CATÓLICA
UNIVERSIDADE CATÓLICA PORTUGUESA
ESCOLA SUPERIOR DE BIOTECNOLOGIA

"Fine mapping of susceptibility loci to malaria clinical episodes in a family-based cohort from Senegal"

Thesis presented to *Escola Superior de Biotecnologia* of the *Universidade Católica Portuguesa* to fulfill the
requirements of Master of Science degree in Microbiology

By

Alison Machado

October, 2010

CATÓLICA
UNIVERSIDADE CATÓLICA PORTUGUESA
ESCOLA SUPERIOR DE BIOTECNOLOGIA

"Fine mapping of susceptibility loci to malaria clinical episodes in a family-based cohort from Senegal"

Thesis presented to *Escola Superior de Biotecnologia* of the *Universidade Católica Portuguesa* to fulfill the
requirements of Master of Science degree in Microbiology

By

Alison Machado

Under the Supervision of:

Richard Paul of Institut Pasteur- Paris

&

Under the Co-Supervision of:

Jean François Bureau of Institut Pasteur- Paris

Thesis performed in:

Laboratoire de Génétique de la réponse aux infections chez l'homme
Unité de Pathogénie Virale
Institut Pasteur- Paris

October, 2010

I want to dedicate this thesis,

To my parents…because everything I am is due to them.
To all my family and my friends.

”Research is to see what everybody else has seen, and to think what nobody else has thought.”
Albert Szent-Györgi

” Savoir s'étonner à propos est le premier pas fait sur la route de la découverte.”
Louis Pasteur

” La science n'a pas de patrie.”
Louis Pasteur

INSTITUT PASTEUR

Acknowledgements

First, I would like to thank Rick Paul for accepting me as student and giving me the opportunity to learn how to "make" science. Your supervision, ideas, patience and your special endless optimism and enthusiasm towards science, contributed to a very rewarding experience that were these last 10 months.

Several thanks to Jean François Bureau, for receiving me and accepting to co-supervise this project. Thank you for introducing me all the concepts, methods and techniques in lab. Thanks for your guidance, opinions, advices and discussions that contributed to make this life time experience. I will never forget all the help given at all times. You were an excellent "teacher" at the bench and at all in this thesis.

I would also like to thank … to all the members of the GRIH laboratory in Institut Pasteur. Anavaj (Golf), thanks for accepting me as student in your lab, for the English lessons and for teaching me more about Thailand, a beautiful country! One day I will go in Thailand! Isabelle Casadémont, there are no words that can express my gratitude for everything. I will never forget all the help that you give me. Cheikh Loucoubar my colleague and friend, thanks for tutoring file, for statistical file and for the known of your culture and country. I sincerely think that we have *one of kind group* of people that is rare to find. I do not take you for granted, so therefore I will always cherish these months and hope to continue to do so with you nearby. Linda Duval thanks for your sympathy, good humor and fantastic chocolate cake!

I also would like to acknowledge Professor Célia Manaia for helping me finding the lab in Pasteur and for co-supervision in Escola Superior de Biotecnologia of University Católica Portuguesa. To Professor José António Couto for all patience with the administrative information's, thank you.

Thanks all the Portuguese friends that I known in Pasteur, special thanks to Ana Mónica Pais and Ana Margarida Almeida. Ana Margarida, there are no words for all that you did for me. Thanks, for your kindness and friendship, for all the support you gave me, to listening to me and for my birthday party. Thank you for the strength that you ever gave me, the company and attention. Never forget my friend of heart.

Thanks to all my family, in special to my mother and my father for the strength and support all the time! My parents are always in my mint and my heart. Thanks a lot!! José Basto, thank you so much for your support all the time. It was very difficult for we are far, but you get ever a good and affectionate word to me. It represented a lot to me, thank you to all persons that help me to realize this thesis. Thanks a lot for everybody. Muito. Obrigado!!

Resumo

O parasita da malária, *P. falciparum*, mata na ordem de um milhão de crianças Africanas em cada ano, e esta é uma pequena fracção do número de pessoas infectadas em todo o mundo. A evolução clínica de uma infecção por este parasita depende em certa medida, da constituição genética do indivíduo infectado. O papel dos factores genéticos que regulam a gravidade da infecção da malária tem sido repetidamente demonstrado em humanos e animais. Os estudos de associação são realizados com o objectivo de identificar os genes implicados na causalidade do resultado da infecção.

Foi detectado anteriormente, linkage no cromossoma humano 5p15 ao número de ataques de *Plasmodium falciparum* (PFA) em Dielmo, uma aldeia senegalesa [48]. Posteriormente, e antes deste estudo, um levantamento usando um ensaio "GoldenGate" da Illumina, com cerca de 1.450 SNPs foi realizada na região de Linkage com o fenótipo PFA. A análise foi realizada com três programas estatísticos baseados na família: Merlin, QTDT e FBAT/PBAT. Estes programas identificaram três genes candidatos associados com o fenótipo PFA: três SNPs (rs4867417, rs7714218 e rs11959398), localizados no gene PDZD2, um SNP (rs11134099) no gene ADAMTS16, e outro (rs3777320) localizado no gene SEMA5A.

O objectivo deste estudo foi investigar estas associações. Os SNPs das regiões destes genes candidatos foram escolhidos por sequenciação de exões situados na região candidata ou por análise bioinformática utilizando dados do HapMap da população Yoruba. O estudado para genotipagem foi através das análises de pré-design ou “Custom” dos SNPs (Applied Biosystems). Os dados foram incluídos num banco de dados e a verificação dos erros de transmissão mendeliana foi efectuada. As análises estatísticas foram realizadas utilizando dois programas de associação familiar, PBAT e QTDT. Foram utilizados diferentes modelos de transmissão de alelos e foi definido como limite de significância p-value = 10^{-3}. As análises de SNPs dos genes PDZD2 e ADAMTS16 não confirmaram a associação, mas encontrou-se associação significativa com SNPs do gene SEMA5A. Um SNP (rs3777325) foi significativamente associado com o fenótipo PFA usando ambos os programas (p-value= -$6.49\text{x}10^{-4}$ usando o programa PBAT e p-value = $2.0\text{x}10^{-4}$ usando o programa QTDT). A análise de haplótipos de dois SNPs adjacentes (rs4541632 e rs1018956), também mostrou uma associação significativa do haplótipo GC (p-value= -$6.82\text{x}10^{-5}$) utilizando o programa PBAT. Este estudo confirma que o locus de susceptibilidade para o fenótipo PFA está localizado no gene SEMA5A. Mais estudos serão necessários para replicar essa associação e identificar o polimorfismo causal.

Abstract

The malaria parasite, *P. falciparum*, kills on the order of a million African children each year, and this is a small fraction of the number of infected individuals world-wide. The clinical outcome of an infection by this parasite depends to some extent on the genetic make-up of the infected individual. The role of genetic factors that regulate the severity of malaria infection has been repeatedly demonstrated in humans and animals. Association studies are conducted with the aim of identifying the causal genes implicated in the outcome of infection.

Linkage was previously detected on human chromosome 5p15 controlling the number of *Plasmodium falciparum* attacks (PFA) in Dielmo, a Senegalese village [48]. Subsequently, and prior to this present study, a fine mapping study using a "GoldenGate assay" from Illumina, with about 1450 SNPs was performed in this region of linkage with PFA phenotype. Analysis was performed with three statistical family-based programs: Merlin, QTDT, and FBAT/PBAT. These programs identified three candidate genes associated with PFA phenotype: three SNPs (rs4867417, rs7714218, and rs11959398) located in *PDZD2*, one SNP (rs11134099) in *ADAMTS16*, and one (rs3777320) in *SEMA5A*.

The aim of this present study was to investigate these associations. Novel SNPs in the candidate regions of these genes were selected either by sequencing exons located in these candidate regions or by bioinformatics analysis using HapMap data from Yoruba population. SNPs were studied using either Pre-design or Custom SNP genotyping assay (Applied Biosystems). Data were included in an Access Database and checked for error of Mendelian transmission. Statistical analyses were performed using two family-based association programs, PBAT and QTDT. We used different models of allele transmission and defined $p=10^{-3}$ as significance threshold. The analyses did not confirm the association with SNPs of *PDZD2* or *ADAMTS16*, but did find significant association with SNPs of *SEMA5A*. One SNP (rs3777325) was significantly associated with PFA phenotype using both programs (p-value= $-6.49\text{x}10^{-4}$ using the PBAT program and p-value=$2.0\text{x}10^{-4}$ using the QTDT program). A haplotype analysis of two adjacent SNPs (rs4541632 and rs1018956) also showed a significant association of the haplotype GC (p-value= $-6.82\text{x}10^{-5}$) using the PBAT program. This work confirms that a susceptibility locus to PFA phenotype is located inside *SEMA5A*. Further studies will be necessary to replicate this association and identify the causal polymorphism.

List of Abbreviations

ADAMTS	A Disintegrin and Metalloproteinase with Thrombospondin motifs
bp	Base pairs
Chr	Chromossome
cM	Centi Morgan
CRD	Cysteine-rich domain
dbSNP	Data base Single Nucléotide Polimorphism
DNA	DeoxyriboNucleic Acid
dNTPs	Deoxynucleotide Triphosphates
EIR	Entomological inoculation rate
EST´s	Expressed Sequence Tag
FBAT	Family-based association tests
FS	Full-sibships
GEE	generalized estimating equations
GWAS	Genome wide association study
H0	Null Hypothesis
H_2O	Water
HS	half-sibships
IRD	Institut de l aRecherche pour le développement
LD	Linkage Disequilibrium
LDL	Lipoprotein
me-PFD	mean *P. falciparum* trophozoite parasite densities
Merlin	Multipoint engine for rapid likelihood inference
MGB	Minor Groove Brinder
MS	mixed full- and half-sibships
mx-PFD	maximum *P. falciparum* trophozoite parasite densities
OC	only-child
PBAT	Power calculations for family-based association tests
PCR	Polymerase chain reaction
PDZD2	*PDZ* domain containing protein 2
PFA	*P. falciparum* clinical attacks
PFA	Plasmodium Falciparum…
Pro-IL-16	Pro-interleukin-16
PSI	Plexin-Semaphorin integrin
PtPF	Prevalence of asymptomatic Plasmodium falciparum infections
QTDT	Quantitative Transmission Disequilibrium Test
rpm	Rotations per minutes
SEMA	Semaphorin
SNP	Single Nucleotide Polymorphism
TDT	transmission disequilibrium test (TDT)
Tm	Temperature Meeting
TSR	Thrombospondin repeats
WHO	World Health Organization
Yri	Yoruba
µl	Microlitre

Contents of Figures

Contents of Tables

General Contents

1- Introduction

1.1- General Introduction

Malaria is one of the most widespread and devastating infectious diseases in the world, [44].In 2008, there were an estimated 243 million cases of malaria worldwide, the vast majority of which (85%) were in Africa and led to 863 000 deaths [64]. The disease is caused by parasites of the genus *Plasmodium* belonging to the apicomplexan phylum, which invade and reproduce in human erythrocytes. The parasites are then transmitted to other humans by hematophagous mosquitoes of the genus *Anopheles*. In Africa, malaria is endemic, constituting a serious public health problem. The main factors for the disease being prevalent in Africa are: 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.

Many projects and studies have been conducted to try to develop a vaccine and to eradicate malaria, but with variable efficiency [55]. In 1989 the Institut Pasteur Paris, the Institut Pasteur Dakar in Senegal, the Institut de Recherche pour le Développement (IRD) and the Senegalese Government officials joined forces to develop a multidisciplinary program in two village-based cohorts in Senegal, where *Plasmodium falciparum*, the etiological agent of lethal human malaria, predominates. This program aimed to improve of our understanding of the acquisition of immunity (non-sterilising) in an area of endemic malaria in order to develop new anti-malaria strategies and particularly vaccines.

The longitudinal study was initiated in 1990 for Dielmo village and in 1993 for Ndiop Village. Malaria transmission intensity differs considerably in these two Villages that are only five kilometers apart. In Dielmo, the population is exposed to perennial and intense transmission, whereas in Ndiop transmission is seasonal and only occurs during the four months of the rainy season [45]. Such detailed long-term longitudinal data enable very precise characterization of malaria epidemiology and individual human variation in the response to infection.

The human host response to infection with *P. falciparum* varies according to environmental and genetic conditions. While some infected individuals die of severe malaria, others survive, and still others are infected without becoming severely ill. It has been estimated that host genetic factors account for approximately 25% of the risk of severe malaria. Indeed, malaria has exerted significant selective pressure on the human genome resulting in the spread of genetic mutations that confer refractoriness to

malaria. This is illustrated by several well studied mutations that protect against different forms of malaria, including sickle cell allele, Duffy group, and alpha thalassaemia [70]. Moreover, the past decade has seen growing evidence of ethnic differences in susceptibility to malaria and of the diverse genetic adaptations to malaria that have arisen in different populations. Differences in susceptibility to malaria have been observed between sympatric populations that share the same environment but suffer different levels of malaria infection and disease burden, strongly suggesting a role for human genetics in determining the outcome of infection. Such population differences in susceptibility to malaria are becoming more amenable to study since the development of high thru-put genetic technology, thereby allowing us to genetically dissect the outcome of infection. The fact that different malaria-resistance alleles have arisen in different places suggests that a great deal of selection by malaria has happened relatively recently in human history and certainly since human migration out of Africa [13].

Genetic association studies of malaria have shown that variation in both innate and adaptive immune pathways condition disease susceptibility and outcome [40]. Association studies allow identification of genes and their allelic variants involved in susceptibility to disease. They are indispensable for identifying susceptibility genes after candidate chromosomal regions have been revealed by genetic linkage study. The basic method of study compares the allele frequency of a genetic marker from sick individuals and control individuals (case-control studies), chosen randomly from a population and who differ only in the phenotype studied. The marker used may be a polymorphism without causal relationship to the phenotype or a mutation in a gene candidate. A positive result suggests that the marker studied is involved either directly or by virtue of being linked to the causal gene (i.e. the marker is in linkage disequilibrium with the causal gene whereby marker and causal alleles co-occur more frequently than they would by chance). The major problem with case-control studies is the possibility of false positive results due to differences in environmental factors that influence the development or the evolution of the disease being studied. Association studies based on families with at least one affected child can overcome this problem.

In a previous study [48], a genome-scan linkage analysis of several malaria-related phenotypes in the Dielmo and Ndiop cohorts identified several susceptibility loci. The authors showed evidence for a strong genetic contribution to the number of clinical *P. falciparum* attacks (PFA) with linkage on chromosome 5p15 (LOD= 2.57

and empirical p=0.001) in Dielmo village. A fine mapping in this candidate region (35 cM) was performed using 1,300 SNPs. Candidate SNPs and genes were defined by using association between the PFA phenotype and these SNPs.

Hence, in our work, we use family-based tests derived from the original transmission disequilibrium test (TDT) and adapted to the study of quantitative trait as proposed by Abecasis et al, in the program QTDT or by Laid and Lange in the program FBAT/PBAT. We performed association studies to find the susceptibility gene(s) associated with the PFA phenotype in both the village cohorts.

1.2- *Plasmodium* parasites

1.2.1- Life Cycle

Anopheles mosquitoes are required for the transmission of the parasite from one human host to another. During a blood meal on an infected individual, the mosquito may ingest sexual gametocyte parasite stages that are essential for successful passage from human to mosquito. Once ingested these gametocytes will male or female gametes. The fertilization of a female gamete by a male takes place within a few minutes and 18 hours later a zygote mobile, the ookinete, is formed. This ookinete traverses the wall of the stomach of the mosquito and develops on its outer surface into an oocyst. Ten to fifteen days later, the oocyst releases thousands of sporozoites that invade the salivary glands of the mosquito. These sporozoites are then injected into the human host during a blood meal, and migrate to the liver wherein they asexually multiply for a week before releasing many thousands of merozoite stage parasites into the bloodstream, which then invade erythrocytes. Within the erythrocyte the parasite grows into a trophozoite, which will either mitotically replicate forming a schizont parasite stage or produce a single gametocyte. Upon rupture of the erythrocyte, the schizont releases 8 merozoites. The erythrocytic cycle from invasion to rupture takes 48 hours for *P. falciparum*. Malaria symptoms and disease occur during the blood-stage part of the lifecycle (the erythrocytic cycle).

Figure 1 show this complex life cycle:

- Within the female *Anopheles*, the mosquito vector and definitive host of the parasite, where the sexual phase of the cycle occurs.

- Within humans, the intermediate host in which the asexual phase occurs, itself divided into two phases: (i) the hepatic or pre-erythrocytic phase corresponding to the

clinically asymptomatic incubation phase; (ii) the erythrocytic phase corresponding to the clinical stage of the disease.

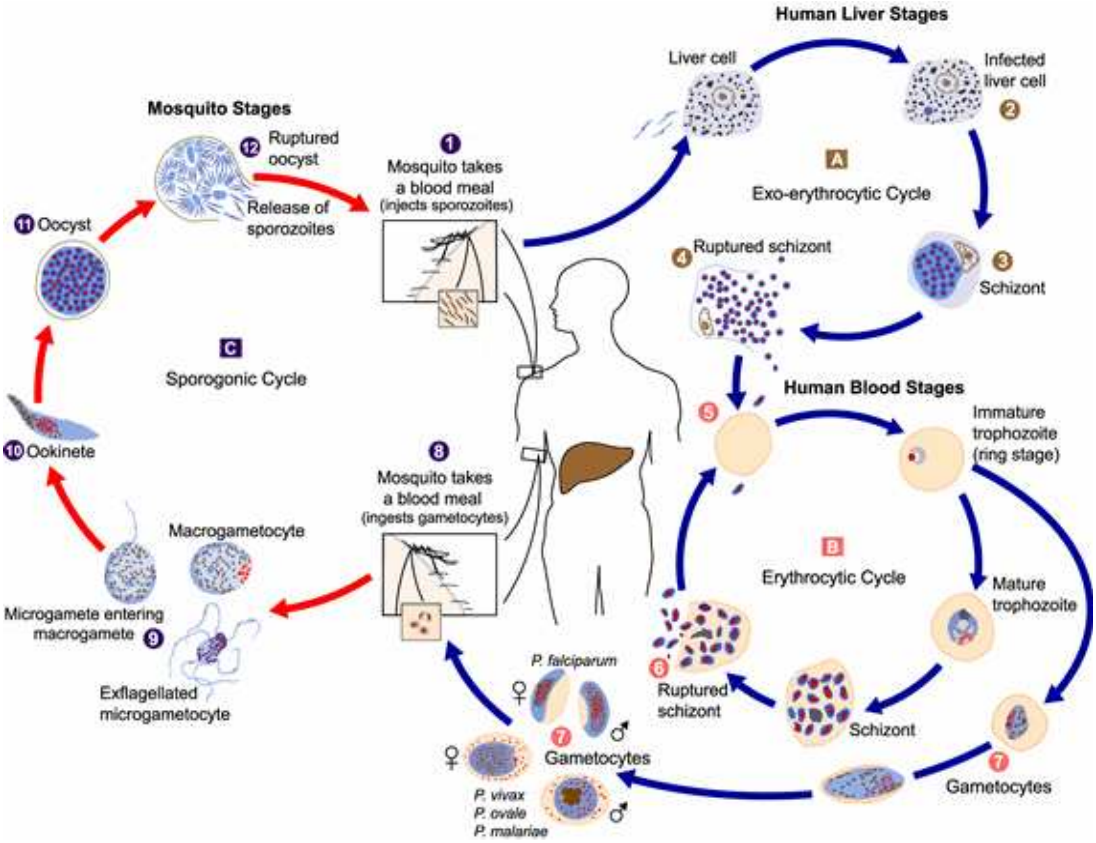


Figure 1- Diagram depicting the life cycle of human malaria, [42].

