Population structure and genome-wide association in the malaria vectors Anopheles gambiae and Anopheles coluzzii

Seth Redmond

Université Pierre et Marie Curie

Complexité des Vivants

Laboratoire Ken Vernick, Institut Pasteur

Population structure and genome-wide association in the malaria vectors *Anopheles gambiae* and *Anopheles coluzzii*

Par Seth Redmond

Thèse de doctorat de Biologie

Dirigée par Ken Vernick

Présentée et soutenue publiquement le 23/02/2015

Devant un jury composé de :

Prof Dominique Higuet	Professeur, Paris VI / UPMC	President
Prof George Christophides	Professeur, Imperial College	Rapporteur
Dr David Weetman	Maître de Conferences, Liverpool School of Tropical Medicine	Rapporteur

Contents

Introduction

Despite reductions in mortality in recent years, malaria remains one of the worlds biggest killers. There were more than 200 million cases of malaria in 2012 (WHO 2013) and more than 600,000 deaths – over two thirds of these children. Over half of the world's population remains at risk of malaria.

The cost in terms of development is significant. In countries with the highest levels of malaria transmission, an estimated 40% of public health budgets is spent on malaria prevention (WHO 2013). Other costs, such as a lack of social development and lost productivity due to illness, are harder to quantify. Although the evidence linking poverty and incidence of malaria within specific regions is mixed (Worrall et al. 2005), there is a well-defined link between malaria incidence and economic retardation (Gallup & Sachs 2001).

Much has been achieved in recent years through the combined use of antimalarials – in particular through the use of artemesinin, both alone and in combination therapies – and insecticide applications such as insecticide-treated bednets, and incidences are falling. However it is clear that this approach alone, as it is currently implemented, can control, but cannot eliminate the disease in the areas of highest transmission. The emergence of resistance to artemesinin in south-east asia (Dondorp et al. 2010) raises fears of the removal of one of the most powerful tools in the antimalarial arsenal. Bednet coverage is also subject to concerns over its poor implementation in many regions and evasion of this control by diverse parasite species. This reinforces the need for a diverse approach to be taken to Malaria control. Improved control of the mosquito vector will be vital.

Control of the transmission vector has, in fact, been the principal method of malaria control ever since Ronald Ross' discovery of transmission by mosquitos, and despite the diversification of methods used today, it remains a cornerstone of the fight against the disease. An estimated 200,000 insecticide-treated bed-nets (ITBNs) will be distributed throughout the course of 2014 – a near tripling of the number disbursed only two years earlier (WHO 2013). Indoor residual spraying is also widely employed to control transmission, with an estimated 135 million people protected by this method in 2012. Indeed all of the countries to have eliminated malaria have either mainly or wholly based their efforts on vector control.

During the course of this thesis I will attempt to show some of the remaining issues with antimalarial interventions. In particular how vector control has failed in the past, and how behavioural heterogeneity and genetic diversity can impair these interventions. I will also demonstrate how novel applications of these basic technologies, or how novel interventions such as population scale genetic interventions can help in the future. During the course of this I will attempt to show the major scientific questions that will need to be answered on the pathway to elimination, in particular the detection of population structure and the phenotypic typing of wild populations that can inform future work.

I will present a review of the current knowledge of the innate immune barriers presented to the mosquito. That is, those genes that might be affected by genetic interventions and the context for the genetic mapping we have undertaken.
This will be followed by a description of the current knowledge of *Anopheles gambiae* / *coluzzii* population structure. How the vector population divides along species and sub-species lines against a background of considerable ecological and genetic diversity.

The first experiment I will show relates to the population structure. Using two large genomic datasets I have identified highly accurate genetic barcodes for two chromosomal inversions. These structures show patterns of association in the wild with particular ecological contexts, and are innately related to ecotypification and speciation events. They are therefore crucial markers for phenotypes that can affect control implementation and can impair our ability to perform genomic mapping (such as that presented in chapter 5).

The principal experiment I have undertaken is a genomic mapping of innate immune factors relating to *Plasmodium falciparum* transmission in the wild. This has been performed using mosquito colonies that are genetically analogous to wild populations. But are controlled for population structure to avoid many of the potential mapping errors that are common in association studies. I have developed a new method of genomic mapping by loss-of-heterozygosity and have used this to identify a novel locus containing two immune genes. These genes, despite having never previously been ascribed an immune function, are amongst the cohort of potential genes described in Chapter 2.

Finally, I will show how the results of these studies, and perhaps more importantly the novel techniques that have been developed, can provide a pathway for future studies of the vector. How the could, in the future, illuminate differences in phenotype and behaviour between structured populations in the field; differences that can affect vector competence and, ultimately, our ability to apply vector control.

Chapter 1: A Brief History of Malaria 'Eradication'

Following the often spectacular successes of eradication efforts in Europe at the national level (Livadas 1952), the years following world war two saw the first concerted global campaign against malaria. The Global Malaria Eradication Program (GMEP), launched in 1955, relied largely on a single 'wonder-insecticide' DDT (dichloro-diphenyl-trichloroethane) delivered above all via indoor residual spray (Brown 2002; Nájera et al. 2011), alongside its widespread use in agricultural pest control. They were large 'vertical' campaigns, with little knowledge of the precise epidemiology of individual regions. These campaigns were undoubtedly highly successful in eradicating malaria in regions of low to medium transmission. However, despite an initial reduction in morbidity, the long term effects on the areas of highest transmission were negligible, with significant rebounds of malaria in the 1960s and 70s (Nájera et al. 2011). As a result of numerous factors, including increasing insecticide resistance, a lack of economic support amidst a number of global financial crises, and political difficulties, the program was eventually abandoned in 1975 and global efforts were scaled back to control rather than eliminate the parasite (Nájera et al. 2011).

Control efforts in the intervening years concentrated on post-infection treatment rather than vector control, and were characterised by a gradual rise in malaria mortality as resistance to antimalarials (chloroquine in particular) began to rise (White et al. 1999).

In recent years, and particularly since the 2007 Gates Malaria Forum, eradication has once again been back on the agenda (Roberts & Enserink 2007). The Roll Back Malaria initiative (launched by WHO in 1998) aims to eliminate malaria "as a global health burden", before proceeding to the ultimate goal of eradicating the parasite entirely (RollBackMalaria 2014).

In contrast to the GMEP there is greater recognition in today's eradication efforts that the precise dynamics of malaria transmission in program areas should be much better understood; the program needs to be "*a synchronous global effort, locally adapted in all endemic areas*" (Tanner & Savigny n.d.). Although there are some striking similarities between the two programs, it is hoped that through more precise surveillance of both insecticide and antimalarial resistance, as well as the use of a wider variety of control methods tailored to specific regions, the outcomes of the current program will be more successful.

To this end there has been recognition that a directed research should go hand-in-hand with any public health program. The Malaria Eradication Research Agenda (malERA) initiative, established by a consortium of malariologists and endorsed by the WHO's Roll Back Malaria program has highlighted a number of areas in which current research is lacking. These

include the development of novel insecticides, a better understanding of the vector biology, and efforts to understand and control transmission resulting from outdoor biting (Alonso et al. 2011).
A poorly-directed eradication campaign is often not simply a waste of resources but may negatively impact on future eradication attempts. Vector interventions have the potential for dramatic off-target effects through biological and behavioural modification of the mosquito. Some of these potential effects will be explored in the next sections.

1.1: Adult Vector Control:

The majority of vector controls target the mosquito's adult stage, almost exclusively through the application of insecticides and barrier methods. This control can be ineffective for a number of reasons. It often fails as a result of poor or incomplete application, however innate or developing factors within the vector can also impact the proper function of control. This includes innate physical and biochemical responses such as insecticide resistance and the more subtle behavioural resistance. Both of these were a factor at times during the GMEP.

Insecticide resistance

Resistance to DDT first emerged in Greece in 1951 and was directly attributed to it's uniform use in house spraying (Livadas 1952). In the years that followed, DDT resistance emerged independently in numerous countries and by 1975 256 million people were living in regions where DDT or dieldrin resistance made effective malaria control impossible; resistant vector strains were the major factor behind the abandonment of the malaria eradication program this same year (Hemingway & Ranson 2000). The problem of resistance didn't abate with the abandonment of the WHO program; by 1986 56 species of Anopheline mosquitos displayed resistance to either DDT or Dieldrin, and *Anopheles gambiae* itself had demonstrated DDT resistance in Liberia, Niger, Togo, Cameroon, DRC (Zaire), and South Africa, and dieldrin resistance in Mauritania, Mali, DRC, Kenya and Madagascar (Brown 1986). The broad geographical spread of these resistance outbreaks suggests multiple independent foci of resistance.
Modern insecticides suffer from the same problems. Pyrethroids recommended by the WHO in current efforts are faster acting than DDT, enabling their application via methods with shorter exposure times, such as insecticide treated nets (ITNs). However resistance to these too has been widely reported in a number of major disease vectors, including *An. gambiae* (Hargreaves et al. 2000; Protopopoff et al. 2013; Ahoua Alou et al. 2012; Kawada et al. 2011; Chandre et al. 1999; Fanello et al. 2003).

Anopheline mosquitos are well suited to the development of resistance to insecticides; their rapid generation times and large generation sizes assisting the rapid expansion of resistance phenotypes (Hemingway & Ranson 2000). However the methods of application must also take some of the blame. The one-size-fits-all approach, concentrating on single insecticides and broad spraying with little tailoring of approaches to individual settings has certainly exacerbated the problem.

Target-site alterations, such as the mutation in the VGSC (voltage gated sodium channel) gene that is responsible for much resistance to both DDT and pyrethroids, typically lead to stable phenotypes with only moderate loss of fitness for the mosquito (Hemingway & Ranson 2000). Though these can be uncompetitive in the absence of control, blanket coverage programs will provide strong positive selection to encourage their spread throughout a vector population. However such strong resistance phenotypes are comparatively rare. Rather, the most common cause of insecticide resistance is via the stimulation of detoxification mechanisms in the insect (Hemingway & Ranson 2000).

The increased expression of detoxification genes, whether by gene duplication or alterations in gene regulation, is a hallmark of insecticide resistance. There are three gene families in particular that are associated with this effect; all three have been directly implicated in insecticide resistance in anopheline vectors, two of these in *A.gambiae*:

Esterases have been shown to bind a broad spectrum of organophosphate neurotoxins used as insecticides, preventing the toxin binding (and thus inhibiting) cholinesterase. Resistant *Anopheles stephensi* (Hemingway 1983b; Hemingway 1982) and *Anopheles arabiensis* (Hemingway 1983a) mosquitos have been identified with esterase-mediated resistance to organophosphates such as malathion. Expression of glutathione-s-transferases (GSTs) is significantly elevated in resistant populations, indicating a role in DDT resistance in A.gambiae (Prapanthadara & Ketterman 1993). Cytochrome P450 monoxygenases (cyP450s) in insects play a role in the detoxification of plant toxins (Schuler 2011), they are also frequently highly expressed in insecticide resistant mosquitos, and cyP450 mediated resistance to pyrethroids has been detected in three anopheline species (Hemingway et al. 1991; Brogdon et al. 1997), including *A.gambiae* (Vulule et al. 1994).

Since all three of these gene families are already constitutively expressed in response to plant defense compounds their over-expression carries only a minor fitness cost on the mosquito; incomplete or minor exposure to insecticides may therefore encourage the emergence of successively more resistant generations.

Perhaps more important than the precise mechanisms of these responses is the fact that they are dosage dependent. The level of detoxification required to evade the insecticidal

effect will depend entirely on the amount of exposure to the chemical and it's concentration at the time of the encounter.
Poor or incomplete coverage during insecticide spraying has been recognised as a potential cause of insecticide resistance ever since the first publication on the subject by A.L. Melander in 1914 (Melander & Experiment 1914). This remarkably prescient article also proposed the use of combination therapies as a solution to emerging insecticide resistance, the solution currently recommended by the WHO in cases of resistance to pyrethroids (WHO 2014a).
With the current reinvigorated focus on the eradication of malaria there has been renewed interest in learning the lessons of the previous attempts to eradicate malaria, in particular how to avoid the 'bounce-back' effect often associated with the cessation of control programs (Cohen et al. 2012). Delivering insecticides in the right quantities, at the right time, and to the right mosquito will require detailed knowledge of the vector populations at hand.

Control targeting:

The reliance on indoor residual spraying was recognised as a significant weakness of the GMEP, particularly since the irritant effect of DDT acted to reduce the overall contact time with the mosquito (Coosemans & Carnevale 1995). It is largely for this reason that current efforts complement IRS with the faster acting insecticides and barrier methods of ITBNs. However, despite efforts to diversity the control measures that can be used for malaria eradication, employing antimalarials, malarial vaccines, and transmission blocking approaches in addition to insecticide treatments, it is still the case that the majority of vector interventions are designed to attack endophagic / endophilic mosquitoes – those that both bite and rest indoors. Any significant alterations in one or other of these interventions will have a dramatic effect on our ability to interrupt transmission.
IRS remains a mainstay of preventative measures, particularly in areas where bednet usage is imperfectly applied, and its weaknesses are well documented. The use of ITBNs represents a significant improvement on DDT spraying. Pyrethroids such as permethrin and deltamethrin do not possess the same degree of irritant effect as DDT and are considerably faster acting (at 4% insecticide concentration DDT LT50 = 108 mins, at 0.25% concentration permethrin LT50 = 14 mins) (Dialynas et al. 2009), and calculations of biting times in anopheline species indicate that over 75% of bites will be prevented by truly mosquito proof bednets (Pates & Curtis 2005).
Indeed practical confirmation of the efficacy of this approach has been shown in Hainan province in china; DDT spraying from 1959 had blocked transmission by the endophagic / endophilic *An. minimus*, however transmission was maintained at high levels by the

endophagic/exophilic *A.dirus*. A failure of control by DDT indoor spraying was reversed by the introduction of deltamethrin treated bednets (Curtis 1990). However it should be noted that, even in concert with IRS, ITBNs will offer no significant protection against entirely exophagic mosquitoes.

Behavioural Immunity

It is clear that knowledge of mosquito behaviour in a region is essential for the implementation of effective interventions. Advance knowledge of the prior presence of exophilic species will save valuable time and money in misdirected control measures. However this view, one that treats vector populations as static and homogeneous, is sorely lacking, and we should consider instead vectors as phenotypically and behaviourally variant. Even a nominally 'endophilic' species can at times exhibit exophilic behaviours. Therefore just as changes in ecology or geography will result in measurable alterations to the genetic makeup of vector populations, the sudden imposition of strong selective pressures such as insecticides, pressures that act against particular inherent behaviours, will select for alternate, evasive, behaviours if they are present in a population.
Whether the emergence of behaviourally resistant vectors is due to gradual adaptation over time, or rapid selection of common pre-existing behaviours remains an open question. However, the fact remains that, in the presence of behaviourally heterogeneous populations, any intervention that relies on a subset of that behaviour will select for it's own obsolescence.

Population replacement

There are more than 3500 species of mosquito, at least 462 anophelines, and at least 70 of these are known to be potential malaria vectors (Hay et al. 2010). Endophily and anthropophagy tend to go hand in hand, so control methods tend to concentrate only on the major endophilic species in each region: *An. gambiae* and *An. funestus* in Africa, *An. culicifacies* in India and *An. minimus* in south-east Asia.
This is understandable, since their strong anthropophily is a significant factor in making them highly efficient vectors. However the landscape of transmission is complex, and the removal of the dominant vector species in any one region can leave an open niche that secondary vectors can fill. Since these secondary vectors are frequently exophilic and/or exophagic, gains made in vector competence can be lost in vector control.
Evidence of species replacement has been seen as a result of control methods in the southwest pacific. Three major vectors were supported in this region; *An. punctulatus, An. koliensis, and An. farauti*. After the introduction of DDT-IRS in the 60s and 70s, the two highly endophagic species *An. punctulatus and An. koliensis* were almost entirely wiped out

from a region yet transmission was maintained by the third vector *An. farauti*, and eventually rebounded to pre-intervention levels (Russell et al. 2013).
Similar responses to control measures have been seen to occur in regions supporting both *An. gambiae*. and *An. arabiensis spp.* in east Africa. As a result of surveys in vector controlled regions of Tanzania, Russell *et al.* have noted a shift in species composition from the endophagic *An. gambiae* to the exophagic *An. arabiensis* after the introduction of widespread bednet use (Russell et al. 2013). Similar results have been seen in Kenya (Bayoh et al. 2010; Mwangangi et al. 2013). At the time of writing, species replacement has not yet been reported in the more heterogeneous populations of West Africa though this may reflect the difficulty of defining species boundaries in this region, rather than a genuine absence of population or species replacement.

