UNIVERSITÉ PARIS DESCARTES

UFR Mathématique et Informatique

Université Cheikh Anta Diop de Dakar

LUX - MEA - LEX

Faculté de Médecine de Pharmacie et d'Odontologie

No d'ordre:

THÈSE

En cotutelle, présentée pour obtenir
LE GRADE DE DOCTEUR EN SCIENCES

DE L'UNIVERSITÉ PARIS DESCARTES V
ET DE L'UNIVERSITÉ CHEIKH ANTA DIOP DE DAKAR

Spécialité: Statistique Génétique
Par Cheikh LOUCOUBAR

Statistical genetic analysis of infectious disease (malaria) phenotypes from a longitudinal study in a population with significant familial relationships

Soutenue le 21 Mars 2012 devant la Commission d'examen :

M.	Avner Bar-Hen	(Directeur de thèse)
M.	Aliou Diop	(Co- Directeur)
Mme	Margaret Mackinnon	(Rapporteur)
M.	Christophe Rogier	(Rapporteur)
M.	Alioune Dièye	(Examinateur)
M.	Richard E. Paul	(Examinateur, Président du jury)

Thèse préparée à
L'UNITÉ DE GÉNÉTIQUE FONCTIONNELLE DES MALADIES INFECTIEUSES
Institut Pasteur
75015 Paris

Et à

L'UNITÉ D'ÉPIDÉMIOLOGIE DES MALADIES INFECTIEUSES
Institut Pasteur de Dakar
36, Avenue Pasteur, Dakar

Et à l'EHESP
ECOLE DES HAUTES ETUDES EN SANTE PUBLIQUE
Rennes-Sorbonne Paris Cité - Avenue du Professeur Léon-Bernard - CS 74312 - 35043 Rennes cedex

Abstract

Long term longitudinal surveys have the advantage to enable several sampling of the studied phenomena and then, with the repeated measures obtained, find a confirmed tendency. However, these long term surveys generate large epidemiological datasets including more sources of noise than normal datasets (e.g. one single measure per observation unit) and potential correlation in the measured values. Here, we studied data from a long-term epidemiological and genetic survey of malaria disease in two family-based cohorts in Senegal, followed for 19 years (1990–2008) in Dielmo and for 16 years (1993–2008) in Ndiop. The main objectives of this work were to take into account familial relationships, repeated measures as well as effect of covariates to measure both environmental and host genetic (heritability) impacts on the outcome of infection with the malaria parasite *Plasmodium falciparum*, and then use findings from such analyses for linkage and association studies. The outcome of interest was the occurrence of a *P. falciparum* malaria attack during each trimester (*PFA*). The two villages were studied independently; epidemiological analyses, estimation of heritability and individual effects were then performed in each village separately. Linkage and association analyses used family-based methods (based on the original Transmission Disequilibrium Test) known to be immune from population stratification problems. Then to increase sample size for linkage and association analyses, data from the two villages were used together.

We adopted several different approaches to find main risk factors associated with the occurrence of *PFA*. The main risk factors found by all used methods in both cohorts were the age of the individual and the period of survey, the most commonly known variables influencing the burden of malaria in endemic areas. On the one hand, two data mining methods, Classification and Regression Tree (CART) and HyperCube®, identified similar disease susceptibility groups defined by these two variables: almost 3 to 4 times more risk to develop *PFA* for individuals having young age (~1 to 5 years old in both cohorts by HyperCube®; ~1 to 5 in Dielmo and ~1 to 15 in Ndiop by CART) and being exposed during periods before the use of efficient drugs (periods before 2004, the year of change in drug treatment from *Chloroquine*, against which malaria parasites developed resistance, to a new and more efficient drug, *Fansidar* and later in 2006 artemisinin-based combination therapy). Whereas CART retained only these variables having strong predictive value via its “pruning tree” procedure in which the objective is to optimize the misclassification rate, HyperCube® also included hemoglobin type and cumulative experience of *P. malariae* infections that significantly increase the relative risk of *PFA*. On the other hand, regression analysis by

Generalized Estimating Equations (GEE) method found not only those variables with a strong contribution in defining highest risk groups, but also other important variables showing significant association with *PFA*. Thus, GEE added variables sex, season of the year, hemoglobin type, blood group, Glucose-6-phosphate dehydrogenase (G6PD), cumulative experience to infections by *P. falciparum*, *malariae* and *ovale*, and exposure.

In addition to these epidemiological factors, malaria infection and disease are strongly influenced by human host factors. To quantify these sources of variation, correlated random effects such as those due to genetic relationships among individuals and repeated measures within individuals should be taken into account in statistical models. Thus, we evaluated the heritability of malaria phenotypes known to be influenced by human genetics, the number of clinical malaria episodes or *P. falciparum* malaria attacks (*PFA*) and the proportion of these episodes being positive for gametocytes (*Pfgam*), the specific stages of the parasite responsible for parasite transmission to the mosquito. We performed Generalized Linear Mixed Models (GLMM) that account for familial relationships and repeated measures and have adjusted the models on the significant environmental variables identified in the epidemiological analysis, to estimate and separate the variance of the phenotypes among four sources: host additive genetics (heritability), intra-individual effects or permanent environmental effects including other personal effects like genetics non-additive, house and unexplained residuals. We found a significant additive genetic effect underlying *PFA* during the first drug period of study; this was lost in subsequent periods. There was no additive genetic effect for *Pfgam* analyzed in Dielmo only. By contrast, the intra-individual effect increased significantly. The complex basis to the human response to malaria parasite infection likely includes dominance/epistatic genetic effects encompassed within the intra-individual variance component. There were no house or maternal effects.

We then performed genetic studies that focus on candidate genes for susceptibility/ resistance to malaria. We used family-based methods with a multi-locus model, more powerful and better adapted, for multifactorial diseases such as malaria, to test for genetic linkage and association at any number of independent loci simultaneously. We used 45 Single Nucleotide Polymorphisms (SNPs) on candidate genes as genetic variables and the adjusted individual effects on PFA as the phenotype of interest. Simulation studies showed a gain of power from single locus to multi-locus models in detecting a genetic effect on a phenotype suspected to be influenced by several independent loci. Then, multi-locus models should be appropriate for malaria phenotypes supposed to be the results of actions from many different genes having weak marginal effects. We then applied this method to our real malaria data by analyzing the SNPs one by one in a first step and SNPs showing at least a weak significance (P-value $\leq$ 0.10) for association with the phenotype were selected in a second step for a multi-locus

model that analyzes simultaneous transmission of alleles from those SNPs. Five SNPs showed weak marginal protective effects against malaria after correction for multiple testing: three SNPs on the *SLC4A1* (AE1) gene (Band 3) located on chromosome 17 (ae1_20_21, P = 0.0005; ae1_117_118, P = 0.0598; ae1_174_187, P = 0.0995), one SNP on the γ-globin gene (Xmn1) located on chromosome 11 (Xmn1, P = 0.0598) and one other on the gene *ABO* located on chromosome 9 (abo297, P = 0.0854). We then analyzed these five loci together and obtained more significant protective effects (P-values were distributed from 10^{-2} to 10^{-8} for joint effects corresponding to different ways of combining these five loci).

Key words: Malaria, Repeated measures, Family based, Genetics, Heritability, Multi-locus, Linkage, Association.

Résumé

Les études longitudinales sur une longue période permettent d'échantillonner plusieurs fois le phénomène étudié et ainsi, avec des mesures répétées, dégager une tendance confirmée. Mais, dès lors, elles produisent de très larges bases de données épidémiologiques accompagnées de plus de sources de bruit par rapport aux études à observation unique ; et souvent, contiennent de la corrélation dans les mesures. Ici, nous avons présenté à travers cette thèse une étude de long terme des facteurs épidémiologiques et génétiques du paludisme menée dans deux cohortes familiales du Sénégal, l'une dans le village de Dielmo suivi pendant 19 années consécutives (1990 – 2008) et l'autre dans le village de Ndiop suivi pendant 16 années consécutives (1993 – 2008). L'objectif de ce travail de thèse a été de développer des méthodes d'analyse statistique pour identifier des gênes de susceptibilité / résistance au paludisme prenant en compte les relations familiales, les mesures répétées et des potentielles interactions génotypes – environnement dans l'évaluation des phénotypes. Par la suite, de tels phénotypes corrigés des facteurs identifiés comme potentielles sources de confusion et/ou de bruit ont été alors utilisés pour les tests de liaison et d'association génétique. Le phénotype principal étudié chez chaque volontaire a été la survenue ou non d'accès palustre, attribué à une infection au parasite *Plasmodium falciparum*, durant chaque trimestre de présence (*PFA*). Les études ont été menées de manière indépendante dans chacun des deux villages, de même que les analyses descriptives, l'estimation de la contribution génétique humaine et des effets individuels. Les tests de liaison et d'association génétique ont été réalisés par des méthodes familiales basées sur l'analyse de la transmission d'allèles des parents aux enfants (Transmission Disequilibrium Test). Ces méthodes sont connues pour être robustes par rapport au problème de la stratification de population et donc nous permettent d'augmenter la taille de notre échantillon dans les études de liaison et d'association génétique en analysant les deux villages en même temps.

Différentes approches ont été adoptées pour l'identification des facteurs épidémiologiques liés à la survenue d'accès palustres. L'âge et les années de suivi ont été les principaux facteurs liés au risque de faire un accès palustre, identifiés par toutes les approches et dans les deux villages. Ces deux variables sont connues pour être déterminant dans l'incidence des épisodes en zone d'endémie. D'une part, les méthodes exploratoires (data mining) à savoir CART (Classification and Regression Tree) et HyperCube®, ont identifié des groupes semblables de susceptibilité au paludisme se basant sur les variables âge et année : le risque relatif de faire un accès palustre est 3 à 4 fois plus élevé chez les jeunes enfants (~1 à 5 ans à Dielmo comme à Ndiop selon les résultats de HyperCube® ; ~1 à 5 ans à Dielmo et ~1 à 15 ans à Ndiop selon

CART) et durant les années avant l'introduction de traitements plus efficaces (i.e. la période avant 2004, année de changement de la chloroquine contre lequel les parasites avaient développé une résistance à un médicament plus efficace, le *Fansidar* et plus tard en 2006 les combinaisons à base d'artémisinine, ACT). CART a choisi un arbre de décision final par validation croisée en optimisant à chaque fois l'erreur de reclassement. Par conséquent CART a gardé dans ces arbres finaux que les variables âge et année qui ont une haute valeur prédictive pour le paludisme, en général quelle que soit l'origine des données étudiées. Cependant, HyperCube® recherchait le facteur ou la combinaison de facteurs qui maximiserait le risque de développer un *PFA* et par conséquent a permis d'identifier en plus de ces deux variables le type d'hémoglobine et le nombre d'infections à *P. malariae* expérimenté auparavant, qui ajoutaient des risques supplémentaires. D'autre part, la régression par GEE (Generalized Estimating Equations) a également identifié âge et année aussi bien que toutes les autres variables associées à la survenue ou non de *PFA* au seuil qu'on s'est fixé ($\alpha = 0.05$). De ce fait les modèles GEE ont ajouté les variables sexe, saison de l'année, type d'hémoglobine, groupe sanguin, Glucose-6-phosphate dehydrogenase (G6PD), durée de présence dans le trimestre et les nombres d'infections à *P. malariae* et *P. ovale* expérimentés auparavant.

En plus des facteurs épidémiologiques déterminant dans les infections et accès palustres, les facteurs génétiques humains ont aussi une influence très importante, surtout dans le devenir d'une infection. Pour évaluer proprement la part des facteurs génétiques et non génétiques, la corrélation des effets individuels, due aux forts liens de parenté entre les personnes suivies, et les corrélations dans les mesures répétées doivent être prises en compte dans les modèles statistiques. L'étape suivante de notre étude a été l'évaluation de la contribution génétique humaine dans les phénotypes comme le nombre d'accès palustres par trimestre et la proportion de ces accès positive aux gamétocytes, la forme transmissible du parasite. Nous avons donc adapté le modèle mixte linéaire généralise (GLMM) pour tenir compte des liens de parenté et des facteurs épidémiologiques et avons évalué la part de chacune de ces quatre sources de variabilité des phénotypes : les effets génétiques additifs (héritabilité), les effets intra-individus contenant les autres effets individuels tels que génétiques non additifs, les effets maison et le résiduel non expliqué. Nous avons trouvé des effets génétiques additifs durant les premières années de suivi (pendant le traitement à la quinine et à la chloroquine) qui, par la suite, ont été réalloués aux effets intra-individus. En effet, la composante polygénique de la réponse aux infections palustres chez l'homme comprend des effets génétiques additifs, mais aussi d'autres effets génétiques non additifs, tels que des effets de dominance/espitasis, qui sont compris dans les effets intra-individus. Aucun effet maison ou encore maternel était significatif.

Nous nous sommes alors intéressés aux gènes candidats pour la dernière partie de cette thèse en essayant de tester lesquels seraient potentiellement impliqués dans la susceptibilité/ résistance au paludisme. Nous avons proposé une méthode basée sur la famille, avec un model multi-locus plus puissant et mieux adapté au contexte de maladie multifactorielle telle que le paludisme, pour tester la liaison et l'association à plusieurs gènes conjointement. Nous disposions de 45 SNPs candidats comme variables génétiques et de l'ensemble des effets individuels ajustés sur les facteurs épidémiologiques comme phénotype. Les études de simulation ont confirmé le gain de puissance avec notre approche multi-locus par rapport à une approche simple locus, quand le phénotype pouvait être influencé par plusieurs gènes en même temps. Le model multi-locus serait alors adéquat pour les phénotypes du paludisme qui sont supposés être les résultantes d'actions de plusieurs gènes à modestes effets marginaux. Nous avons donc analysé les 45 SNPs un par un dans une première étape et ceux qui étaient significatifs au seuil d'erreur de 0.10 ont été sélectionnés dans une deuxième étape pour les modèles multi-locus. A la première étape, 5 SNPs ont été significatifs au seuil de 0.10 après corrections aux multiple tests : 3 SNPs sur le gène *SLC4A1* (AE1), Band 3, situé sur le chromosome 17 (ae1_20_21, P = 0.0005; ae1_117_118, P = 0.0598; ae1_174_187, P = 0.0995), 1 SNP sur le gène γ-globin (Xmn1) situé sur le chromosome 11 (Xmn1, P = 0.0598) et un autre sur le gène *ABO* situé sur le chromosome 9 (abo297, P = 0.0854). A la deuxième étape, ces 5 SNPs ont alors été analysés conjointement et leurs effets protecteurs conjoints ont été beaucoup plus significatifs (P-values distribuées entre 10^{-2} to 10^{-8} pour les effets conjoints correspondant à différentes façons de les combiner).

Mots clés: Malaria, Repeated measures, Family based, Genetics, Heritability, Multi-locus, Linkage, Association.

Remerciements

Je tiens à remercier Avner Bar-Hen mon directeur de thèse pour son encadrement exemplaire tout au long de ces trois années. Son esprit scientifique et le "concrètement" dans sa façon de travailler m'ont appris le raisonnement clef d'un chercheur. Je le remercie de m'avoir fait travailler sans stress et de m'avoir fait découvrir quelques bons plans resto sur Paris.

Je témoigne de ma reconnaissance à Aliou Diop, mon co-directeur de thèse, pour la confiance qu'il a toujours manifestée à mon égard depuis le DEA et pour m'avoir fait découvrir l'Institut Pasteur. Je le remercie pour ces idées pertinentes et son soutien.

Je souhaite remercier et également faire part de ma reconnaissance à Avanaj Sakuntabhai et Richard Paul pour m'avoir accueilli au sein de leur unité à l'Institut Pasteur. Je les remercie pour leur disponibilité et pour avoir fait plus que superviser mes travaux, notamment la recherche de financement, leurs conseils et leur soutien pour la rédaction de cette thèse. Je remercie également Jean François Bureau pour ses bons conseils et son dynamisme dans les travaux qu'on a menés ensemble, travailler avec lui m'a été très bénéfique. Je remercie aussi Christian Roussilhon pour remarques constructives.

Je remercie Laurence Baril pour son soutien, ses encouragements dès le début de ma thèse et également de m'avoir permis de développer des collaborations scientifiques avec Paris.

Je remercie Christophe Rogier, Margaret Mackinnon et Alioune Dièye d'avoir bien voulu accepter de faire partie de mon jury de thèse.

Je tiens à remercier mes collègues de bureau, Isabelle Casadémont, Laura Grange, Olivier Telle, Alison Machado, Sumonmal Uttayamaku, pour leurs remarques constructives et pour la bonne ambiance de travail qu'ils ont apportée.

Je remercie toute l'équipe de l'Unité d'Epidémiologie des Maladies Infectieuses de l'Institut Pasteur de Dakar, en particulier Adama Tall pour la pertinence scientifique de son point de vue dans mes travaux et pour ses conseils, Fatoumata Diène Sarr, Fatou Bintou Badji et Marie Louise Senghor, Joseph Faye et Baba Diakhaby, Ndjido Ardo Bar, Khadijetou Diop, Abdoulaye Badiane et l'équipe de terrain travaillant dans les stations de recherche de Dielmo et de Ndiop ; leur rigueur a permis de fournir des données de qualité et de partir ainsi sur des problématiques scientifiquement intéressantes.

Je remercie toute l'équipe de Jean-François Trape de l'URMITE à l'IRD Dakar pour leur excellente collaboration. Je remercie également toutes les autres équipes ayant travaillé sur ce

projet, particulièrement les unités d'Immunologie et d'Entomologie de l'Institut Pasteur de Dakar.

Je témoigne de mes remerciements à l'ensemble des membres du Laboratoire de Mathématiques appliquées Paris 5 (MAP5), en particulier Marie Hélène Gbaguidi pour son aide précieux et sa sympathie, ceux qui sont encore présents comme ceux qui sont déjà partis, Nicolas Capian, Chirstophe Denis, Mahendra Mariadassou, Anne Cecile Dragon, Adriana Gogonel, Djénaba Thiam, Imen Hammami, Wilson Tousil, Abdul Razzaq (LIPADE), pour leur très grande sympathie.

Je remercie toute l'équipe de EffiScience Research pour leur excellente collaboration, en particulier Augustin Huret, Nicolas Levilain et François d'Ormesson.

Je remercie les membres des écoles doctorales ED-420 et ED-SEV ainsi que le réseau doctoral de l'Ecole des Hautes Etudes en Santé Publique (EHESP), plus particulièrement le personnel administratif.

Je témoigne de ma reconnaissance à la Direction Internationale de l'Institut Pasteur et au réseau doctoral de l'EHESP d'avoir bien voulu financer ma thèse ; je suis également reconnaissant vis-à-vis du SCAC (Service de Coopération et d'Action Culturelle) de l'ambassade de France à Dakar et de la STAFAV (Statistique pour l'Afrique Francophone et Applications au Vivant) d'avoir participé au financement de ma thèse.

Je tiens à remercier chaleureusement l'UCAD et sa direction de la coopération, ainsi que la direction de l'ED-SEV, qui ont bien voulu présenter ma demande de bourse au SCAC.

Je tiens à remercier ma famille de Paris, les Diop, grand Abdoulaye Diop, Christelle qui m'a fait visiter les musées de Paris pour la premiere fois, et la petite Ava ; merci pour les belles photos prises lors de la soutenance.

Je remercie mes amis Ibrahima, Magali et Vincent, Sabah, Julie et Mac, Nimesh et Karnika, Marième, Charlie, Lamine - Astou et les filles, les deux Serignes (celui de Mously et celui de Binette), Ousmane et Seynabou, Bruce - Laura et les enfants, Giles - Jess et les garçons, Fatou, Maguette - Babacar et petit Khass, Elhadj Omar (EON), Matar, Khassim, Amadou Moussa, Papis, William. Les longs débats autour d'un "Tiébou Dieune", d'un "Yassa" discutant de politique et de souvenirs de Dakar m'ont beaucoup aidé à surmonter la nostalgie du Sénégal. Merci également à Marie Josephe Granier (Mamie Joe) pour sa gentillesse, sa bonne humeur, ses encouragements, et surtout pour ses bons "bourikas" au fromage.

Je remercie les villageois de Dielmo et de Ndiop, en espérant que ces résultats leur soient expliqués et que l'incidence du paludisme continue de diminuer rapidement grâce aux mesures de prévention et aux nouveaux moyens de diagnostic et de traitement.

Je remercie particulièrement et dédie ce travail à mon père et à ma mère, ma femme Amy et notre fils Serigne Moustapha, mes frères et sœurs, particulièrement Ndawa, de m'avoir soutenu sur tous les plans, surtout durant les moments difficiles.

Contents

Preface

Context

One of the main goals of genetic epidemiology is to search for molecular pathways implicated in pathogenesis and in immune response. Finding and understanding these pathways can be usefull to treat diseases and to develop vaccines.

Presently, *Plasmodium* infection as well as malaria disease has been shown through several studies to be influenced by environmental factors and also by human geentic factors. Therefore, before genetic analysis using genome-wide approaches for linkage and association studies, it is of great interest to evaluate relative contribution of genetic and non-genetic to the disease phenotypes. Estimation of heritability (variability of the outcome attributable to additive genetics) based on a good knowledge of family structure is essential to estimates how much in the disease is attributable to the human genetics.

Long-term malaria phenotypes, pedigree and genotypes data exist for two cohorts in Senegal. Preliminary genetic analyses have proved informative and yet several major statistical issues have arisen that are not currently developed in the field of infectious disease research and will be a major obstacle in the future. These issues are the effect of genetic relationships (non-independence between individuals), the incorporation of repeated measures that depend on the individual (non-independence of observations within individuals) and potential gene-gene and gene-environment interactions.

Studied populations

We studied a large dataset from a long-term epidemiological and genetic survey of two sub-Saharan African family-based cohorts, followed for 19 years (1990-2008) in Dielmo and for 16 years (1993-2008) in Ndiop. Dielmo is the village with holoendemic transmission (perennial and high intensity) and Ndiop with mesoendemic transmission (seasonal and at a lower intensity compared to Dielmo). Malaria transmission intensity differs between the two villages because of the presence of a river in Dielmo (see location and maps of the study sites, Figures 1.A – C. below) that offers a mosquito breeding site all-year round. These sites are managed by a tripartite agreement between the Institut Pasteur de Dakar (IPD), the Institut de Recherche pour le Développement (IRD) and the Ministry of Health and Prevention of Senegal. A field research station, with a dispensary run by nurses and paramedical personnel, was built for the program in each village and is open 24 hours a day, 7 days a week.

Therefore, almost all fever episodes had been reported to the clinics with blood smears checked for malaria parasites. The health care is free-of-charge for the volunteers. Every person satisfying adhesion conditions could become a volunteer and every volunteer could leave the study at any time, therefore forming a dynamic open cohort. Further details of the study sites and adhesion criteria are previously described (Trape, Rogier et al. 1994; Rogier, Tall et al. 1999).

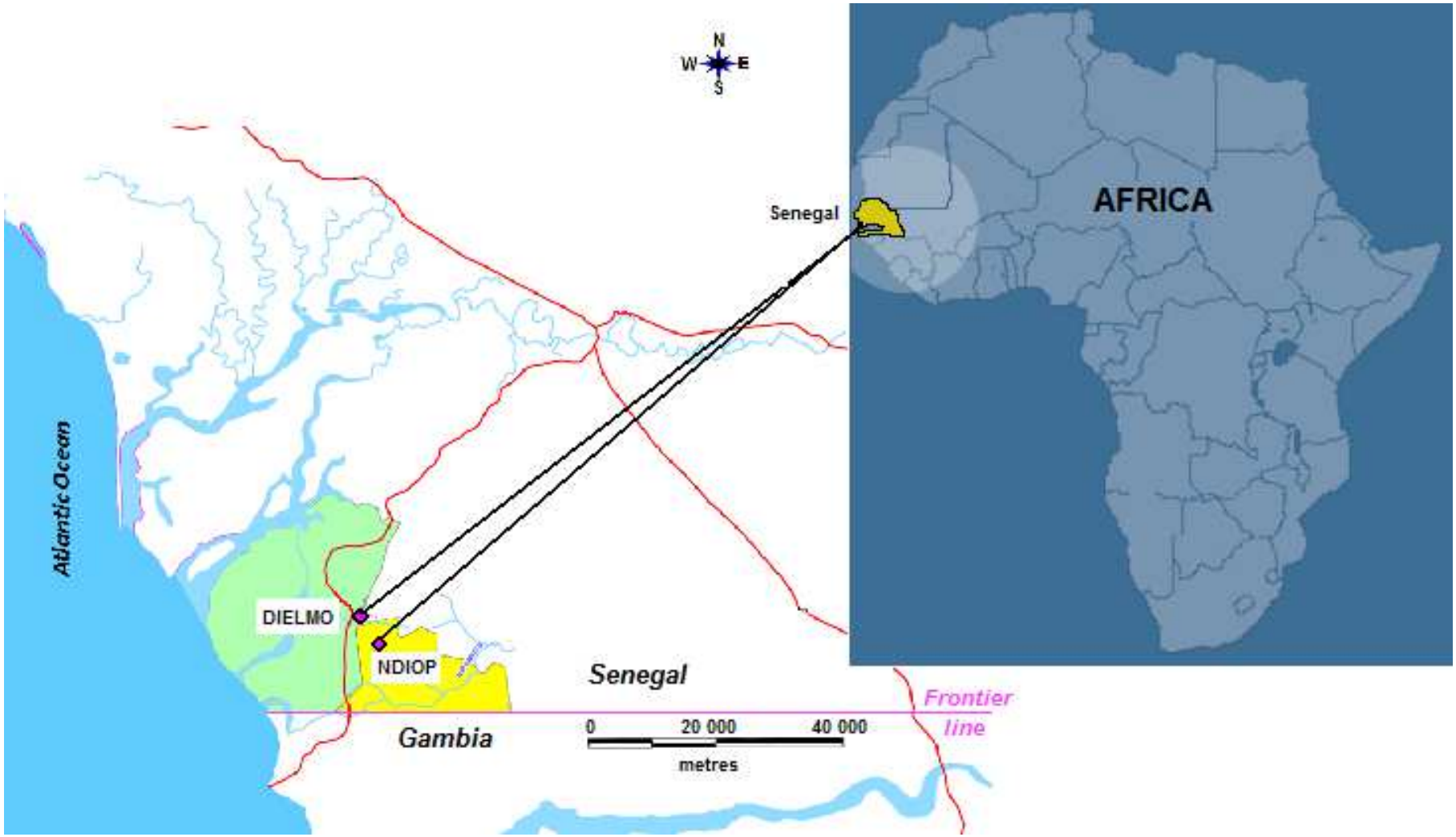


FIG. 1.A. Geographical location of the study sites (Dielmo and Ndiop villages).

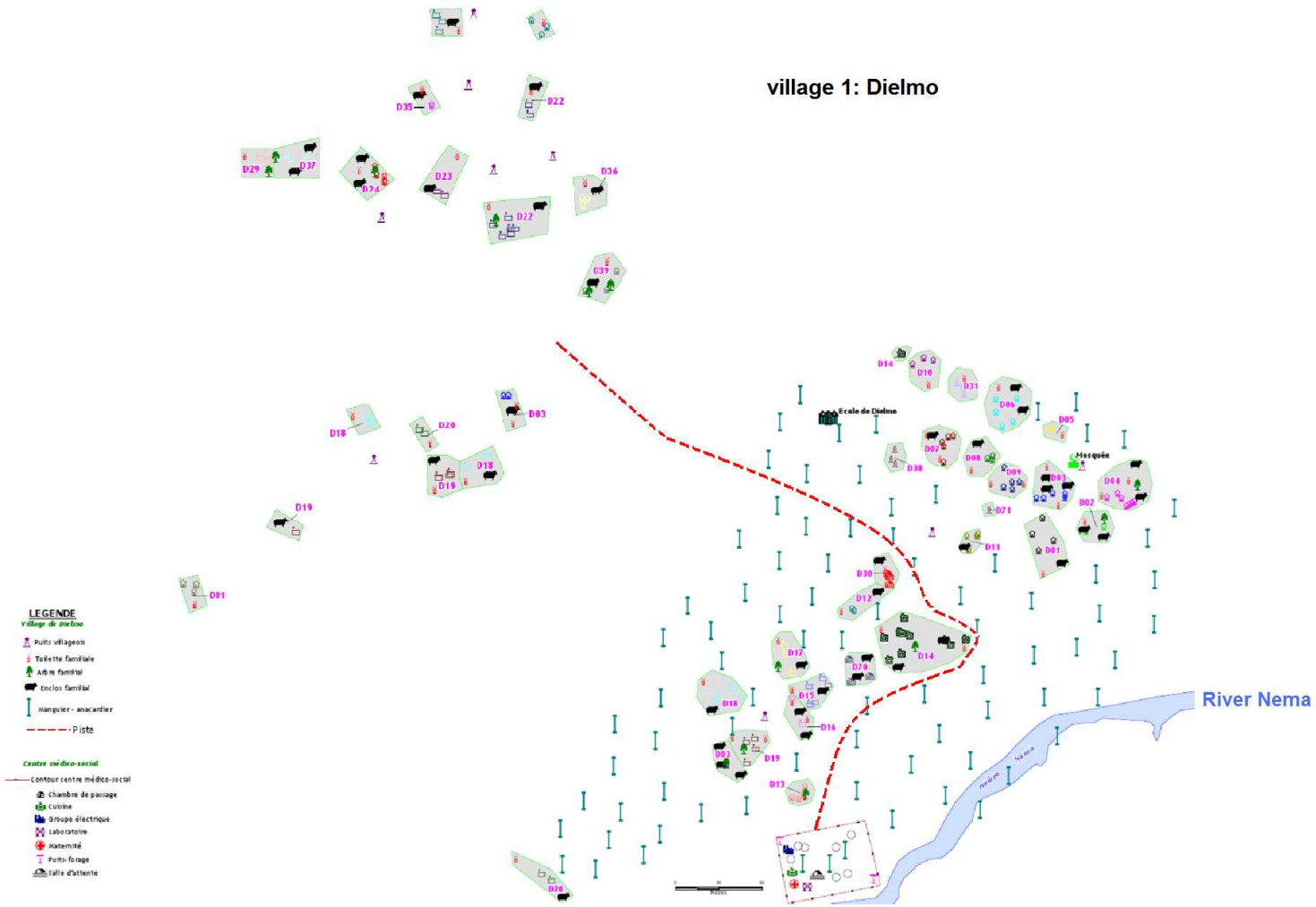


FIG. 1.B. Map of the village of Dielmo.

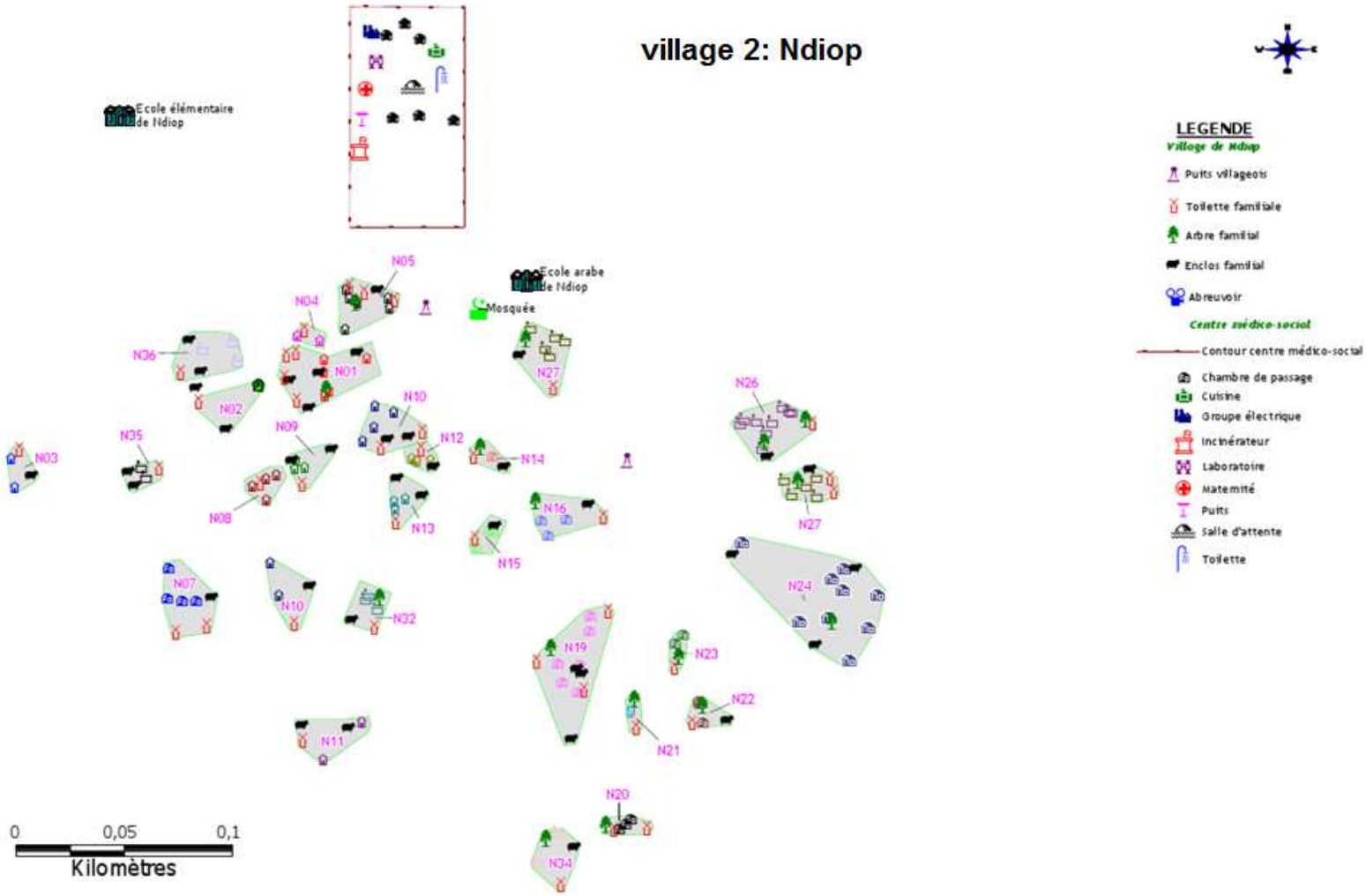


FIG. 1.C. Map of the village of Ndiop.

Objective

The aim of the thesis was to develop and apply appropriate statistical analyses to identify key factors contributing to *Plasmodium falciparum* malaria phenotypes in two long-term family-based longitudinal data sets in Senegal. The challenge was to implement rigorous statistical methods that can take into account familial relationships, repeated measures as well as the effect of covariates, to generate heritability values for specific phenotypes and to then perform linkage and association studies of specific candidate genes in single and multi-locus models using the residual malaria phenotype. The final goal was to obtain fine measures of both environmental and host genetic impacts on malaria phenotypes within a population context of related individuals.

This subject and the design of the study provide novel challenges in statistical modeling, especially in the research field of statistical genetics. Inferences from statistical models assuming basic sample designs with independence among observations or absence of interactions among variables have been more generally addressed. Here, it is not the case with a longitudinal study where the repeated measures of a same individual are not independent, and thus require application of generalized models as Generalized Estimation Equations (GEE) or Mixed Models. Also, it is challenging for statistical genetics methods that use familial relationships when testing for genetic effects underlying diseases. Here, the outcome of *Plasmodium* infections (the phenotype) varies within the same individual from one observation to another depending on many factors, intrinsic (like host genetics) as well as extrinsic (like environment). Then for malaria, the ways to find the most likely category for the disease status (susceptible or resistant) of an individual with such variation on the phenotype always need research efforts in statistical methods. Most of the methods previously developed to test for genetic effects have been designed for Mendelian diseases and not directly applicable for complex infectious diseases.

Thus, this motivates us to do this thesis for the study of human genetics and environmental aspects underlying malaria disease by focusing on statistical methods adequate for such a multifactorial disease.

Plan of the thesis

The key environmental factors determining the outcome of infection with the malaria parasites, *Plasmodium falciparum*, will be evaluated by analyses of family-based longitudinal survey. The overall human additive genetic contribution (i.e. heritability) to malaria

phenotypes will be estimated and the role of candidate genes assessed. For this, our study will be presented in five chapters.

The first part of Chapter 1 “General Introduction” presents malaria disease and the last part presents the main statistical issues in the analysis of epidemiological and genetic data from malaria survey.

Chapter 2 “Descriptive Methods” is the epidemiological analysis part preceding genetic analyses. The methods section of this chapter will start by reviewing some data mining methods usually performed to handle very large datasets and, then, the new HyperCube® approach combining regression and optimization techniques will be presented. Another part of this section will present regression method by Generalized Estimating Equations (GEE) to find significant population effects influencing the burden of the disease. Results and discussion sections gives the application of these methods to the two studied cohorts.

Chapter 3 “Heritability” begins the genetic study part as a first step and presents a method to estimate the overall genetic contribution to malaria disease. A first section presents the methods used to calculate kinship between relative pairs of individuals in the population. Methods of inference of the genetic relatedness among individuals in a population are explained in detail. A second section presents the use of Mixed Models to estimated heritability (additive genetic contribution) via variance components analysis; and simultaneously, association analysis in a valid case-control like design from family data by incorporating the kinship information. Result and discussion sections give the applications of these methods to the two studied cohorts.

Chapter 4 “Linkage and Association Analysis” is the second part of the genetic study and presents family based linkage and association tests using allelic transmission count based on the Transmission Disequilibrium Test (TDT). A first part of the methods section presents some useful definitions in genetics and in multiple testing contexts that will be discussed frequently through this chapter. Using the multinomial distribution, the second part presents the likelihood version of the TDT to test for linkage and association between phenotypes and each of the considered loci in a single-locus approach. A third part shows how to generalize TDT in a Multi-locus and Multi-allelic Approach to test disequilibrium in the simultaneous transmission of alleles from multiple unlinked loci, extending the method proposed by Andrew Morris and John Whittaker for two loci (Morris and Whittaker 1999). This method is powerful to find multiplicative or epistatic effects between several independent genes having weak marginal effects.

Chapter 5 “General Conclusion” summarizes all findings and provides some research perspectives in the field of statistical genetics of multifactorial diseases.

In the annex, some basic notions of metric, e.g. Euclidean and Mahalanobis distances and the influence of their choice when measuring similarities/dissimilarities between observations, are

presented as a preliminary to the “Descriptive Methods” chapter, for interested readers. Next, the R scripts used to simulate data and to analyze our real data are provided. Lastly, the publications related to the thesis (and the list for other publications) are presented.

1. General Introduction

Statistical analysis in malaria genetic epidemiology has always been a challenge due to the fact that the disease phenotypes are difficult to define and are influenced by several known sources such as host genetics, individual's immune state, parasite genetics, environmental factors and their interactions. Obtaining reliable conclusions on factors underlying the outcome of malaria infections needs robust study designs like family-based longitudinal survey to distinguish between the parts of each source of variation.

Malaria infection and disease are strongly influenced by human host and environmental factors and may vary considerably in their severity and clinical manifestations. Previous studies indicated an important contribution of host genetics to the outcome of malaria disease (Mackinnon, Mwangi et al. 2005). Some known genetic and biological markers, most especially those linked to the host immune response, have been implicated in the frequency and severity of malaria disease (Phimpraphi, Paul et al. 2008; Sakuntabhai, Ndiaye et al. 2008). Before going on statistical analysis, some aspects of malaria disease are presented here.

1.1. Presentation of malaria disease

Malaria is a multifactorial infectious disease that has affected human populations since the beginning of mankind and is still the major parasite disease affecting and killing humans. It also affects animals, including monkeys, rodents, birds, and reptiles. Malaria is caused by parasites of the genus *Plasmodium* belonging to the apicomplexan phylum, which invade and reproduce in erythrocytes. Hematophagous mosquitoes of the genus *Anopheles* are required for the transmission of the parasite from one human host to another. The four most prevalent *Plasmodium* species implicated in human malaria are: *Plasmodium falciparum* (the most virulent, more frequent in Africa), *P. malariae*, *P. ovale* and *P. vivax* (absent in sub-Saharan Africa, more frequent in Asia and Southern America). Among the three species present in Africa *P. falciparum* is the most prevalent and is responsible for most morbidity and mortality. The main aspects of malaria can be summarized in three points: (i) the mechanism of transmission through the parasite life cycle between host and vector, (ii) the clinical symptoms, showing illness, that depend on a specific stage of this life cycle and (iii) the burden of morbidity and mortality. The high prevalence of malaria in developing countries underlines the extent to which it represents a public health challenge.

The parasite life cycle

The parasite needs two hosts to complete its life cycle: a mosquito vector and a vertebrate host (in our study a human host).

A – In the human: The female *Anopheles* mosquitoes whilst taking a blood meal, injects the malaria parasites in the form of sporozoites (1). The sporozoites migrate to the liver, invade hepatocytes and multiply. These hepatic merozoites (2) are then liberated into the bloodstream, and invade red blood cells, starting the asexual proliferation cycle, grow into trophozoites and for the most part undergo asexual replication to form a schizont (3). This schizont contains many merozoites that rupture the red blood cell and then seek to invade new red blood cells; this asexual cycle of the parasite is responsible for illness. A small fraction of the merozoites develop into sexual stages of the parasite, namely gametocytes (4); the sexual form is necessary for transmission of the parasite to the mosquito. Gametocytes, or gamete pre-cursors, are either male or female.

B – In the mosquito: Once ingested by mosquitoes, a female gametocyte forms 1 female macrogamete (5.f) and a male gametocyte forms up to 8 male microgametes (5.m). Zygotes (6) are formed by the fusion of gametes (5.f and 5.m). Zygotes become ookinetes (7) that infiltrate the midgut wall and form oocysts (8). These oocysts expand over time and finally release sporozoites (1) after 10-14 days. The sporozoites move into the mosquito salivary gland, making the mosquito infectious for humans during her next blood meal. Figure 1.1 below from Teun Bousema and Chris Drakeley (Bousema and Drakeley 2011) shows the life cycle of the *P. falciparum* parasite between human host and mosquito.

For researchers, an appreciation of this life cycle is necessary to focus on specific stages when developing drugs for treatment or insecticides, vaccine as well as eradication policies.

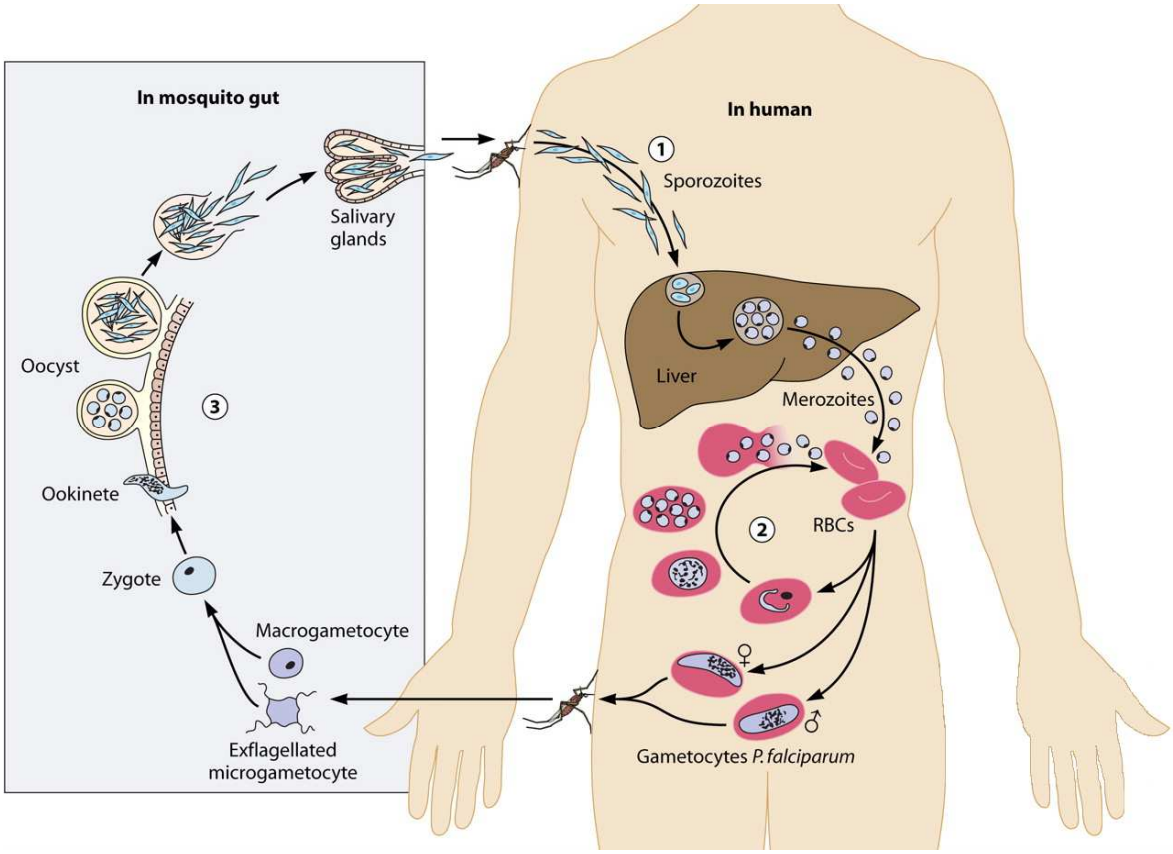


FIG. 1.1. Life cycle of *Plasmodium falciparum* (Source: Bousema and Drakeley, 2011).

Clinical symptoms

Several clinical manifestations can occur from different infections or a same infection by malaria parasites depending on host genetic, parasite genetic, environment and their interactions. Symptoms would include fever, chills when fever is high, sweating, headaches, cough, abdominal pain, diarrhoea, nausea, vomiting, enlarged liver and spleen (sometimes not palpable), loss of appetite, orthostatic hypotension, myalgia (limbs and back), asthenia, etc., that overlap with many other diseases. Clinicians are therefore faced with the challenge of correct diagnosis in an environment where the fraction of fevers (or other malaria symptoms) attributable to malaria will alter. Determining the specific cause of a clinical episode during co-infections with other diseases needs reliable accurate methods of diagnosis. However, children with advance illness, often present for several clinical symptoms that can be due to several different diseases (English, Berkley et al. 2003). In malaria endemic areas, several clinical manifestations due to *Plasmodium* infection occur and overlap with those of many other disease (Kallander, Nsungwa-Sabiiti et al. 2004). Indeed, malaria is so difficult a disease to diagnose by clinical examination alone, that algorithms are not considered useful (Mwangi, Mohammed et al. 2005) and lead to over-diagnosis of malaria (Amexo, Tolhurst et al. 2004; Reyburn, Mbatia et al. 2004). Although the use of rapid diagnostic tests (RDTs) has the potential improve malaria differential diagnosis (Bell, Wongsrichanalai et al. 2006), asymptomatic parasite prevalence can be very high in areas endemic for malaria, leading to misdiagnosis and failure to treat the pathogen responsible for the episode in question.

Prevalence

Plasmodium falciparum is the most common plasmodial parasite invading humans. Malaria is endemic in 108 countries in 2010 making about 3.3 billion people (half of the world population) at risk of infection as shown in Figure 1.2 from the World Health Organization (WHO). World malaria Report for the year 2009 estimated malaria to cause about half a billion episodes per year and is responsible for over 800,000 deaths per year (WHO 2009). Children under 5 years old are the major “at-risk” group for malaria morbidity and mortality. Malaria represents a serious public health problem in Africa, where one in every five (20%) childhood deaths is due to the effects of the disease. The main factors maintaining the disease highly prevalent in Africa are: the propitious climatic conditions, the existence of the vector *Anopheles gambiae*, the socio-economic conditions, the development of resistance to most anti-malarial drugs and the lack of a vaccine.

The situation in Senegal according to the "World Malaria Report 2010" (website: "www.who.int/malaria/world_malaria_report_2010/en"): Throughout Senegal where we performed this study, malaria is endemic with seasonal transmission occurring from June to November; and almost all cases are caused by *P. falciparum*. Inpatient malaria cases and deaths declined markedly between 2007 and 2008 and again in 2009. During the transmission season, 100% of the population is at risk of infection according to the 2010's WHO report, with heterogeneity in the distribution as shown by Figure 1.3 below. The national malaria control program delivered 4.5 million long-lasting insecticidal-treated nets (LLINs) during 2007–2009 covering 73% of the population at risk, and over 661 000 people (5% of the population at risk) were protected with indoor residual spraying (IRS). In the post-campaign national survey in 2009, 82% of households had an insecticide-treated mosquito net (ITN). The program delivered about 320 000 artemisinin-based combination therapy (ACT) treatment courses in 2008 and 184 170 in 2009, sufficient to treat about half the reported malaria cases (probable + confirmed cases) in the public sector.

In order to control malaria, tropical countries such as Senegal have scaled up their intervention strategies combining prevention, via implementation of LLINs, with improved diagnostic techniques (rapid diagnostic tests - RDT) and the introduction of an efficacious treatment using ACT. In addition, intermittent preventive therapy is implemented in specific groups such as the pregnant women. ACTs present several advantages: i) high efficacy and no naturally occurring resistance reported in sub-Saharan Africa; ii) effectiveness against sexual stage parasites (gametocytes) with the potential to reduce parasite transmission (Okell, Drakeley et al. 2008); iii) effective reduction of the asexual parasite population (Adjuik, Babiker et al. 2004; Nosten and White 2007). Thus, ACTs are expected to reduce overall malaria transmission and to impede parasite resistance to the drug combined with the artemisinin derivative (amodiaquine in Senegal).

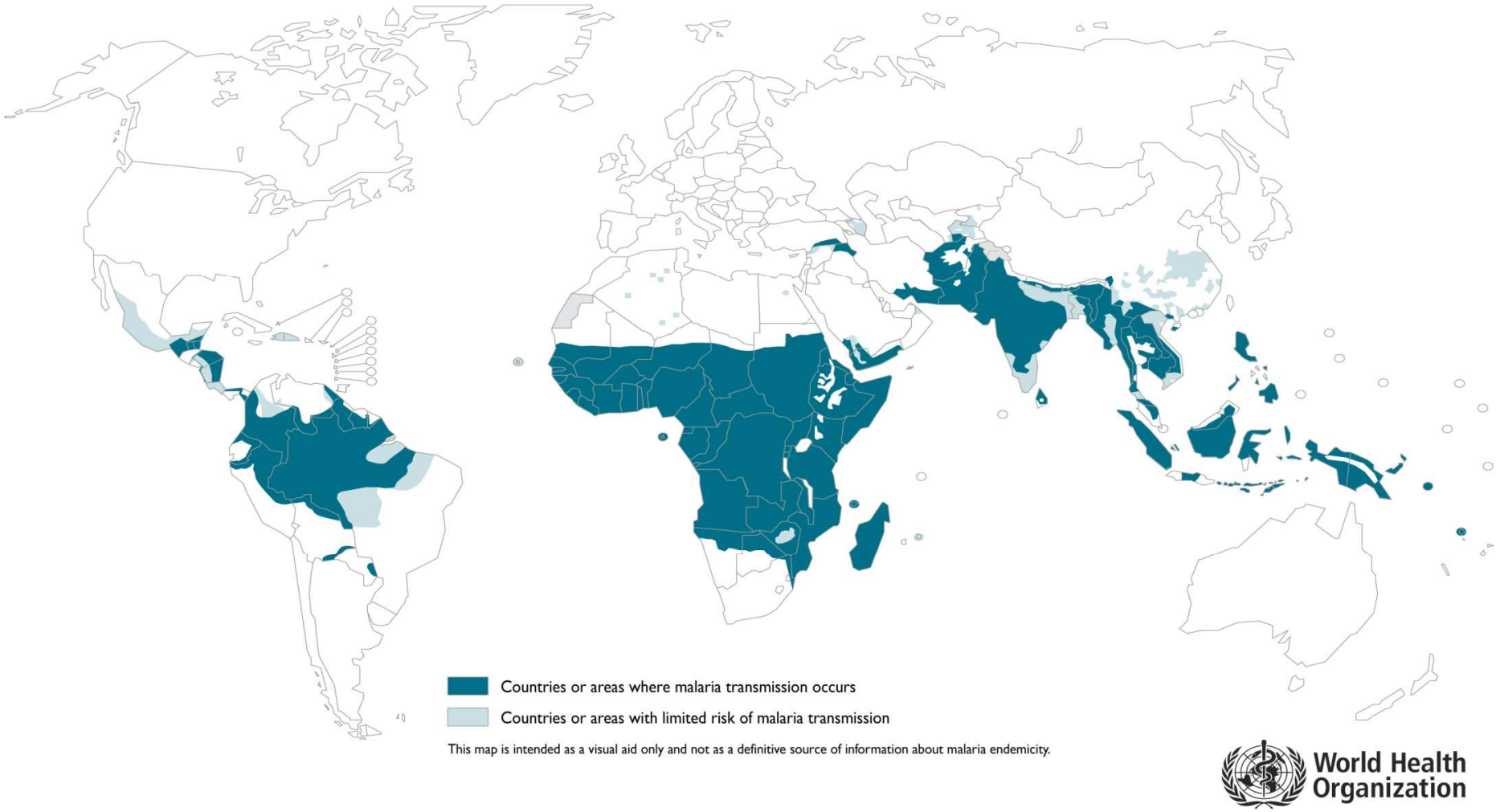


FIG. 1.2. Malaria, countries or areas at risk of transmission in 2010 (source: WHO, 2011).

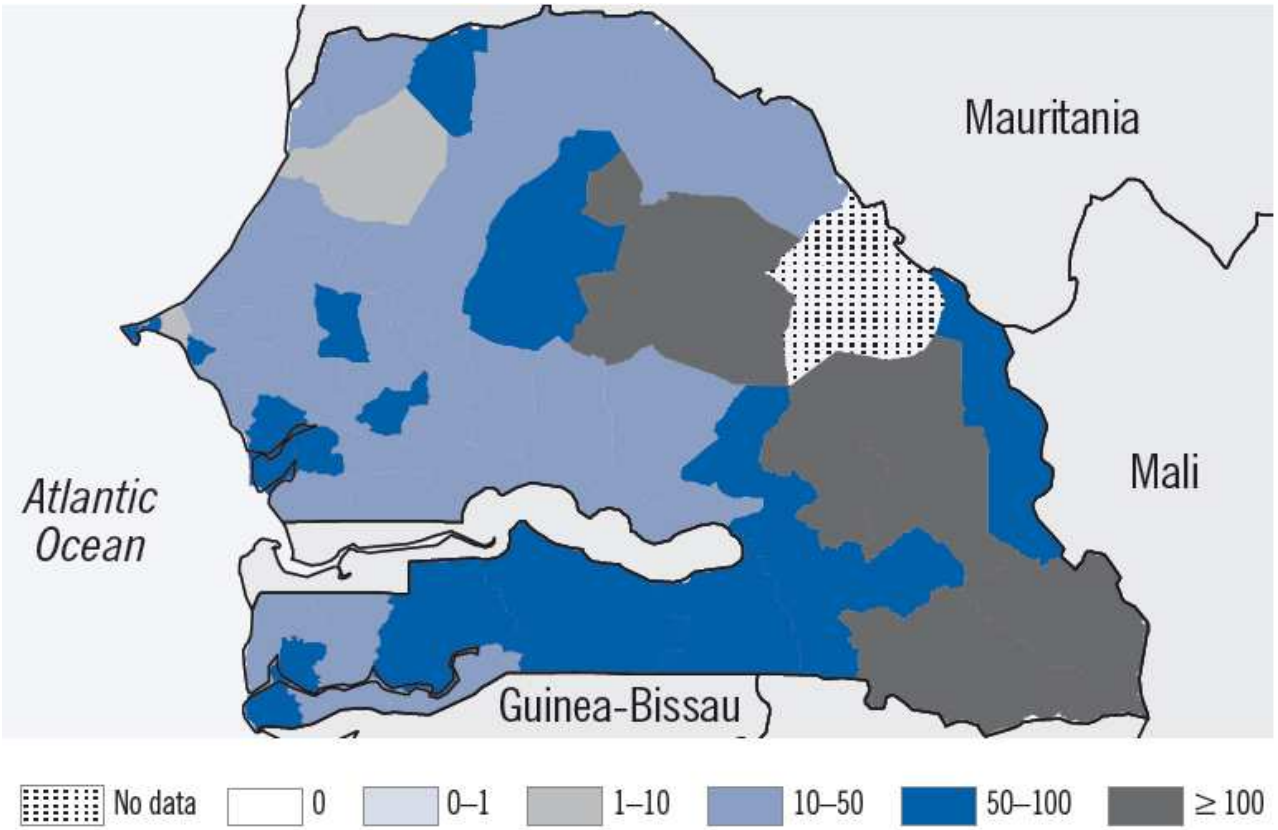


FIG. 1.3. Geographical distribution of confirmed malaria cases in Senegal, per 1000 population (Source: World Malaria Report 2010).

1.2. Genetic susceptibility to malaria

The study of the contribution of human genetics to the risk of severe malaria has a long history, with Haldane in the 1950s reporting a major role of the sickle cell mutation (HbS), in the protection against severe disease (Haldane 1949). Since then, genetic variants of β-globin: HbE (Hutagalung, Wilairatana et al. 1999), HbC (Agarwal, Guindo et al. 2000), HbS (Aidoo, Terlouw et al. 2002; Williams, Mwangi et al. 2005); α-globin (Weatherall 1997; Mockenhaupt, Ehrhardt et al. 2004; Williams, Wambua et al. 2005); Band 3 protein (AE1) (Foo, Rekhraj et al. 1992); HLA (Hill, Allsopp et al. 1991) and several cytokine loci: Tumor Necrosis Factor-alpha (McGuire, Hill et al. 1994; Wilson, Symons et al. 1997; Knight, Udalova et al. 1999), Interleukin-12 (Morahan, Boutlis et al. 2002), Interferon-alpha receptor-1 (Aucan, Walley et al. 2003), Interleukin-4 (Gyan, Goka et al. 2004) have been demonstrated to confer protection to severe malaria. To date, the majority of studies have been case/control association studies, comparing severe malaria to uncomplicated cases. However, there is still a gap of study in this genetic susceptibility field for uncomplicated malaria.

1.3. Main statistical issues for analysis of malaria data

Identifying main risk factors and their interactions in studies of multifactorial diseases always induce statistical and bioinformatics challenges. For malaria disease, there are several epidemiological, environmental, biological and genetic variables that underlie the outcome of infection and their interactions are difficult to understand.

Several statistical methods have been proposed for multivariate analysis and to test interactions among variables. Without prior hypothesis, it is almost impossible to test all possible combinations of variables in a model and all possible interactions among them. Even if a combination of variables is considered, the interaction terms to test need to be specified *a priori* in the model formula. Traditional statistical methods have limitations in dealing with this complexity, especially when large numbers of variables are analyzed simultaneously. In addition, most variables may not be distributed the way most regression methods assume. In our context of long-term study (16 and 19 years of survey) and family-based design, many variables are implicated and are different in their type. Individuals are not independent and data are correlated due to within family similarities, shared environment, as well as repeated measures on a same individual. The number of measures per individual is not the same due to self-presentation of persons making models for repeated measure more complex. The successive measures can be influenced by the different parasite species implicated in previous infections or by actions of medical staff on the latest presentation (e.g. effect of drug

administration: the efficacy and the dosage can induce total or only partial clearance of parasites and can impact on the time to the next episode). These complexities in the data induced by the method of survey and by the characteristics of the population may provide only weak or even false evaluation of epidemiological (e.g. effect of environmental variables) and genetic parameters (e.g. heritability estimates, genetic effects sizes and significance) if they are not taken into account.

To handle all these difficulties, non parametric data mining methods is increasingly used in analyzing very large epidemiological and genetic malaria datasets (Protopopoff, Van Bortel et al. 2009; Loucoubar, Paul et al. 2011) to evaluate importance of non-genetic variables that can confound genetic effects. The HyperCube® method we introduce in chapter 2 can detect all significant interactions among a large number of variables without prior hypotheses or knowledge of their existence. For more details see our published results in "An Exhaustive, Non-Euclidean, Non-Parametric Data Mining Tool for Unraveling the Complexity of Biological Systems – Novel Insights into Malaria" (Loucoubar, Paul et al. 2011). This aspect can be of great interest in analyzing genome-wide data on malaria where phenotypes are known to be the results of several genes and their epistatic effects. Thus, this method would help to identify the main chromosomal regions showing a promising signal from the hundreds thousands single nucleotide polymorphisms (SNPs) typed all along the genome. This could be advantageous to handle the multiple testing problem induced in genome-wide association studies (GWAS).

After the computational management of large datasets and the identification of relationships among variables, estimation of the overall genetic contribution, the heritability, is the next step. Estimation of the heritability of phenotypes can adopt different methods: a more general approach that estimates the human genetic contribution to the phenotype on the whole genome and a more specific approach that estimates the contribution of one specific genomic location or a set of distinct locations. Whatever the method, information on familial relationships (kinships) among studied individual is necessary to estimate the heritability. Several studies on genetic susceptibility/ resistance to malaria have first provided the overall human genetic contribution to the disease (Stirnadel, Beck et al. 1999; Mackinnon, Gunawardena et al. 2000; Mackinnon, Mwangi et al. 2005; Phimpraphi, Paul et al. 2008; Sakuntabhai, Ndiaye et al. 2008; Lawaly, Sakuntabhai et al. 2010; Loucoubar, Goncalves et al. 2011) before focusing on genes potentially responsible to the heritability signal in their studied populations.

Another challenge of great interest in family based studies is the polygenic aspect of malaria disease, with the improvements in traditional linkage and association methods they create. One should allow for hypotheses that assume a cumulative and/or interactive force of several distinct genes, each having a weak marginal effect on the outcome of malaria infections; this point of view differs from the one used in the study of monogenic diseases for which one

single gene determines the disease phenotype. As explained above in the thesis objectives, new methods for tackling polygenic infectious diseases are required.

Part I:

Epidemiological Analysis

2. Descriptive Methods

Abstract

We studied data from a longitudinal survey of malaria in two Senegalese cohorts, followed from 1990 in Dielmo and 1993 in Ndiop to 2008. The outcome of interest was the occurrence of a *Plasmodium falciparum* malaria attack during each trimester (*PFA*). Data were analyzed independently in each village and we adopted different approaches to find main risk factors associated with the occurrence of *PFA*. The risk factors identified by all used methods in both cohorts were the age of the individual and the period of survey. Data mining methods showed, relatively to the general population, almost 3 to 4 times more risk to develop *PFA* for young people (~1 to 5 years old in both villages by HyperCube®; ~1 to 5 in Dielmo and ~1 to 15 in Ndiop by CART) and exposed during periods before the use of efficient drugs (i.e. before 2004, the year of change from *Chloroquine*, for which malaria parasites developed resistance, to a more efficient drug treatment, *Fansidar*). Whereas CART retained only these variables having strong predictive value via its "pruning tree" procedure, in which the objective is to optimize the misclassification rate, HyperCube® also included hemoglobin type and cumulative experience of *P. malariae* infections that significantly increase the relative risk of *PFA*. Analysis by Generalized Estimating Equations (GEE) method found not only those variables with a strong contribution in defining highest risk groups, but also other important variables showing significant association with *PFA*. Thus, GEE added variables sex, season of the year, hemoglobin type, blood group, Glucose-6-phosphate dehydrogenase (G6PD), cumulative experience to infections by *P. falciparum, malariae* and *ovale*, and exposure.

2.1 Introduction

Before genetic study for the identification of resistance/susceptibility genes to a given disease, one should start by trying to understand the epidemiology of the disease in the studied population, and then, perform the adequate method of analysis for further genetic investigations. First investigations should identify the existing links and influences among considered variables. This step is all the more important in the case of multifactorial diseases, like malaria, where confusions can occur because of the fact that observed or measured phenotypes are simultaneously influenced by different factors, environmental, human non-genetic and genetic. This work will be done by data mining and also by regression methods handling repeated measures.

Additional difficulties arise in populations living in highly endemic areas where people can tolerate the parasite in the blood at a certain level because of the development of clinical immunity (Rogier, Commenges et al. 1996). A major challenge is to determine the fraction of clinical manifestations attributable to *P. falciparum* malaria. Phenotype definition is therefore primordial and the impact of non-genetic factors on any defined phenotype for malaria (and several other multifactorial diseases) needs to be disentangled prior to genetic analysis of resistance/susceptibility.

The first subsection presents several data mining methods used to identify relationships among variables in a dataset. We start by reviewing some methods usually performed to handle large datasets, i.e. large number of variables and large sample size. Subsequently, we will compare them with a new exhaustive, non-Euclidean and non-parametric approach combining regression and optimization techniques, dealing with hypercube forms in a multi-dimensional space (Loucoubar, Paul et al. 2011). As data mining methods are not always appropriate for repeated measure designs, a second subsection presents the use of Generalized Estimation Equations (GEE) introduced in 1986 by Kung-Yee Liang and Scott L. Zeger (Zeger and Liang 1986) to describe longitudinal data. It highlights the robustness and advantage of their estimation technique in presence of unknown correlation within multiple measurements of a same subject, which is often the case in real data.