1.2.2- Epidemiology of Malaria

1.2.2.1- Distribution Worldwide

Malaria is concentrated in tropical and subtropical climates (Figure 2). Indeed, the development of *Plasmodium* in the mosquito vector requires temperature above 18 ºC, which has limited the expansion of malaria in temperate regions, [49]. Vector control developed during the 20th century led to eradication of the disease in temperate regions but failed in tropical areas.

In 2008, there were an estimated 243 million cases of malaria (5% and 95% centiles, 190–311 million) worldwide. The vast majority of cases (85%) were in the African region, followed by the South-East Asia (10%) and Eastern Mediterranean Regions (4%). The totals are similar to those reported in the *World Malaria Report 2008* [64].

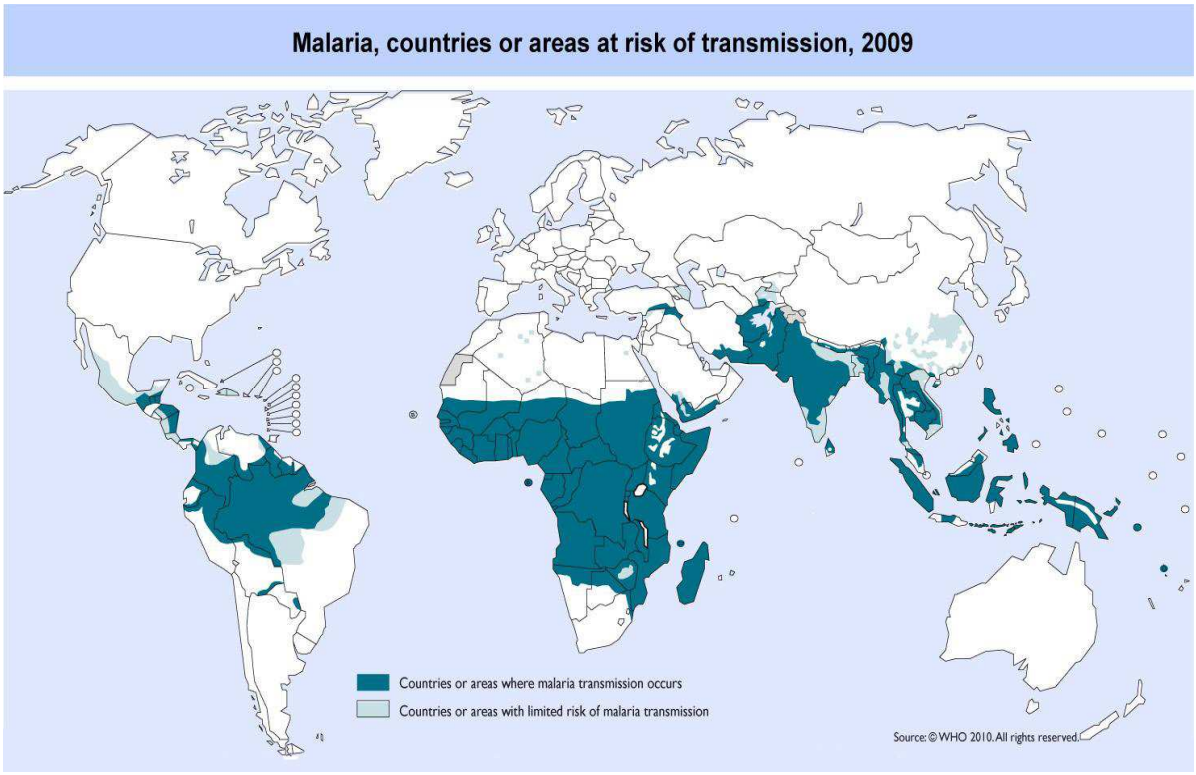


Figure 2- World distribution of malaria in 2009 by countries at risk of transmission, [65].

In endemic areas, the level of endemicity of malaria varies. A classification of the level of endemicity based on the prevalence of the parasite defines: hypoendemic (areas in which childhood infection prevalence is less than 10%), mesoendemic (areas with infection prevalence between 11% and 50%), hyperendemic and holoendemic (areas with an infection prevalence of more than 50%).

1.2.2.2- Environmental Factors

The prevalence of malaria in a community depends on entomological inoculation rate (EIR), which is the number of infectious bites received per individual in the community in a given time. The incidence of clinical cases depends on the rate of inoculation and is modulated by the development of premonition, which develops more rapidly with increasing transmission intensity. In areas of high transmission (hyper- and holoendemic), the estimated EIR is 150-300 infective bites per person per year [36]. In such areas malaria is considered stable. Clinical episodes occur mainly during early childhood; the development of non-sterilising immunity protects older individuals. In areas of low transmission, few infectious bites are received annually, immunity develops more slowly and clinical episodes occur at all ages.

The level of malaria transmission is measured by the entomological inoculation rate (H) per unit time (day, month, and year).

$$H=ma.S$$

S = sporozoite rate is the percentage of mosquitoes with sporozoites in their salivary glands.

m = the number of mosquitoes per person

a= individual mosquito biting rate (a mosquito bites an estimated once every 2 to 3 days)

ma depends on three elements:

1 - The number of potential mosquito breeding sites, whether natural (ponds, lakes...) or anthropogenic (pits, irrigation dams, boreholes).

2 - The mosquito breeding productivity: it depends first of the water supply, through rain or supply of irrigated areas. For species living in running water, outbreaks occur when the level of rivers in dry season creates residual pools that are very productive. The larval ecology determines the rate of occurrence of malaria. In the savannahs of West Africa the main vectors, *Anopheles funestus* and *Anopheles gambia*, live on rainwater collection, thus transmission will occur during the rainy season and at this period most cases were observed.

3 - Exposure of man to the mosquito: 2500 years ago Hippocrates advised to build the villages away from the marshes to avoid fevers. This advice is still valid except that in tropical countries it should be extended to include the modern extensive development of irrigated land for crops. In Africa, the cultivation of valley bottoms was accompanied by the creation of villages near the scene of culture and that it is in these areas where most of the transmission occurs. Transmission is exacerbated by dilapidated houses that are dark and poorly ventilated, well suited to these vectors.

Malaria is very complex since the epidemiology of the disease is modulated by the ecology of vectors, very different from one region to another, but also by other factors such as the socio-economic status of the areas concerned. Indeed, the current distribution of malaria is correlated with the level of development of endemic areas [20]. These are the underdeveloped countries, mainly located in inter-tropical areas that suffer the heaviest toll of the disease. Despite international aid, malaria control must necessarily pass through a government policy of improving living conditions of

populations in these countries. Apart from mass chemoprophylaxis of populations, prevention can be achieved through reducing transmission.

1.2.2.3- Factors related to parasite

Molecular epidemiology studies have highlighted the diversity of *P. falciparum*. This diversity is the source of some of the variability in the outcome of the infection in humans and is one of the main obstacles to the development of immunity as well as anti-malarial vaccines and drugs. The parasite has a tremendous capacity to evolve. *Plasmodium* is haploid throughout its life cycle in the human host and the number of parasites generated by mitosis during the course of a single infection is vast. This provides the basis for easy selection of mutants and consequent adaptation of the parasite.

1.2.2.4- The human genetic factors

The methods of genetic epidemiology of single gene disorders have been applied to the analysis of multifactorial diseases. These methods seek to isolate among all risk factors for disease, those which have a genetic basis. Two approaches are generally used.

1 - The full gene-candidate approach in a general population that studies the comparative distribution of a marker in a group of patients and a group of controls.
2 – Linkage analysis performed in families and studying the joint allelic transmission among siblings of a phenotype of interest with a given genetic marker.

In the case of malaria, several genetic factors have been identified as being involved in the susceptibility/resistance to malaria and the severity of infection mainly by a candidate gene approach.

1.2.3- Populations and Phenotypes

Senegal covers an area of 196,200 km2 and has a population of 11.7 million people. Various scientific establishments in Senegal including University of Cheikh Anta Diop, Dakar, Institut Pasteur de Dakar and Institut de la Recherche pour le Développement (IRD), are studying the epidemiology of malaria.

The expansion of farming land and the effects of drought have largely fashioned the local countryside. Completely different conditions can pertain at sites within a few kilometers of one another. For example, at Ndiop, a small village cited by Fontenille et

al., [15] located in the Sahelo-Suudanese zone where malaria is meso-endemic and seasonal, is quite different from the situation at Dielmo, a village on a permanent flowing river where the disease is holo-endemic and perennial. The inhabitants of Dielmo and Ndiop villages are settled agricultural workers. Small herds of domestic animals live in close contact with the houses. The ground is sandy and the original wooded savannah has been almost entirely clears for cultivation.

Since 1990 a longitudinal epidemiologic and entomologic follow-up has been carried out in Dielmo Village and since 1993 in Ndiop village [10]. The protocol to study Dielmo and Ndiop villages was approved by the Ethical Committee of the Institut Pasteur de Dakar and the Ministére de la Santé of Senegal. An agreement between Institut Pasteur de Dakar, Institut de Recherche pour le Développement (IRD) and the Ministére de la Santé et de la Prévention of Senegal defines all research activities in these 2 villages. 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 for all subjects.

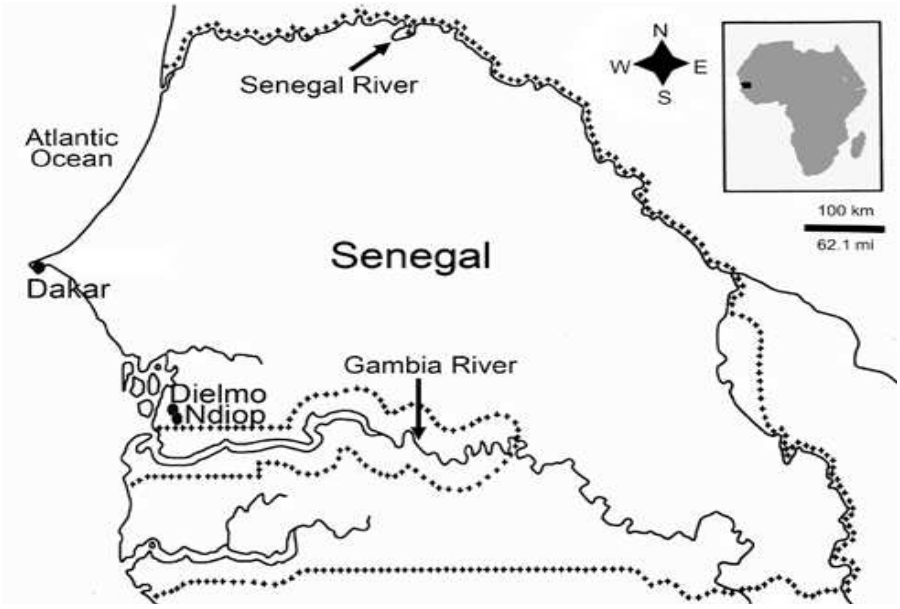


Figure 3- Location of Dielmo and Ndiop in Senegal, Africa. Plus symbol lines define the Senegal frontiers, [14].

1.2.3.1- Dielmo Village

Dielmo village is located about 280 km southeast of Dakar and 15 km north of the Gambia. Malaria transmission is intense and perennial. This is due to the presence

of a small permanent river, the Nema, which allows the persistence of mosquito breeding throughout the year. This leads to high P. *falciparum* prevalence rates of infection in Dielmo village. The village was founded in 1914 and the inhabitants are sedentary farmers. The estimated population is 585 residents consisting of 79% Serere (59% Niominka and 20% Sine/Baol), 11% Mandinka and 10% miscellaneous [48].

Dielmo village is composed of large complex families, with many small family units, some of which, because of multiple marriages, include several half-sibling relationships. The small family units contain either (a) full-sibships (FS) only, (b) mixed full- and half-sibships (MS), (c) half-sibships (HS) only or (d) only-child (OC) family. Thus, for the large complex family in Dielmo village, there were 18 FS, 12 MS, 10 HS units and 52 OC, all of which were linked because at least one of the unit members was a 1st cousin of someone else in this large complex family. There were 2 OC that were linked to the large complex family through a 2nd cousin relationship and 2 FS units and 8 OC that were linked because they were uncles or aunts, [48].

1.2.3.2- Ndiop Village

Ndiop village is 5km from Dielmo. Malaria transmission is meso-endemic, as found almost everywhere else in the region, strictly seasonal and dependent upon the rainy season that occurs from July-September [46]. Thereby, the prevalence of *P. falciparum*, changes from 20% in the dry season to 70% in the rainy season. In Ndiop village, the average incidence of malarial infections was 3 per year per person with a peak incidence in children of between 3 and 6 years of age. Only with adults, the incidence drops below 1, reflecting the development of some protective immunity [53].

The population is composed of 664 individuals associated in 19 large families with 83 nuclear families. Approximately 546 individuals form a very large single family N1. The ethnic groups in Ndiop village are very different from those of Dielmo, consisting of 76% Wolof, 19% Fulani and 5% miscellaneous. Large complex families exist in Ndiop as well as in Dielmo. Following the same classification than the Dielmo village, there were 20 FS, 15 MS, 6 HS units and 24 OC that were linked because at least one of the members was a 1st cousin of someone else in this large complex family. There were 1 OC that was linked to the family through a 2^{nd} cousin relationship and 20 OC that were linked because they were uncles or aunts. In the MS units, the number of

FS units varies. There are more FS units in MS units in Ndiop than Dielmo, which thus gave a higher number of half sib-pair counts in Ndiop, [48].

1.2.3.3- PFA Phenotype

Sakuntabhai et al [48] defined phenotypes with significant genetic contribution in a family-based genome-wide linkage study of Dielmo and Ndiop villages. The malaria-related phenotypes considered by the authors were:

1. The number of clinical episodes of *P. falciparum* (acronym PFA). This phenotype characterizes the individual tendency to become clinically ill following *P. falciparum* infection. The installation of health clinics in each of the study sites enabled passive case detection of malaria episodes. Clinical malaria episodes were defined as measured fever (axillary temperature > 37.5ºC) 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, ii) a *P. falciparum* parasite/leukocyte ratio higher than 0.3 parasite/leukocyte in Ndiop. All positive malaria cases were treated with appropriate antimalarial treatment according to the recommendation of the Malaria Division, Ministry of Public Health, namely quinine until 1995 and then chloroquine in Dielmo and Ndiop for *P. falciparum*, [31];

2- The maximum and mean *P. falciparum* trophozoite parasite densities (i.e. pathogenic parasite stage infecting red blood cells) during clinical episodes (acronym mx-PFD and me-PFD, respectively). Maximum density was chosen to assess whether there were individual human differences in tolerance of parasite burden prior to onset of clinical symptoms.

3- The prevalence of asymptomatic *P. falciparum* infections, which reflects the acquisition of clinical immunity and/or the tolerance of parasitic infection (acronym PtPF).

Thus, the authors [46] defined quantitative phenotypes and also qualitative phenotypes not defined in this report. For most multi-factor diseases, the quantitative phenotypes are more informative than the qualitative phenotypes. These quantitative phenotypes were corrected for well known factors affecting the trait such as time spent in the village, age of the individual, quarter and year of the event. The residual "unexplained" phenotypes were then used to perform the genetic analysis. PFA

phenotype was linked to microsatellite markers of chromosome 5p15 in Dielmo village only (data shown in Figure 4).

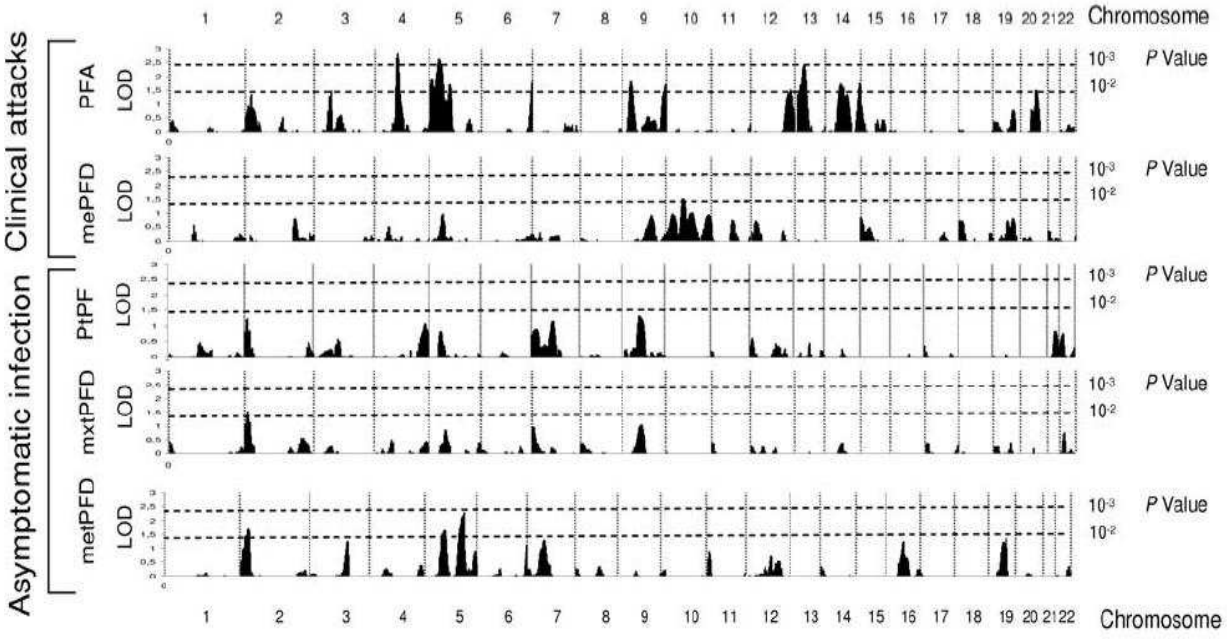


Figure 4- Results of the Genome scan linkage analysis in Dielmo village from Sakuntabhai et al [48].

1.3- Generalities on genetic studies

1.3.1- Association studies

Association studies are used for identifying genes and their common allelic variants involved in predisposition to a disease. Such studies are performed after localization of susceptible loci by linkage analysis. This method compares the allele frequency of a genetic marker of affected and non-affected individuals, chosen at random in a population (case-control study). The marker might be the causal polymorphism or any polymorphism in linkage disequilibrium (LD) with the causal one. A positive association with one marker suggests that this marker is in LD with the causal polymorphism. The LD between two markers is defined by the existence of a combination of alleles of these markers more often than expected by chance.

The choice of the control population is one of the most important problems of case-control study: if the control group are not from the same population as the affected individuals, uncontrolled environmental factors or population stratification might induce false positive association.

Association studies are the most widely used contemporary approach to relate genetic variation to phenotypic diversity. This is due to their higher power and lower cost to detect a susceptibility locus than linkage analysis. Over the past 2 years these studies have identified statistical association between hundreds of loci across the genome and common complex traits. The results of these studies have substantially increased our understanding of the diverse molecular pathways underlying specific human diseases, [16].