Behavioural adaptation

Even in regions of low species diversity, behavioural resistance to control measures can be important. In the previous review, Russell *et al.* note that, as well as replacing the other two species, *A.farauti* demonstrated behavioural modifications that allowed it to evade IRS control. Biting times showed significant rises in both crepuscular and outdoor feeding after DDT spraying was introduced; these behaviours persisted after the cessation of DDT use. Changes in behaviour from within a single species require more work to detect than alterations in species composition, yet finding markers for phenotypically distinct sub-populations will be essential if behaviourally mediated bounceback in transmission is to be prevented.
Indeed behavioural immunity to insecticide spraying appears to be at least as widespread as physiological immunity. As well as the Polynesian examples, behavioural changes have been noted in South American (Trapido 1954; Mattingly 1962; De Zulueta 1959) and African settings, where the incidence of ‘bite-and-run’ behaviour (i.e. endophagic, yet exophilic mosquitos) increased dramatically in Tanzania and Burkina Faso in response to DDT-IRS (Gerold 1977). Similar increases in exophily for endophagic species have been noted in Europe (*An. sacharovi*) (De Zulueta 1959), South America (*An. darlingi*) (Rozendaal 1989) and Australasia (*An sundaicus*) (Sundararaman 1958). It is highly probable that such adaptability in vector behaviour can be found worldwide.
However, even more so than resting behaviour, alterations in feeding behaviour can cause this balance of exposure to alter dramatically. Relatively small changes in biting times can shift the principal target for intervention out of the home and into the field, nullifying the effects of ITBNs as well as IRS.

Perhaps the most dramatic illustration of this effect was on Hainan island in China; after a DDT-based eradication program targeted the endophilic-endophagic *An. minimus* mosquito, and an ITBN program accounted for the exophilic-endophagic *An. dirus*, a later resurgence in the disease was found to be a result of the same *minimus* species, but now with a clear exophilic/exophagic profile (Wu et al. 1993; Pates & Curtis 2005).
Comparable results have been reported on Bioko Island in Equatorial Guinea (Reddy et al. 2011) where the majority of *An. gambiae* sensu stricto and 40% of *sensu lato* were found to be seeking hosts outdoors, as compared to a pre-intervention profile of exclusive endophagy.
It is possible these behavioural changes are only the most extreme examples, yet even behavioural avoidance by a subset of the vector population will nullify eradication efforts. Modelling of behaviourally heterogeneous populations (Eckhoff 2013) suggests that the presence of secondary vectors, or of opportunist exophily in the primary, will be sufficient to maintain transmission in the presence of control. Whilst the greatest exposure to risk has traditionally been indoors, even for the most endophilic mosquitoes a small proportion of exposure will always take place outdoors (Killeen 2013). If overall transmission is reduced due to the introduction of ITBNs this proportion grows in relative importance (principally due to the removal of time spent under a net from the indoor exposure).
It is not entirely necessary that biting times are altered by selection of pre-existing behaviours. Charlwood *et al.*. (Charlwood & Dagoro 1989), who also saw an alteration in biting times of *An. farauti* after bednet introduction in Papua New Guinea, theorised that the alterations in feeding times they saw were a reaction to the lack of available (i.e not bednet covered) bloodmeals during the night, forcing earlier feeding the following day. It is interesting in this context that some environments have shown an increase in exophagy in the early morning rather than the evening. For example, analyses of median catching time (MCT) for *An. funestus* in Benin demonstrated a shift from midnight biting (MCT=2-3am) before the scale-up of control, to morning biting (MCT=5am) after. As a result, 26% of *An. funestus* in this region were found to be biting after 6am - when the majority of the population is not protected by ITBNs (Moiroux et al. 2012).
In addition, Mbogo *et al.* report a shift in biting times and a consequent increase in exophagy for both *gambiae* and *funestus* species complexes after the introduction of bednet control in Kenya (Mbogo et al. 1996).

1.2: Larval control

If indoor measures are impractical, mosquito control becomes difficult to achieve.
Intervention at the moment of biting has the distinct advantage that it ensures that the most anthropophilic species are targeted. In situations where control at the point of infection is not

possible it is necessary to instead identify other points where the vector can predictably be found, and to attempt to apply control here; this is typically limited to larval habitats. Mosquitos exhibit relatively little dispersal, with capture/recapture experiments indicating typical ranges of under ½ km (Midega et al. 2007). As a result the availability of suitable oviposition sites is the major determinant of vector density in a region, and reductions in the number or viability of these ovipositions sites is a good way of interrupting malaria transmission.

Moreover larvae demonstrate some basic advantages as subjects for control methods that the mosquito's adult stage does not. Adult mosquitos are highly mobile and adaptable insects that are able to detect and avoid many control measures. Larvae are confined to single pools for their whole development and are oviposited predictably close to regions of transmission. They are, in comparison, sitting targets for vector control as long as a suitable method can be devised.

Aedes and *Culex spp,* vectors of Dengue, yellow fever or lymphatic filariasis are species that typically exhibit exophilic behaviours (Pates & Curtis 2005). Consequently they are behaviourally resistant to IRS control, and a larval control approach is commonly taken. In addition, African malaria vectors have been twice eradicated from large regions; first when invasive *An.gambiae* was eradicated from Brazil in the late 30s (Soper & Wilson 1943), and again in Egypt in the 1940s (Shousha 1948). Both programs were entirely based on aggressive and coordinated larval control.

Chemical larvicides

Larvicidal control during the Egyptian and Brazilian *An* gambiae eradications, and during control efforts in general during world war two, was characterised by liberal spraying of larvicidal agents across the north African theatre (Service & Office 1963). Indeed 'liberal spraying' might be an understatement: during Egyptian eradication efforts 138 (imperial) tons of larvicide were used over a 15 month period, at its peak fifteen tons were used in a single month (Shousha 1948).

The agent used, Paris Green (copper(II) acetoarsenite), was sprayed across larval habitats as a fine powder that rested on the water surface and was ingested by larvae. It gained its name after being used widely in rat control within the Paris sewers and had found further use as a pigment in clothes, wallpaper and paints; its vivid hue was particularly popular with impressionist painters.

However as its use as rat poison illustrates, it is highly toxic to mammals and linked to a broad variety of ailments including cancer, dermatitis, diabetes and macular degeneration[1]. . Despite its stability, ease of application, and efficacy as a larvicide, Paris Green's indiscriminate toxicity makes it impractical for a continent-wide control effort.
More modern larvicides, in particular temephos, exhibit lower mammalian toxicity and have been used widely for mosquito control in several countries (including india, Mauritius and oman) (Walker & Lynch 2007). However resistance to this compound has begun to be seen (Coosemans & Carnevale 1995).

Biological larvicides

Due to the difficulty of finding suitable persistent toxins, biological control has frequently been used in larval control efforts. Natural predators of larvae include a wide variety of larvivorous fish. Prior to the 1970s the non-native 'mosquitofish' *Gambusia affinis affinis* was widely used, though it has now been phased out in favour of a local species (though several counties in the mosquitofish's native California still supply *gambusia* for domestic ponds (scc.gov.org 2014; mosquitoes.org 2014; contracostamosquito.com n.d.)). Success has been achieved with similar fish chosen from local species across africa: *Aphanus dispar* has been used to suppress *An. culcifacies* populations in Ethiopia (Fletcher et al. 1992); *Oreochromis spilurus* in northern Somalia(Mohamed 2003); And A.gambiae has been controlled by the (non-native) *Poecilia reticulate* in Comoros, leading to a significant reduction in malaria morbidity (Sabatinelli et al. 1991).
The dominant copepod predator *mesocyclops* is also used for control of temporary breeding sites in *Aedes aegypti* in the USA (Marten et al. 1994), Honduras (Marten et al. 1994), and Vietnam (Sinh Nam et al. 2012). It has also been shown to be effective in natural settings against *Anopheles* species in south America (Marten et al. 1989), though it has not yet been used for coordinated control efforts against malarial mosquitoes.
Whilst the use of large predators has been relatively haphazard, two microbial species have been extensively used and demonstrated to be effective larvicides: *B. thuringiensis israeliensis (Bti)* and *B. sphaericus* (*Bs*) are both highly toxic after ingestion by the mosquito larvae (Walker & Lynch 2007). *Bti* has broad application to *Aedes, Culex* and *Anopheles spp.* (Mittal n.d.), though it's duration of efficacy in the field is limited to one or two weeks at most, while *Bs* has a longer duration of efficacy in the field, but a narrower range of toxicities

[1] It is believed that Paul Cezanne's severe diabetes may have been linked to his frequent use of Paris Green – particularly given his habit of applying paint with his fingers (Zieske 1995). Cezanne died of pneumonia whilst in a diabetic coma. Somewhat ironically given the link between Paris Green and diabetes (Maull et al. 2012), this illness was first diagnosed when it robbed him of the ability to distinguish between blue and green in his painting.

(effective against *An. stephensi* and *An. subpictus*, it has limited toxicity to *An culcifacies* (Mittal n.d.) and results are mixed from *An gambiae* (Majori et al. 1987; Karch et al. 1992)). In addition fungal species, including *Metarizium anisopliae* and *Beauvaria bassiana* have been shown to be have potent insecticidal effects against adult *An.gambiae* (Scholte et al. 2005). Whilst these are at an earlier stage of development, they have shown notable advantages in persistence and autodissemination over bacteria (Scholte et al. 2004).

Habitat targeting

Despite the prior success of Paris Green, it is unlikely that such an approach to larval control will ever be taken again, particularly with such an indiscriminate toxin. Subsequent larvicides have fewer off-target effects, but far shorter persistence in the field, rendering a similar scale of control impossible. Finally the kind of campaigns of larvicidal control carried out under dictatorship or military control would be extremely difficult to impose on the post-colonial Africa of today.

Targeting control towards a subset of larval sites is therefore the only realistic option. Models of larval control (Gu & Novak 2005) have estimated that a targeted approach covering 40% of larval sites could dramatically reduce vector populations and the entomological inoculation rate. Whilst identifying larval habitats is time consuming, the tendency for anthropophagic mosquitoes to oviposit near to human habitation should reduce the difficulty of breeding site identification (Walker & Lynch 2007).

However malarial mosquitoes present us with particular challenges. *Aedes* and *Culex* display a preference for permanent, relatively large, larval habitats. As a result identification and control is comparatively easy. In comparison, *An. gambiae* is traditionally characterised as preferring temporary or semi-permanent habits, such as rainpools, hoof prints, and drainage ditches. As these habitats are by nature ephemeral, any control method would have to be able to be implemented in the days or weeks that the habitat will persist, or able to be applied to potential sites and to remain effective after they became viable. Finding a suitable control method can be more difficult than in vectors of other diseases.

Many of the methods that have proven adept at reducing vector numbers for *Aedes* or *Culex* mosquitoes are unsuitable for *Anopheles*. The majority of the field trials so far undertaken are in environments that are either rare or absent in sub-saharan Africa (Walker & Lynch 2007). The classical *An.gambiae* habitat of temporary rainpools are both too shallow and too impermanent for larger predators such as larvivorous fish. Copepod predators demonstrate some resistance to dessication, being able to survive for a period of more than two months without water (Zhen et al. 1994), however dry-period survival in field tests in central America

demonstrated considerably reduced survival (Gorrochotegui-Escalante et al. 1998) and many habitats in sub-saharan Africa will be unsuitable for these predators.
Shorter-acting biocidal controls such as *Bti* and *Bs,* and mosquito-specific larvicides such as temephos also suffer from logistical issues. Since they are not able to survive extensively in the field they cannot be sprayed indiscriminately wherever a mosquito breeding site may appear. Whilst entomopathogenic controls may have a longer application period, their development as larvicides is at an early stage, and it remains to be seen whether they can be adapted into large scale control measures.

Habitat heterogeneity

Mosquito larval habitats are far from homogeneous. Distinct preferences for oviposition sites are exhibited by the two major African vectors, with *An gambiae s.l.* being most commonly found in temporary stagnant rainwater and *An funestus s.l.* being more commonly associated with permanent habitats (Gimnig et al. 2001). Whilst the *funestus* habitat is broadly comparable to *Aedes and Culex spp,* the *gambiae* habitat is not, and is particularly problematic for control.
Although *An. gambiae* and *An. funestus* complexes demonstrate broad differences in habitat preferences, these preferences are not absolute and crossover within them is not uncommon. Within species complexes distinct differences in oviposition sites are found, even between closely related species (see section 3.5). As many of these species cannot be visually distinguished, molecular assays may well be necessary in order to discern whether the species in larval pool ‘A’ is really the same one biting in house ‘B’.
Even within individual species, larval habitats can differ widely between sub-populations. Larval habitat preferences are believed to underlie the major speciation event within *Anopheles gambiae sensu stricto*, and frequently differ in distinct populations of the same species (Coulibaly et al. 2007).
Indeed our knowledge of all larval habitats is clearly lacking; *Anopheles* appears to be highly selective with respect to oviposition sites when plentiful (Mereta et al. 2013), displaying distinct preferences with respect to shade, habitat permanence and predator presence, yet new larval sites for *Anopheles gambiae* are still being discovered (Omlin et al. 2007), and the relative productivity of oviposition sites is variable (Kweka et al. 2012). Moreover there are clear signs of adaptation to particular larval ecosystems between incipient species (Lehmann & Diabate 2008) and – significantly for attempts to control by copepod predation – significant differences in predator avoidance are seen between molecular forms. The actions of inter-form competition on habitat choice may be significant (Diabate et al. 2005).

Species distributions show association with metrics such as water quality, temperature and pH. Similar associations are seen between these metrics and markers of within-species differentiation such as chromosomal inversions (Sanford et al. 2013).
Depriving the vector of oviposition sites, either by drainage, predation or poisoning can have a significant effect on the entomological inoculation rate (Dieter et al. 2012). It may well be a valuable facet to an integrated control program, and has re-emerged as a field of study in recent years. It also, of course, will be equally effective against endo and exophagic mosquitoes. However, although it is possible to avoid many of the problems of behavioural avoidance that can be seen in adulticidal control, identifying these habitats and accurately targeting the control method against the dominant vector is a significant problem.
The recent discovery of a cryptic population of exophilic mosquitoes in larval samplings in West Africa (Riehle et al. 2011) only serves to show how many gaps remain in our knowledge of the relationships between adult and larval vector populations. Notably this population displayed a particularly high susceptibility to malaria parasites. Modelling has indicated that control methods that preferentially targeted endophilic mosquitoes could, in fact, increase transmission where a similar cryptic secondary vector was present (Yakob 2011).

1.3: Genetic Control

Genetic control refers to the introduction of genetic elements into the mosquito population in order to reduce the viable vector population – either by a reduction in real numbers of mosquitoes, or by reducing the capacity of the insect to transmit malaria.
Both in the past and at present, these methods have been more commonly used in *Aedes* mosquitoes, due to the aforementioned issues of exophily and the lack of applicability of ITBNS and IRS. They can be broadly categorised into two approaches: those that require active replacement of a population (i.e. self-limiting), and those that rely on non-Mendelian inheritance to alter a population (and are, as-such, self-sustaining).
As the difficulty of controlling malaria transmission solely with insecticides becomes apparent, these methods are increasingly being looked at for their potential utility in *Anopheles spp.*

Sterile insect technique

Self-limiting genetic control is typified by ‘sterile insect technique’ (SIT). This involves the release of large numbers of infertile males into an area. As mosquitoes are typically monogamous, each coupling with an infertile male will lead to a drop in the total vector

population and successive releases of infertile males will cause a population crash. Sterilization is carried out by irradiation with only moderate loss of fitness to the male vector (although *Anopheles* appear to be less robust to the technique than the larger *Aedes* and *Culex spp.* (Helinski et al. 2009)).

