were performed for analyses on the simulated cohort (Additional file 1: Figure 2). While there are small differences between the QQ-plots, we see that they allow similar diagnostics, that is, an incomplete correction of population structure for MLM analyses, and in contrast an adequate correction for MLR.

Estimation of the SNPs' effects

Figure 5 shows the bias $(\widehat{\gamma} - \gamma)$ obtained for two different values of γ, in three scenarios corresponding to different magnitude of cohort and random effects (A: moderate cohort and random effect; B: moderate cohort effect, large random effect; C: large cohort effect, moderate random effect). Three MAF bins were considered (from 0.05 to 0.10, from 0.20 to 0.25 and from 0.45 to 0.50).

In all situations, the PQL displays no bias, or a very small bias (for example in scenario B, corresponding to large random effects). The two proposed methods tend to have a

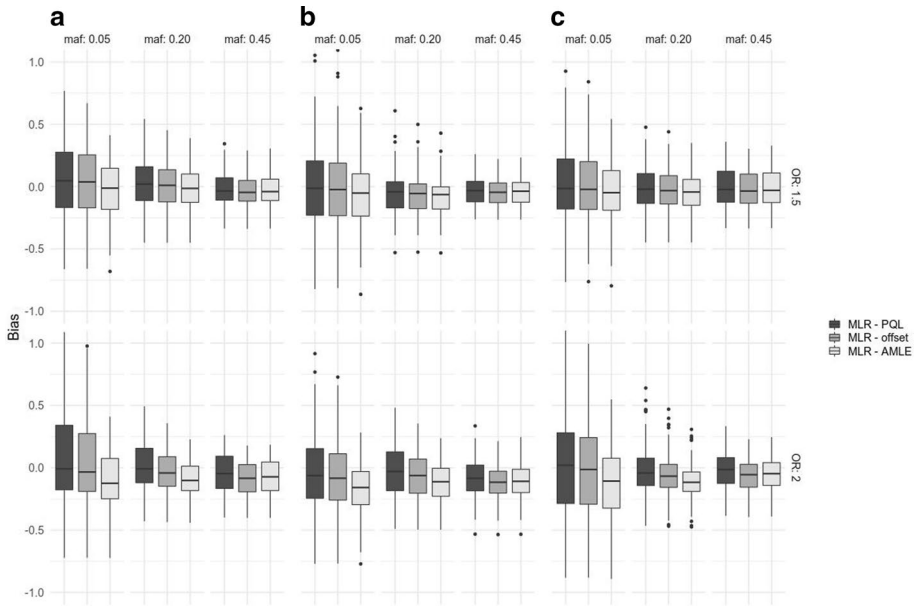


Fig. 5 Bias of the SNP effect estimates $\widehat{\gamma}$ for two values of γ [$\gamma = \log(1.5)$ and $\gamma = \log(2)$], in three scenarios (**a** moderate cohort and random effect; **b** moderate cohort effect, large random effect; **c** large cohort effect, moderate random effect), and three MAF bins (0.05 to 0.10, 0.20 to 0.25, and 0.45 to 0.50)

Table 1 Comparison of powers in simulations based on South Benin data (800 individuals), with $\tau = 1$. p_0 and p_1 are the prevalences in the two strata

Scenario	p_0	p_1	OR	MAF bin	LR	AMLE	Offset
Moderate cohort effect	0.10	0.20	3	(0.20; 0.25]	0.179	0.261	0.268
Moderate cohort effect	0.10	0.20	3	(0.45; 0.50]	0.899	0.920	0.906
No cohort effect	0.30	0.30	2.5	(0.20; 0.25]	0.473	0.496	0.458
No cohort effect	0.30	0.30	2.5	(0.45; 0.50]	0.926	0.928	0.913

Powers at the genome-wide significance level 5×10^{-8} are reported for logistic regression (LR), AMLE (equivalent to the score test) and Offset methods, all analyses including 10 PCs with fixed effect

negative bias. For data simulated with a moderate SNP effect $\gamma = 0.4$ (corresponding to $OR = 1.5$), they both have small negative biases (at most -0.08 corresponding to estimated $OR = 1.4$), independently of the population structure's effect size (scenarios A, B and C).

For larger SNPs effects ($\gamma = 0.7$, corresponding to $OR = 2$), the bias increases, with AMLE having the larger bias, attaining in some situations -0.1 (corresponding to an estimated $OR = 1.8$). The bias is slightly more important for large random effects (scenario B).