1.3.1.1- Powerful Single Nucleotide Polymorphism analysis

The identification of Single Nucleotide Polymorphsism (SNPs) that are associated with the risk for developing complex diseases is an important goal of modern genetics studies. The standard approach for analyzing Genome Wide Association Study (GWAS) in the discovery phase involves individual-SNP analysis. This mode of analysis often involves regressing the phenotype onto each individual typed SNP and generating a parametric p value. The SNPs are then ranked on the basis of their individual p values, and a threshold is set such that all SNPs with a p value less than that threshold will be pushed forward for validation. The threshold can be based on reaching a multiple-comparison-adjusted significance level or established arbitrarily. The level at which each statistical test is conducted must be adjusted. Because of the large number of considered hypotheses, the threshold for genome-wide significance is very high and difficult to obtain ($p \approx 10^{-7}$). Replication in another population is frequently difficult Replication in another population is frequently difficult and causal polymorphisms often have a low effect, increasing, for example, the risk of developing the disease by less than 5-10% [66]. Identifying such SNPs is important and their elimination from being considered as candidate SNPs because of correction for multiple testing may be erroneous. Thus, thresholds of rejection or acceptance must be considered on a case by case basis and in the context of both the genotype and phenotype data in hand.

1.3.1.2- Linkage Disequilibrium and Haplotypes Analysis

A haplotype is the combination of the alleles of many markers from a chromosomal region. The segregation of ancestral haplotypes in a population is accompanied by a reduction in its size due to the occurrence of recombination. However, the recombination between genetic markers separated by a short distance is a very rare event. Thus, individuals who bear susceptible polymorphisms to a disease

from a common ancestral haplotype are supposed to always share a portion of that haplotype. Markers located within that region will be in LD.

1.3.2- Statistical analysis

1.3.2.1- Fine Mapping

A genetic fine map of a specific locus will usually have as its goal the identification and location of markers that flank the targeted region of genome with replicable signals of association, with the aim of localizing the causal variants, [34]. In most cases, markers with positive association and causal polymorphisms are less than one centiMorgan (cM) apart. In some cases, comparisons of map positions can be accurately made with results of other species, examining syntenic regions for defined causal polymorphisms and for similar traits. Thus, fine-mapping is a necessary step to identify causal polymorphisms.

Broadly speaking, fine mapping involves systematic resequencing of the genomic region of interest to identify all common variants, which are then tested for disease association using the largest possible sample size. One of the main problems in regions with high LD is to distinguish causal variants from neighbouring non-functional variants. This has led to growing interest in trans-ethnic studies, which aim to increase the resolution of fine mapping. Studies in Africa could be of particular value because of the low levels of LD found in populations of that continent and because different populations in Africa have different patterns of LD [56].

A great number of strategies and programs have been presently developed to analyze different kinds of data. In our case, we will focus on family-based approaches and statistical programs, such as Merlin, PBAT/FBAT and QTDT programs.

1.3.2.2- Multipoint engine for rapid likelihood inference- Merlin

MERLIN tests for association between SNPs and the quantitative trait in a region previously found to be linked to that phenotype using a maximum likelihood approach. The association test implemented in MERLIN includes an integrated genotype inference feature, which can improve power when some genotypes are missing [62]. MERLIN can construct approximate solutions for dense maps where the probability of observing several recombinants between consecutive markers is close to zero by restricting analysis to gene flow patterns separated by a few recombination

events or less [62]. However, it is important to note that -- in contrast to standard family-based association tests -- the test implemented in MERLIN does not control for population stratification. If population stratification is a concern, population membership should be included as a covariate, [3].

MERLIN estimates haplotypes by finding the most likely path of gene flow or by sampling paths of gene flow at all markers jointly. It can also list all possible nonrecombinant haplotypes within short regions.

1.3.2.3- Power calculations for family-based association tests/ Family-based association tests- PBAT/FBAT

The PBAT software is a powerful tool for complex family-based association analysis. One major interest of this program is that a variety of scenarios can be computed: dichotomous and continuous traits; missing parental information; multiple offspring per family; combinations of different family-types; different genetic models and haplotype analysis, [61].

The data-analysis functions include univariant and multivariant tests for various trait types, procedures for effect-size estimation, and screening techniques to select the most "promising" combinations of markers and phenotypes. All P values can be computed on the basis of both asymptotic theory and permutation tests for covariates. PBAT software provides functions that assess the power of the observed data set, computes the most powerful test statistic, estimates the genetic-effect sizes in different ways, and provides screening techniques/testing strategies that select the optimal combinations of markers and phenotypes for testing. Most of these tests are not available in other programs, which are an advantage for our association studies [61].

PBAT program is based on two key components: the general approach of family-based association tests developed by Lange and Laird [28;30] and the conditional-mean–model approach for family-based association tests (FBATs) [27;29]. The first approach is for nuclear families and extended pedigrees; this function computes the distribution of any FBAT statistic under the null hypotheses (H0) is no association with linkage and alternative hypotheses is association no linkage and [43]. The second approach estimates all parameters of the conditional mean model for FBATs [27;29], using generalized estimating equations (GEE) [69] without biasing the significance level of any FBAT statistic that is computed subsequently. GEE enable

repeated measures from an individual to be analyzed without the bias of pseudo-replication.

PBAT's interactive design allows the user to test different designs, ascertainment conditions, and underlying different genetic model/mode of inheritance [28;30]. All power calculations can be verified by Monte-Carlo simulations.

1.3.2.4- Quantitative Transmission Disequilibrium Test- QTDT

QTDT is a statistical program for family–based association test. QTDT is derived from the transmission disequilibrium test (TDT) initially described by Spielman et al [51]. TDT is a family-based linkage-disequilibrium test that offers a powerful way to test for linkage between a dichotomous phenotype and markers that are either causal (i.e. the marker locus is the disease/trait locus) or in linkage disequilibrium with the causal marker [67]. This principle is robust to the confounding factor of population stratification/admixture that potentially plagues ordinary association tests [50].

Allison extended TDT to quantitative traits [6]. But is the approach proposed by Abecasis et al in 2000 [2] that we were used, because your approach was based on large pedigrees rather than on collections of (usually) unrelated nuclear families. QTDT is based solely on the trait values for informative non-founders [19]. QTDT accommodates data not only from parents and siblings, but also from all available relatives. This test is also robust to population stratification. However, when population stratification is absent, it is possible to utilize even more information, namely the additional information contained in the founder genotypes.

1.4- Candidates genes

1.4.1- *PDZD2* gene

PDZD2 (*PDZ* domain containing protein 2) also known as *PAPIN, AIPC and PIN1*, is a multi-*PDZ* protein of unknown function, [33]. *PIN1/PAPIN/AIPC* is ubiquitously expressed. *PIN1* was isolated from the rat insulinoma cell line INS-1 [57] as a protein that interacted with the basic helix–loop–helix transcription factor E12. Independently, full-length rat and human *PIN1* complementary DNAs were cloned and named *PAPIN* (for plakophilin-related Armadillo-repeat-protein interacting *PDZ* protein [12] and AIPC (for activated in prostate cancer [11], respectively, on the basis of the

protein's binding and expression properties. Recently, the human gene was mapped to 5p13.2 and named *PDZD2* by the Human Genome Nomenclature Committee.

The *PDZ* domain is one of the most common modular protein-interaction domains. The primary function of these domains is to recognize specific ~5'-residue motifs that occur at the C-terminus of target proteins or structurally related internal motifs [18].

PDZD2 shows extensive homology to pro-interleukin-16 (pro-IL-16) and is localized mainly to the endoplasmic reticulum. Pro-IL-16, the encoded protein, is cleaved by a caspase-3-dependent mechanism to generate secreted cytokine IL-16, an atypical cytokine for growth and differentiation. Also, *PDZD2* protein has a secreted peptide generated by the post-translational cleavage at the C-terminus of *PDZD2* by caspase as mentioned above. This gene is up regulated in primary prostate tumors [54] and may be involved in the early stages of prostate tumorigenesis.

1.4.2- *ADAMTS16* gene

The *ADAMTS* (a disintegrin and metalloproteinase with Thrombospondin motifs) family is a group of proteases that are found both in vertebrates and invertebrates [41]. The molecular structure of the *ADAMTS* proteins can be subcategorized into domains, modules and motifs. These enzymes have a complex domain structure consisting of at least a signal peptide, a pro domain, a metalloproteinase domain, a disintegrin domain, thrombospondin type I motifs and cysteine rich domain. The *ADAMTS*s are closely related to the *ADAM* proteinases that are involved in ectodomain shedding or activation of diverse cell surface molecules, including growth factors and adhesion receptors. There is more variability between the different proteins at the C-terminus than at the N-terminus. The ADAMTS proteins are all synthesized initially as inactive pre-proenzymes.

ADAMTS16 gene is a recently described member of the *ADAMTS* gene family [41]. This gene is a member of a family of 19 secreted metalloproteinases, whose function, in contrast to several other family members, is unknown. Few papers have been published on *ADAMTS16* gene and its product. It is expressed in the kidney at a higher level than in other tissues, but expression of *ADAMTS16* mRNA is also high in fetal lung, and adult brain and ovary [52]. *ADAMTS16* gene has recently been

genetically linked to inherited hypertension [23]. Different protein isoforms are coded from *ADAMTS16* through differential splicing of its mRNA [41].

1.4.3- *SEMA5A* gene

The semaphorins are the products of a large family of genes currently containing more than 30 members, all of which share a conserved N-terminal domain called the "sema" domain, a segment of approximately 400–500 amino acids. Based on sequence similarity and distinctive structural features, these genes have been grouped into eight subclasses (1–8) [68]. Class 1 and 2 are present in invertebrates, classes 3-7 are present in vertebrates, and class 8 is found in DNA viruses.

The sema domain shows considerably higher conservation among the different semaphorins and across phyla than do the full-length proteins. In addition to several blocks of conserved amino acids, the sema domain is characterized by highly conserved cysteine residues that have been found to form intrasubunit disulfide bonds. The crystal structure of the sema domain has a β propeller topology with 7 blades. These sema domains have a structure similar to those of β integrins and low-density lipoprotein (LDL) receptors.

As a group, semaphorins are expressed in most tissues and this expression varies considerably with age. Semaphorins are also widely expressed in many organ systems, including cardiovascular, endocrine, gastrointestinal, hepatic, immune, musculoskeletal, renal, reproductive, and respiratory systems. Some members of the semaphorin family were first identified as axonal growth molecules and were later shown to be involved in a variety of functions including cellular migration, immune regulation, apoptosis, angiogenesis, cellular collapse, apoptosis and were associated with autism and cancer. The molecular mechanisms by which semaphorins mediate their functional effects are far from clear. Semaphorin mediated axon repulsion is a result of the modification of the axonal cytoskeleton at the growing tips or growth cones of axons. The control of axon outgrowth or growth-cone motility depends critically upon the dynamics of F-actin polymerization and depolymerization, coupled with the regulation of F-actin translocation and microtubule dynamics.

Semaphorin 5A gene (*SEMA5A*) is located on chromosome 5p15.2 and codes for a transmembrane protein. This protein contains three domains: a sema domain, a PSI domain and a Thrombospondin repeats (TSR) domain. In the carboxy-terminal side of

the sema domain, semaphorins contain a plexin-semaphorin integrin (PSI) domain. This small stretch of cysteine-rich residues has also been referred to as a MET-related sequence (MRS) or a cysteine-rich domain (CRD). With the exception of some viral semaphorins, all examples of proteins containing a sema domain have a PSI domain. The TSR of mammalian Sema5A protein are important in regulating the effect of Sema5A protein on axon guidance. The presence of an inhibitory sema domain and TSR domain suggests that SEMA5A protein is a bifunctional molecule [21].

SEMA5A transcripts were found in muscle, heart, lung, spleen, and at lower levels in brain. *In situ* hybridization analysis of mouse embryos revealed an expression of *SEMA5A* mRNA in mesodermal cells, as well as in the developing brain. Recently, *SEMA5A* expression was found in neuroepithelial cells ensheathing retinal axons, and the protein was shown to inhibit neurite outgrowth and to induce growth cone collapse. In addition, *SEMA5A* gene is one of the genes lost in the Cri-du-chat gene deletion syndrome, associated with severe mental retardation, further suggesting the role of this semaphorin in brain development [8].

1.5- Aim of the Research

The principle aim of this study was to investigate and confirm the results of the fine mapping that identified three candidate genes located on human chromosome 5p and associated with the number of Plasmodium falciparum attacks (PFA) in Dielmo village. This region was previously detected by linkage analysis. Our first aim was to identify by sequencing the SNPs presents in each of the three genes, *PDZD2*, *ADAMTS16* and *SEMA5A*, associated with PFA phenotype. The second aim was to genotype the SNPs identified in our population and test their association to PFA phenotype using two family-based programs, PBAT and QTDT.

2-Materials and Methods

2.1- Senegalese Population

The study participants were a subset of a larger prospective cohort study of malaria conducted in two Senegalese villages, Dielmo and Ndiop. A research program involving the Institut Pasteur in Dakar and in Paris, the Institut de Recherche pour le Développement (IRD), and the Senegalese government authorities, was initiated in the 1990s to study the epidemiology of malaria in populations naturally exposed to repeated infections of *P. falciparum*.

Malaria transmission is different in the two villages: in Dielmo, it is intense and perennial (holoendemic); in Ndiop, it is meso-endemic, strictly seasonal and dependent upon the rainy season that occurs from July to September. Such differing transmission has marked consequences for the epidemiology of malaria in the villages. This is most evident in the higher *P. falciparum* prevalence rates of infection in Dielmo (80%) compared to the seasonal rates in Ndiop that change from 20% in the dry season to 70% in the rainy season [59].

These two villages have been followed longitudinally since 1990, for the Dielmo village, and 1993, Ndiop village. For the present study, data was analyzed up to 1999. The malaria-related phenotype considered was *P. falciparum* clinical attacks (PFA) after correction by age of the individuals, trimester and year of the infection.

Both villages have a family structure with a high rate of consanguinity and polygamy and a large family size can sometimes reach several hundreds of individuals, [48]. In the present study, 585 individuals constitute the population of Dielmo, and are grouped into 10 complex families, while 664 are villagers of Ndiop, forming 19 complex families. DNA samples used in the context of this research were obtained from 421 Dielmo and 457 Ndiop individuals [48].

2.2- Identification of Single Nucleotide Polymorphisms

For the identification of single nucleotide polymorphisms (SNPs) in the genes of interest, we used two different methods. First, we sequenced all the exons and surrounding non-coding regions. Second, we chose the SNPs based on LD (Linkage Disequilibrium) by using the Haploview program.

2.2.1- Polymerase Chain Reaction (PCR)

2.2.1.1- Principle

The polymerase chain reaction (PCR) is an efficient and rapid method for amplification of a specific DNA region. PCR amplification of the DNA is accomplished by the use of primers. A pairs of primers, short single-stranded oligonucleotides, are designed complementary to the ends of the predefined sequence to amplify.

A PCR reaction contains the DNA to amplify, the pair of primers, deoxynucleotide triphosphates (dNTPs), and a thermostable enzyme (DNA polymerase). The PCR product is exponentially amplified by repetitive series of 30 to 40 cycles involving three steps, a template denaturation step, primer annealing step and an extension step (Figure 5) [58].

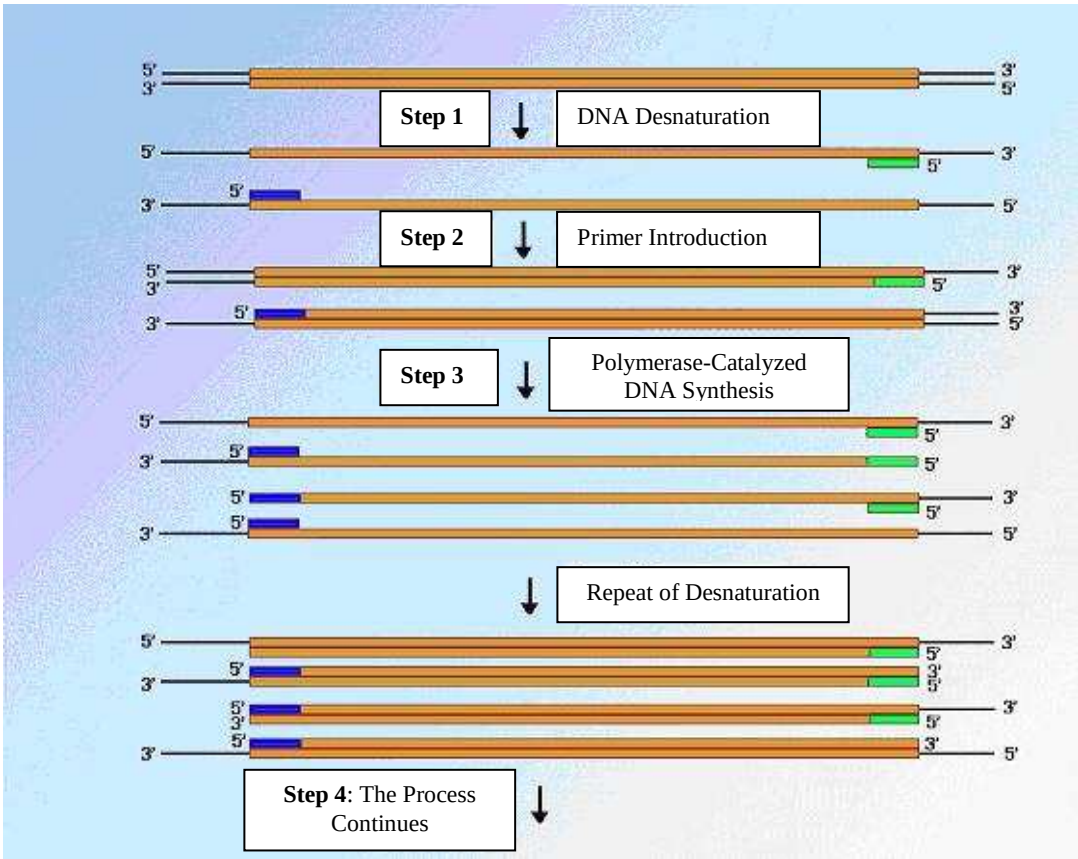


Figure 5- Principle of PCR amplification.

2.2.1.2- PCR conditions

The choice of primers is performed using the Primer3 program available on the website http:www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi and the DNAstrider program. Oligonucleotide primers are approximately 20 base pairs (bp) in length and contain from ten to twelve G+C nucleotides.

For each pair of primers, optimal conditions for DNA amplification are defined by testing four annealing temperatures (50º C, 55 º C, 60 º C and 65 º C). For each test, we used the following reaction buffer: 35.5µl of H_2O; 7.5 µl 10xTampon (Invitrogen); 3.0µl $MgCl_2$ (50mM); 3.73µl of dNTP (4 mM); 0.75µl of Forward and Reverse Primers (100 µM) and 0.50µl Taq Polymerase (5u/µl) (Invitrogen). For each of the four samples, 10µl of this buffer was mixed with 5µl of the DNA sample (1µg/µl of genomic-amplified DNA). The last sample is a control without DNA. PCR conditions are 95ºC for 5min followed by 35 cycles of 95ºC for 30 sec, annealing temperature as defined previously for 30 sec and 72ºC for 30s to 1 min depending of the size of the PCR products (1 min/kb). Efficiency of the PCR amplification was tested by loading 3µl of each sample on a 1% agarose gel and running it at constant voltage (100V) in x1Tris Borate EDTA buffer (Biosolve).

Forty-eight DNA samples from predefined individuals of Dielmo village were PCR amplified in a 96-well plate according to the condition previously defined (Thermo-Fast ABgene).