The technique's application to mosquito control was developed between the mid 50s and mid 70s (Klassen 2009). Small-scale trials with both Aedine and Anopheline mosquitoes were completed and large scale field trials attempted in El Salvador and India (Klassen & Curtis 2005). However work in South America ceased prematurely due to the deteriorating political situation in El Salvador, and in India a public panic that the US-funded project was actually a military experiment to test mosquitoes as bio-weapons (designed to spread yellow fever) caused the Indian government to cancel the project[2] (WHO 1976). Although the common name of *Aedes aegypti* is 'yellow fever mosquito' this is not an endemic disease in India; the actual target of the intervention was dengue fever.

Recent revival of the technique has been encouraged by the availability of transgenic techniques that are able to induce male infertility without any other loss of fitness. Again most advanced in *Aedes aegypti,* the technique has lead to the development of a strain (OX513A) carrying a dominant transgene that is lethal to the pupal stage of development (Harris et al. 2012). Its lethality repressed when grown in the presence of tetracycline in the lab, mosquitoes can develop with no loss of fitness. However after release their wild offspring will inherit the lethal construct; release of large numbers of male OX513A will therefore cause a similar population collapse to irradiated males. Large scale field trials have shown a significant reduction in wild *Aedes* populations in Grand Cayman and Brazil (Harris et al. 2012).

Attempts to develop a similar technique for *anopheline* mosquitoes are underway, with a dominant transgene inserted into *An. stephensi* causing female-specific flightlessness (Marinotti et al. 2013) (equivalent to sterility in an insect that breeds in a swarm). Nevertheless, implementing a successful SIT control is still a significant challenge, requiring extensive testing to ensure mating competitiveness and suitable dispersal. The breeding of

[2] This is not as far fetched as it sounds. Shortly after the Korean war, the US government was, in fact, testing *Aedes aegypti* as a potential bioweapon. 'Operation Big Buzz' involved the release of E14 bombs filled with 330,000 (uninfected) yellow fever mosquitoes, to test dispersal and biting rates on a human population. Fortunately for the residents of New Delhi, the US Army had the good grace to test this closer to home, in Savannah, Georgia.

The price of a yellow fever attack on a battalion was estimated (in 1976) at $26,666; if the mosquitoes are employed against an unarmed target this drops to a very reasonable $10,473 including "truck rental and wages of two semi-skilled people for eight hours".

The files were declassified in 1981(Rose 1981; Lockwood 2008).

the large numbers of mosquitoes required for SIT also requires major financial investment in terms of breeding facilities close to the release site.
It is also apparent that an intervention that is based on matings with sterile males will be unable to breach any barriers to mating in the field. While this is relatively uncommon in aedine populations, Anophelines are known to feature a number of instances of assortative mating (see chapter 3) and released mosquitoes could not be relied upon to breed freely across the vector population. Implementing SIT in malaria control will require extensive knowledge of mosquito population structure and could potentially require the development of numerous transgenic male-sterile mosquitoes from different populations, in order to provide attractive mates for a remarkably choosy set of female mosquitoes.

Gene drive

The second broad class of genetic control consists of those interventions that are self-sustaining. Rather than requiring the regular release of large numbers of mosquitoes that can breed with a proportion of the vector population in order to reduce overall numbers, these techniques instead rely on gene-drive systems to spread genetic elements that aim to reduce the vectorial capacity of the mosquito.
Again much of the work has first been performed in *Aedes aegypti*. The gene drive system is the symbiotic bacterium *Wolbachia pipientis* that was known to actively spread through populations of *Drosophila melanogaster* (Werren et al. 2008) and was responsible for cytoplasmic incompatibility in other mosquitoes (Yen & Barr 1971). The *Wolbachia* bacterium is transmitted vertically, being found in both male and female gametes. Significantly, crosses between infected males and uninfected females produce sterile offspring, yet crosses of two infected parents are viable. The result is that infected males – being able to breed with both infected and uninfected females – are at a distinct evolutionary advantage, and the bacterium will spread rapidly through a given population.
Trans-infection of *Aedes aegypti* with the *wMel* strain, however, gave a serendipitous result; a significant increase in resistance to the dengue virus (DENV) (Hoffmann et al. 2011) (as well as the other arboviruses chikungunya (CHIKV) and yellow fever (YFV) (van den Hurk et al. 2012)). This link was also replicated in nature, where wild *Aedes* infected with *Wolbachia* displayed similar resistance to infection with DENV (Frentiu et al. 2014).
In ongoing large-scale field trials, release of wMel-infected mosquitoes in Queensland, Australia has demonstrated rapid successful invasion of natural populations (Hoffmann et al. 2011). Following the success of this approach in *Aedes* attempts have been made to replicate the work in *Anopheles*. In 2013 a stable, vertically transmitted, infection of *Wolbachia* was established in the Indian malaria vector *Anopheles stephensi* (Bian et al.

2013). *Wolbachia* drive systems rely on a fine balance between the fitness costs of carrying the parasite, balanced with the fitness gains of avoiding cytoplasmic incompatibility. In *Aedes* this goes hand in hand with a drop in vector capacity for DENV.

Alternative approaches have been suggested exploiting selfish genetic elements that would be inherited in a 'super-Mendelian' manner (Sinkins & Gould 2006). If these genetic drive mechanisms could be paired with specific target genes it may allow us to make gene-specific population wide knock-in or knock-outs, enabling highly specific control of the vector capacity in a population.

Maternal-effect selfish genes (MEDEA – maternal effect dominant embryonic arrest) have been suggested as a potential knock-in mechanism for *Anophelines*. First identified in *Tribolium* beetles (Beeman et al. 1992), these systems involve a gene carried on the maternal chromosome that is lethal to her offspring, along with an associated rescue gene on the same chromosome. Any offspring of heterozygous mothers that are subject to the lethal effect but do not inherit the rescue gene will be non viable. If MEDEA elements are introduced at sufficiently high levels they will be rapidly driven to ubiquity. Physically linked genes that are attached to a MEDEA or rescue gene would also be driven to high frequency in a population, providing an effective method of population wide knock-ins. MEDEA elements have been reverse engineered in *Drosophila* by combining a miRNA repressor of the *myd88* gene (a key component of the TOLL pathway, required for activation of the NF-kB transcription factor and vital for embryogenesis), along with a rescuing second copy of *myd88* (Chen et al. 2007). Similar engineered constructs have yet to be developed for any mosquito.

Of particular promise are homing endonuclease genes (HEGs), a class of nuclease originally found in yeast that can spread rapidly through a population even when they impose a fitness cost. The endonuclease proteins encoded by the HEG cleave a site-specific sequence corresponding to the flanking sequence of the HEG itself, after which the cell's DNA repair mechanism uses the HEG+ chromosome as a template for repair, copying the nucleotide into the HEG- chromosome. In most natural systems these are found within introns or inteins, however if the target sequence is found within coding sequence the HEG will interrupt the gene even in the face of countervailing pressure.

Proof of principle experiments have already shown this to generate strong gene drive in *Anopheles gambiae* in controlled colonies (Windbichler et al. 2011), with the number of individuals possessing the element progressing from 19% to 86% within just 12 generations. Further work will be needed to adapt these to different targets sequences (work has been undertaken on adapting their specificity in other organisms (Seligman et al. 2002)), but should HEGs be able to excise specific sequences they could be a highly efficient method for introducing population-wide knockouts into mosquitoes.

Neither MEDEA or HEG systems are close to being ready for use as control, however the potential to genetically manipulate mosquitos at the population level could enable us to specifically target the species of the highest vector capacity, even if they possessed exophagic/exophilic behaviours. Moreover, although this is a research-intensive approach, the release of relatively few individuals should sufficient to have a major effect on vectorial capacity as techniques such as SIT, with far lower overall costs.
However, any gene drive mechanism will suffer from similar problems as SIT (or auto-disseminating entomopathogenic fungi) when faced with populations with strong assortative mating. Methods that have a moderate to high threshold before population spread begins may not achieve that threshold if vectors are segregated into different groups. Suppression of sub-populations with undesirable effects (Yakob 2011) is a distinct possibility.
More basically, for HEG genes in particular, it is unlikely a broad choice of target genes will be available. Therefore it is pertinent to ask which mosquito genes should be interrupted in order to have the greatest effect on malaria transmission, whilst having the least effect on mosquito fitness.
Sreenivasamurthy *et al*. suggest 34 genes that have been confirmed via functional assay to have an effect on either oocyst or sporozooite number in the mosquito (Sreenivasamurthy et al. 2013). Of this 34 a subset will have a sufficiently detrimental effect on fitness to be unsuitable as gene drive targets, and the number may be reduced drastically once we know which genes are suitable for HEG interruption.
It is also important to note that any genetic modification that is achieved by MEDEA/HEG technology is going to take place in the field not in the tightly-controlled conditions found in the lab. Many of the genes implicated in mosquito/*Plasmodium* interactions have been based on highly inbred lab colonies being infected with non-adapted .*P. falciparum,* or the murine malaria parasite *P. berghei*; parasites to which the *gambiae* mosquito has variable and dubious immunity (Holm et al. 2012; Mitri et al. 2009).
Before the potentially considerable expense of gene targeting is undertaken, it will be wise to ensure that target genes identified in the laboratory are indeed relevant to refractoriness in the wild. Methods for examining refractoriness and immunity in the wild are therefore a crucial precursor to genetic control.

1.4: Integrated control programmes

An important feature of control programs is that their effects on the reproductive rate (R0) are multiplicative. Thus even a two-fold reduction in human infectivity will have a significant effect

if combined with a mosquito intervention of greater effect, and easier secondary interventions can overcome diminishing returns as primary control methods are exhausted.
The concept of integrated vector management (IVM), therefore, has been enthusiastically embraced by the WHO (WHO 2014b) as a method of using current interventions to their highest effect, and of ensuring that novel interventions are targeted where they are needed most.

> *Rather than relying on a single method of vector control (e.g. chemical spraying), IVM stresses the importance of first understanding the local vector ecology and local patterns of disease transmission, and then choosing the appropriate vector control tools from the range of options available.*(WHO 2014b)

However, it is not in regions of highest transmission where IVM is practiced most keenly, but in the developed countries. Most obviously in the USA, where public opinion would preclude indiscriminate insecticide spraying, yet nuisance-biting mosquitoes are controlled using an evidence-based combination of techniques, including limited spraying, larval site management and biological control.
It is not simply a matter of public antipathy towards insecticides that should encourage IVM programs. High transmission settings will benefit from similar approaches. The focus on ITBN and IRS as the most effective control methods is well founded – when employed properly they can reduce parasite transmission by 90% (Beier et al. 2008) - however it is also known that this will often be insufficient for an eradication attempt. More comprehensive programs must be implemented if we are to eradicate malaria from regions of the most intense transmission.
In addition, insufficient coverage of insecticide based approaches is a causative factor in the emergence of insecticide resistance; given eradication programs are widely expected to take more than half a century to achieve their goals (Okie 2008) it is more than likely that insecticide resistance will at some point increase to a level that will impair transmission-control efforts. In both these cases a solid understanding of the effects of additional control measures will be vital if evidence-based decision-making is to become a reality.

Modelling and Measuring

Modelling is key to this approach in that it allows us to make decisions on which control method to implement based on prior knowledge of the mosquito ecology, or to monitor the efficacy of a control program as it is implemented. Whilst there is limited application of this in conjunction with control programs, several papers have demonstrated its potential utility. Modelling of transmission in high and low-intensity regions by Griffin *et al.* (Griffin et al. 2010) indicated that in low-to-moderate transmission regions (EIR = 3-81 infected bites per person

per year) ITBN and IRS use alone could reduce malaria prevalence to under 1%. However regions of higher transmission (EIR = 586-756) would necessitate the development of alternative approaches to reach this level.

Similar results have been reported elsewhere. In an alternative model, based on environmental larval capacity and density-dependent competition, White *et al* demonstrated that the availability of larval habitat sites can explain all all of the seasonal variation in malaria transmission (White et al. 2011), and reconfirmed the previous finding that only interventions that target the mosquito at non-feeding stages (i.e. larval or pupal) will have any significant further effect in regions of high ITBN coverage.

Interestingly, however, when investigated using the griffin *et al.* model, moderate regions (EIR=46) with a predominance of the exophilic vector species *An arabiensis* were not amenable to the ITBN/IRS approach (Griffin et al. 2010). This serves to highlight the importance of knowing not only the overall levels of transmission, but the ecology and behaviours of the vector species that give rise to that transmission.

This risk was also highlighted by White *et al.* (White et al. 2011) as a weakness of their model, which considered only a panmictic *An. gambiae s.l.* population. Although An.gambiae and An.arabiensis compete when placed in the same larval breeding sites (Kirby & Lindsay 2009), the two species display distinct preferences for their breeding sites. However the response to ITBN introduction will be different in each case; where breeding sites are separate, control of *An.gambiae* will not significantly affect the numbers of *An.arabiensis*; however where larval competition is a limiting factor on the population size of *An. arabiensis,* ITBN introduction that causes a significant reduction in *An. gambiae* numbers will be expected to cause an increase in the overall number of *arabiensis* individuals (White et al. 2011).

The heterogeneity of larval habitat choices is poorly understood for any of the major vector species, and other factors, such as differences in larval development times and response to temperature (increased larval habitat temperature is associated with shorter larval development times) are also likely to affect density-dependent competition (Lyimo et al. 1992). Increased knowledge of larval and adult ecology will have a direct effect on our ability to predict the effect of intervention methods.

Even with an imperfect understanding of mosquito population dynamic, lower resolution information can also assist in directing control methods. Although mosquitoes travel only over limited distances, in many hypoendemic regions periodic outbreaks of malaria are believed to emerge from nearby mesoendemic regions by limited local migration of the mosquito. Either as a consequence or a corollary with their relative inability to support hypoendemic malaria, these hypoendemic regions are frequently densely populated or

developed. This knowledge can be used to implement 'barrier' control regions to prevent this migration into vulnerable areas.

Control implemented by the Ministry of Health in Zimbabwe has provided a good example of the utility of barrier methods. Low lying areas of the country, below 600m, were holoendemic yet supported only small human populations, whilst the uplands (above ~900m) were too cold in the winter to support dense year-round mosquito populations. Increases in temperature in the spring, however, allowed the mosquito to progressively invade upland areas.

Implementation of intense external spraying in regions between 900-1200m altitude were able to successfully control this seasonal expansion for over 50 years (Taylor & Mutambu 1986). Sadly administrative issues, and a change from centralised to localised control implementation, recently resulted in a significant rebound of malaria transmission in this region, when comprehensive coverage of these barrier regions failed to be implemented (Shiff 2002).

Clearly barrier implementation of control requires detailed knowledge of vector distributions and an understanding of how this is related to overall transmission levels. Since mosquito species and populations have differing abilities to exploit arid regions or permanent larval habitats this knowledge would have to include details of the ecological adaptability of the vector at hand.

1.5: Summary

While great strides have been made by coordinated programs of indoor residual spraying, distribution of bed nets, and the increased availability of antimalarial drugs, these methods have so far failed to eradicate the disease from regions of the highest transmission. This has prompted an admission from many quarters that these tools alone are not sufficient to finish the job. The Malaria Eradication Research Agenda (malERA) was drawn up to identify which areas of research and development were most important for an elimination program (Alonso et al. 2011).