Power of the approximate methods

We compared the power of methods for which no inflation of type I error have been observed for all categories of SNPs. Table 1 reports powers obtained in simulations based on South Benin data for LR, AMLE (equivalent to the score test and GMMAT)

Table 2 Comparison of powers in simulations based on the coalescent model (5000 individuals), with $\tau = 1$

Scenario	p_0	p_1	OR	MAF bin	LR	AMLE	Offset
Large cohort effect	0.05	0.30	1.5	(0.20; 0.25]	0.307	0.251	0.202
Large cohort effect	0.05	0.30	1.5	(0.45; 0.50]	0.597	0.544	0.464
No cohort effect	0.30	0.30	1.5	(0.20; 0.25]	0.871	0.859	0.834
No cohort effect	0.30	0.30	1.5	(0.45; 0.50]	0.996	0.989	0.987

p_0 and p_1 are the prevalences in the two strata. Powers at the genome-wide significance level 5×10^{-8} are reported for logistic regression (LR), AMLE (equivalent to the score test) and Offset methods, all analyses including 10 PCs with fixed effect

and Offset methods. The AMLE performs slightly better than the two other methods, and in particular than LR.

Table 2 is dedicated to the simulations based on the coalescent model. Here LR has the largest power, which was expected as its type I error is largely inflated. Similarly, the Ofset method which was overconservative in this simulation setting, has a lowest power than AMLE.

Illustration on real data

We tested the association with the presence of malaria infection during follow-up in the South Benin data. We focused on SNPs with a minor allele frequency greater than 0.05 on 100 kb segment on chromosome 20, and used both the AMLE and the PQL. The results are displayed in the Additional file 1: Figure 3 (Manhattan plot in panel a). The most associated SNPs (at $p = 10^{-4}$) is the same as in the GWAS (rs6124419). The association signal is weaker than in the original GWAS, which used the recurrence of infections in a Cox model.

The two methods give similar p values (panel b of the figure), but AMLE produces slightly higher p values for the most associated SNPs. Consistently, the largest β values are slightly understimated by AMLE.

Discussion

Our first result is a reproduction of the observation made by Chen et al. that is, in the presence of heterogeneity of disease prevalence between population strata, the mixed linear model (MLM) is inappropriate to analyse binary traits. MLM leads to conservative p values for some SNPs, and to anti-conservative p values for others, depending on the ratio of expected genotype variance in the two strata. The motivation of Chen et al. was a genome-wide association study of asthma including individuals from different Caribbean and Latin American backgrounds, with in particular *ca.* 15% of individuals from Puerto Rico, in which the prevalence of asthma was much higher (25.6%) than in other populations (from 3.9 to 9.6%). We retrieved similar results in an analysis of a simulated phenotype with large differences of prevalence among strata, based on genotype data from two geographically close cohorts (*ca.* 20 km apart) from South Benin [16], but with different self-reported ethnicities. Heterogeneity of prevalence may result from environmental factors (e.g. lifestyle, nutritional behavior, etc.), and could occur frequently in association studies, thus making the analysis of binary traits

with the MLM incorrect, in particular in populations with a high genetic diversity, such as African populations.

A similar result was retrieved with data simulated with a coalescence model on a square grid, the "high-risk strata" consisting in the top left quarter of the grid; these simulations included many first-order related individuals and a random effect based on the kinship matrix. The mixed logistic regression (MLR) including the top PCs as fixed effects is the only method to completely correct for population structure in both simulations. A classical logistic regression including the top PCs can however be worth considering, as this conceptually simpler method was efficient enough for the South Benin data, in which the level of relatedness, though high, is lower than in the simulated data. In terms of power, however, our simulations hints towards a higher power of the MLR even in this case.

It is worth to note that in both simulations sets, it was necessary to include the top PCs as fixed effects in the MLR to obtain the correct type I error. The interest of including top PCs alongside the random components has been noted before [8, 10].

The diagnosis of the correctness of type I error cannot be based on the sole QQ-plot of p values, as the behavior of the test differs in SNP categories defined from the allelic frequencies in the two strata, as mentioned above. When these strata are clearly identified in the study, Chen et al. introduced a QQ-plot stratified on SNP categories based on the allele frequencies in the strata. We propose an extension of this method that can be used when population information is not available, using the first PC as a proxy (or any continuous variable defined at the population level, independently of the phenotype). Our simulations show that this method produces QQ-plot similar to those obtained with full knowledge of the two strata, thus allowing to diagnose better whether the population structure is adequately taken into account or not.