2.2 Material and Methods

2.2.1 Data mining

There is a need for data mining tools to explore large and complex biological datasets to identify combinations of factors that optimally explain the outcome of interest. Hypothesis-free data exploration can potentially generate novel hypotheses that emerge from the data and which are beyond our imagination. These novel hypotheses can subsequently be tested using specific statistical methods or animal models.

In biology, data mining has been essentially focused on sequence alignment algorithms to manage the ever-increasing amount of genetic data. More recently, data mining technology has been proposed as an alternative to traditional statistics to deal with high dimensional data generated by Genome Wide Association studies, in the knowledge that accounting for gene-gene and gene-environment is crucial to understand human genetic susceptibility to disease (Nelson, Kardia et al. 2001; Ritchie, Hahn et al. 2001; McKinney, Reif et al. 2006; Cordell 2009). Factorial[1] approach and *Clustering* are widely used for data mining. In addition to such methods in the field of genetic data analyses, several new heuristic tools have been developed, notably non-parametric modeling techniques such as Classification And Regression Trees (CART) (Breiman, Friedman et al. 1984) and Random Forests (Breiman 2001). These methods present several advantages: models have the capacity to provide accurate fits of the response in a wide variety of situations, enabling fitting of non-linear relationships between explanatory variables and the dependent variable, with no assumption that explanatory variables are independent.

Complementary to these non-parametric methods and to traditional statistical methods, HyperCube® (Augustin Huret, Institute of Health & Science, Paris, France, http://www.institute-health-science.org) uses least general generalized algorithms and genetic algorithms. The underlying idea is to describe a dataset by a group of « local over densities » of a specific outcome with no *a priori* hypothesis or notion of distance, each « over density » being completely independent from every other. This method deals with points in a space with absolutely no assumptions, including those concerning metric and distance or nature of neighborhood essential in classical *Clustering*. Indeed, working with a distance or a defined topology is already an assumption and either is true or not true and, thus, can introduce bias into the model.

[1] Factorial methods represent data from a space with larger dimension, characterized by initial variables as the axes for representation, to a space with lower dimension, characterized by Principal Components (or Factorial Axes) made with linear combinations of initial variables, as the new axes for representation.

These data mining methods can be specified in two distinct approaches: Supervised and Unsupervised. In “Supervised” methods the Y variable, the outcome, is observed and the analyses are guided by this outcome; the other independent variables are selected depending on their capacity to explain the different categories or the distribution of values of the outcome. By contrast, in “Unsupervised” methods, there is no Y variable, and then, all variables play a symmetric role; the analyses are based on techniques that find relationships among variables and/or combinations of variables pertinent to highlight similarities/ dissimilarities among observation units.

2.2.1.1 Supervised

As supervised methods we can cite Factorial Discriminant Analysis (FDA) with a design of *p* quantitative explanatory variables plus one qualitative dependent variable. FDA is a descriptive method based on graphical representation using principal components that are made with linear combinations of initial variables; these principal components explain the dependent variable. We also have Discriminant Analysis (DA) with a same design as previously except for the fact that the qualitative variable is not observed and has to be forecasted. There are also Classification and Regression Tree (CART) (Breiman, Friedman et al. 1984) and Random Forests (RF) (Breiman 2001) methods that can handle any mixture of types of variables. Next, in this chapter, we will use CART as a supervised data mining methods to compare with the new exhaustive, non-Euclidean and non-parametric approach (Loucoubar, Paul et al. 2011). This method also can handle any mixture of types of variables, so adapted for application on our malaria datasets that comprise quantitative and qualitative variables.

Classification and Regression Tree

CART is a rule-based method that allows dichotomization of an explanatory variable into two classes or subsets (called nodes) with significantly different profiles for the response (i.e. maximizing the discrimination); this works in a recursive way applying same splitting in each child class (called sub-nodes) until convergence. Among all partitions of the explanatory variables at a node, the principle of the algorithm is to split the data according to a threshold on one of the variables, such that the reduction of heterogeneity between a node and the two sub-nodes is maximized. Each split is based on a single variable; some variables may be used several times while others may not be used at all. It generates a binary tree through recursive partitioning minimizing heterogeneity criterion computed on the resulting sub-nodes. This splitting algorithm (the growing step), to obtain in a first time the deep maximal tree, is

always followed by a pruning procedure that finally adopts the tree with the minimal expected misclassification error rate, by cutting off insignificant nodes.

In theory there are several functions for the measure of heterogeneity, but the two most widely used are the Gini index and the Shannon entropy that can be easily illustrated when the dependent variable is categorical.

Let Y be a binary dependent variable taking values 0 and 1. Let f_0 and f_1 be the proportions of $y = 0$ and $y = 1$ at a node:

- Gini index = $\Sigma_{i \neq j} f_i \times f_j = \Sigma_{i=0,1} f_i \times (1 - f_i) = 1 - \Sigma_{i=0,1} f_i^{\,2} = 1 - (f_0^{\,2} + f_1^{\,2})$
- Shannon entropy = $-\Sigma_{i=0,1} f_i \times log(f_i) = -[f_0 \times log(f_0) + f_1 \times log(f_1)]$ where $0 \times log(0) \approx 0$

CART uses these indices for convergence criteria of the splitting process. By definition these indices will be close to 0 at a node if that node contains almost only one category (homogeneity), i.e. for one category i, f_i is close to 1 and then all f_j with $j \neq i$ are close to 0.

As it is the case for the growing step, there are criterions to guide the pruning procedure. The two pruning procedures widely used for the minimization of the misclassification error rate are by the control of the minimum number of observations in each node (control of the tree size) and by cross-validation. Decreasing the minimum number of observations at the nodes increases the complexity (number of nodes and leafs, then the size of the tree) and decreases the misclassification error rate. However, this choice leads to overfitting, and then, the final decision tree will perform poorly on new independent data (low *true predictive power* of the tree). So the minimum size needs to be calibrated and cross-validation can help to find the optimal tree size by making a compromise between the complexity and the misclassification error rate of the tree through some complexity cost function. See Breiman *et al* (Breiman, Friedman et al. 1984) for more details.

2.2.1.2 Unsupervised

As unsupervised methods, we can cite Principal Component Analysis (PCA) (Pearson 1901; Hotelling 1933; Jolliffe 2002), Multiple Correspondence Analysis (MCA) (Greenacre 2010) and *Clustering* (Everitt, Landau et al. 2001; Manly 2005). The two first methods, PCA and MCA, use linear algebra to represent data in a space with reduced dimensions via singular value decomposition (Trefethen and Bau 1997). PCA and MCA are very similar and have three major common aspects: (i) Homogeneity in the type of variables to analyze, all are quantitative in PCA while all are qualitative in MCA; (ii) Symmetric role of variables, i.e. non distinction between endogenous and exogenous variables, only relations between variable are important; (iii) Search of factors or principal components by making linear combinations of

initial variables; graphics are made using principal components as for FDA. These methods are also based on Euclidean geometry. For further details in methods and illustrations, see Philippe Besse & Alain Baccini 2007 “Exploration Statistique” (Besse and Baccini 2007). *Clustering*, as unsupervised data mining method, can handle a mixture of variables with different types and metrics other than Euclidean. The simple example on cluster analysis in Annex A (on Figure A.1) illustrates how the results can be influenced by the metric considered. This then encourages us to use data mining method, not only non-parametric, but also without any defined metric, see part *2.2.1.3* below, the HyperCube® method.

Cluster analysis

Cluster analysis (Everitt, Landau et al. 2001; Manly 2005) is a multivariate statistical method that try to categorize a sample of subjects into different groups depending on their profile (or their measures) on a list of variables, such that comparable subjects are placed in the same group.

Cluster analysis can be used on genotypic data to identify genes that characterize a specific population or differentiate many populations (e.g. ethnic groups or different animal races of a same species, or a disease status) by measuring for a gene, represented by a set of maker loci, its capacity to classify similar subjects in a same group.

Limitation: Cluster analysis is sensitive to the metric selected to measure the distance between two subjects (as shown in Figure A.1 of Annex A) and also to the order of clustering. One can obtain different results by using different approaches, thus, the metric and the clustering method should be chosen carefully.

Non-hierarchical clustering methods, or k-means methods introduced by Forgy in 1965 (Forgy 1965), are preferred to hierarchical ones (single, complete or average linkage, Ward’s method). Indeed, k-means algorithm supposes that data will be classified in *k* classes and then work as follows:

(i) *k* points are randomly chosen in the space of individuals as centroids of the *k* initial classes;
(ii) each individual observation is associated to the closest class (distance to the centroids), in the sense of the defined metric;
(iii) barycentres of the clusters that have been formed are found and are set as new centroids;
(iv) steps (ii) and (iii) are repeated until the algorithm converges, i.e. until no change in the clustering between two iterations.

2.2.1.3 The HyperCube® method

We introduce here a new data mining tool using the method of hypercubes. This method belongs to the supervised data mining methods but is based on non-Euclidean geometry, it is assumption-free and proceeds through an exhaustive learning (Loucoubar, Paul et al. 2011). HyperCube® approach combines regression and optimization techniques by searching for all possible stratifications and identifying the best combination of variables to explain a specified outcome.

HyperCube® data mining algorithm

The HyperCube® technology is accessible as a web based software that requires a significant computing power provided through a SaaS architecture (http://www.institute-health-science.org). A hypercube is a subspace defined by a combination of conditions, each condition being either a range or a modality of a continuous or discrete variable (see illustration on Figure 2.2.1). A hypercube has various characteristics: its dimension, the number of variables involved; the "Lift", the measure of the over density compared to the whole database; the "Size", the number of points included in the hypercube.

After defining the dependent variable, HyperCube® program generates a series of rules by exhaustively exploring the space of the random variables, generating optimal subspaces significantly enriched with the occurrence of events, and defining for each interesting subspace, its explicative variables and their corresponding values. A rule is a set of a limited number of continuous and/or categorical variables and their associated values. A search by HyperCube® program is divided in three steps:

(i) A stochastic exploration of the space of random variables: Subspaces are exhaustively generated following this procedure: One point is randomly chosen as a germ (i.e. a starting point) in the *m*-dimensional space defined by the *m* explanatory variables; after, a 2^{nd} point is randomly selected to form a segment. These two points correspond to apical points of a starting subspace having a hypercube design and represent the diagonal of this hypercube (see Figure 2.2.1). This diagonal (jointly the volume of the hypercube) will be optimally increased. Each subspace is selected depending on two constraints: its size, the number of events included in the subspace, and its purity, the percentage of positive events in the subspace. To define explanatory variables, the corresponding axe for each variable delimiting the subspace is suppressed, and the subsequent subspace tested for satisfying the previous constraints. The variables for which the corresponding axe must be present to satisfy these constraints are the

explanatory variables. The subspace is cancelled if it does not satisfy the constraints defined by the user and a new subspace is generated.

(ii) An optimization of the characteristic of the hypercube: The volume of each initial hypercube selected at the first step is locally maximized depending on a Z score using genetic algorithms, and always constrained to a minimum purity.

(iii) Validation of the rule using a non-parametric approach: The Z score of the optimized hypercube is compared to those generated by a random permutation of the dependent variable.

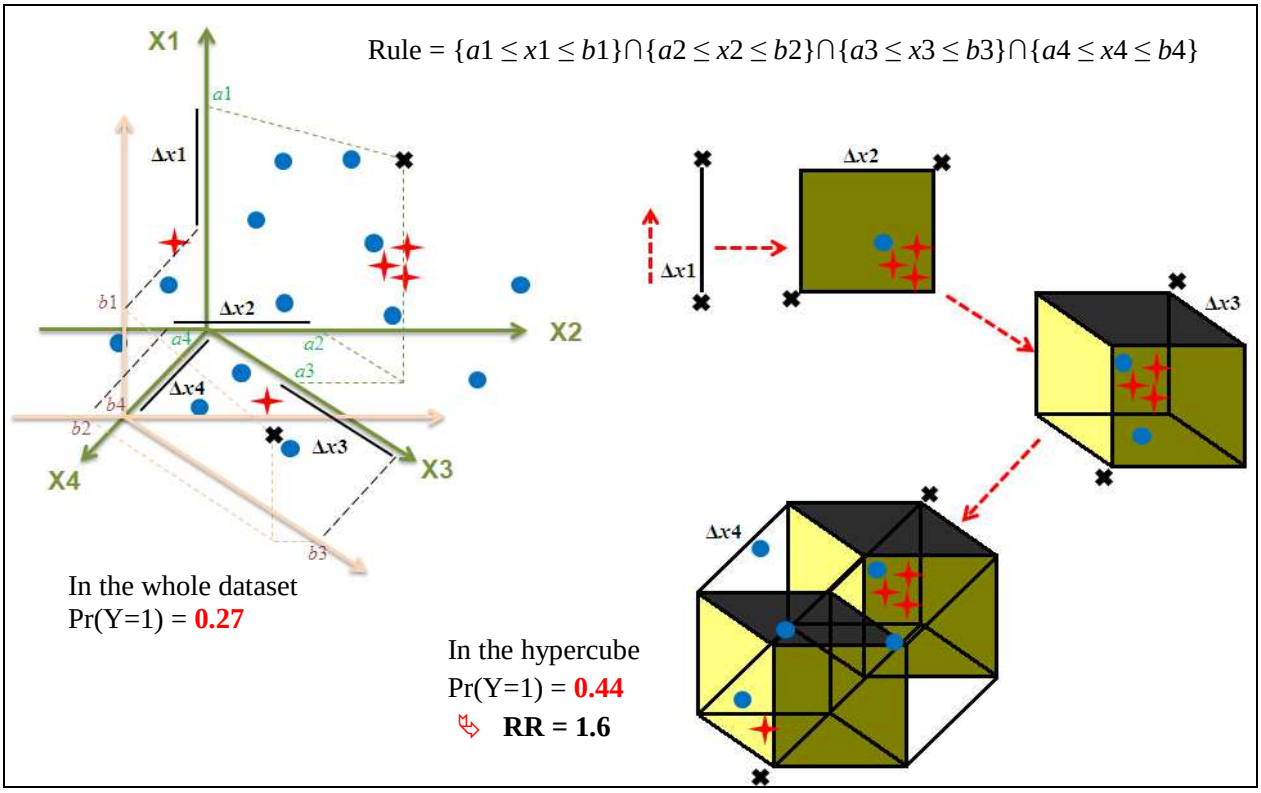


Figure 2.2.1: Principal for selecting a Hypercube (by just selecting the two apical points)

For exhaustiveness, these three steps are repeated until all points have been used as starting point and all the events have been studied; i.e. all the events in the learning dataset have been included in at least one rule. The user can stop the learning process at any time and know the coverage of his exploration. Due to human limitations in understanding complex rules, the maximal number of explanatory variables inside each rule can be fixed, thereby defining

complexity. HyperCube® uses an exhaustive non-parametric and non-Euclidean methodology, it does not use proximity between events but only generates subspaces in which events are present or not and counts occurrences.

One has first to define variables to introduce into the learning dataset and run a simple lift analysis. “Simple lift” classifies variables according to their first order effect and has three major roles: to verify consistency of the data, to detect circular variables and to detect variables with pivot points that define threshold values for their impact on the outcome. “Spearman (or Pearson) Correlation” associated with “Simple lift” help to define which variable to choose amongst the correlated variables. Sometimes, a combined variable from correlated variables is the best choice. The learning process is followed by a validation process. Signal Intensity Graph (SIG) defines the relationship between the two main parameters of a learning process, “purity” and “size”. This graph shows the value of the “purity” for five different “sizes” defined from the database and from a randomized database obtained by permutation. After defining the last parameter, “Complexity”, which defines the maximum number of variables per rule, the learning process is run. From the total number of rules, a set of minimized rules is obtained from an iterative process. In the first step, the rule explaining the most number of events is chosen and at each of the following steps the rule explaining the maximal number of events in the remaining event space not included in the first rule is added. The iterative process is stopped when all the events explained by the total number of rules are explained by the set of minimized rules. The total number of rules and/or the minimized rules is used to perform further analysis.

As mentioned previously, data mining is not always adequate for handling repeated measures and some of them, like the HyperCube® method, do not provide a way to adjust the results on the significant covariates effects. To make up for these weaknesses, appropriate regression techniques like GEE, presented here, or Mixed models presented in the next chapter 3, are used and were developed to handle data from longitudinal surveys.

2.2.2 GEE: estimation of population parameters for repeated measurements data

One of the aims of this prior descriptive analysis is to evaluate effects of the key known non-genetic factors that lead to the illness (a *P. falciparum* attack denoted *PFA*) of a person when he or she is exposed during a trimester. This task can be done using techniques other than data mining, like regression on environmental variables and individual non-genetic variables like age. However, here, the basic assumption of independence between observations in simple regression does not hold. The longitudinal design of the data has the advantage to provide consistence effects but induces several inconveniences such as non-independency; repeated

correlated measures of a same individual should be taken into account using extended regression methods.

This part presents a brief description of the extension of Generalized Linear Models to Generalized Estimation Equations (GEE) (Zeger and Liang 1986) in the context of longitudinal studies to accurately take into account correlation of multiple measurements from the same subject. We present here the model and give the main theory for the estimation of parameters.

2.2.2.1 GEE model

Two main specifications are needed in the context of GEE models:

(a) Measurements on the same subject are allowed to be correlated,
(b) Measurements on different subjects are assumed to be independent.

Specification (b) could be problematic / not met when analyzing family data, but as a first step we are interested in population mean effects of variables; methods presented in Chapter 3 will take into account non-independency among individuals.

Let y_{ij} , i = 1, ..., N and j = 1, ..., n_i , denote the outcome of infection (dependent variable) of the i^{th} individual at his j^{th} episode. There are N individuals and n_i measurements on the individual i and $n = \sum_{i=1}^{N} n_i$ total episodes. Note that the observation times can differ from one individual to another (Zeger and Liang 1986). The presence (y_{ij} = 1) or absence (y_{ij} = 0) of illness in subjects as well as several other epidemiological covariates like level of parasitemia, sex, current age, etc., were recorded at each episode for a subject. We have to consider an individual as a unit. If we take an individual i, his observed data are stored in a vector y_i of dimension (i.e. number of row × number of columns) n_i×1 for the dependent variable and in a matrix X_i of dimension n_i×p for the p covariates:

- $y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix}$ with expected mean $\mu_i = \begin{pmatrix} \mu_{i1} \\ \vdots \\ \mu_{in_i} \end{pmatrix}$

and variance-covariance $V_i = \begin{pmatrix} \text{var}(y_{i1}) & \cdots & \text{cov}(y_{i1}, y_{in_i}) \\ \vdots & \ddots & \\ \text{cov}(y_{in_i}, y_{i1}) & \cdots & \text{var}(y_{in_i}) \end{pmatrix}$

- $\mathrm{X}_i = \begin{pmatrix} \mathrm{X}_{i1}^1 & \mathrm{X}_{i1}^2 & \cdots & \mathrm{X}_{i1}^p \\ \vdots & & & \\ \mathrm{X}_{in_i}^1 & \mathrm{X}_{in_i}^2 & \cdots & \mathrm{X}_{in_i}^p \end{pmatrix}$

Then, for individual *i*, the expected phenotype is modeled as $g(\mu_i) = \mathrm{X}_i\beta$ where *g* is the link function, that express the expected phenotype as a linear function of the explanatory variables, and β the vector containing the effects of the *p* explanatory variables.

The GEEs to solve for estimating β is given by: $\sum_{i=1}^{N} \frac{\partial \mu_i^{\mathrm{T}}}{\partial \beta} V_i^{-1}(Y_i - \mu_i) = 0$ (generalization of quasi-likelihood equations).

If repeated measurements from a same individual *i* were supposed to be independent, V_i would be equal to a matrix A_i with var(y_{ij})'s on the diagonal and 0 elsewhere, i.e. cov(y_{ij},$y_{ij'}$)=0, $\forall\ j \neq j'$, j and j' in {1, ..., n_i}: $A_i = \begin{pmatrix} \mathrm{var}(y_{i1}) & & 0 \\ & \ddots & \\ 0 & & \mathrm{var}(y_{in_i}) \end{pmatrix}$, and in that case GEE would be exactly the simple GLM. However, in most of cases, this independency within individual does never hold because repeated observations are made on each individual, correlation must be anticipated among an individual's measurements. It must be accounted for to obtain a correct statistical analysis. Then, cov(y_{ij},$y_{ij'}$)'s are specified in a "working" correlation matrix $R_i(\alpha)$ that can reflect the type of correlation among samples from a same individual. The α defines the parameterization of the R_i's which are the same for all individuals. Note that "working" refers to the fact that $R_i(\alpha)$ is not expected to be correctly specified, but estimators will be consistent and will have consistent variance estimates even when $R_i(\alpha)$ is misspecified (Zeger and Liang 1986). Therefore, the covariance matrix of repeated phenotypes of a same individual *i* becomes: $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}$.

Note that in quasi-likelihood theory, variance of y_i is expressed as a known function of the expectation of y_i divided by a scale parameter ϕ, $V_i = h(\mu_i)/\phi$, then A_i would be expressed as $diag[h(\mu_{i1}), \ldots, h(\mu_{ini})]/\phi$ and finally $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} / \phi$, expression of the covariance matrix more frequent in the literature.

A useful feature of the GEE approach is that it is not necessary for the "working" correlation matrix to be correctly specified to obtain a consistent and asymptotically Gaussian estimate of β, the effects of explanatory variables on the phenotype. Several working correlation structure had been presented by Liang & Zeger, and choosing the working correlation matrix to be close to the real one, however, increases efficiency (Zeger and Liang 1986). In our study, the outcome of an infection (*PFA* or Not) of two successive clinical episodes for an individual

were assumed to be correlated, implying our choice of an autoregressive of order one, denoted AR(1), “working” correlation structure:

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \dots & \alpha^{n_i} \\ \alpha & 1 & \alpha & \alpha^2 & \dots & \alpha^{n_i-1} \\ \vdots & & & & & \\ \alpha^{n_i} & \dots & \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}$$

The parameter $\alpha \in [0, 1]$, therefore more two episodes are far away, less their correlation is important.

2.2.2.2 GEE iterative estimation

The iterative fitting algorithm used in GEEs can be presented through the following steps:

(i) An initial estimate of effects β is computed using simple GLM, i.e. by assuming independence;
(ii) Current Pearson residuals are deduced on the basis of the current estimate of β;
(iii) An estimate of the working correlation matrix $R(\alpha)$, having the chosen parameterization form, is computed on the basis of the current Pearson residuals and the current estimate of β;
(iv) An estimate of the variance Vi is then computed;
(v) An updated estimation of β is computed taking into account V_i.

Steps (ii) to (v) are repeated until convergence, i.e. until no change in the estimation of β. The final (and stable) estimate of β obtain is the GEE estimation of the effects of explanatory variables.

2.3 Results

In this results section, only main findings are presented. The detailed methodology, of our already published results concerning the descriptive analysis (Loucoubar, Paul et al. 2011), are presented in the Annex.

2.3.1 The measured phenotypes

The main outcome of interest in our study is a *P. falciparum* malaria attack (*PFA*). *PFA* was defined as a presentation with measured fever (axillary temperature >37.5˚C) or fever-related symptoms (headache, vomiting, subjective sensation of fever) associated with i) a *P. falciparum* parasite/leukocyte ratio higher than an age-dependent pyrogenic threshold previously identified in the patients from Dielmo village (Rogier, Commenges et al. 1996), ii) a *P. falciparum* parasite/leukocyte ratio higher than 0.3 parasite/leukocyte in Ndiop village. The threshold was used because of high prevalence of asymptomatic infections in the populations, as occurs in regions endemic for malaria (Sinton 1931; Miller 1958; Richard, Lallemant et al. 1988; Smith, Genton et al. 1994).

Time period of observation was classified as a trimester, and then units of observation were person-trimesters. The dependent variable was defined as a binary trait: individuals with at least one clinical *PFA* during that trimester or without *PFA*. In total, there were 46,837 outcome events of person-trimesters from 1,653 individuals. Almost 20% of the events were *PFA* in both villages.

NB: We were also interested in other phenotypes that reflect frequency and infectiousness of the disease for an individual, see chapters 3 & 4: 1) the number of *P. falciparum* clinical episodes, or malaria attacks, during each trimester (*nbPFA*) and units of observation for this phenotype were person-trimesters; 2) the proportion of clinical episodes that were positive for gametocytes, parasite stages transmissible to mosquitoes (*Pfgam*).

2.3.2 The covariates

Some explanatory variables are time-dependent and then were evaluated for each trimester. These included current age, experience of exposure to other *Plasmodium spp*. (*P. ovale* and *P. malariae*) before the current trimester defined by the cumulated number of previous infections, the corresponding year and trimester, time spent in the village during the current trimester. Other variables are individual-dependent including sex, geographic location (e.g. village, house), and genetic profiles (e.g. blood type, hemoglobin type, Glucose-6-phosphate dehydrogenase (G6PD) deficiency status (by genotypes and by enzyme activity). The list of variables analyzed are presented in the Annex in Publication 1 (Loucoubar, Paul et al. 2011).

2.3.3 The changing epidemiology of malaria in the last decade

We categorized clinical episodes for a volunteer into 3 groups: #0 as absence of episode during a trimester of observation or as having clinical episode(s) without *P. falciparum* infection, but including malaria episodes due to *P. ovale* or *P malariae* (not *PFA*), #1 as having at least one episode with *P. falciparum* infection but not attributed to *P. falciparum*, i.e. parasites density under the threshold (not *PFA*), #2 as having at least one episode attributed to *P. falciparum*, i.e. parasites density above the threshold (*PFA*). For each volunteer and at each trimester of presence, the incidence rate of *P. falciparum* infections (#1) and attacks (#2), corresponding to panels A and B respectively of Figures 2.3.1.a for Dielmo & 2.3.1.b for Ndiop, was estimated as the number of such episodes divided by the number of days of presence for each time period.

The global burden of malaria decreased dramatically over the last decade in both sites (Figures 2.3.1.a & 2.3.1.b) as reported in several other malaria endemic areas (Bhattarai, Ali et al. 2007; Ceesay, Casals-Pascual et al. 2008; O'Meara, Bejon et al. 2008) due to efficacy of combining effective vector control and effective case management. Figures 2.3.1.a and 2.3.1.b thus reflect the decreasing impact on the burden of the ACT (2007) and ACT plus long-lasting insecticidal-treated nets (LLIN) (2008) at a rural community level. The at-risk population for malaria episodes remained the younger children; only a few malaria episodes occurred in adults in either village (Figures 2.3.1.a & 2.3.1.b).

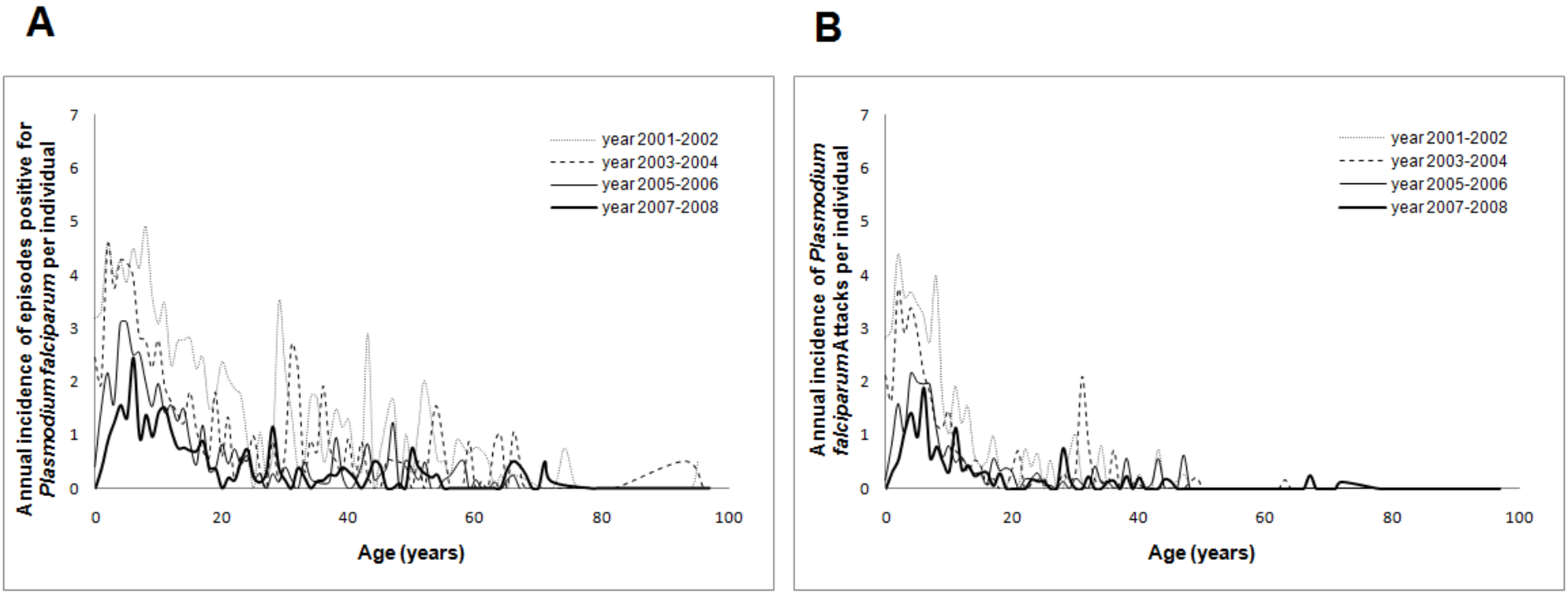


FIG. 2.3.1.a. Incidence rate (per person per year) of malaria infections (A) and attacks (B) between 2001 and 2008 depending on age in Dielmo.

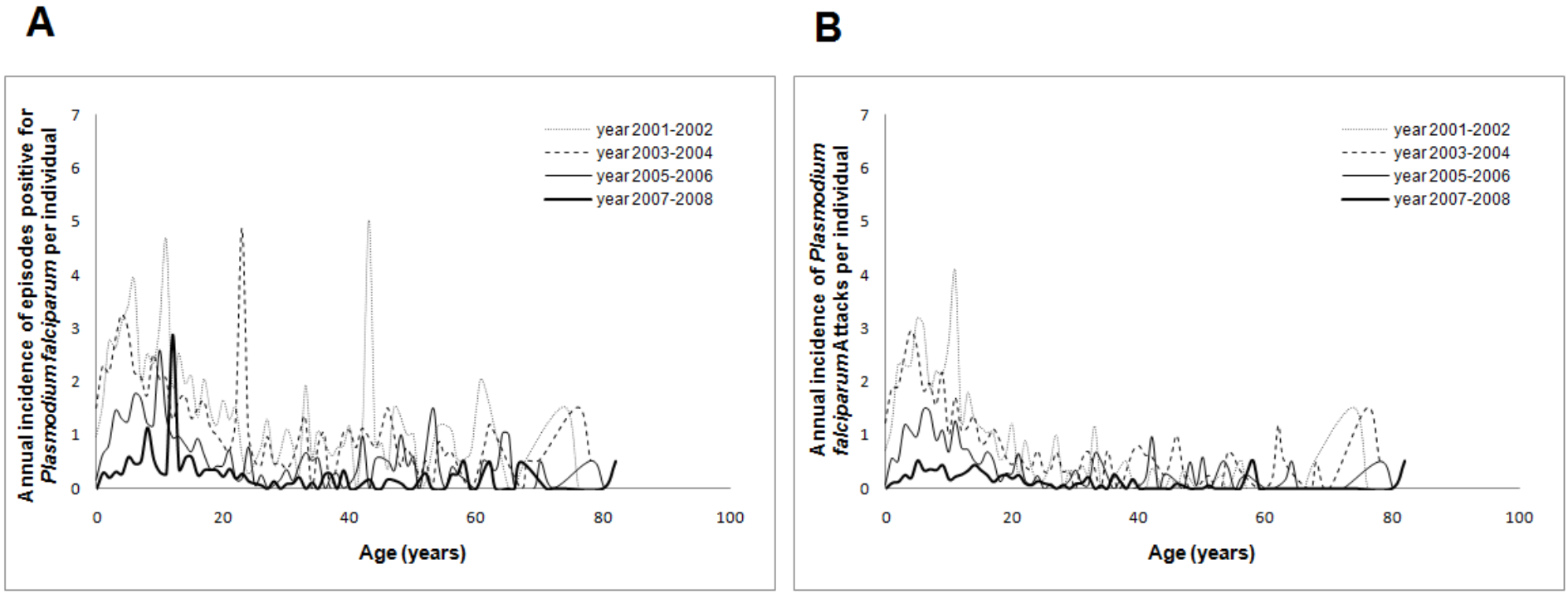


FIG. 2.3.1.b. Incidence rate (per person per year) of malaria infections (A) and attacks (B) between 2001 and 2008 depending on age in Ndiop.

Figures 2.3.2.(A–D) show that the number of malaria episodes per individual decreased over time; this was most notable in the over 12 year old group. This trend was similar for the clinical episodes not related to *P. falciparum*. Figure 2.3.2.(A–D) illustrates the changing epidemiology of clinical malaria following the use of an efficient antimalarial drug therapy such as ACT combined with systematic malaria detection following the onset of clinical symptoms. Notably, in children below 12 years of age, the decrease in the number of malaria episodes reveals an increased number of non malarial clinical episodes. This is probably due to concomitant infections that were previously erroneously classified as malaria episodes, although may reflect release of co-circulating pathogens from the suppressive effect of *P. falciparum* malaria. As in adults, the numbers of persons with no clinical episodes increased between 2001 and 2008; this was more marked in Ndiop rather than in Dielmo.

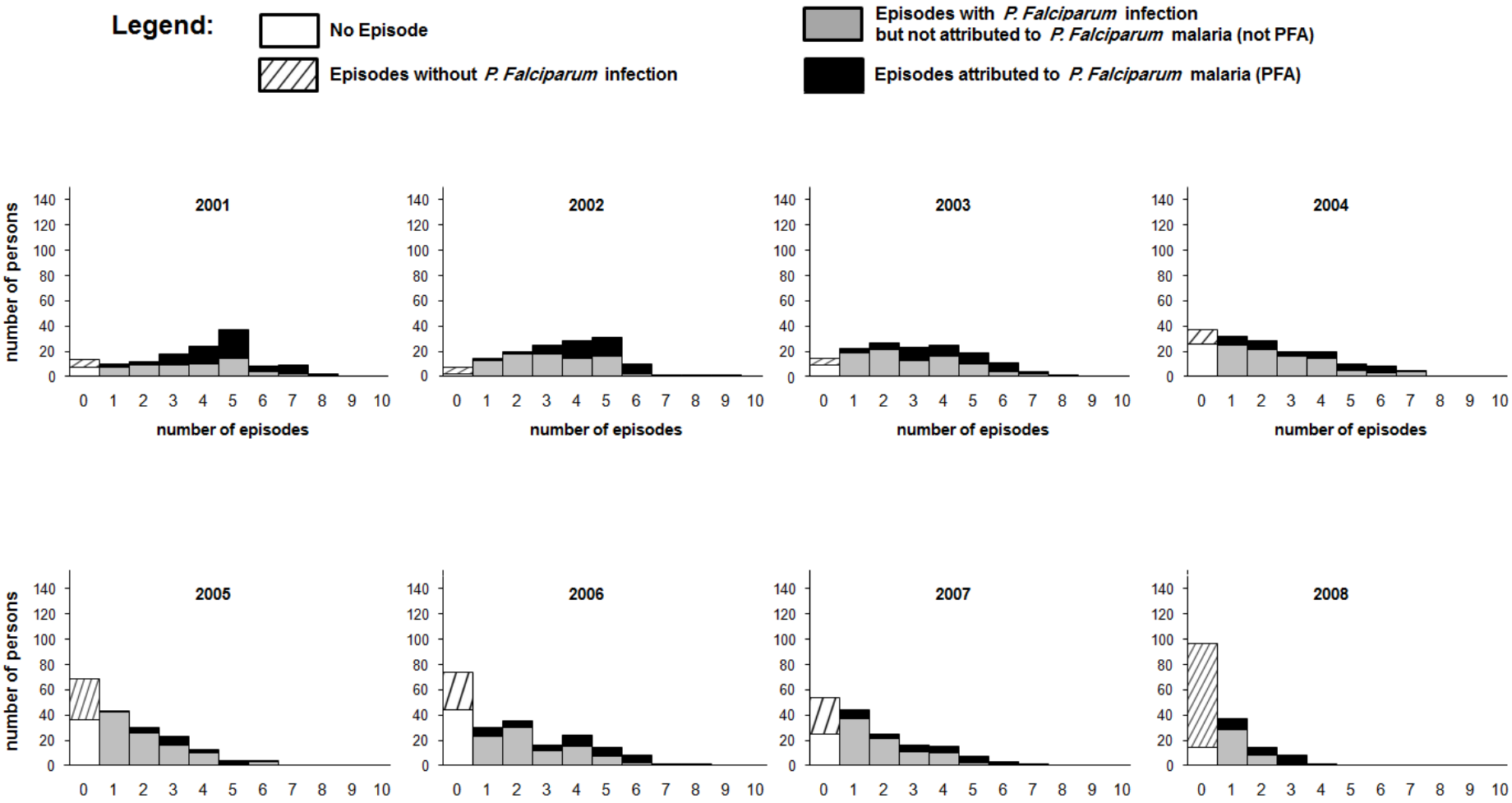


FIG. 2.3.2.(A). Evolution of the number episode types per individuals within group having less than 12 years-old in Dielmo.

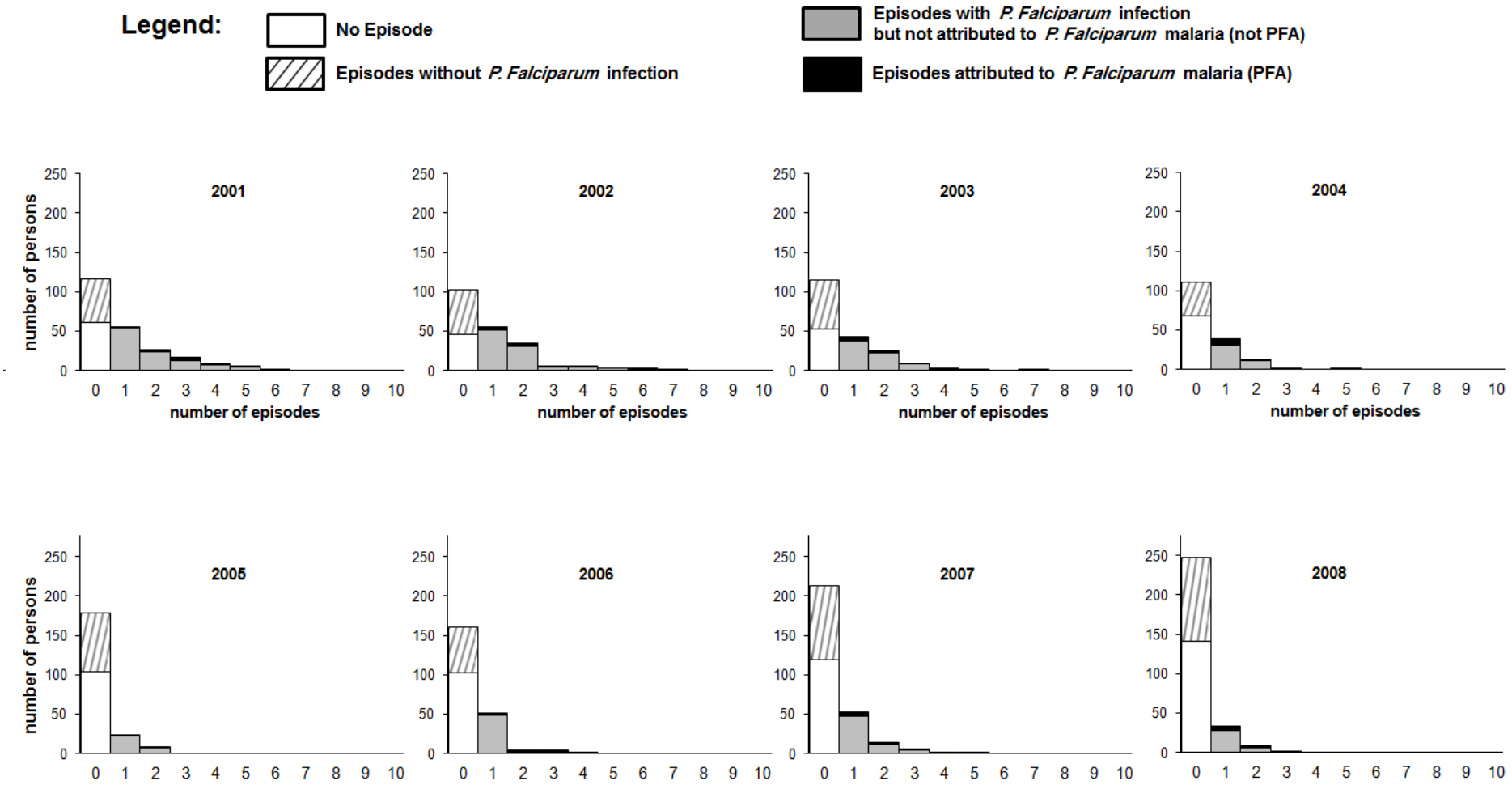


FIG. 2.3.2.(B). Evolution of the number episode types per individuals within group having more than 12 years-old in Dielmo.

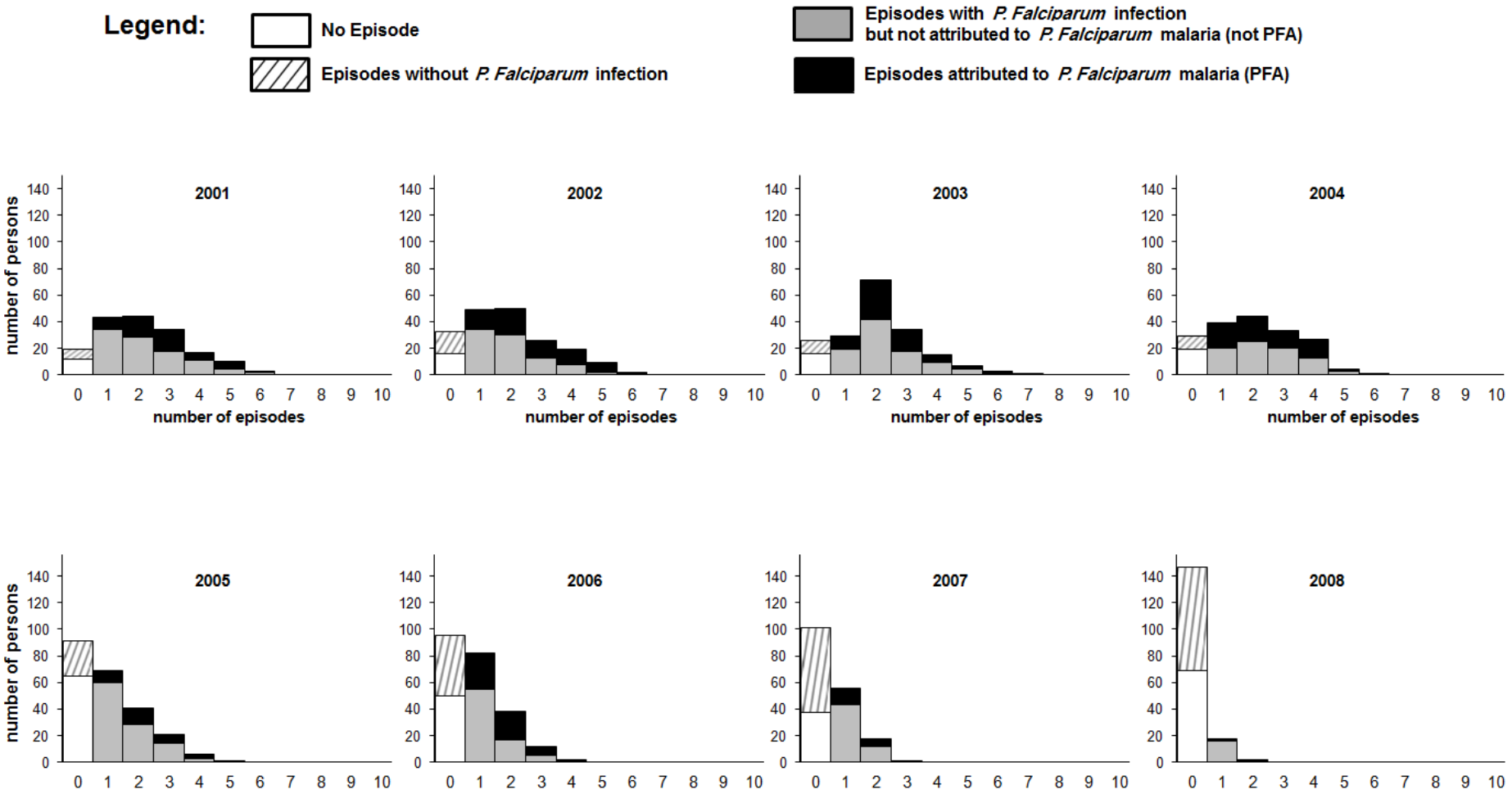


FIG. 2.3.2.(C). Evolution of the number episode types per individuals within group having less than 12 years-old in Ndiop.

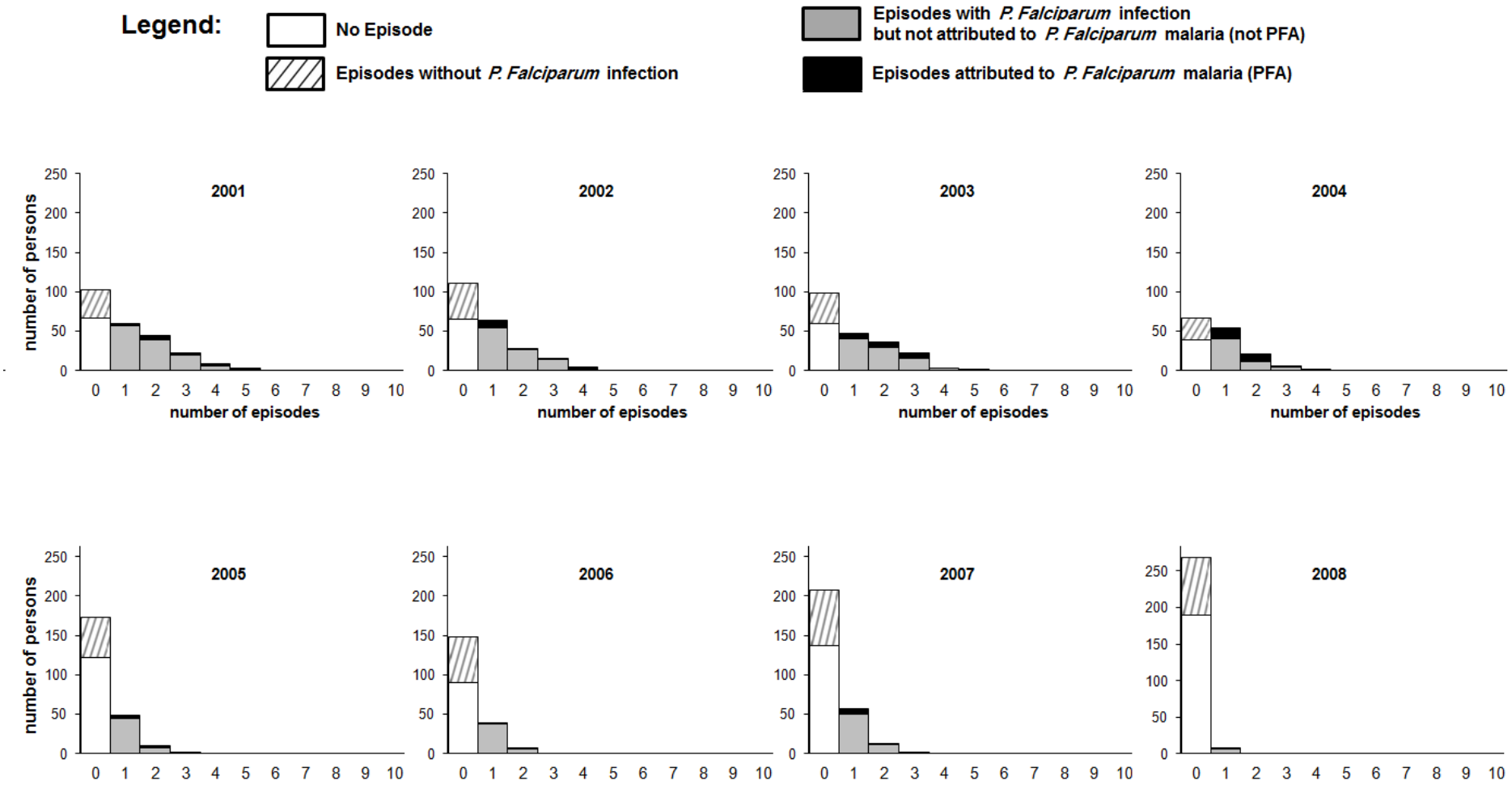


FIG. 2.3.2.(D). Evolution of the number episode types per individuals within group having more than 12 years-old in Ndiop.

2.3.4 Results from data mining using CART

Data mining analyses by CART used Gini index in the splitting step as measure of homogeneity of the nodes and cross-validation in the pruning step to optimize the misclassification error rate, using package *Rpart* from R software version 2.13.2. Thus, CART identified two major variables, Age and Year, which are determinant to predict occurrence of *PFA*. The different leafs correspond to different subpopulations in terms of susceptibility / resistance.

In Dielmo for example, person-trimesters aged from ~8 to 14 years-old whatever the period and their measured values for other variables, had similar risk to develop *PFA* compared to the entire cohort (Figure 2.3.3 (A)); they are no more at risk but not yet protected (RR = 0.95 [95%CI: 0.89 – 1.02]). Individuals having more than ~14 years-old are in general protected whatever their other aspects (RR = 0.23 [95%CI: 0.21 – 0.24]). However, having age between 0.22 and 5.48 and being present during years from 1990 to 2003 defined the high risk group for having *PFA* (RR = 3.26 [95%CI: 3.16 – 3.38]). No other variable or combination of variables yielded a higher Relative Risk by CART method.

In Ndiop, malaria epidemiology is strongly dependent upon season, as expected because mosquito abundance depends on the rains in this village. All individuals are protected (RR = 0.23 [95%CI: 0.21 – 0.25]) during the period of year from January to June (coinciding to dry season, i.e. no rainfall, in this region of Africa) due to absence of the vector and therefore absence of transmission. In this second cohort, even for more than 15-year-old, the protection is weak (RR = 0.85 [95%CI: 0.80 – 0.91]) compared to the same age group from Dielmo because they are not always exposed to malaria infections and hence have developed weaker clinical immunity. All individuals having less than 15-year-old are at risk, confirming that immunity is acquired later in this lower endemic area; the highest relative risk was found for period from 1992 to 2003 (RR = 3.12 [95%CI: 3.02 – 3.23) before decreasing to half the level in 2004 and after (RR = 1.50 [95%CI: 1.39 – 1.61]).

Figures 2.3.3 (A & B) are the classification trees identified by CART for each village. Figures show at each node the cut-off values that divide the dataset into two; at each final leaf are given the Relative Risk (RR) and the number of events associated with that leaf.

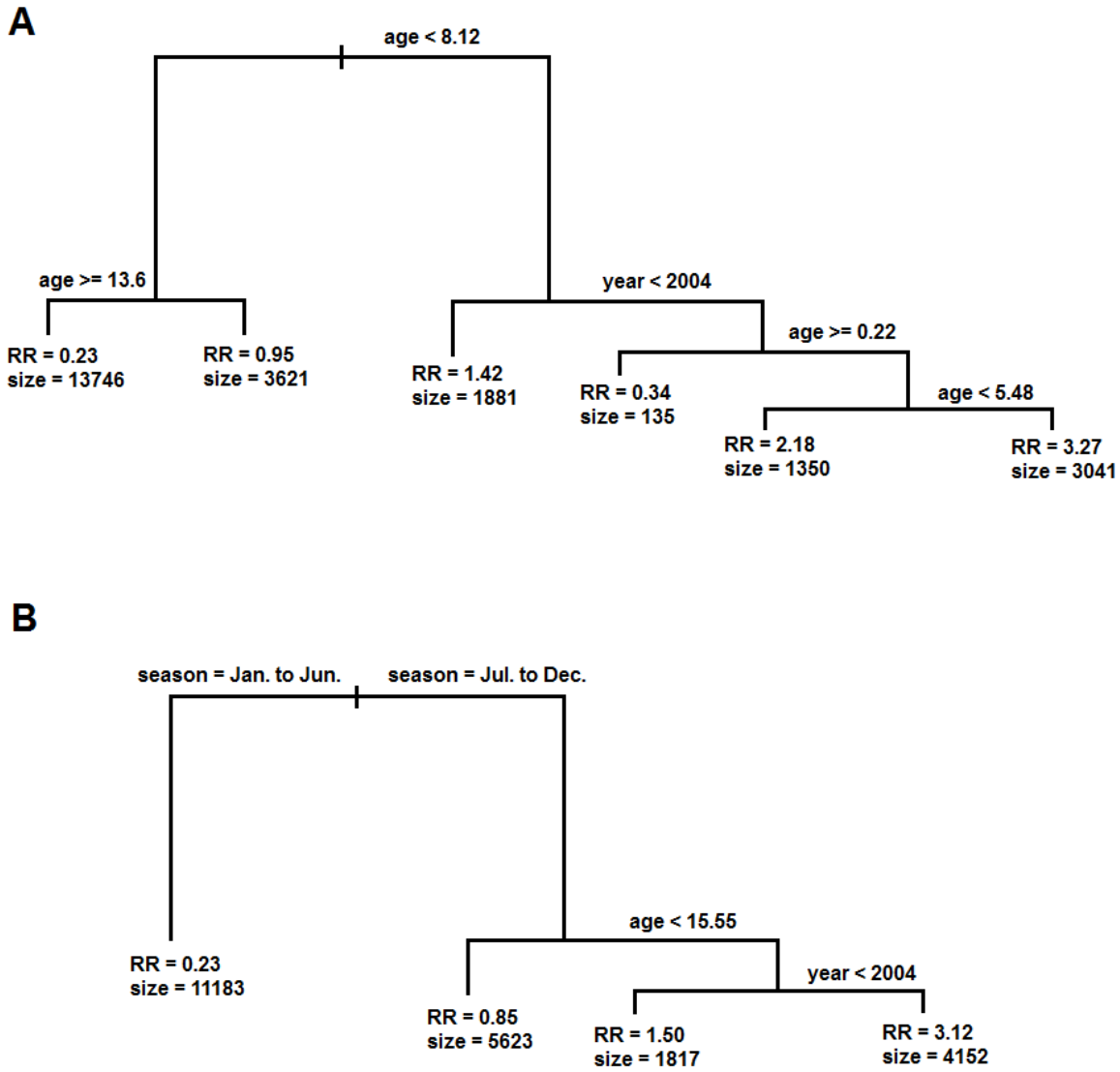


FIG. 2.3.3. Classification tree generated by Classification and Regression Tree (CART) analysis of risk factors determining the occurrence of *P. falciparum* malaria attacks (*PFA*) per trimester in Dielmo (A) and Ndiop (B).

2.3.5 Results from data mining using HyperCube®

We divided our dataset into three phases: learning, validation and replication. We analyzed the two cohorts separately. A random variable was created dividing the data of each cohort into two groups of equal size (in and out samples). The learning phase was carried out using the "in sample" from the first studied cohort. In the validation phase, rules defined in the learning phase were validated in the "out sample" of the same cohort. The effect of each validated rule from the first cohort was studied in the second cohort in the replication phase.

We selected the best predicted rule for further statistical study. The best predictive rule contained 1,689 events from 148 individuals and was defined as: individuals who lived in Dielmo during 1992 to 2003, were of an age between 1 to 5 years old, having hemoglobin type AA , and having had previous *P. malariae* infection (PMI) less than or equal to 10 times. These individuals had 3.71 (95%CI: 3.58 – 3.84) times more *PFA* than the general population; and this sub-population was the most representative (i.e. containing the maximum number of events) among those with a RR of at least equal to 3.71.

2.3.5.1 Replication of the rule in the 2nd cohort

In order to validate the biological and epidemiological aspect of this HyperCube® rule, it was replicated in Ndiop where a sub-population defined as above for Dielmo presented a higher risk to develop *PFA* compared to the general population: (χ^2= 665.96, DF=1, $P < 10^{-16}$), RR of 2.35 (95%CI: 2.22 – 2.48) and OR of 3.50 (95%CI: 3.16 – 3.87). The result was optimal in Dielmo and replicated in Ndiop. Thus, the four variables identified above to be risk factors in Dielmo were also risk factors in Ndiop.

The two cohorts differ in one very pertinent manner: in Dielmo malaria transmission occurs all year round because of the presence of a small stream that enables mosquitoes to breed. In Ndiop, transmission is highly seasonal and occurs during the rainy season (July-December). Hence, we calculated the risk in Ndiop using only the period of year between July to December, a period when environmental factors are quite similar in the two villages. We obtained the same relative risk, RR = 3.78 (95%CI: 3.62 – 3.94), OR of 11.80 (95%CI: 10.11 – 13.77), with a highly significant Pearson chi-square test (χ^2= 1542.50, DF=1, $P < 10^{-16}$). Furthermore, this risk was maximum in Ndiop when age was re-set to 3 to 7 years old (RR = 4.11, 95%CI: 3.97 – 4.27 and OR = 17.31, 95%CI: 14.68 – 20.41) with more events (Size = 932 events from 179 individuals *vs.* of Size of 863 from 157 when using age 1 to 5) and higher significance (χ^2= 2076.17, DF=1, $P < 10^{-16}$).

2.3.5.2 Comparison with other models

We examined whether a classical statistical method could identify the same or better rules. We performed logistic regression analysis and CART using the Dielmo data. We performed logistic regression using several model selection methods: (1) selection based on an exhaustive screening of candidate models in each subset of explanatory variables, selecting the best one in terms of Information Criterion (lowest Akaike Information Criterion (AIC)); (2) forward selection and backward elimination. The results obtained are presented in Table 9 from Publication 1 (Loucoubar, Paul et al. 2011). All sub-groups identified using model selection techniques had lower predictive values for developing *PFA* than the HyperCube® rule. For sub-groups explaining the same or a greater number of events than the one found by HyperCube®, the RR was lower and the 95% confidential intervals of RR did not overlap with those for the HyperCube® rule (Table 9 from Publication 1).

We tested whether the HyperCube® rule predicted the highest risk of developing *PFA*. We used the HyperCube® model as a reference. We modified the reference HyperCube® rule by either removing one of the variables or adding in variables identified by multivariate analysis. As shown in Table 10 from Publication 1, there was no other model that gave higher RR or OR than the one identified by HyperCube® with equal or greater size.

2.3.5.3 Optimality of the rule

We then tested whether the cut-off values delimiting the range of values in the HyperCube® rule (defined as the reference rule) for each variable were the optimal ones. Hemoglobin type was fixed as AA or not. We modified the range of continuous variables of the reference rule. As the cut-off values for continuous variables were considered at integer values, there were a finite number of subsets that we could try for modifying a rule. We tested all possible ranges of the continuous variables (Age, previous PMIs and Year). We first fixed 2 variables and changed one variable at a time. The variable to change was first defined as the range of integer values between its minimum and maximum values, and then reduced from the maximum to smaller integer values covering an ever-decreasing total range until the minimum. This was repeated step by step until each integer value of the variable was set as the minimum for a step. Therefore, the total number of choices for a variable is $1 + 2 + 3 + \ldots + maximum$ = sum of a finite arithmetic sequence = (first value + last value)×(number of values)×(1/2). Each choice corresponds to a specific modification of the reference rule (i.e. a specific interval of values defining the modified rule). Then, for Age, previous PMIs and Year, there are (1+98)×98×0.5 = 4851, (1+45)×45×0.5 = 1035 and (1+19)×19×0.5 = 190

possible choices respectively. We then fixed 1 variable and changed 2 variables simultaneously. When Year is fixed and the couple (Age, previous PMIs) changed simultaneously, there are 4851×1035 = 5,020,785 possible choices. For previous PMIs fixed and (Age, Year) changed and Age fixed and (previous PMIs, Year) changed there are 4851×190 = 921,690 and 1035×190 = 196,650 possible choices. When we selected choices with at least same size as the reference rule, the resulting RR was always lower than the reference RR. Figure 2.3.4 shows the effects of the modified ranges (i.e. the effect of other choices different from the one found by HyperCube®) on the RR. If all 3 variables were allowed to vary simultaneously there would be 4,851 (for Age) ×190 (for Year) ×1035 (for previous PMIs) = 953,949,150 possible choices. The time for running such an analysis on one computer with 2 central processor units (Duo CPU 2.00 GHz 2.00 GHz), Memory (RAM) of 3.00 GB) is estimated at ~ 5678 days (~ 1.94 choices analyzed per second) using function "*system.time(.)*" of R-software, and thus not possible to analyze by this ways of screening exhaustively.

Figure 2.3.4 below shows RRs for all other possible definitions of risk group on the explanatory variables, with equal or greater size than the HyperCube® rule. Y-axis indicates the RR. A) Only ranges of Age are modified: 102 choices among 4,851 possible choices had size equal or greater than 1,689 (size of the HyperCube® rule) and are plotted; B) Only ranges of previous PMIs are modified: 35 choices among 1,035 possible; C) Only ranges of Year are modified: 25 choices among 190 possible; D) Ranges of both Age and previous PMIs are modified simultaneously: 25,040 choices among 5,020,785 possible; E) Ranges of both Age and Year are modified simultaneously: 8,912 choices among 921,690 possible; F) Ranges of both previous PMIs and Year are modified simultaneously: 1,110 choices among 196,650 possible. Filled red triangle represents the RR of HyperCube®'s rule (HyperCube®'s risk group), empty black circles represent the RR of other choices of risk groups.

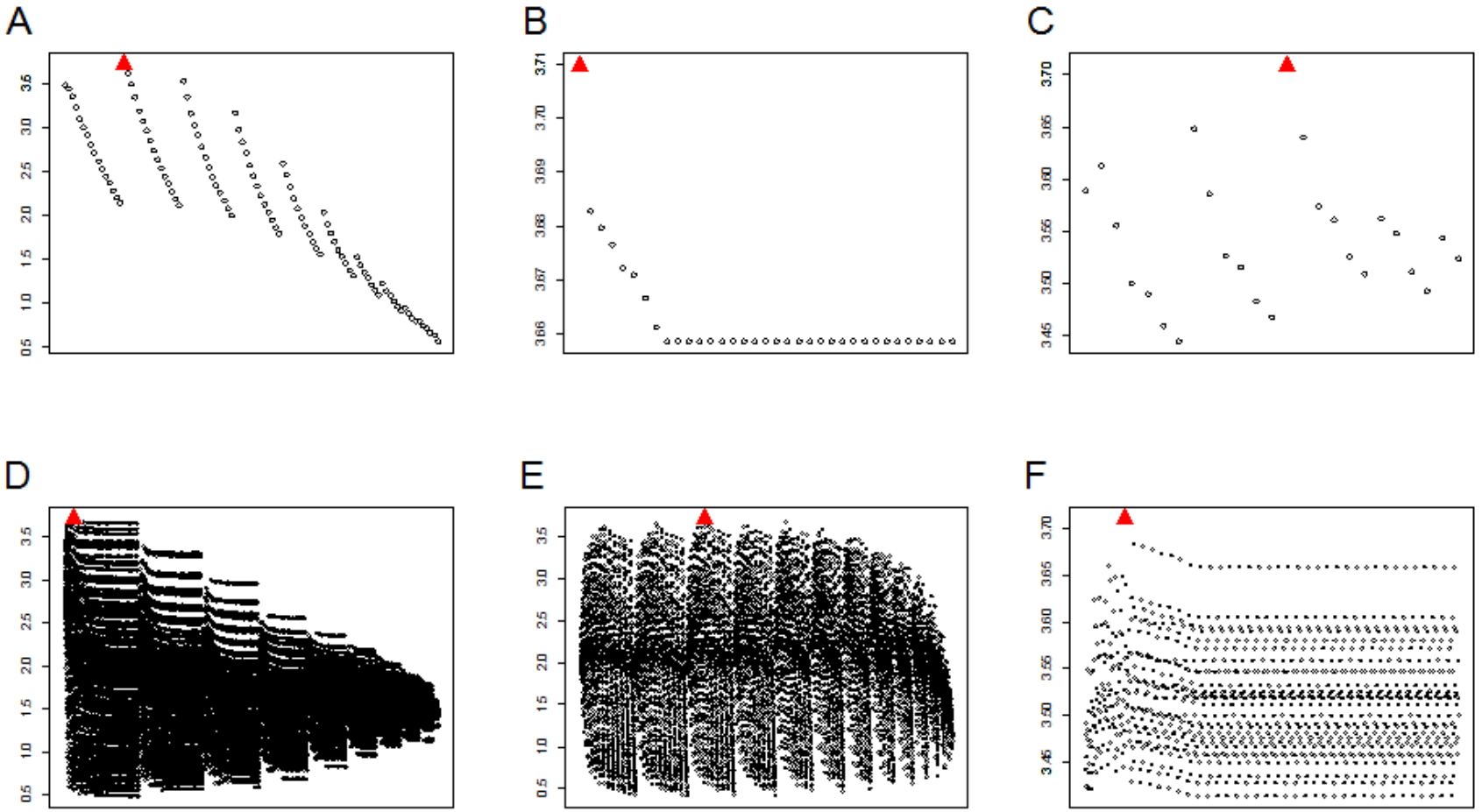


FIG. 2.3.4. Effect on the relative risk (RR) of modifying the ranges of continuous variables found by HyperCube®.