Alongside the development of new classes of insecticides, and improved management of data on malaria transmission, the malERA committee highlight the following foci for research:

1. development of novel methods of control that are effective against exophilic and exophagic mosquitoes;
2. gaining a better understanding of the ecology, behaviour, and population structure of malaria vectors;
3. long-term development of approaches such as genetic manipulation that will permanently reduce the vectorial capacity in regions of highest

> transmission.
> ("The malERA Consultative Group on Vector Control" 2011)

Ultimately, all three of these points rest on increasing our knowledge of vector dynamics in the field. Understanding why and how current methods fail, will aid us in improving the application of those tools in the future, and knowledge of the true diversity of vector species in a given area will enable us to implement locally tailored control measures. Ones that can account for principal vectors without opening niches to secondary vectors that might not be susceptible to the same control.

However assessing the true complexity of a vector population when faced with a species complex of behaviourally divergent, reproductively isolated but *entirely isomorphic* species is a major challenge. Morphological identification will not identify population groupings with sufficient resolution, and financial limits will surely preclude sequencing in the field.

Therefore, before we can understand the population structure in the field, we must first develop methods in order to assay that structure. That is, we must identify markers of genetic separation that are tractable and simple enough to be applied to large numbers of individuals. This, in itself, is a prerequisite for assessing population structure in the field. A subsequent task is then to discover which of these markers / populations are associated with behaviours such as exophily or larval habitat preferences that are important for transmission.

The third focus, the development of novel genetic control, also relies on an understanding of vector population structure and ecology. However, beyond the specific mechanisms of gene-drive systems, it is also important that we understand more about the mosquito immune system. With limited opportunities to intervene at the genetic level, getting the greatest effect from a knock-in/out will require that we have a range of target genes to choose from. In addition we should have a far greater understanding than we do at present which of these genes are actually important for refractoriness to malaria as it is encountered in the field and not in the lab. The development of techniques to assay and discover these genes is therefore also important.

Finally, regular surveillance of vector populations both during and after control (including inter and intra-specific differences) was highlighted by the malERA committee.

Surveillance of this kind, if it is able to show perturbations in the particular proportions of populations that are present in a region, could provide early warning of problems such as insecticide resistance or species replacement. This may enable the implementation of secondary controls before rebound of malaria becomes an issue.

In their review of the lessons learned from the global malaria eradication program, Najera et al. state:

> *"Surveillance should not only aim to detect the last case, it should be an essential instrument from the start, involved in the identification and study of problem areas, beyond the limits of administrative localities. As the elimination programme advances, epidemiological investigations should concentrate successively in the study of outbreaks or clustering of cases and finally of individual case investigations"* (Nájera et al. 2011)

Chapter 2: Mosquito Immunity

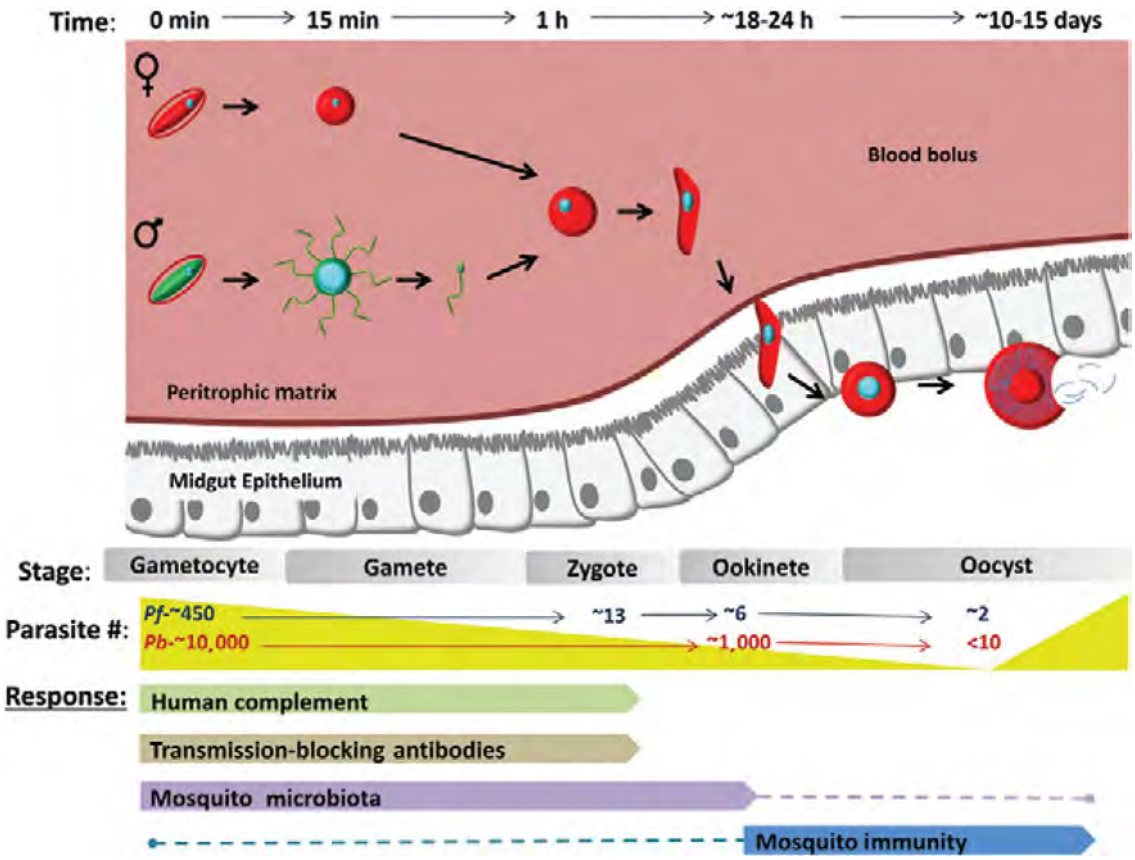
2.1: Introduction

The oocyst bottleneck

The primacy of vector control methods in malaria prevention is unsurprising when considering the *Plasmodium* life cycle (see appendix 1). The complexity of the life cycle may buffer the parasite against intervention measures in one host or the other, however the complexity of this life cycle also presents difficulties for the parasite. Reductions in the overall number of parasites take place at every developmental transition, and an examination of the numbers of individuals at each life stage soon makes it apparent that transmission from one host to the other is a major constraining factor on the transmission of the parasite.
Transmission from mosquito into human is one transition that results in significant reductions in parasite numbers. Relatively little of the mosquito saliva is reinjected into the mammalian host: of the 16,000 sporozoites present in an infectious mosquito only around 10 will be inoculated into the human when the mosquito probes for blood (Sinden 1999).
However despite the larger volumes that are transferred it is, in fact, the ingress into the mosquito host that is the most significant bottleneck. After uptake of the bloodmeal the parasite must undergo the transition into motile gametocytes, all the time dealing with mammalian immune factors still present in the bloodstream, as well as the mosquito innate immune system.
Often unfairly considered less sophisticated than the human immune system, the mosquito innate immune system presents a powerful array of challenges to the invading parasite that successfully kill the vast majority of invading parasites. Previous assays in which naïve mosquitoes were fed on infected human blood demonstrated that, in fact, only 38% of the infected bloodmeals will result in an infected mosquito; a ratio which compares extremely favourably to the human immune response (Gouagna et al. 1998).
Of the many millions of infected erythrocytes that will be carried by a *P. falciparum* infected individual, only around 500 will be ingested by the mosquito; this is rapidly reduced to ~13 zygotes, and around half of that number of ookinetes. Seldom more than one or two ookinetes will proceed to form oocysts (Sinden 1999; Gouagna et al. 1998).
The array of physical and immune barriers acting against the parasite make the mosquito midgut an extremely hostile environment for *Plasmodium.* In addition, due to the similar timescales of Parasite development in the mosquito and the typical mosquito life span (both

are around 14 days), the majority of mosquito infections will have no second encounter with a potential human host. As a result, a small reduction in the efficiency of parasite invasion of the mosquito may interrupt transmission entirely. Despite the obvious fact that mosquitoes are an unfavourable environment for chemical intervention, it may in fact be the best place to intervene in the *Plasmodium* life cycle.

Figure 2.1 :



Parasite losses in the mosquito host represent the single most significant bottleneck within the parasite life cycle (Sinden 1999). The combined challenges of the mosquito immune system and he residual human immune system act to reduce individual parasite numbers by at least two orders of magnitude by the time of oocyst formation.

Source: The Plasmodium bottleneck: malaria parasite losses in the mosquito vector (Smith et al. 2014)

Innate immunity

Due to the fact that insect immunity is largely non-adaptive, it follows that all functions of recognition, along with many of the effector mechanisms, are germline encoded. Therefore, in contrast to an adaptive immune response, where each individual host could mount a

different immune response to the same parasite, in an innate immune system the array of recognition genes and pathways that are activated (though not the actual protein complement) should be near-identical in every individual of a species challenged with the same parasite.

The mosquito response to malaria has been elucidated in recent years by a variety of methods. Most significant in the post-genomic era has been the use of comparative genomics, transcriptomic assays (particularly microarrays) and the development of RNA-mediated knockdowns allowing researchers to identify and test candidate genes that are implicated in infection with malarial parasites. The relatively close relationships between *Drosophila* and *Anopheles* have enabled much of the immune function of the mosquito to be deduced by comparative genomics, however in more recent years a number of *Anopheles*-specific features and genes not shared with *Drosophila* have been uncovered.

2.2: Immune Signalling Pathways:

It is perhaps fortunate that this medically important vector has, in *Drosophila melanogaster*, an extensively annotated model species that is sufficiently related to be used for comparative genomic analysis. This has allowed many gene definitions and pathways to be transferred intact. Homology and gene expression analyses have illuminated the major pathways through which the mosquito acts against infection, and a number of species specific actions of each.

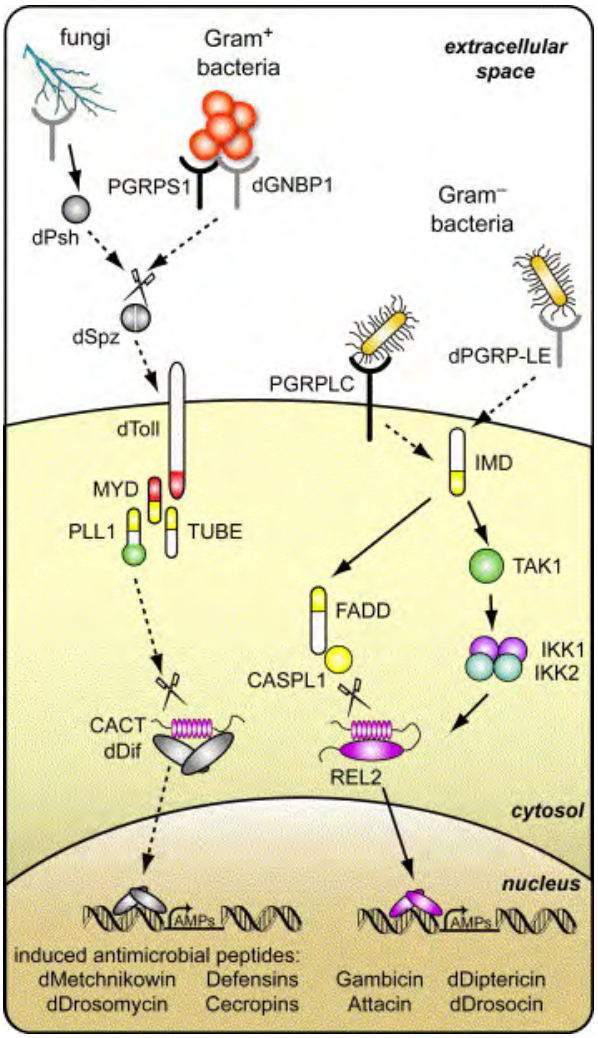


Figure 2.2:

Components of the two principal immune pathways as derived from comparative genomic analyses. As-yet unidentified components that are present only in the Drosophila immune system are prefixed by 'd'.

Source: The Plasmodium parasite--a 'new' challenge for insect innate immunity, (Meister et al. 2004)

The principal signal transduction cascades, TOLL, IMD and JAK-STAT, are known to affect parasite load by the activation of distinct yet overlapping panels of effector molecules (M. a Osta et al. 2004). Each pathway can be stimulated by multiple parasites; pathogens frequently activate both pathways to some degree, thought the relative importance of each pathway is still a source of debate.

TOLL

The TOLL pathway was initially identified as a result of it's effect on dorso-ventral patterning in *Drosophila* (Nüsslein-Volhard & Wieschaus 1980), including the *TOLL* receptor itself and the transcription factor *Dorsal* (a *Rel/NF-kappa-B* homologue). Activation of this cascade was known to be stimulated by the upstream cleavage of a cytokine-like protein spätzle, allowing the cleaved spätzle to bind to the TOLL receptor.

Subsequent analysis of the upstream regions of known antimicrobials in *Drosophila* indicated enrichment for known NF-kB binding sites and ultimately to the identification of *TOLL*'s role in antimicrobial immunity in *Drosophila* (Rosetto et al. 1995). More detailed deconstructions of this pathway were performed in the following years, demonstrating that loss-of-function mutants in *Drosophila* had lower expression of the antifungal protein *Drosomycin* and showed significantly impaired survival under challenge with entomopathogenic fungi (Lemaitre et al. 1996). Later studies expanded the TOLL pathway's repertoire. Indicating it also played a role in combating infection with gram-positive bacteria (Gobert et al. 2003) and viruses (Zambon et al. 2005).

The TOLL receptor itself is a membrane-bound protein, featuring a single transmembrane domain, an external leucine-rich repeat region at the N-terminal and an intracellular TOLL-ILK1 receptor at the C-terminal; signalling is passed through a number of interacting partners, some known: *MyD88, Pelle, Tube* are sequentially next to TOLL in the signalling cascade, and some still obscure. NF-Kappa-B signalling is activated by the phosphorylation and degredation of the *Dif* inhibitor *Cactus*. Upon degredation of Cactus, *Dif* translocation to the nucleus will stimulate the transcription of a broad array of antimicrobial peptides (see section 2.4)

Following the identification of immune function in *Drosophila,* a number of human TLRs (TOLL-like receptors) were subsequently identified, and found to have a role in mammalian innate immunity (Medzhitov et al. 1997). There is clear conservation of the *TOLL* pathways as a signalling cascade (human homologues are known for *TOLL*, *MyD88* and *Pelle,* but not *Tube*), though the functions are not identical: no developmental role for TLRs has been identified in mammals. Differences in the mode of activation of the *TOLL*s and *TLR*s imply

that one of these functions is a more recent co-option of the pathway; cleavage of Spätzle is achieved differently during dorsoventral patterning (where a serine-protease cascade including Nudel and Snake cleave the protein) and immune activation (where cleavage is directed by *Spz*-processing enzyme '*SPE'*) (reviewed in (Valanne et al. 2011)).
There has been some debate over which is the ancestral function of the TOLL Pathway; that is whether the similar immune functions in mammals and *Drosophila* are an example of conservation or of homoplasy. It might be expected that fundamental developmental processes must be the most highly conserved, yet in fact the lack of a known developmental TLR function in mammals (Kimbrell & Beutler 2001), along with the detection of TOLL immune functions in basal metazoans such as the cnidarian *Hydra magnipapillata,* strongly support immune activation as the ancestral function of the TOLL receptors.

The publication of the *Anopheles gambiae* genome in 2002 (Holt et al. 2002) enabled the re-identification of the *TOLL* pathway in mosquitoes by homology (Christophides et al. 2002). The internal *TOLL* pathway was conserved without any duplications or deletions in ether taxa, though loss of one *Rel* was seen (*Drosophila* exhibits both *Dorsal* and *Dif* as *TOLL*-stimulated transcription factors, with *Dif* having a specific immune role and *Dorsal* expressed in both developmental and immune contexts; *Anopheles* has only the *Dorsal* homologue *Rel1*).
Activation of the TOLL pathway by antiviral, antifungal and gram-positive bacteria is also considered to be conserved in *An. gambiae* (Christophides et al. 2002), however, as with other insect species, the activation of the pathway is currently not known. The *Spätzle* processing enzyme (SPE) that cleaves the protein in *Drosophila* does not have a direct orthologue in *Anopheles,* although the spätzle cleavage site does occur in two (unrelated) *CLIP* genes, *CLIPB5* and *CLIPB38* (Waterhouse et al. 2007).

