CLONING AND CHARACTERIZING THE 2RJ INVERSION BREAKPOINTS IN THE MALARIA VECTOR ANOPHELES GAMBIAE S.S.

A Dissertation

Submitted to the Graduate School of the University of Notre Dame in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Mamadou B. Coulibaly, PharmD

Nora J. Besansky, Director

Graduate Program in Biological Sciences

Notre Dame, Indiana

April 2006
A BAC clone already known to map to the vicinity of the proximal breakpoints the 2Rj+(standard) was used as a starting material to identify the 2Rj (inverted) breakpoints in Bamako chromosomal form. The 2Rj distal and proximal breakpoints have been identified, cloned and sequenced. The structure of the breakpoints showed that it is more that a simple cut and flip. A 14.6 kb insertion at each breakpoint but opposite orientation has been identified. This fragment is made of two almost perfect 5.3kb inverted repeats separated by a 4 kb section and is structurally very similar to type 3 fold back transposable elements. Sequence analysis of the flanking regions of the breakpoints revealed the presence of four genes. However we found no evidence of interrupted transcripts by the inversion breaks. A simple PCR assay was designed to diagnose the 2Rj inversion.
DEDICATION

To:

My families

My grand-parents *in memorium*

Especially my lovely wife Sirandou Keita
# TABLE OF CONTENTS

LIST OF FIGURES .............................................................................................................. v

LIST OF TABLES ................................................................................................................. vii

ACKNOWLEDGMENTS ......................................................................................................... viii

CHAPTER 1: INTRODUCTION ................................................................................................ 1
  1.1. Malaria .......................................................................................................................... 1
    1.1.1. Importance .............................................................................................................. 1
    1.1.2. Vector control ........................................................................................................... 2
  1.2. Malaria vector: the Anopheles gambiae complex ....................................................... 4
    1.2.1. Taxonomy ............................................................................................................... 4
    1.2.2. Inversions ............................................................................................................. 6
    1.2.3. Molecular forms in the Anopheles gambiae s.s ................................................... 17
    1.2.4. Rationale and research objectives ....................................................................... 20

CHAPTER 2: MOLECULAR CLONING OF THE 2RJ INVERSION BREAKPOINTS .............. 22
  2.1. Introduction ................................................................................................................ 22
  2.2. Methods ...................................................................................................................... 26
    2.2.1. Genomic libraries ................................................................................................. 26
    2.2.2. Cytogenetics ......................................................................................................... 27
    2.2.3. Genomic DNA isolation and restriction digestion of BAC ................................ 28
    2.2.4. In situ hybridization and polymerase chain reaction ......................................... 28
    2.2.5. Library screening ............................................................................................... 30
    2.2.6. DNA sequencing ............................................................................................... 33
    2.2.7. Bioinformatics ................................................................................................. 34
  2.3. Results ......................................................................................................................... 35
    2.3.1. Cloning and characterization of the 2Rj inversion breakpoints ............................. 35
    2.3.2. Analysis of the surrounding regions of the breakpoints for coding sequences ... 42
  2.4. Discussion and conclusions ....................................................................................... 50

CHAPTER 3: MOLECULAR DIAGNOSTIC OF THE 2RJ INVERSION .................................. 54
  3.1. Introduction ................................................................................................................ 54
  3.2. Methods ...................................................................................................................... 56
    3.2.1. Study sites ............................................................................................................ 56
    3.2.2. Mosquito stock and genomic DNA extraction .................................................... 57
    3.2.3. Karyotyping ......................................................................................................... 57
    3.2.4. Designing primers and polymerase chain reaction ............................................. 58
  3.3. Results ......................................................................................................................... 59
  3.4. Discussion ................................................................................................................... 63
LIST OF FIGURES

Fig 1.1: Karyotypes and polyploid chromosome configurations expected from the random combination of the eight 2R chromosome arrangements most frequently observed as homokaryotypes in Anopheles gambiae from Banambani, Mali

Fig 1.2: The aridity cline (Modified from Powell et al, 1999)

Fig 1.3: Chromosomal relationships between the members of the Anopheles gambiae complex

Fig 1.4: Relationships between chromosomal forms and M and S molecular forms

Fig 2.1 The 2Rj inversion along with inversions b, c, u and d on the 2R chromosome, all in standard orientation

Fig 2.2: Hybridization bottles and meshes

Fig 2.3: Fluorescent in situ hybridization (FISH) of BAC clone 153N7 to a Bamako 2R chromosomes

Fig 2.4: Results of the two library screening experiments showing the plasmid and fosmid clones and their positions relative to the 2Rj+ (Pest) and the 2Rj (Bamako) chromosomes

Fig 2.5: Structure of the 2Rj inversion breakpoints

Fig 2.6 A & B Dot plot of 59I22 (Y-axis) and 35D6 (X-axis)

Fig 2.7 Predicted genes on the fosmid clone 59I22 from the distal breakpoint of the Bamako chromosomal form

Fig 2.8 Predicted genes on the fosmid clone 35D6 from the proximal breakpoint of the Bamako chromosomal form

Fig 2.9 Blastx results (NCBI) using 5.3 kb as query

Fig 2.10 The 14.6 kb insert with the 169 bp sequence similar to CR1 sequences in red present in the 4 kb middle region

Fig 2.11 Blastx of ~ 6kb (containing the 169 bp) against nr at NCBI
Fig 3.1 Study sites……………………………………………………………………………………56

Fig 3.2 A) relative regions where primers were designed  
B) Molecular diagnostic of 2Rj+ (standard orientation) and  
2Rj (inverted orientation)…………………………………………………………………………60
LIST OF TABLES

Table 3.1: Sensitivity and specificity of the 2Rj diagnostic assay…………………………….61

Table 3.2 Sensitivity and the specificity of the 2Rj diagnostic assay across different geographical areas in Mali……………………………………………..62
AKNOWLEDGMENTS

My committee members: Dr. Nora J. Besansky
    Dr. Jeffrey Feder
    Dr. David Severson
    Dr. Frank H. Collins

All Collins/Besansky lab members

All members of the Malaria Research and Training Center, Mali

People from the villages I collected mosquitoes

Special thank to: Dr. Patricia Romans
    Dr. Yeya T. Toure
CHAPTER 1

INTRODUCTION

1.1. Malaria

1.1.1. Importance

Malaria is a re-emerging parasitic infection caused by protozoans in the genus *Plasmodium*. There are four human malaria parasites: *P. falciparum*, *P. ovale*, *P. vivax* and *P. malariae*. The former species is the deadliest. Human malaria is transmitted by female mosquitoes of the genus *Anopheles*.

In 2002 the estimated number of malaria episodes worldwide ranged from 300 to 660 million (Snow et al., 2005). There are more than 1 million deaths every year most of which are children under 5 years old (WHO, 2005). Though 90% of these cases occur in Sub-Saharan Africa malaria is also a burden in other tropical and sub-tropical areas of the world. Despite the discovery of the malaria parasite’s life cycle more than one hundred years ago the disease still remains one the biggest challenges for public health authorities. Resistance of the parasite and the vector to drugs and insecticides respectively has made controlling the disease more challenging than ever. The 2002 breakthroughs that led to the sequencing of the malaria triangle (human, parasite and vector) have shed some light on the genetics of malaria but the problem remains as a whole.
1.1.2. Vector control

Vector control has been a very important component of the malaria control strategies (WHO, 1992). Several different methods have been used to control mosquitoes. Insecticides, larvicides as well as adulticides, have been and are still widely used. In effect the malaria eradication campaign of 1955 primarily targeted vectors by massive dichlorodiphenyltrichloroethane (DDT, an organochlorine) spraying (Mabaso et al., 2004). DDT is now less and less used because anophelines are developing resistance to it (Fanello et al, 2003) and it was found to be toxic, non biodegradable and able to accumulate in the food chain. Other organochlorines and other groups of insecticides such as organophosphates and carbamates are used in malaria vector control programmes. More recent than the above groups of insecticides pyrethroids are widely preferred due to their relatively low mammalian toxicity. They are used for both house spray and for impregnating materials such as insecticide treated bednets (ITNs) and curtains. Their use has considerably reduced the contact of mosquitoes with humans resulting in decrease in parameters such as biting rates (Guillet et al., 2001; N'Guessan et al., 2001). Besides the use of insecticides, effective mosquito control also relies on environmental management which, unfortunately, has drawn less attention (Keiser et al., 2005). This latter strategy coupled with public education about hygiene would efficiently help fight against malaria. A third technique of vector control is the biocontrol. A multitude of organisms have been used against mosquitoes as either larvicide or imagocide; including fishes (Kaneko et al., 2000), bacteria and fungi (Kanzok and Marcelo, 2005; Lawrence and Undeen, 1986). In the last three decades researchers have been investigating the use of genetically modified mosquitoes (refractory to parasites) to control malaria (Atkinson et al., 2001; Collins et al., 2000). Nowadays transgenic mosquitoes have been engineered (Ito et al., 2002) but an effective gene driving system into natural populations is yet
to be found though transposable elements seem to be good candidates (James, 2005; Boete and Koella, 2002; O'Brochta et al., 2003).