2.3.6 Results from GEE regression

Multivariate analyses using GEE (with binomial distribution and LOGIT link function) identified several factors determining the risk of developing *PFA* during a trimester. The results comparing estimations from GEE and estimations considering independence of repeated episodes within a same individual are summarized below in Tables 2.3.1. (A) and (B) for Dielmo and Ndiop respectively. Only interactions of order 2 were tested and the significant (P-value ≤ 0.05) are presented.

The use of LOGIT link unable to have directly the adjusted odds ratios (OR) by taking the exponential of the parameters estimated from the models. Therefore, an additive effect of interaction estimated between two variables is traduced by a multiplicative effect on each of their marginal OR; that's because $exp(a+b) = exp(a) \times exp(b)$. For instance, on Table 2.3.1.(A) for Dielmo, between Hemoglobin and G6PD: the marginal GEE estimates of the adjusted OR of having *PFA* for each of these variables are respectively 4.28 and 3.01 while the interaction effect is 0.33. Then, individuals with "AA" hemoglobin and "Not BB" G6PD are 4.28 times more susceptible to develop *PFA* during a trimester than those with "Not AA" hemoglobin and "Not BB" G6PD. However, having "BB" G6PD additionally, i.e. individuals with "AA" hemoglobin and "BB" G6PD, changes the risk from 4.28 to $4.28 \times 0.33 = 1.41$ compared to the same individuals (i.e. "Not AA" hemoglobin and "Not BB" G6PD).

Table 2.3.1.(A): Risk factors identified by GEE in village of Dielmo.

		Analysis considering independence				GEE Analysis			
Variables		**Adjusted OR**	**95% CI**		**P-values**	**Adjusted OR**	**95% CI**		**P-value**
Sex	*Male (ref.)*	1	-	-	-	1	-	-	-
	Female	3.02	[1.92	4.75]	1.74E-06	3.02	[1.79	5.11]	3.63E-05
Hemoglobin	*Not AA (ref.)*	1	-	-	-	1	-	-	-
	AA	4.21	[2.45	7.24]	1.89E-07	4.28	[2.08	8.82]	7.82E-05
Sex*Hemoglobin	*Female & AA*	0.33	[0.20	0.52]	2.95E-06	0.33	[0.18	0.58]	1.23E-04
G6PD	*Not BB (ref.)*	1	-	-	-	1	-	-	-
	BB	2.96	[1.90	4.62]	1.81E-06	3.01	[1.56	5.78]	9.67E-04
Hemoglobin*G6PD	*AA & BB*	0.33	[0.21	0.52]	2.07E-06	0.32	[0.16	0.64]	1.15E-03
Blood group	*A, B, AB (ref.)*	1	-	-	-	1	-	-	-
	O	1.27	[1.13	1.42]	3.35E-05	1.27	[1.03	1.56]	2.32E-02
Age group (in years)	≤ 4 *(ref.)*	1	-	-	-	1	-	-	-
	5 *to* 14	0.21	[0.18	0.24]	<1.00E-16	0.21	[0.16	0.27]	<1.00E-16
	15 *to* 34	0.05	[0.04	0.06]	<1.00E-16	0.05	[0.04	0.08]	<1.00E-16
	≥ 35	0.03	[0.02	0.04]	<1.00E-16	0.03	[0.02	0.04]	<1.00E-16
Drug treatment period	*Quinine (ref.)*	1	-	-	-	1	-	-	-
	Chloroquine	0.81	[0.70	0.92]	1.86E-03	0.80	[0.66	0.97]	2.44E-02
	Fansidar	0.20	[0.16	0.26]	<1.00E-16	0.20	[0.14	0.29]	<1.00E-16
	ACT	0.11	[0.08	0.15]	<1.00E-16	0.11	[0.07	0.16]	<1.00E-16
Semester	*Jan. - Jun. (ref.)*	1	-	-	-	1	-	-	-
	Jul. - Dec.	1.24	[1.11	1.39]	1.15E-04	1.24	[1.12	1.38]	8.15E-05
cumulated PFA	+1 *attack*	1.06	[1.06	1.07]	<1.00E-16	1.06	[1.05	1.07]	<1.00E-16
cumulated PMI	+1 *infection*	0.95	[0.94	0.97]	3.54E-14	0.95	[0.93	0.97]	1.63E-05
cumulated POI	+1 *infection*	0.84	[0.81	0.87]	<1.00E-16	0.84	[0.78	0.90]	7.05E-07
log(exposure)	+2.72 *days*	2.26	[1.80	2.84]	1.74E-12	2.28	[1.74	2.99]	3.03E-09

- ✓ PMI: *P. malariae* infections
- ✓ POI: *P. ovale* infections
- ✓ Exposure: number of days of presence in the villages by trimester
- ✓ Drug treatment period by *Quinine* was from 1990 to 1994, by *Chloroquine* from 1995 to 2003, by *Fansidar* from 2004 to mid-2006 and by *ACT* from mid-2006 to 2008.

Table 2.3.1.(B): Risk factors identified by GEE in village of Ndiop.

		Analysis considering independence				GEE Analysis			
Variables	***categories***	**Adjusted OR**	**95% CI**		**P-values**	**Adjusted OR**	**95% CI**		**P-value**
Hemoglobin	*Not AA (ref.)*	1	-	-	-	1	-	-	-
	AA	1.33	[1.17	1.51]	7.24E-06	1.32	[1.12	1.55]	8.68E-04
Age group (in years)	≤ 4 *(ref.)*	1	-	-	-	1	-	-	-
	5 *to* 14	0.61	[0.54	0.70]	1.48E-13	0.62	[0.52	0.73]	2.14E-08
	15 *to* 34	0.18	[0.16	0.21]	<1.00E-16	0.18	[0.15	0.22]	<1.00E-16
	≥ 35	0.09	[0.07	0.11]	<1.00E-16	0.09	[0.07	0.12]	<1.00E-16
Drug treatment period	*Quinine (ref.)*	1	-	-	-	1	-	-	-
	Chloroquine	0.59	[0.51	0.68]	2.98E-13	0.58	[0.50	0.67]	3.87E-13
	Fansidar	0.18	[0.15	0.22]	<1.00E-16	0.18	[0.15	0.22]	<1.00E-16
	ACT	0.06	[0.05	0.07]	<1.00E-16	0.06	[0.05	0.07]	<1.00E-16
Semester	*Jan. - Jun. (ref.)*	1	-	-	-	1	-	-	-
	Jul. - Dec.	21.61	[19.21	24.31]	<1.00E-16	21.58	[19.07	24.41]	<1.00E-16
cumulated of PFA	+1 *attack*	1.08	[1.07	1.08]	<1.00E-16	1.08	[1.07	1.09]	<1.00E-16
cumulated of POI	+1 *infection*	0.88	[0.84	0.91]	7.84E-10	0.87	[0.81	0.93]	8.49E-05
log(exposure)	+2.72 *days*	2.29	[1.97	2.67]	<1.00E-16	2.36	[2.00	2.77]	<1.00E-16

2.4 Discussion

All approaches used confirm the general decrease of malaria burden over time and identify almost the same factors underlying the risk of developing *P. falciparum* malaria attacks: increase of age (after 5 years old) led to a decrease of the risk of *PFA*; a decrease RR of *PFA* also occurred from 2004, the year of change in drug treatment from *Chloroquine*, for which malaria parasites developed resistance, to a new and more efficient drug (*Fansidar*), and years after when there was a combined of artemisinin-based combination therapy (ACT) and LLINs. However, different approaches gave slightly different and complementary results.

The two cohorts differ in one very pertinent manner: in Dielmo, malaria transmission occurs all year round because of the presence of a small stream that enables mosquitoes to breed; in Ndiop, transmission occurs only in rainy season from July to December. All methods used confirm this difference in environment between the two villages. Even if environmental factors are much closer between the two cohorts from July to December, we should expect different effects that could be due to genes × environment interactions because of the break of six month in transmission in the 2^{nd} cohort.

When we used different data mining methods, e.g. CART and HyperCube®, variables identified (Age and Year) and their ranges were very similar. Slight differences in results reflect the differences in methodologies of the two techniques. CART uses a sequential approach first splitting the dataset according to the most significant variable and identifying the threshold value of that variable that maximizes the discrimination in the two subsets of data (i.e. least *PFA vs.* most *PFA*). Then, CART will further sub-divide each subset by the next most significant variable that leads to maximum discrimination. This approach thus leads to canalization of the data along different pathways, resulting in a decreased sample size for comparison. The fact that some variables can be used several times at several nodes depending to their importance makes this method to keep in the final tree only variables with strong effects, like Age and Year. In addition, optimization by maximum discrimination at each level may paradoxically lead to an erroneous sub-optimal end-point many levels down. HyperCube®, by contrast, analyses all variables simultaneously with no sequential selection that leads to such loss of power or canalization along a potentially eventual sub-optimal pathway. This aspect unable HyperCube® to catch additional effects of hemoglobin and *P. malariae* infections. Also, the great disadvantage is the impossibility of adjusting on factors making confusion.

While data mining methods keep only variables with strong predictive values in the final results (because of high threshold for effects in HyperCube® and competition between variables at each split in CART), regression methods by GEE can keep factors with weak effects just if they are significant at 0.05 and allows for adjusting on other variables. Another advantage of GEE is the grouping of measures from a same individual; the consequence can be seen as a compromise between the initial sample size (n = total number of episodes from all individuals) and a more realistic sample size (N = number of different individuals). This readjustment of the size is seen in the increase of standard errors of estimates and subsequent increase of P-values from “Analysis considering independence” to “GEE Analysis” (Tables 2.3.1.(A) & (B)).

All these epidemiological aspects of malaria disease discussed in this chapter are important to be understood before genetic analyses presented in the next chapters 3 & 4.

Let us remember the main objectives of the thesis, which are to take into account familial relationships, repeated measures as well as effect of covariates to measure both environmental and host genetic (heritability) impacts on the studied malaria phenotypes, and then use findings from such analyses for linkage and association studies.

Thus, according to these objectives, we have two natural questions. (i) Among this observed variability of malaria disease through these populations, with a great implication of epidemiological variables like age and year periods, what is the overall human genetic contribution? This question will be treated in the following Chapter 3 “Heritability”. (ii) Which of our candidate genes can be suspected to have, independently or jointly, significant

genetic effects on the malaria phenotypes already adjusted on significant epidemiological factors? This question will be treated in the Chapter 4 “Linkage and Association”.

Part II:

Genetic Analysis

3. Heritability

Abstract

In addition to epidemiological factors described in the previous chapter, malaria infection and disease are also strongly influenced by human host factors. To quantify these sources of variation, correlated random effects such as those due to genetic relationships among individuals and repeated measures within individuals should be taken into account in statistical models. Here, we have evaluated the heritability of two *Plasmodium falciparum* malaria parasite phenotypes known to be influenced by human host genetics, the number of clinical malaria episodes or *P. falciparum* malaria attacks (*PFA*) and the proportion of these episodes being positive for gametocytes (*Pfgam*), the specific stages of the parasite responsible for parasite transmission to the mosquito. We performed Generalized Linear Mixed Models (GLMM) that account for familial relationships and repeated measures and have adjusted the models by significant environmental variables identified in the epidemiological analysis, to estimate and separate the variance of the two malaria phenotypes among four sources: host additive genetics (heritability), intra-individual effects or permanent environmental effects including other personal effects like genetics non-additive, house and unexplained residuals. We found a significant additive genetic effect underlying *PFA* during the first drug period of study; this was lost in subsequent periods. There was no additive genetic effect for *Pfgam* analyzed in Dielmo only. By contrast, the intra-individual effect increased significantly. The complex basis to the human response to malaria parasite infection likely includes dominance/epistatic genetic effects encompassed within the intra-individual variance component. There were no house or maternal effects.

3.1 Introduction

After the identification of potential non-genetic variables influencing malaria phenotypes by descriptive methods, estimation of heritability is a second step prior to association studies that use family based methods like allelic transmission counts. The heritability analysis provides an indication of the genetic contribution underlying a specified phenotype and is an important parameter determining the statistical power in gene-mapping studies that use pedigree information. A large heritability implies a strong correlation between phenotype and genotype, so that loci with an effect on the phenotype can be more easily detected (Visscher, Hill et al. 2008). Estimation of the heritability, in this context of family-based longitudinal survey, needs rigorous and adapted statistical model that accounts for repeated measures and disentangles the influence of genetic and environmental factors on the phenotype of interest.

Here, we have collected family data. Therefore, studied individuals are genetically related to each other, so their measured values for the phenotypes are expected to be correlated unless the variability in these values attributable to genetics is null. This chapter presents an extension of GLMM using genetic relatedness among individuals (i) to estimate the effects of covariates free from potential bias induced by non-independence between individuals and (ii) to understand how the phenotypes are genetically and/or environmentally determined by evaluating their variance components. The personal effects of each individual are also evaluated and represent fine phenotypes for genetic linkage and association studies, as these individual effects are already adjusted on potential environmental confusion factors. This extended model, explained in subsection 3.2.2.2 below, will generate appropriate statistics from this family design, e.g. true standard errors of the estimates, independent random individual effects (Vazquez, Bates et al. 2009; Loucoubar, Goncalves et al. 2011).

3.2 Material and Methods

3.2.1 Genetic relatedness

Let us introduce here two main techniques used to quantify genetic relatedness or genetic covariance between relative pairs in a population. The first approach is to use the relationship information from the pedigree and infer kinship between individuals based on the probability of sharing same genomic materials; we will present *inbreeding* and *coancestry* notions which are used to calculate genetic covariance. The second approach that is more accurate is to use pedigree information and individual genotypes to estimate kinship between individuals but relatively to a set of genomic regions for which genotype data are available. This second

method is based on Identity-By-Descent (IBD) of alleles at the considered loci; more the markers loci are dense more the estimates are accurate. A disadvantage of this 2^{nd} method could be the cost for large genotyping coverage in the population to avoid missing data; individuals with missing genotypes at a marker locus will present missing IBD information at that locus for all pairs including those individuals.

3.2.1.1 Pedigree-based genetic relatedness

The Genetic covariance between two individuals can be computed using the pedigree information. For individuals A and B, a given pair in a pedigree, the genetic covariance is computed as $r(A,B) = 2\times coancestry(A,B)$ where the *coancestry* between A and B is calculated referring to the method presented by Falconer and Mackay in 1996 (Falconer and Mackay 1996): $coancestry(A,B) = \Sigma_p(1/2)^{n(p)}\times(1 + I_{Common\ Ancestor})$ where p is the number of paths in the pedigree linking A and B, $n(p)$ the number of individuals (including A and B) for each path p and I_X is the *inbreeding* coefficient of X also equal to the *coancestry* between the two parents of X, I_X is set to 0 if X is a founder.

Illustration: Consider, as an example, the pedigree below (Table 3.2.1 and Figure 3.2.1.(A)) containing 18 individuals named {A, B, …, R} for the calculation of genetic covariance's.

Input: pedigree relationship

Table 3.2.1. Example of simulated pedigree file.

individuals	Father	Mother	Sex
A	.	.	Mal
B	.	.	Fem
C	.	.	Mal
E	A	B	Fem
F	A	B	Fem
H	C	B	Mal
D	.	.	Mal
G	.	.	Mal
I	.	.	Fem
M	D	E	Fem
K	G	F	Mal
L	H	I	Fem
J	.	.	Fem
N	K	J	Mal
O	K	L	Mal
P	.	.	Fem
Q	N	M	Mal
R	O	P	Fem

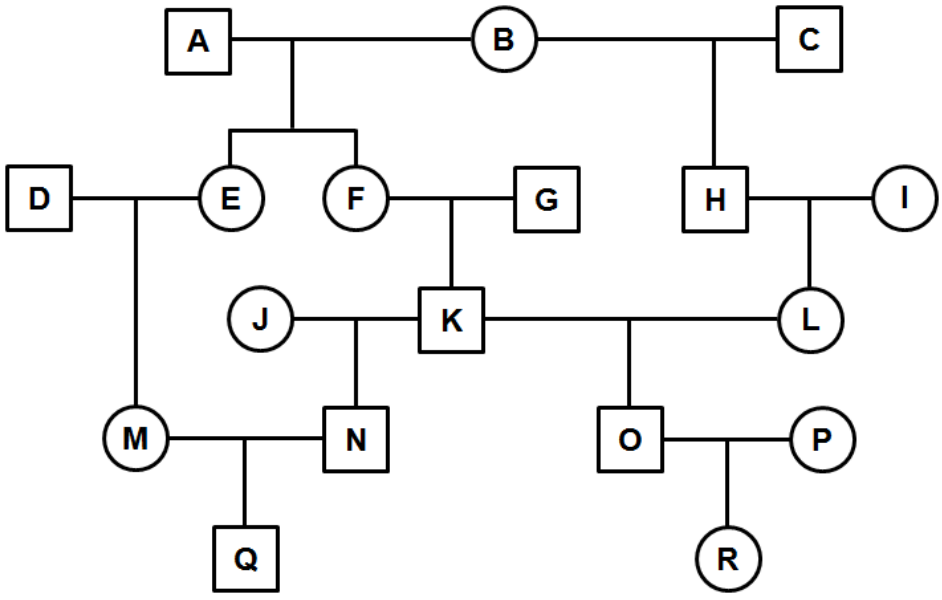


FIG 3.2.1.(A). Pedigree structure derived from Table 3.2.1.

Output: Genetic relationship (or kinship) coefficients derived from the pedigree structure.

Table 3.2.2. Genetic relatedness matrix computes from pedigree structure represented in Figure 3.2.1.(A).

individuals	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**
A	1	0	0	0	0.5	0.5	0	0	0	0	0.25	0	0.25	0.125	0.125	0	0.188	0.063
B	0	1	0	0	0.5	0.5	0	0.5	0	0	0.25	0.25	0.25	0.125	0.25	0	0.188	0.125
C	0	0	1	0	0	0	0	0.5	0	0	0	0.25	0	0	0.125	0	0	0.063
D	0	0	0	1	0	0	0	0	0	0	0	0	0.5	0	0	0	0.25	0
E	0.5	0.5	0	0	1	0.5	0	0.25	0	0	0.25	0.125	0.5	0.125	0.188	0	0.313	0.094
F	0.5	0.5	0	0	0.5	1	0	0.25	0	0	0.5	0.125	0.25	0.25	0.313	0	0.25	0.156
G	0	0	0	0	0	0	1	0	0	0	0.5	0	0	0.25	0.25	0	0.125	0.125
H	0	0.5	0.5	0	0.25	0.25	0	1	0	0	0.125	0.5	0.125	0.063	0.313	0	0.094	0.156
I	0	0	0	0	0	0	0	0	1	0	0	0.5	0	0	0.25	0	0	0.125
J	0	0	0	0	0	0	0	0	0	1	0	0	0	0.5	0	0	0.25	0
K	0.25	0.25	0	0	0.25	0.5	0.5	0.125	0	0	1	0.063	0.125	0.5	0.531	0	0.313	0.266
L	0	0.25	0.25	0	0.125	0.125	0	0.5	0.5	0	0.063	1	0.063	0.031	0.531	0	0.047	0.266
M	0.25	0.25	0	0.5	0.5	0.25	0	0.125	0	0	0.125	0.063	1	0.063	0.094	0	0.531	0.047
N	0.125	0.125	0	0	0.125	0.25	0.25	0.063	0	0.5	0.5	0.031	0.063	1	0.266	0	0.531	0.133
O	0.125	0.25	0.125	0	0.188	0.313	0.25	0.313	0.25	0	0.531	0.531	0.094	0.266	1.031	0	0.18	0.516
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.5
Q	0.188	0.188	0	0.25	0.313	0.25	0.125	0.094	0	0.25	0.313	0.047	0.531	0.531	0.18	0	1.031	0.09
R	0.063	0.125	0.063	0	0.094	0.156	0.125	0.156	0.125	0	0.266	0.266	0.047	0.133	0.516	0.5	0.09	1

The genetic relatedness between individuals N and O is equal to 0.266 from Table 3.2.2. This value is calculated as followed:

The number of paths linking N and O from Figure 3.2.1.(A) is $p = 2$.

- **Path 1** contains $n(1) = 3$ individuals {N, K, O} with K as the common ancestor (Figure 3.2.1.(B)). Inbreeding coefficient of K, I_K , is the *coancestry* between the two parents of K (F and G) and is null because F and G are not genetically linked.
- **Path 2** contains $n(2) = 7$ individuals {N, K, F, B, H, L, O} with B as the common ancestor (Figure 3.2.1.(C)). Inbreeding coefficient of B, I_B , is null because B is a founder.

Therefore, genetic relatedness between individuals N and O is:

$$= 2\times(0.5^{n(1)}\times(1+I_K) + 0.5^{n(2)}\times(1+ I_B))$$

$$= 2\times(0.5^{3}\times(1+0) + 0.5^{7}\times(1+0)) = 0.266$$

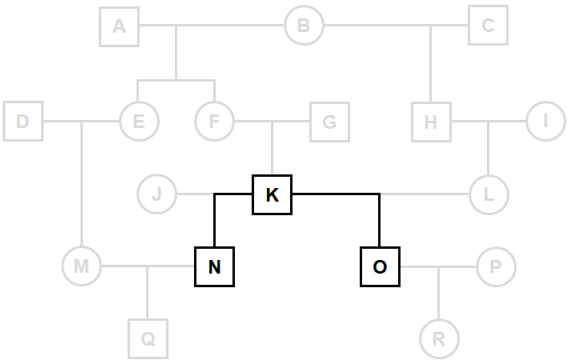


FIG 3.2.1.(B). Path 1 linking N and O.

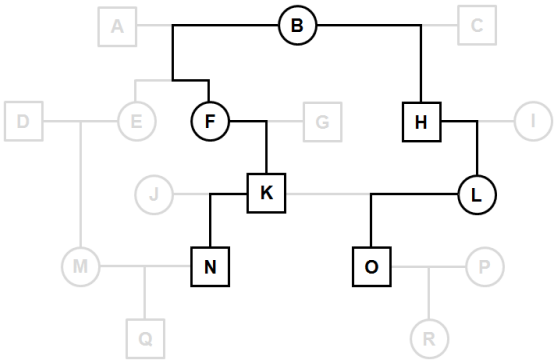


FIG 3.2.1.(C). Path 2 linking N and O.

Remark 3.2.1: In general, the genetic relatedness between parent and child is 0.5, between grandparent and grandson is 0.25, between great grandparent and great grandson is 0.125, and so on, following the series 1/2, 1/4, 1/8, ..., $1/(2^g)$ where g is le number of generations. It is because from generation 1 to g in this kind of direct lineage, the path is unique and the number of individuals making the link goes from $n = 2$ to $n = g+1$ (always the number of current generation +1).

3.2.1.2 IBD-based genetic relatedness

Identity-by-descent (IBD)

Given a pedigree and given a locus, a pair of alleles of two individuals in the pedigree is called identical by descent (IBD) if both alleles have been inherited from a common ancestor (or are "physical copies" of the same founder allele). Remember here that each founder contributes one allele at each given locus, and all non founder alleles are physical copies of founder alleles, the "copying" taking place by segregation during meiosis or a sequence of meioses. IBD-status is determined by the segregation process, not by the nature of the alleles. The two alleles of a single individual are never IBD (unless there is inbreeding in the pedigree) and two individuals may share 0, 1, or 2 alleles IBD, depending on chance and their familial relationship. For instance, a father (respectively a mother) and child always have exactly one allele IBD, if the possibility that the father (respectively a mother) carries the maternal (respectively the paternal) allele of his child is excluded. Thus, a parent and his child always shared 50% of genetic materials at any locus (so in the whole genome). At a locus, a maternal grandmother and grandchild carry 1 gene IBD if the child receives his mother's maternal allele and the child's father is not related to the grandmother (see illustration on Figure 3.2.2 below). The grandmother and grandchild then share 50% of genetic materials at that locus (what arise rarely at many loci simultaneously or if the number of generations between ancestors and descendents increases, due to the transmission of alleles with probability 50/50). This method then joins in some cases the method of Falconer and Mackay, 1996, describe above to measure genetic relationship: IBD approach will be different to the Falconer and Mackay's approach (i) at a locus where the occurrence of allele transmissions is not equilibrated and will be specific to that locus or (ii) if we look at a small number of generations; but will tend to the Falconer and Mackay's approach if we look at a large number of loci simultaneously or in more generations.

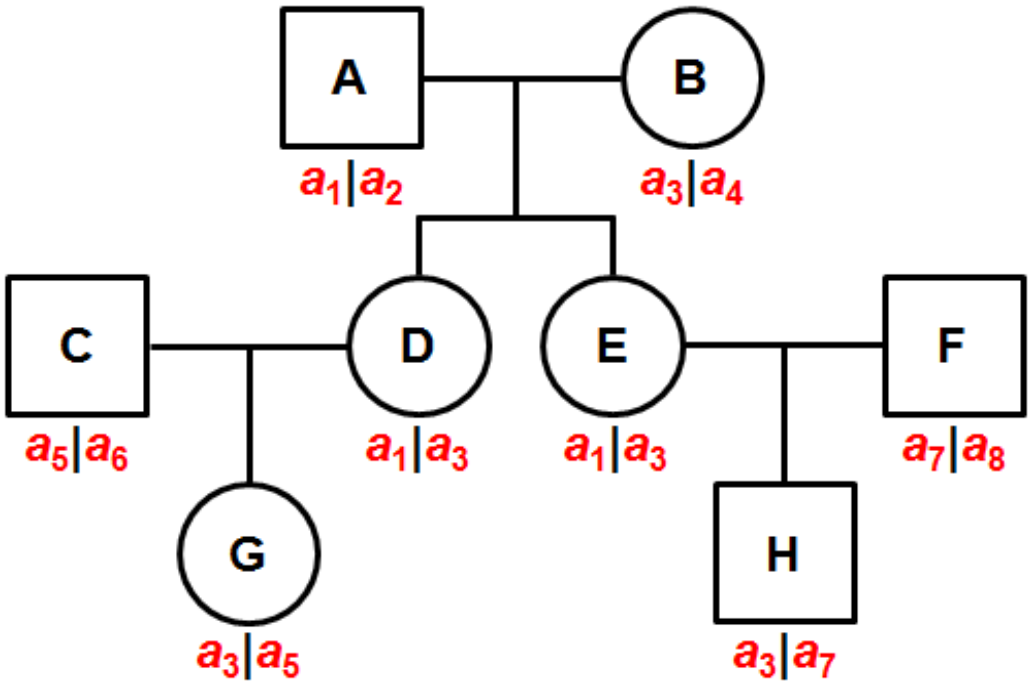


FIG 3.2.2. IBD illustration: Individuals G and H share 1 allele IBD, the allele a_3.

Multipoint IBD can be calculated by MERLIN (Abecasis, Cherny et al. 2002) using genome wide microsatellite genotypes for example. There are three estimated IBD-coefficients between each pair of individuals at each marker: P0 = probability of sharing 0 allele, P1 = probability of sharing 1 allele and P2 = probability of sharing 2 alleles. MERLIN uses information from pedigree, which specifies individual relationships, genotypes and markers location to estimate IBD probabilities. We will not explain here the method used by MERLIN to compute these probabilities; see Abecasis, Cherny et al. 2002 for details on this method. A view of the output file format can be represented as followed:

Table 3.2.3. Presentation of IBD probabilities for each pair of individuals at each marker.

Family	**Individual 1**	**Individual 2**	**Marker**	**P0**	**P1**	**P2**
D1	D1430	D1426	D1S2667	0.00419	0.83123	0.16458
D1	D1430	D1427	D1S2667	0.00083	0.16877	0.83040
D1	D1430	D1433	D1S2667	0.83040	0.16877	0.00083
D1	D1430	D1425	D1S2667	0.00502	0.99498	0
D1	D1430	D1437	D1S2667	0.00083	0.16877	0.8304
D1	D1430	D1430	D1S2667	0	0	1
D1	D9903	D1423	D1S2667	1	0	0
D1	D9903	D9901	D1S2667	0	1	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

We can now use P1 and P2 to define a kinship coefficient, or genetic relatedness, between all relative pairs from genotyped individuals.

Genetic relatedness derived from IBD probabilities

Kinship for a pair of individuals at a marker: The IBD coefficients were computed in each village separately. P1 and P2 are used to measure kinship between two individuals, at a marker *m*, this kinship value is $K_m = P1 \times (1/2) + P2$ and represent the probability of sharing at least one allele identical-by-descent. P1 was divided by 2 because there are two equiprobable ways of sharing one allele identical-by-descent, it can be inherited from the father or from the mother; and we know that when two individuals share one allele identical-by-descent it comes either from the father or from the mother.

Kinship for a pair of individuals through the whole genome: We defined the mean kinship between two individuals in general as the mean of kinship values computed among all markers = $(1/M)\times\Sigma_m K_m = (1/M)\times\Sigma_m(0.5\times P1 + P2)_m$, $m = 1, ..., M$; where M is the number of microsatellite markers.

Remark 3.2.2: The genetic relatedness matrix can be derived, more precisely, from genotypes on genome wide dense SNPs (single nucleotide polymorphisms); then, M is very large making more robust the overall genetic similarity between individuals.

3.2.2 Estimation of covariates effects, individual effects and genetic parameters using Generalized Linear Mixed Models (GLMM) and genetic relatedness matrix

Mixed models are adequate to estimate the effects of explanatory variables on a phenotype in longitudinal survey with case-control design. One problem arises when we are in presence of family data where individuals are genetically linked: their measured values for a given phenotype are expected to be influenced by their correlated random additive genetic effects.

This part presents how to use the additive genetic relatedness matrix derived from the pedigree structure to estimate heritability and to convert the "family design" to an equivalent "case-control" design; and then obtain parameter estimates free from familial correlations. This method has already been applied in several animal breeding models (Henderson 1973; Vazquez, Bates et al. 2009) but not so popular in human genetic studies. For more details concerning general mixed models theory itself, see (Laird and Ware 1982; Henderson 1984; McCulloch 2008).

The name "Mixed Model" comes from the fact that the model contains both fixed effects β parameters, and random-effects γ parameters. Individuals are genetically related to each other, so their measured values for the phenotypes are expected to be correlated unless σ_g^2 (the variability in the phenotype attributable to genetics or the between individual genetic variance) is 0.

3.2.2.1 Design and hypothesis of the GLMM

The design for Mixed Model is the same as the one used for the GEE model in Chapter 2. The y_{ij} , $i = 1, ..., N$ and $j = 1, ..., n_i$, are the measured values for the phenotype of the i^{th} individual

at his j^{th} observation. There are N individuals and n_i measurements on the individual i and $n = \sum_{i=1}^{N} n_i$ total observations. Measured values for the p covariates are stored in a matrix X.

One main additional hypothesis here is the non-independence between related individuals in the studied population. Therefore, we have to take into account both the effect of familial relationships and repeated measurements in the regression models. In the following text, we will use the term "genetic" effects or variances for simplicity but we mean "additive genetic" as we use the information from pedigree to calculate between individuals genetic covariance's.

General formulation of the GLMM

The expectation of the phenotype conditional to the covariates and the random effects is modeled as follows:

$$\begin{aligned} & E(Y \mid \gamma, X, Z) = \mu = l^{-1}(X\beta + Z\gamma + \varepsilon) \\ <=> \quad & l(\mu) = X\beta + Z\gamma + \varepsilon \end{aligned}$$

where Y (n×1) denoted the vector of observed values for the phenotype; μ (n×1) is the expectation of Y conditional to the random effects and the covariates; l is a function that links the expected phenotype μ with a model that is linear in the explanatory variables; β (p×1) is the vector of fixed effects for the covariates; X (n×p) is the design matrix, of rank p, relating fixed effects to μ; γ (N×1) is the vector of random genetic effects of the N individuals; Z (n×N) is the design matrix relating the random effects to μ; ε (n×1) is the vector of random residuals.

Distribution of random genetic effects

For each individual i, the corresponding random genetic effect γ_i is supposed to be normally distributed with mean 0 and variance the unknown between individual genetic variance σ_g^2:

$\gamma_i \sim N(0, \sigma_g^2)$. σ_g^2 is the additive genetic variance component.

Random effects are then identically distributed. However, because of genetic non-independence, for each pair of individuals (i,i') we have cov(γ_i, $\gamma_{i'}$) = σ_g^2×(genetic covariance between i and i') = $a_{i,i'}$×σ_g^2 (= 0 if and only if i and i' are not related). The scalar $a_{i,i'}$ is the element at row i and column i' of A, the genetic relatedness matrix or the matrix of additive genetic covariance's between individuals with dimension (N×N). Genetic covariance's between individuals are derived in this study from the population pedigree structure and stored in a squared matrix A. Therefore, the vector of random genetic effects is distributed as

a Multivariate Normal with mean 0 and covariance matrix $A\sigma_g^2$: $\gamma \sim N(0, A\sigma_g^2)$. Note that if there is no genetic relationship between individuals, A would be equal to I_N, the identity matrix of dimension $N \times N$; and then, the model would be equivalent to a simple mixed model in a context of non-family data.

Distribution of random residuals

The random residuals are supposed to be independent and identically distributed as a Normal with mean 0 and variance the unknown residual variance σ_r^2:

$\varepsilon_{ij} \sim y(0, \sigma_r^2)$. σ_r^2 is the residual variance component.

Then, ε, the vector of random residuals is distributed as a multivariate Normal with mean 0 and covariance matrix $I_n\sigma_r^2$ where I_n is the identity matrix with dimension $(n \times n)$:

$\varepsilon \sim N(0, I_n\sigma_r^2)$.

3.2.2.2 Integrating the genetic relatedness matrix in a family data analysis: How to define an equivalent model design where individual effects are independent

Let us rename $Y^* = l(\mu)$. Y^* can be consider as a linearization of the phenotype through the link function l. The expected mean of Y^* and the variance of Y^* are:

(i) $E(Y^*) = E(X\beta + Z\gamma + \varepsilon)$

$= E(X\beta) + E(Z\gamma) + E(\varepsilon) = X \times E(\beta) + Z \times E(\gamma) + E(\varepsilon)$

$= X\beta$ (asymptomatically).

Proof: Random effects have expected mean equal to 0 as supposed above. In addition, the estimation of β, $\hat{\beta}$ obtained by solving the Henderson's mixed model equations (Henderson 1984), is the "best linear unbiased estimator" (BLUE) if variance components above are known and is "asymptomatically (or empirically) the best linear unbiased estimator" (EBLUE) if variance components above are unknown. Thus, $E(\hat{\beta}) \to \beta$, at least.

Therefore, the expected mean of the phenotype corresponds to the fixed part of the model and is predictable by only observing the covariates and knowing their estimated effects.

(ii) $\mathrm{Var}(Y^*) = \mathrm{Var}(X\beta + Z\gamma + \varepsilon)$

$= \mathrm{Var}(Z\gamma + \varepsilon)$ (as $X\beta$ is the fixed part, thus has variance equal to 0)

$= \mathrm{Var}(Z\gamma) + \mathrm{Var}(\varepsilon)$ (as γ and ε are independent)

$= Z \times \mathrm{Var}(\gamma) \times Z^T + \mathrm{Var}(\varepsilon)$ (Z^T is the transpose of Z)

$= Z(A\sigma_g^2)Z^T + I\sigma_r^2$

$= ZAZ^T\sigma_g^2 + I\sigma_r^2$

If individuals were independent, i.e. $A = I_N$, variance of Y* could be expressed as $ZZ^T\sigma_g^2 + I\sigma_r^2$. However, using linear algebra theory by the method "Cholesky decomposition of a matrix", we can show that there is an equivalent expression of the variance of Y* corresponding to the modeling of data from independent individuals, having γ* as an equivalent vector of random effects and Z* an equivalent design matrix relating γ^* to Y* so that:

$\mathrm{Var}(Y^*) = Z^*(I\sigma_g^2)Z^{*T} + I\sigma_r^2$. $I\sigma_g^2$ is then the covariance matrix of the equivalent independent random individual effects γ^*.

Theorem: *Cholesky decomposition of a matrix*

If A is a symmetric positive-definite matrix, there is a triangular matrix L so that A can be written as $A = LL^T$. L can be seen as the "square root" of the matrix A.

Note that the genetic relatedness matrix A computed using the pedigree information (Falconer and Mackay 1996) is a positive-definite matrix, unless identical twins are in the pedigree in which case it would be positive semi-definite.

Equivalent model with independent random effects: If we set $A = LL^T$ then:

$\mathrm{Var}(Y^*) = Z(A\sigma_g^2)Z^T + I\sigma_r^2$

$= Z(LL^T\sigma_g^2)Z^T + I\sigma_r^2$

$= ZLL^TZ^T\sigma_g^2 + I\sigma_r^2$

$= (ZL)(ZL)^T\sigma_g^2 + I\sigma_r^2$

$= (Z^*)(Z^*)^T\sigma_g^2 + I\sigma_r^2$ (where we set $Z^* = ZL$)

Then, if we define $\gamma^* = L^{-1}\gamma$, we can rewrite the model as:

$Y^* = X\beta + Z^*\gamma^* + \varepsilon$ (because $Z\gamma = Z(LL^{-1})\gamma = (ZL)(L^{-1}\gamma) = Z^*\gamma^*$),

and the γ_i^* are independent, in other terms $Var(\gamma^*) = I\sigma_g^2$, as demonstrated below:

We assumed that $\gamma \sim N(0, A\sigma_g^2)$. Then $\gamma^* = L^{-1}\gamma$ is also distributed as a multivariate Normal with mean $E(\gamma^*) = L^{-1}E(\gamma) = L^{-1}\times 0 = 0$ and variance:

$Var(\gamma^*) = (L^{-1})\times Var(\gamma)\times(L^{-1})^T$

$= (L^{-1})\times A\sigma_g^2\times(L^{-1})^T = (L^{-1})LL^T(L^{-1})^T\sigma_g^2$

$= (L^{-1}L)(L^{-1}L)^T\sigma_g^2$

$= I\sigma_g^2$

The random effects are now independent and then the classical mixed model assuming independence between levels (here individuals) can be applied, and the estimate of fixed effects obtained are fine, i.e. corrected for genetic relationships.

Then, the estimation of fixed effects (effects of covariates) stored in the vector $\hat{\beta}$ and the estimation of random effects (the variance components) stored in the vector $\hat{\gamma}$ are respectively given by:

$$\hat{\beta} = \left(X^T\hat{V}^{-1}X\right)^{-} X^T\hat{V}^{-1}y^*$$

$$\hat{\gamma}^* = \left(\hat{G}Z^{*T}\hat{V}^{-1}\right)\left(y^* - X\hat{\beta}\right)$$

where $\hat{V} = Z^*Z^{*T}\hat{\sigma}_g^2 + I\hat{\sigma}_r^2$, and this illustrates the incorporation of the kinship matrix in the estimation of the effects. See the standard method to solve the mixed model equations (Henderson 1984) for more details on the estimation algorithms.

3.2.2.3 Rewriting GLMM as genetic model

The objective of the model used for the analysis was to estimate and separate different sources of variation underlying the total variation σ_p^2 observed for the phenotype: the relative contributions of human genetics σ_g^2 (additive genetic variance), permanent environment effects σ_{pe}^2, maternal effects σ_m^2, house effects σ_h^2 and unexplained residual variation σ_r^2. The repeated measurements design allows us to separate the two first sources of variation and the occurrence of related individuals living in different houses allows separation of additive genetic variance from that due to shared household.

For reasons of simplicity when writing the algebra in sections above, we presented the case for which the variance of the phenotype was split into genetic and unexplained residual parts only. However, one can explain more by extracting from the residuals, for instance, the permanent environmental, maternal and house effects, or any other evaluable source of variation.

heritability (additive genetic effects)

For a given phenotype in a given population, Heritability (in the broad sense) is by definition the proportion of phenotypic variation that is inherited among individuals. This fraction genetically determined variance is defined as “variance attributable to genetics” divided by “total variance of the phenotype”. In our case for instance, we use the additive approximation through familial relationships, and thus we obtain the fraction of “additively determined variance” or additive heritability (heritability in the narrow sense) equal to:

$$h^{2r} = \sigma_g^2 / \sigma_p^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_{pe}^2 + \sigma_m^2 + \sigma_h^2 + \sigma_r^2).$$

Remark 3.2.3: These variances are measured in a given population, and are dependent on that population. For instance, if a population is genetically very homogeneous, in the extreme case of only one genetic type, then the heritability will be small, because most variation will be environmental. When a single gene is responsible for a disease and the variant of that gene is at fixation, heritability will be zero.

Permanent environmental effects

The random individual effect is included a second time in the model assuming independence between individuals. While the first term will capture the additive genetic variance, this second term will capture the variance between individuals attributable to effects other than additive genetics, e.g., ‘‘permanent environmental’’ effects due to acquired immunity, as well as non-additive genetic effects due to dominance and epistasis (Mackinnon, Gunawardena et al. 2000; Vazquez, Bates et al. 2009). The fraction of variance determined by permanent environmental effects is then equal to:

$$\sigma_{pe}^2 / \sigma_p^2 = \sigma_{pe}^2 / (\sigma_g^2 + \sigma_{pe}^2 + \sigma_m^2 + \sigma_h^2 + \sigma_r^2).$$

Maternal effects

For the individual level, we had the distribution for the vector of random genetic effects as $\gamma \sim N(0, A\sigma_g^2)$ where A reflects the familial relationships between individuals. Using the same approach for the “mother” level, a squared matrix M of dimension the number of mothers reflecting familial relationships between mothers could be derived from the pedigree. Therefore, the vector of random genetic effects for mothers is distributed as a Multivariate Normal with mean 0 and covariance matrix $M\sigma_m^2$: $m \sim N(0, M\sigma_m^2)$. The fraction of variance determined by maternal effects is then equal to:

$$\sigma_m^2 / \sigma_p^2 = \sigma_m^2 / (\sigma_g^2 + \sigma_{pe}^2 + \sigma_m^2 + \sigma_h^2 + \sigma_r^2).$$

House effects

In this step of our study, the two cohorts are analyzed separately and we assume absence of any spatial correlation among houses within a same village. So the vector of random house effects, *c*, contains independent elements and then is assumed to be distributed as a multivariate Normal with mean 0 and covariance matrix $I_H\sigma_h^2$ where I_H is the identity matrix with dimension ($H \times H$): $c \sim N(0, I_H\sigma_h^2)$. The fraction of variance determined by shared house effects is then equal to:

$$\sigma_h^2 / \sigma_p^2 = \sigma_h^2 / (\sigma_g^2 + \sigma_{pe}^2 + \sigma_m^2 + \sigma_h^2 + \sigma_r^2).$$

Residuals variance

The unexplained fraction of variance in the phenotype is equal to:

$$\sigma_r^2 / \sigma_p^2 = \sigma_r^2 / (\sigma_g^2 + \sigma_{pe}^2 + \sigma_m^2 + \sigma_h^2 + \sigma_r^2).$$

Remark 3.2.4: These different variance components are supposed to be independent. Then, the vector of all random effects is assumed to follow a multivariate normal distribution:

$$\begin{pmatrix} \gamma \\ pe \\ m \\ h \\ \varepsilon \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} ; \begin{pmatrix} \mathrm{A}\sigma_g^2 & 0 & 0 & 0 & 0 \\ 0 & \mathrm{I}_N\sigma_{pe}^2 & 0 & 0 & 0 \\ 0 & 0 & \mathrm{M}\sigma_m^2 & 0 & 0 \\ 0 & 0 & 0 & I_H\sigma_h^2 & 0 \\ 0 & 0 & 0 & 0 & I_n\sigma_r^2 \end{pmatrix} \right]$$

I_N is an identity matrix with dimension N, I_H is an identity matrix with dimension the number of houses H, and I_n is an identity matrix with dimension $n = \Sigma_i n_i$, where n_i is the number of measures for individual i.

3.3 Results

For details concerning the findings, see our already published results on the heritability of malaria phenotypes (Loucoubar, Goncalves et al. 2011), presented in the Annex. However, let us present here only main findings.

From 1990 to 2008, four different drug regimens were implemented: *Quinine* from 1990 to 1994, *Chloroquine* from 1995 to 2003, *Fansidar (SP)* from 2004 to mid-2006 and *Artemisinin-based combination therapy (ACT)* from mid-2006 to 2008. The chloroquine drug period was divided into before (CQ1) and after (CQ2) 1999. This was done both to reduce the chloroquine period dataset size and to examine the chloroquine periods prior to and during the observed emergence of parasite resistance to this drug (Noranate, Durand et al. 2007). The statistical analyses were performed independently for each of the five drug treatment periods.

3.3.1 The measured phenotypes

The phenotypes analyzed were: 1) the number of *P. falciparum* clinical episodes, or malaria attacks, during each trimester (*nbPFA*) and units of observation for this phenotype were person-trimesters; 2) the proportion of clinical episodes that were positive for gametocytes, parasite stages transmissible to mosquitoes (*Pfgam*). For *nbPFA* phenotype, we used logarithm of the duration of exposure as offset, therefore results compared between groups the number of *P. falciparum* malaria attacks during each trimester after dividing by the corresponding duration of exposure. These two phenotypes were chosen to be representative of different types of phenotype: *nbPFA* will be strongly influenced by variation in transmission intensity, whereas *Pfgam* will more strongly reflect the host-parasite interaction.

We first excluded any observations of each trimester for which the individual concerned was not present for at least 30 days (=1/3 of the trimester); he or she was considered to be mostly absent. Also, when two clinical episodes were closed, it was probable that most of the observed variability in parasites densities could be attributable to the effect of drug treatment on parasites rather than to human genetics or parasite genetics. Therefore, before statistical analysis, repeated clinical presentations within 15 consecutive days were considered to introduce biais in the study and were excluded from the analyses, unless there was a parasite negative blood smear between two clinical episodes. Only individuals for whom there was pedigree information were included in the analysis.

3.3.2 The covariates

For *nbPFA*, variables found to influence occurrence of clinical malaria episodes in Chapter 2 “Descriptive Methods” were considered as covariates, keeping in final models those significant: sex, age groups, house, season, year (5 categories: 1990 to 1994 for quinine period, 5 categories: 1995 to 1999 for 1st chloroquine period, 4 categories: 2000 to 2003 to the 2nd chloroquine period, 3 categories: 2004 to 2006 for Fansidar period, 3 categories: 2006 to 2008 for ACT period) and logarithm of number of days present in each trimester as offset variable.

For *Pfgam*, we additionally considered the presence of other *Plasmodium* spp. parasites (*P. ovale* and *P. malariae*; 2 categories: yes/no) and time since last treatment. By contrast for *Pfgam*, effect of age was found to be best described when age was a continuous variable in each drug period.

3.3.3 Evolution of heritability of phenotypes with malaria endemicity and drug treatment changes

We applied this specific mixed modeling and estimated the evolution of the variance components with respect to the four successive drug treatment regimens implemented. More details on findings are presented and discussed in Publication 2 “Impact of changing drug treatment and malaria endemicity on the heritability of malaria phenotypes in a longitudinal family-based cohort study” (Loucoubar, Goncalves et al. 2011).

The family structure (pedigree) was available after a demographic census performed for every volunteer at his adhesion in the project. A verbal interview of mothers or key representatives of the household was used to obtain information on genetic relationships between studied individuals, their children, their parents, and to identify genetic links among the population. The total pedigrees, in Dielmo and Ndiop respectively, comprised 828 and 948 individuals, including absent or dead relatives, composed of 206 and 222 nuclear families (father – mother couples with at least one child) with averages of 3.6 and 3.8 children per family.

In addition to calculating the heritability, we estimate the shared environment (here house) and permanent environment effects, including any maternal effects. For each variance component, an estimate was also generated for each individual contributing to the overall component. Thus, for the additive genetic and permanent environment effects, an estimate was established for each person. This predicted individual effect constitutes the individual trend (usually called individual slope) of the phenotype after adjusting on age, transmission season as well as any other significant covariates and also corrected for random variations within individual repeated measurements. Then, individuals can be ranked depending on their personal susceptibility or resistance to the disease; a positive slope corresponds to a positive contribution and a negative slope to a negative contribution to the population’s mean of the phenotype. Therefore, a natural phenotype free from main confounding factors will be this individual trend in the next chapter for genetic linkage and association study. Similarly for house and maternal effects, estimates were established for each house and mother.

3.3.3.1 Studied sample and effects of covariates on number of P. falciparum attacks

The first composite phenotype considered was the number of *P. falciparum* clinical episodes per person per trimester (*PFA*). Over the 19-year study period (1990 to 2008) in Dielmo village, 713 individuals were present between one and 75 complete trimesters generating 22,169 person-trimesters of presence. There were a total of 5,680 clinical *P. falciparum* episodes. In Ndiop village, over the 16-year study period (1993 to 2008), 906 individuals were present between one and 63 complete trimesters generating 20,734 person-trimesters of presence. There were a total of 5,730 clinical *P. falciparum* episodes. The mean (or tendency) of the phenotype is modeled by the fixed part of the mixed model. In both villages, at any drug treatment periods, the number of clinical episodes decreased with age ($P<0.0001$). Year and season also had a consistent influence on the number of clinical episodes ($P<0.0001$) with always a stronger effect of season in Ndiop as expected. The incidence rate of clinical episodes per trimester decreased significantly following the introduction of Fansidar in 2004 as shown in Figures 3.3.1 (A & B); this change in the incidence rate is most evident in the most susceptible age group (<5 years of age in the high and continue transmission area, Dielmo; and <10 years of age in the lower and seasonal transmission area, Ndiop). Results concerning the variance (or fluctuation around tendency) of the phenotype modeled by the random part of the mixed model are variances components presented in Tables 3.3.1 & 3.3.2 below.

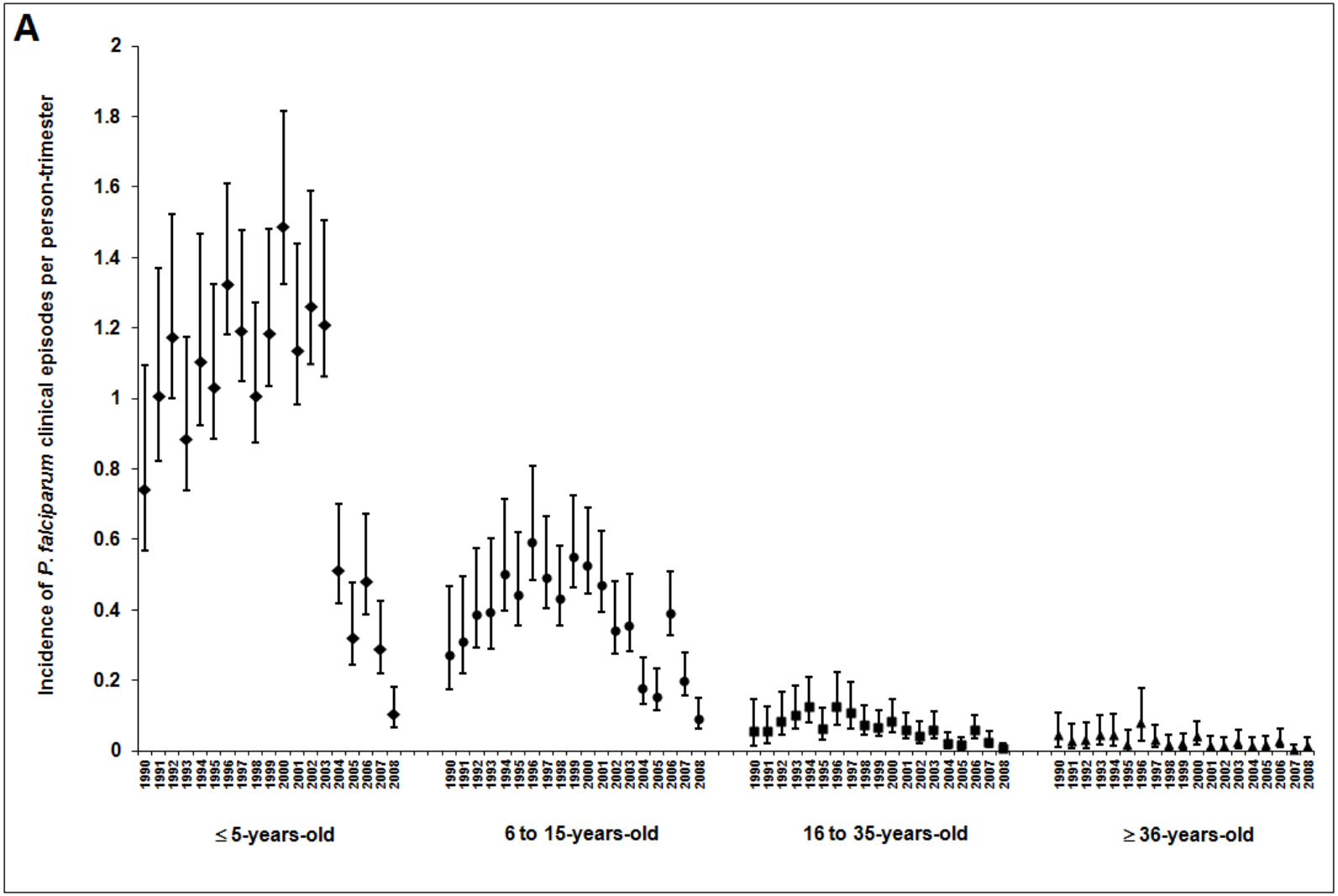


FIG. 3.3.1.(A). The incidence rate (mean ±1.96×SEM) of clinical *P. falciparum* episodes per person-trimester (*PFA*) according to age classes (from left to right on the X-axis) <5, [5–15], [15–35] and ≥35 years that best describe the effect of age on *PFA* in Dielmo.

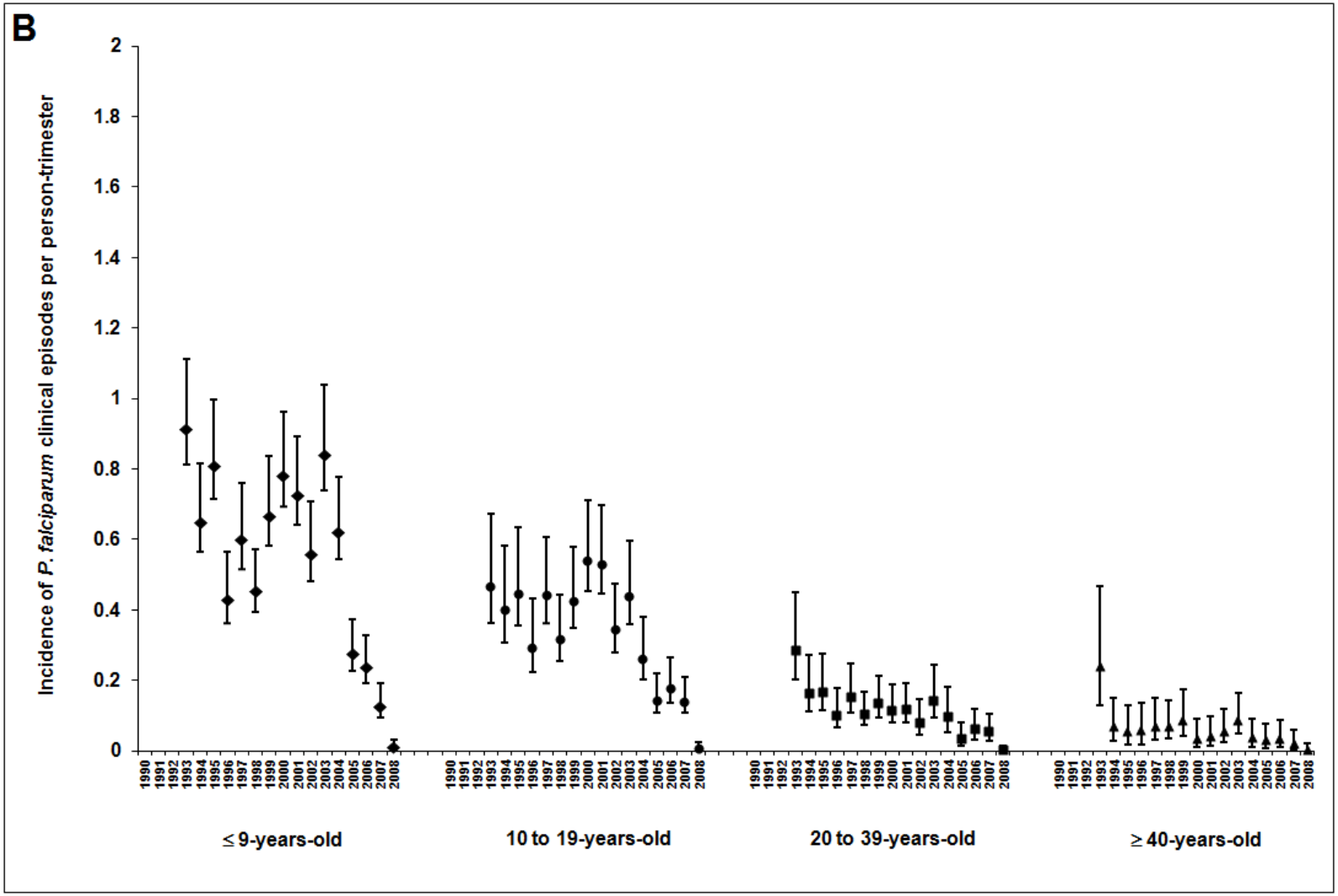


FIG. 3.3.1.(B). The incidence rate (mean ±1.96×SEM) of clinical *P. falciparum* episodes per person-trimester (*PFA*) according to age classes (from left to right on the X-axis) <10, [10–19], [20–39] and ≥40 years that best describe the effect of age on *PFA* in Ndiop.

3.3.3.2 Evolution of heritability for number of P. falciparum attacks

The narrow sense heritability of *PFA* was estimated by drug period. During the quinine period there was significant heritability, estimated at 46%, but which decreased and became non-significant in the subsequent drug treatment periods, in Dielmo village (Table 3.3.1 and Figure 3.3.2 (A) that gives the variance components in percentage). Conversely, the permanent environment effect (PE) increased significantly following the quinine period, accounting for over 50% of the observed variance in *PFA*. There was no house effect during any period (Table 3.3.1 and Figure 3.3.2 (A)).

Table 3.3.1: Variance components of number of *PFA* for village of Dielmo.

Drug period	**var.comp**	**std.err**	**Z**	**Pr > Z**	**95% CI Inf**	**95% CI Sup**
Quinine						
Genetic	0.941	0.384	2.450	**0.014**	0.189	1.693
PE	0.391	0.247	1.580	0.057	0.152	2.343
House	0.030	0.106	0.280	0.390	0.003	8546
residual	0.692	0.016	43.410	<.0001	0.662	0.725
Chloroquine 1						
Genetic	0.257	0.205	1.250	0.211	-0.145	0.658
PE	1.106	0.209	5.300	**<.0001**	0.789	1.664
House	0.039	0.059	0.670	0.252	0.007	85.995
residual	0.603	0.012	50.300	<.0001	0.580	0.627
Chloroquine 2						
Genetic	0.281	0.242	1.160	0.246	-0.193	0.756
PE	1.230	0.229	5.370	**<.0001**	0.880	1.838
House	0.101	0.109	0.930	0.177	0.026	6.787
residual	0.493	0.011	46.870	<.0001	0.473	0.514
Fansidar						
Genetic	0.000	-	-	-	-	-
PE	1.797	0.214	8.380	**<.0001**	1.441	2.304
House	0.036	0.059	0.610	0.272	0.006	392.83
residual	0.395	0.010	41.290	<.0001	0.377	0.415
ACT						
Genetic	0.000	-	-	-	-	-
PE	1.759	0.208	8.450	**<.0001**	1.413	2.250
House	0.125	0.096	1.300	0.098	0.042	1.390
residual	0.357	0.008	43.240	<.0001	0.341	0.374

In village of Ndiop, heritability was not significant from the short survey during the quinine period (1993 and 1995) compared to village of Dielmo (1990 to 1995). During the first half of chloroquine period there was significant heritability, estimated at 19%, but which decreased in the subsequent drug treatment periods; even when it was significant during Fansidar period only, the estimated value was lower (Table 3.3.2 and Figure 3.3.2 (B) that gives the variance components in percentage). The permanent environment effect (PE) was significant during the quinine period, estimated at 15%, decreased during the first years of chloroquine period to 11%, but increased back to 19% during the last years of chloroquine. Both variance components of the phenotype (Genetic and PE) disappear during Fansidar and ACT, periods for which the prevalence of malaria disease was very low in this second village. There was no house effect during any period (Table 3.3.2 and Figure 3.3.2 (B)).

Table 3.3.2: Variance components of number of *PFA* for village of Ndiop.

Drug period	**var.comp**	**std.err**	**Z**	**Pr > Z**	**95% CI Inf**	**95% CI Sup**
Quinine						
Genetic	0.092	0.063	1.460	0.145	-0.032	0.215
PE	0.143	0.067	2.130	**0.017**	0.068	0.474
House	0.000	.	.	.	.	.
residual	0.719	0.023	30.720	<.0001	0.675	0.767
Chloroquine 1						
Genetic	0.253	0.113	2.240	**0.025**	0.032	0.473
PE	0.147	0.088	1.680	**0.046**	0.060	0.764
House	0.032	0.027	1.180	0.119	0.010	0.521
residual	0.934	0.018	51.860	<.0001	0.899	0.970
Chloroquine 2						
Genetic	0.144	0.082	1.760	0.078	-0.016	0.305
PE	0.220	0.070	3.130	**0.001**	0.128	0.464
House	0.020	0.025	0.810	0.208	0.005	4.147
residual	0.786	0.016	49.190	<.0001	0.755	0.818
Fansidar						
Genetic	0.111	0.053	2.090	**0.037**	0.007	0.214
PE	0.000	-	-	-	-	-
House	0.049	0.045	1.090	0.138	0.014	1.187
residual	1.163	0.028	42.210	<.0001	1.111	1.219
ACT						
Genetic	0.031	0.062	0.500	0.618	-0.091	0.154
PE	0.000	.	.	.	.	.
House	0.006	0.031	0.200	0.421	0.001	6.60E+36
residual	1.368	0.032	42.570	<.0001	1.307	1.434

The permanent environment effect (PE) includes, amongst other parameters, any maternal contribution, whether genetic or environmental. In the case of malaria parasite infection, for example, infection during pregnancy can lead to low birth weight with consequent effects on health of the newborn and potentially later in life (Duffy 2007). Thus, as classically performed in heritability analyses, we consequently evaluated the contribution of a maternal effect in addition to the additive genetic and permanent environment effects. There was no maternal effect during any drug period.