2.2.2- Sequencing of PCR products in ABI 3700

2.2.2.1- Principle

The principle of sequencing using the Big-Dye kit is based on stopping DNA synthesis following incorporation of one dideoxynucleotide (either ddATP,or ddCTP, or ddGTP, or ddTTP) labeled with different fluorochromes. Dideoxynucleotides are nucleotides lacking a 3'-hydroxyl (-OH) group on their deoxyribose sugar. The lack of this hydroxyl group means that, after being added by a DNA polymerase to a growing nucleotide chain, no further nucleotides can be added as no phosphodiester bond can be created. Deoxyribonucleoside triphosphates enable DNA synthesis to occur through a condensation reaction between the 5' phosphate of the current nucleotide with the 3' hydroxyl group of the previous nucleotide. The absence of the 3' hydroxyl group,

prevents further chain elongation. The size of newly synthesized DNA fragments depends on which ddNTP has been incorporated. Separation of fragments is made by migration on a polymer gel and detection of the fluorescent ddNTP by a laser.

2.2.2.2- Sequencing conditions

PCR products should be purified before sequencing since remaining primers and nucleotides that have not been incorporated, can interfere with the sequence reaction. Each of the 48 PCR products previously amplified is purified by gel filtration (Bio-Gel P100). 100g of Bio-Gel P100 is mixed with 2L pyrolyzed H_2O and incubated over night at 4 º C. After mixing the gel, 300µl of Bio-Gel P100 is added to each well of a 96-well plate filter Multiscreen (Millipore). The filtration gel is centrifuged at 1800 rpm during 3 min and the excess water removed. Fifteen µl of the PCR product is loaded, and centrifuged at 1800rpm for 4 min. The eluted purified products are collected in a 96-well plate (ThermoQuick).

For sequencing the 48 PCR products, a reaction solution is prepared containing 100 µl of Half-Big Dye (Bioline), 50 µl of Terminator solution (Applied Biosystems, version V1.1), 50 µl of a primer (25 µM) and 250 µl H_2O.Two buffers are prepared containing either the Forward primer of the PCR or the Reverse one. The PCR Sequencing reaction is performed in Thermoquick thermo well by adding 1µl of purified PCR product and 9 µl of the reaction solution. Conditions of PCR amplification are a denaturation step at 96 ° C for 5 min followed by 30 cycles of two steps, a denaturation one at 96 º C for 10 s and an extension one at 60 º C for 4 min.

PCR products need to be purified by gel filtration (Sephadex G50) before migration on the ABI3700 DNA Analyzer. 100g of Sephadex G50 superfine is mixed with 2L of pyrolyzed water overnight at 4 °C. After mixing the gel, 300µl of Sephadex G50 is added to each well of a 96-well plate filter Multiscreen (Millipore). The filtration gel is centrifuged at 1500 rpm for 3 min and the excess water removed. Ten µl of the sequencing product is loaded, and centrifuged at 2000rpm for 2 min. The purified products are collected in 96-well Plate MicroAMP Optical (Applied Biosystems). Sequence products migrate in capillaries containing POP6 polymer (Applied Biosystems) from a 3700 DNA analyzer (Applied Biosystems). Sequences were collected on a Dell computer using the 3700 Data Collection software. Detection of

SNPs was performed by using the multi-alignment program Genalyswin2.8.3b available in http://software.cng.fr/docs/genalys.html.

2.3- Genotyping

After selection of SNPs in each candidate gene, we ordered from Applied Biosystems either a Pre-design SNP Genotyping Assay or a Custom SNP Genotyping Assay depending on the presence of that SNP in the Applied Biosystems catalogue. This method is based on the amplification of a PCR product containing this specific SNP. Each allele of the SNP is detected by the specific hybridization of one probe labeled with different fluorescent probe.

2.3.1- Principle of TaqMan Genotyping

The TaqMan or 5´nuclease assay, genotyping system uses fluorescence-based polymerase chain reaction (PCR) reagents to provide qualitative detection of targets using post-PCR (endpoint) analysis. This assay requires only three components: purified DNA genomic, TaqMan Universal PCR Master mix and SNP genotyping assay. The SNP genotyping assay contains sequence-specific forward and reverse primers and probes.

A fluorogenic probe is an oligonucleotide labeled with both a fluorescent reporter dye (FAM or VIC) and a fluorogenic absorber, the quencher dye, hybridizing to specific allele.

When the probe is intact, the proximity of the transmitter and the quencher results in the absence of fluorescence emission by the transmitter due to the transfer of energy to the quencher. During PCR amplification, each probe that hybridizes specifically to one allele of the PCR product will be degraded by the AmpliTaq Gold DNA polymerase, thus separating the transmitter and quencher, leading to increased fluorescence emission by the transmitter due to the disappearance of the transfer of energy to the quencher. At each PCR cycle, the increase of the concentration of each transmitter (FAM-dye or VIC-dye) not linked to the quencher is directly dependent on the concentration of each allele present in the sample. Therefore at the end of the amplification, fluorescence emission by each of the two fluorochromes depends on the

presence of the allele that had hybridized specifically to its probe in the genomic DNA. A substantial increase in VIC-dye fluorescence indicates homozygosity for allele 1, a substantial increase in FAM-dye fluorescence indicates homozygosity for allele 2 and both VIC and FAM- dye fluorescences indicate heterozygosity [32].

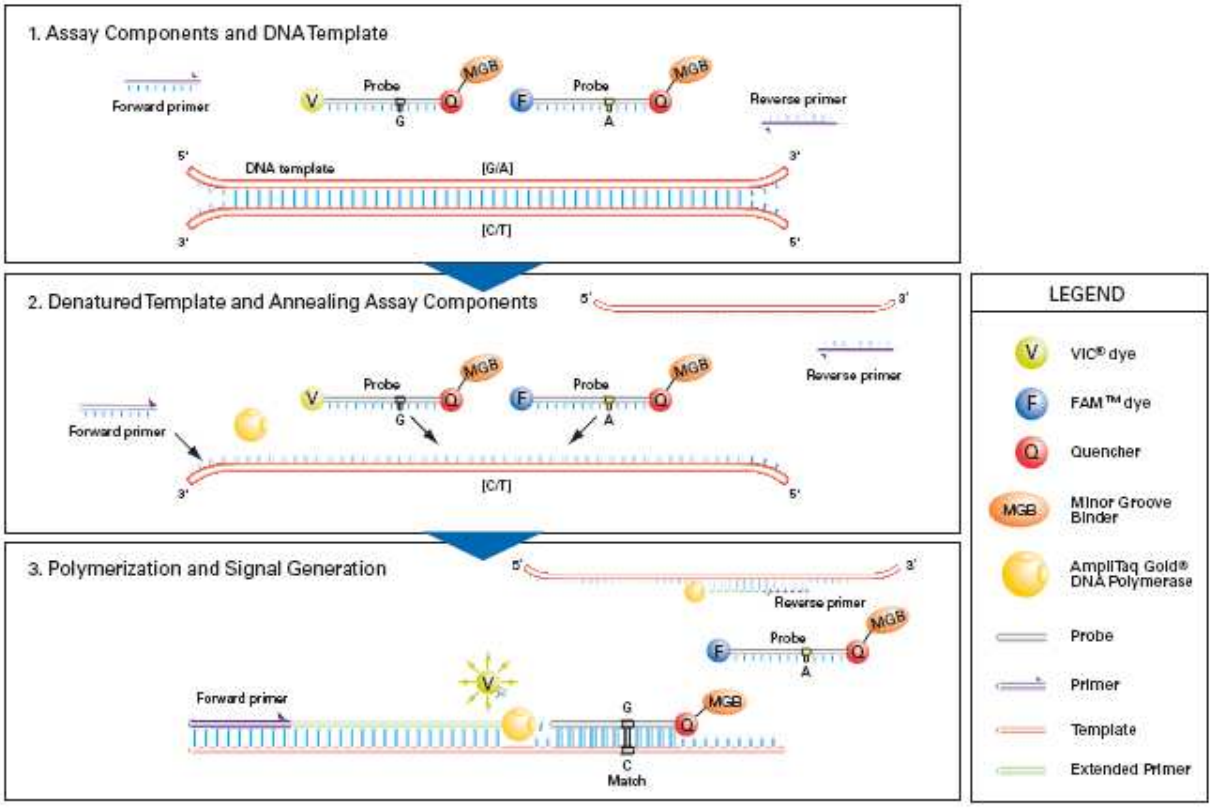


Figure 6- Stepwise representation of the forklike-structure-dependent, polymerization associated, 5´ to 3' nuclease activity of *Taq* DNA polymerase acting on a fluorogenic probe during one extension phase of PCR. Two dyes, a fluorescent reporter (R) and a quencher (Q), are attached to the MGB (fluorogenic probe). When both dyes are attached to the probe, reporter dye emission is quenched. During each extension cycle, the *Taq* DNA polymerase cleaves the reporter dye from the probe. Once separated from the quencher, the reporter dye emits its characteristic fluorescence.

All TaqMan SNP genotyping assays are designed and optimized to work with TaqMan Universal PCR Master mix. The SNP genotyping assay contains sequence-specific forward and reverse primers to amplify the polymorphic sequence of interest, two MGB (minor groove blinder) probes one labeled with VIC-dye and other labeled

with FAM-dye linked to their 5´end. A minor groove binder (MGB) at the 3´ end of each probe increases the melting temperature (Tm) for a given probe length [5;25], which allows the design of shorter probes. Shorter probes result in greater differences in Tm values between matched and mismatched probes, producing robust allelic discrimination. The primers and probes have been designed by Applied Biosystem.

2.3.2-The PCR amplification

The mother stock of the DNA samples of the 421 Dielmo individuals and 457 Ndiop individuals are contained in ten 96-well plates. For each plate, the reaction buffer was performed according to the following conditions: 137.5µl of H_20; 250µl of TaqMan Universal PCR Master Mix; 12.5µl of primers/probes (pre-design or custom SNP genotyping assay). In each well, 4 µl of the reaction buffer was added followed by 1 µl of purified genomic DNA (1 µg/µl) using an HYDRA II automat (MATRIX, Fisher Instrument). For the first plate, the Thermal cycling conditions were: an initial denaturing step at 95ºC for 10 min followed by 40 cycles of two steps, the fist one at 92ºCfor 15 sec and the second one at 60ºC for 1 min. After PCR amplification, fluorescent measurement was performed on the ABI PRISM7000 Sequence Detection System using the allelic discrimination assay. Figure 7 shows a plot of the processed data that is generated by the software. Each point represents a different sample (individual). Four distinct clusters of samples are observed—Allele A Homozygotes, Allele T Homozygotes, A/T Heterozygotes, and samples where no amplification is observed (No Amp). If the four clusters can be easily distinguished, the 5 other plates are amplified as defined. If not, an additional 5 PCR cycles are added and re-read on the ABI PRISM7000 Sequence Detection System were performed until the separation of the 4 clusters is sufficient. If after two additional 5 cycles, the separation is still not sufficient the assay was considered as unsuccessful. A second test was performed increasing the volume of PCR to 10 µl with 2 µl of purified genomic DNA. In most cases, we obtain successful conditions during the first assay and some times during the second assay. PCR assay was exceptionally unsuccessful.

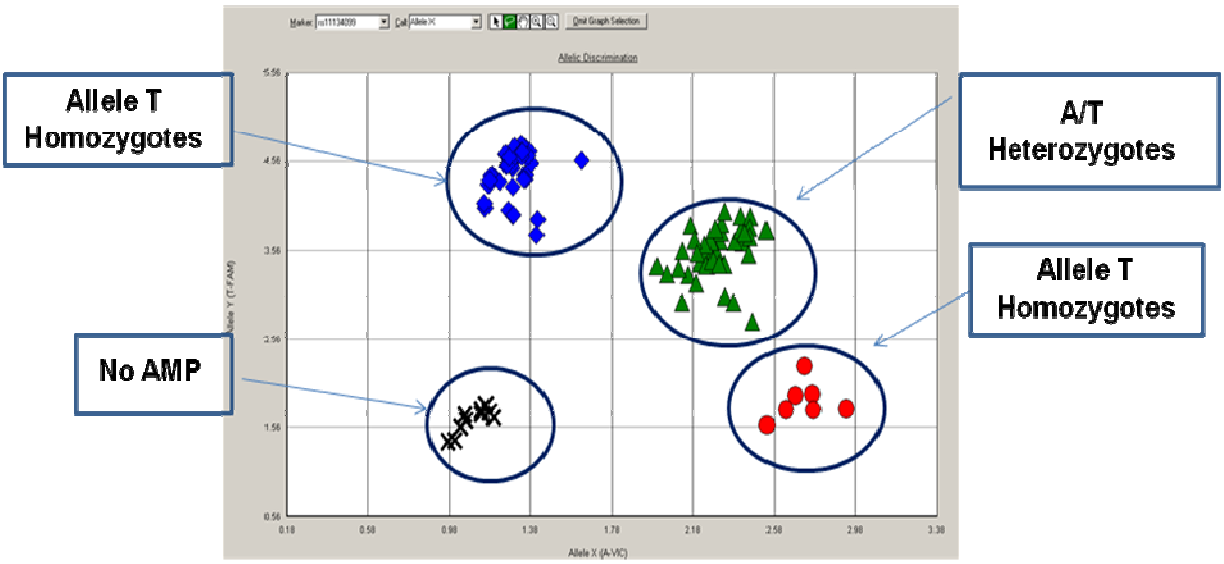


Figure 7- Allelic discrimination results with fluorogenic probes in the 5´ nuclease assay. A segment of exon 17 of the ADAMTS16gene was amplified. This segment contains an A:T single nucleotide polymorphism. The PCR reactions included a FAM-labeled probe for the T allele (Allele T), and a VIC-labeled probe for the A allele (Allele A). After PCR amplification, an endpoint fluorescence reading was made on the ABI PRISM 7000. As explained in the text, the fluorescence spectra were analyzed to generate a normalized Allele 1 value and a normalized Allele 2 value for each sample. Results are shown for the analysis of duplicate samples for 98 individuals.

2.3.3- Quality control of genotyping

Genetic studies require a high quality of genotyping data because any mistake can falsify the results of the analysis. The quality of the results was assessed by three processes. First, to reduce the risk of false results due to contaminant DNA, we performed a “zeros strategy”, using wells without DNA genomic. Second, to verify the specificity of the amplification, several individuals were genotyped more than once. Third, the Mendelian heritability was verified with the PEDCHECK program [38]. Genotype errors were checked by rechecking the genotype data using the allelic discrimination assay on the ABI PRISM7000 Sequence Detection System.

2.4- Bio-analysis programs

2.4.1- Study of Hardy Weinberg equilibrium

After identification of SNPs by sequencing, we proceeded to check the percentage of MAF (minor allele frequency). The frequencies were calculated using the Hardy Weinberg equilibrium equation:

$$p^2+2pq+q^2$$

Where, p= frequency of the minor allele "a"; q= frequency of the major allele "A". Therefore, we calculate the following MAF equation:

$$p= [2x\ (AA) + 1x\ (Aa)] / 2x\ n$$

Where n is the number of individuals.

In this study we selected the SNPs with MAF > 5%, in our population.

2.4.2- Primer Design

Primers were designed using the website primer3 and DNA Strider program.

DNA Strider [35] is a program designed for analysis of molecular sequence data. The program enables alignment of cDNA with its genomic sequence to precisely defined exon junctions, [35]. Primers were chosen to be oligonucleotides of 20 to 24 bases long and optimally containing 10 to 12 G + C nucleotides. The last five 3' residues should contain less than 3 G+C nucleotides.

With the website primer3 we were using the following criteria: melting temperature ranging between 58ºC to 65ºC and absence of dimmers or significant secondary structure.

We also verified that the primers were not in a region containing a known SNP by aligning the sequence of the genomic exon with those from dbSNP database [37] using Blast program.

The sequences of the primers are in supplementary tables I and II.

2.4.3- Sequencing analysis

SNP detection was carried out using the Multiple-Alignment GENALYSwin2.8.3b software by Takahashi M. available in http://software.cng.fr/docs/genalys.html.

GENALYS is a powerful tool to detect SNPs from multiple sequencing data. An advantage of GENALYS is that sequence alignment can be performed with regard to a reference sequence. We first introduced the reference sequence of one candidate exon (classical file text). Secondly, chromatograms directly obtained from the 3700 Data Collection software (ABI Format) are loaded. Sequences from the chromatograms are aligned with the reference sequence by a GENALYS command. Two view-windows are generated, one containing the reference sequence and the sequence differences for each chromatogram using the classical degenerate consensus code; the second view-window a contains each chromatogram. A pointer in the reference sequence allows linking to the corresponding position in the chromatogram. As the two strands of the PCR products have been sequenced, SNPs appear as differences specific for PCR products and not strand specific.

2.4.4- Haploview

The non-random association of alleles at different loci is termed linkage disequilibrium (LD) between markers. As ancestral haplotypes propagate through a population, their physical length is reduced by recombination events. Recombination events between makers separated by very short distances are very rare. Thus, genotypes at nearby markers are not independent and their association may reflect ancestral founder haplotypes.

We used Haploview program in the site of

http://www.broadinstitute.org/haploview/haploview.

Population data from HapMap phase III were downloaded from http://hapmap.ncbi.nlm.nih.gov/ . We select data of Yoruba population (African from Ibadan, Nigeria, West African) for analysis of the 3 candidate genes. After loading these data, the program creates a matrix of pairwise D´ and LOD measures to study LD between SNPs. D´ measures provides a direct measure of the amount of historical recombination that has occurred between two alleles. Output is give as a schematic

colour coded representation: WHITE is LOD < 2 and D´< 1; Blue is LOD < 2 and D´=1; Shades of pink /red is LOD #2 and D´<1 finally the Bright red is LOD #2 and D´=1 (The Broad Institute). Tag SNPs inside haplotypic blocks were defined using Tagger [9]. Thus, we were able to identify SNPs inside region with haplotypic blocks to analyze further. We also generate Haploview files from our own results in which the family structure was not taken into account to compare the haplotypic structure of Senegalese and Yoruba populations. In our analyses, these two populations have similar haplotypic structure differing only by frequency of the different haplotypes.

2.4.5- Study of Mendelian inheritance (PEDCHECK program)

Mendelian inheritance was checked by analyzing the marker data using the PEDCHECK program [38]. This program detects genotype incompatibilities in each family using either the genotypes of parents or those of kindred if genotype data from parents are un-known. Two files, one containing the name of the marker and the other the genotype data (LINKAGE format) are loaded. For each maker and each family, individuals with genotype incompatibilities are listed in a result file (pedcheck.err).

2.4.6- Association study

Family-based association tests used in the present study derived from the original transmission disequilibrium test TDT [51]. Two programs, PBAT/FBAT [30;26] and QTDT [4], adapted to the study of quantitative traits were used.

2.4.6.1- PBAT program

PBAT software is a powerful tool for complex family-based association analysis. This program can handle nuclear families with missing parental genotypes, extended pedigrees with missing genotypic information, analysis of single nucleotide polymorphisms (SNPs), haplotype analysis, quantitative traits, multivariate/longitudinal data and time to onset phenotypes. The data analysis can be adjusted for covariates and gene/environment interactions. Haplotype-based features include sliding windows and the reconstruction of the haplotypes of the probands [30].