However, while the presence on the (stratified) QQ-plot of a deviation of the p values from their expected distribution implies that the association analysis is incorrect, the reverse is not necessarily true. Chen et al. demonstrated that new diagnostic plots can unveil hidden structures in the p values distribution; other new diagnostic plots could unveil other structures. More generally, while the correlation between polygenic effects is well modeled by the GRM, there is no reason that all unaccounted environmental variables have a correlation matrix similar to the GRM, and in theory the mixed model may prove inappropriate in some real studies. However, it seems to us that environmental variables that are not correlated to the genetic background would likely not be confounding variables.

Regarding SNPs's effect estimation, made possible by the two proposed methods, our simulations studies show that both methods are slightly biased downward. The bias of the Offset method is less important than the bias of the AMLE, while the PQL has virtually no bias. It is known that in the presence of unaccounted heterogeneity, logistic regression effect estimates have negative biases [18–20]; our result hints that the heterogeneity between population strata is not fully taken into account by these methods. However, the bias is sensible only for large effects such as $OR = 2$, which makes its impact virtually negligible in GWAS. Note also that these methods are not adapted to the analysis of rare variants in highly unbalanced case-control ratio, as demonstrated in [15] for GMMAT – and thus for AMLE.

Conclusion

We proposed two fast methods to estimate SNPs' effects in mixed logistic regression. Both methods scale to up to at least 10,000 individuals, making them suitable for analysis of most GWAS data. Their implementation in an R package allows flexible use, with for example the possibility to specify a user defined GRM matrix.

The methods are constructed with conceptually simple mathematical principles which could be applied to other models (e.g. mixed Cox model in survival analysis), although literal computations to derive the formulas in the AMLE can be tedious. The Wald test performed with the AMLE is equivalent to the score test of [14], and thus the conclusions drawn for the MLR apply regarding type I error. The second method, which we called the Offset method, have similar performances on the simulations based on the South Benin genotypes, but is slightly over-conservative in the presence of strong familial effects in the simulations based on the coalescent model.

All methods are available in an R package **milorGWAS** based on the R package **Gaston** [21] for data manipulation. The R and C++ source code of **milorGWAS** is available on the CRAN at https://CRAN.R-project.org/package=milorGWAS. An excerpt of the data simulated using the coalescent model is also available, and is used in the package's vignette to illustrate the methods.

Methods

Fast methods for mixed logistic regression

The mixed logistic regression model (or logistic mixed model) considered is

$$\operatorname{logit} E(Y) = \operatorname{logit} P(Y = 1) = X\beta + G\gamma + \omega \tag{1}$$

where Y is an n-dimensional vector of zeroes and ones ($\operatorname{logit} E(Y)$ is the vector of components $\operatorname{logit} E(Y_i)$), X is a $n \times p$ matrix of covariates (including a column of ones for the intercept), G is a vector of genotypes (ususally coded 0, 1 or 2), and ω is a random vector following a multivariate normal distribution $MVN(0, \tau K)$.

The likelihood of this model involves an integral over the random vector ω. There is no closed-form for this integral, and numerical integration schemes are computationally intensive in high dimension. A classical approximate solution is the Penalized Quasi-Likelihood (PQL) algorithm, which is a sequence of approximations of the MLR model by linear mixed models. Even the PQL is too computationally intensive for estimating the effect γ of all SNPs in a GWAS.

We describe below two approximate methods for estimating γ. The first is based on an approximation of the maximum likelihood estimate in the PQL. The second consists in first estimating predicted linear scores $X\widehat{\beta} + \widehat{\omega}$ in the MLR model (1) without the term $G\gamma$, and then incorporating these linear scores as an offset in a (non-mixed) logistic regression model.

Approximate Maximum Likelihood Estimate (AMLE)

We outline here the general principle on which the formula presented in Additional file 1 can be derived. This principle could be applied to any statistical model. Let $\ell(\kappa, \gamma)$ be a log-likelihood, in which κ is a nuisance parameter and γ is the parameter of interest.