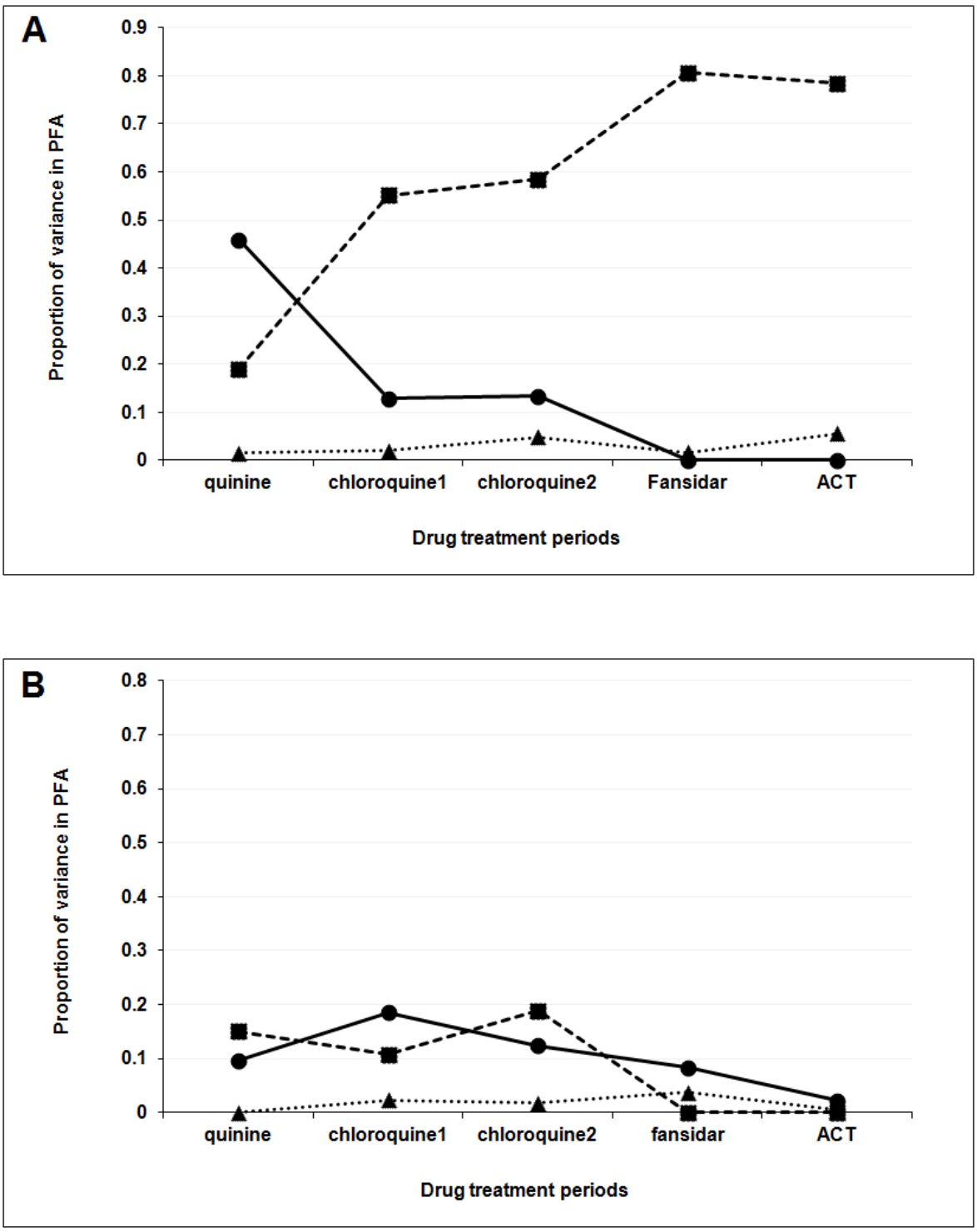


FIG. 3.3.2. Proportion of variance in the number of clinical *P. falciparum* episodes per trimester explained by additive genetic (solid line), intra-individual (dotted line, squares) and house (thin dotted line, triangles) effects in Dielmo (A) and Ndiop (B).

3.3.3.3 Studied sample and effects of covariates on P. falciparum gametocyte positivity

The second composite phenotype considered was the number of *P. falciparum* clinical episodes that were positive for gametocytes, the parasite stages transmissible to mosquitoes. This phenotype was analyzed only in Dielmo due to the lack of positive gametocytes samples in Ndiop due to the low prevalence and seasonal malaria transmission; the small sample size and the high disproportion in the number of observations with presence or absence of gametocytes were note adequate for the non-linear mixed models used here (non convergence of the restricted maximum likelihood algorithm for estimation). In Dielmo, the prevalence of gametocytes at clinical presentation increased from 37% in the quinine period to 48% in the chloroquine periods before decreasing to 17% and 12% in the Fansidar and *ACT* periods respectively (Table 3.3.3 and Figure 3.3.3). The percentage of individuals ever gametocyte positive when having a clinical *P. falciparum* episode likewise increased from 50% in the quinine period to 75% in the second chloroquine period before decreasing to 37% and 25% in the Fansidar and ACT periods respectively. Age, as a continuous variable, was found to be negatively associated with gametocyte presence during the quinine (P=0.02), and the two chloroquine periods (P<0.001). Yearly variation had a significant impact in all periods except ACT. An increasing number of days of individual presence increased gametocyte carriage in the CQ1 period (P=0.02) and increasing time since last drug treatment increased gametocyte carriage in the Fansidar period (P=0.02).

3.3.3.4 Evolution of heritability for P. falciparum gametocyte positivity

Heritability for the prevalence of gametocytes during clinical presentation only approached significance during the Fansidar period (P=0.057), see Table 3.3.3 and Figure 3.3.3 that gives the variance components in percentage. By contrast, the permanent environment effect increased significantly during the chloroquine periods, before becoming non-significant in the Fansidar and ACT periods. There was no house or maternal effects.

Table 3.3.3: Variance components of number of *P. falciparum* gametocyte positivity for village of Dielmo.

Drug period	**var.comp**	**std.err**	**Z**	**P-value**	**95% CI Inf**	**95% CI Sup**
Quinine						
genetic	0.423	0.317	1.340	0.181	-0.197	1.044
PE	0.196	0.272	0.720	0.236	0.040	156.760
House	0.000	.	.	.	.	.
residual	0.932	0.040	23.390	<.0001	0.858	1.015
Chloroquine 1						
genetic	0.164	0.195	0.840	0.401	-0.218	0.545
PE	0.380	0.218	1.750	**0.041**	0.159	1.814
House	0.000	.	.	.	.	.
residual	0.942	0.035	27.300	<.0001	0.878	1.013
Chloroquine 2						
genetic	0.000	.	.	.	.	.
PE	0.530	0.119	4.440	**<.0001**	0.356	0.870
House	0.127	0.090	1.410	0.079	0.045	1.050
residual	0.936	0.031	30.010	<.0001	0.878	1.001
Fansidar						
genetic	0.658	0.346	1.900	0.057	-0.021	1.336
PE	0.000	.	.	.	.	.
House	0.127	0.219	0.580	0.281	0.021	3389.110
residual	0.773	0.055	14.150	<.0001	0.677	0.893
ACT						
genetic	0.570	1.224	0.470	0.641	-1.829	2.970
PE	0.973	1.035	0.940	0.174	0.250	58.229
House	0.070	0.453	0.150	0.439	0.007	2.5E+65
residual	0.593	0.052	11.500	<.0001	0.503	0.708

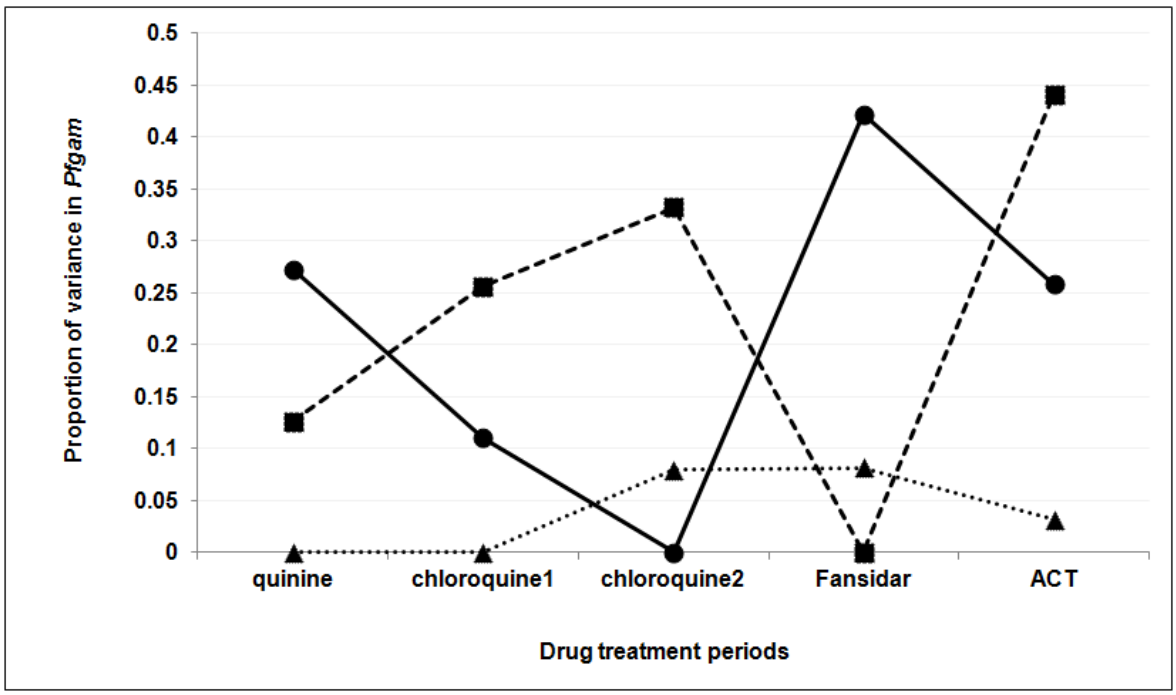


FIG. 3.3.3. Proportion of variance in *P. falciparum* gametocyte positivity explained by additive genetic (solid line), intra-individual (dotted line, squares) and house (thin dotted line, triangles) effects in Dielmo.

3.4 Discussion

Estimation of heritability in its broad sense in natural populations is not possible and hence narrow sense heritability, which estimates the additive genetic contribution, is calculated here. Actual values of heritability are specific for the study populations at a particular time and thus strict comparison is not informative, although broad trends can be inferred. The size of heritability provides an indication of the power to detect the effect of individual genes when performing GWAS. Here it is clear that for several reasons, the choice of the study period for GWAS analysis will affect the quality of the signal. The requirement for large longitudinal datasets to generate sufficient power must therefore be offset by the ever-increasing noise that accompanies long-term datasets – more time means more variance (Lawton 1988).

The peculiarity of the variance component analyses in this study was the replacement of an additive genetic component by a permanent environment component over time. Classical components of permanent environment, such as maternal effects, were not found to be the root cause of this and spatial heterogeneity in exposure seems an insufficient explanation, especially during the quinine and chloroquine periods. There was no significant change in incidence rate, during at least the quinine and chloroquine periods and no difference in the number of different individuals presenting with clinical disease.

From a statistical point of view, insufficient resolution and power of the pedigree matrix may have led to confounding between additive and non-additive genetic components. The replacement of heritability by permanent environment effect could be due just to an important change of genetic relatedness matrix used for the period analyzed. Imagine an individual linked to many others in the cohort such that some individuals have great genetic relatedness only with him and weak relatedness between themselves, as it can be the case for a common grandparent or founder. The absence of this kind of person in the analysis from one period to the next, which can be due to many reasons, would make the sub genetic relatedness matrix concerning individuals analyzed more close to the identity matrix corresponding to the total absence of additive genetics. Hence, all individual effects would be relocated in the permanent environment effects as the total estimate of individual effects stay constant from a model that distinguish between additive genetic and permanent environment to a classic mixed model estimating just the global individual effect. However, in this study it was not the case as the pedigree structure stays stable from period to period (as estimated by the mean genetic relatedness). This suggests that the implementation of the new drug in some way interfered with the human genetic contribution to the outcome of infection.

The loss of an additive genetic effect following implementation of a novel drug treatment may result in significant loss of power to detect genes in a GWA study. Prior genetic analysis of

carefully defined phenotypes, both spatially and temporally delimited, must surely be a pre-requisite for more detailed GWA studies. The temporal changes in the individual genetic and permanent environment estimates are consistent with those expected if there were specific host-parasite genetic interactions. The change in the prevalence of gametocytes at clinical presentation provides additional evidence for there being a change in the parasite population over time. The permanent environment effect contains any non-additive genetic components. The complex, polygenic basis to the human response to malaria parasite infection may well include dominance/epistatic genetic effects that are encompassed within the permanent environment effect. Evaluating their role in influencing the outcome of infection through host genotype by parasite genotype interactions using model systems warrants research effort.

4. Linkage and Association Analysis

Abstract

After the identification of important environmental factors and the evaluation of human factors underlying malaria disease, we performed here genetic studies that focus on candidate genes for susceptibility/ resistance to malaria. We then used family-based methods to test if there was a correlation between alleles' transmission at the genes and the disease status. We used 45 Single Nucleotide Polymorphisms (SNPs) on candidate genes as genetic variables and the adjusted individual effect on PFA as the phenotype of interest. These individual effects, estimated from the Generalized Linear Mixed Models (GLMM) discussed in the previous chapter, represent the individual contributions to the risk of having clinical malaria episode (*PFA*) after adjusting on age and transmission season and also corrected for random variations within individual repeated measurements. Here, we based on an extended Transmission Disequilibrium Test (TDT) for two unlinked disease loci (Morris and Whittaker 1999) and proposed a multi-locus model, more powerful and more adapted, for multifactorial diseases such as malaria, to test for genetic linkage and association simultaneously at any number independent loci. We first detailed the theory of our method and provided simulation studies to compare the power between single locus and multi-locus models in detecting a genetic effect on a phenotype suspected to be influenced by several independent loci. We simulated family data in different configurations depending on the minor allele frequency (MAF) and the sample size. For each configuration, we randomly generated a binary phenotype influenced by each of the simulated loci. In all configurations, the multi-locus models were more powerful to detect genetic effects than the single-locus models. We then applied this method to our real malaria data by analyzing the SNPs one by one in a first step and SNPs showing at least a weak significance (P-value ≤ 0.10) for association with the phenotype were selected in a second step for a multi-locus model that analyzes simultaneous transmission of alleles from those SNPs. Five SNPs showed weak marginal protective effects against malaria after correction for multiple testing: three SNPs on the *SLC4A1* (AE1) gene (Band 3) located on chromosome 17 (ae1_20_21, P = 0.0005; ae1_117_118, P = 0.0598; ae1_174_187, P = 0.0995), one SNP on the γ-globin gene (Xmn1) located on chromosome 11 (Xmn1, P = 0.0598) and one other on the gene *ABO* located on chromosome 9 (abo297, P = 0.0854). We then analyzed these five loci together and obtained stronger protective effect (P-values distributed from 10^{-2} to 10^{-8}) with different combinations of these five loci.

4.1 Introduction

In this chapter we consider family based methods testing deviation from Mendel's Law of allelic inheritance among a sample of offspring. We base on the most widely known method, the Transmission Disequilibrium Test (TDT) for gene-finding (Spielman, McGinnis et al. 1993). Thus, these family based association methods do carry an element of linkage because they make use of related individuals. Ewens and Spielman in 1995 showed that Mendel's Law holds (i.e. equal transmission probability) either when there is no linkage between the marker locus and the disease locus of unknown location, or when there is no association between one specific allele of the marker and the disease's allele (Ewens and Spielman 1995). Therefore, when the null hypothesis of equal transmission probability of alleles is rejected, it is because both linkage and association occur.

In this chapter, we will always adopt as null hypothesis (H_0) and as alternative (H_1) the following:

- H_0: no linkage or no association.
- H_1: association in presence of linkage.

The advantage of these family based methods over regression methods for association is that they give automatic control of confounding: population stratification and/or admixture. The disadvantage is that they require genotyping of cases' parents and more individuals to have power. Usually, it is not possible to have genotypes of cases' parents for a disease that occurs in old ages, but it is not the case in this present study concerning malaria where younger children are the most susceptible to the disease and almost all parents were included in the cohort. Multi-locus family based method can also increase power; several studies using simulated data show more power to detect an effect to a set of loci compared to single locus tests (Ma, Han et al. ; Morris and Whittaker 1999).

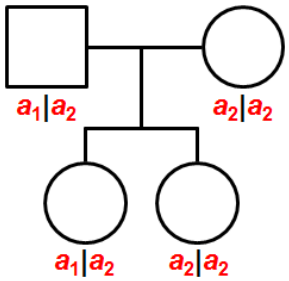
We used some literature from lecture notes by Aad Van der Vaart, 2006 "Statistics in Genetics" (Vaart 2006) and from the book "Handbook of Statistical Genetics", Wiley, 2007 by David J. Balding (Balding, Bishop et al. 2007) to write this chapter. To develop a "Disequilibrium Test for simultaneous transmission of alleles from multiple unlinked multi-allelic loci", we based the work on Sham and Curtis "An extended transmission / disequilibrium test (TDT) for multi-allele marker loci" (Sham and Curtis 1995) and on Andrew Morris and John Whittaker's method for "Generalization of the Extended Transmission Disequilibrium Test (ETDT) to two unlinked disease loci" (Morris and Whittaker 1999).

4.2 Material and Methods

4.2.1 Some useful definitions for linkage and association studies

Mendel's Law of allelic inheritance

If Mendel's Law dictates transmission of alleles, there is equal probability to inherit allele a_1 or a_2 implying equal probability to observe the two genotypes a_1/a_2 and a_2/a_2 among children of two parents having genotypes a_1/a_2 and a_2/a_2.



$$P(a_1/a_2) = P(a_2/a_2) = 0.5$$

Linkage Disequilibrium

An obvious quantitative measure of linkage disequilibrium between loci with alleles a_i and b_j with haplotype frequencies (h_{ij}) and marginal frequencies (p_i) and (q_j) is $D_{ij} = h_{ij} - p_i \times q_j$. These quantities are the difference between the "joint" probability of the alleles at the two loci (the probabilities of the haplotypes a_ib_j) and the probabilities if the loci were independent.

A population is defined to be in "linkage equilibrium" if the alleles at different loci on a randomly chosen haplotype are independent.

Multiple testing

Bonferroni

The Bonferroni correction, when testing many alternative hypotheses at an error rate threshold of α, is to set a new threshold α' corrected for multiple testing such that $\alpha' = \alpha/$(number of tests).

Suppose that we performed m independent tests corresponding to m different alternative hypotheses H_1, H_2, ..., H_m against the same null hypothesis H_0. For example if we test association between a phenotype and m markers M_1, M_2, ..., M_m, we start with the null hypothesis H_0 that any of the makers is associated to the phenotype. While testing each marker for association at a given error rate α, i.e. $P(H_i|H_0)$ = probability to adopt hypothesis i given that H_0 is true = α (the probability to wrongly find marker i positive for the test), an increase in the number of markers tested increase the probability to find at least one of the markers significant, only by chance due to many trials. A natural way to correct this increase in false positive markers is to set a new error α' for each marker such that the probability to have at least one false positive marker is α. Then testing the m markers at an error rate α' will be equivalent to testing a single marker at an error rate of α. Therefore, α' is obtained as followed:

$$\begin{aligned}
\alpha &= P(H_1 \text{ or } H_2, \text{ or ... or } H_m|H_0) \\
&= P(H_1|H_0) + P(H_2|H_0) + ... + P(H_m|H_0) \qquad \text{as } H_1, H_2, ..., H_m \text{ are independent,} \\
&= \alpha' + \alpha' + ... + \alpha' \\
&= m\alpha'
\end{aligned}$$

And then, $\alpha' = \alpha/m$.

False Discovery Rate (FDR)

After performing the m tests as described above, suppose that P are declared positive and N as negative, but in reality m_1 are positive and m_0 are negative as summarized in Table 4.2.1.

Table 4.2.1: Summary of multiples tests

The Truth	Declared Significant by the tests	Not Significant	Total
Null is True	*Fp*	*Tn*	$m_0 = Fp + Tn$
Alternative is True	*Tp*	*Fn*	$m_1 = Tp + Fn$
Total	$P = Fp + Tp$	$N = Tn + Fn$	m

Fp is the number of false positive, *Tp* the number of true positive, *Tn* the number of true negative and *Fn* the number of false negative.

The FDR method provides a control of error rate with a straightforward interpretability for scientists outside of statistics by setting a false discovery rate that satisfy the following condition:

$$\mathrm{E}\left(\frac{Fp}{P} \mid P > 0\right) \leq FDR$$

i.e., given that we obtain a non null number of positive tests, the expectation of error rate which is *Fp*/*P* has to be lower than the FDR. The interpretation is as follows: suppose that *P* tests out of *m* are declared significant at an FDR of 0.05, then 5% of these declarations can be expected to be false positives, on average.

The weak control of FDR proposed by Benjamini and Hochberg in 1995 (Benjamini and Hochberg 1995) follow these three steps:

(i) Order the P-values from the lowest to the highest $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$
(ii) Find the highest rank *k*, let us denoted *k**, that satisfy $P_{(k)} \leq k \times \alpha / m$
(iii)If *k** exists, adopt all hypotheses corresponding to $P_{(1)}, ..., P_{(k^*)}$

Equivalently, we can calculate the adjusted (or corrected) FDR's P-values (P*) as follow:

$P^*_{(m)} = P_{(m)}$

$P^*_{(m-1)} = \min\{P^*_{(m)} ; P_{(m-1)} \times m/(m-1)\}$

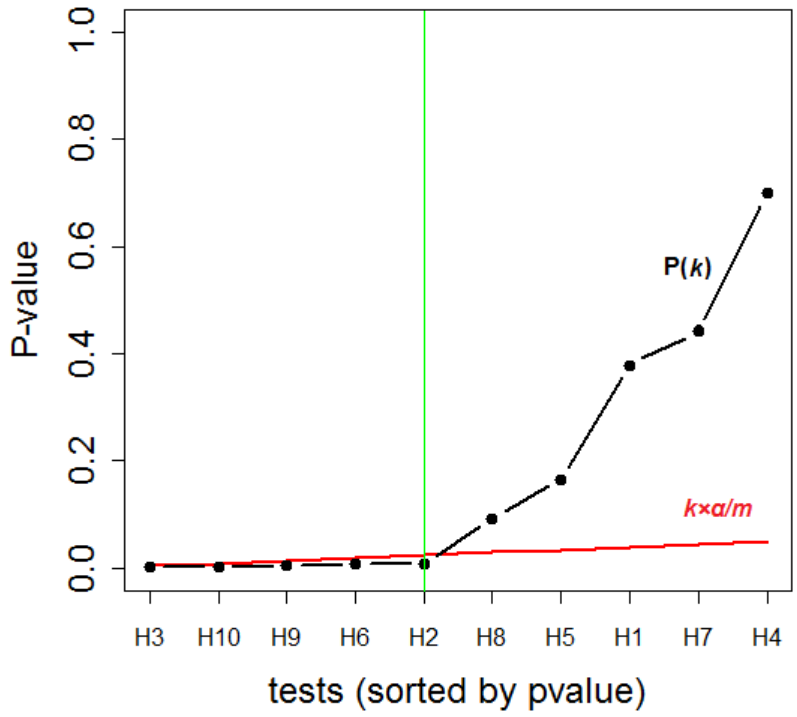
etc. ... until

$P^*_{(1)} = \min\{P^*_{(2)} ; P_{(1)} \times m\}$.

For example $m = 10$ tests and $\alpha = 0.05$ with following P-values:

Tests	P-value
1	0.378
2	0.009
3	0.002
4	0.700
5	0.166
6	0.008
7	0.443
8	0.094
9	0.006
10	0.004

ordered P-value	α	rank k	m	$k \times \alpha / m$	P*
0.002	0.05	1	10	0.005	0.018
0.004	0.05	2	10	0.010	0.018
0.006	0.05	3	10	0.015	0.018
0.008	0.05	4	10	0.020	0.018
0.009	0.05	5	10	0.025	0.018
0.094	0.05	6	10	0.030	0.157
0.166	0.05	7	10	0.035	0.237
0.378	0.05	8	10	0.040	0.473
0.443	0.05	9	10	0.045	0.492
0.700	0.05	10	10	0.050	0.700



In this example, tests 2, 3, 6, 9 and 10 are rejected at a false discovery rate of 0.05.

The Bonferroni correction is a good approximation of what is called the “Family-Wise Error Rate” (FWER) or “Genome-wide Significance Level”. FWER is the probability to obtain at least one false positive result, and is conventionally expected to be equal to 0.05. If α' is the probability for each single test to be found positive wrongly, then:

$$FWER = 1 - \mathrm{P}(\text{number of false positive} = 0 \mid H_0) = 1 - (1 - \alpha')^m \leq \max(m\alpha', 1).$$

However, the Bonferroni approximation of the FWER is less consistent if the *m* tests are not really independent, as could be the case in genome wide studies due to linkage disequilibrium. Another limitation of this method controlling the FWER at 0.05 is its conservativeness; the number of false positive (*Fp*) is evaluated with respect to the total number of tests (*m*) and then not always appropriate for genetic studies where so many genes are often involved.

As an alternative, FDR could be an acceptable way of controlling the inflation of *Fp* in the context of genetic studies by considering the expected number of false positive among the *P* tests declared positive only, instead of referring to all the *m* tests.

4.2.2 Single-locus approach

This section presents linkage and association tests using standard Transmission disequilibrium Test (TDT) (Spielman, McGinnis et al. 1993) and extended TDT (ETDT) (Sham and Curtis 1995) to test markers loci one by one among a set of makers.

4.2.2.1 Transmission Disequilibrium Test (TDT)

The TDT introduced by Spielman in 1993 tests for both linkage and association in families with observed transmissions from parents to affected offspring (Spielman, McGinnis et al. 1993). The TDT can be regarded either as tests of linkage in the presence of association or tests of association in the presence of linkage; in any case we will have linkage and association if the null hypothesis is rejected. The TDT protects against deviations from Hardy–Weinberg equilibrium that could be induced by non-random mating (Balding, Bishop et al. 2007) and is robust against population stratification.

Design for a single biallelic locus

Consider S = {T_1, T_2, ..., T_N} a sample of N trios with affected offspring. Let L be a biallelic locus with alleles coded a_1, a_2. The possible genotypes we can observe among individuals from this sample at this locus are: a_1/a_1, a_1/a_2 and a_2/a_2. Each offspring received two alleles, one inherited from each parent; then, from S we have $2\times N$ transmissions of alleles, N from fathers + N from mothers. However, we will have only n ($\leq 2\times N$) informative transmissions from heterozygous parents: $n = n_{(1)(2)} + n_{(2)(1)}$ where $n_{(1)(2)}$ (resp. $n_{(2)(1)}$) denotes sample frequency for transmission of allele a_1 (resp. allele a_2) from parents having a_1/a_2 genotypes.

Transmission count for one biallelic locus

	$\mathbf{1}_{NT}$	$\mathbf{2}_{NT}$
$\mathbf{1}_{T}$	$n_{(1)(1)}$	$n_{(1)(2)}$
$\mathbf{2}_{T}$	$n_{(2)(1)}$	$n_{(2)(2)}$

The classical Mc Nemar's test

If the disease has nothing to do with the marker locus, then we would expect that heterozygous parents a_1/a_2 transmit a_1 and a_2 alleles with equal probabilities to their affected children. In other words, we expect that the sample frequencies $n_{(1)(2)}$ and $n_{(2)(1)}$ for transmission of alleles are of comparable magnitude. The TDT formalizes this idea by rejecting the null hypothesis of no linkage if $n_{(2)(1)}$ is large relative to $n = n_{(1)(2)} + n_{(2)(1)}$. The test may be remembered as a test for the null hypothesis that given the total number n of heterozygous parents, the number of heterozygous parents who transmit allele a_2 is binomially distributed with parameters n and π =1/2. Under this binomial assumption, given n, the conditional mean and variance of $n_{(2)(1)}$ are $n\pi = n/2$ and $n\pi(1 - \pi) = n/4$, respectively. By applying either the approximation of a binomial by a normal or the Central Limit Theorem we obtain the most popular statistic used for the TDT, the Mc Nemar's statistic:

$$X = \frac{\left(n_{(2)(1)} - n_{(1)(2)}\right)^2}{n_{(2)(1)} + n_{(1)(2)}} \sim \chi^2 \text{ with 1 degree of freedom.}$$

Proof: The approximation of a binomial by a normal states that a random variable distributed as a binomial $B(n, \pi)$ is approximately distributed as a normal with mean and variance equal

to $n\pi$ and $n\pi(1-\pi)$, respectively, under some validity conditions that are: (i) n large, (ii) π not close to 0 or to 1 (the two conditions are translated in practice by $n\pi \geq 5$ and $n(1-\pi) \geq 5$).

Therefore, $n_{(2)(1)} \sim N(n\pi, n\pi(1-\pi)) = N(n/2, n/4)$ under the null hypothesis,

$$\Leftrightarrow \quad \frac{n_{(2)(1)} - n/2}{\sqrt{n/4}} \sim N(0, 1)$$

$$\Leftrightarrow \quad \frac{2\left(n_{(2)(1)} - n/2\right)}{\sqrt{n}} \sim N(0, 1)$$

$$\Leftrightarrow \quad \frac{2 \times n_{(2)(1)} - n_{(1)(2)} - n_{(2)(1)}}{\sqrt{n_{(1)(2)} + n_{(2)(1)}}} \sim N(0, 1)$$

$$\Leftrightarrow \quad \frac{n_{(2)(1)} - n_{(1)(2)}}{\sqrt{n_{(2)(1)} + n_{(1)(2)}}} \sim N(0, 1)$$

$$\Leftrightarrow \quad \frac{\left(n_{(2)(1)} - n_{(1)(2)}\right)^2}{n_{(2)(1)} + n_{(1)(2)}} \sim \chi^2 \text{ with 1 degree of freedom.} \qquad (4.1)$$

The TDT rejects the null hypothesis if $X = (n_{(1)(2)} - n_{(2)(1)})^2/(n_{(1)(2)} + n_{(2)(1)})$ exceeds the appropriate upper quantile of the chi-square distribution with one degree of freedom (equal 3.84 for a type I error 0.05).

Transmission probabilities

Let α_1 be the risk of transmission of allele a_1 and α_2 the risk of transmission of allele a_2 from parents heterozygous at locus L. We can define $\pi_{(1)(2)}$ = Prob (a_1 transmitted | parent genotype is a_1/a_2), the probability of transmitting allele a_1 and not allele a_2:

$$\pi_{(1)(2)} = \alpha_1/(\alpha_1+\alpha_2)$$

and $$\pi_{(2)(1)} = \alpha_2/(\alpha_1+\alpha_2)$$

Likelihood of the transmission model

From the sample of N trios we have $n_{(1)(2)}$ realizations of the event "a_1 is transmitted from a_1/a_2 parents" (denoted "a_1 = T | a_1/a_2") at probability of $\pi_{(1)(2)}$ for each realization; and $n_{(2)(1)}$

realizations of the event "a_2 is transmitted from a_1/a_2 parents" (denoted "a_2 = T | a_1/a_2") at probability of $\pi_{(2)(1)}$ for each realization. Then, the Likelihood of the transmission model is given by:

$$l(\alpha) = \prod_{k=1}^{n} \mathrm{P}(a_1 = \mathrm{T} \mid a_1 / a_2) \times \mathrm{P}(a_2 = \mathrm{T} \mid a_1 / a_2) = \pi_{(1)(2)}^{n_{(1)(2)}} \times \pi_{(2)(1)}^{n_{(2)(1)}}$$

Then, logarithm of the likelihood is:

$$\log l(\alpha) = n_{(1)(2)} \times \log(\pi_{(1)(2)}) + n_{(2)(1)} \times \log(1 - \pi_{(1)(2)}) \quad (4.2)$$

Under the null hypothesis (H_0) of no linkage or no association between disease locus and marker locus, marker alleles are transmitted at random from parents to offspring, regardless of disease status so that $\alpha_1 = \alpha_2 = 0.5$. Thus, the log-likelihood of the null model is given by:

$$\log l_0(\alpha) = n_{(1)(2)} \times \log\left(\frac{1}{2}\right) + n_{(2)(1)} \times \log\left(\frac{1}{2}\right)$$

$$\log l_0(\alpha) = -\log(2) \times (n_{(1)(2)} + n_{(2)(1)})$$

Log-likelihood ratio test

The hypotheses to test for linkage and association between marker locus and disease susceptibility locus of unknown location are:

H_0: $\pi_{(1)(2)} = 0.5$ (or $\pi_{(1)(2)} = \pi_{(2)(1)}$)

H_1: $\pi_{(1)(2)} \neq 0.5$ (or $\pi_{(1)(2)} \neq \pi_{(2)(1)}$)

Ewans and Spielman shown that equal transmission probability occurs either when there is no association, or when there is no linkage between marker and disease, (Ewens and Spielman 1995).

The statistic of the test is given by:

$$X = 2 \times [\max\{\log l_1(\alpha)\} - \log l_0(\alpha)],$$

distributed as a χ^2 with 1 degree of freedom under the null hypothesis. By deriving equation (4.2) we obtain the maximum of the log $l(\alpha)$ when $\pi_{(1)(2)}$ is estimated using sample frequencies (see Figure 4.2.1 below for illustration), i.e. as equal to:

$$\hat{\pi}_{(1)(2)} = \frac{\hat{\alpha}_{(1)(2)}}{\hat{\alpha}_{(1)(2)} + \hat{\alpha}_{(2)(1)}} = \frac{n_{(1)(2)}}{n_{(1)(2)} + n_{(2)(1)}}$$

The null hypothesis is then rejected when the calculated value for X is greater than 3.84, the 95% quantile of the χ^2 with 1 degree of freedom.

For $n_{(1)(2)}$ = 13 and $n_{(2)(1)}$ = 33 (or any chosen values) for instance, we can see that the likelihood of $\pi_{(1)(2)}$ is maximal for $\pi_{(1)(2)}$ = 13/(13+33) = 0.28, by running this R-script just below that screens a sequence of 1000 values for $\pi_{(1)(2)}$ from 0 to 1 and plot the logarithm of the corresponding likelihood (just copy and paste on R-software).

```
# Beginning of the script
n12=13
n21=33
pi12=seq(from=0, to=1, by=0.001)
loglpi12=n12*log(pi12) + n21*log(1-pi12)
plot(pi12,loglpi12, type="l", lwd=5, xlab="pi12", ylab="log-likelihood of
pi12",cex.axis=1.5,cex.lab=1.5)
abline(h=max(loglpi12), col="red", lwd=2)
abline(v=pi12[loglpi12==max(loglpi12)], col="darkgreen", lwd=2, lty=2)
# End of the script
```

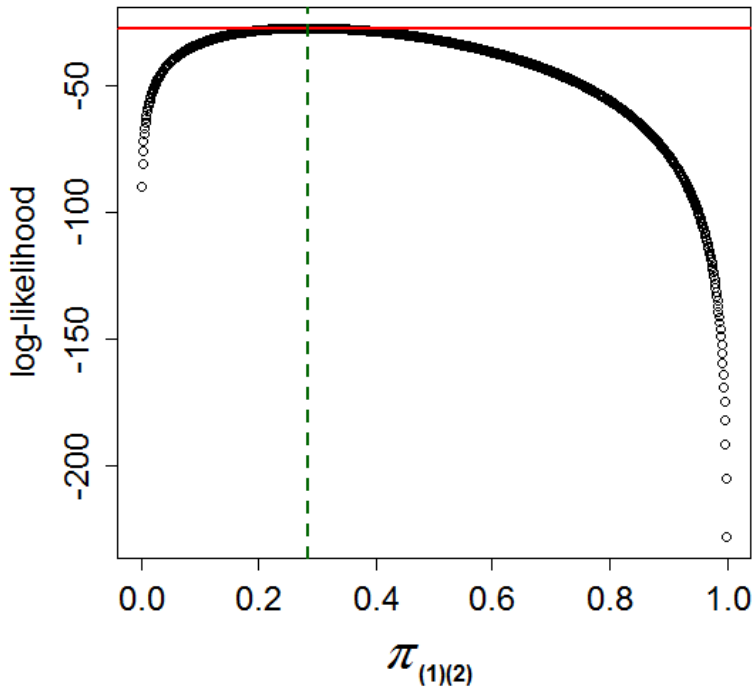


FIG. 4.2.1. Log-likelihood of the transmission model for $n_{(1)(2)}$ = 13 and $n_{(2)(1)}$ = 33, the solid horizontal line is at the maximum of the log-likelihood and the dashed vertical line is at the value of $\pi_{(1)(2)}$ that maximizes the log-likelihood.

4.2.2.2 Extended Transmission Disequilibrium Test (ETDT)

Sham and Curtis developed in 1995 a logistic regression approach that estimates the risk effect for alleles of a microsatellite marker (Sham and Curtis 1995); their method is implemented in the program ETDT.

Like for transmission probabilities in the case of bi-allelic locus, we define in a more general manner $\pi_{(i)(j)} = \alpha_i/(\alpha_i+\alpha_j) = 1-\pi_{(j)(i)}$, if one study a locus that has more than two alleles (as it is the case for microsatellite markers) i and j index the different alleles. If we denote a_1, a_2 ,…, a_l the l alleles of a multi-allelic locus, then the log-likelihood of the transmission model is given by:

$$\log l(\alpha) = \sum_{\substack{i,j=1,\ldots,l \\ i<j}} n_{(i)(j)} \times \log\left(\pi_{(i)(j)}\right) + \sum_{\substack{i,j=1,\ldots,l \\ i<j}} n_{(j)(i)} \times \log\left(1-\pi_{(i)(j)}\right) \quad (4.3)$$

and the log-likelihood of the null transmission model is given by:

$$\log l_0(\alpha) = -\log(2) \times \sum_{\substack{i,j=1,\ldots,l \\ i<j}} \left(n_{(i)(j)} + n_{(j)(i)}\right)$$

The test statistic is $2\times(\log l - \log l_0) \sim \chi^2$ with df = $l - 1$.

For example, for a tri-allelic locus with alleles a_1, a_2, a_3

$$\begin{aligned}\log l(\alpha) = {} & n_{(1)(2)} \log\left(\pi_{(1)(2)}\right) + n_{(2)(1)} \log\left(\pi_{(2)(1)}\right) \\ & + n_{(1)(3)} \log\left(\pi_{(1)(3)}\right) + n_{(3)(1)} \log\left(\pi_{(3)(1)}\right) \\ & + n_{(2)(3)} \log\left(\pi_{(2)(3)}\right) + n_{(3)(2)} \log\left(\pi_{(3)(2)}\right)\end{aligned}$$

and

$$\begin{aligned}\log l_0(\alpha) = -\log(2) \times (& n_{(1)(2)} + n_{(2)(1)} \\ & + n_{(1)(3)} + n_{(3)(1)} \\ & + n_{(2)(3)} + n_{(3)(2)})\end{aligned}$$

and $2\times(\log l - \log l_0) \sim \chi^2$ with df = 2.

There are several other extensions of the basic TDT in the literature: Bickeboller and Clerget-Darpoux (Bickeboller and Clerget-Darpoux 1995), and Spielman and Ewens (Spielman and Ewens 1996) describe extensions for multi-allelic tests. Spielman and Ewens (Spielman and Ewens 1998), Curtis and Sham (Curtis and Sham 1995), Schaid and Li (Schaid and Li 1997), Rabinowitz and Laird (Rabinowitz and Laird 2000), and Fulker et al. (Fulker, Cherny et al. 1999) discuss family tests when parents are missing and/or for general pedigree designs. Martin et al. (Martin, Monks et al. 2000), Horvath and Laird (Horvath and Laird 1998), and Lake et al. (Lake, Blacker et al. 2000) describe methods for general pedigrees that are also valid when testing for association in the presence of linkage. Fulker et al. (Fulker, Cherny et al. 1999), Abecasis et al. (Abecasis, Cardon et al. 2000), Rabinowitz (Rabinowitz 1997), Horvath et al. (Horvath, Xu et al. 2001), and Laird et al. (Laird, Horvath et al. 2000) discuss extensions for quantitative traits.

4.2.3 Multi-locus approach

Analysis methods based on a single SNP have limited power to detect a true genetic effect that requires a combination of specific alleles at several SNPs. In theory it is even possible that two loci might not have "main effects", but do have a joint effect. Including alternatives like multiplicative penetrance or epistasis may make the model more realistic and enable detection of interactions between the loci. This may be detected using haplotype-based methods or multi-locus approaches that consider the joint transmission of alleles at $K = 2, 3, 4, 5$, etc. independent loci, analyzing all SNPs concurrently.

We proposed in this part a generalization of the method proposed by Andrew Morris and John Whittaker for two unlinked loci (Morris and Whittaker 1999) to perform a disequilibrium test for simultaneous transmission of alleles from multiple unlinked multi-allelic loci.

Remark 4.2.1: The K considered loci are not necessary on a haplotype; they can be on different chromosomes. When the loci are on a same chromosome, they should not be in linkage disequilibrium. The advantage of this method is a gain of power through two ways of increasing the sample size:

(i) Nuclear families data (all affected children – father – mother) are considered instead of trios data (one affected child – father – mother). Many offspring of a same family can contribute to the test and the TDT is still valid. The reason is that under the hypothesis of no linkage disequilibrium between the different loci, the transmission or non-transmission of alleles from different loci to each offspring occurs independently.

(ii) In classical TDT, only heterozygous parents contribute to the test. In this multi-locus approach, a parent can be homozygous at many loci from the set of *K*, but will contribute to the test if he/she is heterozygous at least at one locus (for a parent, having two different alleles at one locus is necessary and sufficient to have his set of transmitted alleles different to his set of non transmitted alleles).

4.2.3.1 Design

Consider S = {T_1, T_2, ..., T_N} a sample of *N* trios with affected offspring. As we explained in the remarks above, a nuclear family with several offspring is represented in this sample by as many trios as offspring.

Let L^1, L^2, …, L^K, be *K* independent multi-allelic loci with l_1, l_2, …, l_k alleles respectively. So marker locus L^i has l_i alleles denoted $a_1^i, a_2^i, ..., a_{l_i}^i$. Therefore, the number of possible *K*-tuples of alleles (i.e. a combination set of *K* alleles obtained by sampling one allele from each locus) that we can observe at the *K* loci in the sample is:

$$l = \prod_{k=1}^{K} l_k = l_1 \times l_2 \times ... \times l_K .$$

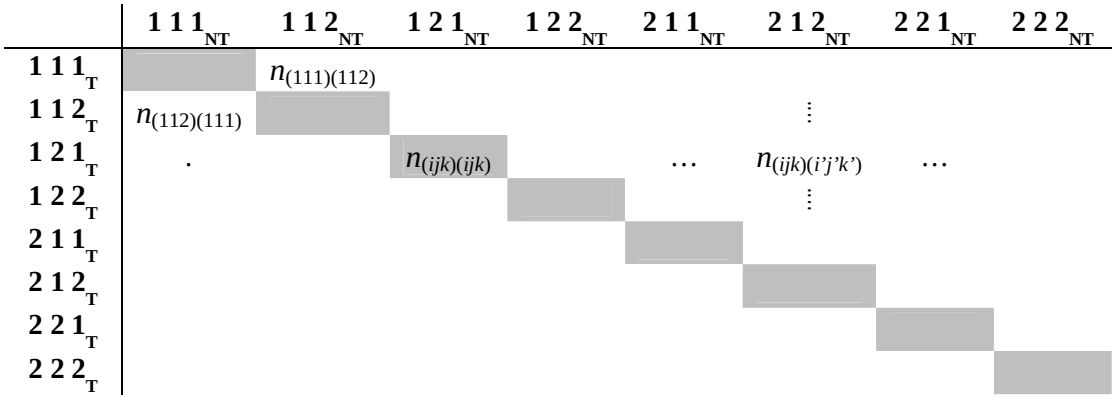
For example if K = 3 loci having 2 alleles each:

	allele 1	**allele 2**
locus1	1	2
locus2	1	2
locus3	1	2

Then $l_1 = l_2 = l_3 = 2$ and the number of possible triplets (or 3-tuples) is $l = 2 \times 2 \times 2 = 2^3 = 8$, and are:

(1 1 1)
(1 1 2)
(1 2 1)
(1 2 2)
(2 1 1)
(2 1 2)
(2 2 1)
(2 2 2)

Transmission count for three biallelic loci



When we consider genotypes at the *K* loci, any sampled offspring has exactly two inherited *K*-tuples of alleles – we are not necessary talking about haplotype – that are $\left(a_u^1 a_v^2 ... a_s^K\right)$ and $\left(a_{u'}^1 a_{v'}^2 ... a_{s'}^K\right)$ where a_u^1 and $a_{u'}^1$ constitute his couple of alleles at locus 1, a_u^1 inherited from one parent and $a_{u'}^1$ from the other parent; $(u,u') \in \{1,2,...,l_1\}$, $(v,v') \in \{1,2,...,l_2\}$, ... ,$(s,s') \in \{1,2,...,l_k\}$. Then, a parent will transmit one *K*-tuple and will not transmit one other. Transmissions from parents homozygous at all of the *K* considered loci, corresponding to parents having two identical *K*-tuples of alleles, are not informative.

For other illustrations related to this multi-locus method, we will show tables for $K = 3$ loci and two alleles each as the number of possible cases and dimension of transmission count tables increase quickly.

4.2.3.1 Simultaneous transmission count

As done for simple TDT, we start by making the squared table that summarizes the transmission counts. Each cell of this table stores the sample frequency for the transmission of one set of *K* alleles while another set of *K* alleles is not transmitted.

Warning: *Under uncertainty on the paternal or maternal origin of an allele, there is no impact at a single locus but it can lead to different choices of transmitted and non transmitted set of alleles at many loci.*

When at the same time father, mother and child are all together heterozygous at one single locus among the *K* loci, it does not have an impact on the transmission count (i.e. the way to fill the squared transmission table) as for the simple TDT method. However, if this situation occurs at two or more loci for a trio there will exist several different ways to fill the table depending on which parent is supposed to give the allele 2 for example (see the illustration below); these are loci of doubt. What we will do is to consider all possible ways as equiprobable. Because for a child there are 2 transmissions, one from each parent, the number 1 has to be divided by the number of possible ways. At the end, for one offspring, the transmission counts from the two parents have to sum to 2.

As we know in genetics and it makes it so nice, counts and number of choices often follow regular sequences. So for this uncertainty on the paternal or maternal origin of an allele there are regular formulae that can be included into the computing scripts to permit automatic dispatching of the 2 transmissions in all possible suppositions that increase with the number of loci of doubt. The strategy we adapt for this kind of trio, father – mother – child are all heterozygous at a number of loci $m \geq 2$, is to replace the child by 2^m fictive children now homozygous for each of the two alleles in question at that *m* loci (so no more doubt for this new children) and keeping the same genotypes at the other loci without doubts. The transmission from one parent to such a fictive child does contribute a count of $1/(2^m)$ instead of one as done for real children. The assignment of genotypes at the loci of doubt for the created children is generating as follows:

By running (just copy and paste) this script on R-software, we obtain the illustration on Table 4.2.2 below for 3 loci of doubt. The number of loci of doubt can be set to any number.

```
# Beginning of the script
nbloci_doubt=3
geno_fictive_children=NULL
for (p in (nbloci_doubt -1):0) {
geno_fictive_children =
c(geno_fictive_children,rep(c(rep("1/1",2^p),rep("2/2",2^p)),2^(nbloci_doubt-p -1)))
}
geno_fictive_children=matrix(geno_fictive_children,nrow=2^nbloci_doubt,ncol=nbloci_doubt)
rownames(geno_fictive_children) = paste("fictive_child",1:2^nbloci_doubt, sep="")
colnames(geno_fictive_children) = paste("locus_doubt",1:nbloci_doubt, sep="")
geno_fictive_children
# End of the script
```

Table 4.2.2: Genotypes generate for 8 created children replacing 1 child who was, as well as his two parents, heterozygous at 3 loci.

	locus_doubt1	**locus_doubt2**	**locus_doubt3**
fictive_child1	1/1	1/1	1/1
fictive_child2	1/1	1/1	2/2
fictive_child3	1/1	2/2	1/1
fictive_child4	1/1	2/2	2/2
fictive_child5	2/2	1/1	1/1
fictive_child6	2/2	1/1	2/2
fictive_child7	2/2	2/2	1/1
fictive_child8	2/2	2/2	2/2

As we can see, these fictive children are all homozygous, thus we will not have with them the problem of trios where all members are heterozygous. The way of assigning their genotypes permits an automatic screening of all possible and equiprobable scenarios that came out with the heterozygous child they replace. To avoid an artificial increase of the sample size, transmission count to each of these fictive children is divided by their number.

Illustrations:

For example if the number of loci of doubt is m = 1, there are only 2^m = 2 possible suppositions contributing equally in the counts, i.e. both suppositions leads to the same count table: supposition 1 is for mother gave allele 2 and supposition 2 is for father gave allele 2.

	father
supposition 1	1 2
supposition 2	1 2

	mother
supposition 1	1 2
supposition 2	1 2

child
1
2

count table for supposition 1

	1=NT	2=NT
1=T		1
2=T	1	

count table for supposition 2

	1=NT	2=NT
1=T		1
2=T	1	

As shown below, we obtain the same contribution in the transmission counts if this child is replaced by 2^m = 2 fictive children homozygous (the first is of marker genotype 1/1 and the second is 2/2, automatically gave by the script above). The transmission from each parent to these 2 new children counts for $1/(2^m)$ = 0.5:

	father
for fictive child 1	1 2
for fictive child 2	1 2

	mother
for fictive child 1	1 2
for fictive child 2	1 2

fictive

child 1	child 2
1	2
1	2

count table

	1=NT	2=NT
1=T		0.5 + 0.5
2=T	0.5 + 0.5	

For example if the number of loci of doubt is $m = 2$, there are $2^m = 4$ possible suppositions contributing now (and for $m > 2$) differently in the counts, i.e. different suppositions can lead to different count tables but that are equiprobable as suppositions are equiprobable.

	father	
supposition 1	1 2	1 2
supposition 2	1 2	1 2
supposition 3	1 2	1 2
supposition 4	1 2	1 2

	mother	
supposition 1	1 2	1 2
supposition 2	1 2	1 2
supposition 3	1 2	1 2
supposition 4	1 2	1 2

child	
1 2	1 2

count table 1

T \ NT	11	12	21	22
11				1
12				
21				
22	1			

count table 2

T \ NT	11	12	21	22
11				
12			1	
21		1		
22				

count table 3

T \ NT	11	12	21	22
11				
12			1	
21		1		
22				

count table 4

T \ NT	11	12	21	22
11				1
12				
21				
22	1			

All the 2^m = 4 suppositions are equiprobable, and then the counts are distributed equally for each one. The child is replaced by 2^m = 4 fictive children homozygous at each of the two loci: child 1 is of genotypes at locus1: 1/1 and locus 2 : 1/1, child 2 is for locus 1: 1/1 and locus 2: 2/2, child 3 is for locus 1: 2/2 and locus 2: 1/1, child 4 is for locus 1: 2/2 and locus 2: 2/2, (automatically gave by the script above). The transmission from each parent to these 4 new children counts for $1/(2^m) = 0.25$:

	father	
for fictive child 1	1 2	1 2
for fictive child 2	1 2	1 2
for fictive child 3	1 2	1 2
for fictive child 4	1 2	1 2

fictive

child 1	
1 1	1 1

child 2	
1 1	2 2

child 3	
2 2	1 1

child 4	
2 2	2 2

	mother	
for fictive child 1	1 2	1 2
for fictive child 2	1 2	1 2
for fictive child 3	1 2	1 2
for fictive child 4	1 2	1 2

count table

T \ NT	11	12	21	22
11				0.25 + 0.25
12			0.25 + 0.25	
21		0.25 + 0.25		
22	0.25 + 0.25			

4.2.3.2 Simultaneous transmission probabilities

Let us denote $\pi_{(u,v,...,s)(u',v',...,s')}$ as the probability that the *K*-tuple of alleles $a_u^1 a_v^2 ... a_s^K$ are transmitted to a child (respectively $\pi_{(u',v',...,s')(u,v,...,s)}$ for the transmission of $a_{u'}^1 a_{v'}^2 ... a_{s'}^K$) , given that the parent is of markers' genotype $\mathrm{L}^1 = a_u^1 / a_{u'}^1$, $\mathrm{L}^2 = a_v^2 / a_{v'}^2$, ..., $\mathrm{L}^K = a_s^K / a_{s'}^K$:

$$\pi_{(u,v,...,s)(u',v',...,s')} = \mathrm{P}\left(a_u^1 a_v^2 ... a_s^K = \mathrm{T}, a_{u'}^1 a_{v'}^2 ... a_{s'}^K = \mathrm{NT} \mid \mathrm{L}^1 = a_u^1 / a_{u'}^1, \mathrm{L}^2 = a_v^2 / a_{v'}^2, ..., \mathrm{L}^K = a_s^K / a_{s'}^K\right)$$

and $\pi_{(u',v',...,s')(u,v,...,s)} = 1 - \pi_{(u,v,...,s)(u',v',...,s')}$; T stands for transmitted and NT for not transmitted.

4.2.3.3 Likelihood of the simultaneous transmission model

If $n_{(u,v,...,s)(u',v',...,s')}$ is the sample frequency of parents transmitting the set of alleles $\left(a_u^1 a_v^2 ... a_s^K\right)$ and not $\left(a_{u'}^1 a_{v'}^2 ... a_{s'}^K\right)$ and $n_{(u',v',...,s')(u,v,...,s)}$ the sample frequency of parents transmitting $a_{u'}^1 a_{v'}^2 ... a_{s'}^K$ and not $a_u^1 a_v^2 ... a_s^K$ from their *K* loci, then the likelihood of the joint transmission model is given by:

$$l\left(\alpha^1, \alpha^2, ..., \alpha^K\right) = \prod_{u<u', v<v', ..., s<s'} \left(\pi_{(u,v,...,s)(u',v',...,s')}\right)^{n_{(u,v,...,s)(u',v',...,s')}} \times \left(\pi_{(u',v',...,s')(u,v,...,s)}\right)^{n_{(u',v',...,s')(u,v,...,s)}}.$$

The log-likelihood is then:

$$\log l\left(\alpha^1, \alpha^2, ..., \alpha^K\right) = \sum_{u<u', v<v', ..., s<s'} n_{(u,v,...,s)(u',v',...,s')} \times \log\left(\pi_{(u,v,...,s)(u',v',...,s')}\right) + \sum_{u<u', v<v', ..., s<s'} n_{(u',v',...,s')(u,v,...,s)} \times \log\left(1 - \pi_{(u,v,...,s)(u',v',...,s')}\right)$$

where α^i is the vector containing the α_j^i's, and each α_j^i corresponds to the single locus risk of transmission of a_j^i, (*j*-th allele of locus *i*) among parents heterozygous for that allele.

Under the null hypothesis of no linkage or no association between the *K* independent makers loci and the *K* independent and unknown disease loci, the *K*-tuples of alleles are transmitted at random from parents to affected offspring so that $\pi_{(u,v,...,s)(u',v',...,s')} = \pi_{(u',v',...,s')(u,v,...,s)} = 1/2$.

The log-likelihood of the null model is then:

$$\log l_0 = -\log(2)\times \sum_{u<u',v<v',\ldots,s<s'} \left(n_{(u,v,\ldots,s)(u',v',\ldots,s')} + n_{(u',v',\ldots,s')(u,v,\ldots,s)}\right)$$

The test statistic is $2\times(\log l(\alpha^1,\alpha^2,\ldots,\alpha^K) - \log l_0) \sim \chi^2$ with a number of free parameters that will depend on the alternative model to test as explained in the following part.

4.2.3.4 The different alternative hypothesis

Now, there are several ways to define the probability to transmit jointly a set of *K* alleles depending on the alternative model we want to test between the *K* markers loci and the *K* disease loci of unknown location. Therefore, the computed likelihood is specific to the alternative one would like to test.

(1) ***Only one of the K markers loci is expected to be linked to one of the disease loci:***

When we want to test for linkage of marker locus *i* to a disease locus ignoring information from the other markers the transmission probabilities are given by:

$$\pi_{(\ldots,j,\ldots)(\ldots,j',\ldots)} = \frac{\alpha^i_j}{\alpha^i_j + \alpha^i_{j'}}$$

among parents having genotype *j*/*j'* at that locus *i* and transmitting *j*, whatever the alleles they transmit elsewhere (*j* and *j'* index the alleles $a^i_1, a^i_2, \ldots, a^i_{l_i}$ of the locus *i*). For example, if we want to test locus 1 only, we have $\pi_{(u,v,\ldots,s)(u',v',\ldots,s')} = \alpha^1_u / \left(\alpha^1_u + \alpha^1_{u'}\right)$ for any heterozygous *u*/*u'*.

The number of free parameters for the corresponding log-likelihood (i.e. the log-likelihood computed using this definition of π) is $l_i - 1$, the number of alleles of locus $i - 1$. This model is exactly the model for a single locus extended TDT (ETDT).

(2) ***A number p (p = 2, 3, …, K) of markers among the K markers loci are expected to be linked, one each, to p of the disease loci:***

When we want to test for linkage of several marker loci to several disease loci, there are two main assumptions:

(a) If we assume multiplicative penetrance across disease loci without interaction (no epistasis), then the risk of transmission of a set of *p* alleles from the *p* markers loci is the product of their marginal risks. Then, the joint transmission probabilities are given by:

$$\pi_{(u,v,\ldots,s)(u',v',\ldots,s')} = \frac{\alpha_u^1 \times \alpha_v^2 \times \ldots \times \alpha_s^p}{\left(\alpha_u^1 \times \alpha_v^2 \times \ldots \times \alpha_s^p\right) + \left(\alpha_{u'}^1 \times \alpha_{v'}^2 \times \ldots \times \alpha_{s'}^p\right)}$$

The number of free parameters for the corresponding log-likelihood is:

$$(l_1 - 1) + (l_2 - 1) + \ldots + (l_p - 1).$$

(b) If we allow for interaction (epistasis) between the disease loci, then the risk of transmission of a set of p alleles from the p markers loci denoted $\gamma_{u,v,\ldots,s}^{1,2,\ldots,p}$ has to be derived from the joint transmission counts' table for set of alleles. The joint transmission probabilities are then given by:

$$\tilde{\pi}_{(u,v,\ldots,s)(u',v',\ldots,s')} = \frac{\gamma_{u,v,\ldots,s}^{1,2,\ldots,p}}{\gamma_{u,v,\ldots,s}^{1,2,\ldots,p} + \gamma_{u',v',\ldots,s'}^{1,2,\ldots,p}}$$

The number of free parameters for the corresponding log-likelihood is:

$$l_1 \times l_2 \times \ldots \times l_p - 1.$$

(a) versus (b): One can test if there is significant deviation from a multiplicative model without interaction to a model with interaction when both are more likely than the null model by using their likelihood's ratio having an approximate chi-squared distribution with the difference of free parameters as the number of degrees of freedom.

Remark 4.2.2: As we can observe, the number of alternative hypotheses increases quickly with the number of loci analyzed as shown in Figure 4.2.2. If we study *K* markers, the number of alternatives is given by:

m = number of models with single locus + number of models with two loci assuming no epistatis + number of models with two loci assuming epistatis + number of models with three loci assuming no epistatis + number of models with three loci assuming epistatis + …etc., until the models with all the *K* loci. Then we have:

$$m = K \; + \; 2\times \sum_{p=2,\ldots,K} C_K^p$$

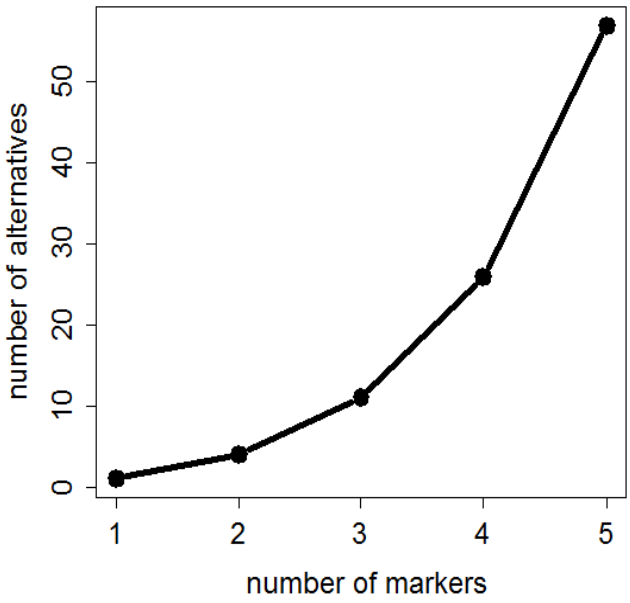


FIG 4.2.2. Number of alternative models by the number of loci tested simultaneously.

To adopt one of the alternatives, the corresponding raw P-value has to be lower than a corrected threshold set by the *Bonferroni* method (i.e. 0.05/*m*) or by the *False Discovery Rate* (*FDR*) method. Equivalently, the adjusted P-values can be compared to 0.05, for example, adjusted P-values by FDR as described in the beginning of this chapter.

4.3 Results

The principle of this method is to compute the table for the transmission of sets of alleles and subsequently compute the likelihoods from that table. One cannot start by computing the 2-by-2 table for alleles of each single locus before, to then deduce the transmission of sets. However, the 2-by-2 transmission table for each given locus can be derived from the multi-locus table by summing over rows and over columns pertinent for that locus. Thus, the single locus models we obtained are derived from the full multi-locus model.

4.3.1 Comparison with results from Family Based Association Test Software (FBAT)

To validate the computations, we used simulated data and compared results of single locus models derived from the multi-locus model to results from single locus model by FBAT Software (Horvath, Xu et al. 2001), a commonly used method for linkage and association analysis. The equivalent TDT model in FBAT was used and without inferring any genotype for an individual (i.e. additive model, transmissions from parents to their affected offspring only are considered, and there are no missing genotypes). For this we simulated 100 trios on three SNPs having minor allele frequencies (MAF) of 0.30. At each SNP we gave random genotypes to a parent by sampling with replacement twice an allele from alleles {1, 2} with occurrence probabilities of 0.70 to take allele 1 and 0.30 to take allele 2. We then sampled one allele from each parent, with probabilities 50/50, to give a random genotype to an offspring. Next we simulated a binary trait, “0” no disease and “1” for disease, associated weakly to each of the three SNPs, by sampling “disease” with higher probabilities for offspring carrying allele 2 . To generate these data, we can run the R script in Annex B. At the end of the script two “.txt” files are saved, the first one is in a format to be analyzed by the R script in Annex C for multi-locus transmissions and the second is in a format for analysis on FBAT after additional changes in the file format and column names (see FBAT’s manual for users).

Table 4.3.1: Result from multi-locus model.

model		log-likelihood	X	DF	P
SNP-1		-21.47	4.19	1	0.0408
SNP-2		-22.74	1.66	1	0.1980
SNP-3		-22.21	2.72	1	0.0992
SNP-1-2	Multiplicative	-20.82	5.50	2	0.0639
SNP-1-3	Multiplicative	-19.33	8.48	2	0.0144
SNP-2-3	Multiplicative	-21.59	3.94	2	0.1391
SNP-1-2	Epistasis	-21.03	5.07	3	0.1670
SNP-1-3	Epistasis	-18.48	10.17	3	0.0172
SNP-2-3	Epistasis	-21.08	4.97	3	0.1738
SNP-1-2-3	Multiplicative	-18.94	9.26	3	0.0260
SNP-1-2-3	Epistasis	-16.88	13.37	7	0.0636

Table 4.3.2: Result from FBAT Software

Marker	Allele	Allele frequency	Informative families	S-E(S)	Var(S)	Z	P
SNP-1	1	0.665	13	-4	4.0	-2.00	0.0455
SNP-1	2	0.335	13	4	4.0	2.00	0.0455
SNP-2	1	0.730	15	-3	5.5	-1.28	0.2008
SNP-2	2	0.270	15	3	5.5	1.28	0.2008
SNP-3	1	0.675	16	-4	6.0	-1.63	0.1025
SNP-3	2	0.325	16	4	6.0	1.63	0.1025

As shown in Tables 4.3.1 and 4.3.2, the results from the two methods are the same for testing one SNP at a time. The multi-locus approach on FBAT is to test association of the disease to a haplotype. It assumes that SNPs are closed on the same haplotype (no recombination). This cannot be compared to our multi-locus approach, which assumes that SNPs are independent, i.e. they can be on different chromosomes and should be far away if they are on a same chromosome (recombination is allowed).

4.3.2 Power study

To compare the power between single locus and multi-locus models to detect genetic effect on a phenotype suspected to be influenced by several independent loci, we simulated 2500 different samples of trios (father – mother – child) and each time on three bi-allelic loci. It consists of 100 repetitions for each of these 25 following configurations: 5 different minor allele frequencies (MAF = 0.10, 0.20, 0.30, 0.40, 0.50) by each of the 5 different sample sizes (100, 200, 300, 400, 500 trios). For each of the 2500 simulations, we generated a random binary phenotype which is influenced by each of the three loci and performed the models for 1, 2 and 3 loci. Figures 4.3.1 – 5 below compare the distributions of P-values between the different models at different settings. The horizontal dashed lines for each type of model (1, 2, and 3 loci) are plotted at 95% quantile of the P-values to avoid comparing outliers

In all configurations, the 3-loci models are more powerful to detect genetic effects than the 2-loci models and the 2-loci models more powerful than the single-locus models (Figures 4.3.1 – 5).

The R script to simulate trios and perform models and plot the P-values is available in Annex C (**warnings**: set a low number of repetitions or fewer configurations before running, otherwise it can take several hours).