IMD

The second major antipathogenic pathway to be identified within *Drosophila* is the Immune Deficiency (IMD) pathway. First identified in 1995, a loss-of-function mutant in the IMD gene was identified in Drosophila, giving rise to reduced survival in mutant flies under bacterial challenge (Lemaitre et al. 1995). This is now known to be principally stimulated in *Drosophila* by gram-negative bacteria. Unlike the TOLL protein, *IMD* is not membrane bound, but is thought to form part of the receptor-adaptor complex with the pathogen recognition protein PGRP-LC (see section 2.3). Downstream, the *Drosophila* IMD pathway culminates in the activation of *Relish,* another NF-kappa-B transcription factor. Between these two ends of the pathway are a series of proteins including FADD (which directly interacts with IMD), TAK1,

IKK-γ / IKK-β and DREDD, which directly cleaves the long-form of Relish, removing its inhibitory C-terminal domain, allowing the active N-terminal to translocate to the nucleus and activate transcription of antimicrobial genes (summarised in Hoffman (Hoffmann 2003)). Despite some superficial similarity to the mammalian TNF-α pathway, the IMD pathway is not thought to be conserved in mammals. However all of the major components of the *Drosophila* pathway are also found in *Anopheles.* One possible novelty in the mosquito appears to be the expression of *Relish.* The *Relish* homologue in *Anopheles, Rel2,* is expressed both in the long cytosolic form, *Rel2-F,* that is activated by DREDD-mediated cleavage, and also in a shorter form *Rel2-S* that can constitutively activate antimicrobial genes (Meister et al. 2005). A constitutively expressed short-form relish has not been detected in *Drosophila*.
Whilst the *Drosophila* IMD pathway is thought to be solely induced in response to gram-negative, and not gram-positive bacteria, the two isoforms of the *Anopheles Rel-2* transcription factor appear to have distinct effects, with *Rel2-S* acting against gram-negative as in Drosophila, but the longer *Rel2-F* form also generating an immune response to the gram-positive *Staphylococcus aureus* (Meister et al. 2005).

JNK

As well as stimulating NF-kappa-B signalling, the IMD pathway is also known to diverge on the actions of *TAK1*, which as well as activating the downstream IMD pathway, also activates the c-Jun-N-terminal-Kinase (JNK) pathway. This was identified by Boutros, et al (Boutros et al. 2002) who noted initially that knockdown of *Relish* prevented activation of a number of known antimicrobial peptides downstream of IMD, but maintained a subset of cytoskeletal proteins that were also affected by IMD. Thought to be a response to septic injury, this implies a close link between immunological and damage-repair functions in the Diptera.

JAK-STAT

The third major immune pathway has, in comparison to TOLL and IMD, not been deeply investigated. The Janus-Kinase / Signal Transducer and Activator of Transcription (JAK-STAT) pathway, is another NF-kappa-B related pathway that is known to function in developmental contexts, including embryonic patterning, and later imaginal disc and wing development (Yan et al. 1996). However it is further seen to be induced in *Drosophila* upon septic injury (Boutros et al. 2002; Agaisse et al. 2003). No specificity to gram-positive or negative bacteria is seen, and Boutros *et al.* hypothesize that the injury itself may be sufficient to induce the activation of JAK-STAT.

Upon activation of the transmembrane receptor DOME, the JAK protein (*hopscotch / hop* in *Drosophila*) is dimerized and activated, phosphorylating the STAT proteins (in *Drosophila, Stat92e*). Phosphorylated STAT, in turn, form dimers and are translocated to the nucleus where they stimulate expression of a variety of cytoskeleton and damage repair genes (Myllymäki et al. 2014). In *Drosophila,* DOME is activated by ligation with one of three 'unpaired' proteins (Upd1, Upd2, Upd3). Whilst localization of two of the Upds (Upd1 and Upd3) to the extracellular matrix has been detected (Wright et al. 2011), the upstream activators of these ligands is not yet known.
The pathway is highly conserved across taxa, being present with few alterations in mammals and dipterans. *An. gambiae* shows a duplication of the STAT protein (STAT-A and STAT-B) when compared to both *Ae. aegypti and D. melanogaster,* with STAT-B acting as a regulator of the ancestral STAT-A transcription factor.
Whilst not as directly associated with parasite killing as the TOLL/IMD pathways, the JAK-STAT pathway has nevertheless been seen to control haemocyte proliferation and differentiation in *Drosophila* (see below) (Minakhina et al. 2011), and it is strongly linked to the expression of important pathways in the opsonisation of parasites for future killing (Agaisse & Perrimon 2004). It should be noted that, since many of these opsonisation pathways are both metabolically costly and toxic, the JAK-STAT pathway also stimulates transcription of the 'Suppressor of cytokine signalling' (SOCS) gene, creating its own negative feedback regulatory loop (Gupta et al. 2009).

2.3: Pathogen recognition:

In comparison to the adaptive immune system, the innate immune system encodes all of its pathogen recognition proteins in the germline. It is therefore no surprise that pathogen recognition receptors (PRRs) are a highly diverse group of proteins. There are 150 putative PRRs in the *An gambiae* genome (Das et al. 2009), most of which are of currently unknown function - perhaps because many of these are presumably highly specific.

PGRPs

Pathogen recognition in both the *TOLL* and *IMD* pathways is mediated by Peptidoglycan Recognition Proteins (PGRPs). Present in broad taxa, including both mammals and diptera, the PGRPs, as the name suggests, bind to the peptidoglycan that is present in bacterial cell walls. There are 13 PGRPs in Drosophila melanogaster, and 7 in *Anopheles* gambiae and cluster into long and short forms, known as PGRP-L and PGRP-S respectively (Royet et al.

2011). PGRP-S proteins are typically secreted, whilst the majority of PGRP-Ls are membrane bound.
The PGRP domain itself is highly specific. Individual PGRPs are able to distinguish fine differences in the composition of peptidoglycan, allowing the PGRP to distinguish between gram-positive bacteria, in which the third amino acid in the peptidoglycan stem peptide is a lysine (lys-type peptidoglycan), from gram-negative, which carry a *meso*-diaminopimelic acid in this locus (DAP-type peptidoglycan) (Royet et al. 2011).

PGRP: IMD Activation

Several PGRP genes are involved in IMD regulation, and one in particular is a constituent part of the canonical IMD pathway in *Drosophila. PGRP-LC* generates three protein products by alternative splicing (y, x and a); IMD activation is stimulated by the direct binding of monomeric peptidoglycan to a membrane-bound complex consisting of PGRP-LCa and PGRP-LCx. PGRP-LCx also forms a homodimer, which activates the pathway by recognising polymeric peptidoglycan. Another member of the family, PGRP-LF, forms heterodimers with PGRP-LCx, reducing the number of proteins available to form either monomeric or polymeric recognition complexes and acting as a negative regulator of IMD activity. Finally PGRP-LE has a dual role; the cleaved form PGRP-LEpg catalyses binding of polymeric peptidoglycan to the homodimer complex of PGRP-LCx, and the full-length form (that is neither secreted nor membrane bound) is able to directly stimulate IMD in reaction to cytosolic peptidoglycan (Royet et al. 2011). PGRP-LC, though not PGRP-LF, has a direct 1:1 orthologue in *Anopheles gambiae. PGRP-LE* has an orthologue in *Ae. aegypti,* but no gene is present in *An. gambiae* (Waterhouse et al. 2013).

The PGRP-LC orthologue in *An. gambiae* has been shown …
The PGRP-LC orthologue in *An. gambiae* has been shown to preserve the drosophilid action, generating a strong immune response to gram-negative bacteria. However in contrast, it has also been shown to induce an immune response when challenged with the gram-positive *S. aureus*, leading to increased expression of the antimicrobial peptides cecropin and defensin indicating a diversification of this role in *Anopheles*. The immune response generated by this PGRP-LC has also been demonstrated to reduce infection intensities with both *P. berghei* and *P. falciparum.*

It is notable that, when challenged with gram-negative bacteria, loss-of-function mutants of PGRP-LC in *Drosophila* have a less severe phenotype than loss of function mutants for IMD (Hoffmann 2003). This may indicate an alternative pathway for stimulation of IMD.

Knockdown of PGRP-LC in *An. gambiae* increases infection intensities of both *P. falciparum* and *P. berghei* (Meister et al. 2009).

PGRP: TOLL activation

Unlike IMD, the TOLL pathway does not contain a constituent PGRP, its transmembrane receptor is instead ligated by the cleaved spätzle protein. However that spätzle protein is cleaved by a serine-protease cascade that is itself stimulated by PGRP action. PGRP-SA and PGRP-SD are both secreted proteins in the *Drosophila* haemolymph, that are known to be upstream of the TOLL receptor. PGRP-SA (also known as *Semmelweiss*) has a 1:1 orthologue in *An. gambiae* (*PGRP-S1*), while PGRP-SD has undergone a gene duplication in both anopheline and aedine mosquitoes (Waterhouse et al. 2013). Loss of function mutations in PGRP-SA in *Drosophila* demonstrate a similar phenotype to TOLL loss of function when challenged with gram-positive bacteria (Hoffmann 2003).

The PRRs associated with melanisation responses are still unclear, however in *Drosophila* stimulation of the PPO cascade has been achieved by overexpression of both the TOLL-associated PGRP-SA (Park et al. 2007) and the IMD-associated PGRP-LE (Takehana et al. 2002) suggesting that the melanisation response is linked to both gram-positive and gram-negative responses, yet is independent of any particular immune pathway.
Seven of the *D. melanogaster* PGRP proteins have amidase function, enabling the catalysis of immune-stimulatory peptidoglycan and act to maintain a low background of peptidoglycan and prevent overstimulation of immune functions (Royet et al. 2011). PGRP-SC and PGRP-LB are secreted and perform this activity specifically in the gut; PGRP-LB has a direct homologue in *An. gambiae*.
PGRP-LCy and PCRP-LD, both of which have orthologues in *Anopheles*, are of unknown function.

GNBPs

The ‘gram-negative binding proteins’ (GNBPs) are one of the other main PRR families in dipteran immunity. Despite the name, they are implicated in defense to both gram-negative and positive bacteria. Indeed the canonical family member, GNBP1 is a co-stimulant of the TOLL pathway in Drosophila (along with PGRP-SA) leading to an immune response to gram-positive bacteria; loss-of-function *D. melanogaster* mutants show a similar loss of immunity to gram-positive bacteria as *TOLL* mutants (Hoffmann 2003).
The *Anopheles* genome contains seven GNBPs split into two distinct families, GNBP-A (2 members) and GNBP-B (five members). Both GNBP-As derive from the TOLL-activating

*Dm*GNBP1, while the GNBP-Bs are a novel expansion in the mosquito (Waterhouse et al. 2007).
A number of GNBPs are upregulated upon *Plasmodium* challenge, and GNBP4 is found to co-localize to *P. falciparum* ookinetes in the midgut (Warr et al. 2008).

CTLs

Collagenous or C-type lectins (CTLs) are a highly diverse set of proteins that share a carbohydrate recognition domain and frequently functions as PRRs across the metazoa. Two of these genes, *CTLMA2* and *CTL4,* have been shown to be essential for the killing of gram-positive bacteria in the haemocoel; RNAi knockdown individuals showing impaired clearance of *E. coli*, but not *S. aureus* (Schnitger et al. 2009).
Knockdowns of the same two genes, CTLMA4 and CTL4, have also been shown to inhibit melanisation of *P. berghei*, with knockdowns of these two genes showing fewer oocysts and a higher degree of melanisation (M. A. Osta et al. 2004). Notably this effect was seen only in *Plasmodium* parasites and not for melanisation of sephadex beads. It has been suggested that this could imply the use of these lectins by the parasite for immune evasion (Warr et al. 2006).

Immunoglobulins

The immunoglobulin superfamily (IgSF) is also active in *An. gambiae* immunity. Of 138 genes found in the mosquito, 85 are upregulated following immune challenge (Garver et al. 2008). Six of these ‘immune responsive immunoglobulin domain’ (IRID) genes were tested for bacterial challenge, with three (IRID-3,5 and 6) being implicated in resistance to both gram-positive and gram-negative bacteria. A further two IRIDs, IRID-4 and IRID-6, are limiting factors for *Plasmodium falciparum* infection (Garver et al. 2008).

One member of the IgSF in particular appears to code for a highly polymorphic set of transcripts and may have a significant role to play in mosquito immunity. The ‘Down syndrome cell adhesion molecule’ AgDSCAM is a hypervariable gene Containing 101 exons, the gene is potentially able to produce 31,000 different isoforms via alternative splicing (Dong, Taylor, et al. 2006). RNAi-mediated knockdowns have demonstrated a role for AgDSCAM in the response to bacterial challenge, binding to invasive bacteria before eventual phagocytosis. Knockdown individuals also demonstrate compromised resistance to *Plasmodium falciparum* and *berghei.* Interestingly, upregulated transcripts appear to be specific to infection type, with different exons upregulated in response to gram-positive and gram-negative bacteria (Dong, Taylor, et al. 2006).

Other PRRs

Given the 150 potential PRRs in the *Anopheles* genome, and the relatively small number that Have been shown to have a specific function, there are undoubtedly many others still to be found. Indeed known PRRs from Drosophila are not present in *Anopheles*, even in cases where the immune response is conserved, implying a substitutive role for other *Anopheles* genes. A good example of this is the *Persephone* protein that initiates the TOLL immune response to fungal infections in *Drosophila; Persephone* has no 1:1 orthologue in *Anopheles,* but has instead expanded to four different CLIP-C genes (see section 2.6) whose interaction with TOLL is not known. Indeed the broad array of pathogens to which the innate immune system can respond, and the apparent fine distinctions that can be made, imply the existence of numerous detection and modulatory pathways. We can therefore expect a plethora of PRRs that are yet to be described.

2.4: Effectors:

AMPs

Anti-microbial peptides (AMPs) refers to low-weight, secreted proteins that have innate antimicrobial activity (as compared to those proteins that opsonise or tag the pathogen for destruction by another method).
There are four principal classes of these within dipterans: Attacins, Defensins, Cecropins and Diptericins, of these Attacins and Diptericins have lost most of their orthologues in *Anopheles* gambiae, demonstrating only one family member each, while the Cecropins and Defensins have both undergone mosquito-specific family expansions (Waterhouse et al. 2013).
All of these genes are upregulated in the fat body upon pathogen challenge, and there is significant overlap in those that are linked to the TOLL and IMD pathways – many are seen to be upregulated by both (Hillyer 2010).
Of those that have been tested for their effects on *Plasmodium* parasites, cecropin appears to have a limiting effect on *P. berghei* infection intensity (Kim et al. 2004) and the mosquito-specific AMP *Gambicin* appears to have anti-*P. berghei* activity when expressed in the midgut (Dong, Aguilar, et al. 2006).

Melanisation

The deposition of melanin on invasive parasites is an immune effector mechanism that is frequently identified in mosquitoes in response to larger pathogens, such as Plasmodia and filarial worms.

The melanisation reaction is begun by a PRR stimulating a serine protease cascade that ends in the conversion of prophenoloxidase (PPO) to phenoloxidase (PO) by site specific cleavage. PO then catalyzes the conversion of Dopa (3,4-dihydroxyphenylalanine) to melanin by one of two pathways via dopamine or dopaquinone; both pathways are thought to be involved in the immune response (Hillyer 2010). Although nine PPOs are present in the *An. gambiae* genome, no activating (i.e. PPO-cleaving) enzyme has been identified. PPO-activating enzymes in other species are typified by a clip domain, however none of the 54 identified CLIPs in the *An. gambiae* genome (Waterhouse et al. 2013) have been confirmed as the PPO-activator. As such the cleavage sites for the 9 PPOs also remain unidentified (Cerenius & Söderhäll 2004).