The null hypothesis to be tested is $H_0 : \gamma = 0$. Denote $\widehat{\kappa}_0$ the Maximum Likelihood Estimator (MLE) of κ under the null hypothesis:

$$\widehat{\kappa}_0 = \arg\max_{\kappa} \ell(\kappa, 0).$$

The score statistic to test for H_0 is the first derivative in γ at the point $(\widehat{\kappa}_0, 0)$:

$$U(0) = \frac{\partial}{\partial \gamma} \ell(\widehat{\kappa}_0, 0).$$

The null hypothesis can be tested using $T = U(0)^2 / \operatorname{var}(U(0))$, which asymptotically follows a $\chi^2(1)$ distribution. A second-order approximation, with $\kappa = \widehat{\kappa}_0$ fixed, gives

$$\ell(\widehat{\kappa}_0, \gamma) \simeq \ell(\widehat{\kappa}_0, 0) + \frac{\partial}{\partial \gamma} \ell(\widehat{\kappa}_0, 0)\gamma + \frac{1}{2} \cdot \frac{\partial^2}{\partial \gamma^2} \ell(\widehat{\kappa}_0, 0)\gamma^2$$

which maximizes in

$$\widehat{\gamma} = -\frac{\frac{\partial}{\partial \gamma} \ell(\widehat{\kappa}_0, 0)}{\frac{\partial^2}{\partial \gamma^2} \ell(\widehat{\kappa}_0, 0)}.$$

This estimator of $\widehat{\gamma}$ is not the MLE of γ, but when the true value of γ is small enough, both estimators are close.

In the context of GWAS, κ is the vector of random term variance and covariates effects, while γ is the effect of the SNP to be tested. This estimator shares with the score test the advantage that $\widehat{\kappa}_0$ has to be estimated only once; the partial derivatives in γ are usually easy to compute, allowing a fast testing and estimating procedure.

However, as mentioned above, in the case of the MLR, the likelihood can't be computed efficiently. We use the PQL to estimate the nuisance parameter $\kappa = (\beta, \tau)$, and the log-likelihood of the last linear approximation used in the PQL to estimate γ. The variance of the resulting $\widehat{\gamma}$ is estimated in this linear approximation; the resulting Wald test is identical to the score test of [14]. All details are given in the Additional file 1.

Offset

The proposed method consists of estimating a vector of individual effects, including both the random components and the covariates in X, which is then incorporated in a logistic regression as an offset:

- First estimate $\widehat{\beta}_0$ and $\widehat{\omega}_0$ under the hypothesis $\gamma = 0$, in the MLR model

  $$\operatorname{logit} E(Y) = X\beta + \omega$$

  with ω as in (1).
- Then, for each vector of genotypes G, fit a linear model for

  $$E(G) = X\delta,$$

let $\widetilde{G} = G - X\widehat{\delta}$ be the residuals of G, and estimate γ in the fixed-effects logistic regression

$$\text{logit}\, E(Y) = X\widehat{\beta}_0 + \widehat{\omega}_0 + G\gamma,$$

in which the vector $X\widehat{\beta}_0 + \widehat{\omega}_0$ is an offset (that is, is held constant).

The motivation of this heuristic is that a similar two-steps method applied to a linear model $E(Y) = X\beta + G\gamma$ would give the same estimator $\widehat{\gamma}$ than the classical regression (cf Additional file 1 for the details).

Asymptotic complexity of the methods and efficiency of the implementation

The initial step of both AMLE and Offset method is to fit the MLR model $\text{logit}\, E(Y) = X\beta + \omega$. Each iterative step of the PQL needs the inversion of an $n \times n$ matrix, where n is the number of individuals; the complexity of this operation is $O(n^3)$. The second step of the AMLE involves, for each SNP, multiplying an $n \times n$ matrix P by the vector of genotypes; this is the most costly operation, and the complexity of this step is thus $O(n^2)$. The second step of the Offset method is an iterative algorithm, each iteration of which is in $O(n)$ (considering the number of covariates as fixed). The complexity of the second step of the Offset is thus $O(n)$.

Our package milorGWAS is implemented in C++ using Rcpp [22], and RcppEigen [23] for matrix arithmetic. We performed simulations with random genotypes for sample size $n = 1000$, $n = 2000$ and $n = 5000$, to assess the performance of the implementation.

