Multi-locus Transmission Disequilibrium Test (ML-TDT) for malaria phenotypes from a longitudinal study in family-based design

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 by infections to *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 ≤ 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.