These techniques (see above) will not be as effective as required without a better understanding of the vector population structure.
1.2. Malaria vector: the *Anopheles gambiae* complex

1.2.1. Taxonomy

*Anopheles gambiae sensu lato* belongs to the *Anopheles* genus which is one of the three genera composing the *Anophelinae* subfamily along with *Bironella* and *Chagasia* (Kirzywinski and Besansky, 2003). Like many other insects of medical importance *Anopheles gambiae s.l.* is a cryptic species complex. For a long time though it was believed to be a single panmictic population until differences in ecological adaptations and behavioral diversities attracted the suspicions of researchers. These suspicions were further sustained by some morphological variations between individuals of this mosquito group. These observations made researchers more determined than ever to solve the *Anopheles gambiae* species puzzle. The first attempts of taxonomic classification were based on morphological, bionomic and ecological data. They were successful in distinguishing the brackish water breeders on the West and East African coasts from the freshwater breeders (Ribbands, 1944 in). Salt water adapted mosquitoes were first reported in 1903 by Dutton (White, 1974). Later these findings were genetically confirmed by Muirhead-Thomson (Muirhead-Thomson, 1948) when he demonstrated the reproductive isolation between salt breeder *Anopheles melas* and the fresh water breeder *Anopheles gambiae* due to hybrid male sterility in the progeny. That was only the first pieces of the puzzle unraveled. In 1955 at the Second African Malaria Conference in Lagos Botha de Meillon regarded the *Anopheles gambiae* taxonomic situation as “charged with explosive” (Coluzzi et al, 1979). Seven years later Davidson inadvertently pushed on the trigger that started the actual explosion and one sibling was discovered after another as in a cascade. In fact Davidson was conducting hybridization tests to reveal inheritance of insecticide resistance traits in *Anopheles gambiae* (Davidson, 1956; Davidson and Jackson, 1962) when he noticed that the hybrid male
sterility that was previously thought to be associated with the insecticide resistance actually resulted from effects of interspecific crossing. In his 1962 paper he reported the discovery of two sibling species of the vector that he designated by species A and species B. Since then thousands of cross-mating tests were conducted on wild material from several countries of Africa. These had resulted in recognition of six siblings: three fresh water breeders, *Anopheles gambiae sensu stricto* (species A), *Anopheles arabiensis* (Species B), *Anopheles quadriannulatus* A (species C) (Davidson, 1964; Paterson et al., 1963), one mineral water breeder found in the Semliki forest in Uganda, *Anopheles bwambae* or species D (White, 1974) and two salt water breeders, *Anopheles melas* in West Africa (Muirhead-Thomson, 1948) and *Anopheles merus* from East Africa (Muirhead-Thomson, 1951). Panoply of markers has been used to study the species composition of the *Anopheles gambiae* complex but not all led to a sustainable success. The morphometric, ecological and bionomic markers and the allozymes (Mahon et al., 1976) achieved some degree of success and remain useful to some extent. Inspite of that, at this point in history the gold standard was cross-mating to test for hybrid sterility. However in the 1960s chromosomal markers such as inversions replaced cross-mating and were proven to be very informative for species recognition in the *Anopheles gambiae* sibling species complex.
1.2.2. Inversions

It is generally accepted that inversions are generated by two independent but simultaneous breaks and the reattachment of the ends in an inverted orientation with respect to the flanking regions (Krimbas and Powell, 1992; Stadler, 1932). If the inverted region contains the centromere the inversion is called pericentric inversion. If the inverted region does not contain the centromere the inversion is called paracentric inversion. Inversions act as recombination suppression mechanisms and may link and protect co-adapted gene complexes (Dobzhansky, 1970; Mayr, 1963). Hence their importance in species evolution.

Alfred H. Sturtevant was the first to detect a paracentric inversion. This is the inversion on chromosome III distinguishing *Drosophila melanogaster* from *Drosophila simulans* (Sturtevant, 1926; Sturtevant, 1931). He had earlier postulated the existence of such rearrangements on chromosomes as a naturally occurring genetic phenomenon in *Drosophila* (Sturtevant, 1917). With the proper recognition of polytene chromosomes (Painter, 1933) the physical basis of inversions could be documented and relatively easily studied. Because of large efforts by Sturtevant and Dobzhansky the inversion polymorphism (i.e. in *Drosophila pseudoobscura* and *Drosophila persimilis* (Sturtevant and Dobzhansky, 1936a; Sturtevant and Dobzhansky, 1936b) served as a cornerstone in the establishment of experimental population genetics (Krimbas and Powell, 1992). Dobzhansky’s work on *Drosophila* inversion polymorphisms expanded to other areas and other species including mosquitoes.

The first cytogenetical investigations on the *Anopheles gambiae* complex were reported in 1956 where Frizzi and Holstein proposed a cytogenetic map for *Anopheles arabiensis* (species
B) (Frizzi and Holstein, 1956). These studies along with others such as the work on the *Anopheles maculipennis* group by Frizzi (Coluzzi and Sabatini, 1967), the work on the *Anopheles punctipennis* group (Baker and Kitzmiller, 1964) were inspirational to Mario Coluzzi and his group who published in 1967 a more complete and detailed cytogenetic description of *Anopheles gambiae* (Coluzzi and Sabatini, 1967). The crossing and cytogenetical studies based on the paracentric inversion polymorphism led to the subdivision of the *Anopheles gambiae* complex into six siblings (see above) until in 1998 when the discovery of a seventh sibling from Ethiopia was reported. It was named *Anopheles quadriannulatus* B because of its similarities with *Anopheles quadriannulatus* A (Hunt et al., 1998). Not only have inversion polymorphisms been important taxonomically for interspecific recognition but they have also been relevant for intraspecific distinction. In effect, in West Africa, fine inversion polymorphism analysis led to the split of *Anopheles gambiae* s.s. into five reproductive units referred to as chromosomal forms. Besides their role in taxonomic recognition between and within species inversions are also relevant for the eco-geographical adaptation of their carriers and for the deduction of phylogenetic relationships in evolutionary studies.
1.2.2.1. Inversions and chromosomal forms

Polytene chromosome investigations conducted on *A. gambiae* s.s. specimens collected from more than 500 different localities in Sub-Saharan Africa showed significant deviations from Hardy-Weinberg expectations. This led to the recognition of five reproductive units in *A. gambiae* s.s., namely Forest, Bissau, Bamako, Mopti and Savanna (Coluzzi, 1992; Coluzzi et al., 1985; Coluzzi et al., 2002; Della Torre et al., 2001; Powell et al., 1999; Toure et al., 1994; Toure et al., 1998). These units are referred to as chromosomal forms. Toure et al. (1998) found eight chromosomal arrangements, on the same arm, as the most frequent homokaryotypes. Those eight arrangements constituted the basis on which they described the Bamako, Mopti and Savanna chromosomal forms and their possible heterokaryotypes from Mali (Fig 1.1). They do interbreed freely under laboratory conditions but the heterozygote deficiency in natural populations support incipient speciation.

The chromosomal forms have different distribution pattern with respect to seasons and eco-climatic zones. But some of them i.e. Mopti, Bamako and Savanna can be found in sympatry.
Fig 1.1 Karyotypes and polytene chromosome configurations expected from the random combination of the eight 2R chromosome arrangements most frequently observed as homokaryotypes in *Anopheles gambiae* from Banambani, Mali. The shaded areas delimit the respective chromosomal forms. “+/+” is shared between Mopti and Savanna. “-“ = standard, “1” = heterozygote, “2” = homozygote. The inset figure represents the relative positions of j, b, c, u inversions on the 2R chromosome (Modified from Toure et al, 1998).
**Bamako chromosomal form**

In Mali the Bamako chromosomal form is found along the upper Niger river basin South of the inner delta but is most prevalent in the riverine areas and localities near the tributaries. Interestingly Bamako, the chromosomal form associated with the 2Rj inversion, particularly breeds in rock pools after the flooded water body has receded. It is almost absent during the dry season whereas it reaches high frequencies in the rainy season with a peak between August and September (Toure et al, 1998). This form has been, so far, recorded in southern Mali and in northern Guinea-Conakry.

Characterized by the homozygote j inversion the Bamako chromosomal form has been recorded in three different karyotypes, jcu/jcu, jbcu/jcu and jbcu/jbcu (Fig 1.1), based on the 2R, with b being the only floating inversion. The Bamako chromosomal form also contains the 2La inversion which is almost fixed in the populations.

**Mopti chromosomal form**

The Mopti chromosomal form breeds in flooded (irrigated) areas such as intense rice cultivation zones and in other sites due to human activities. Although its frequencies fluctuate from season to season the Mopti chromosomal form is present almost all year long with higher frequencies at the beginning of the rainy season compared to Bamako and Savanna (Toure et al, 1998). In Mali its distribution ranges from humid savannas to the Sahel and South Saharan areas displaying a remarkable eco-plasticity. Mopti has been reported from Mali, Burkina Faso, the northern part of Ivory Coast, Ghana, Togo and Benin in West Africa, and from Cameroon in Central Africa.
This form is characterized by the following karyotypes: u/+, u/u, bc/+; bc/u and bc/bc. It shares the +/+ standard homokaryotype with the Savanna chromosomal form (Fig 1.1). As in the Bamako chromosomal form the 2La inversion is nearly fixed in Mopti.

Savanna chromosomal form

Savanna mostly prefers breeding in rain-dependent sites away from big water bodies (e.g. main rivers) and flooded/irrigated areas. It also breeds in the most arid parts of the savanna belt, like A. arabiensis (Toure et al, 1998). However the 2Rb, the only Savanna karyotype, found in East Africa (Petrarca and Beier, 1992) has, seemingly, not been able to spread to the most arid zones of Somalia, Ethiopia and Sudan where A. arabiensis predominates. The temporal variations in the frequencies of Savanna parallel those of Bamako. They are either absent or in very low frequencies at the beginning of the rainy season, reach their peaks in August-September and decline towards the end of the rainy season (Toure et al, 1998).

Savanna is characterized by the following karyotypes: bcu/bcu, cu/bcu, cu/cu, b/bcu, b/cu, b/b, +/bcu, +/-cu, +/b and the standard +/-+. Inversion 2Rj may be present but in very low frequencies (Della Torre et al., 2005). In contrast with the two chromosomal forms above the 2La inversion is polymorphic in Savanna.
Forest chromosomal form

The Forest form is present in almost all rain forest areas and in humid or derived savannas. It apparently intergrades with various savanna chromosomal forms. The forest chromosomal form is characterized by the chromosome 2 monomorphic standard arrangement or few inversions such as 2La, 2Rb, 2Rd (Coluzzi et al, 1985). If *Anopheles gambiae* is believed to have originated from the forest area the inversion polymorphism data suggest this form to be the ancestral chromosomal form (see Powell et al, 1999).