Stratified QQ-plot

One of the contributions of Chen et al. was to show that a QQ-plot of log p values was not sufficient to diagnose an incorrect test procedure, and to propose a "stratified QQ-plot" in which different categories of SNPs are represented separately. This allowed to see that in some of these categories, the test statistics are either inflated or deflated, while the overall distribution of p values was correct. Here is how their categories were defined. Chen et al. consider a population with two strata, indexed by $i = 0$ or 1 according to disease prevalence in ancestry groups, $i = 1$ being the group with a higher risk of desease. The strata are assumed to be panmictic, so that expected variance of a SNP genotype G in stratum i is $\text{var}_i(G) = 2p_iq_i$, p_i and q_i being the SNP allele frequencies. Each SNP is categorized according to the variance ratio $r(G) = \text{var}_1(G)/\,\text{var}_0(G)$ between the two strata as follows (Chen et al. use a threshold $th = 0.8$):

- The SNPs with $r(G) < th$ are category 1,
- the SNPs with $th \leq r(G) \leq 1/\,th$ are category 2,
- the SNPs with $1/\,th < r(G)$ are category 3.

We propose to extend the method to stratify QQ-plots according to any covariate Z. If $G \in \{0, 1, 2\}^n$ is the vector of genotypes, Z a vector with components in the range [0, 1], and $\mathbf{1}$ denotes a vector of ones, we let

$$q_1 = \frac{1}{2}\frac{Z'G}{Z'\mathbf{1}} \text{ and } q_0 = \frac{1}{2}\frac{(\mathbf{1}-Z)'G}{(\mathbf{1}-Z)'\mathbf{1}},$$

and we defined SNP categories as above (with $p_i = 1 - q_i$). If Z is the indicator variable of the strata, q_1 and q_0 are the allelic frequencies in the two strata, and the categories will be identical to those of Chen et al. The point of this extension is that when the relevant sub-strata are unknown, one could use one of the top genomic PCs instead (after rescaling them to [0, 1]).

Simulation studies: type I error in the presence of population structure

We performed two sets of simulations, based on the simulations performed in [14], to assess the efficiency of the different methods to correct for population stratification. In both simulation sets, there are two strata (or two cohorts) with different disease prevalence, and related individuals. Simulations were performed under the null hypothesis of no genetic association [$\gamma = 0$ in Eq. (1)] and were analyzed with

- a logistic regression model (LR)
- a mixed linear model (MLM)
- a mixed logistic regression model, using Chen et al. score test GMMAT, identical to AMLE Wald test (MLR)
- a mixed logistic regression model, using the offset method (Offset)

All analyses were repeated with the top ten PCs included as fixed effects in the model. We assessed the capacity of each test procedure to control type I error rates using Chen's stratified QQ-plot. Moreover, to gauge the interest of our extension, we compared Chen's QQ-plot to the stratified QQ-plot obtained using the first PC instead of the cohort indicator.

Simulations based on South Benin data

We used genotype data from a GWAS on mild malaria susceptibility performed on two cohorts in South Benin [16]. The participants were ascertained in two sites distant of 20 km from each other, in three different health centers for the first cohort and two for the second one. After quality control (QC), the genotypes of 800 individuals were available, 525 in the first cohort, and 275 in the second one. The genotyping was performed with Illumina HumanOmni5 chips (1,847,505 SNPs after QC and filtering out SNPs with minor allele frequency, MAF, less than 5%). This genetic sample presents both population structure and cryptic relatedness. Self-reported ethnic composition differed between the two cohorts, and principal component analysis confirmed the presence of population structure. A sub-structure related to the health center, where the participant was ascertained, was also apparent. Moreover, substantial relatedness was observed in the sample, with levels of relationship corresponding to half-sibs, uncle-nephew or even 3/4 siblings for some pairs (estimated kinship coefficient ϕ from 0.10 to 0.16).

We simulated a binary phenotype with a difference of disease prevalence between the two cohorts, and a random effect modeling both population stratification and relatedness. Specifically, the probability on an individual i of being a case was calculated as:

$$\text{logit}(p_i) = a_0 + a_1 Z_i + \omega_i \quad (2)$$

where Z is an indicator variable for belonging to the second cohort, and ω_i an individual random effect. The coefficients a_0 and a_1 were defined as $a_0 = \text{logit}(0.05)$ and $a_1 = \text{logit}(0.30) - \text{logit}(0.05)$, so as to obtain a prevalence of 0.05 in the first cohort and of 0.30 in the second one (not taking into account the presence of random effects). The vector of random effects was simulated following a multivariate normal distribution:

$$\omega \sim \mathcal{N}(0, \tau K)$$

where K is the Genomic Relationship Matrix (GRM) calculated from all the SNPs. We set $\tau = 1$.