The general "FBAT" statistic *U* [26] is based on a linear combination of offspring genotypes and traits:

$$U = S - E[S], \qquad S = \sum_{ij} T_{ij} X_{ij},$$

in which X_{ij} denotes some function of the genotype of the *j*-th offspring in family *i* at the locus being tested. It depends on the genetic model under consideration. The T_{ij} is the coded trait, depending upon possibly unknown parameters (nuisance parameters). In general, the coding for T_{ij} is specified as Y_{ij} - μ_{ij}. Here, Y_{ij} denotes the observed trait of the *j*-th offspring in family *i*, and μ_{ij} is an offset value.

The expectation in the expression for the general FBAT statistic is calculated under the null hypothesis of no association, conditioning on T_{ij} and on parental genotypes. If parental genotypes are missing, we condition on the sufficient statistics for parental genotypes. Under the same null hypothesis, *U* is unbiased since *E(U)=0*. Using the distribution of the offspring genotypes (treating X_{ij} as random and T_{ij} as fixed), *V =Var(U) =Var(S)* can also be calculated under the null and used to standardize *U*. If X_{ij} is a scalar summary of an individual's genotype, then the large sample test statistic

$$Z = U / \sqrt{V}$$

is approximately *N*(0,1). If X_{ij} is a vector, then

$$\chi^2 = U' V^{-} U,$$

has an approximate χ^2-distribution with degrees of freedom equal to the rank of *V*. Here, V^{-} denotes the inverse of *V* (or a generalized inverse when the inverse does not exist; this generalized inverse is based on the singular value decomposition of *V* [Press *et al.* 1986]. As this test generalized the transmission disequilibrium test, which directly tests transmission of an allele from a parental heterozygote to children, its result is independent of population stratification.

We employed the PBAT version available in:

http://cran.rproject.org/web/packages/pbatR/index.html

Two files, one containing the genotype data (LINKAGE format) and the name of the markers in a first line and the other the phenotype data, are loaded. We defined which trait and which SNPs or haplotypes have to be studied with which model of transmission. As suggested by the authors [28;26], with no *a priori* information, an additive model of transmission is considered the default. All the analysis were performed under the null hypothesis, linkage and no association. We defined significant association at $p < 0.001$.

2.4.6.2- QTDT program

The program QTDT studies the association between a genetic marker and a phenotype in extended families with several generations. This program can be used to analyze quantitative traits in nuclear families. It can be used when the population is not homogeneous and use all the information in a pedigree to construct powerful tests of association that are robust in the presence of stratification.

This program exploits the approach developed by Fulker et al, [17], showing that the genetic effect of a marker is partitioned into two components: one component corresponding to genetic effects within each family (W) and that corresponding to the genetic effect between the different families (B). Thus the average quantitative phenotype is modeled in the equation:

$$M=\mu+B+W \quad ;$$

and the null hypotheses by: $M=\mu+B$

We employed the qtdt-2.6.1 version available in:
http://www.sph.umich.edu/csg/abecasis/QTDT/download/ .

Two files, one containing the genotype data (LINKAGE format) followed by phenotypes and one defining markers, phenotypes, and covariates with their names, are loaded. Additive and dominant model of transmission were tested using the so-called "Orthogonal" model. We defined significant association at $p < 0.001$.

3- Results

3.1-Flowchart of experiments

3.1.1– Defining candidate genes

Recently, a genome-scan linkage analysis of several malaria-related phenotypes was performed in two villages of Senegal [48]. A suggestive susceptibility locus to the number of *Plasmodium falciparum* attacks was localized on chromosome 5p15 only in the holo-endemic Dielmo village (LOD = 2.26 empirical p = 0.0014), [48]. Furthermore, a fine-mapping study using a GoldenGate assay from Illumina was carried out in the candidate region (about 1450 markers located from 5 to 40 cM of human Chr5). A first analysis of the data using MERLIN program had suggested that two genes were associated with the number of *P. falciparum* attacks (PFA): First, the *PDZD2* gene with 3 SNPs, rs4867417, rs7714218 and rs11959398 with respective p-values = $1.7x10^{-5}$, $5.2x10^{-5}$, and $7.0x10^{-5}$; second, the *ADAMTS16* gene with one SNP, rs11134099 (p-value = $8.7x10^{-7}$).

To confirm this analysis, we re-analyzed the data of the fine mapping using two family-based association program QTDT and PBAT/FBAT and different models of allele transmission. This analysis confirmed the association of rs11134099 with the number of PFA, but did not confirm the association with *PDZD2* for any of the 3 SNPs (Table1). In contrast, a SNP of *SEMA5A* gene, rs3777320, was associated with PFA using a non-parametric model with Monte-Carlo re-sampling or a Rank test (Table 1).

We decided to analyze further these 3 genes *PDZD2*, *ADAMTS16*, and *SEMA5A*.

Table 1- Association results of the Fine Mapping with PFA phenotype, using Illumina. Analysis performed using family-based programs and non-parametric models (Monte-Carlo and Rank).

Gene	SNPs	Merlin Threshold $<1.0x10^{-4}$	QTDT Threshold $<1.0x10^{-3}$	PBAT Threshold $<1.0x10^{-3}$	Monte carlo Threshold $<1.0x10^{-3}$	Rank Threshold $<1.0x10^{-3}$
PDZD2	rs4867417	$1.7x10^{-5}$	-	-	-	-
	rs7714218	$5.2x10^{-5}$,	-	-	-	-
	rs11959398	$7.0x10^{-5}$	-	-	-	-
ADAMTS16	rs11134099	$8.7x10^{-7}$	$4.0x10^{-5}$	$8.9x10^{-4}$	-	-
SEMA5A	rs3777320	-	-	-	$1.0x10^{-4}$	$-9.6\ x10^{-4}$

3.1.2- Defining SNPs in candidate region of interest

Candidate SNPs were defined either by sequencing these genes in individuals from Dielmo village or by studying the linkage disequilibrium (LD) region located around the previously associated SNPs with the Haploview program and the data of Yoruba population from HapMap phase 3. Candidate exons of *PDZD2* and *ADAMTS16* genes were PCR amplified from DNA of 48 individuals of Dielmo village. PCR products were directly sequenced. Forward and reverse sequences of each PCR product were aligned with GENALYS program. For each SNP detected, we verified whether it was previously defined in dbSNP database [37]. We defined the candidate SNPs as all the non-synonymous and synonymous SNPs, independently of their minor allele frequency (MAF) and the other SNPs located in introns close to exon borders or in other non-coding regions if their MAF was greater than 5%. In addition, in LD regions, we defined Tag SNPs, which cover more than 90% of all haplotypes. For *SEMA5A* gene, candidate SNPs were defined by studying the linkage disequilibrium (LD) region located around rs3777320 SNP with the Haploview program and the data of Yoruba population from HapMap phase 3.

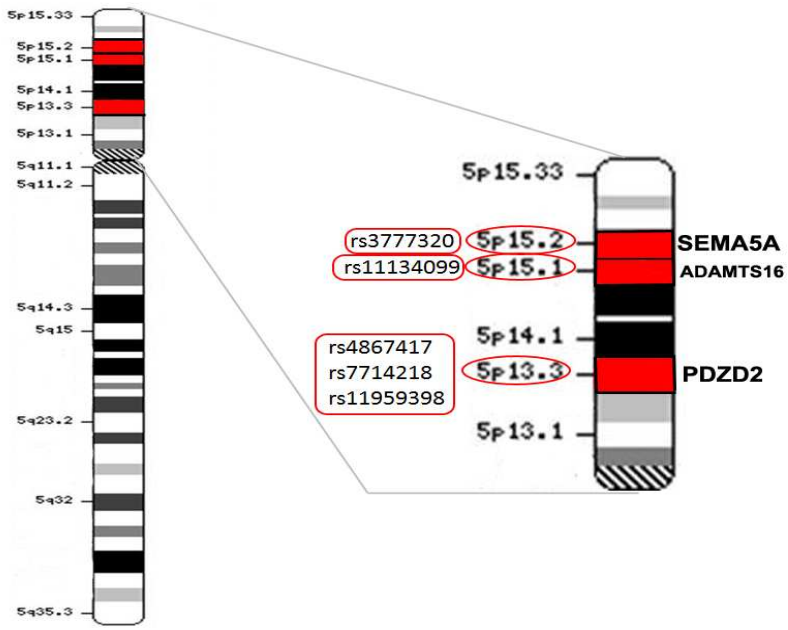


Figure 8- Representation of the SNPs identified by Fine Mapping analysis, and each respective gene.

3.1.3- SNP Genotyping and statistical analysis

SNPs were genotyped on a 7000 Sequence Detection System (Applied Biosystems) in 421 individuals of Dielmo village using either a Pre-design SNP genotyping assay or a Custom SNP genotyping assay from Applied Biosystems and five 96 PCR-plaques. For each SNP, we defined conditions of the PCR such as the number of cycles in the first amplification. In some cases, SNPs were also genotyped in an additional 457 individuals from Ndiop village by analyzing 4 other 96 PCR-plaques. Data from each SNP were loaded in a dbSNP database and in GenDB database.

For each gene, data files for QTDT/Merlin and PBAT/FBAT programs were generated from GenDB database. Genotypes were checked for error of Mendelian transmission using PedCheck program and errors were corrected after looking at the allelic discrimination screen of 7000 System Software (Applied BioSystems). Association between PFA phenotypes and each SNP or haplotype defined by combining different SNPs was studied using PBAT/FBAT program with H_0 hypothesis of linkage and no association and using different models of allele/haplotype transmission. Association between PFA and each SNP was also studied using QTDT program with different model of allele transmission. Association was considered significant if the p-value was below 0.001.

3.2- Analysis of the 3 candidate genes

3.2.1 *PDZD2* gene

PDZD2 is 240 kb long, contains 23 exons and codes for an mRNA of about 11, 650 basepairs long. Some alternative-spliced mRNAs have been described. From the three SNPs associated with PFA phenotype, two (rs11959398 and rs7714218) were located at about 80kb centromeric to the promoter and one other (rs4867417) in intron 16 (Figure 9). Thus, the candidate region for association with PFA phenotype contains the whole *PDZD2* gene.

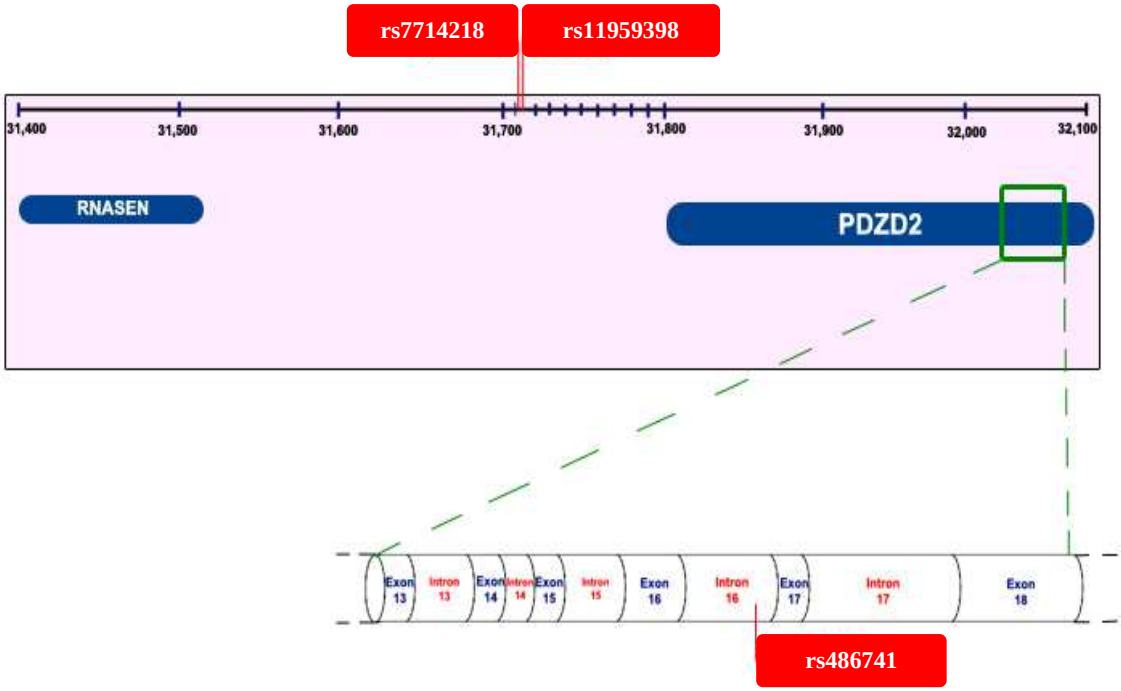


Figure 9- Physical map of the position of the three SNPs (rs7714218, rs119593 and rs4867417) of *PDZD2*, identified in fine mapping.

From 24 SNPs discovered by sequencing 20 of the 23 exons (5 new SNPs), twenty were defined as candidate SNPs (4 new SNPs) and 13 of which were genotyped (Supplementary Table III). No significant association was detected using either QTDT or PBAT/FBAT programs with different models of allele transmission (Supplementary Table 1). Also, no blocks of linkage disequilibrium among these 13 SNPs were detected using data of Dielmo village and Haploview program excepted for two, rs61746949 and rs2291114, which are less than 300 bp apart (Figure 9). Analysis of data from Yoruba population (HapMap Phase 3) confirmed that there are few blocks in this very large gene (data not shown). Under such conditions, the discovery of another SNP significantly associated to PFA phenotype in that region of more than 240 kb would be highly laborious and risky and thus it was considered preferable to concentrate on the two other candidate genes.

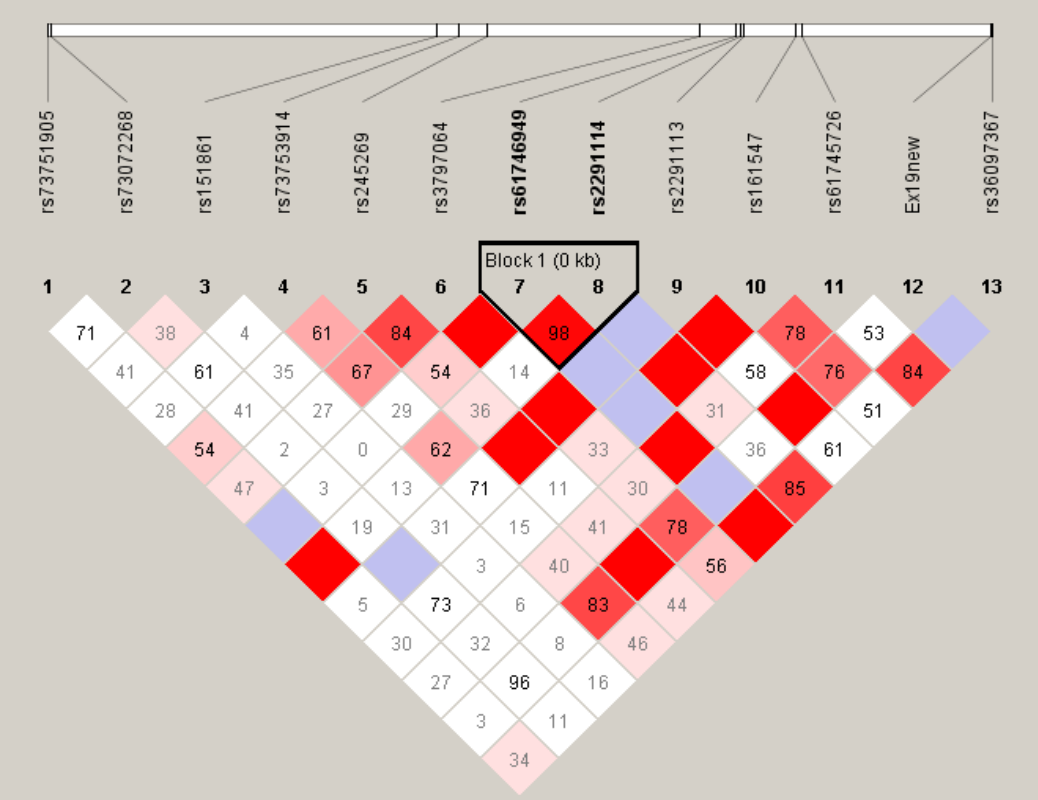


Figure 10- Representation of the LD using Haploview program with the SNPs genotyped in *PDZD2* gene.

3.2.2 *ADAMTS16* gene

ADAMTS16 is 190 kb, contains 23 exons and codes for an mRNA of 4921 base pairs long. The single SNP associated with PFA phenotype, rs11134099 is located in intron 17 of the gene (Figure 11). The choice of the candidate region for association with PFA phenotype was made by defining a region of LD containing 4 exons, from exon 15 to exon 18, and conserved non-coding sequences. Thus we sequenced four exons.

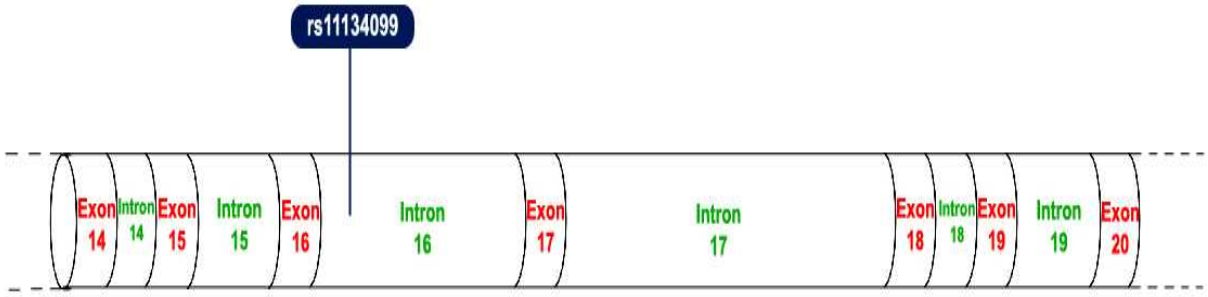


Figure 11- Physical map of the *ADAMTS16* and localization of the SNP (rs3777320) associated with PFA. This figure only represents 7 out of 23 exons existents in the *ADAMTS16* gene.

Three SNPs were discovered by sequencing, rs2964431 in intron 15, rs10475291 in intron 16 and rs2059799 in intron16 (Table 2). Only two of these three SNPs were genotyped (rs1047529 and rs2964431), because rs2059799 is in complete LD with rs2964331. We also decided to genotype two other SNPs (rs2086310 and rs2964433) that have been previously shown to be associated with inherited hypertension [23] and rs11134099, because of its association in the original fine mapping analysis by Merlin. Thus, six SNPs of *ADAMTS16* were genotyped.

Table 2- Characteristics of the SNPs identified in *ADAMTS16*.

Marker	Position	Identification	Minor Allele	MAF (%)
rs2964431	5229533 (Intron 15)	Sequencing	A	30.95
rs10475291	5230174 (Intron 16)	Sequencing	A	15.48
rs2059799	5230194 (Intron 16)	Sequencing	C	13.95
rs2964433	5227597 (Intron 14)	Association Hypertension	T	13.00
rs2086310	5136335 (Exon 3)	Association Hypertension	C	42.30
rs11134099	5235185 (Intron 17)	Previously Fine Mapping	A	33.30

We tested association between PFA phenotype and each of these six SNPs using the family-based programs, QTDT and PBAT/FBAT. This analysis of association was performed under the additive model and null hypothesis (H_0) of linkage but no association (Table 3).

In this analysis, we did not detect any significant association of PFA phenotype with any of the SNPs. Note that rs11134099 SNP, which exhibited a significant association with data from genotyping using Illumina (p value= $4.0x10^{-5}$, previously fine mapping), was not significant using Applied technology (Table 3).