Bissau chromosomal form

The Bissau chromosomal form has been reported from coastal rice cultivated areas of The Gambia, South Senegal (Casamance), Guinée Bissau and Guinée Conakry. It is characterized by the 2Rd inversion and shows a partial intergradation with the Savanna chromosomal form (Bryan et al 1982; Coluzzi et al, 1985). Interestingly Bissau is the only chromosomal form that is salt tolerant as it is able to colonize inland regions seasonally inundated with salt water (Bryan et al., 1982).
1.2.2.1. Inversions are associated with eco-geographical adaptations

The distribution of *A. gambiae* and *A. arabiensis* has been influenced by human activities. This is illustrated by the spread of the former from rain forests to savannas and the introduction of the latter into the forest area due to deforestation and urbanization. The resulting adaptation of *A. gambiae* to dryer areas than it is used to, apparently involves inversions as a mechanism for ecotypic differentiation. Chromosome 2 inversions in these species show frequency variations more or less related to an ecological cline in West Africa. The frequencies of some inversions increase with increasing degrees of aridity. In the study conducted by Coluzzi et al. (1979) the frequencies of some inversions ranged from zero in the moist tropical forest of West Africa to near or 100% in the Sahel in the North (see fig 1.2). Inversions 2Rb, 2Rbc, 2Rd and 2La appear in the savanna-forest zones. In the Guinea savanna there is a fairly sharp increase in the 2Rb and 2La frequencies. The frequencies of the 2Rd, too, sharply increase in Guinea savanna in relation to the microgeographical conditions. In Sudan savanna the 2Rbc and 2Rd show a rather consistent and parallel rise in frequencies (Coluzzi et al. 1979, Powell et al. 1999). A similar pattern was described in Mali specifically concerning 2Rbc (Toure et al. 1998). Interestingly, the frequencies of the standard chromosome in *A. gambiae* decline continuously from forest to sahel. Individuals collected from the forest areas are almost monomorphic for the standard chromosome two arrangements (Coluzzi et al. 1979). In *A. arabiensis*, 2Ra, the inversion most strongly associated with aridity, is apparently absent in the forests while it shows frequencies from less than 40% to up to 80% in Guinea savanna and Sudan and sahel savanna respectively. Incidentally it can reach higher frequencies or even fixation in semi-desert areas (Senegal and Sudan) (Coluzzi et al. 1979). The aridity cline is not only spatial. It is also characterized by a
seasonal pattern as frequencies of such inversion as 2Rbc varied from as low as 30 % in the rainy season to about 90 % in the dry season (Toure et al, 1998, Powell et al, 1999).

Fig 1.2 The aridity cline (Modified from Powell et al, 1999)

Besides the macrogeographical variations inversion polymorphism in *A. gambiae* and *A. arabiensis* shows microgeographical variations (Both between and within villages). Interestingly there are variations in the inversion frequencies even between samples collected indoors and those collected outdoors in the same compound showing a non random distribution of karyotypes. The most significant and direct investigations were conducted on *A. arabiensis* in 1975 in Jirima (Garki district in Nigeria). The results of these studies showed that mosquitoes carrying certain karyotypes have contrasted propensities to bite human indoors and rest indoors. Furthermore the more frequent karyotypes in indoor collections were those favored by drier
environments (e.g. 2Ra, 2La, 2Rd) whilst the karyotypes more frequent in outdoor collections were those adapted to relatively more humid environments (Coluzzi et al 1979, Powell et al, 1999).

1.2.2.3. Inversions are used to infer the phylogenetic relationships in the *Anopheles gambiae* complex

Inversions have been used to tentatively reconstruct the phylogeny of *Anopheles gambiae*. Contrary to the old belief that *A. quadriannulatus* was the likely ancestral species inversion data indicate that the most likely ancestral species among the seven siblings is *A. arabiensis* (Fig 1.3) (Ayala and Coluzzi, 2005). *A. arabiensis* might have originated in the Middle East and expanded to Africa via the Arabian peninsula. Data imply that *A. quadriannulatus* originated from *A. arabiensis*. The former split into two species, *A. quadriannulatus A* in South Africa and *A. quadriannulatus B* in Ethiopia (Hunt et al 1998). Two lineages originated from *A. quadriannulatus*. One led *A. bwamba* restricted to the Semliki forest in Uganda and to *A. melas* whose distribution is limited to the coast of Western Africa. The second lineage gave rise to *A. gambiae s.s.* which is subdivided into three chromosomal forms in Mali, namely Bamako, Mopti and Savanna. The latter is thought to be the ancestral form of the three (Toure et al, 1998). Inversion evidence suggests *A. gambiae* to be the ancestor for *A. merus*. 
Fig 1.3 Chromosomal relationships between the members of the *Anopheles gambiae* complex. (Modified from Ayala and Coluzzi, 2005).
1.2.3. Molecular forms in the *Anopheles gambiae s.s.*

In the last two decades, regions of the ribosomal DNA (rDNA) have been under study to molecularly distinguish apart the chromosomal forms. Variations in the intergenic spacers (IGS) and the internal transcribed spacers (ITS) led to the identification of two main rDNA types in West Africa: the IGS types, M and S (Favia et al., 1994; Favia et al., 2001; Favia et al., 1997) or the ITS types, type I and type II (Gentile et al., 2001; Mukabayire et al., 2001). They are now referred to as M molecular form (type II) and S molecular form (type I) respectively. In the West African countries of Mali and Burkina Faso the M molecular form corresponds to the Mopti chromosomal form and the S molecular form corresponds to Bamako and Savanna together. However this equivalence is not consistent beyond these West African borders. There are two explanations to that: 1) the reproductive isolation between them is independent of the chromosomal constitution (Della Torre et al., 2001; Toure et al., 1998); 2) there could be genetic variation across distant geographic areas (Gentile et al., 2002). If it is true from both approaches that one rDNA type corresponds to Mopti while the other corresponds to Bamako/Savanna, then the question of whether these entities are “good species” arises. This has been the subject of debate for long time. In addition the bigger picture of these chromosomal forms-molecular forms relationships is very complicated as the same chromosomal forms may be shared by different molecular forms (Fig 1.4).
Fig 1.4 Relationships between chromosomal forms and M and S molecular forms. White boxes include most frequent chromosome-2 karyotypes (and less common polymorphisms) typical of each chromosomal form. Shaded areas show the PCR-RFLP characterization of the chromosomal forms, as observed in the study area (Della Torre et al., 2001).

The geographical distribution of the molecular forms shows a general pattern of North-South for the M-form and East-West for the S-form. The distribution of the M-form shows an association with arid areas as it is, so far, the only molecular form reported from the Sahel savanna of Northern Senegal, the Sudan and the areas bordering the desert in Southern Angola. While the M-form shows a latitudinal distribution the S-form has a longitudinal one. Its distribution extends from the Sudan Guinea savannah in Senegal, Mali and Burkina Faso to South Eastern Africa. The S form is the only molecular form reported from East of the Rift Valley with a few exceptions (Della Torre et al, 2005). The distribution of the molecular forms is more complex on a regional scale. In West Africa they are frequently sympatric but in contrast within particular localities only one or the other, instead of both forms, is frequently recorded in a particular zone. This is explained by the alternation of semi-permanent/permanent breeding sites (e.g. rice fields, artificial lakes, excavations etc.) exploited by the M form, and rain-dependent breeding sites exploited by the S-form.
The question of whether or not M and S are good species is as controversial as the debate about the speciation process itself. M and S are thought to be at least partially reproductively isolated. Analysis of other parts of the genome failed to prove consistent differentiation between the forms. However rDNA studies (Favia et al, 2001, Gentile et al, 2001, Della Torre et al, 2002), and microsatellites loci (Wang et al, 2001, Stump et al, 2005) at the X chromosome centromeric region showed differences between M and S forms. This is supported by a significant difference in SINE insertion in the proximal region of the same X chromosome (Barnes et al, 2005) and the fact that the knock down resistance (kdr) gene, in certain areas, has been found in S form but not in the sympatric M form (Awolola et al., 2003; Fanello et al., 2003). In addition the hybrid deficit in natural population (<< 1 %) (Della Torre et al, 2005) indicates the presence of premating barriers between M and S forms. These data strongly suggest an on going process of incipient speciation within A. gambiae s.s. with the involvement of natural selection.
1.2.4. Rationale and research objectives

In Mali *A. gambiae* s.l. is composed of *A. gambiae* s.s. and *A. arabiensis*. *A. gambiae* s.s. consists of chromosomal forms namely Bamako, Mopti and Savanna. These forms, defined by different combinations of inversion j, b, c, d and u on chromosome 2, show a particular distribution pattern (see Bryan et al., 1982; Coluzzi et al., 1985; Toure et al., 1994; Toure et al., 1998). Each of the chromosomal forms is characterized by unique inversion polymorphisms conferring a difference in habitat choice, breeding season, microhabitat selection and in the adaptation to disturbed breeding sites (Coluzzi et al., 1985; Coluzzi et al., 1979). The association of niche partitioning with inversions is very important from an epidemiological standpoint. Inversion 2Rb/+ has been found predominantly in the usual sunlit shallow and rain-dependent *A. gambiae* breeding sites. Meanwhile the 2Rbc/+ characteristic of the Mopti chromosomal form is found mainly in more permanent breeding sites associated with irrigation and agriculture. This habitat shift allows *A. gambiae*, mainly the Mopti form, to breed during dry season, therefore extending the transmission of malaria through a larger part of the year (Toure et al, 1996). The transmission level also could be associated with inversion polymorphism as individual mosquitoes homozygote for inversion are often found indoors. This preference allows them more opportunity to blood-feed than the standard inversion carriers. Not only are inversions relevant for the eco-geographical distribution of *A. gambiae* but they also play a very important role in the genetics of this vector. Because inversions reduce or suppress recombination in heterozygous individuals hence protecting co-adapted gene complexes, they are of special interest in the evolutionary biology of *A. gambiae* s.s.. The degree of reduction or suppression increases near the breakpoints. The identification and characterization of the sequences neighboring or at the breakpoints is of paramount importance. These sequences could provide
researchers with critical information about the genetic structure of the breakpoints, the association of transposable elements with inversion breakpoints and the monophyletic origin of the inversions. In addition breakpoint sequences could also be informative from an epidemiological standpoint as they can be used to design a PCR-based diagnostic assay for the inversion that can avoid some of the difficulties involved in karyotyping. So studying the biology of inversion is a critical step for researchers and control programmes to take in order to better understand the ecology and the genetics of *A. gambiae s.s.*.