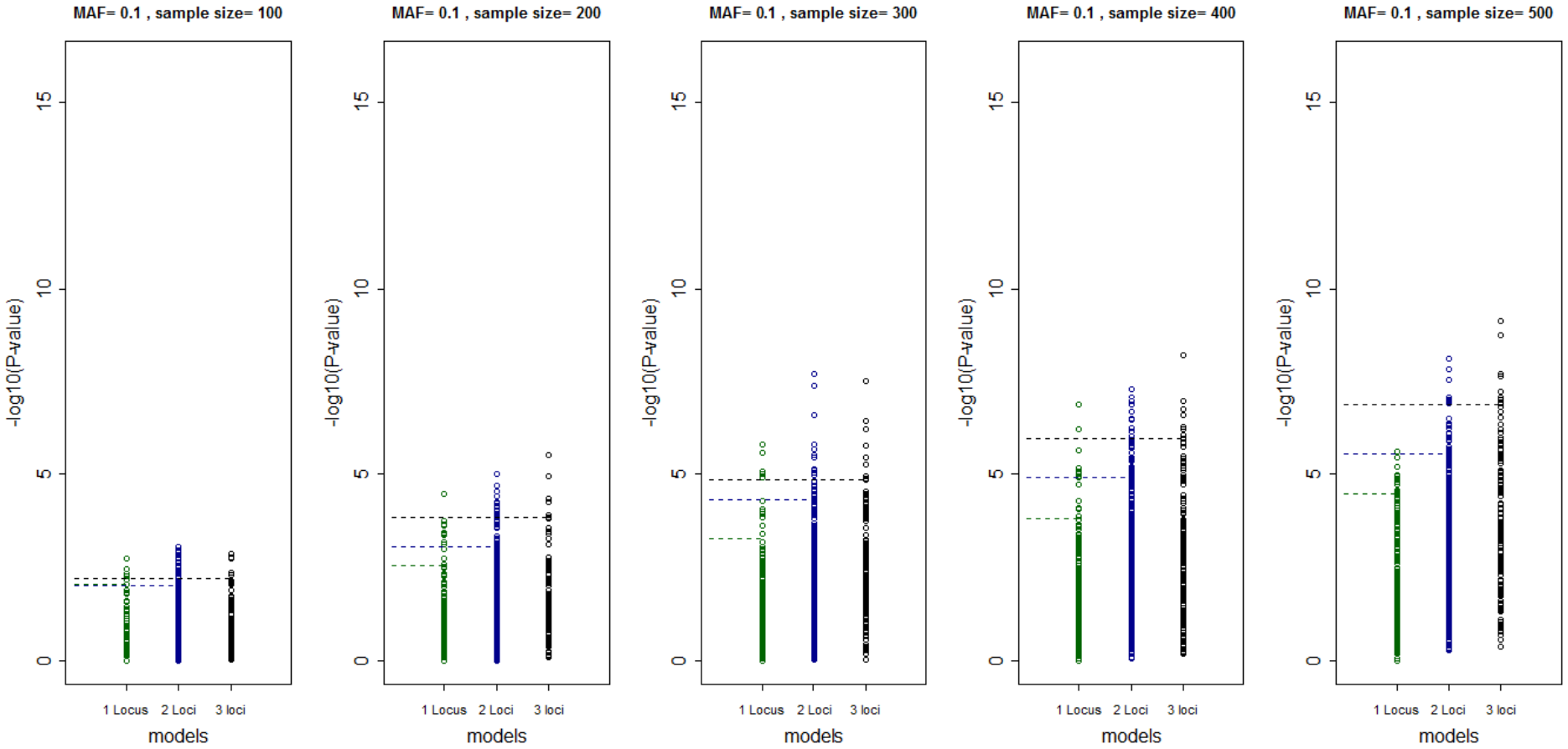


FIG. 4.3.1. Comparison of P-values between single locus and multi-locus for MAF = 0.10 and at sample size of 100, 200, 300, 400, and 500.

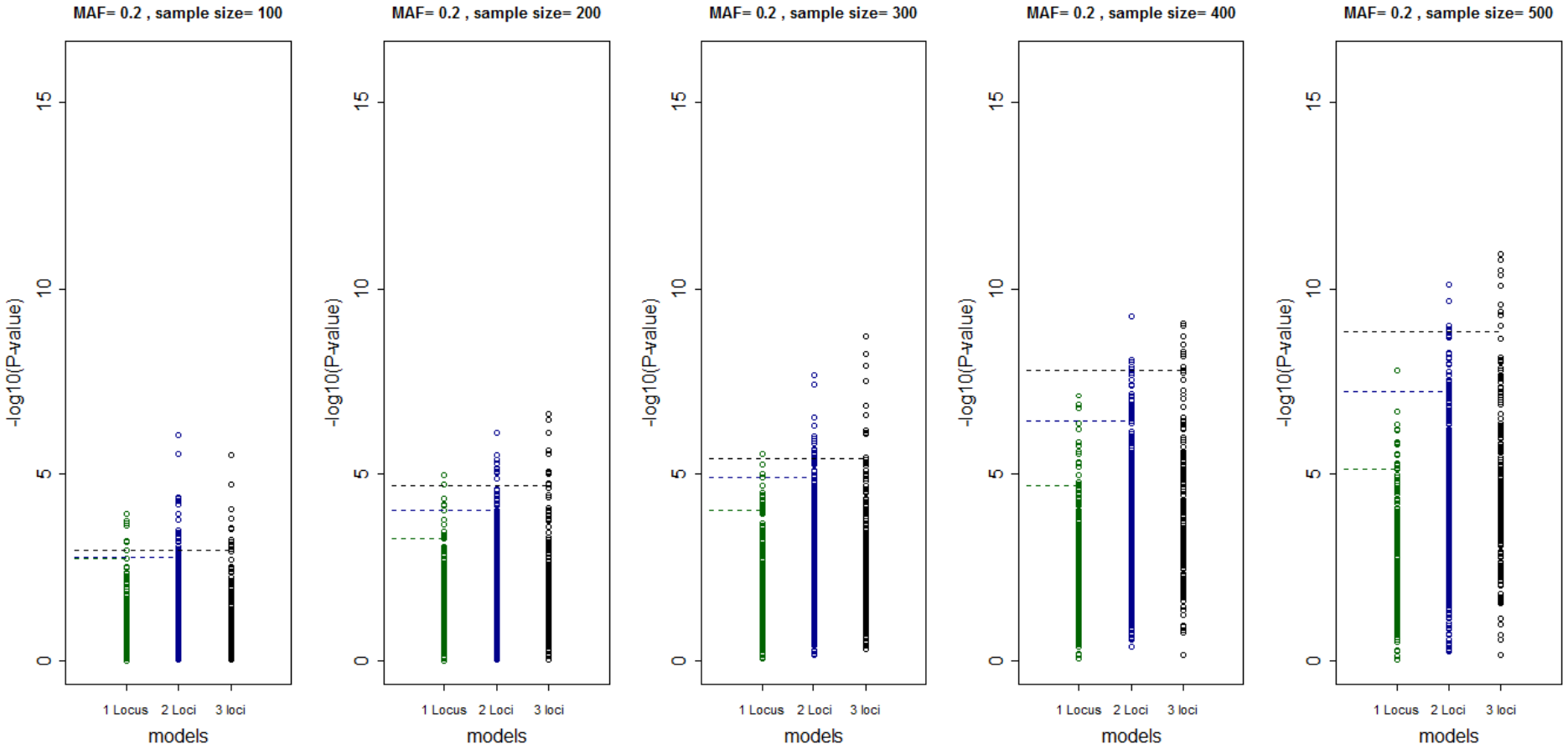


FIG. 4.3.2. Comparison of P-values between single locus and multi-locus for MAF = 0.20 and at sample size of 100, 200, 300, 400, and 500.

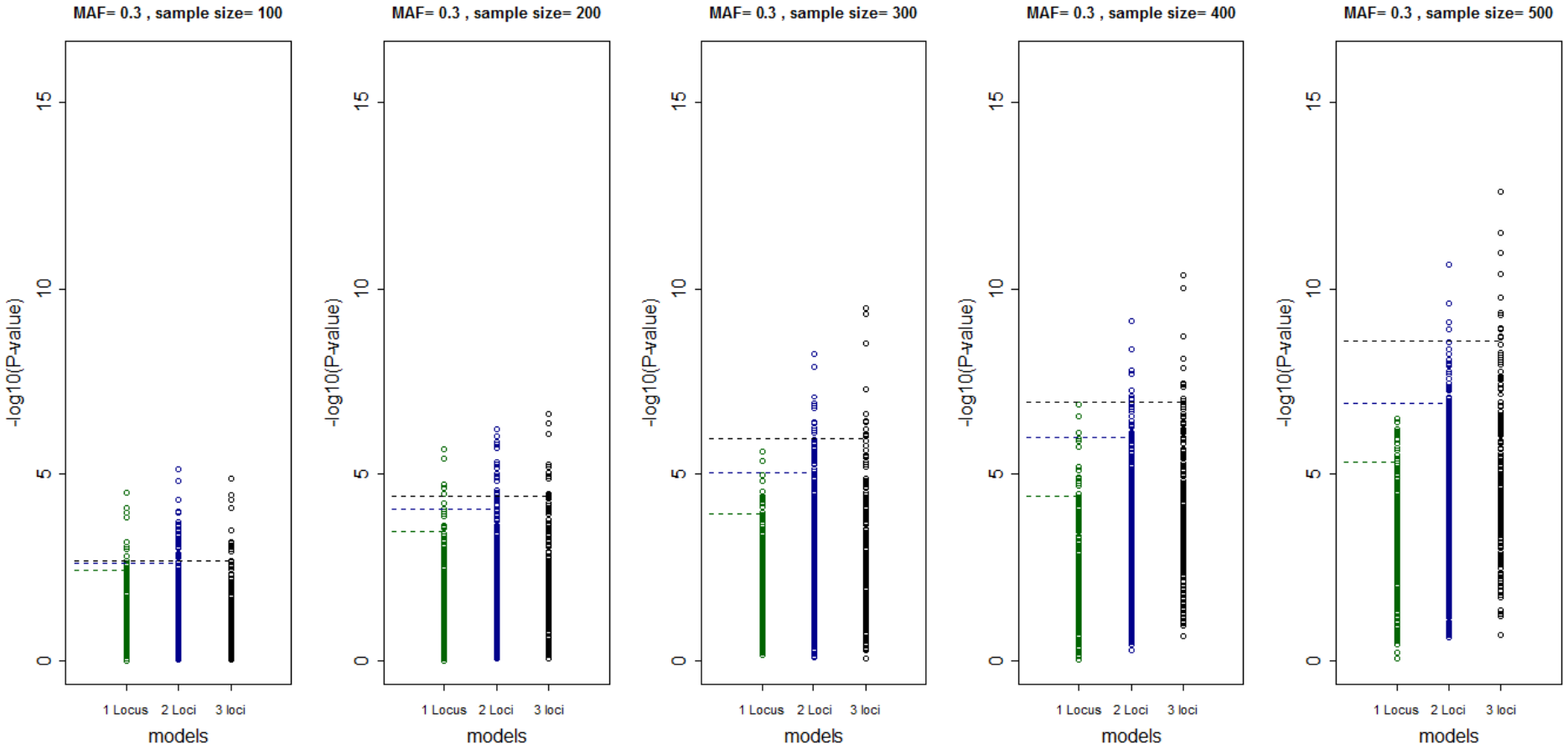


FIG. 4.3.3. Comparison of P-values between single locus and multi-locus for MAF = 0.30 and at sample size of 100, 200, 300, 400, and 500.

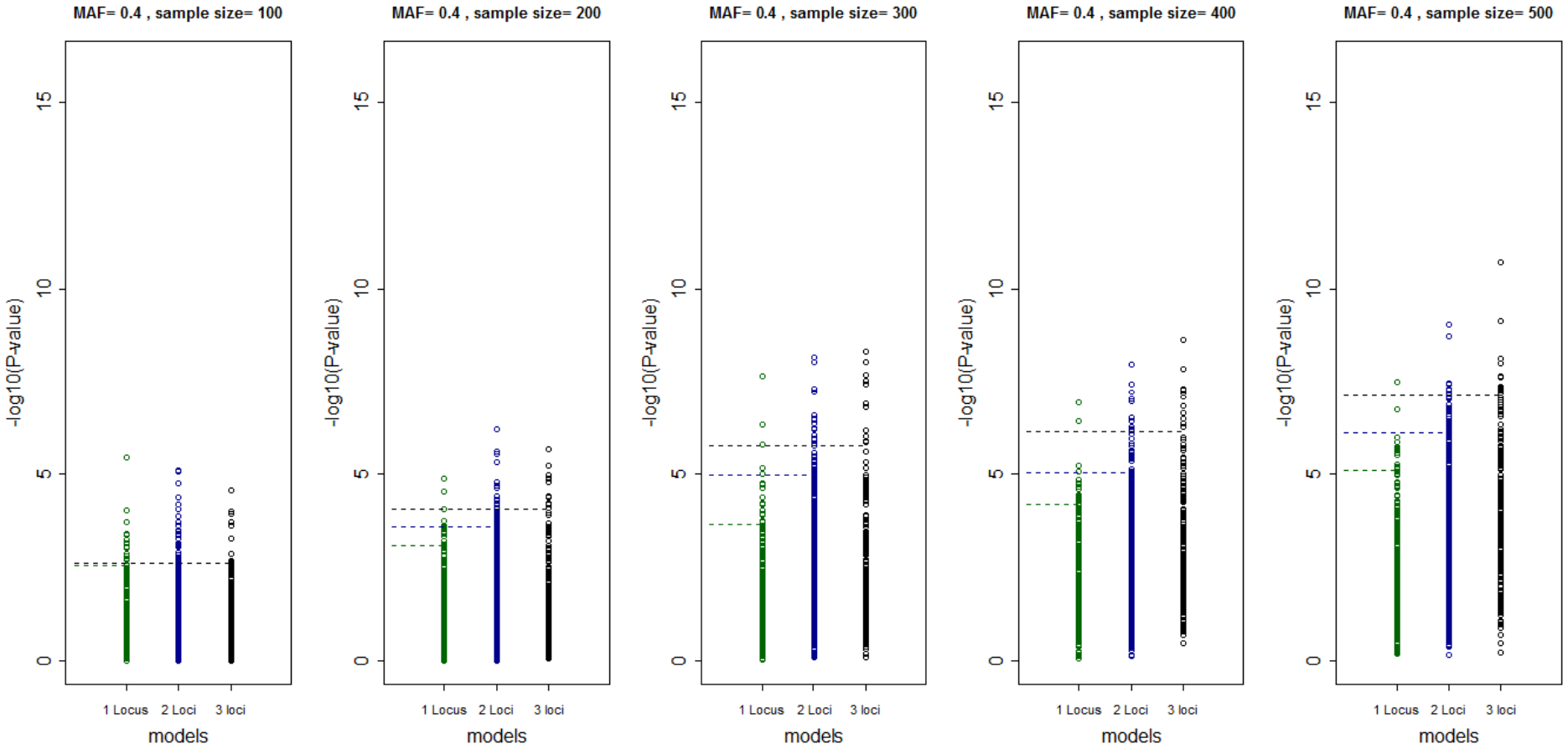


FIG. 4.3.4. Comparison of P-values between single locus and multi-locus for MAF = 0.40 and at sample size of 100, 200, 300, 400, and 500.

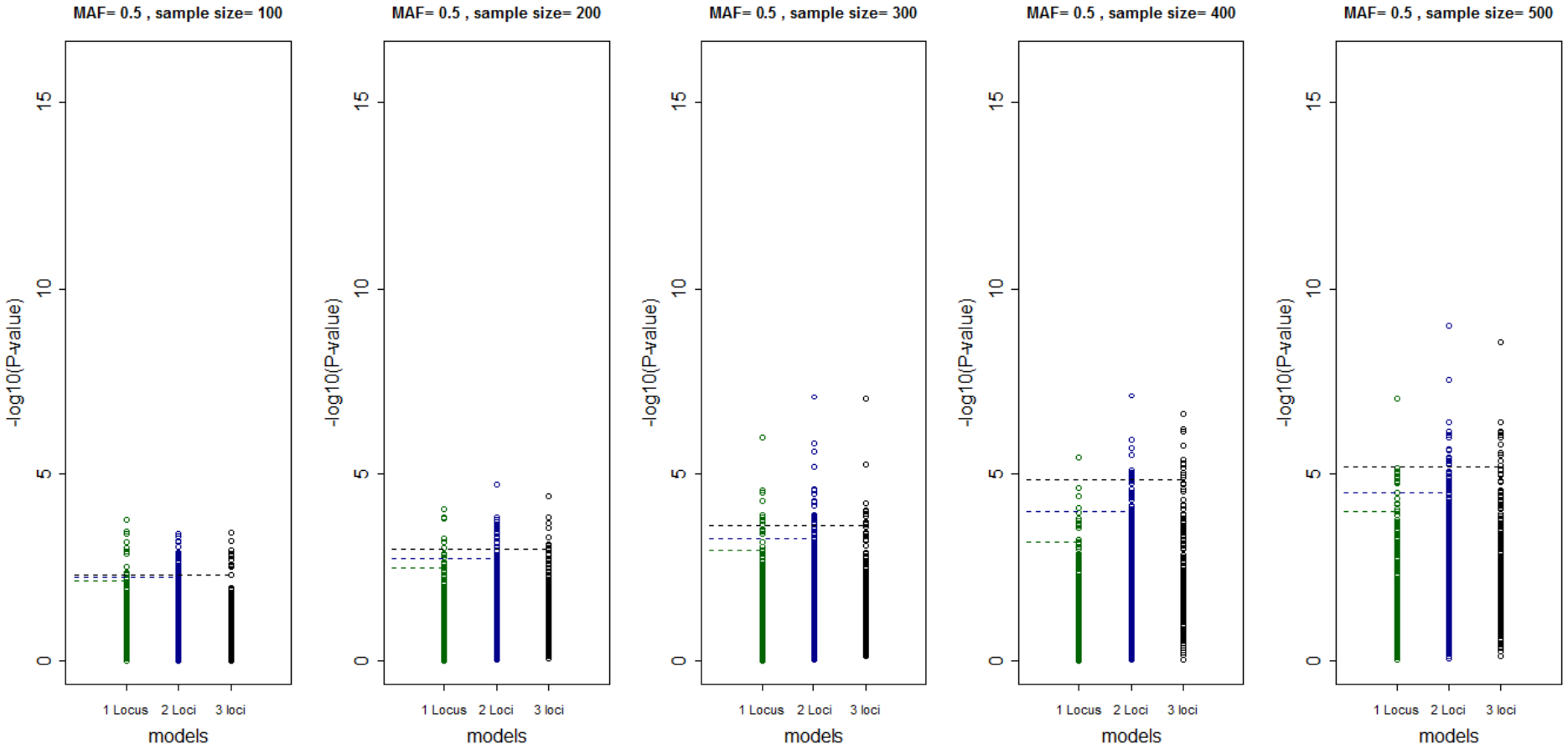


FIG. 4.3.5. Comparison of P-values between single locus and multi-locus for MAF = 0.50 and at sample size of 100, 200, 300, 400, and 500.

4.3.3 Application to the data of Dielmo and Ndiop

We apply the multi-locus method on 45 genes, candidates for association with malaria disease (Table 4.3.3). The malaria phenotype we used is the predicted individual effect from the mixed model in Chapter 3 "Heritability" where we separated the individual additive genetic effect from other personal effects included in the permanent environment effect. However, this permanent environment effect contains any non-additive genetic components. The complex, polygenic basis to the human response to malaria parasite infection may well include dominance/epistatic genetic effects that are encompassed within the permanent environment effect. Thus, the whole individual effect containing additive as well as non-additive genetics is used as the phenotype. As explained previously in the chapter 3, this phenotype is the individual contribution (individual slope or trend) to the risk of having clinical malaria episode (*PFA*) after adjusting on age, transmission season and also corrected for random variations within individual repeated measurements. Individuals having a positive slope correspond to those with a positive contribution to the population's mean risk to develop *PFA* and were classified as susceptible and individuals with negative slope contribute negatively to the population's mean risk to develop *PFA* and were classified as resistant. Transmission of alleles from parents to resistant offspring is then analyzed here to find genes showing protective effects against malaria.

We first analyzed the SNPs one by one and those showing marginal effect after correcting for multiple tests (by False Discovery Rate) were then selected for a joint transmission model. Results from single locus model are presented in Table 4.3.3. SNPs with marginal corrected P-value less than 0.10 were tested for linkage disequilibrium (LD) and, when they showed independency, were subsequently used for multi-locus models. Result of LD are presented in Tables 4.3.4 (A and B) and illustrated by Figures 4.3.6 (A and B). Results from multi-locus model are presented in Table 4.3.5. To limit the number of alternative hypothesis tested, we analyzed all models with one locus, all with two loci, all with three loci as the limit, and additionally the complete set of the *K* loci.

Table 4.3.3: Results for single locus allele transmission models.

Single locus models	log-likelihood	X	DF	P-value	Bonferroni	FDR
SNP-1 = ae1_20_21	-32.86	18.84	1	**0.00001**	**0.0005**	**0.0005**
SNP-2 = Xmn1	-56.04	8.52	1	**0.00351**	0.1579	**0.0598**
SNP-3 = ae1_117_118	-72.10	8.29	1	**0.00399**	0.1795	**0.0598**
SNP-4 = abo297	-53.27	7.13	1	**0.00759**	0.3415	**0.0854**
SNP-5 = ae1_174_187	-54.30	6.46	1	**0.01105**	0.4973	**0.0995**
SNP-6 = abo771	-51.93	5.65	1	**0.01746**	0.7857	0.1309
SNP-7 = ae1_189_190	-5.22	4.82	1	**0.02816**	1	0.181
SNP-8 = ubtf13_14	-78.82	4.55	1	**0.03290**	1	0.1815
SNP-9 = tgeiv_2134	-61.58	4.38	1	**0.03630**	1	0.1815
SNP-10 = HbS	-43.79	3.92	1	**0.04778**	1	0.215
SNP-11 = c9i203v	-39.94	3.30	1	0.06941	1	0.2601
SNP-12 = tgt220m	-20.59	3.18	1	0.07464	1	0.2601
SNP-13 = g6pd376	-30.30	3.17	1	0.07514	1	0.2601
SNP-14 = tga143a	-2.70	2.91	1	0.08798	1	0.2828
SNP-15 = l_30633_34	-71.40	2.76	1	0.09637	1	0.2891
SNP-16 = adarb2_in2_4900	-88.86	2.50	1	0.11382	1	0.3201
SNP-17 = ankl_9_10	-64.66	2.38	1	0.12303	1	0.3257
SNP-18 = rs10074987	-88.29	2.25	1	0.13389	1	0.3281
SNP-19 = hdc	-44.65	2.19	1	0.13855	1	0.3281
SNP-20 = cr1_q981h	-16.34	1.99	1	0.15871	1	0.3537
SNP-21 = g6pd202	-2.50	1.93	1	0.16504	1	0.3537
SNP-22 = abo526	-43.02	1.29	1	0.25603	1	0.507
SNP-23 = ae1_180_181	-66.61	1.25	1	0.26353	1	0.507
SNP-24 = acpl8_9	-69.41	1.20	1	0.27324	1	0.507
SNP-25 = tgq62r	-9.12	1.16	1	0.28169	1	0.507
SNP-26 = spk5k420e	-19.67	0.87	1	0.35196	1	0.5973
SNP-27 = abo467	-52.26	0.84	1	0.35835	1	0.5973
SNP-28 = cr1_r1601g	-60.63	0.73	1	0.39344	1	0.6138
SNP-29 = alpha_37del	-34.30	0.72	1	0.39557	1	0.6138
SNP-30 = c9r5w	-53.05	0.64	1	0.42471	1	0.6264
SNP-31 = cr1_k1590e	-90.49	0.62	1	0.43149	1	0.6264
SNP-32 = phf11b5_2	-53.83	0.46	1	0.49669	1	0.6985
SNP-33 = spkn368s	-18.55	0.33	1	0.56330	1	0.7443
SNP-34 = M_rs1427407	-74.69	0.33	1	0.56360	1	0.7443
SNP-35 = cr1_hind3	-35.89	0.31	1	0.57891	1	0.7443
SNP-36 = tg_862_1	-25.52	0.24	1	0.62168	1	0.7695
SNP-37 = tg_974	-3.37	0.20	1	0.65363	1	0.7695
SNP-38 = dip2c_in2_2051	-90.71	0.19	1	0.66218	1	0.7695
SNP-39 = dip2c_in2_2993	-93.48	0.19	1	0.66692	1	0.7695
SNP-40 = phf11b5_3	-40.82	0.15	1	0.69606	1	0.7831
SNP-41 = rs316414	-64.41	0.10	1	0.75572	1	0.8294
SNP-42 = fcgr2a_r131h	-79.67	0.08	1	0.77966	1	0.8354
SNP-43 = M_rs11154792	-45.72	0.06	1	0.80553	1	0.843
SNP-44 = acpl1_7	-25.63	0.03	1	0.86941	1	0.8892
SNP-45 = abo261	-68.62	0.01	1	0.91994	1	0.9199

Study of linkage disequilibrium (LD) between the SNPs candidate for multi-locus models:

SNPs with marginal P-value ≤ 0.10 were candidate for multi-locus models. Then five SNPs were concerned (Table 4.3.3): three on the same gene (ae1_20_21, ae1_117_118, ae1_174_187) and two on different genes located on different chromosomes (Xmn1 and abo297). Then, LD only between SNPs within the same gene needed to be tested. This was done using the programs SIMWALK2 (watson.hgen.pitt.edu/docs/simwalk2.html) that can test for LD in family data context and GOLD (www.sph.umich.edu/csg/abecasis/GOLD) that provides a graphical summary of LD results.

Table 4.3.4.(A): linkage disequilibrium test between AE1 SNPs among Dielmo families

		ae1_180_181	**ae1_174_187**	**ae1_20_21**	**ae1_189_190**
ae1_117_118	N	62	55	62	62
	X2(DF=1); P	35.4; 2.7E-09	8.2; 0.004	4.6; 0.032	0; 0.977
	Cramer's V	**0.76**	**0.39**	**0.27**	**0.01**
	U	**0.48**	**0.12**	**0.06**	**0.00**
	DELTA^2	**0.57**	**0.15**	**0.07**	**0.00**
ae1_180_181	N		55	62	62
	X2(DF=1); P		5.9; 0.015	4.7; 0.029	1.2; 0.275
	Cramer's V		**0.33**	**0.28**	**0.14**
	U		**0.09**	**0.06**	**0.02**
	DELTA^2		**0.11**	**0.08**	**0.02**
ae1_174_187	N			55	55
	X2(DF=1); P			12.9; 0.0003	0.5; 0.4571
	Cramer's V			**0.49**	**0.10**
	U			**0.19**	**0.01**
	DELTA^2			**0.24**	**0.01**
ae1_20_21	N				62
	X2(DF=1); P				0.1; 0.813
	Cramer's V				**0.03**
	U				**0.00**
	DELTA^2				**0.00**

N is Number of pairs scored; **Cramer's V** is a transformation of the Chi-squared based measures of association into [0,1]; **U** is uncertainty coefficient (How much information on one marker given by the other); **DELTA^2** is the Delta-Squared Measure of disequilibrium.

Table 4.3.4.(B): linkage disequilibrium test between AE1 SNPs among Ndiop families

		ae1_180_181	ae1_174_187	ae1_20_21	ae1_189_190
ae1_117_118	N	87	90	89	90
	X2(DF=1); P	54.3; 1.7E-13	5.9; 0.014	5.0; 0.025	1.6; 0.209
	Cramer's V	**0.79**	**0.26**	**0.24**	**0.13**
	U	**0.52**	**0.05**	**0.05**	**0.03**
	$DELTA^2$	**0.62**	**0.07**	**0.06**	**0.02**
ae1_180_181	N		89	89	90
	X2(DF=1); P		6.6; 0.010	4.2; 0.041	1.2; 0.269
	Cramer's V		**0.27**	**0.22**	**0.12**
	U		**0.06**	**0.04**	**0.02**
	$DELTA^2$		**0.07**	**0.05**	**0.01**
ae1_174_187	N			91	92
	X2(DF=1); P			5.7; 0.017	0.3; 0.603
	Cramer's V			**0.25**	**0.06**
	U			**0.05**	**0.01**
	$DELTA^2$			**0.06**	**0.00**
ae1_20_21	N				92
	X2(DF=1); P				3.6; 0.056
	Cramer's V				**0.20**
	U				**0.06**
	$DELTA^2$				**0.04**

N is Number of pairs scored; **Cramer's V** is a transformation of the Chi-squared based measures of association into [0,1]; **U** is uncertainty coefficient (How much information on one marker given by the other); **$DELTA^2$** is the Delta-Squared Measure of disequilibrium.

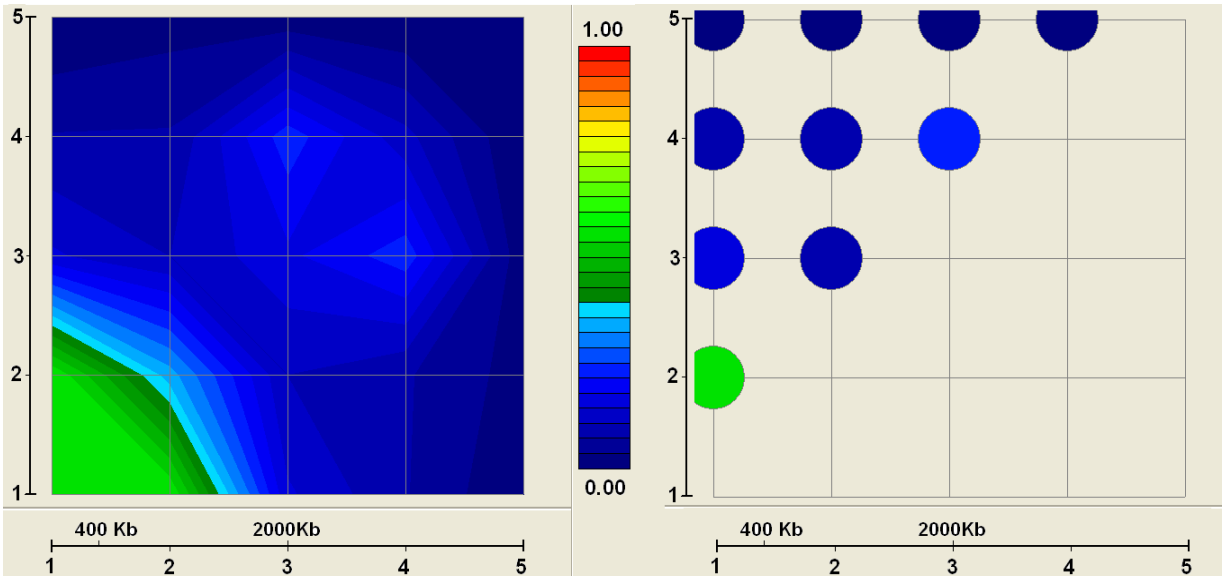


Figure 4.3.6.(A): Disequilibrium map for the AE1 markers among Dielmo families

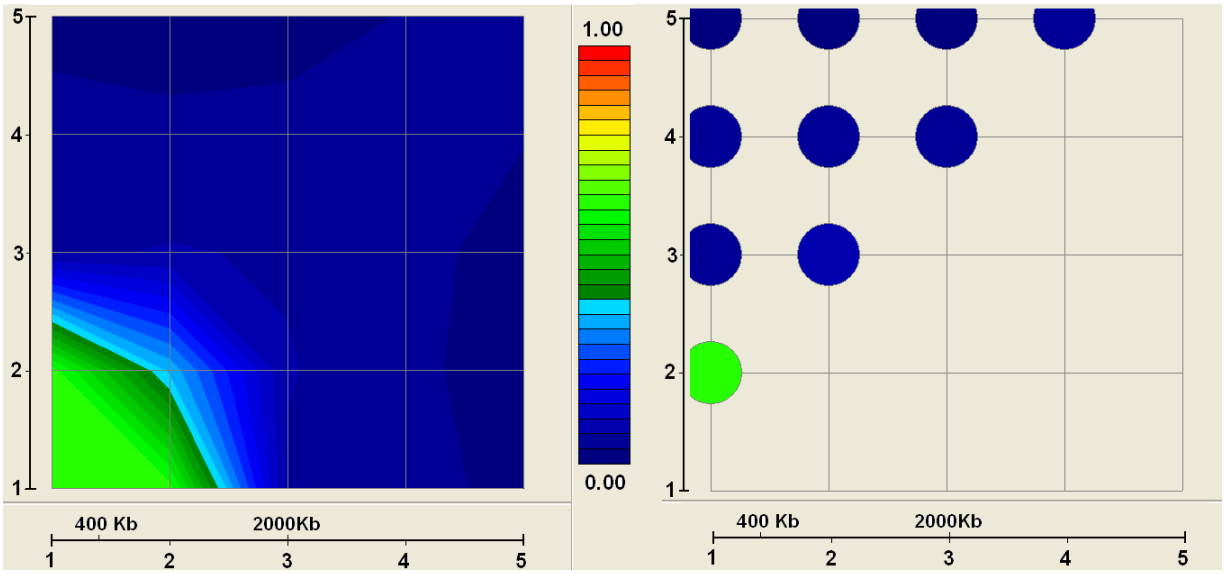


Figure 4.3.6.(B): Disequilibrium map for the AE1 markers among Ndiop families

Table 4.3.5: Results for multi-locus allele transmission models.

models		log-likelihood	X	DF	P-value	Bonferroni	FDR
SNP-1	(ae1_20_21)	-98.07	17.36	1	0.000031	0.001394	0.000082
SNP-2	(Xmn1)	-98.60	16.30	1	0.000054	0.002438	0.000106
SNP-3	(ae1_117_118)	-101.97	9.56	1	0.001991	0.089593	0.002421
SNP-4	(abo297)	-102.59	8.32	1	0.003927	0.176700	0.004531
SNP-5	(ae1_174_187)	-104.50	4.50	1	0.033929	1.000000	0.033929
SNP-1-2	Multiplicative	-91.12	31.26	2	2.00E-07	7.00E-06	3.30E-06
SNP-1-3	Multiplicative	-95.35	22.79	2	1.00E-05	5.10E-04	4.00E-05
SNP-1-4	Multiplicative	-93.59	26.31	2	2.00E-06	9.00E-05	1.00E-05
SNP-1-5	Multiplicative	-97.48	18.54	2	9.00E-05	4.25E-03	1.70E-04
SNP-2-3	Multiplicative	-95.59	22.30	2	1.00E-05	6.50E-04	4.00E-05
SNP-2-4	Multiplicative	-95.13	23.24	2	9.00E-06	4.00E-04	3.00E-05
SNP-2-5	Multiplicative	-96.92	19.64	2	5.00E-05	2.44E-03	1.10E-04
SNP-3-4	Multiplicative	-98.14	17.20	2	1.80E-04	8.29E-03	2.90E-04
SNP-3-5	Multiplicative	-101.58	10.34	2	5.69E-03	0.256230	6.25E-03
SNP-4-5	Multiplicative	-100.64	12.21	2	2.23E-03	0.100380	2.64E-03
SNP-1-2	Epistasis	-93.65	26.18	3	9.00E-06	3.90E-04	3.00E-05
SNP-1-3	Epistasis	-95.16	23.18	3	4.00E-05	1.67E-03	8.00E-05
SNP-1-4	Epistasis	-95.15	23.19	3	4.00E-05	1.66E-03	8.00E-05
SNP-1-5	Epistasis	-95.05	23.38	3	3.00E-05	1.51E-03	8.00E-05
SNP-2-3	Epistasis	-97.19	19.12	3	2.60E-04	0.011640	3.60E-04
SNP-2-4	Epistasis	-93.41	26.66	3	7.00E-06	3.10E-04	3.00E-05
SNP-2-5	Epistasis	-98.56	16.38	3	9.50E-04	0.042700	1.22E-03
SNP-3-4	Epistasis	-99.33	14.82	3	1.98E-03	0.088980	2.42E-03
SNP-3-5	Epistasis	-100.88	11.73	3	8.36E-03	0.376270	8.96E-03
SNP-4-5	Epistasis	-101.71	10.08	3	0.017930	0.806990	0.018340
SNP-1-2-3	Multiplicative	-89.84	33.81	3	2.00E-07	1.00E-05	3.00E-06
SNP-1-2-4	Multiplicative	-87.38	38.72	3	2.00E-08	9.00E-07	9.00E-07
SNP-1-2-5	Multiplicative	-90.94	31.60	3	6.00E-07	3.00E-05	6.00E-06
SNP-1-3-4	Multiplicative	-91.24	31.02	3	8.00E-07	3.79E-05	6.30E-06
SNP-1-3-5	Multiplicative	-96.33	20.83	3	1.10E-04	5.15E-03	2.00E-04
SNP-1-4-5	Multiplicative	-93.32	26.85	3	1.00E-05	2.80E-04	3.00E-05
SNP-2-3-4	Multiplicative	-92.36	28.76	3	3.00E-06	1.10E-04	1.00E-05
SNP-2-3-5	Multiplicative	-95.39	22.70	3	5.00E-05	2.10E-03	1.00E-04
SNP-2-4-5	Multiplicative	-93.64	26.21	3	1.00E-05	3.90E-04	3.00E-05
SNP-3-4-5	Multiplicative	-97.94	17.62	3	5.30E-04	0.023730	7.00E-04
SNP-1-2-3	Epistasis	-91.50	30.48	7	8.00E-05	3.49E-03	1.50E-04
SNP-1-2-4	Epistasis	-	-	-	-	-	-
SNP-1-2-5	Epistasis	-90.08	33.33	7	2.00E-05	1.04E-03	6.00E-05
SNP-1-3-4	Epistasis	-92.49	28.52	7	1.80E-04	7.96E-03	2.80E-04
SNP-1-3-5	Epistasis	-93.74	26.02	7	5.00E-04	0.022510	6.80E-04
SNP-1-4-5	Epistasis	-92.93	27.64	7	2.60E-04	0.011510	3.60E-04
SNP-2-3-4	Epistasis	-92.30	28.88	7	1.50E-04	6.83E-03	2.50E-04
SNP-2-3-5	Epistasis	-96.57	20.35	7	4.86E-03	0.218650	5.47E-03
SNP-2-4-5	Epistasis	-92.64	28.22	7	2.00E-04	9.04E-03	3.00E-04
SNP-3-4-5	Epistasis	-97.70	18.08	7	0.011600	0.522000	0.012140
SNP-1-2-3-4-5	Multiplicative	-87.48	38.53	5	3.00E-07	1.00E-05	3.00E-06
SNP-1-2-3-4-5	Epistasis	-	-	-	-	-	-

4.4 Discussion

By simulation study, we have shown the advantage of multi-locus models over single locus models to find significant genetic effects when the phenotype is influenced by several independent loci. Therefore multi-locus models represent important alternatives in the study of genetic susceptibility and resistance to multifactorial diseases such as infectious diseases. The results obtained by applying multi-locus modeling on the studied cohorts for malaria disease have confirmed these findings.

Among the 45 markers loci for the candidate genes, 5 showed after correction for multiple testing weak protective effect against malaria when we ignored information from other loci: three are on the *SLC4A1* (AE1) gene located on chromosome 17 and are independent, i.e. not in linkage disequilibrium, as shown above (ae1_20_21, P = 0.0005; ae1_117_118, P = 0.0598; ae1_174_187, P = 0.0995), one is on the γ-globin gene (Xmn1) located on chromosome 11 (Xmn1, P = 0.0598) and one on the *ABO* gene located on chromosome 9 (abo297, P = 0.0854). See Table 4.3.3 for single locus models. We then analyzed these five loci together. The sample is reduced to individuals with no missing genotypes at all the five loci otherwise we cannot know which complete set of alleles is transmitted and which other is not transmitted at these five genomic locations for each offspring, unless by inferring genotypes. Then, when we considered simultaneous transmission of alleles from these five loci, the protective effect became stronger as shown on Table 4.3.5. Also the single locus models performed on this sample reduced to individuals with no missing genotypes at all the five loci showed better marginal effects (Table 4.3.5). As the malaria phenotypes are suspected to be influenced by several genes at different location in the human genome, these results suggest that each of these five markers may be causally related to the disease or may not themselves be causal, but may be sufficiently close to five causal loci so as to be in linkage disequilibrium with them. The mutations occurring on *SLC4A1* (AE1) gene are known to be responsible for inherited blood disorders. Interestingly, it has been recognized for over 60 years that this negative effect of inherited blood disorders is compensated by the protection afforded against malaria parasites and yet the mechanism underlying this protection remains unknown (Williams 2006).

Whilst both parasite invasion and growth may be affected by such red cell mutations, there are currently two immunologically based hypotheses for the protective effect of blood disorders. One implicates this *SLC4A1* (AE1) gene having a main effect that is to accelerate red blood cell aging. This is more pronounced in parasitized red blood cells: the parasite causes premature aging of the cell. Band 3 (as known as “Anion Exchanger 1”, AE1) is a membrane ion transporter encoded by the gene *SLC4A1* that serves the additional functions of providing red cell membrane stability and “flagging” red blood cells for destruction. As the

red cell ages, alterations in band 3 lead to increased membrane rigidity and to its exposure on the surface of the red cell. Old red cells are thus removed from the circulation through filtering of the physically rigid cells and through antibody-dependent mechanisms. The protective effect of the blood disorders is thus to accelerate parasite removal using the bodies' own old red cell destruction mechanism (Pantaleo, Giribaldi et al. 2008; Tokumasu, Nardone et al. 2009). The other immunologically based hypothesis concerns the impact on infected red cell sequestration: *Plasmodium falciparum* expresses the *var* gene molecule PfEMP1 on knobs at the surface of the red cell. This molecule enables the parasite to sequester and thus avoid clearance by the spleen. The red cell disorders are believed to disrupt effective expression of PfEMP1 and thus impair parasite sequestration (and enhance the acquisition of immunity).

From a statistical point of view, the assumption of independence among the loci tested represents a great advantage of this multi-locus model for increasing the sample size when nuclear families are used. That is due to the fact that the tests are valid for any number of affected children in the nuclear families. Each offspring with the same parents constitute an independent trio in this case of independence as the Mendel's Law of allelic inheritance implies that the transmission of sets of alleles among offspring occurs independently.

One drawback of this method is the quick increase in the number of alternative hypotheses and in the number of free parameters that rapidly makes the corrected threshold of significance for the P-values at a very low level. Another disadvantage is that individuals should have genotype information on all the loci tested reducing the sample size, particularly when the missing genotypes for each locus occur on different individuals. Also we did not make analyses using inferred genotypes when no parental information is available, but such alternative methods based on the original TDT and using sib information exist in the literature: The sib TDT (S-TDT) of Spielman and Ewens (1998), the sibship disequilibrium test (SDT) of Horvath and Laird (1998) that uses data from all the affected and all the unaffected siblings.

5. General Conclusion

Summary

We followed two malaria cohorts in sub-Saharan Africa where the general burden of malaria has declined from 1990 to 2008. Despite this decrease in the general burden of malaria, from 1990 to 2008, found with several approaches, recent studies on the same population show a quick and recent increase back to the higher transmission levels during 2009 and 2010 (Trape, Tall et al. 2011). Indeed, this increase appears with a positive shift of age for population of susceptibility, from traditional young (less than 5 year-old) to older children (more than 10 year-old). Descriptive analyses allowed us to show that young age and the period of treatment were major factors determining the risk of PFA. This can be seen through results of the different methods we have tried in the first part of the thesis concerning the epidemiological analysis where variables “Age” and “Year” (or “Drug period” when the years were aggregated by drug periods) are variables with stronger predictive values explaining occurrence or not of the disease. Also, environmental factors are determinant in the transmission of the parasite from one individual to another as reflected by the contrasted prevalence of malaria between the two cohorts. The prevalence is high in the village of Dielmo where the transmission of the disease occurs all the year due of the presence of a small stream that enables mosquitoes to breed and seasonal in the village of Ndiop where transmission occurs only during rainy season from July to December.

However, there are many more factors involved at the individual level and as important as were the age and the year at the population level. Those factors are the inherited genetic background of the individual and the interactions between genetic and environment. Then our first step in the second part of the thesis which was the genetic analysis has focussed on variance component analysis to assess the overall genetic contribution to the disease prior to linkage and association studies. The variance component analyses divided the longitudinal study according to drug treatment to consider the impact of the radical selection pressure that would have been exerted on the parasite population at each change in drug treatment (Loucoubar, Goncalves et al. 2011). In addition, the change in transmission intensity occurring over the 19 year enabled us to assess its impact on the genetic contribution of malaria phenotypes. The evolution of anti-malarial drug resistance and the force of infection

have been well studied in the population (Trape, Rogier et al. 1994; Rogier, Tall et al. 1999; Noranate, Durand et al. 2007) and thus we explored heritability in these two cohorts undergoing well-defined environmental changes. However, the pedigree data has enabled just estimation of heritability in the narrow sense that is the additive genetic contribution. Actual values of heritability are specific for the study populations at a particular time and thus strict comparison is not informative. The size of heritability provides an indication of the power to detect the effect of individual genes when performing GWAS. Evolution of variance components in this study showed the replacement of an additive genetic component by a permanent environment component time. The permanent environment effect includes other non additive genetic effects. The total estimate of individual effects (additive and permanent environment) stayed constant over time, suggesting that the loss of the additive genetic effect may be due to absence of sufficient resolution in the pedigree matrix. Hence, the phenotype used in subsequent genetic analyses used the sum of both the individual additive genetic and permanent environmental effects. Hence, a mixed model using data from all the duration of survey and adjusted on the age and the year was performed separately in each village to pick up the global individual effect for each person as phenotype for linkage and association analysis.

As known for infectious disease, the genetic component of susceptibility/ resistance to malaria is very complex, with multiple genes involved. That motivated us to use multi-locus models that remain a relatively poorly developed field in genetic statistics research. Many of the existing methods, like FBAT Software's method, deal with haplotypes assuming an effect of an aggregation of very close loci to avoid hypothesis of recombination between genes, which increases the computational challenge. However, for multifactorial disease like malaria many candidate genes are distributed over the human genome on different chromosomes and up to now have showed weak effects, except for HbS. By simulation study, we have shown the advantage of multi-locus models over single locus models to find significant genetic effects when the phenotype is influenced by several independent loci. Therefore multi-locus models represent important alternatives in the study of genetic susceptibility or resistance to malaria. The results obtained by applying multi-locus modeling on the two studied cohorts for malaria disease have confirmed these findings. The assumption of independence between the loci in the computation of the likelihood of allelic transmission in this multi-locus model is not constrained by the non recombination hypothesis. However, the method is limited by the quick increase in the number of free parameters and in the number of alternatives hypothesis that makes the corrected threshold of significance for the P-values very low. Also, the study sample of individuals should have genotype information on all the loci tested reducing the sample size when no method to infer genotype is included, particularly when missing genotypes for each locus occur on different individuals.

Perspective

Statistical genetics in the studies of infectious disease represents at the moment a very large research field. One of the main challenges for infectious diseases is the problem of phenotypes, which are frequently quantitative and thus need robust definitions to distinguish between disease and no disease; weak genetic effects will be very sensitive to the phenotype resolution. Also, repeated measurements are generally preferred to a single measure for phenotypes to detect a confirmed trend for each patient, but, this choice induces several challenges in statistical modeling. Generalized Estimation Equations (GEE) proposed in 1986 by Kung-Yee Liang and Scott L. Zeger (Zeger and Liang 1986) and/or Mixed Models are then adequate but more adapted to case-control designs than family designs. FBAT-GEE was proposed as a GEE version of family based method but has limitations to deal with multiple independent loci. Other proposed programs allowing for multi-locus analysis do not provide methods for repeated measurements. Thus, in any case, we usually work in two steps, as we did in this thesis, by dealing with the problem of the phenotype in a prior analysis and use the residual phenotype in last genetic analyses.

Therefore, concerning infectious disease, an important gap has to be filled in the development of methods of analysis allowing for repeated and correlated measurements, one locus as well as many loci that could be independent or dependent, and allowing for covariates. Such specific statistical methods that fit the available data can be helpful to empirically confirm or disprove, or to find genes with not necessarily strong effects on malaria disease. Effects of such genes could easily be hidden by phenotype resolution or by the method of analysis or by empirical properties of the of tests' statistics.

Four SNPs of the G6PD gene were typed in our two studied cohorts and only weak protective effects, depending on sex, were found in a subpopulation of the cohorts by using regression methods and survival analyses. These methods are not immune from population stratification problems and yet, family based methods we performed did not provide significant results. Several investigations are being done in that field. Scientists studying genetic association of malaria susceptibility / resistance should consider further study and improvements in the method of G6PD deficiency assessment as well as other inherited blood disorders and also in statistical genetic methods to make advance in malaria genetic researches.

Bibliography

Abecasis, G. R., L. R. Cardon, et al. (2000). "A general test of association for quantitative traits in nuclear families." Am J Hum Genet **66**(1): 279-92.

Abecasis, G. R., S. S. Cherny, et al. (2002). "Merlin--rapid analysis of dense genetic maps using sparse gene flow trees." Nat Genet **30**(1): 97-101.

Adjuik, M., A. Babiker, et al. (2004). "Artesunate combinations for treatment of malaria: meta-analysis." Lancet **363**(9402): 9-17.

Agarwal, A., A. Guindo, et al. (2000). "Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S." Blood **96**(7): 2358-63.

Aidoo, M., D. J. Terlouw, et al. (2002). "Protective effects of the sickle cell gene against malaria morbidity and mortality." Lancet **359**(9314): 1311-2.

Amexo, M., R. Tolhurst, et al. (2004). "Malaria misdiagnosis: effects on the poor and vulnerable." Lancet **364**(9448): 1896-8.

Aucan, C., A. J. Walley, et al. (2003). "Interferon-alpha receptor-1 (IFNAR1) variants are associated with protection against cerebral malaria in the Gambia." Genes Immun **4**(4): 275-82.

Balding, D. J., M. J. Bishop, et al. (2007). Handbook of statistical genetics, John Wiley & Sons.

Bell, D., C. Wongsrichanalai, et al. (2006). "Ensuring quality and access for malaria diagnosis: how can it be achieved?" Nat Rev Microbiol **4**(9 Suppl): S7-20.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. Series B (Methodological) **57**(1): 289-300.

Besse, P. and A. Baccini (2007). Exploration Statistique. Laboratoire de Statistique et Probabilites. Institut de Mathematiques de Toulouse.

Bhattarai, A., A. S. Ali, et al. (2007). "Impact of artemisinin-based combination therapy and insecticide-treated nets on malaria burden in Zanzibar." PLoS Med **4**(11): e309.

Bickeboller, H. and F. Clerget-Darpoux (1995). "Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers." Genet Epidemiol **12**(6): 865-70.

Bousema, T. and C. Drakeley (2011). "Epidemiology and infectivity of *Plasmodium falciparum* and *Plasmodium vivax* gametocytes in relation to malaria control and elimination." Clin Microbiol Rev **24**(2): 377-410.

Breiman, L. (2001). "Random Forests." Machine Learning **45**(1): 5-32.

Breiman, L., J. Friedman, et al. (1984). "Classification and Regression Trees." Chapman and Hall.

Ceesay, S. J., C. Casals-Pascual, et al. (2008). "Changes in malaria indices between 1999 and 2007 in The Gambia: a retrospective analysis." Lancet **372**(9649): 1545-54.

Cordell, H. J. (2009). "Detecting gene-gene interactions that underlie human diseases." Nat Rev Genet **10**(6): 392-404.

Curtis, D. and P. C. Sham (1995). "A note on the application of the transmission disequilibrium test when a parent is missing." Am J Hum Genet **56**(3): 811-2.

Dice, L. R. (1945). "Measures of the Amount of Ecologic Association Between Species." Ecology **26**(3): 297–302.

Duffy, P. E. (2007). "*Plasmodium* in the placenta: parasites, parity, protection, prevention and possibly preeclampsia." Parasitology **134**(Pt 13): 1877-81.

English, M., J. Berkley, et al. (2003). "Hypothetical performance of syndrome-based management of acute paediatric admissions of children aged more than 60 days in a Kenyan district hospital." Bull World Health Organ **81**(3): 166-73.

Everitt, B., S. Landau, et al. (2001). Cluster analysis, Arnold.
Ewens, W. J. and R. S. Spielman (1995). "The transmission/disequilibrium test: history, subdivision, and admixture." Am J Hum Genet **57**(2): 455-64.
Falconer, D. S. and T. F. C. Mackay (1996). Introduction to Quantitative Genetics. London, Longman.
Foo, L. C., V. Rekhraj, et al. (1992). "Ovalocytosis protects against severe malaria parasitemia in the Malayan aborigines." Am J Trop Med Hyg **47**(3): 271-5.
Forgy, E. W. (1965). "Cluster analysis of multivariate data: efficiency vs interpretability of classifications." Biometrics **21**: 768–769.
Fulker, D. W., S. S. Cherny, et al. (1999). "Combined linkage and association sib-pair analysis for quantitative traits." Am J Hum Genet **64**(1): 259-67.
Greenacre, M. J. (2010). "Correspondence analysis." Wiley Interdisciplinary Reviews: Computational Statistics **2**(5): 613-619.
Gyan, B. A., B. Goka, et al. (2004). "Allelic polymorphisms in the repeat and promoter regions of the interleukin-4 gene and malaria severity in Ghanaian children." Clin Exp Immunol **138**(1): 145-50.
Haldane, J. B. (1949). "The association of characters as a result of inbreeding and linkage." Ann Eugen **15**(1): 15-23.
Henderson, C. R. (1973). "Sire evaluation and genetic trends." American Society of Animal Science.
Henderson, C. R. (1984). "Applications of Linear Models in Animal Breeding." University of Guelph.
Hill, A. V., C. E. Allsopp, et al. (1991). "Common west African HLA antigens are associated with protection from severe malaria." Nature **352**(6336): 595-600.
Horvath, S. and N. M. Laird (1998). "A discordant-sibship test for disequilibrium and linkage: no need for parental data." Am J Hum Genet **63**(6): 1886-97.
Horvath, S., X. Xu, et al. (2001). "The family based association test method: strategies for studying general genotype--phenotype associations." Eur J Hum Genet **9**(4): 301-6.
Hotelling, H. (1933). "Analysis of a complex of statistical variables into principal components." Journal of Educational Psychology **24**(7): 498-520.
Hutagalung, R., P. Wilairatana, et al. (1999). "Influence of hemoglobin E trait on the severity of *Falciparum* malaria." J Infect Dis **179**(1): 283-6.
Jaccard, P. (1901). "Etude comparative de la distribution florale dans une portion des alpes et des jura." Bulletin de la Societe Vaudoise des Sciences Naturelles **37**: 547–579.
Jolliffe, I. T. (2002). Principal Component Analysis. New York, Springer-Verlag New York.
Kallander, K., J. Nsungwa-Sabiiti, et al. (2004). "Symptom overlap for malaria and pneumonia--policy implications for home management strategies." Acta Trop **90**(2): 211-4.
Knight, J. C., I. Udalova, et al. (1999). "A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria." Nat Genet **22**(2): 145-50.
Laird, N. M., S. Horvath, et al. (2000). "Implementing a unified approach to family-based tests of association." Genet Epidemiol **19 Suppl 1**: S36-42.
Laird, N. M. and J. H. Ware (1982). "Random-effects models for longitudinal data." Biometrics **38**(4): 963-74.
Lake, S. L., D. Blacker, et al. (2000). "Family-based tests of association in the presence of linkage." Am J Hum Genet **67**(6): 1515-25.
Lawaly, Y. R., A. Sakuntabhai, et al. (2010). "Heritability of the human infectious reservoir of malaria parasites." PLoS One **5**(6): e11358.
Lawton, J. H. (1988). "More time means more variation." Nature: 334, 563.
Loucoubar, C., B. Goncalves, et al. (2011). "Impact of Changing Drug Treatment and Malaria Endemicity on the Heritability of Malaria Phenotypes in a Longitudinal Family-Based Cohort Study." PLoS One **6**(11): e26364.
Loucoubar, C., R. Paul, et al. (2011). "An exhaustive, non-euclidean, non-parametric data mining tool for unraveling the complexity of biological systems - novel insights into malaria." PLoS One **6**(9): e24085.

Ma, L., S. Han, et al. "Multi-locus test conditional on confirmed effects leads to increased power in genome-wide association studies." PLoS One **5**(11): e15006.

Mackinnon, M. J., D. M. Gunawardena, et al. (2000). "Quantifying genetic and nongenetic contributions to malarial infection in a Sri Lankan population." Proc Natl Acad Sci U S A **97**(23): 12661-6.

Mackinnon, M. J., T. W. Mwangi, et al. (2005). "Heritability of malaria in Africa." PLoS Med **2**(12): e340.

Mahalanobis, P. C. (1936). "On the generalised distance in statistics." Proceedings of the National Institute of Sciences of India **2**(1): 49–55.

Manly, B. F. J. (2005). Multivariate statistical methods: a primer, Chapman & Hall/CRC Press.

Martin, E. R., S. A. Monks, et al. (2000). "A test for linkage and association in general pedigrees: the pedigree disequilibrium test." Am J Hum Genet **67**(1): 146-54.

McCulloch, C. E. (2008). Generalized, linear, and mixed models, Hoboken, N.J. : Wiley

McGuire, W., A. V. Hill, et al. (1994). "Variation in the TNF-alpha promoter region associated with susceptibility to cerebral malaria." Nature **371**(6497): 508-10.

McKinney, B. A., D. M. Reif, et al. (2006). "Machine learning for detecting gene-gene interactions: a review." Appl Bioinformatics **5**(2): 77-88.

Miller, M. J. (1958). "Observations on the natural history of malaria in the semi-resistant West African." Trans R Soc Trop Med Hyg **52**(2): 152-68.

Mockenhaupt, F. P., S. Ehrhardt, et al. (2004). "Alpha(+)-thalassemia protects African children from severe malaria." Blood **104**(7): 2003-6.

Morahan, G., C. S. Boutlis, et al. (2002). "A promoter polymorphism in the gene encoding interleukin-12 p40 (IL12B) is associated with mortality from cerebral malaria and with reduced nitric oxide production." Genes Immun **3**(7): 414-8.

Morris, A. and J. Whittaker (1999). "Generalization of the extended transmission disequilibrium test to two unlinked disease loci." Genet Epidemiol **17 Suppl 1**: S661-6.

Mwangi, T. W., M. Mohammed, et al. (2005). "Clinical algorithms for malaria diagnosis lack utility among people of different age groups." Trop Med Int Health **10**(6): 530-6.

Nelson, M. R., S. L. Kardia, et al. (2001). "A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation." Genome Res **11**(3): 458-70.

Noranate, N., R. Durand, et al. (2007). "Rapid Dissemination of *Plasmodium falciparum* Drug Resistance Despite Strictly Controlled Antimalarial Use." PLoS ONE **2**: e139.

Nosten, F. and N. J. White (2007). "Artemisinin-based combination treatment of *falciparum* malaria." Am J Trop Med Hyg **77**(6 Suppl): 181-92.

O'Meara, W. P., P. Bejon, et al. (2008). "Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya." Lancet **372**(9649): 1555-62.

Okell, L. C., C. J. Drakeley, et al. (2008). "Reduction of transmission from malaria patients by artemisinin combination therapies: a pooled analysis of six randomized trials." Malar J **7**: 125.

Pantaleo, A., G. Giribaldi, et al. (2008). "Naturally occurring anti-band 3 antibodies and red blood cell removal under physiological and pathological conditions." Autoimmun Rev **7**(6): 457-62.

Pearson, K. (1901). "On lines and planes of closest fit to systems of points in space." Philosophical Magazine **2**: 559-572.

Phimpraphi, W., R. Paul, et al. (2008). "Heritability of P. *falciparum* and P. *vivax* malaria in a Karen population in Thailand." PLoS One **3**(12): e3887.

Protopopoff, N., W. Van Bortel, et al. (2009). "Ranking malaria risk factors to guide malaria control efforts in African highlands." PLoS One **4**(11): e8022.

Rabinowitz, D. (1997). "A transmission disequilibrium test for quantitative trait loci." Hum Hered **47**(6): 342-50.

Rabinowitz, D. and N. Laird (2000). "A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information." Hum Hered **50**(4): 211-23.

Reyburn, H., R. Mbatia, et al. (2004). "Overdiagnosis of malaria in patients with severe febrile illness in Tanzania: a prospective study." BMJ **329**(7476): 1212.

Richard, A., M. Lallemant, et al. (1988). "[Malaria in the forest region of Mayombe, People's Republic of the Congo. III. The role of malaria in general morbidity]." Ann Soc Belg Med Trop **68**(4): 317-29.

Ritchie, M. D., L. W. Hahn, et al. (2001). "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer." Am J Hum Genet **69**(1): 138-47.

Rogier, C., D. Commenges, et al. (1996). "Evidence for an age-dependent pyrogenic threshold of *Plasmodium falciparum* parasitemia in highly endemic populations." Am J Trop Med Hyg **54**(6): 613-9.

Rogier, C., A. Tall, et al. (1999). "*Plasmodium falciparum* clinical malaria: lessons from longitudinal studies in Senegal." Parassitologia **41**(1-3): 255-9.

Sakuntabhai, A., R. Ndiaye, et al. (2008). "Genetic determination and linkage mapping of *Plasmodium falciparum* malaria related traits in Senegal." PLoS One **3**(4): e2000.

Schaid, D. J. and H. Li (1997). "Genotype relative-risks and association tests for nuclear families with missing parental data." Genet Epidemiol **14**(6): 1113-8.

Sham, P. C. and D. Curtis (1995). "An extended transmission/disequilibrium test (TDT) for multi-allele marker loci." Ann Hum Genet **59**(Pt 3): 323-36.

Sinton, J. A., Harbhagivan, S. and Singh, J. (1931). "The numerical prevalence of parasites in relation to fever in chronic benign tertian malaria." Indian Journal of Medical Research **18**: 871-9.

Smith, T., B. Genton, et al. (1994). "Relationships between *Plasmodium falciparum* infection and morbidity in a highly endemic area." Parasitology **109 (Pt 5)**: 539-49.

Spielman, R. S. and W. J. Ewens (1996). "The TDT and other family-based tests for linkage disequilibrium and association." Am J Hum Genet **59**(5): 983-9.

Spielman, R. S. and W. J. Ewens (1998). "A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test." Am J Hum Genet **62**(2): 450-8.

Spielman, R. S., R. E. McGinnis, et al. (1993). "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)." Am J Hum Genet **52**(3): 506-16.

Stirnadel, H. A., H. P. Beck, et al. (1999). "Heritability and segregation analysis of immune responses to specific malaria antigens in Papua New Guinea." Genet Epidemiol **17**(1): 16-34.

Tokumasu, F., G. A. Nardone, et al. (2009). "Altered membrane structure and surface potential in homozygous hemoglobin C erythrocytes." PLoS One **4**(6): e5828.

Trape, J. F., C. Rogier, et al. (1994). "The Dielmo project: a longitudinal study of natural malaria infection and the mechanisms of protective immunity in a community living in a holoendemic area of Senegal." Am J Trop Med Hyg **51**(2): 123-37.

Trape, J. F., A. Tall, et al. (2011). "Malaria morbidity and pyrethroid resistance after the introduction of insecticide-treated bednets and artemisinin-based combination therapies: a longitudinal study." Lancet Infect Dis.

Trefethen, L. N. and D. Bau (1997). Numerical linear algebra, Society for Industrial and Applied Mathematics.

Vaart, A. V. d. (2006). Statistics in Genetics.

Vazquez, A. I., D. M. Bates, et al. (2009). "Technical note: an R package for fitting generalized linear mixed models in animal breeding." J Anim Sci **88**(2): 497-504.

Visscher, P. M., W. G. Hill, et al. (2008). "Heritability in the genomics era--concepts and misconceptions." Nat Rev Genet **9**(4): 255-66.

Weatherall, D. J. (1997). "Thalassaemia and malaria, revisited." Ann Trop Med Parasitol **91**(7): 885-90.

WHO. (2009). "http://www.who.int/features/factfiles/malaria/en/index.html."

Williams, T. N. (2006). "Red blood cell defects and malaria." Mol Biochem Parasitol **149**(2): 121-7.

Williams, T. N., T. W. Mwangi, et al. (2005). "Sickle cell trait and the risk of *Plasmodium falciparum* malaria and other childhood diseases." J Infect Dis **192**(1): 178-86.

Williams, T. N., S. Wambua, et al. (2005). "Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal *Plasmodium falciparum* malaria on the coast of Kenya." Blood **106**(1): 368-71.

Wilson, A. G., J. A. Symons, et al. (1997). "Effects of a polymorphism in the human tumor necrosis factor alpha promoter on transcriptional activation." Proc Natl Acad Sci U S A **94**(7): 3195-9.

Zeger, S. L. and K. Y. Liang (1986). "Longitudinal data analysis for discrete and continuous outcomes." Biometrics **42**(1): 121-30.

Annexes

Annex A

Metric

The first step for measuring variability in a population is to understand notions of similarity and dissimilarity, and then define a measure of distance between each pair of observations, such that the same group will be attributed to similar subjects, based only on their observations. A variety of metrics can be used to calculate similarities and their choice may affect the results.

Consider $\Omega = \{1,...,i,...,n\}$ the set of n individuals. We can present several metrics on $\Omega\times\Omega$ from less to more structured.