Table 3- Results from association of SNPs for *ADAMTS16* gene with PFA phenotype using PBAT/FBAT program with the null hypothesis (H_0) of linkage but no association and QTDT programs. Both analyses implemented an additive model and used a significance threshold of 10^{-3}.

Analysis with PBAT		**frequency**	**Marker**	**Analysis with QTDT**	
p-value	Allele			Allele	**p-value**
-0.0609	C	0.307	**rs2964431**	C	Not significant
-0.229	G	0.0906	**rs2964433**	G	0.0432
0.356	G	0.121	**rs10475291**	G	0.0478
0.0494	T	0.331	**rs11134099**	T	0.0192
-0.924	C	0.438	**rs2086310**	C	Not significant

How to explain this discrepancy? The first hypothesis was that the two methods generated different genotypes for a great number of individuals. We were able to detect only eight individuals out of 429 with different genotypes with the two methods (Supplementary Table IV). However, these two methods generated positive genotypes for different individuals. We therefore merged the two genotyping data sets after taking into account the eight genotyping differences and reanalyzed the data (Table 4). We accounted for the eight genotyping differences by two different methods; first, the eight genotypes were considered as unknown (Applied + Illumina 2); second we re-examined the Applied results for the eight discrepancies. For five of them, we confirmed the result and as previously the genotype of these individuals was unknown, these genotypes were accepted. For the three other SNPs we estimated that the Illumina result was more probable (Applied + Illumina 1). Results in Table 4 show that the merged data of Applied and Illumina by each of the two methods are not significant albeit close to the defined threshold. The previous significant association between PFA phenotype and Illumina data of rs11134099 was likely due to chance and a bias in the partial genotyping of Dielmo individuals.

Table 4- Association analysis of rs11134099 SNP with PFA phenotype using four different genotype data sets, Applied genotyping, Illumina genotyping and two merged methods.

Marker	Allele	Frequency	PBAT/FBAT (p-value)	QTDT (p-value)
rs11134099 (Applied)	T	0.331	0.0494	0.0192
rs11134099 (Illumina)	T	0.324	0.000908	0.00004
rs11134099 (Applied + Illumina 1)	T	0.323	0.00121	0.0019
rs11134099 (Applied + Illumina 2)	T	0.326	0.00120	0.0022

The figure 12 shows that some LD exists between markers close to rs11134099 as also found in Yoruba population in HapMap (data not shown). This region of LD is of a small size containing 4 exons of *ADAMSTS16*, exon14, 15, 16, and 17. No SNP located inside these 4 exons was, however, associated with PFA phenotype.

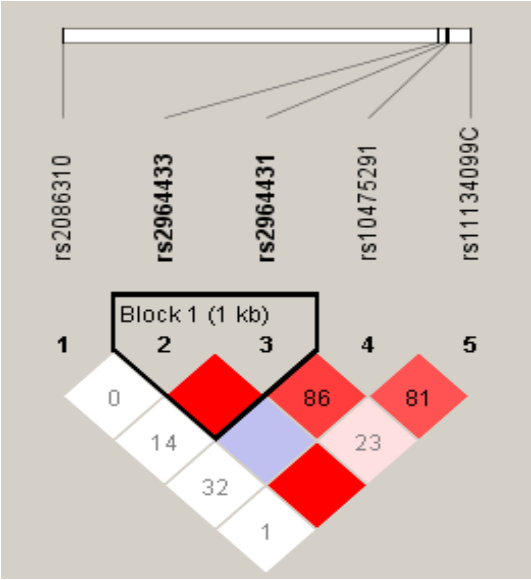


Figure 12- Representation of the LD using Haploview program with SNPs genotype in *ADAMTS16*. In the LD diagram are indicated the five SNPs where two SNPs formed one LD block. The rs11134099C represents the merged genotype classification between data using Illumina and Applied (method 1).

3.2.3 *SEMA5A* gene

SEMA5A gene is 500 kb long, contains 23 exons and codes for an mRNA of about 11,821 base pairs. The SNP associated (rs3777320) with PFA phenotype in the original fine mapping analysis, is located in intron 5.

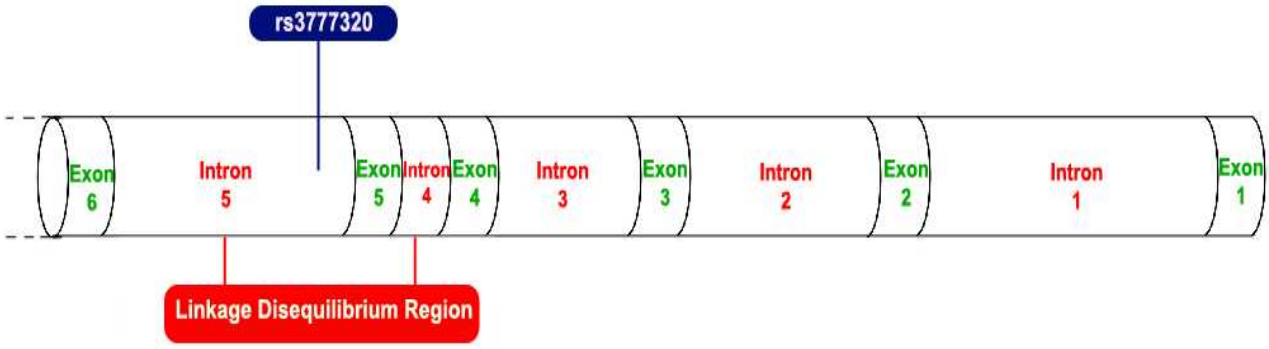


Figure 13- Physical map of LD region and position of rs3777320 within *SEMA5A* gene. This figure only represents 6 out of 23 exons of *SEMA5A* and the region of the Linkage Disequilibrium with rs3777320 (between block 45 and block 48).

In contrast to the other two genes (*PDZD2* and *ADAMTS16*), the selection of the SNPs of SEMA5A gene was not performed by sequencing. Sequencing of *PDZD2* and *ADAMTS16* genes to detect SNPs did not result in a large number of new SNPs: only five new SNPs from which only four had a MAF higher than 5%. Thus, to identify the SNPs of *SEMA5A* gene, we analyzed data of the Yoruba population from HapMap using Haploview program version 4.2. The rs3777320 is located within a 10 kb block of high LD; pairwise D´values between SNPs in this block higher than 0.95, MAF greater than 5% and the level of recombination inside the block lower than 0.1. In this block, we selected three SNPs rs3777311, rs3777316, and rs1557879 in partial linkage disequilibrium with rs3777320. As these three SNPs segregated identically amongst haplotypes only rs3777311 was selected. Interestingly, two other SNPs rs1018956 and rs451632 defined haplotypes that segregated similarly to alleles of rs3777320. They were also selected.

In the second step, we defined SNPs in linkage disequilibrium with rs3777320 in more distant blocks (Figure 14). The four haplotypes with the highest frequency in the block containing rs377320 are well conserved in the three adjacent blocks. We decided to test other SNPs bearing the same haplotypic structure as rs3777320 in these four

haplotypes (i.e. allele in the first haplotype different from those of the 3 others). Five SNPs were identified rs377325, rs3777327, rs1005934, rs3797980, and rs9313273. As rs3777325, rs3797980, and rs9313273 segregated identically, only rs377325 was genotyped.

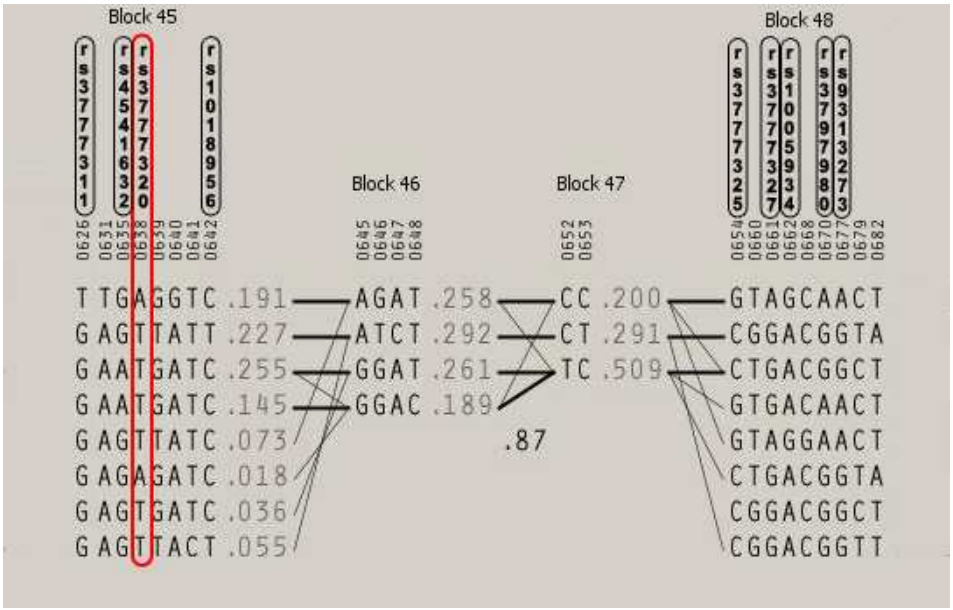


Figure 14- Haplotypic blocks surrounding rs3777320 SNP. Only selected SNPs have their name given in the upper part. SNP number 631: rs3777316; SNP number 640: rs1157879; SNP number 661: rs3777327; SNP number 670: rs3797980; SNP number 677: rs931273.

In summary, five SNPs were selected, three rs3777311, rs3777325, rs10059341 because they segregated similarly to rs3777320 and rs1018956, and rs451632, because the haplotypes they defined segregated similarly to rs3777320 (Table 5).

Table 5- Characteristics of the five selected SNPs of *SEMA5A* gene.

Marker	Position in Chromosome	Position in LD Block	Gene position
rs10059341	9315205	Block 48	Intron 4
rs3777325	9309778	Block 48	Intron 4
rs1018956	9300216	Block 45	Intron 5
rs3777311	9291324	Block 45	Intron5
rs4541632	9296125	Block 45	Intron 5

All the five SNPs were genotyped using pre-design genotyping assays (Applied Biosystems) in 421 individuals of Dielmo village. Results were analyzed using two family-based association programs PBAT/FBAT and QTDT. The rs3777311 SNP was significantly associated with PFA phenotype using PBAT program (p = 0.000986) but not using QTDT one. In contrast, rs377325 SNP was significantly associated to PFA phenotype using both programs (p = 0.000649 and p = 0.0002, respectively for PBAT and QTDT program). Individuals of Ndiop village were also genotyped for this SNP, but, the association between rs3777325 and PFA phenotype was not found to be significant (data not shown). In Dielmo village, no other SNP was significantly associated to PFA phenotype (Table 6). In addition, the haplotype analysis with rs4541632 and rs1018956 using PBAT/FBAT program showed a significant association with PFA phenotype (Table 7).

Table 6– Association between the five selected SNPs of the SEMA5A gene and PFA phenotype. The null hypothesis (H_0) in PBAT program was linkage and no association. Both analyses were performed with the additive model using a threshold of 1x $10e^{-03}$. An allele decreasing PFA phenotype has a negative p value in contrast to those increasing the risk.

Analysis with PBAT/FBAT		**frequency**	**Marker**	**Analysis with QTDT**	
p-value	Allele			Allele	**p-value**
-0.002671	G	0.166	rs10059341	A	**0.0170**
0.470041	T	0.194	rs1018956	C	**0.0506**
0.006709	A	0.438	rs4541632	G	**0.0105**
-0.000986	T	0.169	rs3777311	G	**0.0126**
-0.000649	G	0.260	rs3777325	C	**0.0002**

In conclusion, suggestive results obtained with rs3777320 from Illumina analysis were confirmed with rs3777325 and the haplotype analysis from rs4541632 and rs1018956 data (table 7).

Table 7- Analysis for haplotypes defined by combining data of rs4541632, and rs1018956 using PBAT/FBAT program. The null hypothesis (H_0) is linkage and no association implementing an additive model and threshold of significance P< 0.001.

Haplotype rs4541632-rs1018956	**Haplotypic frequency**	**p-value**
G:C	0.383	$-6.820e^{-05}$
G:T	0.1764	$5.460e^{-01}$
A:C	0.426	$2.806e^{-03}$
A:T	0.0141	$2.349e^{-01}$

5-Discussion

Plasmodium falciparum infection is the major cause of disease and mortality in humans in the tropical regions of the world, especially in West Africa. Although it is recognized that human genetics contributes to infection outcome, the role of genes coding for immune effectors remains poorly understood.

The present study follows on from a previous fine mapping study that used a GoldenGate assay from Illumina in the candidate region (from 5 to 40 cM of human Chr5p), identified by linkage mapping [48]. The number of *P. falciparum* attacks in the population of Dielmo, a village in Senegal, was linked to microsatellite markers of chromosome 5p.

To date, association studies have been centered on populations of European descent, and the degree to which knowledge gained from these studies is transferable to other populations has not been extensively investigated [46]. As we studied an African population, this present work could contribute to a fine mapping project in such populations [22]. The fact that we studied an African population brought us some problems due to their specific characteristics. The problems we faced concerned the HapMap data, LD mapping of our SNPs linked to the phenotype PFA and minor allele frequencies (MAF), and the structure and mode of living of the families in Africa, particularly in the two villages of our study.

The data used for association analysis were from HapMap Project phase 3. About 1.5 million genetic markers from 1,115 individuals and 11 populations were described during Phase 3 of HapMap. A fundamental limitation of HapMap data comes from the relatively small number of individuals studied, predominantly of European descent. This is particularly problematic for our studies because SNP diversity is higher in African populations than in European or Asian populations. Therefore, with the HapMap coverage of the African populations, there is presently a fifty percent probability of finding a SNP in linkage disequilibrium with the causal SNP associated to phenotype PFA (data from Illumina not shown). In the HapMap data, the Yoruba population from Ibadan, Nigeria is the closest population to the Senegalese populations; there is no available data for a West African population. It remains uncertain to what extent data from the Yoruba population can be extrapolated to other populations in Africa, given the high level of haplotypic diversity and the complex population structure [56]. The 1000 genomes project [1] is presently developing for a better understanding of the recent human population history, but little data have been hitherto published. For this reason, at the beginning of this study, we have decided to sequence all the exons

and the conserved non-coding sequences in the defined candidate regions. Results obtained from *PDZD2* and *ADAMTS16* exon sequencing showed that most of the SNPs were already in the database, except for some of those with low MAF (< 0.05). The LD between SNP was also similar between the studied populations and the Yoruba HapMap population. Thus, we consequently decided to use data from the Yoruba HapMap population to define candidate SNPs for *SEMA5A*.

The potential confounding effect of population structure on genetic association studies in Africa is illustrated by the existence of more than 2,000 distinct language groups, most of which correspond to a specific ethnic group [56]. There is growing evidence that these ethnic differences correlate with genetic differences and that levels of population structure are much greater within Africa than in other parts of the world. Failure to account for population structure in a community with multiple ethnic groups, as in Dielmo and Ndiop villages, can increase the false-discovery rate and reduce the power of the study. These confounding effects can be minimized by ethnic matching of cases and controls, but accurate matching can be difficult in communities in which there is substantial mixing between groups [56]. In Dielmo and Ndiop villages, the population forms a large complex family; some of which, because of multiple marriages, include several half-sibling relationships. This fact and the diversity of ethnic groups in both villages is a large limitation for an association study. To eliminate the confounding effects of population structure in Dielmo and Ndiop villages, we performed statistical analysis with family-based programs such as FBAT/PBAT and QTDT programs. These programs use statistical tests derived from TDT, which was developed to be independent of the population stratification.

Human genetic variants are typically referred as either common or rare, to denote the frequency of the minor allele in the human population. Common variants are synonymous to polymorphisms, defined as genetic variants with a MAF of at least one percent in a population, whereas rare variants have a MAF of less than 1%. There is presently an important debate among human geneticists about the respective role of common and rare variants in controlling complex traits [7]. Susceptibility to common diseases might be explained by association with multiple common variants. Alternatively, many rare slightly deleterious variants might also be responsible for common diseases, each of them explaining susceptibility for a small group of families. Thus, selection of SNPs with MAF greater than 5%, in this study, will only test the common variant hypothesis.

The efficiency of fine-mapping association depends on the genetic structure of the candidate region in the population studied. Fine-mapping association methods designed to identify risk variants seek to consider allelic shared descent, and are informative at a small scale [46]. The *PDZD*2 gene is 240 kb and the 3 SNPs associated with PFA phenotype were spread over the entire gene. To analyze this big candidate region, we decided to sequence all exons of *PDZD2* gene in a group of 48 individuals from Dielmo village. Most of the SNPs were tested for association with PFA phenotype and were negative. No LD blocks among the thirteen tested SNPs were detected. The lack of any significant SNPs, the size of the candidate region and the low LD complicate further analysis and made continued study of this gene untenable for this study period. For *ADAMTS*16 gene, in the initial fine mapping, only one SNP, rs11134099, was associated with PFA phenotype. A candidate region of small size was defined by using weak LD data. This region contains only four exons. They were sequenced to identify the candidate SNPs. As no SNPs were associated with PFA phenotype, we decide to re-genotype rs11134099. The initial association between rs11134099 and PFA phenotype in the fine-mapping study was not confirmed after extending the genotyping to all individuals of Dielmo by using the Taqman technology. In this case, the low LD around rs11134099 was sufficient to perform an extensive analysis of the region of interest, albeit with no significant success. The candidate region of *SEMA5A* gene is similar to that of *ADAMTS16* gene. Rs3777320 was located in a well defined block of LD and close to a second block in partial LD with the first one. After easily defining 3 candidate SNPs, we were able to rapidly confirm the putative association between rs3777320 and PFA phenotype. The results for the three genes studied, illustrate that the level of LD affects the efficiency to identify SNPs associated with a phenotype.

At the end of the present statistical analysis and in contrast to the previous fine mapping study, none of the SNPs genotyped for *PDZD2* gene was significantly associated to PFA phenotype. The fine mapping study was only analyzed by MERLIN program. A new analysis with two other programs, FBAT/PBAT and QTDT, did not confirm the fine mapping results of MERLIN program. We decided to test if some parameters had not been optimally taken into account in the first analysis with MERLIN. This program is able to separate the effect of linkage to that of association for each marker analyzed. During fine mapping, markers are SNPs and each of these biallelic markers measures a small part of the linkage effect. If SNPs are not clustered

using LD, MERLIN will underscore the effect of linkage. We confirmed that hypothesis by reanalyzing the fine mapping data with haplotypes defined from groups of SNPs with r >0.8. For the 4 SNPs of *PDZD*2 gene, the effect of linkage greatly increased but with a strong decrease of the effect of association. No other SNPs of *PDZD2* associated with PFA phenotype were discovered. This result coupled with the size of the candidate region, 240 kb, and the low LD makes the discovery of another SNP linked to PFA phenotype very unlikely with the present strategy. Other strategies will be more efficient, such as sequencing the whole *PDZD2* gene from two groups of individuals, one with high number of malaria episodes and the other with a low number. In conclusion, our results do not confirm the association of a SNP of *PDZD2* gene with PFA phenotype but further analysis will be necessary to evaluate the role of this gene.