The present work, specifically aiming at characterizing a particular inversion, 2Rj, will provide tools to shed light on the ecological behavior of *A. gambiae s.s.* subpopulations associated with that inversion. This way more specific measures can be taken towards a directed control strategy. This work will also allow designing a PCR-based assay to diagnose the 2Rj inversion. To achieve the goal set above we have fixed the following objectives:

**Objectives:**

1. **Identify, clone and sequence the 2Rj breakpoints of *Anopheles gambiae s.s.***

2. **Analyze the surrounding regions of the breakpoints for transposable elements or any other repeated sequences**

3. **Develop and validate a diagnostic PCR for the 2Rj inversion**
CHAPTER 2

MOLECULAR CLONING OF THE 2RJ INVERSION BREAKPOINTS

2.1. Introduction

Chromosomal inversion polymorphisms, first described by Alfred Sturtevant in 1917 (see chapter 1), play an important role in the evolution of closely related species of insects. In heterozygous individuals there is an increasing topological constraint on homolog chromosome pairing near the inversion breakpoints (Matzkin et al., 2005; Novitski and Braver, 1954). Consequently, the homologous chromosome bearing the inverted sequences forms a loop called inversion loop for the best base-pairing to occur. Then a cross-over results in significant reduction or suppression of viable recombinants as some gametes are lost in the process because they are unbalanced. This possibly results in the protection of co-adapted gene complexes and eventually leads to genetic differentiations in a population. Generally believed to be maintained in natural population by balancing selection (Dobzhansky, 1970) inversions have neither clearly known functions (Andolfatto et al., 2001) nor known origin. However, regarding their origin, transposable elements are associated with their breakpoints and are thought to be responsible for the generation of several of them which have been characterized molecularly (Càceres et al., 2001; Càceres et al., 1999; Evgen'ev et al., 2000; Gray Y. H., 2000; Ladevèze et al., 1998; Mathiopoulous et al., 1999; Wolf-Ekkehard and Saedler, 2002; Zelentsova et al., 1999). Regarding their functions they are associated with the eco-geographical distribution of insects.
populations such as *Drosophila* and *A. gambiae*. Their origin may not be crystal clear but inversions remain strongly associated with ecological adaptations in the *Anopheles gambiae* complex, the primary vector of malaria in Sub-Saharan Africa. Because particular inversion or combinations of inversions are associated with particular groups of individuals in particular habitats inversions are, therefore, an important tool in the determination of insect population structures and behaviors. Such inversion derived information about natural populations is critical for effective disease vector control programs such as malaria.

In Mali *Anopheles gambiae s.s.* is composed of three chromosomal forms (see chapter 1), Bamako, Mopti and Savanna. The j inversion is among the five polymorphic paracentric inversions on the right arm of chromosome 2 that are used in different combinations to cytologically identify the chromosomal forms (Toure et al., 1998). It has a higher frequency in the Bamako chromosomal form where it is found in the following karyotypic combinations: jbcu/jbcu, jcu/jcu and jbcu/jcu. It is also found in Savanna but in very low frequencies (Della Torre et al., 2005). The 2Rj inversion, unlike b, c, d and u, is a very large inversion (see fig 2.1). It stretches from division 7C to division 10C on the Pest (Pink Eye Standard) physical map of chromosome 2. The 2Rj inversion is of special interest because it is associated with the Bamako chromosomal form that is very particular in its ecology. It is found in high frequencies in Southern Mali in the banks of River Niger. Bamako is unique in the fact that it breeds in rock pools when the water body recedes. Also interesting is its breeding season which is in the middle of the rainy season peaking in August-September along with Savanna chromosomal form. Therefore studying the 2Rj inversion will provide very useful information on the eco-
geographical behavior of the Bamako chromosomal form. This can, eventually, help decision-takers design a more targeted control strategy.

Here we cloned and characterized the 2Rj inversion breakpoints of the *A. gambiae s.s.*. We analyzed the surrounding sequences for any transposable elements or any repeats. We also looked for gene structures at the breakpoints.
Fig 2.1 The 2Rj inversion along with inversions b, c, u and d on the 2R chromosome, all in standard orientation (Modified from the *Anopheles gambiae* physical map in *Science* 2002)
2.2. Methods

2.2.1. Genomic libraries

High molecular weight DNA was extracted from laboratory reared mosquitoes of the jcu/jcu karyotype (Bamako chromosomal form from Rome) and sent to Lucigen® (http://www.lucigen.com/) for library constructions. Two libraries were made: a plasmid (pSmart LC Kan) library with an average insert size of 12-15kb and a fosmid (pCC1) library with an average insert size of ~40kb. The plasmid library was in a 48µl ligation solution and was kept at -20°C until needed. 4µl of the ligation solution were used to transform competent *Escherichia coli* cells for a 3x genome coverage. The genome size of *Anopheles gambiae* is approximately 278 Mb (Holt et al, 2002). Therefore 23,166 colony forming units (CFU) from transformations with plasmids of about 12 kb insert size corresponded to one genome coverage. 3x genome coverage worth of transformed cells were plated in 24 Petri dishes (diameter: 150mm). After 1hr incubation with shaking (250 rpm) at 37°C. Each plate contained about 3000 CFU. After 24hrs of incubation at 37°C the plates were ready for lifting. The fosmid library was received already arrayed in 120 384-well plates corresponding to 10x genome coverage. Plates were kept at -80°C until needed.
2.2.2. Cytogenetics

A slightly modified protocol from Crampton and Beard (1997) was used as follows. One third, sometimes even one fourth of an ovary, was used for karyotyping as opposed to using the entire ovary. The piece was placed into a drop of 50 % propionic acid deposited on a labeled (e.g.: mosquito ID, collection date, dissection date) microscope slide. The portion of the ovary was put in the solution. They remained immersed at room temperature for about 2-5min or until they had swollen to approximately twice their original size. Another drop of propionic acid was added while the preparation dried out (This gets rid of the fat tissues). Using a dissecting needle the ovaries were spread out and squashed to free the nurse cells. A fresh drop of propionic acid was added and the preparation was covered with a coverslip. The assembly was sandwiched between filter papers (the coverslip side up). The back of a pen or a pencil (with eraser) was used to gently tap on the coverslip to break the follicles and spread the chromosomes. At this point care was taken to not move the coverslip by using a fingertip lined with the filter paper. To flatten the chromosomes the preparation was blotted between two layers of filter paper and firmly but carefully pressed using the palm of the hand (Any movement of the coverslip at this stage may affect the quality of the preparation). The slide was then checked under a phase contrast microscope (Olympus BX40) using 20x, 40x or 100x objectives.

Nota Bene: Preparations can be stored at room temperature, when unsealed, for a few days, longer at -4°C and for months at -20°C. Air infiltration can be prevented by sealing the preparations with small amounts of lactic acid. Care should be taken to avoid the decolorizing effect of the lactic acid.
2.2.3. Genomic DNA isolation and restriction digestion of BAC

Genomic DNA extraction from individual mosquitoes was done using the Qiagen kit (DNeasy®) with slight modifications. The elution volume was cut down to 35µl from initial 50µl in the kit to obtain higher DNA concentration. The bacterial artificial chromosome DNAs (BAC) from *Anopheles gambiae* PEST (2Rj+) and the fosmid DNAs from the Bamako chromosomal form were isolated using either the midiprep kit from Eppendorf (Perfectprep® Plasmid Midi kit) or a standard phenol-chlorophorm isoamyl extraction protocol. To further localize the 2Rj breakpoints in BAC clone 153N7 its DNA was simultaneously digested with Not I and Apa I into smaller fragments. Fragments were visualized on a 0.4 % gel, immediately gel extracted (QIAquick gel extraction kit from Qiagen®). To screen for those still containing the breakpoint sequences the fragments were in situ hybridized to Bamako 2R chromosome.

2.2.4. *In situ* hybridization and polymerase chain reaction

Laboratory reared mosquitoes were bloodfed 2-3 times and let to develop ovaries for 22-26 hours at 27°C and 83 % relative humidity. Half gravid females were knocked out by freezing them at -20°C for 1 min. The ovaries were extracted using forceps and kept in modified Carnoy’s solution (methanol: glacial acetic acid, 3:1) for 24 hrs at room temperature. One third of an ovary was processed on a dust free slide as stated in the cytogenetics section (see section 2.2.2). Good chromosome spreads were fixed by placing the slides on a warmer at 60°C for 1 min. The slides were then placed in a humid chamber containing 2xSSC and kept at 4°C for 12 hours to further flatten the chromosomes and prevent evaporation. The preparations were dipped in liquid nitrogen then immediately placed in -20°C 50 % ethanol and kept at 4°C for 2 hrs. They were then dehydrated by soaking for 5 min in a series of ethanol dilutions: 70 % (room
temperature), cold 90 % (4°C) and 100 % (room temperature). After air drying they were ready for in situ hybridization. 1 µg of BAC DNA or 100 ng of PCR products used as probes were labeled with Cy5-dUTP or Cy3-dUTP (Amersham) using the BioNick™ Labelling System (GibcoBRL). The latter manufacturer’s protocol was followed to perform the in situ hybridization. Preparations were then washed with 0.2xSSC, counterstained with Yoyo-1 (Sigma) and mounted in an antifade solution, DABCO (1, 4-diaza-bicyclo-[2,2,2]-octane). They were placed under a coverslip and observed using a confocal microscope (ref).

In situ hybridization of restriction fragments (from BAC 153N7) to Bamako 2R chromosome indicated a 13 kb piece containing the breakpoint. To isolate the latter primers A (forward): 5’-GTTGACGTTTTGGTCGTTT-3’ and C (reverse): 5’-CTCGGTTTGGGAAAAAGTCA-3’ were used. The PCR product was hybridized to a 2Rj inverted chromosome for confirmation. The PCR reaction contained 1x of PCR buffer 16mM of dNTPs, 3mM of Mg²⁺ 1pmol of each primer 1-2.5 U of Taq polymerase. The volume was topped up to 50 µl with deionized water. The PCR conditions were 94°C for 3 min, 35 cycles of 94°C for 1 min, 60°C for 15 s, 72°C for 4 min, and 72°C for 10 min followed by a 4°C hold.
2.2.5. Library screening

2.2.5.1. Plasmid library

Plating and DNA binding to Nitrocellulose membranes

Colonies (see section 2.2.1. for plasmid library) were transferred onto nylon membranes (137 mm; NitroPure, supported, Osmonic, Inc), denatured, neutralized and baked according to the manufacturer’s recommended protocol. To remove the cell debris for hybridization the membranes were treated with proteinase K (> 600 units/ml, 14-22 mg/ml). The protocol used for this was a modified version of that from Roche®, (http://www.roche-applied-science.com/pack-insert/1699075a.pdf). 0.5ml of a 1:10 ratio proteinase K:2xSSC solution was recommended for a single membrane which would be 50µl (1:10) of proteinase K per membrane. The total number of membranes was then multiplied by 50µl to get the required volume for all the membranes. That volume was then added to 200-250µl of 2xSSC. The membranes were incubated in that solution at 37°C for 1 hr by shaking very gently. Any visible cell debris or clumps were stripped off by hands. Cleaned membranes were washed in 5xSSC/0.5 % SDS/1mM EDTA at 50°C for 30min and ready for pre-hybridization.