Definition *Similarity*: A similarity index is an application s on a pair of individuals having positive real values (s: $\Omega\times\Omega \rightarrow R_+$) and verifying the following conditions:

(i) $s(i,j) = s(j,i), \ \forall (i,j) \in \Omega\times\Omega,$
(ii) $s(i,i) = S > 0, \ \forall i \in \Omega$ and S independent of i,
(iii) $s(i,j) \leq S, \quad \forall (i,j) \in \Omega\times\Omega.$

Remark: $s^*(i,j) = (1/S)\times s(i,j)$ is a normed similarity index,

with $s^* : \Omega\times\Omega \rightarrow [0,1]$ and $S^* = 1$.

Definition *Dissimilarity*: A dissimilarity index is an application d on a pair of individuals having positive real values (d: $\Omega\times\Omega \rightarrow R_+$) and verifying the following conditions:

(i) $d(i,j) = d(j,i), \quad \forall (i,j) \in \Omega\times\Omega,$
(ii) $d(i,i) = 0, \quad \forall i \in \Omega,$ (or $i = j \Rightarrow d(i,j) = 0 \ \forall (i,j) \in \Omega\times\Omega$).

Remark: $d^*(i,j) = (1/D) \times d(i,j)$ is a normed dissimilarity index, where $D = \max\{d(i,j)\}$,

with $d^*: \Omega \times \Omega \rightarrow [0,1]$ and $D^* = 1$.

These two dual notions, similarity and dissimilarity, have weak properties due to the generality to construct indices satisfying their conditions.

Definition *Distance*: A distance is a dissimilarity index verifying in addition these two conditions: $\{d(i,j) = 0 \Rightarrow i = j, \forall (i,j) \in \Omega \times \Omega\}$ and {*The Triangle Inequality*}. Then a distance is an application $d: \Omega \times \Omega \rightarrow R_+$ verifying:

(i)	$d(i,j) = d(j,i)$,	$\forall (i,j) \in \Omega \times \Omega$,	(Symmetry)
(ii)	$d(i,j) = 0 \quad <=> i = j$	$\forall (i,j) \in \Omega \times \Omega$,	(Positive definiteness)
(iii)	$d(i,j) \leq d(i,k) + d(k,j)$	$\forall (i,j,k) \in \Omega \times \Omega \times \Omega$.	(The Triangle Inequality)

Several dissimilarity indices exist and can be used to make distance between subjects. Some widely used are Jaccard index (Jaccard 1901) or Dice & Zekanowski index (Dice 1945).

In statistical analysis, Euclidean distance is implicitly considered in almost all methods used to measure variability and tendencies, when the user knows or does not know. An alternative metric widely used in descriptive analysis (e.g. Discriminant Analysis) is Mahalanobis distance (Mahalanobis 1936) that takes into account correlation of the dataset, and therefore is robust in handling outliers or most noisy observations for which lower weights are assigned. The assignment of lower weights to most correlated pairs of observations can be perceived through the definition of Mahalanobis distance where the inverse of the covariance matrix integrate the formula; then, larger covariance means larger denominators leading to a lower weight, see equations below.

Euclidean distance

Consider $\Omega = \{1,...,i,...,n\}$ the set of n individuals represented in a m-dimensional space. The Euclidean distance between two individuals is the length of the segment joining them. This length is given by:

$$d(i,j) = \sqrt{\sum_{k=1}^{m} (i_k - j_k)^2} = \sqrt{(i_1 - j_1)^2 + (i_2 - j_2)^2 + ... + (i_m - j_m)^2}$$

where m is the number of dimensions, i_k and j_k are the coordinates on the k^{th} dimension of individual i and j.

Mahalanobis distance

Consider $\Omega = \{1,...,i,...,n\}$ the set of n individuals represented in a m-dimensional space. The Mahalanobis distance between two individuals is given by:

$$d(i,j) = \sqrt{(p-q) \times \Sigma^{-1} \times (p-q)^{\mathrm{T}}}$$

where $p = (i_1\ i_2\ ...\ i_m)$ and $q = (j_1\ j_2\ ...\ j_m)$ are the vector of coordinates of individuals i and j; m is the number of dimensions, i_k and j_k are the coordinates on the k^{th} dimension; Σ is the covariance matrix of the data with dimension $m \times m$, Σ^{-1} is the inverse of Σ; x^{T} is the transpose of x.

NB: The parallel is done with a dataset containing n observations and m variables after some standardization necessary when the variables' scales differ and when the variables are qualitative. Most statistical software automatically process to the standardization of data prior to analysis. Standardized values on variables are used as Cartesian coordinates in a space where each variable represents an axis. For data that show linear relationships, Euclidean distance is a useful measure of distance.

When Σ is the identity matrix (matrix with only values 1 on the diagonal and 0 elsewhere), corresponding to the case where all variables in the dataset are independent, Mahalanonis and Euclidean distances are equivalent.

Illustration: Clustering of *N* individuals around *k* centroids using different metrics.

Consider X and Y, two random quantitative variables observed on a sample of *N* individuals to be classify in *k* given groups represented by their centroids. Let us introduce some correlation in the data by taking observations of some samples as combinations of the others (see the R script below). The result of clustering will not be always the same using Euclidean or Mahalanobis metric. This can be illustrated by running several times the script, applied for a simulation of $k = 3$ classes and $N = 9$ individuals. Here is presented one realization of both random variables X and Y on the sample:

	X	**Y**
Centroid_1	**-2.746**	**-1.372**
Centroid_2	**0.473**	**-1.685**
Centroid_3	**2.823**	**-0.997**
individual_1	-4.328	-0.124
individual_2	1.251	2.158
individual_3	2.878	2.189
individual_4	-2.209	-3.143
individual_5	1.304	3.422
individual_6	-1.406	-0.356
individual_7	-5.573	-9.514
individual_8	0.816	-1.773
individual_9	-2.937	-0.543

Figure A.1 illustrates how the choice of different metrics can leads to different results, and then, encourages the use of hypothesis- free methods.

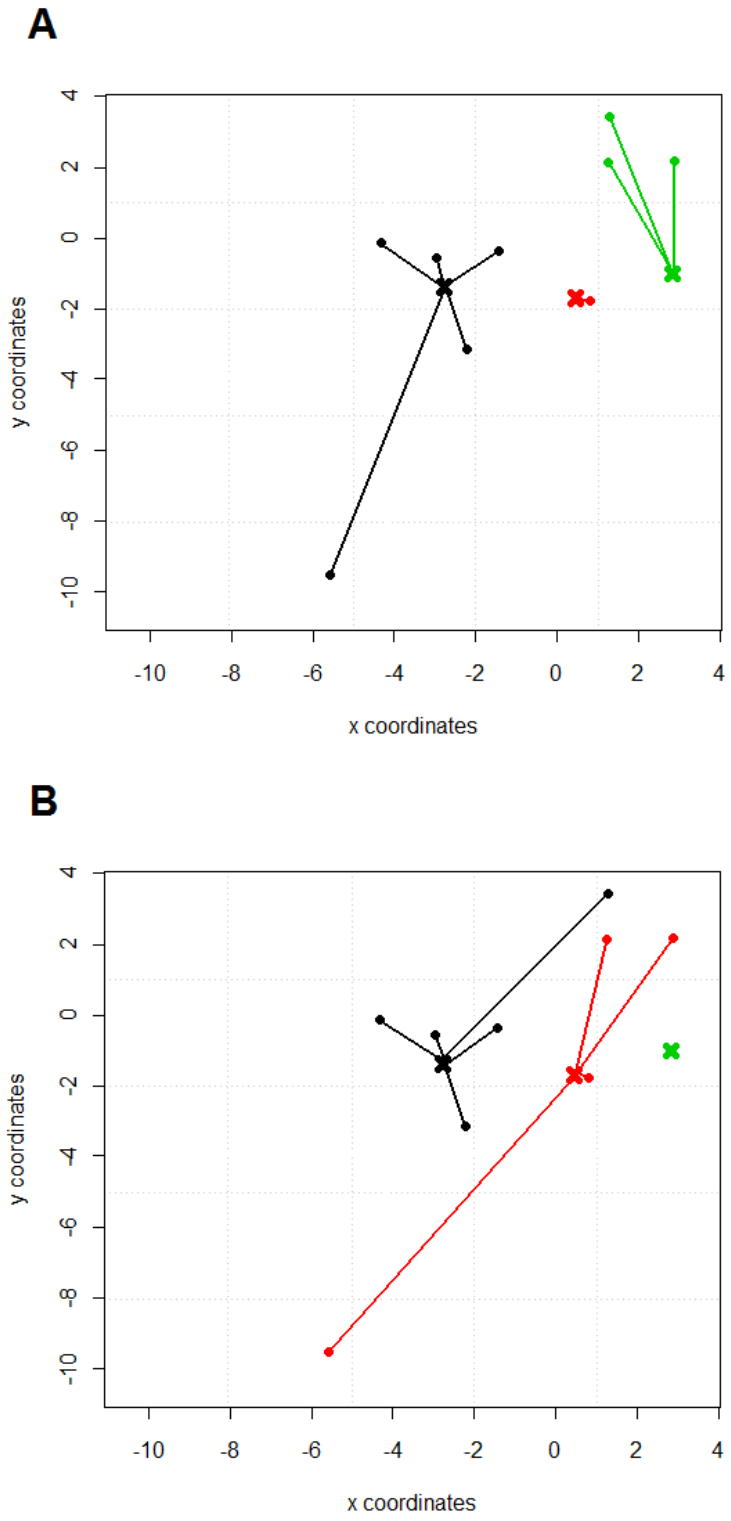


FIG. A.1. Clustering around three centroids using (A) Euclidean distance and (B) Mahalanobis distance, on the same data.

These differences occurring in the clustering can be illustrated by running several realizations of X and Y using this R script:

```
# ----- Beginning of the R script ----- #
library(MASS)
mahalanobis=function(a,b,metric){
 dist_maha = sqrt((a-b)%*%ginv(metric)%*%(a-b))
 return(dist_maha)
}
Euclide <- function(X1=c(0,0), X2=c(0,0)){
return(sqrt((X1[1]-X2[1])^2 + (X1[2]-X2[2])^2))
}
n=3; N=3*n
s1 = cbind(runif(n,-5,5),runif(n,-5,5))
s2 = 0.28*s1 + matrix(rnorm(n*2,0,1.54),n,2)
s3 = 1.33*s2 -0.54*s2^2
Points = rbind(s1,s2,s3)
colnames(Points) = c("X","Y")
i01=c(runif(1,-5,5),runif(1,-5,5))
i02=c(runif(1,-5,5),runif(1,-5,5))
i03=c(runif(1,-5,5),runif(1,-5,5))
par(mfrow=c(1,2))
plot((min(Points,i01,i02,i03)-1):(max(Points,i01,i02,i03)+1),(min(Points,i01,i02,i03)-
1):(max(Points,i01,i02,i03)+1), type="n", panel.first = grid(5,5),frame.plot=T, axes=T, xlab="x
coordinates",ylab="y coordinates")
for(i in 1:N){
if (min(Euclide(Points[i,],i01),Euclide(Points[i,],i02),Euclide(Points[i,],i03))==Euclide(i01, Points[i,]))
{points(Points[i,1],Points[i,2],col=1,pch=19); segments(Points[i,1],Points[i,2],i01[1],i01[2],lwd=2,col=1)}
        else{
        if (min(Euclide(Points[i,],i01),Euclide(Points[i,],i02),Euclide(Points[i,],i03))==Euclide(i02, Points[i,]))
        {points(Points[i,1],Points[i,2],col=2,pch=19);
segments(Points[i,1],Points[i,2],i02[1],i02[2],lwd=2,col=2)}
                else{
                if (min(Euclide(Points[i,],i01),Euclide(Points[i,],i02),Euclide(Points[i,],i03))==Euclide(i03,
Points[i,]))
                {points(Points[i,1],Points[i,2],col=3,pch=19);
segments(Points[i,1],Points[i,2],i03[1],i03[2],lwd=2,col=3)}
                }
        }
}
points(i01[1],i01[2],col=1, pch=4,lwd=5)
points(i02[1],i02[2],col=2, pch=4,lwd=5)
points(i03[1],i03[2],col=3, pch=4,lwd=5)

plot((min(Points,i01,i02,i03)-1):(max(Points,i01,i02,i03)+1),(min(Points,i01,i02,i03)-
1):(max(Points,i01,i02,i03)+1), type="n", panel.first = grid(5,5),frame.plot=T, axes=T, xlab="x
coordinates",ylab="y coordinates")
for(i in 1:N){
if
(min(mahalanobis(Points[i,],i01,cov(Points)),mahalanobis(Points[i,],i02,cov(Points)),mahalanobis(Points[i,],i03,
cov(Points)))==mahalanobis(i01, Points[i,],cov(Points)))
```

```
{points(Points[i,1],Points[i,2],col=1,pch=19); segments(Points[i,1],Points[i,2],i01[1],i01[2],lwd=2,col=1)}
        else{
        if
(min(mahalanobis(Points[i,],i01,cov(Points)),mahalanobis(Points[i,],i02,cov(Points)),mahalanobis(Points[i,],i03,
cov(Points)))==mahalanobis(i02, Points[i,],cov(Points)))
        {points(Points[i,1],Points[i,2],col=2,pch=19);
segments(Points[i,1],Points[i,2],i02[1],i02[2],lwd=2,col=2)}
                else{
                if
(min(mahalanobis(Points[i,],i01,cov(Points)),mahalanobis(Points[i,],i02,cov(Points)),mahalanobis(Points[i,],i03,
cov(Points)))==mahalanobis(i03, Points[i,],cov(Points)))
                {points(Points[i,1],Points[i,2],col=3,pch=19);
segments(Points[i,1],Points[i,2],i03[1],i03[2],lwd=2,col=3)}
                }
        }
}
points(i01[1],i01[2],col=1, pch=4,lwd=5)
points(i02[1],i02[2],col=2, pch=4,lwd=5)
points(i03[1],i03[2],col=3, pch=4,lwd=5)
# ----- End of the R script ----- #
```

Annex B

```
# to initialize parameters
alleles=c(1,2)
nbtrios=100
MAF=c(0.3,0.3,0.3)
fm=sort(rep(1:nbtrios,3))
offspring=rep(c(0,0,1),nbtrios)
ch=1:(3*nbtrios)
fa=NULL
mo=NULL
sex=NULL

fal1a1=fal1a2=NULL
fal2a1=fal2a2=NULL
fal3a1=fal3a2=NULL

mol1a1=mol1a2=NULL
mol2a1=mol2a2=NULL
mol3a1=mol3a2=NULL

chl1a1=chl1a2=NULL
chl2a1=chl2a2=NULL
chl3a1=chl3a2=NULL

# to generate a sample
p=seq(1,3*nbtrios,3)

for (i in 1:length(p))
{
        fa=c(fa,c(0,0,p[i]))
        mo=c(mo,c(0,0,p[i]+1))
        sex=c(sex,1,2,sample(c(1,2),1))

        gfa=sort(sample(alleles,2,replace=TRUE, prob=c(1-MAF[1],MAF[1])))
        gmo=sort(sample(alleles,2,replace=TRUE, prob=c(1-MAF[1],MAF[1])))
        gch=sort(c(sample(gfa,1),sample(gmo,1)))
        fal1a1=c(fal1a1,0,0,gfa[1]); fal1a2=c(fal1a2,0,0,gfa[2])
        mol1a1=c(mol1a1,0,0,gmo[1]); mol1a2=c(mol1a2,0,0,gmo[2])
        chl1a1=c(chl1a1,gfa[1],gmo[1],gch[1]); chl1a2=c(chl1a2,gfa[2],gmo[2],gch[2])

        gfa=sort(sample(alleles,2,replace=TRUE, prob=c(1-MAF[2],MAF[2])))
        gmo=sort(sample(alleles,2,replace=TRUE, prob=c(1-MAF[2],MAF[2])))
        gch=sort(c(sample(gfa,1),sample(gmo,1)))
        fal2a1=c(fal2a1,0,0,gfa[1]); fal2a2=c(fal2a2,0,0,gfa[2])
        mol2a1=c(mol2a1,0,0,gmo[1]); mol2a2=c(mol2a2,0,0,gmo[2])
        chl2a1=c(chl2a1,gfa[1],gmo[1],gch[1]); chl2a2=c(chl2a2,gfa[2],gmo[2],gch[2])

        gfa=sort(sample(alleles,2,replace=TRUE, prob=c(1-MAF[3],MAF[3])))
        gmo=sort(sample(alleles,2,replace=TRUE, prob=c(1-MAF[3],MAF[3])))
        gch=sort(c(sample(gfa,1),sample(gmo,1)))
```

```
        fal3a1=c(fal3a1,0,0,gfa[1]); fal3a2=c(fal3a2,0,0,gfa[2])
        mol3a1=c(mol3a1,0,0,gmo[1]); mol3a2=c(mol3a2,0,0,gmo[2])
        chl3a1=c(chl3a1,gfa[1],gmo[1],gch[1]); chl3a2=c(chl3a2,gfa[2],gmo[2],gch[2])
}

locus1=paste(chl1a1,chl1a2)
locus2=paste(chl2a1,chl2a2)
locus3=paste(chl3a1,chl3a2)


phen1=phen2=phen3=NA

phen1[offspring==1 & locus1=="1 1"]=rbinom(length(offspring[offspring==1 & locus1=="1 1"]),1,0.20)
phen1[offspring==1 & locus1=="1 2"]=rbinom(length(offspring[offspring==1 & locus1=="1 2"]),1,0.70)
phen1[offspring==1 & locus1=="2 2"]=rbinom(length(offspring[offspring==1 & locus1=="2 2"]),1,0.90)

phen2[offspring==1 & locus2=="1 1"]=rbinom(length(offspring[offspring==1 & locus2=="1 1"]),1,0.20)
phen2[offspring==1 & locus2=="1 2"]=rbinom(length(offspring[offspring==1 & locus2=="1 2"]),1,0.70)
phen2[offspring==1 & locus2=="2 2"]=rbinom(length(offspring[offspring==1 & locus2=="2 2"]),1,0.90)

phen3[offspring==1 & locus3=="1 1"]=rbinom(length(offspring[offspring==1 & locus3=="1 1"]),1,0.20)
phen3[offspring==1 & locus3=="1 2"]=rbinom(length(offspring[offspring==1 & locus3=="1 2"]),1,0.70)
phen3[offspring==1 & locus3=="2 2"]=rbinom(length(offspring[offspring==1 & locus3=="2 2"]),1,0.90)


phen=NULL
phen[(phen1+phen2+phen3)==0 | (phen1+phen2+phen3)==1]=0
phen[(phen1+phen2+phen3)==2 | (phen1+phen2+phen3)==3]=1

SimulatedData2=data.frame(fm,ch,fa,mo,sex,phen,locus1,locus2,locus3)
SimulatedData3=data.frame(fm,ch,fa,mo,sex,phen,chl1a1,chl1a2, chl1a1,chl2a2, chl3a1,chl3a2,fal1a1,fal1a2,
fal1a1,fal2a2, fal3a1,fal3a2,mol1a1,mol1a2, mol1a1,mol2a2, mol3a1,mol3a2)

# to save simulated data on a file for further use on FBAT after some changes in format and column names
write.table(SimulatedData2, file="C:/ ... give the path here ... /SimulatedData2.txt", sep="\t", quote=F,
row.names=F, col.names=T)
```

Annex C

```
# ----- Beginning of the R script ----- #
# ------------------------------------------------------ #
# ---------- CLEAN OBJECTS AND LOAD PACKAGES
# ------------------------------------------------------ #
ls()
rm(list=ls())
library(foreign)

# ------------------------------------------------------ #
# ---------- LOAD DATA FILES ---------------- #
# ------------------------------------------------------ #
mydata=read.dta("C:/mydata.dta")

# ------------------------------------------------------ #
# ------- CHOICE OF MARKERS TO ANALYZE AND PHENOTYPE
# ------------------------------------------------------ #
mydata$idlocus1=mydata$ae1_20_21
mydata$idlocus2=mydata$xmn1
mydata$idlocus3=mydata$ae1_117_118
mydata$idlocus4=mydata$abo297
mydata$idlocus5=mydata$ae1_174_187
mydata$phen=mydata$pfaidbin
l=5
locus_on_X=c(0) # Put between brackets the list number of loci localized on X chromosome, separated by ",".

# ------------------------------------------------------ #
# ---------- DATA FRAME OF GENOTYPES
# ------------------------------------------------------ #

father=data.frame(unique(mydata$fatherid))
names(father)=c("id")
father=unique(merge(father,mydata[,c("fm","id",paste("idlocus",1:l,sep=""))], by="id"))
names(father)=c("fatherid","fm",paste("falocus",1:l,sep=""))

mother=data.frame(unique(mydata$motherid))
names(mother)=c("id")
mother=unique(merge(mother,mydata[,c("fm","id",paste("idlocus",1:l,sep=""))], by="id"))
names(mother)=c("motherid","fm",paste("molocus",1:l,sep=""))

gendata=unique(mydata[,c("fm","id","fatherid","motherid","sex","phen",paste("idlocus",1:l,sep=""))])
gendata=merge(gendata, father, by=c("fm","fatherid"), all.x=T)
gendata=merge(gendata, mother, by=c("fm","motherid"), all.x=T)
rm(father,mother)

gendata[,(dim(gendata)[2]-3*l+1):dim(gendata)[2]][is.na(gendata[,(dim(gendata)[2]-
3*l+1):dim(gendata)[2]])==TRUE]="0 0"
gendata=gendata[,c("fm","id","fatherid","motherid","sex","phen",paste("idlocus",1:l,sep=""),paste("falocus",1:l,
sep=""),paste("molocus",1:l,sep=""))]
```

```
gendata2=unique(gendata[is.na(gendata$phen)==FALSE & gendata$phen==0 &                    # 0 to select
resistant and 1 to select suceptible
                              gendata$falocus1!="0 0" & gendata$molocus1!="0 0" & gendata$idlocus1!="0 0" &
                              gendata$falocus2!="0 0" & gendata$molocus2!="0 0" & gendata$idlocus2!="0 0" &
                              gendata$falocus3!="0 0" & gendata$molocus3!="0 0" & gendata$idlocus3!="0 0" &
                              gendata$falocus4!="0 0" & gendata$molocus4!="0 0" & gendata$idlocus4!="0 0" &
                              gendata$falocus5!="0 0" & gendata$molocus5!="0 0" & gendata$idlocus5!="0 0" ,
                              c("falocus1","molocus1","idlocus1",
                                "falocus2","molocus2","idlocus2",
                                "falocus3","molocus3","idlocus3",
                                "falocus4","molocus4","idlocus4",
                                "falocus5","molocus5","idlocus5",
                                "fatherid","motherid","id","sex")])

# ------------------------------------------------------ #
write.table(gendata2, file="C:/gendata2.txt", sep=" ", quote=F, row.names=F, col.names=F)
gendata2=read.table("C:/gendata2.txt", sep=" ")
names(gendata2)=c("fal1a1","fal1a2","mol1a1","mol1a2","chl1a1","chl1a2",
                              "fal2a1","fal2a2","mol2a1","mol2a2","chl2a1","chl2a2",
                              "fal3a1","fal3a2","mol3a1","mol3a2","chl3a1","chl3a2",
                              "fal4a1","fal4a2","mol4a1","mol4a2","chl4a1","chl4a2",
                              "fal5a1","fal5a2","mol5a1","mol5a2","chl5a1","chl5a2",
                              "father","mother","child","sex")

# ------------------------------------------------------ #
# ---------- LISTE OF POSSIBLE K-UPLET - #
# ------------------------------------------------------ #

nbloci=5
nballeles=2
taballeles=matrix(NA,nbloci,nballeles)
rownames(taballeles)=c(paste("locus",1:nbloci,sep=""))
colnames(taballeles)=c(paste("allele",1:nballeles,sep=""))
for (l in 1:nbloci){
for (a in 1:nballeles){
taballeles[l,a]=unique(sort(c(as.matrix(gendata2[,((l-1)*6+1):(6*l)])))))[a]
}}
rm(a,l)

kuplet=NULL
l=0
for (l1 in taballeles[1,]){
for (l2 in taballeles[2,]){
for (l3 in taballeles[3,]){
for (l4 in taballeles[4,]){
for (l5 in taballeles[5,]){
        if (is.na(l1)==FALSE & is.na(l2)==FALSE & is.na(l3)==FALSE & is.na(l4)==FALSE &
is.na(l5)==FALSE){
                l=l+1
                kuplet[l]=paste(l1,l2,l3,l4,l5, sep="")
        }
}}}}}
nbkuplet=length(kuplet)
rm(l,l1,l2,l3,l4,l5)

# ------------------------------------------------------ #
```

```
# SIMULATION OF POSSIBLE CHILDREN FOR AMBIGUOUS TRANSMISSIONS #
# ---------------------------------------------------------- #
gendata2$countw=1
gendata2$realchild=1

for (n in 1:nrow(gendata2)){
nbdoubt=0
locusdoubt=0
for (l in 1:nbloci){
        if (gendata2[n,6*(l-1)+1]==gendata2[n,6*(l-1)+3] & gendata2[n,6*(l-1)+1]==gendata2[n,6*(l-1)+5] &
           gendata2[n,6*(l-1)+2]==gendata2[n,6*(l-1)+4] & gendata2[n,6*(l-1)+2]==gendata2[n,6*(l-1)+6] &
           gendata2[n,6*(l-1)+1]!=gendata2[n,6*(l-1)+2] &
           gendata2$realchild[n]==1){
        nbdoubt=nbdoubt+1
        locusdoubt[nbdoubt]=l
        }
}

if (nbdoubt>0) {
gensimchild=NULL
for (p in (nbdoubt-1):0) {
gensimchild = c(gensimchild,rep(c(rep(1,2^p),rep(2,2^p)),2^(nbdoubt-p-1)))
}
gensimchild=matrix(gensimchild,2^nbdoubt,nbdoubt)
for (i in 1:2^nbdoubt){
gendata2=rbind(gendata2,gendata2[n,])
gendata2[nrow(gendata2),6*(locusdoubt-1)+5]=gensimchild[i,]
gendata2[nrow(gendata2),6*(locusdoubt-1)+6]=gensimchild[i,]
}
gendata2$countw[n]=0
gendata2$countw[(nrow(gendata2)-2^nbdoubt+1):nrow(gendata2)]=1/2^nbdoubt
gendata2$realchild[(nrow(gendata2)-2^nbdoubt+1):nrow(gendata2)]=0
}}
rm(n,l,p,i)

# ------------------------------------------------------ #
# --- TO COMPUTE MATRIX OF SUMULTANEOUS TRANSMISSION
# ------------------------------------------------------ #
transmat=matrix(0,length(kuplet),length(kuplet))
rownames(transmat)=kuplet
colnames(transmat)=kuplet

for (n in 1:nrow(gendata2)){
kuplet_fa_T=NULL; kuplet_mo_T=NULL; kuplet_fa_NT=NULL; kuplet_mo_NT=NULL

for (l in 1:nbloci){

if ((length(setdiff(locus_on_X,l))==length(locus_on_X)) | (length(setdiff(locus_on_X,l))!=length(locus_on_X)
& gendata2$sex[n]==2)){

for (i in taballeles[l,][is.na(taballeles[l,])==FALSE]){
for (j in i:max(taballeles[l,][is.na(taballeles[l,])==FALSE])){
for (u in taballeles[l,][is.na(taballeles[l,])==FALSE]){
for (v in u:max(taballeles[l,][is.na(taballeles[l,])==FALSE])){
```

```
        if (gendata2[n,6*(l-1)+1]==i & gendata2[n,6*(l-1)+2]==j & gendata2[n,6*(l-1)+3]==u &
gendata2[n,6*(l-1)+4]==v & ((gendata2[n,6*(l-1)+5]==i & gendata2[n,6*(l-1)+6]==u) | (gendata2[n,6*(l-
1)+5]==u & gendata2[n,6*(l-1)+6]==i))){
        kuplet_fa_T=paste(kuplet_fa_T,i, sep=""); kuplet_mo_T=paste(kuplet_mo_T,u, sep="");
kuplet_fa_NT=paste(kuplet_fa_NT,j, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,v, sep="")}
        else {
                if (gendata2[n,6*(l-1)+1]==i & gendata2[n,6*(l-1)+2]==j & gendata2[n,6*(l-1)+3]==u &
gendata2[n,6*(l-1)+4]==v & ((gendata2[n,6*(l-1)+5]==i & gendata2[n,6*(l-1)+6]==v) | (gendata2[n,6*(l-
1)+5]==v & gendata2[n,6*(l-1)+6]==i))){
                kuplet_fa_T=paste(kuplet_fa_T,i, sep=""); kuplet_mo_T=paste(kuplet_mo_T,v, sep="");
kuplet_fa_NT=paste(kuplet_fa_NT,j, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,u, sep="")}
                else {
                        if (gendata2[n,6*(l-1)+1]==i & gendata2[n,6*(l-1)+2]==j & gendata2[n,6*(l-
1)+3]==u & gendata2[n,6*(l-1)+4]==v & ((gendata2[n,6*(l-1)+5]==j & gendata2[n,6*(l-1)+6]==u) |
(gendata2[n,6*(l-1)+5]==u & gendata2[n,6*(l-1)+6]==j))){
                        kuplet_fa_T=paste(kuplet_fa_T,j, sep=""); kuplet_mo_T=paste(kuplet_mo_T,u,
sep=""); kuplet_fa_NT=paste(kuplet_fa_NT,i, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,v, sep="")}
                        else {
                                if (gendata2[n,6*(l-1)+1]==i & gendata2[n,6*(l-1)+2]==j & gendata2[n,6*(l-
1)+3]==u & gendata2[n,6*(l-1)+4]==v & ((gendata2[n,6*(l-1)+5]==j & gendata2[n,6*(l-1)+6]==v) |
(gendata2[n,6*(l-1)+5]==v & gendata2[n,6*(l-1)+6]==j))){
                                kuplet_fa_T=paste(kuplet_fa_T,j, sep="");
kuplet_mo_T=paste(kuplet_mo_T,v, sep=""); kuplet_fa_NT=paste(kuplet_fa_NT,i, sep="");
kuplet_mo_NT=paste(kuplet_mo_NT,u, sep="")}
                                }
                        }
                }
}}}}
}

if (length(setdiff(locus_on_X,l))!=length(locus_on_X) & gendata2$sex[n]==1){

if (gendata2[n,6*(l-1)+1]==1 & gendata2[n,6*(l-1)+2]==1 & gendata2[n,6*(l-1)+3]==1 & gendata2[n,6*(l-
1)+4]==2 & gendata2[n,6*(l-1)+5]==2 & gendata2[n,6*(l-1)+6]==2){
        kuplet_fa_T=paste(kuplet_fa_T,1, sep=""); kuplet_mo_T=paste(kuplet_mo_T,2, sep="");
kuplet_fa_NT=paste(kuplet_fa_NT,1, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,1, sep="")}

if (gendata2[n,6*(l-1)+1]==1 & gendata2[n,6*(l-1)+2]==1 & gendata2[n,6*(l-1)+3]==1 & gendata2[n,6*(l-
1)+4]==2 & gendata2[n,6*(l-1)+5]==1 & gendata2[n,6*(l-1)+6]==1){
        kuplet_fa_T=paste(kuplet_fa_T,1, sep=""); kuplet_mo_T=paste(kuplet_mo_T,1, sep="");
kuplet_fa_NT=paste(kuplet_fa_NT,1, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,2, sep="")}

if (gendata2[n,6*(l-1)+1]==2 & gendata2[n,6*(l-1)+2]==2 & gendata2[n,6*(l-1)+3]==1 & gendata2[n,6*(l-
1)+4]==1 & gendata2[n,6*(l-1)+5]==1 & gendata2[n,6*(l-1)+6]==1){
        kuplet_fa_T=paste(kuplet_fa_T,1, sep=""); kuplet_mo_T=paste(kuplet_mo_T,1, sep="");
kuplet_fa_NT=paste(kuplet_fa_NT,1, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,1, sep="")}

if (gendata2[n,6*(l-1)+1]==2 & gendata2[n,6*(l-1)+2]==2 & gendata2[n,6*(l-1)+3]==1 & gendata2[n,6*(l-
1)+4]==2 & gendata2[n,6*(l-1)+5]==2 & gendata2[n,6*(l-1)+6]==2){
        kuplet_fa_T=paste(kuplet_fa_T,1, sep=""); kuplet_mo_T=paste(kuplet_mo_T,2, sep="");
kuplet_fa_NT=paste(kuplet_fa_NT,1, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,1, sep="")}

if (gendata2[n,6*(l-1)+1]==2 & gendata2[n,6*(l-1)+2]==2 & gendata2[n,6*(l-1)+3]==1 & gendata2[n,6*(l-
1)+4]==2 & gendata2[n,6*(l-1)+5]==1 & gendata2[n,6*(l-1)+6]==1){
        kuplet_fa_T=paste(kuplet_fa_T,1, sep=""); kuplet_mo_T=paste(kuplet_mo_T,1, sep="");
kuplet_fa_NT=paste(kuplet_fa_NT,1, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,2, sep="")}
```

```
if (gendata2[n,6*(l-1)+1]==1 & gendata2[n,6*(l-1)+2]==1 & gendata2[n,6*(l-1)+3]==2 & gendata2[n,6*(l-1)+4]==2 & gendata2[n,6*(l-1)+5]==2 & gendata2[n,6*(l-1)+6]==2){
	kuplet_fa_T=paste(kuplet_fa_T,1, sep=""); kuplet_mo_T=paste(kuplet_mo_T,2, sep="");
kuplet_fa_NT=paste(kuplet_fa_NT,1, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,2, sep="")}

if (gendata2[n,6*(l-1)+1]==2 & gendata2[n,6*(l-1)+2]==2 & gendata2[n,6*(l-1)+3]==2 & gendata2[n,6*(l-1)+4]==2 & gendata2[n,6*(l-1)+5]==2 & gendata2[n,6*(l-1)+6]==2){
	kuplet_fa_T=paste(kuplet_fa_T,1, sep=""); kuplet_mo_T=paste(kuplet_mo_T,2, sep="");
kuplet_fa_NT=paste(kuplet_fa_NT,1, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,2, sep="")}

if (gendata2[n,6*(l-1)+1]==1 & gendata2[n,6*(l-1)+2]==1 & gendata2[n,6*(l-1)+3]==1 & gendata2[n,6*(l-1)+4]==1 & gendata2[n,6*(l-1)+5]==1 & gendata2[n,6*(l-1)+6]==1){
	kuplet_fa_T=paste(kuplet_fa_T,1, sep=""); kuplet_mo_T=paste(kuplet_mo_T,1, sep="");
kuplet_fa_NT=paste(kuplet_fa_NT,1, sep=""); kuplet_mo_NT=paste(kuplet_mo_NT,1, sep="")}
}}

if (length(setdiff(kuplet,kuplet_fa_T))!=length(kuplet) & length(setdiff(kuplet,kuplet_fa_NT))!=length(kuplet) & length(setdiff(kuplet,kuplet_mo_T))!=length(kuplet) &
length(setdiff(kuplet,kuplet_mo_NT))!=length(kuplet)){
transmat[kuplet_fa_T,kuplet_fa_NT] = transmat[kuplet_fa_T,kuplet_fa_NT] + gendata2$countw[n]
transmat[kuplet_mo_T,kuplet_mo_NT] = transmat[kuplet_mo_T,kuplet_mo_NT] + gendata2$countw[n]}
}
rm(i,j,l,n,u,v,kuplet_fa_NT,kuplet_fa_T,kuplet_mo_NT,kuplet_mo_T)

sum(transmat)	# this has always to be equal to 2*number of offspring analyzed, i.e. the number of row of the dataset "gendata2"

# ------------------------------------------------------ #
# - TRANSMISSION INTENSITY OF ALLELES AT SINGLE LOCUS -- #
# ------------------------------------------------------ #
alpha=matrix(NA,nbloci,nballeles)
rownames(alpha)=c(paste("locus",1:nbloci,sep=""))
colnames(alpha)=c(paste("allele",1:nballeles,sep=""))

for (l in 1:nbloci){
for (a in 1:nballeles){
	alpha[l,a]=sum(transmat[substr(rownames(transmat),l,l)==paste(a),substr(colnames(transmat),l,l)!=paste(a)])/(sum(transmat[substr(rownames(transmat),l,l)==paste(a),substr(colnames(transmat),l,l)!=paste(a)])+sum(transmat[substr(rownames(transmat),l,l)!=paste(a),substr(colnames(transmat),l,l)==paste(a)]))
}}
rm(a,l)

# ------------------------------------------------------ #
# ----- NUMBER OF TRANSMITTED AND NOT-TRANSMITTED ------ #
# ------------------------------------------------------ #
k=0
nT=0
nNT=0
for (i in 1:(dim(transmat)[2]-1)){
for (j in (i+1):dim(transmat)[1]){
k=k+1
nT[k]=transmat[i,j]
nNT[k]=transmat[j,i]
}}
rm(i,j,k)
```

```
# ------------------------------------------------------------ #
# ----- The Number of possible alternative hypotheses
# ----- but we will ignore combination of loci over 3
# ------------------------------------------------------------ #
fact=function(m){
 fm=1
 while (m>=2){
  fm=fm*m
  m=m-1}
 return(fm)}
# ------------------------------------------------------------ #
comb=function(n,p){
 while (n>=p){
  return(fact(n)/(fact(p)*fact(n-p)))}}
# ------------------------------------------------------------ #
nbmodel= nbloci+2
if(nbloci>=2){
for (i in 2:3){
 nbmodel= nbmodel+ 2*comb(nbloci,i)
}}
nbmodel
rm(comb,fact,i)

# ---------------------------------------------------------- #
# ------- SINGLE TRANSMISSION PROBABILITIES ------------ #
# ---------------------------------------------------------- #
tau=matrix(0,nbmodel,length(nT))
ddl=0
for (m in 1:nbloci){
ddl[m]=length(taballeles[m,])-1
k=0
for (i in 1:(nbkuplet-1)){
for (j in (i+1):nbkuplet){
            k=k+1
            a=as.numeric(substr(rownames(transmat)[i],m,m))
            b=as.numeric(substr(colnames(transmat)[j],m,m))
            tau[m,k]=alpha[m,a]/(alpha[m,a]+alpha[m,b])
}}}
rm(a,b,i,j,k)

# ---------------------------------------------------------- #
# ------- 2-UPLET TRANSMISSION PROBABILITIES
# ---------------------------------------------------------- #

# MULTIPLICATIVE #
for (lm in 1:(nbloci-1)){
for (ln in (lm+1):nbloci){
m=m+1
ddl[m]=length(taballeles[lm,])-1 + length(taballeles[ln,])-1
k=0
for (i in 1:(nbkuplet-1)){
for (j in (i+1):nbkuplet){
            k=k+1
            a=as.numeric(substr(rownames(transmat)[i],lm,lm))
            b=as.numeric(substr(colnames(transmat)[j],lm,lm))
```

```
            c=as.numeric(substr(rownames(transmat)[i],ln,ln))
            d=as.numeric(substr(colnames(transmat)[j],ln,ln))
            tau[m,k]=alpha[lm,a]*alpha[ln,c]/(alpha[lm,a]*alpha[ln,c]+alpha[lm,b]*alpha[ln,d])
}}}}
rm(a,b,c,d,i,j,k,lm,ln)

# EPISTASIS #
for (lm in 1:(nbloci-1)){
for (ln in (lm+1):nbloci){
m=m+1
ddl[m]=length(taballeles[lm,])*length(taballeles[ln,])-1
k=0
for (i in 1:(nbkuplet-1)){
for (j in (i+1):nbkuplet){
            k=k+1
            a=as.numeric(substr(rownames(transmat)[i],lm,lm))
            b=as.numeric(substr(colnames(transmat)[j],lm,lm))
            c=as.numeric(substr(rownames(transmat)[i],ln,ln))
            d=as.numeric(substr(colnames(transmat)[j],ln,ln))

            x=sum(transmat[substr(rownames(transmat),lm,lm)==paste(a) &
substr(rownames(transmat),ln,ln)==paste(c),
                                    substr(colnames(transmat),lm,lm)!=paste(a) |
substr(colnames(transmat),ln,ln)!=paste(c)])/
             (sum(transmat[substr(rownames(transmat),lm,lm)==paste(a) &
substr(rownames(transmat),ln,ln)==paste(c),
                                    substr(colnames(transmat),lm,lm)!=paste(a) |
substr(colnames(transmat),ln,ln)!=paste(c)])
              +sum(transmat[substr(rownames(transmat),lm,lm)!=paste(a) |
substr(rownames(transmat),ln,ln)!=paste(c),
                                    substr(colnames(transmat),lm,lm)==paste(a) &
substr(colnames(transmat),ln,ln)==paste(c)]))

            y=sum(transmat[substr(rownames(transmat),lm,lm)==paste(b) &
substr(rownames(transmat),ln,ln)==paste(d),
                                    substr(colnames(transmat),lm,lm)!=paste(b) |
substr(colnames(transmat),ln,ln)!=paste(d)])/
             (sum(transmat[substr(rownames(transmat),lm,lm)==paste(b) &
substr(rownames(transmat),ln,ln)==paste(d),
                                    substr(colnames(transmat),lm,lm)!=paste(b) |
substr(colnames(transmat),ln,ln)!=paste(d)])
              +sum(transmat[substr(rownames(transmat),lm,lm)!=paste(b) |
substr(rownames(transmat),ln,ln)!=paste(d),
                                    substr(colnames(transmat),lm,lm)==paste(b) &
substr(colnames(transmat),ln,ln)==paste(d)]))

            tau[m,k]=x/(x+y)
}}}}
rm(a,b,c,d,i,j,k,x,y,lm,ln)

# ------------------------------------------------------ #
# ------- 3-UPLET TRANSMISSION PROBABILITIES#
# ------------------------------------------------------ #

# MULTIPLICATIVE #
for (lm in 1:(nbloci-2)){
```

```
for (ln in (lm+1):(nbloci-1)){
for (lo in (ln+1):nbloci){
m=m+1
ddl[m]=length(taballeles[lm,])-1 + length(taballeles[ln,])-1 + length(taballeles[lo,])-1
k=0
for (i in 1:(nbkuplet-1)){
for (j in (i+1):nbkuplet){
        k=k+1
        a=as.numeric(substr(rownames(transmat)[i],lm,lm))
        b=as.numeric(substr(colnames(transmat)[j],lm,lm))
        c=as.numeric(substr(rownames(transmat)[i],ln,ln))
        d=as.numeric(substr(colnames(transmat)[j],ln,ln))
        e=as.numeric(substr(rownames(transmat)[i],lo,lo))
        f=as.numeric(substr(colnames(transmat)[j],lo,lo))
        tau[m,k]=alpha[lm,a]*alpha[ln,c]*alpha[lo,e]/(alpha[lm,a]*alpha[ln,c]*alpha[lo,e] +
alpha[lm,b]*alpha[ln,d]*alpha[lo,f])
}}}}}
rm(a,b,c,d,e,f,i,j,k,lm,ln,lo)

# EPISTASIS #
for (lm in 1:(nbloci-2)){
for (ln in (lm+1):(nbloci-1)){
for (lo in (ln+1):nbloci){
m=m+1
ddl[m]=length(taballeles[lm,])*length(taballeles[ln,])*length(taballeles[lo,]) -1
k=0
for (i in 1:(nbkuplet-1)){
for (j in (i+1):nbkuplet){
        k=k+1
        a=as.numeric(substr(rownames(transmat)[i],lm,lm))
        b=as.numeric(substr(colnames(transmat)[j],lm,lm))
        c=as.numeric(substr(rownames(transmat)[i],ln,ln))
        d=as.numeric(substr(colnames(transmat)[j],ln,ln))
        e=as.numeric(substr(rownames(transmat)[i],lo,lo))
        f=as.numeric(substr(colnames(transmat)[j],lo,lo))

        x=sum(transmat[substr(rownames(transmat),lm,lm)==paste(a) &
substr(rownames(transmat),ln,ln)==paste(c) & substr(rownames(transmat),lo,lo)==paste(e),
                substr(colnames(transmat),lm,lm)!=paste(a) |
substr(colnames(transmat),ln,ln)!=paste(c) | substr(colnames(transmat),lo,lo)!=paste(e)])/
          (sum(transmat[substr(rownames(transmat),lm,lm)==paste(a) &
substr(rownames(transmat),ln,ln)==paste(c) & substr(rownames(transmat),lo,lo)==paste(e),
                substr(colnames(transmat),lm,lm)!=paste(a) |
substr(colnames(transmat),ln,ln)!=paste(c) | substr(colnames(transmat),lo,lo)!=paste(e)])
          +sum(transmat[substr(rownames(transmat),lm,lm)!=paste(a) |
substr(rownames(transmat),ln,ln)!=paste(c) | substr(rownames(transmat),lo,lo)!=paste(e),
                substr(colnames(transmat),lm,lm)==paste(a) &
substr(colnames(transmat),ln,ln)==paste(c) & substr(colnames(transmat),lo,lo)==paste(e)]))

        y=sum(transmat[substr(rownames(transmat),lm,lm)==paste(b) &
substr(rownames(transmat),ln,ln)==paste(d) & substr(rownames(transmat),lo,lo)==paste(f),
                substr(colnames(transmat),lm,lm)!=paste(b) |
substr(colnames(transmat),ln,ln)!=paste(d) | substr(colnames(transmat),lo,lo)!=paste(f)])/
          (sum(transmat[substr(rownames(transmat),lm,lm)==paste(b) &
substr(rownames(transmat),ln,ln)==paste(d) & substr(rownames(transmat),lo,lo)==paste(f),
```

```
                                substr(colnames(transmat),lm,lm)!=paste(b) |
substr(colnames(transmat),ln,ln)!=paste(d) | substr(colnames(transmat),lo,lo)!=paste(f)])
            +sum(transmat[substr(rownames(transmat),lm,lm)!=paste(b) |
substr(rownames(transmat),ln,ln)!=paste(d) | substr(rownames(transmat),lo,lo)!=paste(f),
                                substr(colnames(transmat),lm,lm)==paste(b) &
substr(colnames(transmat),ln,ln)==paste(d) & substr(colnames(transmat),lo,lo)==paste(f)]))
        tau[m,k]=x/(x+y)
}}}}}
rm(a,b,c,d,e,f,i,j,k,x,y,lm,ln,lo)

# ------------------------------------------------------ #
# ------- L-UPLET TRANSMISSION PROBABILITIES
# ------------------------------------------------------ #

# MULTIPLICATIVE #
m=m+1
k=a=b=0
for (i in 1:(nbkuplet-1)){
for (j in (i+1):nbkuplet){
        k=k+1
        x=y=1
        df=0
        for (l in 1:nbloci){
        a[l]=as.numeric(substr(rownames(transmat)[i],l,l))
        b[l]=as.numeric(substr(colnames(transmat)[j],l,l))
        x=x*alpha[l,a[l]]
        y=y*alpha[l,b[l]]
        df = df + length(taballeles[l,])-1
                                        }
        tau[m,k]=x/(x+y)
        ddl[m]=df
}}
rm(a,b,i,j,k,l,x,y,df)

# EPISTASIS #
m=m+1
k=0
for (i in 1:(nbkuplet-1)){
for (j in (i+1):nbkuplet){
        k=k+1
        tau[m,k]=(sum(transmat[i,-i])/(sum(transmat[i,-i])+sum(transmat[-i,i])))/((sum(transmat[i,-
i])/(sum(transmat[i,-i])+sum(transmat[-i,i])))+(sum(transmat[j,-j])/(sum(transmat[j,-j])+sum(transmat[-j,j]))))
}}
rm(i,j,k)

df=1
for (l in 1:nbloci){df=df*length(taballeles[l,])}
ddl[m]=df-1
rm(l,df)

# ------------------------------------------------------ #
# ------------------ LOG-LIKELIHOODS ------- #
# ------------------------------------------------------ #
# MODEL0: WHITE MODEL
LL0= -log(2)*sum(nT+nNT)
LL0
```

```
# MODEL 1 to m
LL=0
ETDT=0
pvalETDT=0
for (l in 1:m){
        LL[l]= sum(nT*log(tau[l,]/(1-tau[l,]))) + sum((nT+nNT)*log(1-tau[l,]))
        ETDT[l]=2*(LL[l]-LL0)
        pvalETDT[l]=1-pchisq(ETDT[l],ddl[l])
}
rm(l)

# ------------------------------------------------------ #
# ---------------- TO DISPLAY RESULTS ----- #
# ------------------------------------------------------ #
models=c("L1","L2","L3","L4","L5","Multiplicative_L1L2","Multiplicative_L1L3","Multiplicative_L1L4","Mu
ltiplicative_L1L5","Multiplicative_L2L3","Multiplicative_L2L4","Multiplicative_L2L5","Multiplicative_L3L4"
,"Multiplicative_L3L5","Multiplicative_L4L5","Epistasis_L1L2","Epistasis_L1L3","Epistasis_L1L4","Epistasis
_L1L5","Epistasis_L2L3","Epistasis_L2L4","Epistasis_L2L5","Epistasis_L3L4","Epistasis_L3L5","Epistasis_L
4L5","Multiplicative_L1L2L3","Multiplicative_L1L2L4","Multiplicative_L1L2L5","Multiplicative_L1L3L4","
Multiplicative_L1L3L5","Multiplicative_L1L4L5","Multiplicative_L2L3L4","Multiplicative_L2L3L5","Multipl
icative_L2L4L5","Multiplicative_L3L4L5","Epistasis_L1L2L3","Epistasis_L1L2L4","Epistasis_L1L2L5","Epis
tasis_L1L3L4","Epistasis_L1L3L5","Epistasis_L1L4L5","Epistasis_L2L3L4","Epistasis_L2L3L5","Epistasis_L
2L4L5","Epistasis_L3L4L5","Multiplicative_L1L2L3L4L5","Epistasis_L1L2L3L4L5")

Result_of_thesis=data.frame(models,LL,ETDT,ddl,pvalETDT)
Result_of_thesis
# ----- End of the R script ----- #
```

Publications

Publications in relation to the thesis

- **Loucoubar C**, Paul R, Bar-Hen A, Huret A, Tall A, Sokhna C, trape JF, Ly AB, Faye J, Badiane A, Diakhaby G, Sarr FD, Diop A, Sakuntabhai A, Bureau JF: "An Exhaustive, Non-Euclidean, Non-Parametric Data Mining Tool for Unraveling the Complexity of Biological Systems – Novel Insights into Malaria", 2011. Plos ONE.

- **Loucoubar C**, Goncalves B, Tall A, Sokhna C, Trape JF, Sarr FD, Faye J, Badiane A, Ly AB, Diop A, Bar-Hen A, Sakuntabhai A, Paul R: "Impact of changing drug treatment and malaria endemicity on the heritability of malaria phenotypes in a longitudinal family-based cohort study", 2011. Plos ONE.

Other publications

- Berghout J, Higgins S, **Loucoubar C**, Sakuntabhai A, Kain KC and Gros P: "Genetic diversity in human erythrocyte pyruvate kinase", 2011. Genes and Immunity.

- Book Chapter:
  Machado A, **Loucoubar C**, Grange L, Bureau JF, Sakuntabhai A, Paul R: "Human genetic contribution to the outcome of infection with malaria parasites", 2012, in MALARIA PARASITES, ISBN: 979-953-307-072-7, Editor: Pr. Omolade Okwa, Publisher: INTECH.

Conferences and Seminars

- "3 rd EMBO Conference on Host Genetic Control of Infectious diseases": Evidence of gene-environmental interaction on the risk of *Plasmodium falciparum* attacks in a highly endemic area of Senegal; 28^{th} – 30^{th} September, 2011. Institut Pasteur, Paris.

- "Evolution of heritability of *Plasmodium falciparum* gametocytes carriage and *Plasmodium falciparum* density with changes in treatment protocols"; 12^{th} – 13^{th} May, 2011. Hôtel DIEU, Paris.

- "5th MIM (Multilateral Initiative on Malaria) Conference"; 02^{nd} – 07^{th} November, 2009. Nairobi (Kenya).

OPEN ACCESS Freely available online

PLoS one

An Exhaustive, Non-Euclidean, Non-Parametric Data Mining Tool for Unraveling the Complexity of Biological Systems – Novel Insights into Malaria

Cheikh Loucoubar[1,2,3], Richard Paul[1], Avner Bar-Hen[2,4], Augustin Huret[5], Adama Tall[3], Cheikh Sokhna[6], Jean-François Trape[6], Alioune Badara Ly[3], Joseph Faye[3], Abdoulaye Badiane[3], Gaoussou Diakhaby[3], Fatoumata Diène Sarr[3], Aliou Diop[7], Anavaj Sakuntabhai[1,8]*, Jean-François Bureau[1]

1 Institut Pasteur, Unité de Pathogénie Virale, Paris, France, **2** Laboratoire de Mathématiques Appliquées Paris 5 (UMR 8145), Université Paris Descartes, Paris, France, **3** Unité d'Épidémiologie des Maladies Infectieuses (UR 172), Institut Pasteur de Dakar, Dakar, Sénégal, **4** Ecole des Hautes Etudes en Santé Publique, Rennes, France, **5** Institute of Health and Science, Paris, France, **6** Unité de Paludologie Afro-Tropicale (UMR 198), Institut de Recherche pour le Développement, Dakar, Sénégal, **7** Laboratoire d'Études et de Recherche en Statistique et Développement, UGB, Saint-Louis, Sénégal, **8** Center of Excellence for Vectors and Vector-Borne Diseases, Faculty of Science, Mahidol University, Bangkok, Thailand

Abstract

Complex, high-dimensional data sets pose significant analytical challenges in the post-genomic era. Such data sets are not exclusive to genetic analyses and are also pertinent to epidemiology. There has been considerable effort to develop hypothesis-free data mining and machine learning methodologies. However, current methodologies lack exhaustivity and general applicability. Here we use a novel non-parametric, non-euclidean data mining tool, HyperCube®, to explore exhaustively a complex epidemiological malaria data set by searching for over density of events in m-dimensional space. Hotspots of over density correspond to strings of variables, rules, that determine, in this case, the occurrence of *Plasmodium falciparum* clinical malaria episodes. The data set contained 46,837 outcome events from 1,653 individuals and 34 explanatory variables. The best predictive rule contained 1,689 events from 148 individuals and was defined as: individuals present during 1992–2003, aged 1–5 years old, having hemoglobin AA, and having had previous *Plasmodium malariae* malaria parasite infection ≤10 times. These individuals had 3.71 times more *P. falciparum* clinical malaria episodes than the general population. We validated the rule in two different cohorts. We compared and contrasted the HyperCube® rule with the rules using variables identified by both traditional statistical methods and non-parametric regression tree methods. In addition, we tried all possible sub-stratified quantitative variables. No other model with equal or greater representativity gave a higher Relative Risk. Although three of the four variables in the rule were intuitive, the effect of number of *P. malariae* episodes was not. HyperCube® efficiently sub-stratified quantitative variables to optimize the rule and was able to identify interactions among the variables, tasks not easy to perform using standard data mining methods. Search of local over density in m-dimensional space, explained by easily interpretable rules, is thus seemingly ideal for generating hypotheses for large datasets to unravel the complexity inherent in biological systems.

Citation: Loucoubar C, Paul R, Bar-Hen A, Huret A, Tall A, et al. (2011) An Exhaustive, Non-Euclidean, Non-Parametric Data Mining Tool for Unraveling the Complexity of Biological Systems – Novel Insights into Malaria. PLoS ONE 6(9): e24085. doi:10.1371/journal.pone.0024085

Editor: Fabio T. M. Costa, State University of Campinas, Brazil

Received May 31, 2011; **Accepted** July 29, 2011; **Published** September 9, 2011

Funding: Funding was provided by Institut Pasteur and the Ecole des Hautes Etudes en Santé Publique. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: anavaj@pasteur.fr

Introduction

Identifying the key variables of a biological system that determine the outcome of interest is difficult. Not only are there potentially many factors involved, but they also do not work independently. Testing for all possible interactions is almost impossible both with respect to statistical validation and biological interpretation. There is a need for data mining tools to explore large and complex biological data sets to identify combinations of factors that optimally explain the outcome of interest. Hypothesis-free data exploration can potentially generate novel hypotheses that emerge from the data and which are beyond our imagination. These novel hypotheses can subsequently be tested using standard statistical methods.

To date, data mining tools have been primarily developed for data retrieval through search engines. In biology, this has been essentially focused on sequence alignment algorithms to manage the ever-increasing amount of genetic data. More recently, data mining technology has been proposed as an alternative to traditional statistics to deal with high dimensional data generated by Genome Wide Association studies, in the knowledge that accounting for gene-gene and gene-environment is crucial to understand human genetic susceptibility to disease [1,2,3,4]. In addition to such methods in the field of genetic data analyses, several new heuristic tools have been developed, notably non-parametric modeling techniques such as Classification And Regression Trees (CART) [5] and Random Forests [6]. These methods present several advantages: models have the capacity to

provide accurate fits of the response in a wide variety of situations, enabling fitting of non-linear relationships between explanatory variables and the dependant variable, with no assumption that explanatory variables are independent. CART is a rule-based method that generates a binary tree through recursive partitioning. This splits a subset (called a node) of the data set into two subsets (called sub-nodes) according to minimization of a heterogeneity criterion computed on the resulting sub-nodes. Random forests is a procedure that generates a large number of tree predictors and then selects the most popular class. Despite the analytical advances of all of these techniques, none perform exhaustive exploration of the data [4] and to date, there is no algorithm that can search for all possible stratifications and identify the best combination of variables to explain a specified outcome.

Complementary to these non-parametric methods and to traditional statistical methods, a new approach, HyperCube® (Institute of Health & Science, Paris, France) is based on the latest research in artificial intelligence, using least general generalized algorithms and genetic algorithms. The underlying idea is to describe a dataset by a group of « local over densities » of a specific outcome with no *a priori* hypothesis or notion of distance, each « over density » being completely independent from every other. The breakthrough is the ability to deal with points in a space with absolutely no assumptions, including those concerning metric and distance or nature of neighborhood. Indeed, working with a distance or a defined topology is already an assumption and either is not true or introduces bias into the model.

This method has been applied to various topics, mainly in the financial and business sectors, but remains unvalidated in the field of biology [7]. Through exhaustive exploration of m-dimensional space, HyperCube® will classify subsets of the study population into high and low risk groups and pinpoint not only the key explanatory variables and their interactions, but also the key range of values within each explanatory variable. Whilst this approach has evident value for risk factor analysis critical for clinical decision making, it also offers a tool with which to explore complexity, potentially revealing unimaginable combinations of explanatory variables underpinning the observed outcome.

We report here a rigorous assessment of the performance of this novel HyperCube® method. The aim of the study is to test whether the rules identified by HyperCube® give the best predictive value. We use HyperCube® to explore a large longitudinal epidemiological data set of malaria. We compare the predictive value of the rules identified by HyperCube® with models generated using classical statistical methods, binomial regression and CART. We demonstrate that HyperCube® can identify the best combination of factors predicting the outcome of malaria infection in our dataset.

Results

Populations, outcome and explanatory variables

We studied a large dataset from a long-term epidemiological study of two family-based cohorts in Senegal, followed for 19 years (1990–2008) in Dielmo and for 16 years (1993–2008) in Ndiop [8,9]. Time period of observation was classified as a trimester. The dependant variable was defined as a binary trait: individuals with at least one clinical *Plasmodium falciparum* malaria attack (PFA) during that trimester or without PFA. In total, there were 46,837 outcome events of person-trimesters from 1,653 individuals. Almost 20% of the events were PFA in both villages. Thirty-four explanatory variables for association with the occurrence of PFA were considered. Twenty one variables were qualitative (eight nominal and 13 ordered) and 13 were quantitative (Table 1 and 2).

HyperCube® analysis

We first analyzed the data using HyperCube®. We divided our dataset into 3 phases: Learning, Validation and Replication. We analyzed the two cohorts separately. A random variable was created dividing the data of each cohort into two groups of equal size (in and out samples). The learning phase was carried out using the "in sample" from the first studied cohort. In the validation phase, rules defined in the learning phase were validated in the "out sample" of the same cohort. The learning set contained 11,893 events and the validation set had 11,939 in Dielmo, while in Ndiop there were 11,530 events in the learning set and 11,475 in the validation set. The effect of each validated rule from the first cohort was studied in the second cohort in the replication phase.

We defined three parameters for running the learning process, "Lift", "Size" and "Complexity". "Lift" is the ratio of the prevalence of positive PFA events within a rule over the prevalence of positive PFA events in the entire population; this is equivalent to relative risk (RR). "Size" is the minimum number of events described by the rule. "Complexity" describes the maximum number of variables in a rule. Choice of "Lift" and "Size" parameters are optimized using the "Signal Intensity Graph" (see Material and Method). The "Complexity" parameter is here fixed to six factors, of which two are forced, the "in sample" and the cohort. Table 3 summarizes the parameters used and results obtained from the HyperCube® analyses.

After 27 and 23 hours of analyses, we obtained 4,853 and 6,860 rules in Dielmo and Ndiop, respectively. We calculated the probability for the occurrence of a rule with identical "Lift" and "Size" parameters from randomization of the entire dataset to obtain an empirical *P* value (emp*P*). We selected minimized rules (see materials and methods) with emp*P* less than 10^{-80} in Dielmo and Ndiop, for the validation phase (Table 3). We used this high threshold emp*P* for selection to minimize the risk of over-fitting. We were able to validate 51 of 52 minimized rules (98%) and 36 of 36 (100%) in Dielmo and Ndiop respectively. Of these, all 51 (100%) rules from Dielmo were replicated in Ndiop and all 36 (100%) rules from Ndiop were replicated in Dielmo with emp*P* less than 10^{-3}. We selected the best predicted rule for further statistical study (Figure 1). The best predictive rule contained 1,689 events from 148 individuals and was defined as: individuals who lived in Dielmo during 1992 to 2003, were of an age between 1 to 5 years old, having hemoglobin type AA, and having had previous *Plasmodium malariae* infection (PMI) less than or equal to 10 times. These individuals had 3.71 (95%CI: 3.58–3.84) times more PFA than the general population; and this sub-population was the most representative (i.e. containing the maximum number of events) among those with a RR of at least equal to 3.71.

Confirmation of the HyperCube® rule with traditional statistical methods

We sought to replicate the HyperCube® rule using logistic regression. We redefined continuous variables as binary variables according to the HyperCube® rule: The "Year" variable was defined as after 1991 and before 2004 or else; Age variable as between 1 and 5 years old or else; Hemoglobin type AA or else and cumulative number of previous PMIs as $\leq$10 times or else. By multivariate analysis, we tested all possible interactions between two variables and dropped interaction terms with $P>0.05$ until all had $P\leq0.05$. The variables showed highly significant marginal effect ($P<0.0001$) except age (Table 4). Age was highly significant ($P<10^{-4}$) when taking into account other criteria including year (between 1992 and 2003) and previous PMIs ($\leq$10). Analysis incorporating all possible interaction terms (i.e. with more than 2 variables) generated considerable over-dispersion and was difficult

Table 1. List of explanatory categorical variables.

Categorical (nominal) Variables	**No of levels**
House	67 (36 in Dielmo and 31 in Ndiop)
Independent Family	36 (12 in Dielmo and 24 in Ndiop)
Sex	2
Hemoglobin Type	7 (5 in Dielmo and 7 in Ndiop)
ABO blood group	4
G6PD Haplotype (on 4 SNPs: G6PD-376*, G6PD-202*, G6PD-968* and G6PD-542*)	11
PMI	2
POI	2
Categorical (ordered) Variables	**No of levels**
Drug treatment period	4
Year	19 (19 in Dielmo and 16 in Ndiop)
Trimester	4
ABO-261*: rs8176719	3
ABO-297*: rs8176720	3
ABO-467*: rs1053878	3
ABO-526*: rs7853989	3
ABO-771*: rs8176745	3
Alpha globin-3.7deletion	3
G6PD-202*: rs1050828	3
G6PD-376*: rs1050829	3
G6PD-542*: rs5030872	3
G6PD-968*: rs76723693	3

G6PD: Glucose-6-phosphate dehydrogenase, PMI: *Plasmodium malariae* infection, POI: *Plasmodium ovale* infection.
*: Position on the gene.
doi:10.1371/journal.pone.0024085.t001

to interpret. This result demonstrates that even though age is a major factor influencing development of PFA, without considering other variables, this effect would have been missed.

In order to replicate precisely the HyperCube® rule and determine the relative risk for comparison with other models/rules, we estimated the overall effect of the four key variables and all their possible interactions by defining a dummy variable X to represent the two sub groups of the population: X = 1 for a sub-population defined by the observations in the rule (i.e. living in Dielmo during 1992 to 2003, age 1 to 5 years old, having hemoglobin type "AA" and having had previous PMIs≤10); X = 0, otherwise (Table 5). Table 5 shows 1,232 PFA+457 not PFA in the rule = 1,689 events via HyperCube®. The Pearson chi-square test confirmed the strongly significant probability to develop PFA ($\chi2 = 2740.55$, DF = 1, $P<10^{-16}$), yielding a RR of 3.71 (95%CI: 3.58–3.84) and odds ratio (OR) of 11.02 (95%CI: 9.87–12.29). Using logistic regression, we confirmed the results of HyperCube®.