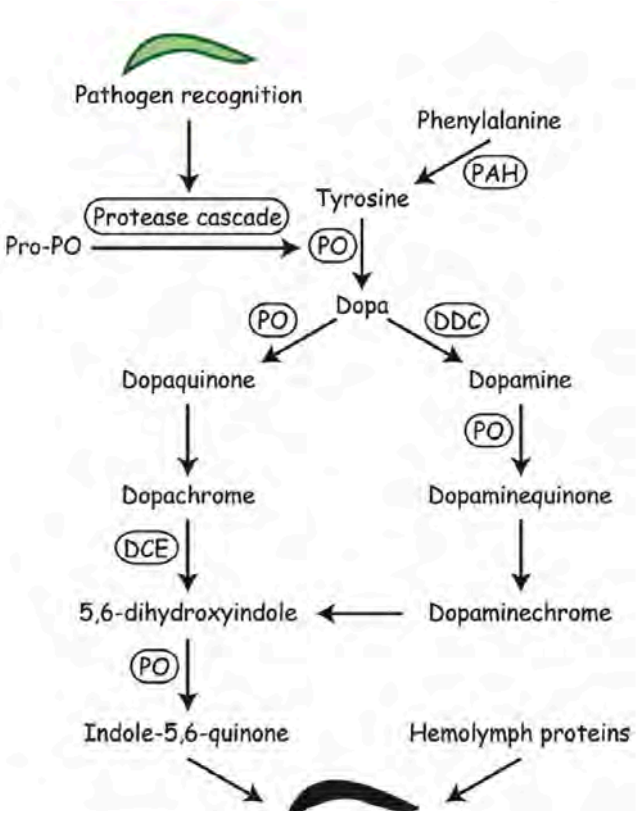


Figure 2.3 :

A proposed model for the PPO Cascade leading to melanisation of larger parasites in *Anopheles gambiae*.

Source: Invertebrate immunity: ch12: Mosquito Immunity (Hillyer 2010)

Although melanisation is frequently associated with pathogen killing, there is also some debate over the relative importance and the precise method of this. Volz *et al.* have demonstrated that, in the *Plasmodium*-susceptible G3 strain of *Anopheles,* melanisation is required for parasite killing. While in the refractory L3-5 strain killing is achieved by another method, and melanisation serves only as a clearance mechanism to dispose or sequester dead parasites (Volz et al. 2006).

Various methods have been suggested for the melanin/PPO-mediated killing of parasites, including cellular damage due to oxidising intermediates of the PPO cascade, or of starvation as the parasite is encapsulated from the nutrient-rich hemolymph (Hillyer 2010). However the large number of modulators of this pathway (see section 2.6) and the apparently complex relations between melanisation and parasite and vector species make reconstruction of this mechanism problematic.

It should be noted that, although the melanisation cascade is extracellular, many of the constituent enzymes are, in fact, expressed and secreted from hemocytes.

Cellular responses

Phagocytosis is a major component of the antimicrobial immune response, and one that is highly conserved across the metazoa. In all species, once identified by a PRR, the pathogen is bound by the plasma membrane and internalised into a phagosome; the phagosome then fuses with lysosomes and lytic enzymes dispose of the pathogen. The precise cells reported to be phagocytic differs between species, and frequently contradictory reports can be found within the same species (Lavine & Strand 2002), most likely due to the difficulties in identifying different cell types. Further difficulties also arise, since the nomenclature of Drosophila haemocytes differs from other insects for historical reasons.
Nevertheless studies of hemolymph from *Anopheles* and *Aedes* mosquitoes have identified three principal hemocyte types: granulocytes, oenocytoids, and prohemocytes. Prohemocytes, due to their undifferentiated morphology and lack of labelling by functional markers are believe to be progenitor cells for the other two forms of hemocytes.

The majority of haemocytes in anopheline and aedine species are granulocytes, comprising over 90% of the total hemocyte population (Castillo et al. 2006), and these are known to be phagocytic, and required for the encapsulation of foreign targets. Hemocytes are efficient parasite killing mechanisms: individual hemocytes are believed to be able to phagocytose over 1000 bacteria per hemocyte within 24 hours of infection (Hillyer 2010).

The hemocytes are also responsible for the expression of many of the components of the serine protease cascades required for immune signalling to the fat body, and for the expression of proteases associated with the melanisation response. Oenocytoids are responsible for constitutive phenoloxidase activity, whilst the phenoloxidase activity of granulocytes is heavily upregulated following immune challenge (Castillo et al. 2006).

Hemocyte proliferation is strongly stimulated after the mosquito bloodmeal, and has been implicated in the phenomenon of immune priming: despite not having a classically ‘adaptive’ immune system, mosquitoes subjected to bacterial challenge will demonstrate reduced susceptibility to a second challenge. Examinations of the hemocyte types before and after primary challenge demonstrated a 3.2-fold increase in the granulocyte population, and a 10% fall in the prohemocyte population; strongly suggesting a process of prohemocyte differentiation into granulocytes (Rodrigues et al. 2010).

It has also been suggested that the granulocytes actually differ on the basis of their serine protease expression, and may form finer groupings of cell types than we can currently detect (Castillo et al. 2006). This may underlie the apparent pathogen-specificity of the hemocyte priming effect, as has been observed in other insect species (Pham et al. 2007; Roth et al. 2009).

2.5: Opsonisation

While hemocytes themselves remain somewhat underexamined, the signalling and opsonisation mechanisms that stimulate parasite killing have been greatly illuminated in recent years.

LRRs

The thioester-containing protein TEP1 was originally identified due to it's homology to human complement protein C3. Experiments in a hemocyte-derived cell line indicated it localised to both gram-positive and gram-negative bacteria, and this localisation was seen to promote phagocytosis *in vitro* (Levashina et al. 2001).
It was subsequently found to be upregulated after *Plasmodium* (*berghei*) challenge, and RNAi-mediated knockdown of the gene caused a significant increase in parasite load (Blandin et al. 2004). It is further seen to localise to killed and melanised *Plasmodium* parasites (Povelones et al. 2009) acting as an opsonising agent for an as-yet-unknown killing mechanism.

In both mammals and dipterans the thioester domain mediates binding to non-self and self proteins (complement factors). This highly reactive domain, once exposed, can spontaneously hydrolyse leading to covalent binding to its substrate. Regulation of the cleavage and localisation of the TEP1 protein is therefore paramount. In order to prevent the premature or overactivation of immune processes, the thioester domain is obscured by an inactive tail, that must be cleaved before binding can occur. TEP1 is therefore secreted as a long-form TEP1-f and cleaved to TEP1-c before binding.

Cleavage of TEP1 is mediated by a pair of leucine-rich-repeat (LRR) proteins. LRIM1 (leucine rich immune protein) and APL1C (*Anopheles Plasmodium*-responsive leucine rich repeat) circulate in the hemolymph as a disulphide-bonded complex, and isolation of this complex from the hemolymph yields a fluid highly enriched for TEP1.

Knockdown of either of these regulating proteins causes a drop in TEP1-c, but not a corresponding rise in TEP1-f, as might be seen if the LRIM1/APL1C complex was required for cleavage. Instead is it believed that the complex is required for stabilisation of the highly reactive TEP1-c after cleavage, preventing its hybridisation to self surfaces and directing its activity to the parasite surface (Povelones et al. 2009).

Transcriptional regulation of this complex and the cascade that activates TEP1 cleavage is still unclear. Comparisons of rel1 and rel2 knockdowns, both individually and together, did not show any significant impairment of TEP1 induction upon parasite challenge (Frolet et al. 2006). However Frolet *et al.* also determined that the gene has a distinct pre and post-invasion profile. Knockdowns of NF-kappa-B transcription factors prior to invasion had a drastic effect on the ability of the mosquito to counter infection, and knockdown of the *rel1* inhibitor *cactus* rendering the mosquito entirely refractory, however knockdowns of *rel1* and *rel2* post invasion did not have any effect on *Plasmodium* development. The authors conclude that basal expression of TEP1, prior to parasite invasion, is more important than it's replenishment post-invasion, and that the secretion or activation of previously expressed TEP1 is, in fact, the more important immune response (Frolet et al. 2006). Indeed depletion of cactus, despite causing an increase in TEP1 expression, did not lead to any significant impact on oocyst development, strongly indicating that the TEP1-mediated immunity is effective against ookinetes, but does not affect oocysts.

TEP1 is constitutively expressed and found permanently in the haemolymph in its full form; that constitutive expression has been shown to be reliant on combined *rel1* and *rel2* activity (S. A. Blandin et al. 2008). Transcription of TEP1 is also significantly upregulated following septic injury, and loss-of-function mutations in *DmHop,* (*Hopscotch* - the JAK-STAT pathway's janus kinase) demonstrate significantly decreased *TEP1* expression, indicating a likely JAK-STAT stimulation of the pathway under immune challenge (Agaisse & Perrimon 2004). In addition, JNK pathway silencing leads to severed reductions in *TEP1* mRNA levels, but not of the other hemocyte specific genes in the complex (LRIM1/APL1C). It could be surmised that the downregulation under JAK-STAT knockdown is related to a lack of hemocyte proliferation, as has been seen in drosophila (Minakhina et al. 2011). Further dissection of this pathway will no doubt highlight the interrelatedness of these functions.

ROS / RNS

The high reactivity of the TE domain might suggest that TEP1 should bind indiscriminately once cleaved. That it doesn't implies the existence of a further regulatory framework for its activation or localisation.
It is now clear that TEP1 circulates as a stable complex along with LRIM1/APL1C, however neither of those proteins appear to have specific binding domains for any pathogen-associated patterns. It is unclear, therefore, what determines TEP1 release from the APL1C/LRIM1 complex, or binding to the surface of parasite.

The parallels with mammalian immunity, in which binding to IFN-γ and stimulates production of nitric oxide synthase (NOS) via the JAK-STAT pathway, has led researchers to investigate the role of reactive oxygen and reactive nitrogen species (ROS and RNS) in dipteran immunity.
Cells that have been physically damaged mount a powerful immune response involving the expression of high levels of nitric oxide synthase (NOS) – an enzyme that catalyzes the production of reactive nitrous oxide from L-arginine. In the mosquito midgut the further expression of heme peroxidase (HPX2) and 'nicotinamide adenine dinucleotide phosphate oxidase 5' (NOX5) act to increase the oxidative pressure and potentiate the toxicity of the NOS-derived nitric oxide.
HPX2 is also seen to be upregulated in the mosquito midgut after uptake of a bloodmeal regardless of whether that bloodmeal is infected or not; strongly indicating a general immune / defence response. At high levels ROS / RNS can be naturally cytotoxic, causing widespread damage to both self and non-self cells, however its deployment in *Anopheles* appears to be more complex than simply generating oxidative stress.
Both cellular rupture and the immune response act to increase the levels of NO toxicity in the mosquito and will damage bacterial and larger pathogens that attempt to traverse the midgut wall. It is not surprising therefore that the RNAi-mediated knockdown of HPX2 would increase parasite numbers under immune challenge with *Plasmodium* – providing a more amenable environment for their survival. Knockdowns of *immunomodulatory peroxidase* (IMPer) that increase oxidative stress also lead to a similar reduction in parasite numbers (Kumar et al. 2010).
In comparison, knocking down TEP1 interrupts a specific immune response and consequently has an effect of greater magnitude. Yet if these processes were independent we might expect their effects to be additive, however knockdown of HPX2 along with TEP1 has an identical effect to TEP1 alone, strongly suggesting that this is an essential step in the

TEP1-mediated immune response. This effect was confirmed by dual knockdowns of IMPer / TEP1 that entirely ablated the previous reduction in parasite numbers (Oliveira et al. 2012).

The combination of these effects appears to indicate that TEP1 activity – whether upstream of TEP1 cleavage, or at the point of TE-binding – is dependent on prior nitration of the parasite surface. Indeed high correlation is seen between the levels of midgut nitration and the number of surviving parasites. Moreover the L3-5 strain of mosquito – known to be highly resistant due to a powerful TEP1-mediated melanising response – is actually in a state of permanent oxidative stress; an effect that could provide an explanation for its more aggressive immune response (Kumar et al. 2003).
In this model the NOS system would therefore act as a tagging system, increasing the visibility of the invading pathogen and opsonising it for eventual TEP1 binding leading to either lysis or melanisation. As a result, regardless of the severity of the TEP1-mediated response, the degree of immunity would instead depend on the severity of the cellular ROS response, along with the amount of time the parasite spent in the NO-rich environment.

This would also indicate that the complement response does not begin when encountering TEP1, upon first contact with the hemolymph, but instead begins with the first interruption of the midgut cell wall. Ookinete invasion of the midgut epithelia rapidly increases the level of NOS production, and causes cellular damage that ends in apoptosis; the nitration of the ookinete tags the parasite as having been closely associated with cellular damage or death and marks it for destruction once complement factors are encountered.

ROS / NOS regulation

This also highlights the regulation of ROS production as a constitutive immune response and implicates the JAK-STAT and JNK immune responses in pathogen resistance. The JNK pathway can be activated by multiple input streams – being stimulated directly by cellular damage, or as a ‘branch’ of the IMD signalling cascade. This pathway is strongly stimulated by bacterial infection in Drosophila and has been directly associated with the upregulation of HPX2 and HOX5 in the midgut, beginning the opsonisation process (Garver et al. 2013). When combined with the previously noted stimulation of TEP1 expression this provides a coordinated JNK-derived response of nitration and opsonisation that is stimulated at the very earliest stages of pathogen invasion of the midgut epithelium.

However it has been widely believed that, once the oocyst matures and modifies its surface it is relatively invulnerable to the mosquito immune system. While it certainly appears that JNK-mediated immunity takes no further part in the immune response, the JAK-STAT pathway

has recently been implicated in a separate immune response against late-stage oocysts. Knockdown of either STAT-A or STAT-B does not lead to a significant drop in survival under bacterial challenge, nor does it reduce early stage oocyst or ookinete numbers. However the STAT-A knockdown does decrease NOS production after bacterial challenge and demonstrates a significant increase in the survival of late stage oocysts as compared to GPF knockdowns.
Strangely, not only do JAK-STAT deficient mosquities show no decrease in early-stage oocysts, they in fact show an increase, suggesting perhaps that JAK-STAT response is beneficial for epithelial traversal (Gupta et al. 2009). This effect may be related to the stimulation of SOCS, the 'suppression of cytokine signalling' gene that restricts the cytotoxic effects of the JAK-STAT effectors; after immune challenge SOCS mRNA levels are seen to peak at 3 hours post infection, earlier than NOS levels which peak at 6 hours post infection (Gupta et al. 2009).

2.6: Modulation

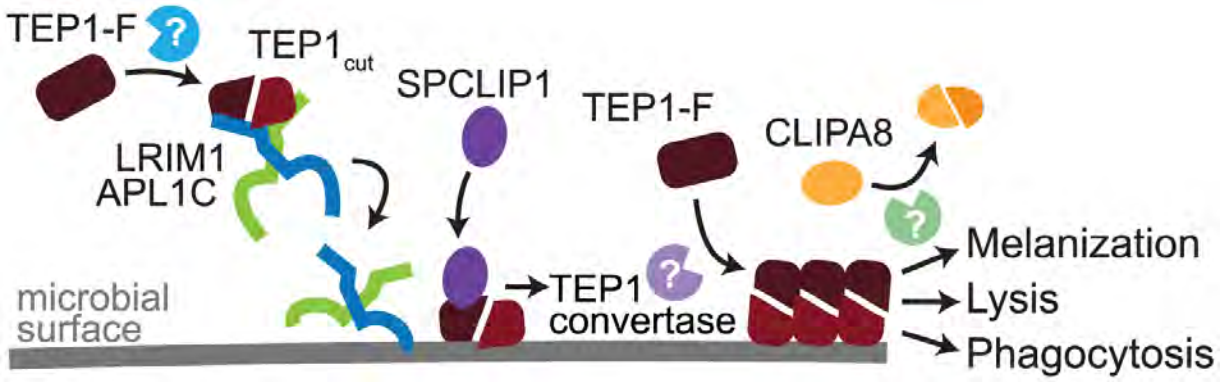
Srpns / clips

What might be termed the modulatory families, SRPNs and CLIPs will almost inevitably include some proteins that are key initiators of some of the major signalling cascades.
The Serine Protease Inhibitors (SRPNs) are a diverse group of proteins found in all higher eukaryotes, frequently with immune functions. 18 SRPNs exist in *Anopheles gambiae*, most of which are conserved across the culicidae, but with few direct orthologues in *Drosophila*. Much like the CLIPs, SRPNs are pressed into service in positive and negative regulatory functions, with SRPN2 appearing to repress the PPO cascade, and SRPN6 promoting it (S. a Blandin et al. 2008).