Pre-hybridization and Hybridization

In the pre-hybridization step the pre-washed membranes were rolled with meshes (Labnet International: http://www.labnetlink.com/index.cfm). In a pan containing 2xSSC membranes were placed, colonies facing up; interspaced with a mesh. The assembly was rolled together, flipped and placed in a hybridizing bottle (see figure below) containing about 10 ml of 2xSSC. The bottle was horizontally held and turned slowly to unroll the assembly. The 2xSSC was replaced with 10 ml of pre-hybridization solution (6xSSC/0.1 % SDS/5x Denhardt’s pre-warmed
at 45-60°C). The bottle was then placed in an incubator with tube holders and a rotation system (the direction of rotation should not oppose the rolling orientation of the meshe-membrane assembly otherwise the latter will roll again and the membranes surfaces may not get properly soaked). The pre-hybridization was done at 68°C for 2-3 hrs.

While the pre-hybridization was going on the probe (1.3 kb) was prepared by incorporating 50 µci (microcurries) of $^{32}$P dCTP (5 µl) (Amersham) into a PCR mixture containing 50 ng/0.5 µl of template, 1x/5 µl of 10x Taq buffer, 1.5 mM/3 µl of MgCl$_2$, 25 pmol of each of primers A and C (0.5 µl each), 5mM/0.2 µl of dXTP, 1-2.5U/0.2 µl of Taq polymerase and water for total volume of 50 µl. The amplification conditions consisted of an initial denaturation step of 94°C for 4 min, five cycles of 94°C for 1 min, 60°C for 15 s, 72°C for 4 min and a hold at 72°C for 5 min the probe was then purified (Removal of the non-incorporated radioactive material) by using the CENTRI-SEP protocol.
The purified radio-labeled solution was added to the hybridization solution (pre-warmed at 68°C). The pre-hybridization solution was then replaced with 10 ml of the hybridization solution (6xSSC/0.1 % SDS/5x Denhardt’s plus 1 % of Salmon Sperm DNA) and membranes were incubated at 68°C overnight.

**Post hybridization washing**

The membranes were transferred into a plastic dish container containing 300-350 ml of a first wash solution (wash solution 1 = 2xSSC/0.1 % SDS) and incubated at room temperature for 5 min. The wash step was repeated two more times replacing the solution at each time. Following this step was a stringent wash. The wash solution 1 was replaced with 500 ml of wash solution 2 (0.2xSSC/0.1 % SDS) and the membranes were incubated at 68°C for 1 hr. This was repeated one more time with solution replacement. After the second wash the membranes were briefly blotted between two pieces of 3MM Whatmann paper then exposed to films (Kodak Biomax MS) for 24 hrs.
2.2.5.2. Screening the Bamako fosmid library by PCR

The Bamako fosmid library was screened for clones containing the breakpoint sequences by PCR amplification of plate, row and column pools using primers designed off the unique regions of two plasmid clones each spanning a different breakpoint of the 2Rj inversion. Primers fosF1: 5’-AAGCAAGACGCGAGATTGC-3’ and BfosR1: 5’-TTTTCCAAACAATCGGTGTC-3’ were used for the distal breakpoint. Primers BfosF5_2: 5’-TTCCCATTTTCACCGTGTGCCT-3’ and BfosR5_2: 5’-CATTATGCATTTCTGTTGCA-3’ were used for the proximal breakpoint. A single PCR reaction contained 1x of PCR buffer, 8 mM of dNTPs, 1.5 mM of MgCl2, 5 pmol of each primer, 1 µl of the template DNA, 1U of Taq polymerase and water qs for a total volume of 25 µl. The PCR conditions were 94°C for 2 min, 30 cycles of 94°C for 30 s, 50°C for 30 s and 72°C for 30 s followed by a 72°C for 5 min and a 4°C hold.

2.2.6. DNA sequencing

All sequencing reactions were done using the Big Dye Treminator version 3.1 protocol and the 3700 DNA Analyser (ABI Prism®). Whenever the DNA molecule to be sequenced contained repetitive regions the genome priming system (GPS™ -1) was used. The GPS-1 is a Tn7 transposon-based in vitro system that uses TnsABC transposase to insert primers (Transprimer™, New England Biolabs® Inc, version 2.5) into the DNA to be sequenced.
2.2.7. Bioinformatics

The generated sequences were assembled using Seqman software from DNASTAR (DNSTAR, INC.) and analyzed for homology using similarity search tools such as Blastn, Blastx and Blastp at NCBI (http://www.ncbi.nlm.nih.gov/) and at Ensembl (http://www.ensembl.org/). Repeat sequences were identified using alignments tools such as Blast 2 sequences at NCBI, two web-based algorithms, Repeat finder, RepeatMasker and ClustalX. Ab initio gene prediction programs such as FGENESH (http://www softeberry.com), GenScan (http://genes.mit.edu/GENSCAN.html) and GeneID (http://www1.imim.es/geneid.html) were used to predict gene structures. Statistical data were entered and analyzed in SPSS 12.0.
2.3. Results

2.3.1. Cloning and characterization of the 2Rj inversion breakpoints

2.3.1.1. Isolation of the probe spanning the breakpoints

The starting material was a BAC library (pECbAC1) that contained clones that were known to map to the vicinity of the 2Rj+ proximal breakpoint in the Pink Eye STandard strain (PEST) of *Anopheles gambiae*. Candidate clones were hybridized to chromosomes extracted from *Anopheles gambiae* Bamako (Bko) chromosomal form. BAC clone 153N7 revealed signals at both the distal and the proximal breakpoints of the 2Rj (see figure below), indicating that it contains the 2Rj inversion breakpoints.
Fig 2.3 Fluorescent in situ hybridization (FISH) of BAC clone 153N7 to a Bamako 2R chromosomes. Signals in red indicates the distal and the proximal breakpoints of the j inversion.

The BAC clone 153N7 was 124 kb long. To further localize the breakpoint within this large BAC smaller fragments produced through Not I and Apa I double digests or by PCR were purified and used as probes for further FISH to Bko chromosomes. Ultimately the breakpoint was localized within a 1.3 kb fragment that hybridized to both distal and proximal breakpoints of the 2Rj inversion. To isolate the corresponding sequences from the rearranged 2Rj chromosomes this fragment was used as a probe to screen a plasmid genomic library from Bko (Lucigen®) which carries the jcu/jcu arrangement.
2.3.1.2. Screening Bamako plasmid library

The 1.3 kb fragment was labeled with $^{32}$P and used to interrogate nylon membranes containing 3x genome coverage of the Bko chromosomal form. Two positive plasmid clones, Clone1 and Clone5 were isolated and end sequenced. End sequences were Blasted against the *A. gambiae* genome (standard for 2Rj) at Ensembl (http://www.ensembl.org). One end of each clone hit unique sequences on the 10C region of chromosome 2R. The alignment of these ends with the BAC clone 153N7 showed that the similarity between the BAC and one end (clone 1) starts where the similarity between the BAC and the other end (clone 5) starts. In addition the end sequences mapped to opposite sides of the 1.3 kb probe. The two other ends (one for each clone) returned multiple unmapped scaffolds (i.e. likely containing repetitive DNA). After that confirmation the clones were completely sequenced and assembled. The sizes were 11.2 kb and 11.5 kb for clone 1 and clone 5 respectively. Each clone was characterized by a small unique region and a big repeat region. The unique region was inside the inversion for clone 1 and outside the inversion for clone 5 (see fig below). So despite 11kb of sequence neither clone crossed the breakpoint. A library of bigger insert size was then needed in order to cross the breakpoints. In effect a fosmid library (average insert size of ~ 40 kb) was screened.
2.3.1.3. Screening the Bamako fosmid library by PCR

From the unique region of each plasmid clone (blue and green in Fig 2.4), primers (forward and reverse) were designed to screen the fosmid library. Two clones were isolated 59I22 and 35D6. The former was picked up by primers designed off plasmid clone 1 and the latter by primers designed off plasmid clone 5. The two ends of each fosmid clone were sequenced and blasted against the PEST genome. The results showed that the fosmid clones crossed the breakpoints into unique regions (see fig below).

Fig 2.4 Results of the two library screening experiments showing the plasmid and fosmid clones and their positions relative to the 2Rj+ (Pest) and the 2Rj (Bamako) chromosomes. The solid black arrows on the chromosomes indicate the orientation of the j sequences. The long dash-dot arrows show the physical sequences involved in the recombination. *Note: drawing not to scale.*
The results of the blast also showed that the plasmid clones overlapped completely with the corresponding fosmids. That meant that the fosmids also included the plasmid sequences. The distal fosmid clone, 59I22, was entirely sequenced, using the GPS protocol, and assembled. The actual size was 35 kb. Because of the similarity between the plasmid clones we hypothesized that the fosmid clones will also be similar in the repetitive region and therefore we adopted a more targeted sequencing approach, employing the same primer pairs to sequence intervening repetitive DNA within fosmid 35D6. So instead of sequencing the whole fosmid (35D6) we sequenced across the breakpoint and 3-4 kb flanking sequences on each side for a total of ~20 kb covering the proximal breakpoint was recovered. The sequences around the distal breakpoint was indeed very similar to that around the proximal (see fig below in the next section).

2.3.1.4. The structure of the 2Rj breakpoints

The 2Rj inversion is characterized by a 14.6 kb insertion (fig 2.5) present at each breakpoint which is apparently absent from the corresponding locations on 2R (in PEST) as an intact single piece. However it is not entirely absent from the genome as Blastn against the A. gambiae genome at NCBI returns hits to different places including the breakpoints. This showed the repetitive nature of the insert. This Blast also showed that at least a 9 kb piece of the insert is present as a whole piece in the genome though on an unmapped scaffold (gb|AAAB01008935.1). That 14.6 kb consists of two almost perfect inverted repeats of 5.3 kb each separated by a 4 kb region. The two 5.3 kb at each breakpoint showed 99% of identity not only to each other but also to the ones on the other breakpoint (fig 2.6). Reiterative Blasting showed that the breakpoint is made of small fragments of described and non-described transposable elements of both class I and II. Two examples illustrate that; a 169 bp region inside the 4 kb is part of CR1 elements, a non-LTR (non-long terminal repeat) retrotransposon family
described in *Anopheles gambiae*, and a ~ 2 kb of the 5.3 kb returned, from similarity search, hits related to a transposase though the E-values were around 4e-8 (see section 2.3.2.3).