Replication of the rule in the 2nd cohort

In order to validate the biological and epidemiological aspect of this HyperCube® rule, it was replicated in Ndiop where a sub-population defined as above for Dielmo presented a higher risk to develop PFA compared to the general population: ($\chi2 = 665.96$, DF = 1, $P<10^{-16}$), RR of 2.35 (95%CI: 2.22–2.48) and OR of 3.50 (95%CI: 3.16–3.87). The result was optimal in Dielmo and replicated in Ndiop. The four variables identified above to be risk factors in Dielmo were thus also risk factors in Ndiop. Keeping the same settings as in Dielmo for time period (from 1992 to 2003), previous PMIs (≤10) and hemoglobin ("AA"), risk was maximum when age was re-set to 3 to 7 years old, with a RR of 2.53 (95%CI: 2.41–2.66) and OR of 4.04 (95%CI: 3.67–4.45) with more events (size = 1,761 events from 181 individuals) and more strongly significant ($\chi2 = 933.93$, DF = 1, $P<10^{-16}$) than when using the Dielmo age range of 1–5 years old (Size of 1,607 events from 158 individuals). This risk in Ndiop was, however, still lower than in Dielmo.

The two cohorts differ in one very pertinent manner: in Dielmo malaria transmission occurs all year round because of the presence of a small stream that enables mosquitoes to breed. In Ndiop, transmission is highly seasonal and occurs during the rainy season (July–December). Hence, we calculated the risk in Ndiop using only the period of year between July to December, a period when environmental factors are similar in the two villages. We obtained the same relative risk, RR = 3.78 (95%CI: 3.62–3.94), OR of 11.80 (95%CI: 10.11–13.77), with a highly significant Pearson chi-square test ($\chi2 = 1542.50$, DF = 1, $P<10^{-16}$). Furthermore, this risk was maximum when using age 3 to 7 years old (RR = 4.11, 95%CI: 3.97–4.27 and OR = 17.31, 95%CI: 14.68–20.41) with more events (Size = 932 events from 179 individuals *vs.* of Size of 863 from 157 when using age 1 to 5) and higher significance ($\chi2 = 2076.17$, DF = 1, $P<10^{-16}$).

Comparison with other models

We examined whether a classical statistical method could identify the same or better rules. We performed logistic regression

Table 2. List of explanatory continuous variables.

Continuous Variables	Mean	Median	Min	Max
Age	21.35 (23.14 in Dielmo and 19.46 in Ndiop)	15.90 (17.06 in Dielmo and 14.97 in Ndiop)	0	97.88 (97.88 in Dielmo and 83.25 in Ndiop)
Mean genetic relatedness (Pedigree-based)	0.012 (0.012 in Dielmo and 0.012 in Ndiop)	0.011 (0.012 in Dielmo and 0.008 in Ndiop)	0.001	0.041 (0.028 in Dielmo and 0.041 in Ndiop)
Mean genetic relatedness IBD*-based)	0.008 (0.008 in Dielmo and 0.007 in Ndiop)	0.007 (0.008 in Dielmo and 0.007 in Ndiop)	0.002	0.029 (0.025 in Dielmo and 0.029 in Ndiop)
No. of previous PMI	2.53 (4.10 in Dielmo and 0.82 in Ndiop)	1 (1 in Dielmo and 0 in Ndiop)	0	44 (44 in Dielmo and 9 in Ndiop)
Time since first PMI (year)	6.07 (6.67 in Dielmo and 5.03 in Ndiop)	5.25 (5.95 in Dielmo and4.32 in Ndiop)	0	18.51 (18.51 in Dielmo and 15.25 in Ndiop)
No. of previous POI	1.09 (1.33 in Dielmo and 0.83 in Ndiop)	0	0	11 (11 in Dielmo and 10 in Ndiop)
Time since first POI (year)	5.52 (6.20 in Dielmo and 4.72 in Ndiop)	4.88 (5.55 in Dielmo and 4.25 in Ndiop)	0	18.51 (18.51 in Dielmo and 15 in Ndiop)
Exposure (number of days present in the village) per trimester	80.76 (81.65 in Dielmo and 79.87 in Ndiop)	91 (91 in Dielmo and 90 in Ndiop)	1	92
Distance to animal enclosure (meters)	322 in Dielmo and 147 in Ndiop	271 in Dielmo and 139 in Ndiop	1 in Dielmo and 2 in Ndiop	765 in Dielmo and 393 in Ndiop
Distance to toilets (meters)	326 in Dielmo and 149 in Ndiop	280 in Dielmo and 143 in Ndiop	1 in Dielmo and 2 in Ndiop	774 in Dielmo and 401 in Ndiop
Distance to house's tree (meters)	344 in Dielmo and 152 in Ndiop	311 in Dielmo and 149 in Ndiop	1 in Dielmo and 1 in Ndiop	759 in Dielmo and 386 in Ndiop
Distance to wells (meters)	365 in Dielmo and 195 in Ndiop	453 in Dielmo and 174 in Ndiop	17 in Dielmo and 17 in Ndiop	719 in Dielmo and 483 in Ndiop
Distance to all (animals, toilets, house's tree, wells) together (meters)	329 in Dielmo and 150 in Ndiop	288 in Dielmo and 143 in Ndiop	1 in Dielmo and 1 in Ndiop	774 in Dielmo and 483 in Ndiop

*IBD: Identity-By-Descent.

doi:10.1371/journal.pone.0024085.t002

analysis and CART using the Dielmo data. We first tested the effect of each variable on PFA by univariate analysis. When two or more variables were correlated, the most explicative variable was chosen. Continuous explanatory variables were categorized to enable comparison with HyperCube®, by grouping the range of values having similar values for the dependant variable. Searching for the cut-off values for continuous variables was guided by Classification and Regression Trees (CART) methods [5]. CART identified cut-off values to categorize Age and Exposure variables, but did not find significant cut-off values for previous PMIs or any other continuous variable. Therefore, median was chosen as the cut-off value for each of these other variables. We then selected variables that showed ≤ 0.10 type I error for multivariate analysis (Table 6 and 7). As HyperCube® dichotomizes any variable, being in or out of the rule; we redefined each variable in a similar way. Categorical, ordinal and interval variables that had more than 2 levels were redefined by regrouping levels for which their partial effects were in the same direction. Trimester variable was redefined as semester (January–June and July–December) since the first two trimesters had decreasing effects and the last two had increasing effects on PFA when we adjusted on the other variables. Year variable was redefined in two levels (period 1: "year$\leq$2003" and period 2: "year$\geq$2004") according to the effect of each year. Age variable was classified into two levels (having between 0.4 and 8.1 years-old or else) according to CART analysis, ABO blood group in two levels (O or not O). Table 8 shows the result of univariate analysis after redefinition. For multivariate analysis we used the binary explanatory variables from Tables 6–8 and analyzed by logistic regression using several model selection methods: (1) selection based on an exhaustive screening of candidate models in each subset of explanatory variables, selecting the best one in terms of Information Criterion (lowest Akaike Information Criterion (AIC)); (2) forward selection and backward elimination. Model selection was computed using Package

Table 3. Parameters used and rules obtained from the HyperCube® analyses.

Cohort	Total number of events	Learning Set	Validation Set	*Purity*	*Lift*	*Size*	Time of run	*Coverage*	Number of Total *rules*	Number of minimized *rules*	Number of validated *rules*	Number of replicated *rules*
Dielmo	23,832	11,893	11,939	0.73	4.00	400	27 h	67%	4,853	52	51	51
Ndiop	23,005	11,530	11,475	0.74	3.49	400	23 h	72%	6,860	36	36	36

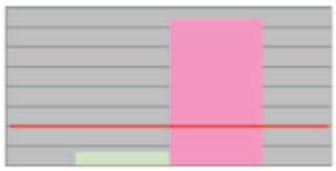
Purity: prevalence of events {PFA = 1} in the rule; Lift: Relative Risk of belonging to the rule compared to the total population; Size: number of events in the rule; Coverage: percentage of events {PFA = 1} in all rules found by HyperCube® compared to the total number of events {PFA = 1} in the whole dataset.

doi:10.1371/journal.pone.0024085.t003

A

Keys Indicators	
Purity: **0.73**	Size: **1689 samples**
Average number of PFA: **1.39**	Not *Plasmodium falciparum* malaria attack
Plasmodium falciparum malaria attack	

B



C

Rule space					
Patients that match all the following conditions:			**Loss**	**Coverage**	**Size**
Village	is	*Dielmo*	-0.67	50.9	+1607
Cumulative numbers of *Plasmodium malariae* infections	between	0 and 10	-0.06	93.4	+63
Year	between	1992 and 2003	-0.27	64.7	+508
Age	between	1 and 5	-2.53	14.2	+9135
Hemoglobin type	is	*AA*	-0.11	86.6	+268
Had 3.71 more often *Plasmodium falciparum* malaria attacks					

D

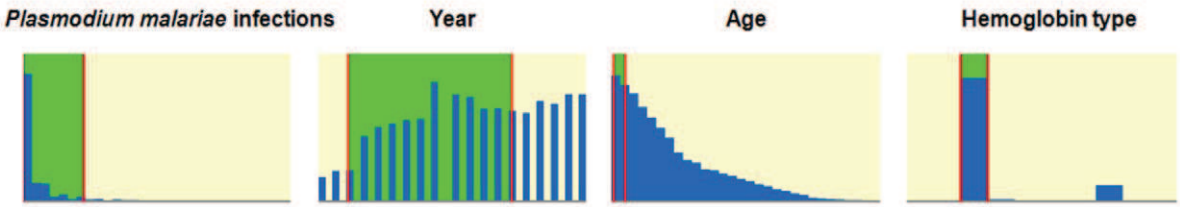


Figure 1. Typical result from HyperCube®. A) Table "*Key Indicators*" shows Lift: 1.39; Size: 1,689; Purity: 0.73. B) Graph showing comparative proportion of events within the rule and events in the entire population, pink: affected (PFA positive), green unaffected (PFA negative). Both pink and green bars would reach the horizontal red line if there was same proportion of positive PFA in the rule and in the entire population. C) Table "*Rule space*" shows marginal contribution of each variable to the lift. Loss: gives partial decreases of lift when removing each variable (or risk factor) from the rule; Coverage: percentage of events {PFA = 1} defined by the corresponding variable alone compared to the total number of events {PFA = 1} in the whole dataset; Size: increase of events in a rule when the constraint defined within a variable is cancelled or by dropping the variable. D) Graphs showing distribution (in blue) of each variable, and the range of values (in green) within the rule.
doi:10.1371/journal.pone.0024085.g001

"glmulti" of R software [10]. The results obtained are presented in Table 9.

According to the results of the multivariate regression model selection (Table 9), we defined for each selected model a sub-group X = 1 when all risk factors are present, otherwise X = 0. For each model, we gave RR, p-value, and number of events for the sub-group having all identified risk factors. All sub-groups identified using model selection techniques had lower predictive values for developing PFA than the HyperCube® rule (Table 9). For sub-groups explaining the same or a greater number of events than the one found by HyperCube®, the RR was lower and the 95% confidential intervals of RR did not overlap with those for the HyperCube® rule (Table 9).

Table 4. Multivariate analysis of risk factors associated with clinical *P. falciparum* malaria attacks in Dielmo using the HyperCube® rule.

Parameters			**DF**	**Estimate**	**SE**	**χ2**	**Pr>χ2**	**OR**	**Wald 95%CL**	
Intercept			1	−3.43	0.16	483.4	<.0001	-	-	-
Age group (years)	1 to 5		1	0.38	0.28	1.8	0.178	1.46	[0.84	2.53]
Type of hemoglobin	AA		1	0.38	0.07	27.8	<.0001	1.46	[1.27	1.68]
Year	After 1991 and Before 2004		1	1.80	0.15	139.4	<.0001	6.07	[4.50	8.19]
Number of previous *P. malariae* infections	≤10		1	0.80	0.15	29.4	<.0001	2.23	[1.67	2.97]
Age group **P. malariae* infections	1 to 5	≤10	1	1.62	0.27	36.5	<.0001	5.06	[2.99	8.56]
Age group* Year	1 to 5	Before 2004	1	0.77	0.10	55.8	<.0001	2.15	[1.76	2.63]
P. malariae infections*Year	≤10	Before 2004	1	−1.38	0.16	72.2	<.0001	0.25	[0.18	0.35]

DF: degree of freedom; Estimate: effect of explanatory variable's levels on *logit*(Probability of {PFA = 1}); SE: standard error; χ2: chi-square DF = 1; OR: Odds ratio; CL: confidential level.
doi:10.1371/journal.pone.0024085.t004

Table 5. Number of positive/negative PFA events (*P. falciparum* malaria attacks) in subgroups of individuals in and out of the HyperCube® rule.

	PFA positive	No PFA
In the *rule*	1232	457
Out of the *rule*	7977	37171
Total population	9209	37628

doi:10.1371/journal.pone.0024085.t005

We tested whether the HyperCube® rule predicted the highest risk of developing PFA. We used the HyperCube® model as a reference. We modified the reference HyperCube® rule by either removing one of the variables or adding in variables identified by multivariate analysis. Using the same method to define subsets of the population and construct contingency tables, we calculated RR, OR and *P* values for each model. As shown in Table 10, there was no other model that gave higher RR and/or OR than the one identified by HyperCube® with equal or greater size.

In contrast to the regression analyses, CART found that age (between 0.22 and 5.48) and year (from 1990 to 2003) defined the high risk group for having PFA (RR = 3.26 [95CI: 3.16–3.38], OR = 7.34 [95CI: 6.80–7.93] and size = 3,041 with $\chi 2 = 3268.85$, DF = 1 [$P<10^{-16}$]) (Figure 2). No other variable or combination of variables yielded a higher Relative Risk.

Optimality of HyperCube® choice

We then tested whether the cut-off values delimiting the range of values in the HyperCube® rule (defined as the reference rule) for each variable were the optimal ones. Hemoglobin type was fixed as AA or not. We modified the range of continuous variables of the reference rule. As the cut-off values for continuous variables were considered at integer values, there were a finite number of subsets that we could try for modifying a rule. We tested all possible ranges of the continuous variables (Age, previous PMIs and Year) with constraint of minimum "Size" of ≥400 events in the rules. We first fixed 2 variables and changed one variable at a time. The variable to change was first defined as the range of integer values

Table 6. Univariate logistic regression analysis of each categorical risk factor for clinical *falciparum* malaria (PFA) attacks in Dielmo.

		No of Person-trimesters						
		N = 23832						
		PFA = 0	PFA = 1	Estimate (Std. Error)	Crude OR	Wald 95%CL	*P*-values	Global *P*
		N(%) = 19475	N (%) = 4357					
Age group (years)	[0–0.4]	303 (84.17)	57 (15.83)	Ref.	1			
	[0.4–6.7]	2344 (46.72)	2673 (53.28)	1.80 (0.15)	6.06	[4.54–8.09]	<.0001	
	[6.7–8.12]	692 (67.13)	338 (32.82)	0.95 (0.16)	2.6	[1.9–3.55]	<.0001	<.0001
	[8.12–13.6]	2943 (81.28)	678 (18.72)	0.20 (0.15)	1.22	[0.91–1.65]	0.1782	
	≥13.6	13138 (95.58)	608 4.42)	−1.40 (0.15)	0.25	[0.18–0.33]	<.0001	
	Missing data	55	3	-	-	-	-	
Sex	Male	9663 (80.77)	2301 (19.23)	Ref.	1			
	Female	9812 (82.68)	2056 (17.32)	−0.13 (0.03)	0.88	[0.82–0.94]	-	<.0001
Blood group	O	7597 (79.56)	1952 (20.44)	Ref.	1			
	A	5131 (83.65)	1003 (16.35)	−0.27 (0.04)	0.76	[0.70–0.83]	<.0001	
	AB	920 (90.20)	100 (9.80)	−0.86 (0.11)	0.42	[0.34–0.52]	<.0001	<.0001
	B	4496 (82.40)	960 (17.60)	−0.19 (0.04)	0.83	[0.76–0.91]	<.0001	
	Missing data	1331	342	-	-	-	-	
Type of hemoglobin	AA	16304 (81.28)	3756 (18.72)	Ref.	1			
	AC/AS/SS	2007 (87.53)	286 (12.47)	−0.48 (0.07)	0.62	[0.54–0.70]		<.0001
	Missing data	5196	1438	-	-	-	-	
G6PD	Normal alleles	6448 (84.0)	1228 (16.0)	Ref.	1			
	Mutated allele	7865 (82.30)	1691 (17.70)	−0.12 (0.04)	0.89	[0.82–0.96]		0.0032
	Missing data	5162	1438	-	-	-	-	
P. malariae infections	≤1 (median)	9348 (81.99)	2099 (18.34)	Ref.	1			
	>1	8983 (79.91)	2258 (20.09)	0.11 (0.03)	1.12	[1.04–1.20]	-	0.0008
	missing	1144	0	-	-	-		
P. ovale infections	≤0 (median)	9946 (81.54)	2251 (18.46)	Ref.	1			
	>0	8385 (79.93)	2106 (20.07)	0.10 (0.03)	1.11	[1.04–1.19]	-	0.002
	missing	1144	0	-	-	-		

Estimate: effect of explanatory variable's levels on *logit*(Probability of {PFA = 1}); SE: standard error; OR: Odds ratio; CL: confidential level; Ref.: reference level.
Age and Exposure were categorized using CART and previous PMIs and previous POIs using median since CART did not find significant cut-off values.
doi:10.1371/journal.pone.0024085.t006

Table 7. Univariate logistic regression analysis of each temporal risk factor for clinical *falciparum* malaria (PFA) attacks in Dielmo.

		No of Person-trimesters						
		N = 23832						
		PFA = 0	PFA = 1	Estimate (Std. Error)	Crude OR	Wald 95%CL	*P*-values	Global *P*
		N(%) = 19475	N (%) = 4357					
Year	1990	587 (82.21)	127 (17.79)	Ref.	1			
	1991	740 (81.59)	167 (18.41)	0.04 (0.13)	1.04	[0.81–1.35]	0.7457	
	1992	717 (77.18)	212 (22.82)	0.31 (0.13)	1.37	[1.07–1.75]	0.0126	
	1993	790 (78.61)	215 (21.39)	0.23 (0.12)	1.26	[0.99–1.61]	0.0653	
	1994	774 (75.44)	252 (24.56)	0.41 (0.12)	1.50	[1.19–1.91]	0.0008	
	1995	796 (77.06)	237 (22.94)	0.32 (0.12)	1.38	[1.08–1.75]	0.0093	
	1996	853 (72.23)	328 (27.77)	0.58 (0.12)	1.78	[1.41–2.24]	<.0001	
	1997	818 (73.3)	298 (26.7)	0.52 (0.12)	1.68	[1.33–2.13]	<.0001	
	1998	1179 (80.2)	291 (19.8)	0.13 (0.12)	1.14	[0.91–1.44]	0.2632	
	1999	1137 (78.09)	319 (21.91)	0.26 (0.12)	1.30	[1.03–1.63]	0.0258	<.0001
	2000	1151 (76.84)	347 (23.16)	0.33 (0.12)	1.39	[1.11–1.75]	0.0041	
	2001	1019 (77.91)	289 (22.09)	0.27 (0.12)	1.31	[1.04–1.65]	0.0222	
	2002	1061 (80.75)	253 (19.25)	0.1 (0.12)	1.10	[0.87–1.40]	0.4188	
	2003	1055 (80.47)	256 (19.53)	0.11 (0.12)	1.12	[0.89–1.42]	0.3396	
	2004	1153 (87.81)	160 (12.19)	−0.44 (0.13)	0.64	[0.50–0.83]	0.0006	
	2005	1312 (91.11)	128 (8.89)	−0.8 (0.13)	0.45	[0.35–0.59]	<.0001	
	2006	1228 (83.2)	248 (16.8)	−0.07 (0.12)	0.93	[0.74–1.18]	0.5663	
	2007	1495 (90.44)	158 (9.56)	−0.72 (0.13)	0.49	[0.38–0.63]	<.0001	
	2008	1610 (95.72)	72 (4.28)	−1.58 (0.16)	0.21	[0.15–0.28]	<.0001	
Season	Jan–Mar	4749 (82.62)	999 (17.38)	Ref.	1			
	April–June	4912 (82.03)	1076 (17.97)	0.04 (0.05)	1.04	[0.95–1.14]	0.4029	
	July–Sept	4841 (80.38)	1182 (19.62)	0.15 (0.05)	1.16	[1.06–1.27]	0.0017	0.0128
	Oct–Dec	4973 (81.89)	1100 (18.11)	0.05 (0.05)	1.05	[0.96–1.16]	0.2973	
Exposure	≤66.5 days	2978 (94.33)	179 (5.67)	Ref.	1			
	>66.5 days	15745 (81.57)	3558 (18.43)	1.32 (0.08)	3.76	[3.22–4.39]	-	<.0001
		752	620	-	-	-		

Estimate: effect of explanatory variable's levels on *logit*(Probability of {PFA = 1}); SE: standard error; OR: Odds ratio; CL: confidential level; Ref.: reference level. Age and Exposure were categorized using CART and previous PMIs and previous POIs using median since CART did not find significant cut-off values.

doi:10.1371/journal.pone.0024085.t007

between its minimum and maximum values, and then reduced from the maximum to smaller integer values covering an ever-decreasing total age range until the minimum. This was repeated step by step until each integer value of the variable was set as the minimum for a step. Therefore, the total number of choices for a variable is 1+2+3+…+*maximum* = sum of a finite arithmetic sequence = (first value+last value)*(number of values)*(1/2). Each choice corresponds to a specific modification of the reference rule (i.e. a specific interval of values defining the modified rule). Then, for Age, previous PMIs and Year, there are (1+98)*98*0.5 = 4851, (1+45)*45*0.5 = 1035 and (1+19)*19*0.5 = 190 possible choices respectively. We then fixed 1 variable and changed 2 variables simultaneously. When Year is fixed and the couple (Age, previous PMIs) changed simultaneously, there are 4851*1035 = 5,020,785 possible choices. For previous PMIs fixed and (Age, Year) changed and Age fixed and (previous PMIs, Year) changed there are 4851*190 = 921,690 and 1035*190 = 196,650 possible choices. When we selected choices with at least same size as the reference rule, the resulting RR was always lower than the reference RR. Figure 3 shows the effects of the modified ranges (i.e. the effect of other choices different from the one found by HyperCube®) on RR. If all 3 variables were allowed to vary simultaneously there would be 4,851(Age) *190(Year) *1035(previous PMIs) = 953,949,150 possible choices. The time for running such an analysis on one computer with 2 central processor units (Duo CPU 2.00 GHz 2.00 GHz), Memory (RAM) of 3.00 GB) is estimated at ~5678 days (~1.94 choices analyzed per second) using function "*system.time(.)*" of R-software, and thus not possible to analyze.

Discussion

We describe here a new data mining algorithm that can identify the combinations of variables that give the optimal prediction of the outcome of interest. We demonstrate that the model identified by HyperCube® has better predictive value than any other model tested. HyperCube® was able to identify the best cut-off value and range for continuous variables. It classified the population into high and low risk groups and made the results easier to interpret in

Table 8. Univariate analysis of each risk factor (redefined in only two levels) for clinical *P. falciparum* malaria attacks (PFA) in Dielmo.

		No of Person-trimesters					
		N = 23832					
		PFA = 0	PFA = 1	Estimate (Std. Error)	Crude OR	Wald 95%CL	*P*-values
		N (%) = 19475	N (%) = 4357				
		(81.72)	(18.28)				
Age group (years)	<0.4 or ≥8.12	16384 (92.42)	1343 (7.58)	Ref.	1		
	[0.4–8.12]	3036 (50.21)	3011 (49.79)	2.49 (0.04)	12.1	[11.22–13.04]	<.0001
	Missing data	55	3	-	-	-	
Blood group	A or B or AB	10547 (83.64)	2063 (16.36)	Ref.	1		
	O	7597 (79.56)	1952 (20.44)	0.27 (0.04)	1.31	[1.23–1.41]	<.0001
	Missing data	1331	342	-	-	-	
Year	≥2004	6798 (89.87)	766 (10.13)	Ref.	1		
	<2004	12677 (77.93)	3591 (22.07)	0.92 (0.04)	2.51	[2.31–2.73]	<.0001
Semester	Jan–Jun	9661 (82.32)	2075 (17.68)	Ref.	1		
	Jul–Dec	9814 (81.13)	2282 (18.87)	0.08 (0.03)	1.08	[1.16–1.16]	0.0179

Estimate: effect of explanatory variable's levels on *logit*(Probability of {PFA = 1}); SE: standard error; OR: Odds ratio; CL: confidential level; Ref.: reference level.
doi:10.1371/journal.pone.0024085.t008

terms of biology than the probability estimates generated by most statistical methods.

The principle of this method is to explore all possible combinations of predictor variables and to find, through stochastic parallel computing exploration, the optimal hypercubes (or sub-spaces) defined by a combination of these variables, without making any assumptions. This method allows generation of rules, sets of variables and ranges of variable values that define subpopulations with high risk for the outcome of interest and that best predict the outcome. Inspired from latest research in artificial intelligence, Least General Generalized algorithms and Genetic Algorithms, HyperCube® SaaS software generates local hypercubes and stabilizes each local hypercube to a local optimum, each optimum being new and independent. By doing so, it is possible to describe and understand local configurations without there being necessarily any global effect, i.e. some specific combination of factors that are only found in a sub-set of the population may increase the risk of outcome for that sub-population, but which are not detectable when averaged across the entire population. HyperCube® enables us to describe the range of values and the combination of variables that can trigger the events. Although the statistics aims to reject, or not, a predefined assumption according to given risks, these complex event intelligence techniques allow us to generate assumptions on rules without any prerequisite. A hypercube is expressed in a simple formal way as a rule, directly readable and comprehensible.

As correction for multiple testing is not possible when using HyperCube®, statistical validation and replication in independent cohorts are crucial, even prior to biological validation. We randomly divided the population in one cohort into the learning set and the validation set. We used the other cohort for replication. In addition, we calculated an empirical *P* value from whole randomized data. We demonstrated that using a high threshold of empirical *P* value (10^{-80}), 98–100% of the rules could be validated and 100% of validated rules could be replicated in another cohort despite their differences in human ethnicity and malaria endemicity [11].

Biological validation of the rule is most important. Here three of the variables are known *a priori* to increase the risk of developing PFA: young children (i.e. lack of clinical immunity), normal hemoglobin Hb AA, and living during a period of intense malaria transmission. However, HyperCube® allowed us to identify the range of continuous variables, such as age and year, which enable us to define high and low risk groups. In addition, the effect of these three variables alone did not reach our stringent acceptance threshold. Identifying an additional variable using classical techniques would be a big challenge due to the number of possible choices. HyperCube® added a fourth one "number of previous PMIs at ranges less than or equal to 10" to define a rule containing 1,232 events with PFA and 457 events without PFA (prevalence = 72.9%) compared to 19.7% prevalence of the whole population (RR 3.71 (95%CI: 3.58–3.84). This RR is the highest of all models containing this number of events. This rule explained 28.28% of total events with PFA in the dataset.

The effect size of each variable was estimated by removing each variable and calculating the loss in "Lift" (Figure 1c). The strongest effect is age (68%), then village (18%), followed by year (7.3%). Hemoglobin type explained 3% of the "Lift" while previous PMIs had only 1.6% effect. There was 1.8% of the "Lift" that could not be explained by each of these variables individually (Table 11) and thus reflects interaction among the variables. In Dielmo, malaria transmission is holoendemic with an average of more than 200 infectious bites per person per year, 10 times more than Ndiop [12]. Therefore, individuals living in Dielmo have more chance to develop PFA. Age is a well known factor of PFA due to rapid development of clinical immunity in high malaria transmission regions. Using variance component analysis, age explained 29.8% of total variation in number of PFA in Dielmo [11]. The year effect is almost certainly yearly variation in transmission intensity. Indeed in 2003, the HyperCube® rule threshold for year, a new drug for PFA treatment was introduced and malaria transmission decreased in following years. Hemoglobin type is one of the best known genetic factors protecting against malaria. In our and other studies, sickle cell mutation explained 2–5% of risk in development of severe and clinical *falciparum* malaria [13], similar to that estimated by HyperCube® (Table 11). The new variable that HyperCube® identified is previous *P. malariae*

Table 9. Multivariate model selection for risk factors associated with clinical *P. falciparum* malaria attacks (PFA) in Dielmo using factors identified from univariate logistic analysis.

		Best model (with lowest AIC) when number of explanatory variables is equal to:											
Variables		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Forward***	**Backward***
Sex	Female										√	NSE	1
Age group (years)	0.4 to 8.1	√	√	√	√	√	√	√	√	√	√	1	NSR
Blood Type	O									√	√	9	NSR
Type of hemoglobin	AA							√	√	√	√	7	NSR
G6PD	Normal					√	√	√	√	√	√	6	NSR
Year	Before 2004		√	√	√	√	√	√	√	√	√	2	NSR
Semester	Jul–Dec								√	√	√	8	NSR
Exposure	>66.5				√	√	√	√	√	√	√	4	NSR
P.malariae infections	≤1			√	√	√	√	√	√	√	√	3	NSR
P.ovale infections	≤0						√	√	√	√	√	5	NSR
RR		2.53	2.96	3.22	3.22	3.15	2.98	3.08	3.27	3.32	2.95	3.32	3.32
(95% CI)		(2.45–2.61)	(2.86–3.05)	(3.10–3.35)	(3.09–3.37)	(2.93–3.38)	(2.71–3.27)	(2.79–3.40)	(2.89–3.70)	(2.83–3.91)	(2.18–3.99)	(2.83–3.91)	(2.83–3.91)
p-value		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Size of subset defined by all risk factors		6044	4277	2000	1520	507	316	261	143	78	31	78	78

√ : For selected variables.
NSE: No (additional) effects met the 0.05 significance level for entry into the model.
NSR: No (additional) effects met the 0.05 significance level for removal from the model.
*: Both Forward and Backward methods selected the best (in terms of AIC) model with 9 explanatory variables.
doi:10.1371/journal.pone.0024085.t009

Table 10. Predictive values of modified HyperCube® rule.

Variable	Size	RR	95%CL		OR	95%CL		χ^2	DF	Pr>χ^2
M.ref:		3.71	3.58	3.84	11.02	9.87	12.29	2741	1	<.0001
P. malariae infections+Year+Age+Hemoglobin	**1689**									
M.ref−*P. malariae* infections	**1752**	3.65	3.52	3.77	10.35	9.30	11.51	2705	1	<.0001
M.ref−Year	**2197**	3.44	3.33	3.56	8.58	7.82	9.40	2843	1	<.0001
M.ref−Age	**10824**	1.18	1.14	1.23	1.24	1.18	1.30	71	1	<.0001
M.ref−Hemoglobin	**1957**	3.60	3.48	3.73	9.94	9.00	10.99	2898	1	<.0001
M.ref+Sex−*P. malariae* infections	879	3.69	3.53	3.86	10.82	9.31	12.57	1475	1	<.0001
M.ref+Sex−Year	1031	3.59	3.44	3.75	9.82	8.57	11.25	1592	1	<.0001
M.ref+Sex−Age	**5377**	1.16	1.10	1.22	1.20	1.13	1.29	29	1	<.0001
M.ref+Sex−Hemoglobin	990	3.62	3.46	3.78	10.06	8.75	11.56	1562	1	<.0001
M.ref+Blood Type−*P. malariae* infections	784	3.61	3.44	3.79	10.03	8.58	11.72	1249	1	<.0001
M.ref+Blood Type−Year	966	3.46	3.30	3.63	8.69	7.57	9.96	1351	1	<.0001
M.ref+Blood Type−Age	**4312**	1.29	1.22	1.36	1.38	1.29	1.49	78	1	<.0001
M.ref+Blood Type−Hemoglobin	852	3.66	3.50	3.83	10.48	9.01	12.19	1399	1	<.0001
M.ref+G6PD−*P. malariae* infections	651	3.76	3.58	3.95	11.56	9.69	13.79	1162	1	<.0001
M.ref+G6PD−Year	717	3.72	3.55	3.91	11.17	9.46	13.20	1244	1	<.0001
M.ref+G6PD−Age	**4840**	1.17	1.11	1.23	1.22	1.13	1.31	30	1	<.0001
M.ref+G6PD−Hemoglobin	661	3.84	3.66	4.02	12.59	10.53	15.05	1249	1	<.0001
M.ref+Semester−*P. malariae* infections	884	3.77	3.62	3.94	11.76	10.09	13.69	1574	1	<.0001
M.ref+Semester−Year	1117	3.56	3.41	3.72	9.54	8.38	10.86	1677	1	<.0001
M.ref+Semester−Age	**5458**	1.23	1.17	1.30	1.31	1.23	1.40	64	1	<.0001
M.ref+Semester−Hemoglobin	988	3.76	3.61	3.92	11.62	10.06	13.42	1734	1	<.0001
M.ref+Exposure−*P. malariae* infections	1403	3.66	3.25	3.80	10.46	9.29	11.78	2228	1	<.0001
M.ref+Exposure−Year	**1804**	3.44	3.31	3.57	8.51	7.69	9.42	2367	1	<.0001
M.ref+Exposure−Age	**8729**	1.15	1.11	1.20	1.20	1.14	1.27	42	1	<.0001
M.ref+Exposure−Hemoglobin	1535	3.62	3.49	3.76	10.14	9.05	11.35	2361	1	<.0001
M.ref+*P. ovale* infections−*P. malariae* infections	729	3.88	3.71	4.06	13.13	11.06	15.60	1410	1	<.0001
M.ref+*P. ovale* infections−Year	759	3.87	3.71	4.05	13.05	11.02	15.44	1459	1	<.0001
M.ref+*P. ovale* infections−Age	**4256**	1.52	1.44	1.59	1.73	1.62	1.86	246	1	<.0001
M.ref+*P. ovale* infections−Hemoglobin	768	3.85	3.69	4.03	12.79	10.82	15.10	1456	1	<.0001

M.ref: reference model; Size: number of events; RR: risk ratio; OR: Odds ratio; χ2: chi-square DF = 1; CL: confidential level.
doi:10.1371/journal.pone.0024085.t010

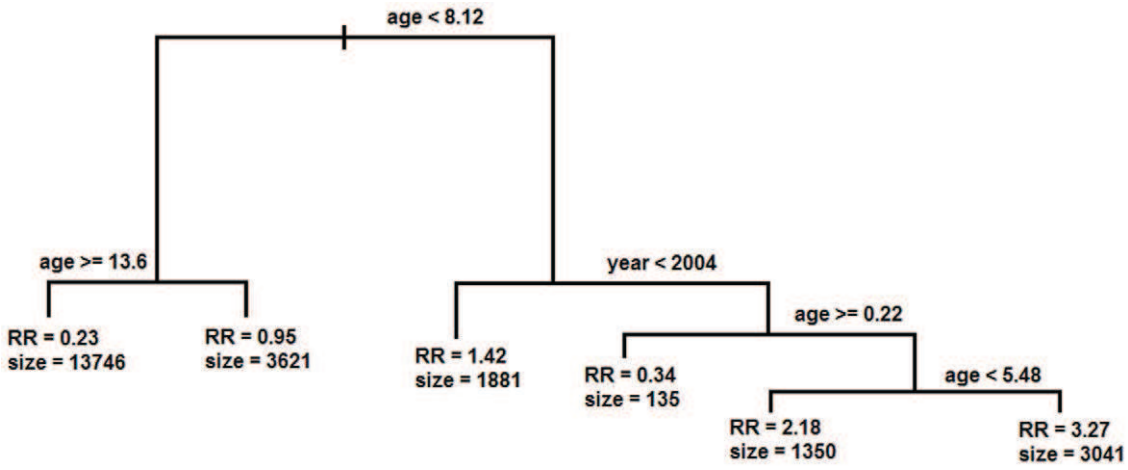


Figure 2. Decision tree generated by Classification and Regression Tree (CART) analysis of risk factors determining the occurrence of *P. falciparum* malaria attacks (PFA) per trimester. Figure shows the cut-off values identified by CART that divide the dataset into two. At each leaf are given the Relative Risk (RR) and the number of events associated with that leaf.
doi:10.1371/journal.pone.0024085.g002

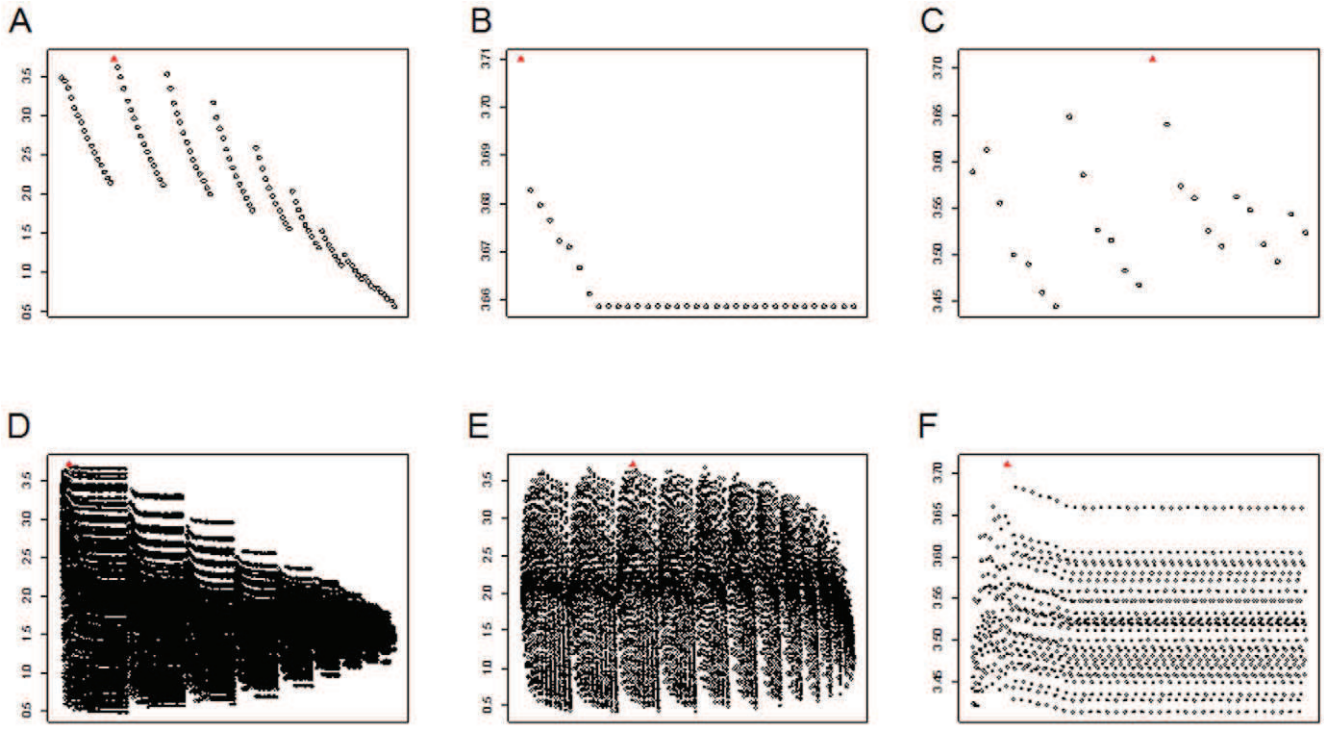


Figure 3. Effect on relative risk (RR) of modifying the ranges of continuous variables. Graphs show RR for all other possible definitions of risk group on the explanatory variables, with equal or greater size than the HyperCube® rule. Y-axis indicates the RR. A) Only ranges of Age are modified: 102 choices among 4,851 possible choices had size equal or greater than 1,689 (size of the HyperCube® rule) and are plotted; B) Only ranges of previous PMIs are modified: 35 choices among 1,035 possible; C) Only ranges of Year are modified: 25 choices among 190 possible; D) Ranges of both Age and previous PMIs are modified simultaneously: 25,040 choices among 5,020,785 possible; E) Ranges of both Age and Year are modified simultaneously: 8,912 choices among 921,690 possible; F) Ranges of both previous PMIs and Year are modified simultaneously: 1,110 choices among 196,650 possible. Filled red triangle represents the RR of HyperCube®'s rule (HyperCube®'s risk group), empty black circles represent the RR of other choices of risk groups.
doi:10.1371/journal.pone.0024085.g003

infection - PMI. Although CART did not identify any significant threshold for previous PMI, using the median as the cut-off value gave a significant effect for previous PMI is the univariate logistic regression, whereby above median previous PMI increased risk of PFA (P = 0.0008, Table 6). Interestingly in the HyperCube® rule the reverse was found and this is because of the interaction of previous PMI with age: being young and having previous PMI decreased risk.

Table 11. Effect size of each variable in the rule.

	DIELMO		NDIOP			
			All year		July	December
	Loss	% Loss	Loss	% Loss	Loss	% Loss
Initial Lift	3.71	100%	2.35	100%	3.78	100%
Age	−2.53	−68.2%	−0.82	−34.9%	−1.26	−33.3%
Village	−0.67	−18.1%	−0.7	−29.8%	0.05	1.3%
Year	−0.27	−7.3%	−0.07	−3.0%	−0.06	−1.6%
Hb	−0.11	−3.0%	−7.0%	−3.0%	−0.09	−2.4%
Previous PMIs	−0.06	−1.6%	−0.13	−5.5%	−0.12	−3.2%
Semester	-	-	-	-	−1.43	−37.8%
Total Loss	−3.64	−98%	−1.79	−76%	−2.91	−77%
Residual Lift	0.07	1.9%	0.56	23.8%	0.87	23.0%

Loss: partial decreases of lift when removing each variable from the rule.
doi:10.1371/journal.pone.0024085.t011

Cross-species immunity among different *Plasmodium* species has long been suspected and there is evidence of among-species negative interactions during concomitant infection [14,15]. An influence of *P. malariae* carriage on subsequent *P. falciparum* infection has been observed before. In Gabon, children infected with *P. malariae* presented more often with a *P. falciparum* infection and at higher parasite densities [16]. During the follow-up, subjects who were infected by *P. malariae* were reinfected by *P. falciparum* more rapidly. Such a relationship was also observed in the Garki project [15,17,18]. Although small scale variation in mosquito biting rate could generate similar levels of exposure to each parasite spp., the species infection association was found to be related to differences in acquired immunity and not to differences in exposure, suggesting that the levels of immunity to *P. falciparum* and to *P. malariae* were inter-related [18]. More recently, a family-based study found a strong relationship between *P. falciparum* parasite density and frequency of *P. malariae* infections [19]. *P. falciparum* parasite density has previously been shown to be under human genetic control and linked to the chromosomal region 5q31 in four independent studies [11,20,21,22]. These results suggest that individuals genetically susceptible to *P. falciparum* are also genetically pre-disposed to *P. malariae* [19]. Little is known on the impact of infection by one species on the incidence of disease of another. The relationships between parasite density and risk of attributable disease were found to be similar for *P. falciparum*, *P. vivax* and *P. malariae* in Papua New Guinea, compatible with the

hypothesis that pan-specific mechanisms may regulate tolerance to different *Plasmodium* spp. [23]. Pertinent to our finding here, Black *et al.* found that children with symptomatic episodes not only presented with fewer mixed species infections, but also had fewer previous *P. malariae* infections than symptom-free children, as demonstrated by serology [24]. The induced infection experiments also provide evidence of the development of some cross-protective immunity [25]. Interestingly, previous infection with *P. malariae* has been previously shown to impact upon a *P. falciparum* infection, but with respect to the production of transmission stages and not clinical presentation [26,27].

Many other rules used this variable confirming that previous infection by *P. malariae* is associated with protection against development of PFA. It is presently impossible to conclude if this association is a causal one or is due to a correlation to an unknown factor affecting the risk to develop PFA. As both parasites are transmitted by the same mosquito species, increased exposure to one species (*P. malariae*) might be expected to correlate with increased exposure to the other (*P. falciparum*). Hence, spatial heterogeneity in the exposure to infection could simultaneously result in increase risk of infection by both parasite *spp*. Our analysis did not take into account "number of previous *P. falciparum* attacks" (nbpPFA) and so it is possible that the variable previous PMIs replaces this information. However, in another HyperCube® analysis, we found that both previous PMIs and nbpPFA are used in different rules (data not shown), indicating that the previous infection by the two parasite species is not perfectly correlated. Thus, it seems probable that the parasite species effect reflects some impact of *P. malariae* infection on the development of immunity against *P. falciparum*. In our study, there were from 0 to 44 *P. malariae* infections per person prior to a clinical *P. falciparum* episode. Hypercube® identified that having few *P. malariae* infections (less than 10) was a potent risk factor, which excluded about 10% of events from those individuals who were often infected with *P. malariae*. The fact that a threshold of ten infections was identified as eliminating this risk factor is clearly not an exact threshold, but generally reflects the weakly immunising effect of *P. malariae* infection, reminiscent of that induced by *P. falciparum* infection. Furthermore, whereas eighteen out of 51 rules used the number of previous *P. malariae* infections, none used the number of previous *P. ovale* infections, illustrating that infection by the two *Plasmodium* species differently affects susceptibility to *P. falciparum* attacks. However, it should be noted that the absence of an effect of *P. ovale* on clinical *P. falciparum* attacks does not mean that *P. ovale* definitively has no effect. It may be the case that additional variables may be required to be taken into account. Indeed, in the multivariate model selection analysis (Table 9), previous P. ovale infection is significantly as a risk factor when a minimum of 6 explanatory variables are used. In our HyperCube® analyses, we limited the number of variables in a single rule to four. This differential species effect is currently under investigation.

We compared the rule with the model identified by classical logistic regression method. Although we aimed to include all possible interaction terms among variables studied in multivariate analysis, over-dispersion of the data made this unstable. In addition, the running time would have been unacceptably long, taking ~5678 days for one a common computer to analyze about 10^9 models (3 variables with around 10^3 cases for each). With HyperCube®, it took 23 to 27 hours to analyze 35 variables. In addition, the results of testing interaction among more than 2 variables by classical methods are difficult to interpret. We demonstrated that by omitting or adding other variables identified by other statistical methods or varying the cut-off value of continuous variables, the rule still performed best. Although some rules had higher RR, they have lower "Size" or more complexity and less significant *P* value. Among rules with "Size" equal to or greater than 1,689, the same as the reference rule, the reference rule gave the highest RR.

Interestingly, the rule identified by classical method covered 0.67% of total positive events whereas one HyperCube® rule explained 13.4%. When considering the minimized rules, we could identify risk factors that could explain 67% of total positive events, a percentage of coverage that would never be achieved by classical methods. While the classical method looked at events in 2 dimensions, HyperCube® identified rules in multi-dimensional space. Although all factors identified by the classical method are risk factors for development of PFA, different groups of people developed PFA for different reasons. The rule identified by the classical method involved only individuals who had all the risk factors. We could only separate groups of individuals with different risk factors when looking at the events in multi-dimensional space.

Analysis by CART identified a combination of variables, Age and Year, that increased risk of PFA. Both of these variables and the range of these variables were very similar to those identified by HyperCube®. That CART failed to detect Hemoglobin or previous PMIs likely reflects the differences in methodologies of the two techniques. CART uses a sequential approach first splitting the data set according to the most significant variable and identifying the threshold value of that variable that maximizes the discrimination in the two subsets of data (i.e. least PFA *vs.* most PFA). Then, CART will further sub-divide each subset by the next most significant variable that leads to maximum discrimination. This approach thus leads to canalization of the data along different pathways, resulting in a decreased sample size for comparison. In addition, optimization by maximum discrimination at each level may paradoxically lead to an erroneous sub-optimal end-point many levels down. HyperCube®, by contrast, analyses all variables simultaneously with no sequential selection that leads to such loss of power or canalization along a potentially eventual sub-optimal pathway.

One limitation arises when studying qualitative variables with more than two levels. It is not possible for HyperCube® to combine levels having a similar effect in the same rule. One alternative would be to use analysis of variance, as we previously did in our classical analysis for qualitative variables with more than 2 levels, to detect modalities having a similar effect on the dependant variable and group them *a priori*.

Another more practical problem comes from the efficiency of the learning process. This process is more efficient in explaining the minor outcome, which is sometimes not the standard way of thinking. For instance, we could identify only factors increasing the risk of PFA, but not those conferring protection against malaria, which is the classical choice in malaria field. The positive events for PFA made up ~15% of the total number of events. To identify factors conferring protection (negative PFA), of which the prevalence was 85%, would have presented a vastly increased analytical challenge and yielded many, many more rules.

The choice of minimum group size for the outcome variable can, however, generate problems for biological interpretation. For example, here we observe that hemoglobin AA (normal hemoglobin) increases risk for development of PFA compared to the mutated sickle form, AS, which is known to confer protection. Importantly, we cannot conclude from our analysis that AS confers protection. In general, care must be taken in interpreting the direction of the effect and further specific analyses should be performed prior to establishing formal conclusions.

Repeated measures and potential pseudo-replication of events from the same individual are difficult to take into account. Whilst

this can be accounted for *a posteriori* in confirmatory classical analyses, this cannot be currently taken into account in HyperCube®. For the rules obtained, the full information on the number of events and the number of people contributing to those events can be provided, as done here. In addition, with regard to use of human genetic factors as explanatory variables, bias due to population stratification is difficult to take into account in HyperCube®. Such a bias needs to be secondarily tested on validated rules using classical methods.

A final limitation is that HyperCube® requires huge computational power and needs to use massive parallel processing. Today, HyperCube® is accessible as a web based software that requires no specific learning skills, though it requires significant computing power provided through SaaS architecture. Currently HyperCube® is used on various complex problems [7]; we now report an analysis of epidemiological data using this algorithm. HyperCube® classified events or individuals into high and low risk groups defined by combinations of variables. It efficiently sub-stratified quantitative variables to optimize the effect. In addition, it was able to identify interactions among the variables. These tasks are not easy to perform using standard data mining methods. HyperCube® is very useful in handling large datasets with complexity of the dependant variable, such as found in large epidemiological studies and genetic studies. We have proved that the rules identified by HyperCube® are the optimal in the dataset and that no other methods can find them in a reasonable time. Search of local over density in m-dimensional space, explained by easily interpretable rules, is thus seemingly ideal for generating hypotheses for large datasets to unravel the complexity inherent in biological systems. Hypotheses generated by this data mining program should be validated using classical statistical methods and/or by biological experimentation. Further statistical analyses, to provide adequate description and inference on the sub-population identified in a rule, have to be performed by using specific models (e.g. Generalized Estimating Equations [28] or Generalized Linear Mixed Models [29] to take into account repeated measures and/or genetic covariance between individuals, or distribution of the dependent variable).

Materials and Methods

Ethics statement

The project protocol and objectives were carefully explained to the assembled village population and informed consent was individually obtained from all subjects either by signature or by thumbprint on a voluntary consent form written in both French and in Wolof, the local language. Consent was obtained in the presence of the school director, an independent witness. For very young children, parents or designated tutors signed on their behalf. The protocol was approved by the Ethical Committee of the Pasteur Institute of Dakar and the Ministry of Health of Senegal. An agreement between Institut Pasteur de Dakar, Institut de Recherche pour le Développement and the Ministère de la Santé et de la Prévention of Senegal defines all research activities in the study cohorts. Each year, the project was re-examined by the Conseil de Perfectionnement de l'Institut Pasteur de Dakar and the assembled village population; informed consent was individually renewed from all subjects.

Populations

The populations studied come from two family-based village cohorts, Dielmo and Ndiop, in Senegal. These populations have been recruited for a long-term immunological and epidemiological study [8]. Malaria transmission intensity differs between the 2 villages because of the presence of the river in one of them that offers a mosquito breeding site all-year round.

Research stations have been installed in the villages with full-time nurses and paramedical personnel. Almost all fever episodes were reported to the clinics with blood smears checked for malaria parasites. The outcome of interest is a *Plasmodium falciparum* malaria attack (PFA). PFA was defined as a presentation with measured fever (axillary temperature $>37.5^{\circ}C$) or fever-related symptoms (headache, vomiting, subjective sensation of fever) associated with i) a *P. falciparum* parasite/leukocyte ratio higher than an age-dependent pyrogenic threshold previously identified in the patients from Dielmo [30], ii) a *P. falciparum* parasite/leukocyte ratio higher than 0.3 parasite/leukocyte in Ndiop. The threshold was used because of high prevalence of asymptomatic infections in the populations, as occurs in regions endemic for malaria.

Some explanatory variables are time-dependent and were therefore evaluated for each trimester. These included current age, experience of exposure to other *Plasmodium spp.* (*Plasmodium ovale* and *Plasmodium malariae*) before the current trimester defined by the cumulated number of previous infections, the corresponding year and trimester, time spent in the village during the current trimester. Other variables are individual-dependent including sex, geographic location (e.g. village, house), and genetic profiles (e.g. blood type, hemoglobin type, Glucose-6-phosphate dehydrogenase (G6PD) deficiency status (genotype and Enzyme activity). All variables are summarized in Table 1 and 2.

Mutation characterization

Sickle cell mutation and alpha-globin 3.7 deletion were typed as described [31]. G6PD mutations and ABO polymorphisms were typed by PCR-RFLP, SNaPshot® (Applied Biosystems, Foster City, USA) or TaqMan SNP genotyping assays (ABI Prism®-7000 Sequence Detection System, Applied Biosystems, Foster City, USA) according to the manufacturer recommendation. Primers, probes and restriction enzymes used are shown in Table 12. PCR conditions will be sent on request. ABO polymorphisms were selected to differentiate the A, B and O alleles [32].

HyperCube® data mining algorithm

The HyperCube® technology is accessible as a web based software that requires no specific learning skills, though it requires a significant computing power provided through a SaaS architecture (Institute of Health & Science, Paris, France). A hypercube is a subspace defined by a combination of conditions, each condition being either a range or a modality of a continuous or discrete variable. A hypercube has various characteristics: its dimension, the number of variables involved; the "Lift", the measure of the over density compared to the whole database, the "Size", the number of points included in the hypercube.

After defining the dependent variable, HyperCube® program generates a series of rules by exhaustively exploring the space of the random variables, generating optimal subspaces significantly enriched with the occurrence of events, and defining for each interesting subspace, its explicative variables and their corresponding values. A rule is a set of a limited number of continuous and/or categorical variables and their associated values. A search by HyperCube® program is divided in 3 steps:

(i) *A stochastic exploration of the space of random variables*: Subspaces are exhaustively generated following this procedure: One point is randomly chosen as a germ (a starting point) in the *m*-dimensional space defined by the *m* explanatory variables; after a 2nd point is randomly selected to form a segment.

Table 12. Primer sequences probes, restriction enzymes and rs numbers used for typing Glucose-6-phosphate dehydrogenase (G6PD) and ABO blood group single nucleotide polymorphisms.

Polymorphism name	rs number	Genotyping method	Forward primer (5′-3′)	Reverse primer (5′-3′)	Probe (5′-3′)	Restriction enzyme
G6PD						
G6PD-202	rs1050828	PCR-RFLP	GTGGCTGTTCCGGGATGGCCTTCTG	CTTGAAGAAGGGCTCACTCTGTTTG		*FokI*
G6PD-376	rs1050829	PCR-RFLP	CGTGAATGTTCTTGGTGACG	CCCCAGAGGAGAAGCTCA		*NlaIII*
G6PD-542	rs5030872	TaqMan®	ACCGCATCATCGTGGAGAAG	AGATCTGGTCCTCACGGAACA	probe 1-AGAGCTCTGACCGGCTG	
					probe 2-AGAGCTCTGTCCGGCTG	
G6PD-968	rs76723693	TaqMan®	TGTGGTCCTGGGCCAGTA	GACGACGGCTGCAAAAGT	probe 1-CCAAAGGGTACCTGGACGA	
					probe 2-CAAAGGGTACCCGGACGA	
ABO						
ABO-261	rs8176719	PCR-RFLP	GCCTCTCTCCATGTGCAGTA	TCCACAGTCACTCGCCACT		*RsaI*
ABO-297	rs8176720	TaqMan®	TGGCTGGCTCCCATTGTC	CCTGAACTGCTCGTTGAGGAT	probe 1-CGATGTTGAATGTGC	
					probe 2-CGATGTTGAACGTGC	
ABO-467	rs1053878	PCR-RFLP	TGCAGATACGTGGCTTTCCT	CGCTCGCAGAAGTCACTGAT		*EagI*
ABO-526	rs7853989	PCR-RFLP	TGCAGATACGTGGCTTTCCT	CGCTCGCAGAAGTCACTGAT		*BsaHI*
ABO-771	rs8176745	SNaPshot®	CGGGAGGCCTTCACCTAC	CACAAGTACTCGGGGGAGAG	AAAAAACAGTCCCAGGCCTACATCCC	

doi:10.1371/journal.pone.0024085.t012

These two points correspond to apical points of a starting subspace having a hypercube design and represent the diagonal of this hypercube. This diagonal (jointly the volume of the hypercube) will be optimally increased. Each subspace is selected depending on two constraints: its size, the number of events included in the subspace, and its purity, the percentage of positive events in the subspace. To define explanatory variables, the corresponding axe for each variable delimiting the subspace is suppressed, and the subsequent subspace tested for satisfying the previous constraints. The variables for which the corresponding axe must be present to satisfy these constraints are the explanatory variables. The subspace is cancelled if it does not satisfy the constraints defined by the user and a new subspace is generated.

(ii) *An optimization of the characteristic of the hypercube*: The volume of each initial hypercube selected at the first step is locally maximized depending on a Z score using genetic algorithms, and always constrained to a minimum purity.

(iii) *Validation of the rule using a non-parametric approach*: The Z score of the optimized hypercube is compared to those generated by a random permutation of the dependant variable.

For exhaustiveness, these three steps are repeated until all points have been used as starting point and all the events have been studied; i.e. all the events in the learning dataset have been included in at least one rule. The user can stop the learning process at any time and know the coverage of his exploration. Due to human limitations in understanding complex rules, the maximal number of explanatory variables inside each rule can be fixed, thereby defining complexity. HyperCube® uses an exhaustive non-parametric and non-Euclidean methodology, it does not use proximity between events but only generates subspaces in which events are present or not.

We have first to define variables to introduce into the learning data set. If necessary, the outcome variable is transformed into a dichotomous variable. In our case, the number of clinical *P. falciparum* attacks by trimester was divided into two groups: "no attack during the trimester", and "at least one attack during the trimester". This is done on a local computer using MATRIX program with two main functions: "Simple lift" and "Correlation". "Simple lift" classifies variables according to their first order effect and has 3 major roles: to verify consistency of the data, to detect circular variables and to detect variables with pivot points that define threshold values for the impact of a variable on the outcome. "Spearman (or Pearson) Correlation" associated with "Simple lift" will help to define which variable to choose amongst the correlated variables. Sometimes, a combined variable from correlated variables is the best choice. The matrix is loaded onto the supercomputer after defining on which part of the database the learning process will be performed. In our case, we chose the learning set of Dielmo cohort. We defined on which group of the dichotomous variable the learning process would be carried out, in our case "at least one attack during the trimester". First, we constructed a Signal Intensity Graph (SIG), which defines the relationship between the two main parameters of a learning process, "purity" and "size". This graph shows the value of the "purity" for 5 different "sizes" defined from data of the database and of a randomized database. This graph can be downloaded onto the local computer. After defining the last parameter, "Complexity", which defines the maximum number of variables per rule, the learning process is run. From the total number of rules, a set of minimized rules is obtained from an iterative process. In the first step, the rule explaining the most number of events is chosen and at each of the following steps the rule explaining the maximal number of events in the remaining event space not included in the first rule is added. The iterative process is stopped when all the events explained by the total number of rules are explained by the set of minimized rules. The total number of rules and/or the minimized rules can be downloaded onto the local computer to perform further analysis.

Statistical analysis

We used Classification and Regression Trees (CART) methods [5] to split continuous explanatory variables to categories. We performed a Logistic Regression Model to estimate overall RR and OR of combinations of factors [33,34].

Identity-by-descent (IBD)

We estimated multipoint IBD using genome wide microsatellite genotypes by MERLIN [35]. We defined "IBD-based mean genetic relatedness" for an individual to the rest of the population, based on IBD probabilities, as the mean of his kinship coefficients with all other individuals $= (1/(N-1)) \times (1/M) \times \sum_i \sum_m [0.5 \times P1 + P2]_{i,m}$, $i = 1, \ldots, N-1$ and $m = 1, \ldots, M$ where N is the number of individuals genotyped for the microsatellite markers in the population, M the number of microsatellite markers P1 = probability of sharing 1 allele and P2 = probability of sharing 2 alleles.

Pedigree-based mean genetic relatedness

The genetic covariance is computed as $r(A,B) = 2 \times coancestry(A,B)$ where the *coancestry* between A and B is calculated referring to this following method (Falconer and Mackay 1996) [36]: $coancestry(A,B) = \sum_p (1/2)^{n(p)} \times (1 + I_{Common\ Ancestor})$ where p is the number of paths in the pedigree linking A and B, $n(p)$ the number of individuals (including A and B) for each path p and I_X is the *coancestry* between the two parents of X, which is set to 0 if X is a founder. We defined the mean relatedness coefficient for an individual to the rest of the population, based on the pedigree, as the mean of his kinship coefficients with all other individuals. The variable named "Pedigree-based mean genetic relatedness" was defined by this measure.

Acknowledgments

We are grateful to the villagers of Dielmo and Ndiop for their participation and continued collaboration in this project. We thank the administrative authority of Institut Pasteur of Dakar for their continuous support. We particularly thank Christophe Rogier, Andre Spiegel and Laurence Baril as well as the field workers for their sustained contribution to the project and in generating and maintaining the malaria databases.

Author Contributions

Conceived and designed the experiments: JFB AS AH. Performed the experiments: CL RP. Analyzed the data: CL AD ABH JFB. Contributed reagents/materials/analysis tools: AT CS JFT ABL JF AB GD FDS. Wrote the paper: CL RP ABH AH AS JFB.

References

1. Nelson MR, Kardia SL, Ferrell RE, Sing CF (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 11: 458–470.

2. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, et al. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69: 138–147.