The results for the SNPs of *ADAMTS16* were very different from those of *PDZD2*. Only one SNP, rs11134099, was found to be associated with PFA phenotype during the fine mapping. The candidate region was defined as a 5kb long block of weak LD. Three SNPs detected by sequencing regions surrounding Exon 15 to 18 were not associated with PFA phenotype. A new analysis of rs11134099 using Taqman technology did not, however, confirm the results of the original fine-mapping Illumina technology. This difference was not due to genotyping error. After merging both results and correcting a few discrepancies, p-values for rs11134099 were respectively equal to 0.00121 and 0.0019 with FBAT/PBAT and QTDT programs. This result is not significant, but close to the threshold of $p=10^{-3}$. As a great number of markers partially linked to each other have been studied during the fine mapping (about 1450 SNPs), the threshold of 10^{-3} is not very stringent (equivalent with Bonferonni's correction to a p-value of 5% with 50 independent tests). Our results strongly suggest that a locus of susceptibility to PFA phenotype is not located in the 5kb candidate region and likely not in the *ADAMTS16* gene at all. For *SEMA5A*, we found a strong block of LD around rs3777320, an LD structure quite different to that observed in *PDZD2* and *ADAMTS16*. In addition, sequencing *PDZD2* and *ADAMTS16* yielded no new SNPs with MAF > 5% and in the population of Dielmo. For these two reasons, we decided to choose as candidate SNPs those partially in linkage disequilibrium with rs3777320 in HapMap database (see the supplementary figure 3). We detected an association between rs3777325 and PFA phenotype with p-value= -0.000649 (PBAT) and p-value= 0.0002 (QTDT). Based on haplotype analysis, we found a strong association between two SNPs (rs4541632-rs1018956) with GC haplotype that segregated similarly to

rs3777320. These results strongly suggest that a locus of susceptibility to PFA phenotype is located in the blocks around rs377320. However, we cannot presently exclude that it is located in another region of the *SEMA5A* gene.

To definitely confirm that association, we have to replicate this result in another population. We first tried to replicate that association in the population of Ndiop village but this experiment was negative. This result was expected because the linkage with PFA phenotype and markers of chromosome 5p was detected in Dielmo village only and not in Ndiop [48]. This difference may be related to important differences in both the ethnic backgrounds and the prevailing transmission conditions in these villages. Another possibility of replication would be to use the same population of Dielmo but with data from a different time. In holoendemic region, such as those of Dielmo village, high numbers of PFA attacks occur in children of age lower than 10 years. So, two non-overlapping datasets analyzing clinical data for two periods of 10 years would enable replication. This analysis is presently underway.

What function could be altered by mutation of our putatively causal SNP in *SEMA5A*? *SEMA5A* is expressed in a great number of tissues. SEMA5A protein is well known to control axonal guidance [21], angiogenesis [47] but also recent association for autism [63]. Its expression is altered in many tumors of different cell types [39]. The rs3777320 is located in intron 5, near the end of exon 5 of *SEMA5A*. This gene has three domains: a Sema domain, a PSI domain and Thrombospondin domains. We verified that the rs3777320 is located inside of the SEMA domain. This domain is more conserved among different semaphorins of human species and across phyla than any other domains of this protein. The SEMA domain has been show to have a particular 3D-structure: a β-propeller with seven blades [68]. However, no non-synonymous or synonymous mutation have been described in exon 5, the only exon of *SEMA5A* gene inside the candidate region associated with PFA phenotype. Therefore, we searched for Human ESTs that have been described for this gene in the web site http://www.genome.ucsc.edu [60]. We found three groups of ESTs with undescribed exons inside intron 5 of *SEMA5A*.

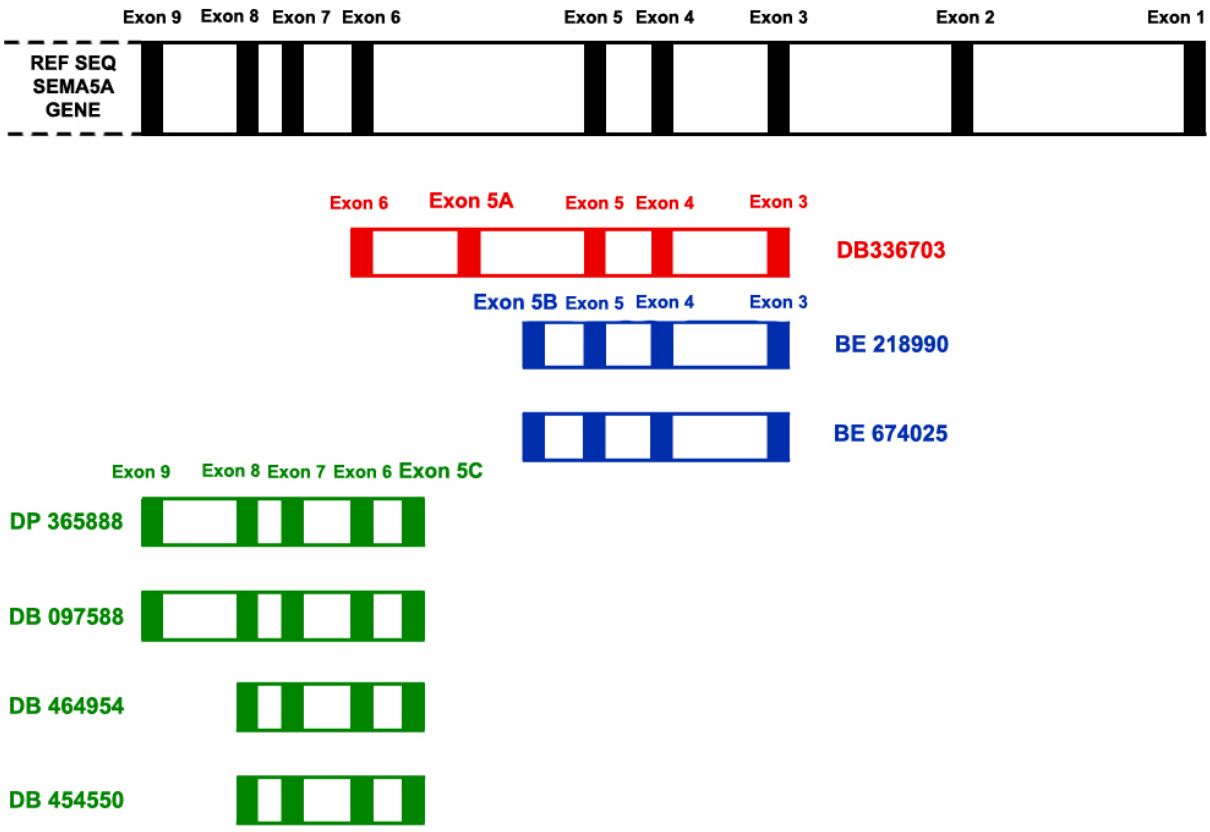


Figure 15- Representation of the exons of the ESTs suggesting alternative splicing among exon 5 and exon 6 in SEMA5A.

Alternative Splicing is important for regulation of protein activity and explains some tissue specificity. Many monogenic disorders have been associated to a problem of splicing. The first group of ESTs contains 4 ESTs BP365888, DB097588, DB454550, and DB464954, [60]. They all start by an additional exon located at the end of intron 5 close to exon 6 suggesting that this new exon is a first one controlled by a specific promoter. This exon is far away our from our candidate region and has not presently retained our attention. The second group of ESTs contains only EST DB336703. This EST has a new exon in the middle of intron 5 compared to the reference sequence of the *SEMA5A*. The open reading frame of *SEMA5A* is broken when this putative alternative spliced exon is used. This additional exon is, however, relatively distant from rs3777320 and is not located in the blocks of LD defining the candidate region. The third group contains two ESTs: BE218990 and BE674025. These have an additional exon located 182 bases after the last base of the classical exon 5. They contain, at their 3' end, a putative signal of polyadenylation AATAAA, but which

finish's abruptly after a few bases without any polyA tail. However, we think that this exon is a terminal one. Interestingly, SNPs have been described in that region inside an intron and just after this putative last exon. They constitute candidate SNPs to control the differential utilization of this last exon. All these hypotheses need to be confirmed, but suggest that a functional polymorphism affecting the efficiency of an alternative splicing controls expression of *SEMA5* gene.

Splicing signals are major contributors to evolutionary change and have evolved dramatically. Moreover, there has been a selective expansion of splicing regulatory proteins, such as serine/arginine proteins (SR proteins) in metazoans and heterogenous nuclear ribonucleoproteins (hnRNPs) in vertebrates, which may have assisted the basal splicing machinery in finding short exons in large intronic sequences, [24]. The crucial issue is to show if rs3777320 or another SNP in linkage disequilibrium affects the alternative splicing of one of these 3 putative exons. Presently, the most interesting exon is the putative terminal one.

6- Conclusion and Future Work

Studies of African population are challenging due to their specific characteristics and different factors. However, the study of Senegalese populations, from Dielmo and Ndiop villages, contributes to our understanding of the susceptibility/resistance of humans to malaria. Malaria caused by *Plasmodium falciparum* infection is more prevalent in Africa than elsewhere. This major infectious disease has been a strong force for recent evolutionary selection in the human genome. Repeated infection and exposure to the malaria parasites leads to the development of non-sterilizing acquired immunity, whereby the individual controls, but not eliminates, the parasite and shows no clinical symptoms. Discovering the human genetic factors that confer resistance to malaria would provide clues to the molecular basis of protective immunity. These studies might also be of great help for vaccine developments. In African populations, the low levels of linkage disequilibrium make association studies more difficult. However, after detection of an association, characterizing the causal polymorphism is likely to be easier. Presently, the genome diversity between African populations has not been as well evaluated as for European and Asiatic populations. The 1000 Genomes Project will be an important first step towards reliable multicentre GWA studies in Africa and the fine mapping of causal variants. One interest of the present study was to show that the Senegalese populations are not so different from the Yoruba population studied in the HapMap project. A second interest is that it suggests that the coverage of the previous fine mapping was too low because it identified ADAMTS16 but in fact its SEMA5A associated to PFA phenotype.

We found an association between SNPs of *SEMA5A* gene and PFA phenotype in the population of Dielmo village, and defined a 30kb long candidate region around rs3777325 SNP. One of the first future goals will be to verify this association in a second study. Replication in the population of Ndiop village was unsuccessful. This negative result is in agreement with the linkage study, which shows linkage only in the population of Dielmo village and not in that of Ndiop. One possibility will be to perform a replication over time in the population of Dielmo. In this holoendemic region of malaria, *Plasmodium falciparum* attacks occur mostly in children of less than 10 years old. As the population of Dielmo has been studied since 1989, it is possible to analyze two non-overlapping periods of ten years.

A second future goal will be to verify the hypothesis defined by the ETS analysis. First, test the existence of the 3 putative exons and define their expression in different tissues; second, verify that the first of the three alternatively splice exons is a

terminal exon and the last alternatively spliced exon is a first exon. These hypotheses will be tested by RT-PCR using cDNAs of different tissues (Real Time RT-PCR for RNA quantification, 5' and 3' Race to test the existence of a first and a terminal exon). This analysis will be followed by search of candidate SNPs for controlling different expression of these 3 exons Two aspects have to be developed in parallel: a bio-informatic analysis to define if a SNP polymorphism is located in a region, which might modify alternative splicing efficiency and Real Time RT-PCR studies to quantify the differential expression of these different exons depending on the haplotype of the candidate region.

To characterize the causal polymorphism, it will be interesting to study the evolution of these different alternative-splicing events in primate and mammal species. One goal of these studies will be to define regions that might control occurrence of these splicing events [24].

As mentioned previously, the susceptibility locus inside *SEMA5A* gene most probably does not explain the whole effect of the chromosome 5p on linkage with PFA phenotype. To define other genes controlling that phenotype, a second fine mapping analysis will be performed with new data from HapMap and 1000 genomes projects.

7-Bibliography

1- 1000 genomes. 2008. A Deep Catalog of Human Genetics Variation. Available: www.1000genomes.org [Date visited 02/09/2010].

2- Abecasis, G. R., L. R. Cardon & W. O. Cookson. 2000. A general test of association for quantitative traits in nuclear families. *Américan Journal of Human Genetics 66:279-92.*

3- Abecasis, G. R., S. S. Cherny, W. O. Cookson & L. R. Cardon. 2001. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics 30:97-101.*

4- Abecasis, G. R., W. O. Cookson & L. R. Cardon (2000b) Pedigree tests of transmission disequilibrium. *European Journal of Human Genetics 8:545-51.*

5- Afonina, I., M. Zivarts, I. Kutyavin, E. Lukhtanov, H. Gamper & R. B. Meyer. 1997. Efficient priming of PCR with short oligonucleotides conjugated to a minor groove binder. *Nucleic Acids Research 25:2657-60.*

6- Allison, D. B. 1997. Transmission-Disequilibrium Tests for Quantitative Traits. *The American Society of Human Genetics* 60**:**676-690.

7- Antonarakis, S. E., A. Chakravarti, J. C. Cohen & J. Hardy. 2010. Mendelian disorders and multifactorial traits: the big divide or one for all? *Nature Review Genetics 11:380-384.*

8- Artigiani, S., P. Conrotto, P. Fazzari, G. F. Gilestro, D. Barberis, S. Giordano, P. M. Comoglio & L. Tamagnone. 2004. Plexin-B3 is a functional receptor for semaphorin 5A. *EMBO Reports 5:710-714.*

9- Barrett, J. C., B. Fry, J. Maller & M. J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics 21:263-265.*

10- Cancre, N., A. Tall, C. Rogier, J. Faye, O. Sarr, J. F. Trape, A. Spiegel & F. Bois. 2000. Bayesian analysis of an epidemiologic model of Plasmodium falciparum malaria infection in Ndiop, Senegal. *American Journal of Epidemiology 152:760-770.*

11- Chaib, H., M. A. Rubin, N. R. Mucci, L. Li, J. M. G. Taylor, M. L. Day, J. S. Rhim & J. A. Macoska. 2001. Activated in prostate cancer: a PDZ domain-containing protein highly expressed in human primary prostate tumors. *Cancer Research 61:2390-2394.*

12- Deguchi Maki, Toshihiko Iizuka, Yutaka Hata, Wataru Nishimurai, Kazuyo Hirao, Ikuko Yao, Hiroshi Kawabe, and Yoshimi Takai. 2000. A novel multiple PSD-95/Dlg-A/Z-1Proten interacting with neural Plakophilin-related *Armadillo*

repeat Protein/δ-Catenin and p0071. *The American Society for Biochemistry and Molecular Biology* 275:29875–29880.

13- Eid, N. A., A. A. Hussein, A. M. Elzein, H. S. Mohamed, K. A. Rockett, D. P. Kwiatkowski & M. E. Ibrahim. 2010. Candidate malaria susceptibility/protective SNPs in hospital and population-based studies: the effect of sub-structuring. *Malaria Journal 9:119-129.*

14- Fenollar, F., J. F. Trape, H. Bassene, C. Sokhna & D. Raoult. 2009.Tropheryma whipplei in fecal samples from children, Senegal. *Emergency Infectious Diseases 15:922-924.*

15- Fontenille, D., L. Lochouarn, N. Diagne, C. Sokhna, J. J. Lemasson, M. Diatta, L. Konate, F. Faye, C. Rogier & J. F. Trap. 1997. High annual and seasonal variations in malaria transmission by anophelines and vector species composition in Dielmo, a holoendemic area in Senegal.. *American Journal of Tropical Medicine and Hygiene 56:247-253.*

16- Frazer, K. A., S. S. Murray, N. J. Schork & E. J.Topol. 2009. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics 10:241-251.*

17- Fulker, D. W., S. S. Cherny, P. C. Sham & J. K. Hewitt. 1999. Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics 64:259-267.*

18- Harris, B. Z. & W. A. Lim. 2001. Mechanism and role of PDZ domains in signaling complex assembly. *Journal of Cell Science 114:3219-3231.*

19- Havill, L. M., T. D. Dyer, D. K. Richardson, M. C. Mahaney & J. Blangero. 2005. The quantitative trait linkage disequilibrium test: a more powerful alternative to the quantitative transmission disequilibrium test for use in the absence of population stratification. *BMC Genetics 6 Suppl 1:S91.*

20- Hay, S. I., C. A. Guerra, A. J. Tatem, A. M. Noor & R. W. Snow. 2004. The global distribution and population at risk of malaria: past, present, and future. *The Lancet Infectious Diseases 4:327-336.*

21- Hilario, J. D., L. R. Rodino-Klapac, C. Wang & C. E. Beattie. 2009. Semaphorin 5A is a bifunctional axon guidance cue for axial motoneurons in vivo. *Developmental Biolology.326:190-200.*

22- Jallow, M., Y. Y. Teo, K. S. Small, K. A. Rockett, P. Deloukas, T. G. Clark, K. Kivinen, K. A. Bojang, D. J. Conway, M. Pinder, G. Sirugo, F. Sisay-Joof, S. Usen, S. Auburn, S. J. Bumpstead, S. Campino, A. Coffey, A. Dunham, A. E. Fry, A. Green, R. Gwilliam, S. E. Hunt, M. Inouye, A. E. Jeffreys, A. Mendy, A. Palotie, S. Potter, J. Ragoussis, J. Rogers, K. Rowlands, E. Somaskantharajah, P. Whittaker, C. Widden, P. Donnelly, B. Howie, J. Marchini, A. Morris, M. SanJoaquin, E. A. Achidi, T. Agbenyega, A. Allen, O. Amodu, P. Corran, A. Djimde, A. Dolo, O. K. Doumbo, C. Drakeley, S. Dunstan, J. Evans, J. Farrar, D. Fernando, T. T. Hien, R. D. Horstmann, M. Ibrahim, N. Karunaweera, G. Kokwaro, K. A. Koram, M. Lemnge, J. Makani, K. Marsh, P. Michon, D. Modiano, M. E. Molyneux, I. Mueller, M. Parker, N. Peshu, C. V. Plowe, O. Puijalon, J. Reeder, H. Reyburn, E. M. Riley, A. Sakuntabhai, P. Singhasivanon, S. Sirima, A. Tall, T. E. Taylor, M. Thera, M. Troye-Blomberg, T. N. Williams, M. Wilson & D. P. Kwiatkowski.2009. Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genetics 41:657-665.*.

23- Joe, B., Y. Saad, N. H. Lee, B. C. Frank, O. H. Achinike, T. V. Luu, K. Gopalakrishnan, E. J. Toland, P. Farms, S. Yerga-Woolwine, E. Manickavasagam, J. P. Rapp, M. R. Garrett, D. Coe, S. S. Apte, T. Rankinen, L. Perusse, G. B. Ehret, S. K. Ganesh, R. S. Cooper, A. O'Connor, T. Rice, A. B. Weder, A. Chakravarti, D. C. Rao & C. Bouchard. 2009. Positional identification of variants of Adamts16 linked to inherited hypertension. *Human Molecular Genetics 18:2825-2838.*

24- Keren, H., G. Lev-Maor & G. Ast. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics 11:345-355.*

25- Kutyavin, I. V., E. A. Lukhtanov, H. B. Gamper & R. B. Meyer. 1997. Oligonucleotides with conjugated dihydropyrroloindole tripeptides: base composition and backbone effects on hybridization. *Nucleic Acids Research 25:3718-3723.*

26- Laird, N. M., S. Horvath & X. Xu. 2000. Implementing a unified approach to family-based tests of association. *Genetics Epidemiology19 Suppl 1:S36-42.*

27- Lange C, L. H., DeMeo DL, Raby B, Silverman EK, Weiss ST. 2003b. A new powerful non-parametric two-stage approach for testing multiple phenotypes in family-based association studies. *Human Heredity* 56:10-17.