Simulations based on a coalescent model

We also performed coalescent simulations, reproducing closely the simulations described in [14], to obtain genotypes for a large cohort of 10,000 individuals, with both population structure and relatedness, using the `ms` software [17]. This procedure, which is based on a stepping stone model with symmetric migration between adjacent cells of the grid, is commonly used to simulate a population with a spatially continuous population structure [24, 25]. We use a 20×20 grid, in which the migration rate between adjacent cells was set to 10; this parameter produces a Wright's fixation index $F_{st} < 0.01$ when dividing the simulated grid into two equal sub-populations, a level comparable to what is observed within Europe [24]. We simulated a total of 10 million independent SNPs. After filtering out SNPs with MAF lower than 5%, 2,840,903 SNPs were available. The full command line arguments for `ms` are included in Additional file 1. We also created a R data package containing part of the simulated data (link in Additional file 1).

To obtain related individuals, we first simulated the genotypes for 8000 founders (20 on each of the 400 cells). We then sampled 10 pairs of individuals in each cell, forming 4000 couples, and simulated two offsprings by gene dropping. Thus we obtained 16,000 individuals (founders and offspring) from which 10,000 individuals were randomly selected to obtain the cohort.

The phenotype was simulated as before using Eq. (2), where Z_i was set to one when individuals were sampled in the top left 10×10 grid, corresponding to strata with a higher risk. The values of a_0, a_1 and τ were set as before; in this simulation set, $K = 2\Phi$ were Φ is the matrix of kinship coefficients (entries are 0.5 for first order relatives, 1 on the diagonal, 0 elsewhere). Data analyses were subsequently performed using a GRM calculated from 100,000 random SNPs.

Simulations studies: estimation of the SNPs' effects

Simulations including a SNP effect were performed using the South Benin data set, using the model

$$\text{logit}(p) = a_0 + a_1 Z + G\gamma + \omega$$

with G, a genotype picked at random in the data, and Z and ω as described above. We considered four different scenarios:

A Moderate cohort effect (respective prevalence $p_0 = 0.10$ and $p_1 = 0.20$) and moderate random effect ($\tau = 0.3$).
B Moderate cohort effect (respective prevalence $p_0 = 0.10$ and $p_1 = 0.20$) and large random effect ($\tau = 1$).
C Large cohort effect (respective prevalence $p_0 = 0.05$ and $p_1 = 0.30$) and moderate random effect ($\tau = 0.3$).

The coefficients a_0 and a_1 are computed as $a_0 = \text{logit}(p_0)$ and $a_1 = \text{logit}(p_1) - \text{logit}(p_0)$; G is centered to ensure that the expected prevalence is as prescribed. For each scenario, we considered SNPs with MAF in intervals (0.05; 0.10], (0.20; 0.25] and (0.45; 0.50], and SNP effect $\gamma = \log(1.5)$ and $\gamma = \log(2)$ (corresponding to $OR = 1.5$ and 2). One hundred replicates were performed for each condition, redrawing a vector of random effects each time, and analyzed with the PQL, the Offset and the AMLE, including the top 10 PCs.

Simulations studies: comparison of powers

To compare the power of the methods that have shown a correct type I error, we performed additional simulations in a similar setting as in the previous section, with a large random effect ($\tau = 1$), a moderate cohort effect (as defined above) and $OR = 3$, and without cohort effect and $OR = 2.5$. Other simulations were performed based on 5000 individuals extracted from the cohort generated under the coalescent model, with either a large cohort effect or no cohort effect, and an $OR = 1.5$. In each scenario, 1000 replicates were performed, redrawing a vector of random effects each time.

Illustration on real data

To illustrate the method, we applied it on the data from the GWAS in South Benin described in [16]. We used as binary phenotype the presence/absence of any malaria infection during the follow-up (there were 229 individuals with no infection and 546 with at least one infection). We tested all SNPs with a minor allele frequency greater than 0.05 on a 100 kb segment on chromosome 20, which contains one of the strongest association signals discovered in [16]. We included as covariates the site of ascertainment, the duration of the wollow-up, and mean infection exposure. We performed the testing with AMLE and, for comparison, with the PQL.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03862-2.

Additional file 1. Including details on the AMLE and Offset methods, commands used for the simulations with the coalescent model and supplementary figures.

Acknowledgements

We thank the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources.