3. McKinney BA, Reif DM, Ritchie MD, Moore JH (2006) Machine learning for detecting gene-gene interactions: a review. Appl Bioinformatics 5: 77–88.
4. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392–404.
5. Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and Regression Trees Chapman and Hall.
6. Breiman L (2001) Random Forests. Machine Learning 45: 5–32.
7. Institute of Health & Science website. Available: http://www.institute-health-science.org. Accessed 2011 May 30.
8. Trape JF, Rogier C, Konate L, Diagne N, Bouganali H, et al. (1994) The Dielmo project: a longitudinal study of natural malaria infection and the mechanisms of protective immunity in a community living in a holoendemic area of Senegal. Am J Trop Med Hyg 51: 123–137.
9. Rogier C, Tall A, Diagne N, Fontenille D, Spiegel A, et al. (1999) Plasmodium falciparum clinical malaria: lessons from longitudinal studies in Senegal. Parassitologia 41: 255–259.
10. Calcagno V, Mazancourt CD (2010) glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models. Journal of Statistical Software 34: 1–29.
11. Sakuntabhai A, Ndiaye R, Casademont I, Peerapittayamongkol C, Rogier C, et al. (2008) Genetic determination and linkage mapping of Plasmodium falciparum malaria related traits in Senegal. PLoS One 3: e2000.
12. Sokhna CS, Faye FBK, Spiegel A, Dieng H, Trape JF (2001) Rapid reappearance of Plasmodium falciparum after drug treatment among Senegalese adults exposed to moderate seasonal transmission. Am J Trop Med Hyg 65: 167–170.
13. Mackinnon MJ, Mwangi TW, Snow RW, Marsh K, Williams TN (2005) Heritability of malaria in Africa. PLoS Med 2: e340.
14. Bruce MC, Donnelly CA, Alpers MP, Galinski MR, Barnwell JW, et al. (2000) Cross-species interactions between malaria parasites in humans. Science 287: 845–848.
15. McKenzie FE, Bossert WH (1999) Multispecies Plasmodium infections of humans. J Parasitol 85: 12–18.
16. Domarle O, Migot-Nabias F, Mvoukani JL, Lu CY, Nabias R, et al. (1999) Factors influencing resistance to reinfection with Plasmodium falciparum. Am J Trop Med Hyg 61: 926–931.
17. Molineaux L (1980) The Garki project : research on the epidemiology and control of malaria in the Sudan savanna of West Africa/by L. Molineaux and G. Gramiccia. .
18. Molineaux L, Storey J, Cohen JE, Thomas A (1980) A longitudinal study of human malaria in the West African Savanna in the absence of control measures: relationships between different Plasmodium species, in particular P. falciparum and P. malariae. Am J Trop Med Hyg 29: 725–737.
19. Domarle O, Migot-Nabias F, Pilkington H, Elissa N, Toure FS, et al. (2002) Family analysis of malaria infection in Dienga, Gabon. Am J Trop Med Hyg 66: 124–129.
20. Flori L, Kumulungui B, Aucan C, Esnault C, Traore AS, et al. (2003) Linkage and association between Plasmodium falciparum blood infection levels and chromosome 5q31–q33. Genes Immun 4: 265–268.
21. Garcia A, Marquet S, Bucheton B, Hillaire D, Cot M, et al. (1998) Linkage analysis of blood Plasmodium falciparum levels: interest of the 5q31–q33 chromosome region. Am J Trop Med Hyg 58: 705–709.
22. Rihet P, Traore Y, Abel L, Aucan C, Traore-Leroux T, et al. (1998) Malaria in humans: Plasmodium falciparum blood infection levels are linked to chromosome 5q31–q33. Am J Hum Genet 63: 498–505.
23. Muller I, Genton B, Rare L, Kiniboro B, Kastens W, et al. (2009) Three different Plasmodium species show similar patterns of clinical tolerance of malaria infection. Malar J 8: 158.
24. Black J, Hommel M, Snounou G, Pinder M (1994) Mixed infections with Plasmodium falciparum and P malariae and fever in malaria. Lancet 343: 1095.
25. Collins WE, Jeffery GM (1999) A retrospective examination of sporozoite- and trophozoite-induced infections with Plasmodium falciparum in patients previously infected with heterologous species of Plasmodium: effect on development of parasitologic and clinical immunity. Am J Trop Med Hyg 61: 36–43.
26. Bousema JT, Drakeley CJ, Mens PF, Arens T, Houben R, et al. (2008) Increased Plasmodium falciparum gametocyte production in mixed infections with P. malariae. Am J Trop Med Hyg 78: 442–448.
27. McKenzie FE, Jeffery GM, Collins WE (2002) Plasmodium malariae infection boosts Plasmodium falciparum gametocyte production. Am J Trop Med Hyg 67: 411–414.
28. Zeger SL, Liang KY (1986) Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42: 121–130.
29. McCulloch CE (2008) Generalized, linear, and mixed models; statistics Wsipa, editor: Hoboken, N.J. : Wiley.
30. Rogier C, Commenges D, Trape JF (1996) Evidence for an age-dependent pyrogenic threshold of Plasmodium falciparum parasitemia in highly endemic populations. Am J Trop Med Hyg 54: 613–619.
31. Lawaly YR, Sakuntabhai A, Marrama L, Konate L, Phimpraphi W, et al. (2010) Heritability of the human infectious reservoir of malaria parasites. PLoS One 5: e11358.
32. Yamamoto F, McNeill PD, Hakomori S (1995) Genomic organization of human histo-blood group ABO genes. Glycobiology 5: 51–58.
33. Cox DR, Snell EJ (1969) The analysis of binary data Chapman and Hall.
34. Hosmer DW, Lemeshow S (2000) Applied logistic regression. Wiley Series in Probability and Mathematical Statistics.
35. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97–101.
36. Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics; 4, editor. London: Longman.

OPEN ACCESS Freely available online

PLoS one

Impact of Changing Drug Treatment and Malaria Endemicity on the Heritability of Malaria Phenotypes in a Longitudinal Family-Based Cohort Study

Cheikh Loucoubar[1,2,3,4], Bronner Goncalves[1], Adama Tall[3], Cheikh Sokhna[5], Jean-François Trape[5], Fatoumata Diène Sarr[3], Joseph Faye[3], Abdoulaye Badiane[3], Alioune Badara Ly[3], Aliou Diop[6], Avner Bar-Hen[2,4], Jean-François Bureau[1], Anavaj Sakuntabhai[1,7], Richard Paul[1]*

1 Institut Pasteur, Unité de Génétique Fonctionnelle des Maladies Infectieuses, Paris, France, **2** Laboratoire de Mathématiques Appliquées Paris 5 (UMR 8145), Université Paris Descartes, Paris, France, **3** Unité d'Épidémiologie des Maladies Infectieuses (UR 172, ED-SEV, Université Cheikh Anta Diop), Institut Pasteur de Dakar, Dakar, Senegal, **4** Ecole des Hautes Etudes en Santé Publique, Rennes, France, **5** Institut de Recherche pour le Développement, Laboratoire de Paludologie, Dakar, Senegal, **6** Laboratoire d'Études et de Recherche en Statististique et Développement, Université Gaston Berger, Saint-Louis, Senegal, **7** Center of Excellence for Vectors and Vector-Borne Diseases, Faculty of Science, Mahidol University, Bangkok, Thailand

Abstract

Despite considerable success of genome wide association (GWA) studies in identifying causal variants for many human diseases, their success in unraveling the genetic basis to complex diseases has been more mitigated. Pathogen population structure may impact upon the infectious phenotype, especially with the intense short-term selective pressure that drug treatment exerts on pathogens. Rigorous analysis that accounts for repeated measures and disentangles the influence of genetic and environmental factors must be performed. Attempts should be made to consider whether pathogen diversity will impact upon host genetic responses to infection. We analyzed the heritability of two *Plasmodium falciparum* phenotypes, the number of clinical malaria episodes (*PFA*) and the proportion of these episodes positive for gametocytes (*Pfgam*), in a family-based cohort followed for 19 years, during which time there were four successive drug treatment regimes, with documented appearance of drug resistance. Repeated measures and variance components analyses were performed with fixed environmental, additive genetic, intra-individual and maternal effects for each drug period. Whilst there was a significant additive genetic effect underlying *PFA* during the first drug period of study, this was lost in subsequent periods. There was no additive genetic effect for *Pfgam*. The intra-individual effect increased significantly in the chloroquine period. The loss of an additive genetic effect following novel drug treatment may result in significant loss of power to detect genes in a GWA study. Prior genetic analysis must be a pre-requisite for more detailed GWA studies. The temporal changes in the individual genetic and the intra-individual estimates are consistent with those expected if there were specific host-parasite interactions. The complex basis to the human response to malaria parasite infection likely includes dominance/epistatic genetic effects encompassed within the intra-individual variance component. Evaluating their role in influencing the outcome of infection through host genotype by parasite genotype interactions warrants research effort.

Citation: Loucoubar C, Goncalves B, Tall A, Sokhna C, Trape J-F, et al. (2011) Impact of Changing Drug Treatment and Malaria Endemicity on the Heritability of Malaria Phenotypes in a Longitudinal Family-Based Cohort Study. PLoS ONE 6(11): e26364. doi:10.1371/journal.pone.0026364

Editor: Jose Antonio Stoute, Pennsylvania State University College of Medicine, United States of America

Received July 29, 2011; **Accepted** September 25, 2011; **Published** November 3, 2011

Funding: Doctoral finances for Cheikh Loucoubar awarded by the Ecole des Hautes Etudes en Santé Publique. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rpaul@pasteur.fr

Introduction

The genomics era has heralded a plethora of Genome Wide Association (GWA) studies that have successfully identified genetic determinants of many medical disorders [1–4]. Heritability analyses provide an indication of the genetic contribution underlying a specified phenotype. Whereas in the case of monogenic diseases genetic determinants in GWA studies account for the estimated heritability, there is considerable missing heritability in more complex diseases [5]. This had led to an intense debate of the potential causes for this, citing amongst others the potentially important roles of epistasis, gene-environmental interaction and the confounding effect of population specific genetic architecture [6]. In addition to genetic explanations, one potential source contributing to the missing heritability concerns the phenotype; poorly resolved phenotypes lower the power to detect genetic variants [7].

The application of GWA studies to infectious diseases has only more recently developed [8–10], but is likely to become increasingly implemented [11]. Infectious disease phenotypes are, however, composite phenotypes reflecting both the human and pathogen genetics and their interactions. Thus, the phenotype "problem" is likely to be much greater than in non-infectious diseases. Over the long-term, host-pathogen co-evolution will maintain genetic variation if the additive genetic value of a host genotype changes when parasites evolve in response to the

selection induced by the host [12]. This, thus, may be apparent in the local genetic architecture of the human genetics determining specific traits, where populations have undergone widely different exposure to the pathogen. In addition, despite the current efforts to untangle the genetic basis to complex diseases [13], no attention has been paid to the impact of radical short-term changes in the pathogen population genetic structure, such as those induced by drug pressure, on the human genetic contribution to infection phenotypes.

In recent years, particular attention has been paid to addressing the human genetic susceptibility and resistance to *Plasmodium falciparum* malaria [14–16]. Sickle cell trait has long been recognized as having a protective effect against severe disease [17,18] and this provided a positive control for the first GWA study of severe malaria [19]. Following this success and in the knowledge that the human genetic response to malaria parasite infection is complex and polygenic [20], it is now widely admitted that well-conducted epidemiological studies that take into account confounding environment factors are required [21]. In general, the requisite large sample size for GWA studies necessarily means combining participants from many sites. Whilst among-site variation in human population sub-structure and in the intensity of transmission can in principle be taken into account, such confounding variation may have more subtle effects. Variation in the intensity of transmission, for example, not only has discernable effects on the development of immunity, it also influences parasite genetic diversity [22].

To date genetic analyses have implicitly assumed that any variation brought about by parasite diversity will only have a minor impact, especially with very broad binary phenotypes such as severe versus mild malaria. This has been to some extent confirmed in animal models, but significant host-by-parasite interactions have been observed [23]. In contrast to such extreme binary disease phenotypes, there has been increasing interest in quantitative phenotypes that describe the outcome of infection [16,24–27]. Such phenotypes focus on the actual biology of the parasite within the human host, rather than the extreme disease phenotype, but may be more affected by changes in parasite diversity. Parasite genetic variation in growth rate, transmissibility and other biological phenotypes is well recognized [28] and thus quantitative malaria phenotypes may be influenced strongly by parasite genetics. Indeed, it was recently demonstrated that there was a parasite genetic contribution to time to clearance following treatment [29]. Transmission intensity influences the number of different parasite clones within an infection, which itself can impact on quantitative phenotypes [30]. Moreover, malaria parasites exhibit extensive phenotypic plasticity and quantifiable parasite phenotypes are affected by the immunological and hematological state of the host [31]. Finally, parasite populations evolve over time, especially in the face of persistent drug pressure and there has been recent suggestion that drug resistance is linked to or will select for virulence of the parasite [32,33]. All such sources of variation in the parasite population may significantly alter the observed outcome of infection and thus cloud the signal in the genetic analyses.

Here we address the extent to which malaria phenotypes in a longitudinal family-based epidemiological study are influenced by the changes in anti-malarial drug treatment and in transmission intensity from 1990 to 2008. We estimate the heritability of two *P. falciparum*-related phenotypes: the number of clinical malaria episodes (*PFA*) [16] and the proportion of infections carrying gametocytes (parasite stages that can infect mosquitoes) (*Pfgam*) [27,34]. Heritability is an important parameter that determines statistical power in gene-mapping studies that use pedigree information. A large heritability implies a strong correlation between phenotype and genotype, so that loci with an effect on the phenotype can be more easily detected [35]. These two phenotypes were chosen to be representative of different types of phenotype: *PFA* will be strongly influenced by variation in transmission intensity, whereas *Pfgam* will more strongly reflect the host-parasite interaction. In addition to calculating the heritability, we estimate the shared environment (here house) and intra-individual (also known as "permanent environment") effects, including maternal effects.

Materials and Methods

Ethics statement

The project protocol and objectives were carefully explained to the assembled village population and informed consent was individually obtained from all subjects either by signature or by thumbprint on a voluntary consent form written in both French and in Wolof, the local language. Consent was obtained in the presence of the school director, an independent witness. For very young children, parents or designated tutors signed on their behalf. The protocol was approved by the Ethical Committee of the Institut Pasteur de Dakar and the Ministère de la Santé et de la Prévention of Senegal. An agreement between Institut Pasteur de Dakar, Institut de Recherche pour le Développement and the Ministère de la Santé et de la Prévention of Senegal defines all research activities in the study cohorts. Each year, the project was re-examined by the Conseil de Perfectionnement of the Institut Pasteur de Dakar and the assembled village population; informed consent was individually renewed from all subjects.

Study site and study population

The study was conducted in the malaria research project carried out since 1990 in a family-based cohort in Senegal, which has perennial holoendemic transmission (high force of infection). This site is managed by a tripartite agreement between the Institut Pasteur de Dakar, the Institut de Recherche pour le Développement and the Ministère de la Santé et de la Prévention of Senegal. A field research station with a dispensary run by nurses was constructed for the program and the health care is free-of-charge for the volunteers. All participants were asked to come to a study clinic for all their healthcare needs. Every person satisfying adhesion conditions could become a volunteer and every volunteer could leave the study at any time, therefore forming a dynamic open cohort. Further details of the study sites and adhesion criteria are previously described [36,37].

The family structure (pedigree) was available after a demographic census performed for every volunteer at his adhesion in the project. A verbal interview of mothers or key representatives of the household was used to obtain information on genetic relationships between studied individuals, their children, their parents, and to identify genetic links among the population. The total pedigree comprised 828 individuals, including absent or dead relatives, composed of 206 nuclear families (father – mother couples with at least one child) with an average of 3.6 children each. In addition, previous typing with microsatellites has enabled the construction of a pedigree based on Identity-by-Descent (IBD) using MERLIN [16,38].

Data collection - P. falciparum malaria phenotypes

The parasite phenotypes analyzed were: (i) the number of *P. falciparum* clinical episodes per trimester (*PFA*) and (ii) the proportion of clinical episodes that were positive for gametocytes, parasite stages transmissible to mosquitoes (*Pfgam*). A malaria

episode is defined as a clinical presentation with measured fever (axillary temperature $>37.5^{\circ}$C) or fever-related symptoms (headache, vomiting, subjective sensation of fever) and with a blood smear positive for *P. falciparum* at a parasite/leukocyte ratio higher than the age-dependent pyrogenic threshold previously defined by Rogier et al. [39]. For *PFA*, we first excluded any observations of each trimester for which the individual concerned was not present for at least 30 days ($=1/3$ of the trimester). Individuals satisfying presence conditions without any *P. falciparum* clinical episode in a trimester were classified as *PFA = 0*; individuals satisfying presence conditions with 1 or more malaria clinical episodes in a trimester correspond to person-trimester with *PFA* = {1, 2, 3, 4, or 5}. Repeated clinical presentations within 15 consecutive days were not considered to be independent and were excluded from the analyses, unless there was a parasite negative blood smear between two clinical episodes. In all cases parasite positivity was established as follows. Thick and thin blood films were prepared and stained by 3% Giemsa stain. Blood films were examined under an oil immersion objective at ×1000 magnification by the trained laboratory technicians and 200 thick film fields were examined to count the number of asexual and gametocyte parasite stages. The proportion of clinical episodes carrying gametocytes excluded any repeated clinical presentations within 15 days of previous treatment.

The following covariates were considered: sex, house, season (4 categories: Jul–Sep; Oct–Dec; Jan–Mar; Apr–Jun) nested within year, year (5 categories: 1990 to 1994 for quinine period, 5 categories: 1995 to 1999 for 1st chloroquine period, 4 categories: 2000 to 2003 to the 2nd chloroquine period, 3 categories: 2004 to 2006 for fansidar period, 3 categories: 2006 to 2008 for ACT period) and logarithm of number of days present in each trimester. For *Pfgam*, we additionally considered the presence of other *Plasmodium* spp. parasites (*Plasmodium ovale* and *Plasmodium malariae*; 2 categories: yes/no) and time since last treatment. For *Pfgam*, age was found to be best described as a continuous variable in each drug period. By contrast, age classes <5 years, [5–15[, [15–35[and ≥35 years best described the effect of age on *PFA*. Only individuals for whom there was pedigree information were included in the analysis.

Data analyses

From 1990 to 2008, four different drug regimens were implemented: *Quinine* from 1990 to 1994, *Chloroquine* from 1995 to 2003, *Fansidar* from 2004 to mid-2006 and *Artemisinin-based combination therapy (ACT)* from mid-2006 to 2008. The chloroquine drug period was divided into before (CQ1) and after (CQ2) 1999. This was done both to reduce the chloroquine period data set size and to examine the chloroquine periods prior to and during the observed emergence of parasite resistance to this drug [40]. The statistical analyses were performed independently for each of the five drug treatment periods.

We implemented Generalized Linear Mixed Models (GLMM) using SAS 9.1.3 (SAS Institute Inc., Cary, NC, USA) procedures GLIMMIX, MIXED and INBREED [41–43]. GLMM allows fitting of mixed models with correlated random effects, such as those due to genetic relationships. Random effects are assumed to be normally distributed, and conditional on these random effects, the exogenous variable had (i) a Poisson distribution when the studied phenotype was number of *P. falciparum* clinical episodes per trimester (*PFA*) or (ii) a Binomial distribution when the studied phenotype was the proportion of clinical episodes that were positive for gametocytes (*Pfgam*). Genetic covariance, or relationship among all pairs of individuals in the study and among their parents or more distant ancestors, were stored in a squared matrix, the Pedigree-based genetic relatedness matrix, of dimension $K \times K$ where K is the total number of individuals in the pedigree including those with missing phenotypes. Genetic covariance between two individuals was computed using the pedigree information as described below:

For A and B, a given pair in a pedigree, the genetic covariance is computed as $r(A,B) = 2\times$ *coancestry*(A,B) where the *coancestry* between A and B is calculated using the method presented in Falconer and Mackay (1996) [44]: *coancestry*$(A,B) = \Sigma_p (1/2)^{n(p)} \times (1 + I_{\text{Common Ancestor}})$ where p is the number of paths in the pedigree linking A and B, $n(p)$ the number of individuals (including A and B) for each path p and I_X is the inbreeding coefficient of an individual X, which is equal to the *coancestry* between the two parents of X. I_X is set to 0 if X is a founder. This matrix was built using INBREED procedure of SAS and then integrated into the models [42].

The objective of the model used for the analysis was to estimate and separate different sources of variation underlying the total variation observed for the phenotype: the relative contributions of human genetics (additive genetic variance), intra-individual variance, maternal effects, house effects and unexplained residual variation. The repeated measurements design allows us to separate additive genetic variance from intra-individual variance. The occurrence of related individuals living in different houses allows separation of additive genetic variance from that due to shared household. Therefore, the random part of the mixed models included (i) the house identification variable as random effect assuming independence between houses to capture variance due to houses, (ii) the individual identification variable twice: a first time to capture the additive genetic variance by assuming non-independence between individuals and using the subpart of the Pedigree-based genetic relatedness matrix concerning individuals for which the phenotype was observed as covariance matrix between all pairs and a second time to capture other individual variances (e.g. intra-individual effects) assuming independence between individuals and (iii) the mother identification variable to capture maternal effects, assuming non-independence between mothers and offspring, using the subpart of the Pedigree-based genetic relatedness matrix concerning mothers of individuals for which the phenotype was observed. The unexplained residual variation was then deduced.

PFA was analyzed using a Poisson regression model, which explicitly takes into account the non-negative integer-valued aspect of the dependent count variable. Therefore a GLMM with a Poisson distribution was fitted using SAS proc GLIMMIX and *log* as the link function between E(*PFA* | *covariates*) and a predictor that is linear. Initially a maximal model with all covariates was fitted and a minimal adequate model including only significant covariates was obtained. The effect of each covariate on the outcome variable was estimated taking into account both inbreeding, via the genetic relatedness matrix integrated in the SAS Proc GLIMMIX using the LDATA option, and repeated measures, as well as house effects.

The vector of random effects was assumed to follow a multivariate normal distribution:

$$\gamma = \begin{pmatrix} g \\ m \\ b \\ c \\ \varepsilon \end{pmatrix} \sim N\left[0\,;\ \begin{pmatrix} A_N\sigma_g^2 & 0 & 0 & 0 & 0 \\ 0 & A_M\sigma_m^2 & 0 & 0 & 0 \\ 0 & 0 & I_N\sigma_b^2 & 0 & 0 \\ 0 & 0 & 0 & I_H\sigma_c^2 & 0 \\ 0 & 0 & 0 & 0 & I_n\sigma_\varepsilon^2 \end{pmatrix}\right]$$

where g is the additive genetic effect, m is the maternal effect, b is the intra-individual effect, c is the house effect and ε is a random

residual; σ_g^2, σ_m^2, σ_b^2, σ_c^2, σ_ε^2 are the additive genetic, maternal, intra-individual, house and residual variances, respectively. A_N represents the matrix of additive genetic relationships between individuals, with dimension $N \times N$, A_M represents the matrix of additive genetic relationships of mothers to offspring, with dimension $M \times M$, I_N is an identity matrix with dimension $N \times N$, I_H is an identity matrix with dimension $H \times H$, and I_n is an identity matrix with dimension $n \times n$; and $n = \Sigma_i n_i$ where n_i is the number of measure for individual i, N is the number of individuals for which the phenotype was observed and M the number of their mothers.

The heritability is defined by $\sigma_g^2/(\sigma_g^2 + \sigma_m^2 + \sigma_b^2 + \sigma_c^2 + \sigma_\varepsilon^2)$

For each variance component, an estimate was also generated for each individual contributing to the overall component. Thus, for the additive genetic and intra-individual effects, an estimate was established for each person. Similarly for house and maternal effects, estimates were established for each house and mother.

Pfgam was analyzed by fitting a GLMM with a Binomial distribution, using SAS proc GLIMMIX [41]. The distribution of random effects and corresponding indices were defined as for *PFA* in the first analysis.

Results

Data description and epidemiological analyses of key environmental factors

The first composite phenotype considered was the number of *P. falciparum* clinical episodes per person per trimester (*PFA*). Over the 19-year study period, 713 individuals were present from between one and 75 complete trimesters generating 22,169 person-trimesters of presence. There were a total of 5,680 clinical *P. falciparum* episodes. The maximum number of *PFA* per person-trimester was five and the median was one. 485 individuals had at least one *PFA* positive trimester during the study period. The maximum number of clinical episodes per person per drug period was 40 and the median was two. Table 1 summarizes the data by drug period and additionally gives the mean relatedness (by IBD) of the individuals present in each period. The number of clinical episodes decreased with age ($P<0.0001$) and this decrease was most accurately described by 4 groups (<5 years, 5–14 years, 15–34 years and >35 years old). Year and season also had a consistent influence on the number of clinical episodes ($P<0.0001$). The incidence rate of clinical episodes per trimester decreased significantly following the introduction of Fansidar; this change in the incidence rate is most evident in the most susceptible age group (<5 years of age) (Figure 1).

The second composite phenotype considered was the number of *P. falciparum* clinical episodes that were positive for gametocytes, the parasite stage transmissible to mosquitoes. The prevalence of gametocytes at clinical presentation increased from 37% in the quinine period to 48% in both the chloroquine periods before decreasing to 17% and 12% in the Fansidar and ACT periods respectively (Table 1). The percentage of individuals ever gametocyte positive when having a clinical *P. falciparum* episode likewise increased from 50% in the quinine period to 75% in the second chloroquine period before decreasing to 37% and 25% in the Fansidar and ACT periods respectively. Age, as a continuous variable, was found to negatively associate with gametocyte presence during the quinine ($P = 0.02$), and the two chloroquine periods ($P<0.001$). Yearly variation had a significant impact in all periods except ACT. An increasing number of days of individual presence increased gametocyte carriage in the CQ1 period ($P = 0.02$) and increasing time since last drug treatment increased gametocyte carriage in the Fansidar period ($P = 0.02$).

Heritability analyses – (i) number of P. falciparum clinical episodes per trimester

A. Additive genetic, intra-individual, maternal and house variance components. The narrow sense heritability of *PFA* was estimated by drug period. During the quinine period there was significant heritability, estimated at 46%, but which decreased and became non-significant in the subsequent drug treatment periods (Figure 2 and Table 2). Conversely, the intra-individual effect increased significantly following the quinine period, accounting for over 50% of the observed variance in *PFA*. There was no house effect during any period (Figure 2 and Table 2).

The intra-individual effect includes, amongst other parameters, any maternal contribution, whether genetic or environmental. In the case of malaria parasite infection, for example, maternal antibodies protect the newborn during the first few months of life and thus the mother transfers her acquired immunity. In addition, infection during pregnancy can lead to low birth weight with consequent effects on health of the newborn and potentially later in life [45]. Thus, as classically performed in heritability analyses, we consequently evaluated the contribution of a maternal effect in addition to the additive genetic and intra-individual effects. There was no maternal effect during any drug period.

B. Additive genetic and intra-individual estimates for individuals. Estimates for the additive genetic variance strongly correlated for all the three drug periods for which the total additive genetic variance was not zero (i.e. thus for which there were non-zero genetic estimates per individual). There were only individual significant estimates for the additive genetic effect during the quinine period. Nineteen individuals had significant estimates during the quinine period; fourteen of these were present during more than one drug period but none had significant estimates subsequent to the quinine period. By contrast, five of them had significant estimates for the intra-individual effect in periods subsequent to the quinine period. Overall, individual estimates of genetic effects were highly correlated with intra-individual effects by drug period when non-zero (i.e. for quinine, CQ1 and CQ2 periods, Table 3) ($r = 0.73$, 0.71 and 0.65 respectively).

By definition, major components of the intra-individual variance are features that are particular to each individual. Pertinent to malaria parasite infection would be heterogeneity in exposure to mosquitoes but that which is independent of any detectable household spatial effect; i.e. specific individual behaviors that lead to differential exposure to mosquitoes. We examined how the intra-individual estimates for each individual were correlated over the drug periods. Estimates always correlated in the drug period that followed, but decreasingly so in subsequent drug periods (Table 3). Estimation of the individual contributions to the overall intra-individual effects revealed that 54, 47, 91 and 76 individuals had significant estimates in the CQ1, CQ2, Fansidar and ACT periods respectively. There were no individuals with significant estimates during the quinine period. The majority of these individuals (129 of 191) had a significant estimate in only one drug period. Fifteen and 47 individuals had significant estimates in three and two drug periods respectively.

Of the 210 individuals present throughout the 19 year period, 69 had significant intra-individual estimates: fifty individuals in only one treatment period and the remainder in two ($n = 15$) or three ($n = 4$) different periods. Figure 3 displays a comparative scatter plot of intra-individual estimates in all drug periods. For simplicity, only the 50 individuals with significant estimates during a single drug period are highlighted: individuals with a significant estimate in a specific period are denoted as red stars (CQ1), green squares (CQ2), blue triangles (Fansidar) and yellow circles (ACT)

Table 1. Data summary for analyses of the number of *P. falciparum* clinical episodes per person per trimester (*PFA*) and the number carrying gametocytes (*Pfgam*).

Drug Period	Person-Trimesters	Individuals present	Mean relatedness	Number of Pf episodes	Individuals Pf positive	Range	% Pfgam positive	Individuals Pfgam positive	Range
Quinine	4080	338	0.0082	1454	234	1–40	37.2	117	1–16
CQ1	5469	405	0.0080	1950	245	1–38	47.1	151	1–26
CQ2	4800	423	0.0081	1481	205	1–38	48.6	155	1–28
Fansidar	3753	417	0.0084	466	148	1–11	17.1	55	1–5
ACT	4067	487	0.0083	329	135	1–10	12.2	34	1–3

Shown are the total number of person-trimesters per drug treatment period in which the number of *P. falciparum* clinical episodes occurred, the number of individuals present, their overall genetic relatedness (IBD), the number having a clinical episode, the range in the number of episodes per person, the percentage of these episodes that were positive for gametocytes, the number of individuals ever carrying gametocytes during a clinical episode and the range in the number of times individuals carried gametocytes.

doi:10.1371/journal.pone.0026364.t001

in every graph. In the vertical quinine box column, all points cluster around zero with respect to the x-axis – there is no intra-individual effect in the quinine period. This negligible intra-individual variance component in the quinine period and the subsequent increase in the following periods can be clearly seen in Figure 3: the data points are increasingly spread out along the x-axis from the quinine column through the CQ1, CQ2 and Fansidar columns. The extreme significant values in the CQ1 (red stars), CQ2 (green squares), Fansidar (blue triangles) and ACT (yellow circles) periods clearly separate from the rest in their respective drug periods: thus for example the individuals represented by yellow circles have much larger values than the others in the ACT Y-axis row, whereas these same individuals do not differ from the rest in the CQ1, CQ2 and Fansidar Y-axis rows. This shows in detail how individuals with much higher or lower numbers of *P. falciparum* episodes (very positive or very negative values) have so in only single drug periods. Interestingly, the degree to which the significant points separate from the rest appears to increase with time (i.e. from CQ1 through ACT); the blue triangles (Fansidar) and yellow circles (ACT) are more clearly separated from the rest in their respective Y-axis rows. This increase in the intra-individual variance component as displayed though individual estimates over time is reflected in the summarized intra-individual variance component in Table 2. This

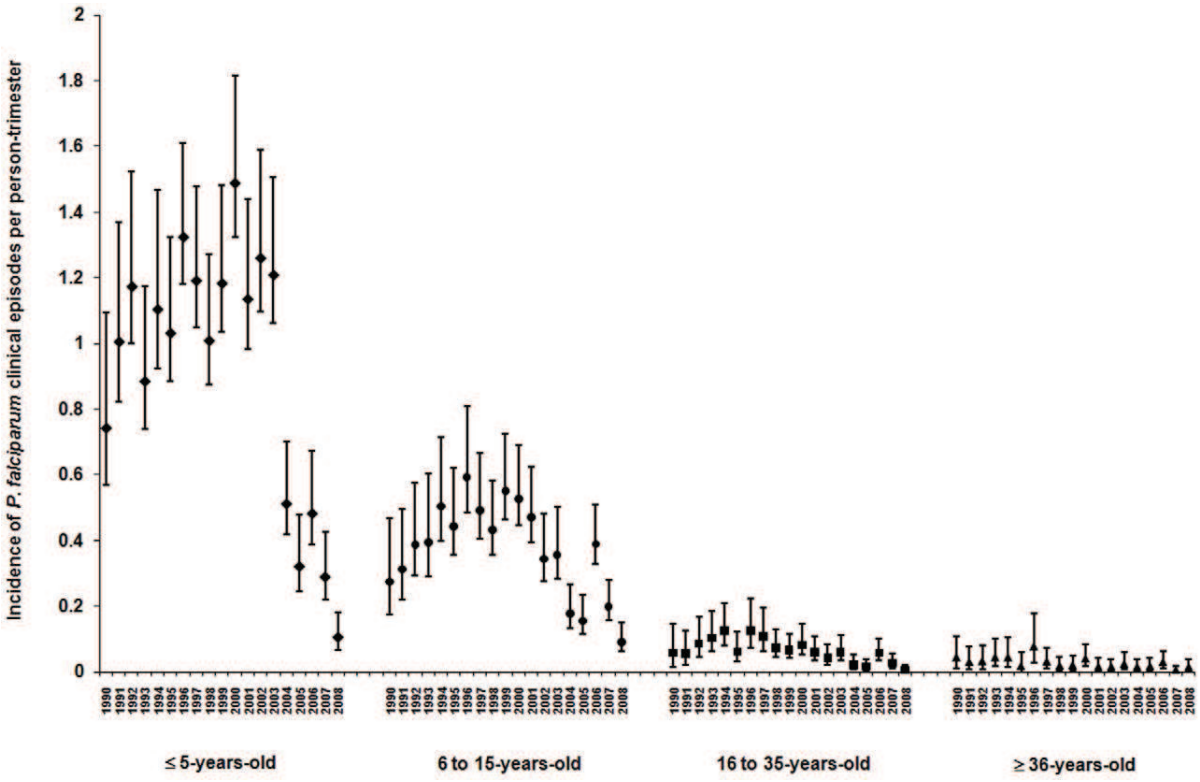


Figure 1. The incidence rate (mean and SEM) of clinical *P. falciparum* episodes per person-trimester (*PFA*) according to age classes (from left to right on the X-axis) <5 years, [5–15], [15–35] and ≥35 years that best describe the effect of age on *PFA*.

doi:10.1371/journal.pone.0026364.g001

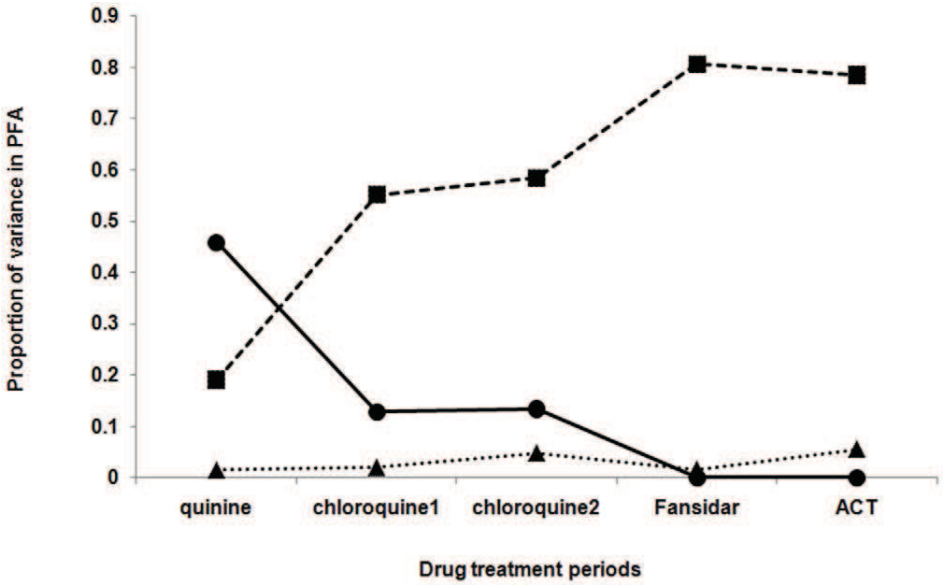


Figure 2. Proportion of variance in the number of clinical *P. falciparum* episodes per trimester explained by additive genetic (solid line), intra-individual (dotted line, squares) and house (thin dotted line, triangles) effects.
doi:10.1371/journal.pone.0026364.g002

shows that as the overall incidence rate drops, there is a growing gap between certain individuals having a high numbers of episodes and the rest. Comparing across drug periods, not only do period-specific significant individual estimates become non-significant in subsequent periods, they seemingly take on increasingly opposed values. This is most evident for CQ1, where the significant estimates for this period, denoted by red stars, decrease in value during the CQ2 and Fansidar periods (Figure 3, horizontal row "ACT"). Similarly for CQ2, significant estimates (green squares) became less than zero in the Fansidar and ACT periods. This suggests that individuals with previously very high numbers of clinical episodes have increasingly fewer numbers of episodes than the rest. One explanation for this would simply be the acquisition of clinical immunity due to repeated exposure to the parasite.

As can be seen in Figure 1, age is a reasonable proxy of the acquisition of immunity and both age and time spent within the site impact upon incidence rate. However, no single factor was found to be shared by individuals with significant intra-individual estimates. I.e. Age, gender and time spent within the village since inception of the study or during the six months prior to the episode were not significant variables determining the intra-individual estimate.

In the knowledge that resistance to chloroquine and then Fansidar emerged during the respective drug treatment periods, a potentially confounding factor would clearly be repetitive presentation of a single infection because of treatment failure. To evaluate whether the observed increases in the intra-individual variance was a result of drug treatment failure, we examined whether individuals with significant individual intra-individual estimates had a shorter time since previous treatment in the quinine and chloroquine periods, when incidence rate remained high and stable. Although the time since previous treatment for those individuals having significant intra-individual estimates at any time was shorter than for those never having significant estimates ($P<0.001$), drug period *per se* had no effect ($P=0.31$). Thus, there was no difference in time between infections in the quinine and 2 chloroquine periods, suggesting that treatment failure was not causing this significant increase in the intra-individual variance component.

Table 2. Variance component analyses of the number of *P. falciparum* clinical episodes (*PFA*) according to drug period.

Drug period	var.comp	std.err	Z	Pr >Z	95% CI Inf	95% CI Sup
Quinine						
Genetic	0.941	0.384	2.450	**0.014**	0.189	1.693
Intra	0.391	0.247	1.580	0.057	0.152	2.343
House	0.030	0.106	0.280	0.390	0.003	8546
residual	0.692	0.016	43.410	<.0001	0.662	0.725
Chloroquine 1						
Genetic	0.257	0.205	1.250	0.211	−0.145	0.658
Intra	1.106	0.209	5.300	**<.0001**	0.789	1.664
House	0.039	0.059	0.670	0.252	0.007	85.995
residual	0.603	0.012	50.300	<.0001	0.580	0.627
Chloroquine 2						
Genetic	0.281	0.242	1.160	0.246	−0.193	0.756
Intra	1.230	0.229	5.370	**<.0001**	0.880	1.838
House	0.101	0.109	0.930	0.177	0.026	6.787
residual	0.493	0.011	46.870	<.0001	0.473	0.514
Fansidar						
Genetic	0.000	-	-	-	-	-
Intra	1.797	0.214	8.380	**<.0001**	1.441	2.304
House	0.036	0.059	0.610	0.272	0.006	392.83
residual	0.395	0.010	41.290	<.0001	0.377	0.415
ACT						
Genetic	0.000	-	-	-	-	-
Intra	1.759	0.208	8.450	**<.0001**	1.413	2.250
House	0.125	0.096	1.300	0.098	0.042	1.390
residual	0.357	0.008	43.240	<.0001	0.341	0.374

Genetic – additive genetic effect; Intra – Intra-individual effect; House – House effect.
doi:10.1371/journal.pone.0026364.t002

Table 3. Correlation of individual estimates of (i) the intra-individual and (ii) additive genetic effects underlying the variation in the number of *P. falciparum* clinical episodes according to drug period.

PFA					
(i) Intra	**Quinine**	**CQ1**	**CQ2**	**Fansidar**	**ACT**
Quinine		0.49***	0.04	−0.01	0.04
CQ1			0.30***	0.002	0.04
CQ2				0.29***	0.18*
Fansidar					0.16*
(ii) Genetic	**Quinine**	**CQ1**	**CQ2**		
Quinine		0.51***	0.23***		
CQ1			0.44***		

*P<0.05,
**P<0.01,
***P<0.001.
doi:10.1371/journal.pone.0026364.t003

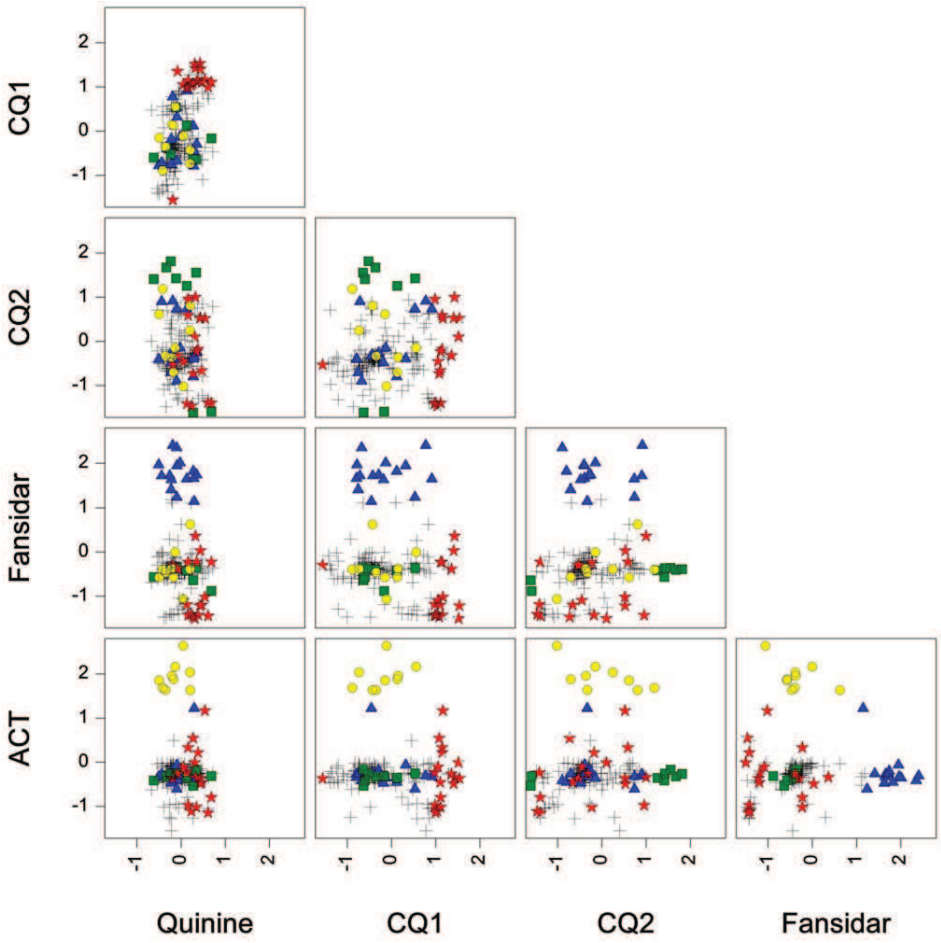


Figure 3. Comparative scatter plot of the Intra-individual estimates per individual per drug period for those individuals present throughout the study period. Individuals with significant intra-individual estimates at any period are shown in color: red stars (significant in CQ1), green squares (significant in CQ2), blue triangles (significant in Fansidar) and yellow circles (significant in ACT).
doi:10.1371/journal.pone.0026364.g003

Heritability analyses – (ii) prevalence of gametocytes during clinical P. falciparum episodes

A. Additive genetic, intra-individual, maternal and house variance components. Heritability for the prevalence of gametocytes during clinical presentation only approached significance during the Fansidar period (P = 0.057) (Table 4, Figure 4). By contrast, the intra-individual effect increased significantly during the chloroquine periods, before becoming non-significant in the Fansidar and ACT periods. There were no house or maternal effects.

B. Additive genetic and intra-individual estimates for individuals. Correlation between estimates for the individual intra-individual and genetic effects revealed a similar pattern to *PFA*: there was significant correlation between estimates in consecutive drug periods, both with respect to estimates of individual intra-individual and additive genetic effects, but no correlation between more distantly related periods (Table 5). Moreover, individual estimates of the genetic and intra-individual effects by drug period were again highly correlated when non-zero (i.e. for Quinine, CQ1, and ACT periods, Table 5) (r = 0.79, 0.77 and 0.80 respectively).

Table 4. Variance component analyses of the prevalence of gametocytes in treated clinical episodes (*Pfgam*) according to drug period.

Drug period	var.comp	std.err	Z	P-value	95% CI Inf	95% CI Sup
Quinine						
genetic	0.423	0.317	1.340	0.181	−0.197	1.044
Intra	0.196	0.272	0.720	0.236	0.040	156.760
House	0.000	.	.	.	.	.
residual	0.932	0.040	23.390	<.0001	0.858	1.015
Chloroquine 1						
genetic	0.164	0.195	0.840	0.401	−0.218	0.545
Intra	0.380	0.218	1.750	**0.041**	0.159	1.814
House	0.000	.	.	.	.	.
residual	0.942	0.035	27.300	<.0001	0.878	1.013
Chloroquine 2						
genetic	0.000	.	.	.	.	.
Intra	0.530	0.119	4.440	**<.0001**	0.356	0.870
House	0.127	0.090	1.410	0.079	0.045	1.050
residual	0.936	0.031	30.010	<.0001	0.878	1.001
Fansidar						
genetic	0.658	0.346	1.900	0.057	−0.021	1.336
Intra	0.000	.	.	.	.	.
House	0.127	0.219	0.580	0.281	0.021	3389.110
residual	0.773	0.055	14.150	<.0001	0.677	0.893
ACT						
genetic	0.570	1.224	0.470	0.641	−1.829	2.970
Intra	0.973	1.035	0.940	0.174	0.250	58.229
House	0.070	0.453	0.150	0.439	0.007	2.5E+65
residual	0.593	0.052	11.500	<.0001	0.503	0.708

Genetic – additive genetic effect; Intra – Intra-individual effect; House – House effect.

doi:10.1371/journal.pone.0026364.t004

The strongly significant intra-individual variance component in CQ2 was due to 12 individuals, eight of whom repeatedly had gametocytes and four who rarely presented with gametocytes. Although the time since previous drug treatment was shorter in these significant individuals, there was no difference between those frequently carrying gametocytes and those rarely doing so (mean 32.4 days SEM 2.5 *vs.* 34.8 days SEM 2.02). There is thus no indication that previous drug treatment is causing this intra-individual effect. No obvious factor, such as age or sickle cell trait, was shared by such individuals. Five of these individuals had significant intra-individual estimates for PFA. Only one individual had a significant intra-individual estimate in the CQ1 period and was not significant in the CQ2 period.

Correlations between malaria phenotypes

There were no significant correlations in either individual additive genetic or intra-individual effects between *PFA* and *Pfgam* at any period where non-zero estimates were available.

Discussion

Here we have made an initial study of the heritability of two *P. falciparum* malaria-related phenotypes in a single population over time. The analyses divided the longitudinal study according to drug treatment to examine the impact of the radical selection pressure that would have been exerted on the parasite population at each change in drug treatment. In addition, the change in transmission intensity occurring over the 19 year enabled us to assess its impact on the heritability of the malaria phenotypes. The evolution of anti-malarial drug resistance and the force of infection have been well studied in the population [36,37,40] and thus we explored heritability in a single population undergoing well-defined environmental changes.

Firstly, it was notable that for *PFA*, a phenotype known to be influenced by human genetics, significant heritability was lost following the change in drug treatment from quinine to chloroquine and in subsequent drug periods. There was no significant change in incidence rate, at least during the quinine and chloroquine periods, no difference in the number of different individuals presenting with clinical disease, or in the pedigree structure (as estimated by the mean genetic relatedness). This suggests that the implementation of the new drug in some way interfered with the human genetic contribution to the outcome of infection. In direct contrast, the intra-individual variance component increased following the implementation of chloroquine.

Intra-individual variance encompasses effects specific to each individual, classically including maternal effects and dominance (non-additive) genetic effects [35,46]. There was no maternal effect for the number of *P. falciparum* clinical episodes in our cohort at any time period. The very high correlation of the individual genetic and intra-individual estimates within each drug period suggests that the two effects are highly confounded. This might be a result of insufficient resolution of the relatedness matrix within each drug period – i.e. either not enough relative-pairs were present within each period and/or the IBD matrix was not sufficiently resolved. This would lead to confounding between shared environmental, additive and non-additive genetic effects [47] and might explain the loss of heritability. However, given the similarity in mean genetic relatedness of individuals in the quinine (when the genetic effect was significant) and other periods, this seems an insufficient explanation. One potential source of variation would be local heterogeneity in individual exposure to mosquitoes. The increase in the intra-individual variance component as the transmission intensity decreased is consistent with

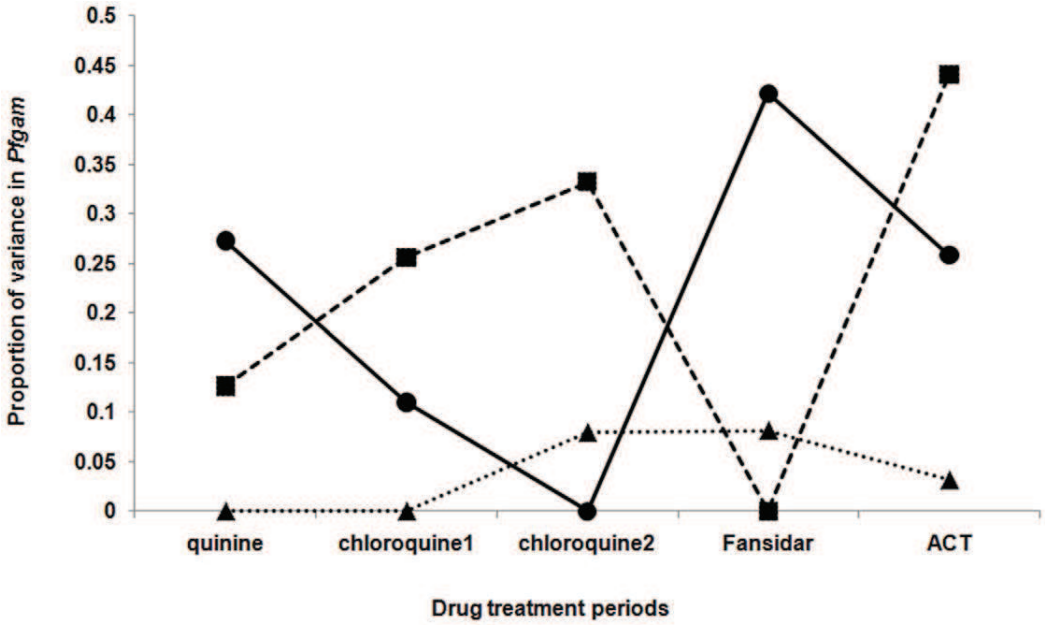


Figure 4. Proportion of variance in the prevalence of *P. falciparum* gametocytes during clinical *P. falciparum* episodes (*Pfgam*) explained by additive genetic (solid line), intra-individual (dotted line, squares) and house (thin dotted line, triangles) effects.
doi:10.1371/journal.pone.0026364.g004

heterogeneity in mosquito biting. Although there was no evidence for a significant impact of shared environment (house), heterogeneity in exposure may occur at a finer level of spatial resolution and/or that reflecting individual behavioral differences ([48] including commentary). One possible source of differential exposure would come from bednet use. However, long-lasting insecticidal-treated nets were not actively promoted until the summer of 2008. Individuals showing extreme intra-individual estimates shared no particular feature, whether it be age, sex or time present in the study site. This argues against any particular behavior or state of immunity contributing to the observed increase in estimates. The intra-individual variance component also includes environmental effects on an individual's phenotype that are constant across (or common to) repeated measures on that individual [46]. It is notable that not only do individual estimates correlate only with those from the subsequent drug period, but also that the majority of the extreme values per individual occurred in one drug period. One explanation for this concerns the impact of the differing drug treatments on the parasite population.

The most evident change in the parasite population during the study was the development of resistance first to chloroquine and then to Fansidar [40]. Treatment failure would result in the same individual presenting more than once for the same infection, thus artificially increasing that individual's number of malaria episodes and hence the estimated intra-individual effect. However, there was no evidence for treatment failure biasing the number of malaria episodes per person. The second effect of drug pressure

Table 5. Correlation of individual estimates of (i) the intra-individual and (ii) additive genetic effects underlying the variation in the proportion of *P. falciparum* clinical episodes positive for gametocytes according to drug period.

Pfgam					
(i) Intra	**Quinine**	**CQ1**	**CQ2**	**Fansidar**	**ACT**
Quinine		0.23*	0.42***	-	0.33
CQ1			0.26**	-	0.11
CQ2				-	0.34**
(ii) Genetic	**Quinine**	**CQ1**	**CQ2**	**Fansidar**	**ACT**
Quinine		0.31**	-	0.40*	0.27
CQ1			-	0.33**	−0.02
CQ2				-	-
Fansidar					0.25*

*P<0.05,
**P<0.01,
***P<0.001.
doi:10.1371/journal.pone.0026364.t005

would be to radically reduce parasite diversity and select for a sub-population of parasites. This process would not be instantaneous, because the majority of the parasite population at any one time in this cohort resides in untreated, asymptomatic infections. Thus, the positive correlations of individual intra-individual and indeed additive genetic estimates in consecutive drug periods might reflect the slowly changing parasite population, implicitly suggesting the existence of specific human-parasite interactions. Drug pressure would result in a stochastic loss of particular parasite genotypes, selection for drug resistant genotypes and potentially selection of parasites more pathogenic for particular individuals. The changing drug regimens would be expected to differentially select for parasite genotypes at each instance, thus making it highly unlikely that the same individuals would be continually susceptible. Whilst an attractive hypothesis, a combination of immune state, behavior and random focal transmission for specific periods of time could generate the observed increase in the intra-individual effect. Our study can not provide the immunological and parasite genetic data that demonstrate changes in the parasite population that would likely have clinical implications for a sub-set of individuals. Moreover, given the complexity and uncertainty of the key parasite antigens that are implicated in the development of clinical immunity [49], such data might not be simple to interpret.

In contrast to the immeasurable effect of very fine scale spatial heterogeneity in exposure to infection that will impact on *PFA*, variability in gametocyte production in an infection will reflect the influence of the host-parasite interaction. Both parasite and host genetics can influence gametocyte production [27,50]. In this study we found no additive genetic effect underlying the proportion of clinical infections with gametocytes, confirming our previous observations [27]. Interestingly, however, there was a similar increase in the intra-individual effect to that observed for *PFA* and the two phenotypes were not correlated. Moreover, as for *PFA*, there was good correlation in estimates across only consecutive periods. These comparable effects to *PFA* were particularly notable during the period when transmission intensity was stable. Subsequently, the decrease in intensity in the Fansidar and ACT periods was accompanied by an even more significant decrease in gametocyte prevalence, resulting in perilously small sample sizes for reliable analysis.

Here, the period of drug treatment strongly influenced this phenotype. Such an influence has been well documented following treatment. Chloroquine increases gametocyte production [51] and Fansidar has also been suggested to increase gametocyte production [52] and/or longevity of gametocyte carriage in a single infection with drug resistant parasites [53]. By contrast, ACT has a gametocytocidal activity and reduces gametocyte carriage [54]. Here, there were no indications that previous treatment contributed to gametocyte presence at presentation, thereby inflating the intra-individual effects in the chloroquine periods. During the Fansidar period, a longer time since treatment was associated with gametocyte presence. The variation in the prevalence of gametocytes at presentation strongly suggests that the parasite population altered according to drug period and the correlated individual intra-individual estimates over successive drug periods are similar to those seen for *PFA*. This would support the hypothesis that changes in the parasite population diversity are contributing to the observed phenotype.

Estimation of heritability in its broad sense in natural populations is not possible and hence narrow sense heritability, which estimates the additive genetic contribution, is calculated. Actual values of heritability are specific for a study population at a particular time and thus strict comparison is not informative, although broad trends can be inferred. The size of heritability provides an indication of the power to detect the effect of individual genes when performing GWA studies. Here it is clear that for several reasons, the choice of the study period for GWA study analysis will affect the quality of the signal. The requirement for large longitudinal data sets to generate sufficient power must therefore be offset by the ever-increasing noise that accompanies long-term data sets – more time means more variance [55].

The peculiarity of the variance component analyses in this study was the replacement of an additive genetic component by an intra-individual component over time. Classical components of the intra-individual component, such as maternal effects, were not found to be the root cause of this and spatial heterogeneity in exposure seems an insufficient explanation, especially during the quinine and chloroquine periods. Insufficient resolution and power of the pedigree matrix may have led to confounding between additive and non-additive genetic components, but again this seems an inadequate explanation given the mean genetic relatedness of the individuals implicated. Observed patterns of individual estimates were consistent with there being specific host-parasite interactions. Although relatives might be expected to respond similarly to an identical parasite, this might not be detectable as an additive genetic component. To what extent changes in the parasite population can impact upon genetic studies is important to understand, both on a practical level of study sampling strategy and at a fundamental level to ask whether candidate genes should be expected to have an effect under whatever circumstances. In the hypothetical case of population fixation of a protective gene, heritability will be zero. What will be the expected heritability in a diverse human population if parasite diversity approaches zero? Will certain genes only be protective against a sub-set of parasites?

In this study we have found suggestive evidence that the parasite population may impact upon estimates of heritability. Whereas a review of theory and data have led to the suggestion that additive genetic variance will represent the majority of genetic variance in complex traits [56], this conclusion averages across populations and may not therefore be the case within a single population [6], especially in the case for infectious diseases. The complex, polygenic basis to the human response to malaria parasite infection may well include dominance/epistatic genetic effects that are encompassed within the intra-individual effect. Evaluating their role in host genotype by parasite genotype interactions in model systems will surely be fruitful. In conclusion, prior genetic analysis of carefully defined phenotypes, both spatially and temporally delimited, must surely not only be a pre-requisite to more detailed GWA studies, but also may be informative for the potential importance of pathogen genetics and the occurrence of host-pathogen interactions.

Acknowledgments

We are grateful to the villagers of Dielmo for their participation and continued collaboration and to the field workers for their active contribution in this project.

Author Contributions

Conceived and designed the experiments: AS RP. Analyzed the data: CL BG RP. Contributed reagents/materials/analysis tools: AT CS J-FT FD-S JF AB AL. Wrote the paper: CL BG AD AB-H J-FB AS RP.

References

1. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41: 703–707.
2. Sullivan PF, de Geus EJC, Willemsen G, James MR, Smit JH, et al. (2009) Genome-wide association for major depressive disorder: a possible role for the presynaptic protein picollo. Mol Psych 14: 359–375.
3. The Wellcome Trust Case Control Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464: 713–720.
4. National Human Genome Research Institute, National Institutes of Health (2011) A Catalog of Published Genome-Wide Association Studies. Available: http://www.genome.gov/gwastudies.
5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461(7265): 747–53.
6. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11(6): 446–50.
7. van der Sluis S, Verhage M, Posthuma D, Dolan CV (2010) Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. PLoS One 5(11): e13929.
8. Thye T, Vannberg FO, Wong SH, Owusu-Dabo E, Osei I, et al. (2010) Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. Nat Genet 42(9): 739–41.
9. Davila S, Wright VJ, Khor CC, Sim KS, Binder A, et al. (2010) Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. Nat Genet 42(9): 772–776.
10. Zhang FR, Huang W, Chen SM, Sun LD, Liu H, et al. (2009) Genomewide association study of leprosy. N Engl J Med 361(27): 2609–2618.
11. Davila S, Hibberd ML (2009) Genome-wide association studies are coming for human infectious diseases. Genome Med 1(2): 19.
12. Haldane JBS (1949) Disease and evolution. Ric Scientifica 19: 68–76.
13. Rao DC (2008) An overview of the genetic dissection of complex traits. Adv Genet 60: 3–34.
14. Kwiatkowski DP (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet 77: 171–192.
15. Timmann C, Evans JA, Konig IR, Kleensang A, Ruschendorf F, et al. (2007) Genome-wide linkage analysis of malaria infection intensity and mild disease. PLoS Genet 3: e48.
16. Sakuntabhai A, Ndiaye R, Casadémont I, Peerapittayamonkol C, Rogier C, et al. (2008) Genetic determination and linkage mapping of *Plasmodium falciparum* malaria related traits in Senegal. PLoS ONE 3: e2000.
17. Haldane JB (1949) The association of characters as a result of inbreeding and linkage. Ann Eugen 15: 15–23.
18. Allison AC (1954) Protection afforded by sickle-cell trait against subtertian malarial infection. Br Med J 1: 290–294.
19. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, et al. (2009) Wellcome Trust Case Control Consortium; Malaria Genomic Epidemiology Network. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet 41(6): 657–665.
20. Cooke GS, Hill AV (2001) Genetics of susceptibility to human infectious disease. Nat Rev Genet 2(12): 967–977.
21. Ntoumi F, Kwiatkowski DP, Diakité M, Mutabingwa TK, Duffy PE (2007) New interventions for malaria: mining the human and parasite genomes. Am J Trop Med Hyg 77(6 Suppl): 270–275.
22. Babiker HA, Walliker D (1997) Current views on the population structure of *Plasmodium falciparum*: Implications for control. Parasitol Today 13(7): 262–267.
23. Grech K, Watt K, Read AF (2006) Host-parasite interactions for virulence and resistance in a malaria model system. J Evol Biol 19(5): 1620–30.
24. Garcia A, Cot M, Chippaux JP, Ranque S, Feingold J, et al. (1998) Genetic control of blood infection levels in human malaria: evidence for a complex genetic model. Am J Trop Med Hyg 58: 480–488.
25. Rihet P, Traore Y, Abel L, Aucan C, Traore-Leroux T, et al. (1998) Malaria in humans: *Plasmodium falciparum* blood infection levels are linked to chromosome 5q31-q33. Am J Hum Genet 63: 498–505.
26. Phimpraphi W, Paul R, Witoonpanich B, Turbpaiboon C, Peerapitt-ayamongkol C, et al. (2008) Heritability of *P. falciparum* and *P. vivax* malaria in a Karen population in Thailand. PLoS ONE 3: e3887.
27. Lawaly YR, Sakuntabhai A, Marrama L, Konaté L, Phimpraphi W, et al. (2010) Heritability of the human infectious reservoir of malaria parasites. PLoS ONE 5(6): e11358.
28. McKenzie FE, Smith DL, O'Meara WP, Riley EM (2008) Strain theory of malaria: the first 50 years. Adv Parasitol 66: 1–46.
29. Anderson TJ, Nair S, Nkhoma S, Williams JT, Imwong M, et al. (2010) High heritability of malaria parasite clearance rate indicates a genetic basis for artemisinin resistance in western Cambodia. J Infect Dis 201(9): 1326–30.
30. Nassir E, Abdel-Muhsin AM, Suliaman S, Kenyon F, Kheir A, et al. (2005) Impact of genetic complexity on longevity and gametocytogenesis of *Plasmodium falciparum* during the dry and transmission-free season of eastern Sudan. Int J Parasitol 35: 49–55.
31. Paul REL, Ariey F, Robert V (2003) The evolutionary ecology of *Plasmodium*. Ecology Letters 6: 866–880.
32. Gandon S, Mackinnon MJ, Nee S, Read AF (2001) Imperfect vaccines and the evolution of pathogen virulence. Nature 414(6865): 751–6.
33. Schneider P, Chan BH, Reece SE, Read AF (2008) Does the drug sensitivity of malaria parasites depend on their virulence? Malar J 7: 257.
34. Gouagna LC, Bancone G, Yao F, Yameogo B, Dabiré KR, et al. (2010) Genetic variation in human HBB is associated with *Plasmodium falciparum* transmission. Nat Genet 42: 328–331.
35. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era–concepts and misconceptions. Nat Rev Genet 9(4): 255–266.
36. Trape JF, Rogier C, Konate L, Diagne N, Bouganali H, et al. (1994) The Dielmo project: a longitudinal study of natural malaria infection and the mechanisms of protective immunity in a community living in a holoendemic area of Senegal. Am J Trop Med Hyg 51: 123–137.
37. Rogier C, Tall A, Diagne N, Fontenille D, Spiegel A, et al. (1999) *Plasmodium falciparum* clinical malaria: lessons from longitudinal studies in Senegal. Parassitologia 41: 255–259.
38. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2001) Merlin– rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97–101.
39. Rogier C, Commenges D, Trape JF (1996) Evidence for an age-dependent pyrogenic threshold of *Plasmodium falciparum* parasitemia in highly endemic populations. Am J Trop Med Hyg 54: 613–619.
40. Noranate N, Durand R, Tall A, Marrama L, Spiegel A, et al. (2007) Rapid dissemination of *Plasmodium falciparum* drug resistance despite strictly controlled antimalarial use. PLoS ONE 2(1): e139.
41. SAS (2010) The GLIMMIX Procedure, SAS/STAT User's Guide. SAS 9.1.3. SAS Institute Inc.
42. SAS (2010) The INBREED Procedure, SAS/STAT User's Guide. SAS 9.1.3. SAS Institute Inc.
43. SAS (2010) The MIXED Procedure, SAS/STAT User's Guide. SAS 9.1.3. SAS Institute Inc.
44. Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics. 4th Edn. London: Longman.
45. Duffy PE (2007) *Plasmodium* in the placenta: parasites, parity, protection, prevention and possibly preeclampsia. Parasitology 134(Pt 13): 1877–81.
46. Kruuk LE, Hadfield JD (2007) How to separate genetic and environmental causes of similarity between relatives. J Evol Biol 20(5): 1890–903.
47. Lee SH, Goddard ME, Visscher PM, van der Werf JHJ (2010) Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. Genetics Selection Evolution 42: 22.
48. Mackinnon MJ, Mwangi TW, Snow RW, Marsh K, Williams TN (2005) Heritability of malaria in Africa. PLoS Med 2: e340.
49. Bull PC, Marsh K (2002) The role of antibodies to *Plasmodium falciparum*-infected-erythrocyte surface antigens in naturally acquired immunity to malaria. Trends Microbiol 10(2): 55–8.
50. Graves PM, Carter R, McNeill KM (1984) Gametocyte production in cloned lines of *Plasmodium falciparum*. Am J Trop Med Hyg 33: 1045–1050.
51. Ali E, Mackinnon MJ, Abdel-Muhsin AM, Ahmed S, Walliker D, et al. (2006) Increased density but not prevalence of gametocytes following drug treatment of *Plasmodium falciparum*. Trans R Soc Trop Med Hyg 100(2): 176–83.
52. Barnes KI, White NJ (2005) Population biology and antimalarial resistance: The transmission of antimalarial drug resistance in *Plasmodium falciparum*. Acta Trop 94(3): 230–40.
53. Barnes KI, Little F, Mabuza A, Mngomezulu N, Govere J, et al. (2008) Increased gametocytemia after treatment: an early parasitological indicator of emerging sulfadoxine-pyrimethamine resistance in falciparum malaria. J Infect Dis 197(11): 1605–13.
54. Price RN, Nosten F, Luxemburger C, ter Kuile FO, Paiphun L, et al. (1996) Effects of artemisinin derivatives on malaria transmissibility. Lancet 347: 1654–8.
55. Lawton JH (1988) More time means more variation. Nature 334: 563.
56. Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4(2): e1000008.