The two families of CLIP-domain serine proteases (CLIPs), CLIP-A and CLIP-B, are similarly diverse, adapting to a number of roles in each pathway. The prophenoloxidase cascade is largely composed of CLIPs, meaning that their level of expression can have a drastic effect on the immune response by interruption the melanisation response. However it is difficult to predict function from sequence homology alone.
CLIPA8 appears to be essential for PO cascade activation, whereas the cascade is suppressed by CLIPA2, CLIPA5 and CLIPA7. Other members of the family have more moderate effects; CLIPB3, CLIPB4 and CLIPB17 are not essential for melanisation, but knockdowns of these genes reduces its intensity. Outside of the PPO cascade, CLIPB14 and CLIPB15 appear to promote parasite lysis (Barillas-Mury 2007).

A novel CLIP-E family member, SPCLIP1, has recently been identified from EST sequence and is not represented in the PEST genome. Interestingly *SPCLIP1* appears to have a crucial role in recruiting further TEP1-c to parasite surfaces after initial binding of TEP1. That is, after the binding of TEP1 via the TEP1/APL1C/LRIM1 complex, SPCLIP1 is recruited to the parasite surface and, via an unknown intermediate, stimulates the further cleavage of TEP1-f – enhancing the TEP1 optimisation and tagging the pathogen for killing. In this sense it performs a role analogous to mammalian convertase (Povelones et al. 2011).

Figure 2.4:



Proposed model of the convertase-like complex of TEP1/LRIM1/APL1C: the LRIM1/APL1C pair act to stabilise the highly reactive TEP1 in the hemolymph and direct it's action to the parasite surface. SPCLIP1, via a currently undetermined method, induces additional cleavage of TEP1-F stimulating further localisation of TEP1 to the parasite surface and leading to parasite killing by various methods.

Source: The CLIP-Domain Serine Protease Homolog SPCLIP1 Regulates Complement Recruitment to Microbial Surfaces in the Malaria Mosquito *Anopheles gambiae* (Povelones et al. 2013)

2.7: Other effects:

Part of the difficulty of identifying immune pathways arises due to the inherent complexity of the relationships involved. Although *Plasmodium* quite naturally forms the focus of *Anopheles* research, the immunological arms race between these two species is played out in a background of parallel infections and alternative requirements for the mosquito.

Nutrient availability / vitellogenin

It shouldn't be forgotten that the bloodmeal is both source of infection and essential for the mosquito reproduction, and that an immune reaction that severely depresses reproductive capacity is likely to prove more detrimental than the presence of the parasite. It is also important to note that, since the major expansion occurs after traversal of the midgut epithelium, the growth of the parasite is dependent to a large degree on the availability of nutrients in the hemolymph and not in the bloodmeal.
The nutrient transport proteins Lipophorin (LP) and Vitellogenin (VG) are both upregulated after a bloodmeal, and vitellogenin – as the name suggests – is implicated in the oogenesis cascade. As a result, both of these genes have both been investigated for their roles in supporting Oocyst growth. Lipophorin, the first of these genes to be investigated, has been shown to affect oocyst growth in both *Plasmodium berghei* (Vlachou et al. 2005) and *Plasmodium falciparum* (Mendes et al. 2008).
Subsequent similar studies on vitellogenin demonstrated a similar effect, though of greater magnitude and with no additive effect from dual LP + VG knockdowns (Rono et al. 2010). As well as a reduction in the number of oocysts, lipophorin knockdowns cause a reduction in the mean size of oocysts; it is probable therefore that LP is a lipid source for growing oocysts. Silencing of LP also caused a noticeable reduction in VG mRNA expression – suggesting that nutrient availability is an important stimulant of oogenesis, and reduced the number of mature eggs produced by the mosquito. Knockdown of *vitellogenin* had a similar effect on mosquito oogenesis.

However whilst this could be explained through a simple lack of lipid availability for the parasite, there are also indications that a more nuanced interaction with the mosquito immune system is taking place. Although it affected Oocyst number, knockdown of VG alone did not significantly affect oocyst size, implying nutrient availability was not at the root of the reduction in oocyst numbers. Moreover knockdowns of VG/LP along with TEP1 demonstrated an ability to rescue the phenotype, generating comparable oocyst numbers to knockdown of TEP1 alone. This suggests that the action of VG impinges more closely, possibly directly, on the TEP1-mediated killing mechanism.

Vitellogenin has shown significant pleiotropy in other species, and has demonstrated a number of roles that, if maintained in the mosquito, could greatly affect the immune response. In the honeybee, *Apis meliffera,* vitellogenin has been shown to be capable of reducing oxidative stress by scavenging free radicals (Li et al. 2008), which could impact upon known opsonisation effects of NOS (see above). However, a more direct opsonisation

effect is also seen in other species. Vitellogenin from the greenling fish *Hexagrammos otakii has been shown to* bind to pathogen-associated macromolecules including peptidoglycan, glucan and laminarin. Moreover VG binding was shown to have an opsonising effect under a variety of immune challenges – increasing phagocytosis under gram-positive, gram-negative and fungal challenge (Li et al. 2008). If maintained in the mosquito, it is possible that this VG binding could misdirect the immune response or non-productively compete with TEP1 binding to the oocyst surface. Alternatively, the parasite could recruit available lipids to alter its lipid membrane, or increase the rate of membrane renewal, reducing the visibility of the oocyst to TEP1-mediated opsonisation. As noted above, the oocyst binding partner of TEP1 is not known, however hydroxyl residues on surface lipids could covalently bind to the TEP1 thioester domain.

It is clear, therefore, that the mosquito must balance its vector capacity with its reproductive capacity, and linkage between the post-bloodfeeding immune and the post-bloodfeeding oogenesis mechanisms have been seen. Overactivation of the TOLL pathway by knockdown of the rel1-repressor cactus leads to a reduction in vitellogenin expression in the yellow fever mosquito *Aedes aegypti* (Bian et al. 2005), suggesting that the mosquito undergoes a reduction in oogenesis after detection of an infected bloodmeal. TOLL activation therefore affects *Plasmodium* and other parasites in two ways: through an increase in the expression of the directly opsonising TEP1, and via a reduction of the levels of vitellogenin in the hemolymph that might interfere with that TEP1 binding. This may provide a selective pressure to reduce TOLL activation for intense infections. Whether this phenomenon also represents a ‘starve a fever’ response from the infected mosquito bears further investigation.

2.8: Anti-*Plasmodium* immunity

Balance & specificity:

I: Which pathway?

It is clear that the mosquito immune system has adapted to a variety of pathogens, and that, depending on what protein initially recognises the pathogen it can activate one of several different immune responses.
However the classical model that is often put forward; that one distinct pathway is activated for each type of pathogen might be an oversimplification. In *Drosophila*, double TOLL and IMD (loss-of-function) mutants are more susceptible than a single IMD mutant to a challenge with *E. coli.* Similarly loss of TOLL function renders Drosophila susceptible to (gram-positive)

Pseudomonas spp. (Brennan & Anderson 2004). Both of which would suggest, at least, a broader acting role for TOLL immunity than merely gram-negative and fungal infections. Indeed the antimicrobial peptides stimulated by the TOLL and IMD pathways overlap significantly, and the expression of the TEP1 complement factor has been linked with the action of each of the major immune pathways.

Direct comparisons of the efficacy of the TOLL/IMD pathways have been performed via comparative knockdowns of the REL1/REL2 transcription factors. Despite reconfirming the roles of the JAK-STAT and TOLL pathways in pathogen resistance, these comparisons indicate that IMD response is by far the more effective at reducing oocyst numbers (Garver et al. 2012).
This reaction was particularly effective against the early-stage infection; against ookinete and possibly gametocyte stages of the parasite, concurring with the earlier results of Frolet *et al*. Indeed the concept of IMD as a rapid-reacting pathway is not new; expression studies of bacterial challenge in *Drosophila* show an enrichment of IMD components in the early stages of the immune response, with TOLL taking over as the dominant response afterwards (Boutros et al. 2002).
The anti-*Plasmodium* response should perhaps more correctly be seen as a succession of responses: basal immune expression providing the initial response, before the classical immune pathways, along with the JNK pathway, are activated upon penetrance of the peritrophic matrix and midgut epithelia, and JAK-STAT being stimulated by later oocyst formation.

The response stimulated appears also to vary depending on the circumstances; the high oxidative potential of the L3-5 strain leading to a melanisation-dominant response, whereas the response in the G3 strain is lytic. Further dissection of the modulatory cascades will surely uncover some of the triggers that underlie the selection of one clearance mechanism over another, and perhaps most importantly, which of these is genuinely of primary importance in parasite killing.

II: Which malaria?

One of the reasons these relationships are so difficult to tease out is the wide variety of response that are seen with different parasites – not only gram-positive / gram-negative bacteria, but different species of *Plasmodium* can stimulate vastly different immune responses with apparently distinct regulatory frameworks.

The majority of the reverse genetics screening that has been performed in this system over the past decade has been in the *Anopheles gambiae* – *Plasmodium berghei* pair. However results frequently differ when experiments are performed in the medically important *A gambiae* – *P. falciparum* pair. Of the various genes mentioned above that have been tested with both pathogens, most have not shown the same effect in both species. CTL4, CTLMA2, SRPN2 and even LRIM1 demonstrate a clear immune role against *P. berghei*, but no role in *P. falciparum* killing. In fact, only TEP1 and SPCLIP1 – what might be termed the core genes in the complement system – maintain the same roles in both parasites (Levashina et al. 2001).

It is not just these 'modulatory' genes that differ in their effects to *Plasmodium* species, differing responses are apparent in entire pathways when challenged with human and murine malaria parasites. The TOLL pathway is seen to be activated in response to both plasmodia, having been tested in response to *P. berghei* (Frolet et al. 2006) and *P. falciparum* (Garver et al. 2009), yet comparative knockdowns affecting the REL1 and REL2 transcription factors demonstrate that it only appears to *dominate* the immune response when the mosquito is challenged by *P. berghei*. In contrast, the immune response to *P. falciparum* is dominated by the IMD pathway (Garver et al. 2012).
The two isoforms of rel2, as well as reacting independently to gram-positive and negative bacteria, appear to demonstrate different responses to human and murine malarias, with only the short-form rel2-S conferring protection to *P. falciparum* (Mitri et al. 2009).
Moreover, the APL1C/TEP1/LRIM1 cascade is also apparently not effective against a wide variety of pathogens. Although TEP1 and LRIM1 maintain their effectiveness against both species of *Plasmodium*, APL1C does not. Knockdowns of the three members of the gene family (APL1A / APL1B / APL1C) show differing effect on each parasite: with APL1C active against *P. berghei* and *P. yoelii. P. falciparum* appears to be targeted by APL1A, and the pathogen targeted by APL1B is as yet unknown.
It is tempting to speculate that the combinations of APL1 genes and the other LRIMs (part of a gene family encompassing at least 20 genes (Waterhouse et al. 2010) might seem to enable the mosquito immune system to discriminate between infections – directing the reactive TE opsonin towards a variety of different pathogens. Definitive identification of the APL1A binding partner would lend credence to this theory.

Further complexity creeps in (and casts some doubt on the specificity of the APL1A/C pairs) when investigations are performed on infections of different intensity. By infecting mosquitoes with differing concentrations of gametocytes Garver *et al.* were able to generate predictably varying numbers of oocysts using the same parasite clone.

Far from showing uniformity of effect against this single clone, clear differences were apparent between APL1 family members at different infection intensities – APL1C being most effective at low-level infections, APL1B being most effective at median infection intensities, and APL1A being ineffective at any level (Garver et al. 2012). Whilst this in no way negates the conclusions of Mitri *et al.* in deriving the species-specificity of APL1 paralogues (APL1A having shown no effect on *P. berghei* in either experiment), it does highlight the difficulty of drawing definitive conclusions based on small numbers of experiments.

Allelic differences

Even within individual effector genes themselves, and within single parasite challenges, significant differences in efficacy have been seen between different alleles. The APL1A gene has been found to have three distinct alleles within one colony. These alleles, termed APL1A-1,2,and 3, were not only seen to have different protective profiles ($APL1A^3$ having a greater protective effect than the other two alleles) but the transcripts were sufficiently different to have different cellular localisations, with the largest allele localising within cytoplasmic vesicles rather than being secreted.

Perhaps the most drastic allelic difference is within the TEP1 gene itself. The TEP1 locus is highly polymorphic, and may in fact have derived from successive gene conversion events, with other TEP loci (Obbard et al. 2008). This has resulted in alleles falling into two broad classes depending on which gene conversion event they are most closely related to. Comparisons between TEP1 alleles within the reference PEST strain and the refractory L3-5 strain indicated that these two groups have significantly different protective profiles when challenged with *P. berghei.* Indeed, the resistant and susceptible alleles (TEP1-R and TEP1-S respectively) are sufficiently diverged that RNAi primers can be designed to knockdown specific alleles whilst leaving the other untouched.

In a series of F1 and F2 crosses of R/S lines, *Blandin et al.* illustrated the significant differences in protective capacity of these two TEP1s when challenged with *P. berghei* – presumably related to the differing binding efficiencies of the TE domain to this parasite.

It should be noted, however, that assays of TEP1 alleles in the wild actually show that the nominally susceptible allele is still maintained at high frequencies in many populations (White et al. 2011); it seems probable that the nominally resistant allele is of lower efficiency when faced with alternative pathogens in the wild.

2.9: Immune families

As well as illuminating the various pathways that are employed in the immune process against the parasite, the past ten years has seen a massive leap in our understanding of what we don't know. The precise mechanism of TEP1-mediated killing remains a mystery, along with the factors that potentiate one killing mechanism over another. There is also a wealth of immune factors that have been identified in the *berghei* response, but are yet to be identified under *falciparum* challenge. Perhaps chief amongst these are the TEP1-convertase mechanisms for *Plasmodium falciparum* and other parasites.

The prior concentration on comparative genomics contains a distinct weakness when investigating host-pathogen interactions that are not replicated in the *Drosophila* model. Although great advances have been made dissecting the specific actions of the innate immune system that are unique to *Anopheles*, this does not obviate the need for hypothesis-free methods of investigating innate immunity.

Nevertheless, whilst not ruling out the possibility of important non-orthologous genes, it should be borne in mind that the typical modes of gene expansion and adaptation make it likely that novel mechanisms will be found in related gene families. Some of the larger families of immune genes, many with as-yet unknown function are described in Appendix A2.

2.10: Future Approaches

While a great deal has been learnt about the innate immune system over the past decade, there are still significant holes in our understanding. The killing mechanism for *Plasmodium* is not known; the dominant pathway used against malaria is still not clear; and the potential reasons for immune evasion by some parasites over another have only just begun to be discerned.

Perhaps more importantly, the apparent major differences in immune processes between mosquito colonies and parasite species makes the use of models problematic. It is difficult to be sure that results derived from infections with murine malaria are applicable to human malaria. More subtly, there have been indications that many of the immune processes modelled in the lab are in fact less important in the field (Cohuet et al. 2006). Although assaying directly in the wild is infeasible, this highlights the importance of testing results in colonies that are at least close to the wild; that is, colonies that are recently founded or of high diversity (see section 5.4). For the same reasons, whilst QTL approaches between

inbred colonies are valuable, and have identified some of the major interacting partners identified above (Blandin et al. 2009; Riehle et al. 2007), they will not necessarily identify the same genes that are important for mosquito immunity in the field; mapping in wild or semi-wild samples will also be of importance.