Fig 2.5 Structure of the 2Rj inversion breakpoints.
Fig 2.6 A) Dot plot of 59I22 (Y-axis) and 35D6 (X-axis). Unbroken diagonal line represents 99% sequence identity (in plus/minus orientation) across 14,056 bp. Broken diagonal lines represent inverted repeats. B) Dot plot of 59I22 against itself. Note the broken diagonals, representing the perfect flanking inverted repeats.
2.3.2. Analysis of the surrounding regions of the breakpoints for coding sequences

2.3.2.1. Predicted genes at and around the distal breakpoint of 2Rj

Ab initio gene finding softwares such as Fgenesh, GeneID and GenScan, along with reiterative Blasting were used to predict genes. Only genes predicted by at least two softwares were retained. Near the breakpoint of the inverted sequence two genes were predicted (see figure below). The first, obtained by reiterative Blasting to the Ensembl peptide database, is related to the Ensembl two-exon novel gene ENSANGP00000016927 in the Pest genome (position 3261405-3260284) with 357 amino-acid residues. The Ensembl reciprocal Blast analysis indicates that this gene has putative orthologues in Drosophila melanogaster (CG4553) and in Apis mellifera (ENSAPMG00000017015). Homology searches indicate that this gene may be a putative adenine triphosphate binding cassette transporter (ABC transporter) nucleotide binding domain. ABC transporters constitute a large family of proteins involved in the transport of compounds such as sugars, ions, peptides and more complex organic molecules. (Schmitt and Tampe, 2002). It lies 400 bp upstream of the 2Rj breakpoint

At about 7 kb from the 2Rj breakpoint (in region 3/division 10C) lies the second predicted gene which was obtained by the softwares cited above. It is related to a protein kinase specifically a serine-threonin kinase, catalytic domain. The gene contains 2 exons and is about 1.4 kb at Ensembl although 4 were predicted by ab initio software. This gene corresponds to the Ensembl gene ENSANGP00000007766 (Blastx score = 365 with E-value = 2e-99) in the PEST strain. The Blastp also showed a conserved domain (score = 127, E-value = 4e-30). Serine/threonine kinases have been shown to be involved in ATP binding. They also play an
important role for the catalytic activity of the enzymes. It is similar to the *Drosophila melanogaster* gene CG32742.

Fig 2.7 Predicted genes on the fosmid clone 59I22 from the distal breakpoint of the Bamako chromosomal form (figure not drawn to scale). Thick colored solid bars indicate exons. Thin solid bars indicate introns.
2.3.2.2. Predicted genes at and around the proximal breakpoint of 2Rj

Using the same methods as above gene prediction was undertaken with the 2Rj proximal breakpoint sequences. Two genes were predicted (see Fig below). One of the predicted genes (in region 4, division 10C) appears to be related to a zinc finger of type CCCH with G patch domain, a glycine rich nucleic binding protein and is present in the PEST genome as ENSANGP00000017646. The G-patch, also called D111 domain, is a short conserved region of about 40 amino acids which occurs in a number of putative RNA-binding proteins, including tumor suppressor and DNA-damage-repair proteins. This suggest that it has an RNA binding function (Aravind and Koonin, 1999). This domain has seven highly conserved glycines, hence its name. The putative G-patch lies 868 bp downstream of the breakpoint in region 4 of the 2Rj and has 8 exons predicted in contrast to 2 predicted by Ensembl. Upstream of the breakpoint in region2 of 2Rj, sequences overlap with the untranslated region of a dynein light chain type I (Fig 2.8). Dynein light chain type I belongs to the dynein light chain intermediate family. It might serve in dynein binding to membranous organelles or chromosome; or it probably regulates enzyme activities in association with the heavy chains in the dynein head. Dynein is a large multimeric complex with ATPase activity, responsible for the movement of eukaryotic cilia and flagella (axonemal dynein) and for the intracellular retrograde motility of vesicles, organelles and chromosomes along microtubules (cytosolic dynein) (Ilangoivan et al., 2005) (also see UniProt at http://ca.expasy.org/cgi-bin/get-entries?KW=Dynein).

Apparently the 2Rj rearrangement interrupts no known or predicted transcripts.
2.3.2.3. Predicted genes inside the 14.6kb insert

The analysis of the 14.6 kb inserts involved multiple steps. First smaller fragments were blasted (blastn) against nr at NCBI. Then if any was found to be repetitive or to hit a coding region the top 10-20 top partial hits were selected along with 500 bp to 1 kb flanking sequences. Multiple alignments were then preformed using Clustal X to determine the full length of the repeat or the gene. Ab initio software were also used to predict gene structures within the insert. The results showed a fragment similar to a transposase in the 5.3 kb repeat (see Fig 2.9). Blastx of the fragment at NCBI against nr returned hits from which the top ones were either not annotated or hypothetical proteins. However the fragment showed 28 % of amino-acid similarity with transposae from Neisseria meningitidis with an E-value of 4e-08 though the scores were relatively low, 64.7 (see Fig 2.9). The best hits corresponded to the Ensembl predicted gene...
ENSANGP00000026120 in the PEST strain. This putative transposase, if intact, would yield a single exon with 295 aa residues (ab initio prediction). However the similarity only extends to a fragment of this transposase suggesting it (transposase) may not be active on the repeat.

In the 4 kb middle section results showed that a 169 bp piece is similar to CR1 TE families (see Fig 2.10). This piece showed 98 % nucleotide similarity with CR1 elements. CR1 elements are non-LTR retrotransposons mainly characterized by two ORFs, one corresponding to a nucleic binding domain and the other coding for both an apurinic/apyrimidic endonuclease domain and a reverse transcriptase (RT) (see Biedler and Tu, 2003). Blastx of the reconstructed TE against nr at NCBI showed similarity to reverse transcriptase (see Fig 2.11).

Another feature of this repeat is its AT richness, about 56 %. Interestingly the middle section (4 kb) is also AT rich, about 69 %.
### Fig 2.9 Blastx results (NCBI) using 5.3 kb as query.
gamb = *gambiae*; Caeno = *Caenorhabditis*

<table>
<thead>
<tr>
<th>Sequences producing significant alignments:</th>
<th>Hit/organisms</th>
<th>Score (Bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>55249671</td>
<td>gb</td>
<td>EAL40575.1</td>
</tr>
<tr>
<td>gi</td>
<td>55235819</td>
<td>gb</td>
<td>EAL31875.1</td>
</tr>
<tr>
<td>gi</td>
<td>55246635</td>
<td>gb</td>
<td>EAL42119.1</td>
</tr>
<tr>
<td>gi</td>
<td>55247118</td>
<td>gb</td>
<td>EAL43206.1</td>
</tr>
<tr>
<td>gi</td>
<td>55245431</td>
<td>gb</td>
<td>KAA04841.3</td>
</tr>
<tr>
<td>gi</td>
<td>55237921</td>
<td>emb</td>
<td>CAE74945.1</td>
</tr>
<tr>
<td>gi</td>
<td>55246085</td>
<td>gb</td>
<td>EAL09999.2</td>
</tr>
<tr>
<td>gi</td>
<td>55243503</td>
<td>gb</td>
<td>EAL42158.1</td>
</tr>
<tr>
<td>gi</td>
<td>55229501</td>
<td>gb</td>
<td>EAL10638.2</td>
</tr>
<tr>
<td>gi</td>
<td>55251355</td>
<td>emb</td>
<td>CAE80001.1</td>
</tr>
<tr>
<td>gi</td>
<td>55592861</td>
<td>emb</td>
<td>CAE71697.1</td>
</tr>
<tr>
<td>gi</td>
<td>55591385</td>
<td>emb</td>
<td>CAE71161.1</td>
</tr>
<tr>
<td>gi</td>
<td>55590191</td>
<td>emb</td>
<td>CAE7341.1</td>
</tr>
<tr>
<td>gi</td>
<td>55591651</td>
<td>emb</td>
<td>CAE71514.1</td>
</tr>
<tr>
<td>gi</td>
<td>55297957</td>
<td>emb</td>
<td>CAE76144.1</td>
</tr>
<tr>
<td>gi</td>
<td>55256961</td>
<td>gb</td>
<td>EAL34467.1</td>
</tr>
<tr>
<td>gi</td>
<td>55580382</td>
<td>emb</td>
<td>CAE71742.1</td>
</tr>
<tr>
<td>gi</td>
<td>55260804</td>
<td>emb</td>
<td>CAI2120.11</td>
</tr>
<tr>
<td>gi</td>
<td>55264110</td>
<td>ref</td>
<td>XP 696050.1</td>
</tr>
<tr>
<td>gi</td>
<td>55267784</td>
<td>emb</td>
<td>CAE57092.1</td>
</tr>
<tr>
<td>gi</td>
<td>55550131</td>
<td>emb</td>
<td>CAE71741.1</td>
</tr>
<tr>
<td>gi</td>
<td>55583527</td>
<td>emb</td>
<td>CAE73985.1</td>
</tr>
<tr>
<td>gi</td>
<td>55258572</td>
<td>emb</td>
<td>CAE53987.1</td>
</tr>
<tr>
<td>gi</td>
<td>55584463</td>
<td>emb</td>
<td>CAE73518.1</td>
</tr>
<tr>
<td>gi</td>
<td>55585351</td>
<td>emb</td>
<td>CAE73518.1</td>
</tr>
<tr>
<td>gi</td>
<td>55585364</td>
<td>emb</td>
<td>CAE74021.1</td>
</tr>
<tr>
<td>gi</td>
<td>55584969</td>
<td>emb</td>
<td>CAE84443.1</td>
</tr>
<tr>
<td>gi</td>
<td>55335179</td>
<td>gb</td>
<td>AFF44038.3</td>
</tr>
<tr>
<td>gi</td>
<td>5533706</td>
<td>emb</td>
<td>CAI41166.1</td>
</tr>
<tr>
<td>gi</td>
<td>55270324</td>
<td>gb</td>
<td>EAN78470.1</td>
</tr>
<tr>
<td>gi</td>
<td>55252904</td>
<td>ref</td>
<td>EF 802182.1</td>
</tr>
<tr>
<td>gi</td>
<td>57423034</td>
<td>gb</td>
<td>AB045394.1</td>
</tr>
<tr>
<td>gi</td>
<td>55362461</td>
<td>ref</td>
<td>EF 0022541.2</td>
</tr>
<tr>
<td>gi</td>
<td>55259240</td>
<td>gb</td>
<td>EAL40577.1</td>
</tr>
<tr>
<td>gi</td>
<td>55250256</td>
<td>gb</td>
<td>AAD14006.1</td>
</tr>
<tr>
<td>gi</td>
<td>55258359</td>
<td>gb</td>
<td>EAL13208.1</td>
</tr>
<tr>
<td>gi</td>
<td>55249342</td>
<td>gb</td>
<td>AJN52837.1</td>
</tr>
<tr>
<td>gi</td>
<td>55229361</td>
<td>ref</td>
<td>NF 009203.1</td>
</tr>
<tr>
<td>gi</td>
<td>553931</td>
<td>emb</td>
<td>CAI42428.1</td>
</tr>
</tbody>
</table>
Fig 2.10 The 14.6 kb insert with the 169 bp sequence similar to CR1 sequences in red present in the 4 kb middle region.
<table>
<thead>
<tr>
<th>Sequences producing significant alignments:</th>
<th>Hit/oranisms</th>
<th>Score (Bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>5538611</td>
<td>gb</td>
<td>EAL9926.1</td>
</tr>
<tr>
<td>gi</td>
<td>100135</td>
<td>ps</td>
<td>EAL3751</td>
</tr>
<tr>
<td>gi</td>
<td>132431</td>
<td>gb</td>
<td>AA58938.1</td>
</tr>
<tr>
<td>gi</td>
<td>5523949</td>
<td>gi</td>
<td>EAL9465.1</td>
</tr>
<tr>
<td>gi</td>
<td>55234291</td>
<td>gb</td>
<td>EAL1058.1</td>
</tr>
<tr>
<td>gi</td>
<td>17490719</td>
<td>gb</td>
<td>AA25248.1</td>
</tr>
<tr>
<td>gi</td>
<td>55230250</td>
<td>gb</td>
<td>EAL9986.1</td>
</tr>
<tr>
<td>gi</td>
<td>52008493</td>
<td>ref</td>
<td>XF 78507.1</td>
</tr>
<tr>
<td>gi</td>
<td>6762570</td>
<td>gi</td>
<td>CAJ0239.1</td>
</tr>
<tr>
<td>gi</td>
<td>55235501</td>
<td>gb</td>
<td>EAL9358.1</td>
</tr>
<tr>
<td>gi</td>
<td>552442461</td>
<td>gb</td>
<td>EAL8229.1</td>
</tr>
<tr>
<td>gi</td>
<td>52638665</td>
<td>gb</td>
<td>AAO2626.1</td>
</tr>
<tr>
<td>gi</td>
<td>52637891</td>
<td>ref</td>
<td>XF 79847.1</td>
</tr>
<tr>
<td>gi</td>
<td>7215503</td>
<td>ref</td>
<td>XF 79284.1</td>
</tr>
<tr>
<td>gi</td>
<td>17529698</td>
<td>gb</td>
<td>EAL40415.1</td>
</tr>
<tr>
<td>gi</td>
<td>50248081</td>
<td>emb</td>
<td>CAE53561.1</td>
</tr>
<tr>
<td>gi</td>
<td>6762570</td>
<td>gb</td>
<td>EAL9358.1</td>
</tr>
<tr>
<td>gi</td>
<td>5705997</td>
<td>ps</td>
<td>T3706</td>
</tr>
<tr>
<td>gi</td>
<td>5707973</td>
<td>ps</td>
<td>T3505</td>
</tr>
<tr>
<td>gi</td>
<td>55236451</td>
<td>gb</td>
<td>EAL9307.1</td>
</tr>
<tr>
<td>gi</td>
<td>17215561</td>
<td>ref</td>
<td>XF 795504.1</td>
</tr>
<tr>
<td>gi</td>
<td>52637891</td>
<td>ref</td>
<td>XF 79847.1</td>
</tr>
<tr>
<td>gi</td>
<td>52638665</td>
<td>ref</td>
<td>XF 79847.1</td>
</tr>
<tr>
<td>gi</td>
<td>6762570</td>
<td>gi</td>
<td>CAJ0239.1</td>
</tr>
</tbody>
</table>