28- Lange C, Laird NM. 2002b. On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power and optimality considerations. *Genetics Epidemiology* 23**:**165-180.

29- Lange C, Silverman E.K, Xu X, Weiss ST, Laird NM. 2003c. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* 4**:**195-206.

30- Lange, C., D. L. DeMeo & N. M. Laird. 2002. Power and design considerations for a general class of family-based association tests: quantitative traits. *American Journal of Human Genetics 71:1330-1341.*

31- Lawaly, Y. R., A. Sakuntabhai, L. Marrama, L. Konate, W. Phimpraphi, C. Sokhna, A. Tall, F. D. Sarr, C. Peerapittayamongkol, C. Louicharoen, B. S. Schneider, A. Levescot, A. Talman, I. Casademont, D. Menard, J. F. Trape, C. Rogier, J. Kaewkunwal, T. Sura, I. Nuchprayoon, F. Ariey, L. Baril, P. Singhasivanon, O. Mercereau-Puijalon & R. Paul. 2010. Heritability of the human infectious reservoir of malaria parasites. *PLoS One 5:e11358.*

32- Livak, K. J. 1999. Allelic discrimination using fluorogenic probes and the 5' nuclease assay. *Genet Anal 14:143-149.*

33- Ma, R. Y., T. S. Tam, A. P. Suen, P. M. Yeung, S. W. Tsang, S. K. Chung, M. K. Thomas, P. S. Leung & K. M. Yao. 2005. Secreted PDZD2 exerts concentration-dependent effects on the proliferation of INS-1E cells. *The International Journal of Biochemistry & Cell Biology 38:1015-1022.*

34- Marchini, J. & B. Howie. 2010. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics 11:499-511.*

35- Marck, C. 1988. 'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. *Nucleic Acids Research 16:1829-36.*

36- Mouchet, J. 1999. Facteurs environnementaux lies au paludisme. Tansfusion Clinique et Biolologique. 6:35-43.

37- NCBI. 2008. Data base Single Nucleotide Polymorphism. Available: http://www.ncbi.nlm.nih.gov/snp [Date visited 5/11/2009].

38- O'Connell, J. R. & D. E. Weeks. 1998. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *American Journal of Human Genetics 63:259-66.*

39- Pan, G., H. Lv, H. Ren, Y. Wang, Y. Liu, H. Jiang & J. Wen. 2010. Elevated expression of semaphorin 5A in human gastric cancer and its implication in carcinogenesis. *Life Science 86:139-44.*

40- Phawong, C., C. Ouma, P. Tangteerawatana, J. Thongshoob, T. Were, Y. Mahakunkijcharoen, D. Wattanasirichaigoon, D. J. Perkins & S. Khusmith. 2010. Haplotypes of IL12B promoter polymorphisms condition susceptibility to severe malaria and functional changes in cytokine levels in Thai adults. *Immunogenetics 62:345-56.*

41- Porter, S., I. M. Clark, L. Kevorkian & D. R. Edwards. 2005. The ADAMTS metalloproteinases. *Biochemistry Journal 386:15-27.*

42- Prof. Dr. M.Eichner. 2009. Malaria life cycle. Available: http://www.uni-tuebingen.de/modeling/Mod_Malaria_Cycle_en.html [Date visited: 11/08/2010]

43- Rabinowitz, D. & N. Laird. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity 50:211-23.*

44- Rais Akhtar, Ashok K. Dutt, Vandana Wadhwa. 2010. Malaria in South Asia: eradication and resurgence during the second half of the twentieth century. London Springer, 241 pages.

45- Rogier, C., A. Tall, N. Diagne, D. Fontenille, A. Spiegel & J. F. Trape. 1999. Plasmodium falciparum clinical malaria: lessons from longitudinal studies in Senegal. *Parassitologia 41:255-259.*

46- Rosenberg, N. A., L. Huang, E. M. Jewett, Z. A. Szpiech, I. Jankovic & M. Boehnke. 2010. Genome-wide association studies in diverse populations. *Nature Reviews Genetics 11:356-66.*

47- Sadanandam, A., Erin G. Rosenbaugh, Seema Singh, Michelle Varney, Rakesh K. Singh. 2010. *Microvascular Research* 79:1-9.

48- Sakuntabhai, A., R. Ndiaye, I. Casademont, C. Peerapittayamongkol, C. Rogier, P. Tortevoye, A. Tall, R. Paul, C. Turbpaiboon, W. Phimpraphi, J. F. Trape, A. Spiegel, S. Heath, O. Mercereau-Puijalon, A. Dieye & C. Julier. 2008. Genetic determination and linkage mapping of Plasmodium falciparum malaria related traits in Senegal. *PLoS One 3:e2000.*

49- Snow, R. W., C. A. Guerra, A. M. Noor, H. Y. Myint & S. I. Hay. 2005. The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature 434:214-217.*

50- Spielman, R. S. & W. J. Ewens. 1996. *The TDT and other family-based tests for linkage disequilibrium and association*. American Journal of Human Genetics 59:983-989.

51- Spielman, R. S., R. E. McGinnis & W. J. Ewens. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Jpurnal of Human Genetics 52:506-16.*

52- Surridge, A. K., U. R. Rodgers, T. E. Swingler, R. K. Davidson, L. Kevorkian, R. Norton, J. G. Waters, M. B. Goldring, A. E. Parker & I. M. Clark. 2009. Characterization and regulation of ADAMTS-16. *Matrix Biology 28:416-24.*

53- Sylvie Manguin, Pierre Carnevale, Jean Mouchet, Marc Coosemans, Jean Julvez, Dominique Richard-Lenoble et Jacques Sircoulon 2008. Biodiversity of malaria in the world. J. Libbey Eurotext, Paris, pp 427.

54- Tam, C. W., V. W. Liu, W. Y. Leung, K. M. Yao & S. Y. Shiu. 2008. The autocrine human secreted PDZ domain-containing protein 2 (sPDZD2) induces senescence or quiescence of prostate, breast and liver cancer cells via transcriptional activation of p53. *Cancer Letter 271:64-80.*

55- Tatem, A. J. & D. L. Smith. 2010. International population movements and regional Plasmodium falciparum malaria elimination strategies. *Proceedings of the National Academy of Sciences 107:12222-7.*

56- Teo, Y. Y., K. S. Small & D. P. Kwiatkowski. 2010. Methodological challenges of genome-wide association analysis in Africa. *Nature Reviews Genetics 11:149-60.*

57- Thomas, M. K., K. M. Yao, M. S. Tenser, G. G. Wong & J. F. Habener. 1999. Bridge-1, a novel PDZ-domain coactivator of E2A-mediated regulation of insulin gene transcription. *Molecular Cell of Biology. 19:8492-504.*

58- Tom Strachan and Andrew P. Read. 2004. Human molecular genetics3. 3rd edition. Gardan Science New York, pp 674.

59- Trape, J.-F. 1996. Combating Malaria Morbidity and Mortality by Reducing Transmission. *Parasitology Today* 12**,** 236-240.

60- UCSC Genome Bioinformatics. 2006. UCSC Genome Browser on Human. Available: http://www.genome.ucsc.edu. [date visited : 26/05/2010]

61- Van Steen, K. & C. Lange. 2005. PBAT: a comprehensive software package for genome-wide association analysis of complex family-based studies. *Human Genomics 2:67-69.*

62- Weimin Chen and Goncalo Abecasis, Center for Statistical Genetics. MERLIN Tutorial -- Association Analysis. Available: http://www.sph.umich.edu/csg/abecasis/merlin/tour/assoc.html [Date visited: 28/01/2010].

63- Weiss, L. A., D. E. Arking, M. J. Daly & A. Chakravarti. 2009. A genome-wide linkage and association scan reveals novel loci for autism. *Nature 461:802-808.*

64- WHO. 2009. World malaria report 2009. Publisher World Health Organization. Geneva, pp 190

65- WHO. 2009. Malaria, Countries or qreas at risk of transmission. Available: http://gamapserver.who.int/mapLibrary/Files/Maps/Global_Malaria_ITHRiskMap.JPG. [Date visited: 29/07/2010].

66- Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter & X. Lin. 2010. Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *American Journal of Human Genetics 86:929-942.*

67- Xiong, M. M., J. Krushkal & E. Boerwinkle. 1998. TDT statistics for mapping quantitative trait loci. *Ann Hum Genet 62:431-452.*

68- Yazdani, U. & J. R. Terman. 2006. The semaphorins. *Genome Biology 7:211.*

69- Zeger, S. L. & K. Y. Liang. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics. 42:121-130.*

70- Zhang, L., D. Prather, J. Vanden Eng, S. Crawford, S. Kariuki, F. ter Kuile, D. Terlouw, B. Nahlen, A. A. Lal, L. Slutsker, V. Udhayakumar & Y. P. Shi. 2010. Polymorphisms in genes of interleukin 12 and its receptors and their association with protection against severe malarial anaemia in children in western Kenya. *Malaria Journal 9:87.*

Annex

Suplementary Table I- Primers used for polymorphism screening by sequencing of the gene *PDZD2*.

Primers Sequences		Length of PCR fragment (bp)	Position
Forward	**Reverse**		
GCAGGGACTTGTTAGACACT	GAAGATGTGATGCTGCGTTGA	906	Exon 1
CTCCTGATCCTACGGTACTT	GTTCTGAGCCCAGCTTGTAT	867	Exon2
CCGTCTTGGATGGCTCAAAT	GCCAGGGCTTTGAACAAGT	478	Exon3
CCAGAATGTCTTGACAAGGGA	CTGTGGAAGCAAGGCTGTTT	325	Exon 4
GCCTTTGAGGTAACCTTGCA	CGCCCAGGACTTCAGAAAAA	362	Exon 5
CTGAAGGGCTGTGCTTTGAT	ACACAGGCAGGAGGCTAAAT	441	Exon 6
GCCTGACACTAATGGCTTCA	GGCTGACTTCTTTGTAGCCT	515	Exon 7
CCAGCACCTGCGATAGTATT	AGGGTCATTCATTGCACCCT	766	Exon 8
CAGAACCAAGGGTGCAATGA	CTGACTATGCCAGGCAGATT	834	Exon 9
GTGACATTCCTGGGAGCTAA	CAGAGACCACGTGTTGTTGA	738	Exon 10 and 11
GTTAGCCATGGGCTCATTCT	GTCTTGGCTACGTTCTTGGA	449	Exon 12
CTCTGAGCCTGTGTTCCTAT	CCTATCCCCAGCTGGGAAAA	631	Exon 13
GCAGATATCAGCTGTGTTGT	GACTGTCTTTGACAGCTACT	479	Exon 14
			Exon 15
CTGGCAGCAGTCACCTTTTT	AGGTCTGGTTTAGCCACCAT	864	Exon 16
GGCCAAGTGGGAAGCTCAAT	GGCAGCTTACCCATGCTTTA	956	Exon 17
CCCTCACTACCAAACTAAGGA	GGTATCTCCCCACCTACTTT	678	Exon 18
GAGGGGCTGCTTTCTTATAT	ATGCTTCTGCAGCAAGGTCT	726	Exon 19 (1)
GAGCAGATTCTGTGTCCTCA	CTCTCTCGGAGTGCTGTATT	783	Exon 19 (2)
CCTGCCTCAGCCAAAGTTCT	GACCACACTTTGCCTTCTGA	775	Exon 19 (3)
CCTGCTGCGAATGCTGTGAA	CGCCAGTTTTGCCATGCTGA	804	Exon 19 (4)
CCAGGCAGAGCAGGAAATGT	AGCTGCGTCTCAACAACCTT	734	Exon 19 (5)
GGTCTTCCGGCAGCATTGTT	GAGATCCTGGGCAGCTTCA	587	Exon 19 (6)
			Exon 20
CCAAGAGGACTGACATCAGA	GGGATTGGAAATGGCCTATA	486	Exon 21
GGGGCTTAACAAGAGCACTT	CACGTGGATGGCATGAGTTA	623	Exon 22
			Exon 23

Suplementary Table II- Primers used for polymorphisms screening by sequencing of the gene *ADAMTS16*.

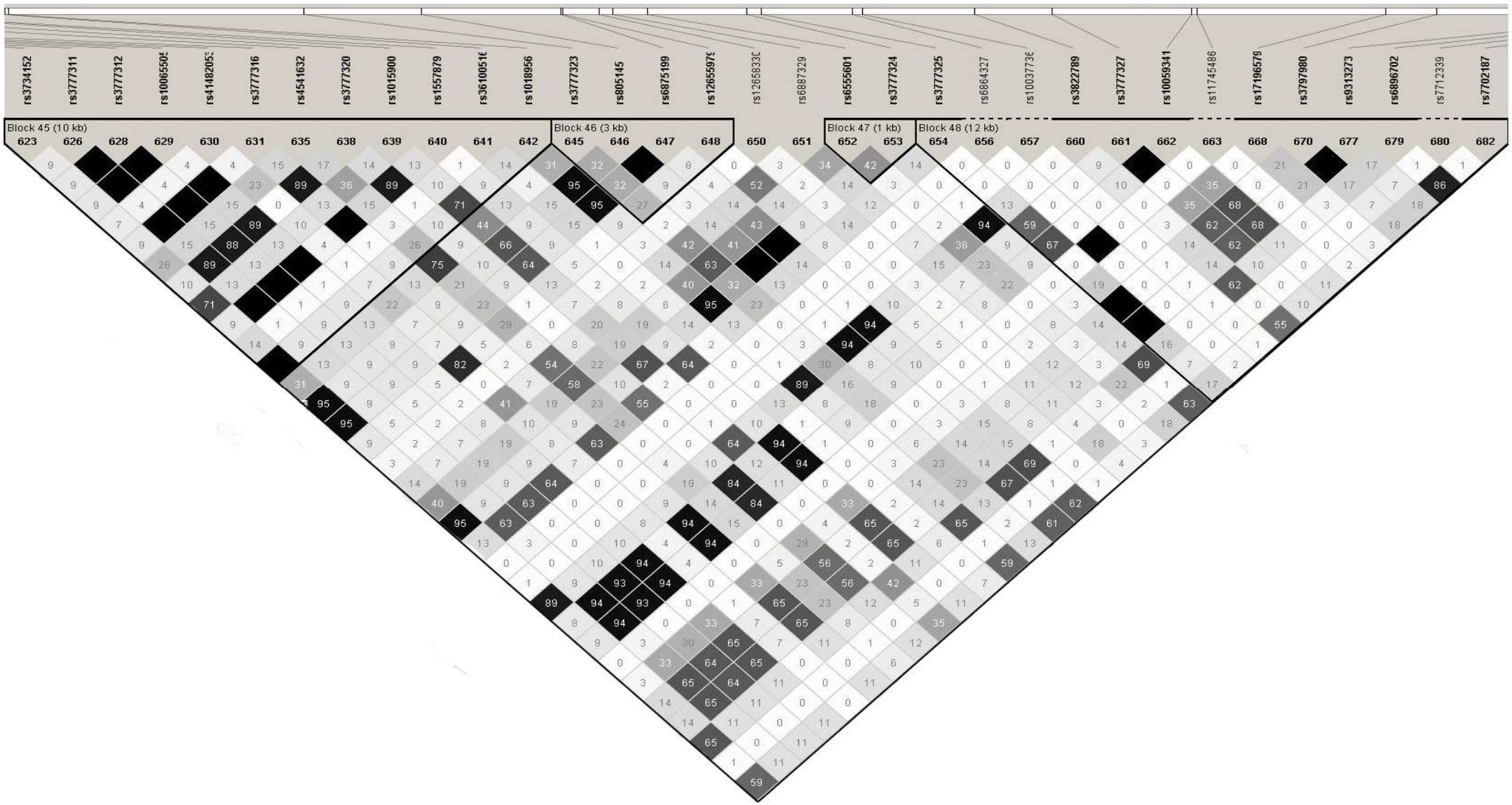
Primers Sequences		Length of PCR fragment (bp)	Position
Forward	**Reverse**		
GGGGGAAGAAAGAAAGTCTA	GCAGCATGTCATCAATGTGT	387	Exon 15
CCGCTGTGCACATATTAGGT	CCTGTGATGGTCACTTTTCCT	507	Exon 16
CACGTCTTCACCCAGTTCCA	CATCTCACAGCCATCGAGCT	656	Exon 17
CTAGATGGGAGGGTTCAGAA	GCCAGTGCAATGTACGCAGT	387	Exon 18

Supplementary Table III- The SNPs identified by sequencing of *PDZD2* gene. In Shown are the MAF (minor allele frequency) and the results of the association analysis using two family-based programs, PBAT/FBAT and QTDT. From 21 identified SNPs we studied 13(identified in blue). QTDT program only gives p-values when significant at P<0.05, otherwise the output is not significant (NS).

	Gene Position	MAF (%)	Type Mutation	frequency	PBAT/FBAT (p-value)	QTDT (p-value)
NewEx1	Exon 1	5.21	-	-	-	-
NewInt8	Intron8	2.17	-	-	-	-
NewEx5	Exon 5	2.04	-	-	-	-
NewInt6	Intron 6	11.25	-	-	-	-
Rs59628971	Intron 14	17.02	-	-	-	-
Rs57158698	Exon17	17.05	Non-Synonymouse Glu -» Gly	-	-	-
Rs16889405	Intron 17	6.67	-	-	-	-
Rs2279232	Intron 21	21.70	-	-	-	-
Rs161547	Intron17	7.78	-	0.103	0.0731	NS
Rs2291113	Exon 18	WWW	-	0.0505	0.463	NS
Rs36097367	Exon 18	16.67	Non-Synonymouse Thr -» Met	0.213	-0.297	NS
Rs151861	Exon 11	9.57	Synonymouse	0.0692	0.641	NS
Rs73751905	Intron 5	18.40	-	0.176	0.275	NS
Rs7372268	Intron 6	11.13	-	0.0779	-0.582	NS
Rs73753914	Intron 11	38.30	-	0.383	0.432	NS
Rs245269	Intron 12	33.30	-	0.392	-0.766	NS
Rs3797064	Intron 16	23.96	-	0.257	0.593	NS
Rs61746949	Exon17	15.56	Synonymouse	0.136	-0.531	NS
Rs2291114	Exon 17	22.09	Synonymouse	0.231	-0.509	NS
Rs61745726	Exon 18	40.43	Non-Synonymouse Ser- » Asn	0.376	-0.353	0.0425
EX19New	Exon 19	5.21	Synonymouse	0.0527	0.459	NS

Supplementary Table IV- Errors detected after merging genotyping of the results from Illumina and genotyping with marker from Applied Biosystems of *ADAMTS16* gene.

ID	**Illumina Genotyping**	**Applied Genotyping**	**Merged C**	**Merged D**
D0610 (D99129-D3102)	11	14	00	00
D0711(D0701-D0702)	14	11	14	00
D1514 (D99044-D99043)	11	14	00	00
D1526 (D1516-D1503)	11	14	00	00
D1812 (D991812-D1833)	11	14	11	00
D2208	14	44	00	00
D2226 (D2201-D2202)	11	14	00	00
D3208 (D99019-D99020)	14	44	14	00



Supplementary Figure I- Representation of the linkage disequilibrium region of the rs3777320 from the SEMA5A gene in the Yoruba population. The LD values are shown with R-squared, *r*2 values calculated using Haploview version 4.0 software, based on data from HapMap of the Yoruba population. The r^2= 1 is represented in Black, 0< r^2<1 is represented in grey and white is r^2=0.