Authors' contributions

HP designed the AMLE method and wrote the R package. JM designed and performed the simulations studies. JM and HP both designed the Offset method, the extension of the stratified QQ-plot, and wrote the manuscript. DC and AG designed the GWAS project on mild malaria susceptibility and provided the data from the South Benin cohort. All authors read and approved the final manuscript.

Funding

Not applicable

Availability of data and materials
The genotype datasets from South Benin used in the current study are available in the DataSuds repository (https://doi.org/10.23708/EXSQTM). The dataset are not publicly available due to patient confidentiality but are available for researchers who meet the criteria for access to confidential data. Data to reproduce the simulations based on a coalescent model are included in this published article and its supplementary information files.

Ethics approval and consent to participate
Genotype data come from a GWAS on mild malaria susceptibility [16]. The protocols of the two cohorts studies were approved by both the Beninese Ethical Committee of the Faculté des Sciences de la Santé (FSS) and the IRD Consultative Committee on Professional Conduct and Ethics (CCDE)

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Université de Paris, MERIT, IRD, 75006 Paris, France. [2] Université Paris-Saclay, UVSQ, Inserm, CESP, 94807 Villejuif, France.

Received: 25 May 2020 Accepted: 4 November 2020
Published online: 23 November 2020

References

1. Lander ES, Schork NJ. Genetic dissection of complex traits. Science. 1994;265(5181):2037–48.
2. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993;52(3):506.
3. Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered. 2000;50(4):211–23.
4. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997–1004.
5. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.
6. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000;67(1):170–81.
7. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904.
8. Zhang Y, Pan W. Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? Genet Epidemiol. 2015;39(3):149–55.
9. Dandine-Roulland C, Perdry H. The use of the linear mixed model in human genetics. Hum Hered. 2015;80(4):196–206.
10. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11(7):459.
11. Aulchenko YS, De Koning D-J, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. 2007;177(1):577–85.
12. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8(10):833.
13. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993;88(421):9–25.
14. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedón JC, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am J Hum Genet. 2016;98(4):653–66.
15. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50(9):1335.
16. Milet J, Boland A, Luisi P, Sabbagh A, Sadissou I, Sonon P, Domingo N, Palstra F, Gineau L, Courtin D, et al. First genome-wide association study of non-severe malaria in two birth cohorts in Benin. Hum Genet. 2019;138(11–12):1341–57.
17. Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002;18(2):337–8.
18. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika. 1984;71(3):431–44.
19. Cramer JS. Robustness of logit analysis: Unobserved heterogeneity and mis-specified disturbances. Oxf Bull Econ Stat. 2007;69(4):545–55.
20. Ayis S. Quantifying the impact of unobserved heterogeneity on inference from the logistic model. Commun Stat Theory Methods. 2009;38(13):2164–77.
21. Dandine-Roulland C, Perdry H. Genome-wide data manipulation, association analysis and heritability estimates in R with Gaston 1.5. Hum Hered. 2018;83:6.

22. Eddelbuettel D, François R. Rcpp: Seamless R and C++ integration. J Stat Softw. 2011;40(8):1–18. https://doi.org/10.18637/jss.v040.i08.
23. Bates D, Eddelbuettel D. Fast and elegant numerical linear algebra using the RcppEigen package. J Stat Softw. 2013;52(5):1–24.
24. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012;44(3):243.
25. Bradburd GS, Ralph PL, Coop GM. A spatial framework for understanding population structure and admixture. PLoS Genet. 2016;12(1):1005703.

Publisher's Note

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Résumé : Malgré les moyens importants de prévention et de lutte mis en place ces dernières années, le paludisme reste dévastateur avec près d'un demi-million de décès par an (405 000 en 2018, d'après le dernier rapport de l'OMS). Le rôle clé joué par les facteurs génétiques de l'hôte dans la susceptibilité et la sévérité de la maladie est admis aujourd'hui. Cependant, les bases moléculaires de la sensibilité/résistance au paludisme restent encore mal connues. Ces dix dernières années, les efforts de recherche pour l'identification de gènes impliqués dans la sensibilité au paludisme à *P. falciparum* se sont concentrés sur les formes graves de paludisme, avec plusieurs études d'association sur l'ensemble du génome (*Genome-Wide Association Study* ou GWAS) publiées. Ce manuscrit porte sur l'extension de cette approche aux formes simples du paludisme, au travers de l'étude d'association génome entier de deux cohortes de nouveau-nés au Sud Bénin (au total 800 enfants), suivis pendant 18-24 mois par l'UMR261 (MERIT IRD/Université de Paris).