Fig 2.11  Blastx of ~ 6kb (containing the 169 bp) against nr at NCBI
2.4. Discussion and conclusions

This is the first time the breakpoints of a major inversion involved in the incipient speciation process of *Anopheles gambiae* s.s. have been cloned and sequenced. The analysis of the two fosmid clones revealed the molecular structure of the 2Rj breakpoints. They are not the product of a simple two-breaks and a flip event as inversions are very often defined. In addition to the break-flip events they are also characterized by a large insertion, 14.6 kb, made of almost perfect inverted repeats separated by a middle section. It has been demonstrated that TEs are associated with inversion breakpoints (Càceres et al., 2001; Càceres et al., 1999; Ladevèze et al., 1998; Mathiopoulos et al., 1998; Mathiopoulos et al., 1999). However every inversion is not necessarily associated with a TE (Cirera et al., 1995; Wesley C.S. and Eanes W.F., 1994). We found no complete class I or II transposable element associated with the breakpoints though the analysis of the 5.3 kb returns a putative transposase and the breakpoint regions are made of several pieces of DNA related to transposable elements from both class I and II. This is illustrated by the sequences in the 5.3 kb and the 169 bp in the middle section of the 14.6 kb which appears to be related to CR1, a non-LTR retrotransposon (Biedler and Tu, 2003). Nevertheless, the reiteration of an almost 15 kb insertion on both sides of the chromosomal rearrangement strongly suggests its direct involvement in generating the inversion.

The 14.6 kb insert, taken as a whole, has a structure similar to fold back transposable elements (FTs). FTs are elements of variable size, generally made of two inverted terminal repeats (IR or limbs or FB for fold back) that may be different in length, and a middle sequence that also varies between elements. Some present internal tandem repeats contiguous to the IR sequences. Specifically our insert’s structure is more similar to that of the *Newton* elements (Casals et al, 2005) or the type 3 fold back elements described by Rebatchouk and Jonathon
In the 14.6 kb from Bamako the limbs have the same size of 5.3 kb and are separated by a middle section. The IRs may contain a few internal repeats and are AT rich (i.e. 75%) (Casals et al 2005). The 5.3 IRs in our case are 56% AT rich each. Interestingly the 4 kb middle section is also AT rich. It’s been reported that the middle sections can be mobilized by the FBs (Paro et al, 1983). However it’s also been shown to have a coding potential suggesting its probable role in the elements’ transposition (Harden and Ashburner, 1990). No coding motif has been identified in the 4 kb middle section of the 14.6 insert which is in contrast with the Drosophila melanogaster FB-NOF elements whose 4 kb middle section codes for at least one protein (unknown function) (Harden and Ashburner, 1990). However sequences in the IRs have shown similarities to transposases. This suggests that the inverted repeats may, in certain cases, have coding potential but does not infer any activities of transposition. This is more likely the mechanism involved in the transposition of the insert here. This mechanism was suggested by Paro et al (1983) to argue the transposition mechanism of an FB element. Though our data show no evidence of a transposition mechanism for the insert its ability to move is supported by its presence in identical copies in, at least, two different genomic locations. Another proposition of transposition mechanism for FB element consists of using other elements’ machinery however elements have moved after having lost the foreign DNA they are associated with (Paro et al, 1983). The presence of only 169 bp sequences similar to CR1 elements may indicate that, if such an element was early associated with the insert, it has been lost over time. Therefore it seems unlikely that the insert has recently used the CR1’ machinery to move. While their transposition mechanism remains unknown, FB elements, as well as other TEs, have shown high frequencies of insertions in inversion breakpoints (Casals et al, 2005). This in silico analysis of
sequences needs to be confirmed by biological experiments such as in situ hybridization or Southern analysis which will also give an estimate of the copy numbers of the putative element.

The presence, at both 2Rj breakpoints, of identical copies of the 14.6 kb, suggests a relative recent origin of the j inversion. This is supported by: 1) its geographical distribution; so far it’s been reported from southern Mali and northern Guinea Conakry only; and 2) the precision of the inversion breaks. Inversions are known to be generated by ectopic recombination between inverted copies of TE inserted at each breakpoint (Gray, 2000; Caceres et al, 1999; Ladevèze et al, 1998). Our data clearly indicate the association of the insert with the breakpoints but its direct implication in the generation of the 2Rj inversion can not be inferred from this study. Therefore more experiments should be conducted in order to determine the generation mechanism of the 2Rj inversion.

Besides their association with TEs inversion breakpoints are also known to be interrupting coding sequences (Matzkin et al, 2005, Sharakov et al, 2006 in press). Homolog genes from the PEST strain have been found close to each breakpoint but no evidence of interrupted transcript was found. The G-patch domain and the non-annotated gene, ENGSANGP00000016927 lies 794 bp and 400 bp respectively from the respective breakpoints. Though we used ab initio gene prediction tools there is cDNA evidence (see http://www.ensembl.org) these two genes in PEST. Because they are close to the breakpoints sequences of these genes may lead to evolutionary important information. In effect their sequences can eventually be isolated from non j carriers and other j carriers and analyzed to infer the relative age of the inversion. This will also shed light on the monophyletic origin of the 2Rj. After these genes have been confirmed from actual
biological experimentations, their expression levels could be determined in order to assess the impact of the inversion.