Dans une première partie nous présentons les résultats de la première GWAS réalisée sur les formes simples de paludisme dans ces deux cohortes. L'association a été testée avec la récurrence des accès palustres et la récurrence de l'ensemble des infections (incluant les accès palustres et les infections asymptomatiques) en prenant en compte un risque environnemental estimé au niveau individuel. Elle met en évidence plusieurs signaux d'association forts, en lien avec des gènes dont la fonction biologique est pertinente pour le paludisme (notamment *PTPRT*, *MYLK4*, *UROC1* et *ACER3).*

La forte variabilité génétique présente au sein des populations africaines a nécessité de prendre en compte l'effet de confusion potentiel de la structure de population. Dans l'étude des formes simples de paludisme, une approche en deux étapes a été utilisée, le modèle de Cox mixte, utilisé pour l'analyse des données longitudinales, n'étant pas applicable à l'ensemble du génome du fait du temps de calcul nécessaire. Un modèle de Cox mixte a été appliqué pour construire un « effet individuel » ajusté sur les covariables, puis un modèle mixte linéaire pour tester l'association avec les polymorphismes du génome. Ceci nous a conduits à nous intéresser plus généralement aux modèles mixtes non-linéaires. Deux méthodes permettant l'estimation de l'effet des polymorphismes avec le modèle logistique mixte sont proposées, qui pourront être dans le futur généralisées à d'autres modèles, dont le modèle de Cox.

Dans une dernière partie, le paludisme ayant constitué une des plus fortes pressions de sélection que l'Homme ait connue dans son histoire récente, nous explorons la possibilité d'exploiter l'information de sélection naturelle pour augmenter la puissance de l'analyse, et améliorer la détection des signaux d'association. L'analyse des signaux de sélection positive récente sur l'ensemble du génome a été réalisée avec plusieurs méthodes basées sur les haplotypes longs (iHS, nsL and XP-EHH). Celle-ci met en évidence plusieurs régions chromosomiques d'intérêt potentiel où les signaux d'association et de sélection co-localisent ; mais confirme également la difficulté à mettre en évidence les signaux de sélection liés au paludisme avec les outils disponibles actuellement.

Abstract : In spite of numerous prevention and control efforts in recent years, malaria remains a major global public health problem with nearly half a million deaths per year (405,000 in 2018). The key role played by genetic factors of the host in the susceptibility and severity of the disease is admitted nowadays. However, the molecular basis of susceptibility/resistance to malaria has not been elucidated to date. Over the past decade, research efforts to identify genes involved in malaria susceptibility have focused on severe malaria, with several genome-wide association studies (GWAS) published. This manuscript is about the extension of this approach to uncomplicated forms of malaria, through the genome wide association study of two birth cohorts in South Benin (800 children), followed for 18-24 months by UMR261 (MERIT IRD/ University of Paris).

In the first part, we present the results of the first GWAS performed on simple forms of malaria in these two cohorts. The association was tested with the recurrence of malaria attacks and the recurrence of all infections (including malaria attacks and asymptomatic infections) taking into account an environmental risk estimated at the individual level. It highlights several strong association signals, linked to genes whose biological function is relevant for malaria (in particular *PTPRT*, *MYLK4*, *UROC1* and *ACER3*).

The high genetic diversity within African populations has made it necessary to take into account the potential confounding effect of population structure. In this study we proceeded with a two-step strategy as the Cox mixed model, used for the analysis of longitudinal data, is not applicable to the whole genome due to computational burden. In a first step, an analysis was performed with a Cox mixed model to build an "individual effect" fitted on the covariates, then a linear mixed model were used to test the association with genome polymorphisms. This led us to focus more generally on non-linear mixed models. Two methods allowing the estimation of the effect of polymorphisms with the mixed logistic model are proposed, which may in the future be generalized to other models, including the Cox model.

In a final part, malaria having been one of the strongest selection pressures that man has known in recent history, we explore the possibility of exploiting natural selection information to increase the power of analysis, and improve the detection of association signals. The analysis of recent positive selection signals were performed using several genome-scan methods focusing on patterns of long-range haplotype homozygosity (iHS, nsL and XP-EHH). This analysis revealed several chromosomic region of potential interest, where the signals of association and selection co-localized but confirms also the difficulty of highlighting the selection signals linked to malaria with tools currently available.

.