The characterization of the 2Rj inversion breakpoints is not only important for understanding the genesis of inversions but also for the fact that they provide information on an inversion that is associated with a chromosomal form, Bamako, which has a particular eco-geographical distribution pattern. Bamako breeds in rock pools when the water front recedes. So this study could provide tools to study questions related to the molecular basis of the ecological behavior of the Bamako chromosomal form of the *Anopheles gambiae s.s.*. Finally the breakpoint sequences could also be used to design a PCR-based assay to diagnose the 2Rj inversion just as in previous studies with different inversions in different organisms (Mathiopoulos et al., 1998; Matzkin et al., 2005).
MOLECULAR DIAGNOSTIC OF THE 2RJ INVERSION

3.1. Introduction

The use of molecular techniques has revolutionized the field of malaria vector control. rDNA-based PCRs have been used to distinguish between sibling taxa of the *Anopheles gambiae* complex (Fanello et al., 2002; Scott et al., 1993) and between chromosomal forms (Fanello et al., 2002; Favia et al., 2001). Even though molecular typing is not yet perfect at the chromosomal form level it still has some advantages over cytogenetics that remains the gold standard for the identification of the *A. gambiae* members. Cytogenetic based identification consists of recording each individual inversion within a combination that is associated with a sibling species or a chromosomal form. This can be very tedious, time consuming and limit the sample size as half gravid females only are processed. Because inversions play a central role in the speciation and ecological adaptation of the *A. gambiae* members it is necessary to design molecular methods to diagnose them in order to overcome the weaknesses of cytogenetics. Very few studies have developed PCR-based assays to identify inversions. In *Drosophila* an assay was designed for inversion *In(3R) Payne* (Matzkin et al., 2005). In *Anopheles arabiensis* an assay was designed to diagnose inversion 2Rd’(Mathiopoulos et al., 1998). So far no inversion in *Anopheles gambiae s.s.* is diagnosed with molecular means (but 2La and 2Rb,c and u are under investigation).
The 2Rj inversion homozygote, in combination with 2Rb, c and u, is associated with the Bamako chromosomal form, one of the three chromosomal forms in Mali. Studying the biology of this inversion might lead to the understanding of the latter chromosomal form’s ecological behavior. One of the initial steps for such investigation consists of getting the breakpoints sequence information of the inversion under study. The inversion’s breakpoints have been cloned and characterized (see chapter 2). Here we propose to design a simple PCR-based assay to diagnose the 2Rj inversion in natural populations.
3.2. Methods

3.2.1. Study sites

Field mosquitoes used for this study were collected from six villages in Mali, namely Bancoumana and Kela in the South, Banambani, Moribabougou and N’Gakororo in the center, and Douna and Fanzana in the South East (see fig 3.1 in green).

![Fig 3.1 Study sites](image-url)
3.2.2. Mosquito stock and genomic DNA extraction

The specimens were collected by spray catch and sorted morphologically to *Anopheles gambiae s.l.* and by repletion stage. Only the half gravid females were selected and dissected. The ovaries were pulled and kept in labeled individual micro-tubes in modified Carnoy solution (1:3 glacial acetic acid : 100 % ethanol) for 24h at room temperature then placed at -70°C – 80°C until needed. The head/thorax was kept on silicagel in individual 0.5 ml tubes with labels matching that of the tubes containing the corresponding ovaries. The labeling consisted of the village code (e.g. bc for Bancoumana), an identification number and the collection date. DNA extracted from head/thorax was subjected to identification by an rDNA PCR assay to determine which species of *A. gambiae* (Only *A. gambiae* were analyzed further) and the molecular forms (Fanello et al, 2002, Favia et al, 2001).

3.2.3. Karyotyping

Ovaries from individual mosquitoes were squashed in 50 % propionic acid on a dust free microscopic slide to free the nurse cells. Cells were covered with a coverslip then gently tapped to break the follicles, and spread and flatten the chromosomes. The preparation was finally checked under a phase contrast microscope (Olympus BX40) to determine the karyotype (See chapter 2, section 2.2.2).
3.2.4. Designing primers and Polymerase Chain Reaction

Flanking regions of the 2Rj breakpoints were aligned with the corresponding sequences from the PEST genome. Two PCR primers oriented across region 1-2, named Rgn2U-F (5’-GGCGGATTCTAGCAATGTC-3’) and Rgn1-R (5’-CGGCCAAGGTGCTTACATAG-3’), were designed to reliably detect the uninverted 2Rj+ arrangement. These primers produced a 458 bp product. However, because of the 14.6 kb insertion, it was physically impossible to design a PCR assay that could cross the entire breakpoint to detect the 2Rj arrangement. A third primer was then designed to be 2Rj-specific and anneal to region 2 within ~100 bp of the breakpoint. That primer, named Rgn2J_R (5’-CGAAGCGTTAAGTCTCATCG-3’), produced a 186 bp product in conjunction with Rgn2U_F (see fig 3.2 A). Each PCR reaction contained 1x of PCR buffer, 8 mM of dNTPs, 1.5 mM of Mg²⁺, 2.5 pmol of primers Rgn2U_F and Rgn1_R, 1 pmol of Rgn2J_R, 17-20 ng of template, 1.25 U of Taq polymerase and deionized water qs for a total volume of 25 µl. The template was then amplified under the following PCR conditions: 94°C for 2 min, 40 cycles of 94°C for 20 s, 60°C for 15 s, 72°C for 20 s, and 72°C for 5 min followed by a hold at 4°C.
3.3. Results

Validation of the PCR

A triplex PCR strategy with primers Rgn2U_F, Rgn2J_R and Rgn1_R was designed (see fig 3.2-A) and set up. The results showed excellent correspondence between karyotypes and molecular types for the 2Rj (see Fig 3.2-B).
Fig 3.2 A) Illustration of PCR assay primer location sites (figure not drawn to scale). B) Molecular PCR diagnostic of 2Rj+ (standard orientation) and 2Rj (inverted orientation). Specimens shown in this figure are from different villages. The two letters at the beginning of the identification numbers indicate the villages where the mosquitoes were caught; bn = bamanabani, mb = Moribabougou, ng = N’Gakororo, dn = Douna, kl = Kela, fz = Fanzana. The outermost lane on either side contain the DNA reference marker (100 bp). Lanes 2-5 are karyotyped 2Rj+ samples from different villages. Lanes 6-9 are karyotyped 2Rj samples from different villages. Lane 10 is an actual heterozygote individual. Lane 11 is an artificial heterozygote (1:1 ratio). In lane 12 is amplification from the Bamako fosmid clone 35D6 that crossed the proximal breakpoints (used as positive control for the inverted samples). In lane 13 is a laboratory strain, 4arr, known to be standard for 2R (used as positive control for the standard samples). In lane 14 is the negative control for the PCR reactions. Included above are the respective karyotypes for each specimen.
To validate the 2Rj diagnostic assay 272 field collected specimens were fully karyotyped and tested. The 2Rj inverted and standard carriers represented 28 % and 72 % respectively. The sensitivity, which is defined as the ability to yield a PCR product, was 92 % for the 2Rj and 95 % for the 2Rj+ arrangement. The specificity, here defined as the correspondence between the karyotypes and the molecular types, was 98.5 % for 2Rj and 97.3 % for 2Rj+. The global specificity, 2Rj and 2Rj+ together, was 96.5 % (see Table 3.1). Only a single heterozygote 2Rj/j+ has been encountered and the PCR assay identified it (see fig above).

TABLE 3.1

SENSITIVITY AND SPECIFICITY OF THE 2RJ DIAGNOSTIC ASSAY

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2Rj/j</td>
<td>76</td>
<td>70 (92 %)</td>
<td>69 (98.5 %)</td>
</tr>
<tr>
<td>2Rj+/j+</td>
<td>196</td>
<td>186 (95 %)</td>
<td>181 (97.3 %)</td>
</tr>
<tr>
<td>Total</td>
<td>272</td>
<td>256 (94.1 %)</td>
<td>247 (96.5 %)</td>
</tr>
</tbody>
</table>

Sensitivity is the ability to yield a product regardless of the size. Specificity is the correspondence of molecular types with the respective karyotypes.

Table 3.2 shows the sensitivity and the specificity of the 2Rj diagnostic assay across different geographical areas in Mali. Specimens were from seven villages from Mali. The specificity varies from 97 % to 100 % for the 2Rj/j and from 89 % to 100 % for the 2Rj+/j+.
TABLE 3.2

SENSITIVITY AND THE SPECIFICITY OF THE 2Rj DIAGNOSTIC ASSAY ACROSS DIFFERENT GEOGRAPHICAL AREAS IN MALI.

<table>
<thead>
<tr>
<th></th>
<th>Kela</th>
<th>Bancoumana</th>
<th>Douna</th>
<th>Moribabougou</th>
<th>N’Gabakoro</th>
<th>Banambani</th>
<th>Fanzana</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>35</td>
<td>31</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Sn</td>
<td>31</td>
<td>31</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Sp</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>35</td>
<td>31</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Sn</td>
<td>31</td>
<td>31</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Sp</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>35</td>
<td>31</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Sn</td>
<td>31</td>
<td>31</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Sp</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>35</td>
<td>31</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Sn</td>
<td>31</td>
<td>31</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Sp</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

T = total
Sn = sensitivity
Sp = specificity
3.4. Discussion

The presence of the 14.6 kb (see chapter 2) insert at the breakpoints was unexpected. This has however made it impossible to design primers across the breakpoints that amplify reliable PCR products. In this approach we exploited the positions of regions 1 and 2 in standard (division 7C) relative to the inverted chromosomes to design the diagnostic primers. This is responsible for the exhibited success rate of the assay. The few instances where the molecular types have not corresponded to the respective karyotypes can be explained by several factors. Firstly a contamination is likely to happen during the genomic DNA extraction as another individual’s leg or wing can inadvertently be carried over in the wrong tube. Secondly karyotyping errors can occur. A third explanation may be DNA polymorphism at the primer binding site in natural populations. A fourth explanation is very rare but possible recombination or gene conversion between inverted and standard arrangements near the breakpoints. A possible fifth explanation could be primer mis-annealing: i.e. the 2Rj specific primer may be able anneal to 2Rj+ if conditions are not sufficiently stringent. The results show the robustness of the test even though many natural heterozygote specimens have not been tested. This is due to the very low frequency of such karyotype in both our samples and natural populations. The assay was applied to specimens collected from different locations in Mali, West Africa. The successful amplification of the expected products for each karyotyped specimen from the different locations indicates the high reproducibility of the test across broad geographical areas. It will be very informative to test samples collected beyond the Malian borders to determine the range of applicability of the assay across larger geographical regions. Similar assays diagnosing inversions have been
designed in *Drosophila melanogaster* (Matzkin et al, 2005), in *A. arabiensis* (Mathiopoulos et al, 1998) and in *A. gambiae* as well (Sharakov et al, 2006, in press) but this is the first time an inversion associated with a chromosomal form (Bamako) and present in only the S-molecular form (Della Torre, 2005) has been diagnosed by molecular means. The development of such a tool has a biological and epidemiological implication in the control of malaria. The biological implication involves better understanding of the 2Rj inversion’s behavior in natural populations which could also lead to more information about the biology of the Bamako chromosomal form. From an epidemiological perspective, by circumventing some of the limiting factors of current cytogenetic based diagnostic molecular PCR karyotyping may lead to an easy and quick identification of vectors in their microhabitats leading to a targeted control strategy when the inversion is strongly associated with a population such as the 2Rj is with the Bamako chromosomal form.
LITERATURE CITED