Chapter 3: Population structure

3.1: Introduction

As well as being a vector of public health importance, over the past decade *Anopheles gambiae* has become an important model for the study of speciation. *An. gambiae* exhibits a highly complex population structure both at the species (Lawniczak et al. 2010; Della Torre et al. 1997; Wang-Sattler et al. 2007) and population level (Riehle et al. 2011; Coluzzi et al. 2002) and this can have a drastic effect on control methods and particularly on attempts at genomic mapping.

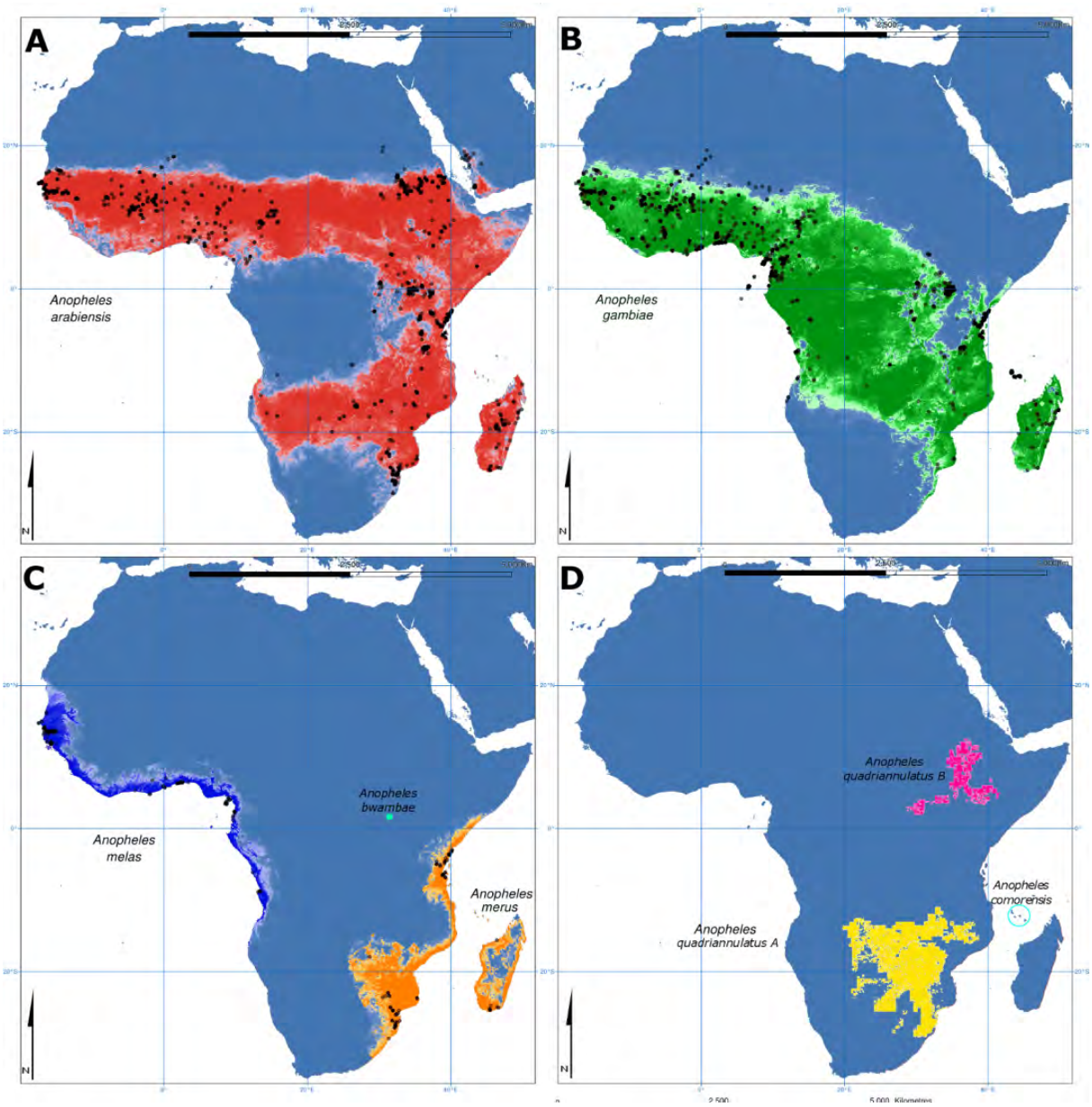
An. gambiae is one of a number of closely related species, though to have diverged extremely recently. Apparent species and population-specific niche expansion has allowed it to occupy almost the whole of sub Saharan Africa. In addition to this, Anopheles gambiae is also found to be undergoing an active speciation event, with the lineage dividing into *An. gambiae s.s.* and *An. coluzzii* (Lawniczak et al. 2010). Finally, at finer scale the mosquito appears to have propensity for developing complex population structure along ecological or behavioural lines – frequently associated with large structural variations (White et al. 2009). The interrelations between these speciation events and the ecological specialisation are currently uncertain, although examination of the genome is beginning to illuminate some of the connections.

3.2: *Anopheles gambiae* species complex

At the species level, *gambiae* exists at the centre of a cluster of recently diverged species. Believed to be a single species up until 1956 (Davidson 1956), F1 hybrid infertility between chromosomally and environmentally differentiated mosquitoes eventually identified 7 distinct species. Most of the seven, referred to collectively as *Anopheles gambiae sensu lato*, are indistinguishable as adults, but they exhibit distinct larval morphologies, behaviours, and - particularly prominently - habitat preferences. *Anopheles merus* and *melas* are coastal species that are resident in (saline) mangroves in East and West Africa respectively. *An. quadriannulatus* are found in the south of Africa in freshwater larval habitats, whilst *An.amharicus* (formerly known as *quadriannulatus* 'B') is found in East Africa only, almost exclusively in Ethiopia (Coetzee et al. 2013). Both of the former *quadriannulati* are zoophilic. *An. bwambae* has an even more highly restricted range, existing only nearby geothermal springs in the semliki forest in Uganda (White 1985). Finally there are the two most

widespread species: *An. arabiensis* and *An. gambiae*. Both found continent wide, frequently in sympatry, and are respectively highly and near-exclusively anthropophilic. Both are major vectors of malaria. (See Figure 3.1).

Figure 3.1: Continental Distributions of *Anopheles gambiae s.l.* species



Continental distributions of the species comprising the Anopheles gambiae species complex, showing the vastly expanded distribution of the species with the greatest number of inversion polymorphisms: A) Anopheles arabiensis, B) Anopheles gambiae / coluzzii C) The saltwater tolerant Anopheles merus, melas and bwambae D) Anopheles quadriannulatus, amharicus (previously known as quadriannulatus B) and comorensis.

Source: Lee / Lanzaro - The Distribution of Genetic Polymorphism and Patterns of Reproductive Isolation Among Natural Populations (Lee & Lanzaro 2013)

It is not just the lack of morphological distinction that has prevented many of these relationships being discerned until now. Genetic relationships are also far from clear. Whilst species boundaries are well established in the literature, they are also genetically porous: introgression is ongoing between a number of taxa (Besansky et al. 2003). The less populous species are, in the main, not sympatric - aside from a small crossover in the ranges of *quadriannulatus* and *merus* in South Africa. However it is entirely possible that they hybridise occasionally with *An. gambiae* or *An. arabiensis* where they are in sympatry. More significantly *Anopheles gambiae* and *arabiensis* are known to have considerable historic and at least some ongoing genetic introgression. F1 crosses of *arabiensis* and *gambiae* produce sterile males but viable females, therefore although hybrids are comparatively rare in nature (found at frequencies of around 0.75% in the wild) they are frequent enough to indicate ongoing gene transfer (Besansky et al. 2003). Hybridisation experiments and genomic evidence support this as being a viable and historically important mechanism of gene flow. In laboratory conditions hybrid *agambiae* / *arabiensis* populations were found to generate stable hybrid autosomes, with only the x chromosomes showing severe underdominance (Della Torre et al. 1997) (in this case the agambiae x chromosome was eliminated by the F3 generation). The genome also shows signs of introgression across almost its entire autosomal length. Comparative genomic studies between *gambiae, arabiensis*, *merus* and *melas* indicate that, outside of the sex chromosomes, introgression has introduced significant anomalies to the phylogenetics of the genome; some regions on autosomes indicating *gambiae* and *arabiensis* to be sister species, and others (principally x-chromosomal) suggesting *gambiae* and *merus* are most closely related (Wang-Sattler et al. 2007; Besansky et al. 2003). Heterogeneities in the genome, in particular the large structural variants, make it difficult to state definitively which of these is the correct hypothesis.

3.3: *An. gambiae* population structure

What is more interesting about *gambiae* from an evolutionary standpoint, is that it also exhibits complex population structure at finer grained scales - leading to the eventual definition of two separate species in late 2013 (Coetzee et al. 2013). The frequent changes in nomenclature have led to a number of inconsistencies within the literature as the *Anopheles gambiae* taxon itself has been iteratively refined. In attempting to describe the population structure within *An.gambiae* '*sensu media*' (i.e. the species pair that will eventually be called *An.gambiae* / *An. coluzzii*) it should be borne in mind that these were defined as, and indeed were, a single species until very recently. When dealing with population structure within this

section the definition prior to 2013 will be used, that is *Anopheles coluzzii* = 'M-form' and *Anopheles gambiae* = 'S-form'.
The genetic investigation of population structure in *An. gambiae* has been extremely challenging. The mosaic genome structure described above, in which most of the genome will generate misleading phylogenies, makes it difficult to choose reliable markers for population studies. However studying the distributions of isozymes, chromosomal inversions, microsatellite and lately SNP genotypes has elucidated much about population structure within the taxa.

Geographical variation

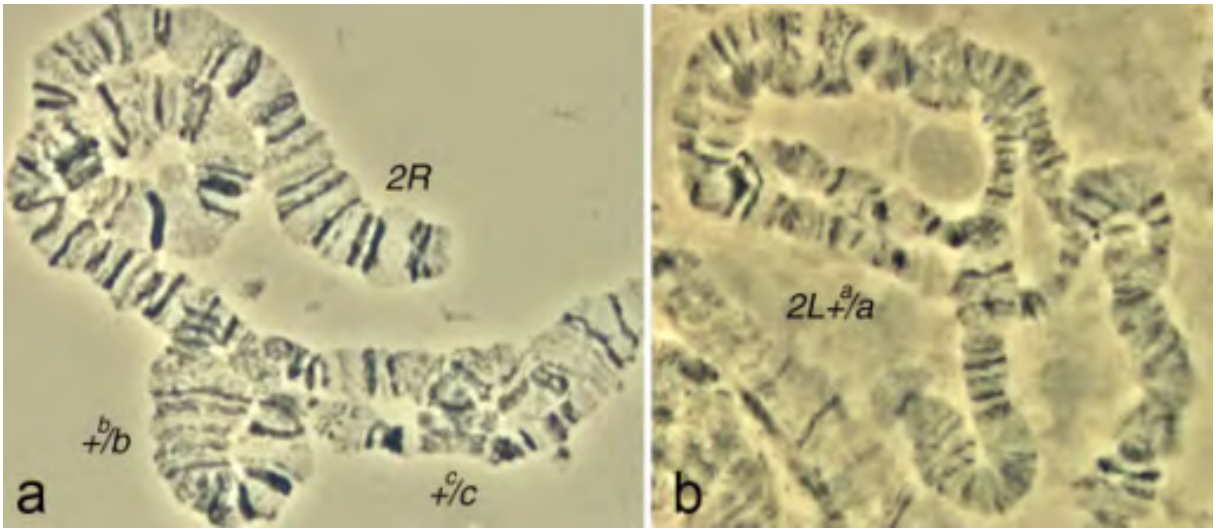
A panmictic model on a continent-wide level would predict a steadily increasing degree of divergence as distance between samples increased. That is, due to the limited range of *Anopheles gambiae*, the homogenizing effect of breeding would break down over all but the shortest distances. Initial studies involving large structural variants suggested this was, indeed the case (Taylor et al. 2001), however subsequent genetic studies failed to confirm this, with isozyme and microsatellite results instead suggesting that populations as far away as Kenya and Senegal (> 6000km apart) were highly similar, showing an F_{ST} between 0.01 to 0.1 (Besansky et al. 1997; Lehmann et al. 1996). Yet the low dispersal rates for An. gambiae make this extremely difficult to believe and explanations offered at the time – of long-range hybridisations related to human migration – are perhaps unlikely.
Subsequent studies using a higher density of markers and broader geographic sampling indicate that these results may have suffered from errors due to low sample size: Lehmann *et al,* using a sample set from 10 countries and 11 microsatellite loci, uncovered a strong relationship between distance and diversity and strongly indicated that the earlier studies had derived their misleading results from a restrictive sample size (Lehmann 2003).
In addition to confirming the relationship between distance and divergence, Lehmann *et al..* (ibid) also noted a distinct geographical separation when comparing samples continent-wide. Countries split into two pan-national categories, with a northwest group consisting of Senegal, Ghana, Nigeria, Cameroon, Gabon, DRC and Western Kenya, and a south-eastern group consisting of Tanzania, Malawi, Zambia and Eastern Kenya. F_{ST} between these groups was greater than 0.1 – far higher than could be seen within those groups.
The split between Eastern and Western Kenya is instructive; Kenya is bisected by the Great Rift Valley (GRV) and all of the southeastern group with the partial exception of Zambia lie to the east of this feature. The GRV, a 6000km long trench running from Eritrea to Mozambique (and extending beneath the Red Sea as far as Syria), presents a major barrier to gene flow between mosquito populations. Studies within Kenya using RFLPs showed that populations

within 10km could be highly differentiated, and that distances of of a few hundred kilometres showed complete genetic isolation (McLain et al. 1989) (far from the homogeneity suggested by Besansky *et al.* (Besansky et al. 1997) across thousands of kilometres).
The inconsistencies between results highlights one of the major issues with variation in this species in general; that, due to the relatively recent dispersal of this species (thought to be linked to a human expansion between 2-5000 years ago (Coluzzi et al. 1985)), the majority of variants are widely shared and can be uninformative for population structure studies. Poor choice of markers, particularly where ascertainment biases may favour one result over another, can render a study ineffective. Moreover the unequal levels of introgression seen in the presence of chromosomal inversions and other structural features, can mean that even whole-genome marker sets can produce apparently erroneous results (Besansky et al. 2003).

3.4: Chromosomal forms

Chromosomal forms are defined on the basis of their representation of a small number of frequent chromosomal inversions (see section 3.7 for a more complete description of the genomic effects of inversions). Whilst they have often shown inconsistencies and non-linear phylogenetic relationships, they are nevertheless informative of demography; for instance, assuming results showing *merus* / *gambiae* as sister taxa (Besansky et al. 2003) are confirmed by subsequent continent wide sampling, it will underline the fact that the original analysis of diversity based on chromosomal inversions was, in fact, the correct one.
There are ten inversions that are fixed between species in the An gambiae complex (e.g. Xag: this inversion is not polymorphic within any species; it is inverted and fixed in *Anopheles gambiae* and *merus*, and fixed in its standard orientation in the other species). In addition within *An gambiae* there are seven frequent inversions, all of which are on chromosome 2, and most of those on the 'right' chromosome arm: (2La, 2Rj, 2Rb, 2Rc, 2Rd, 2Ru, and the rarely seen 2Rk) (Coluzzi et al. 2002). *Anopheles*, like other dipterans, demonstrates endoreplication – the replication of individual chromatids without mitosis; the result of which is giant polytene chromosomes. When identified in certain tissues (in mosquitoes this is particularly clear in larval salivary glands and adult female nurse cells, other tissues having a lower degree of endoreplication) these can be viewed under a light microscope and enable the identification of banding patterns within the chromosome. Chromosomal inversions are therefore identifiable as reversals of the expected banding pattern, and heterokaryotypic individuals (containing one standard and one inverted chromosome) by an obvious and characteristic loop (della Torre 1996).

Figure 3.2: heterokaryote loop of inversions 2La, 2Rb, 2Rc



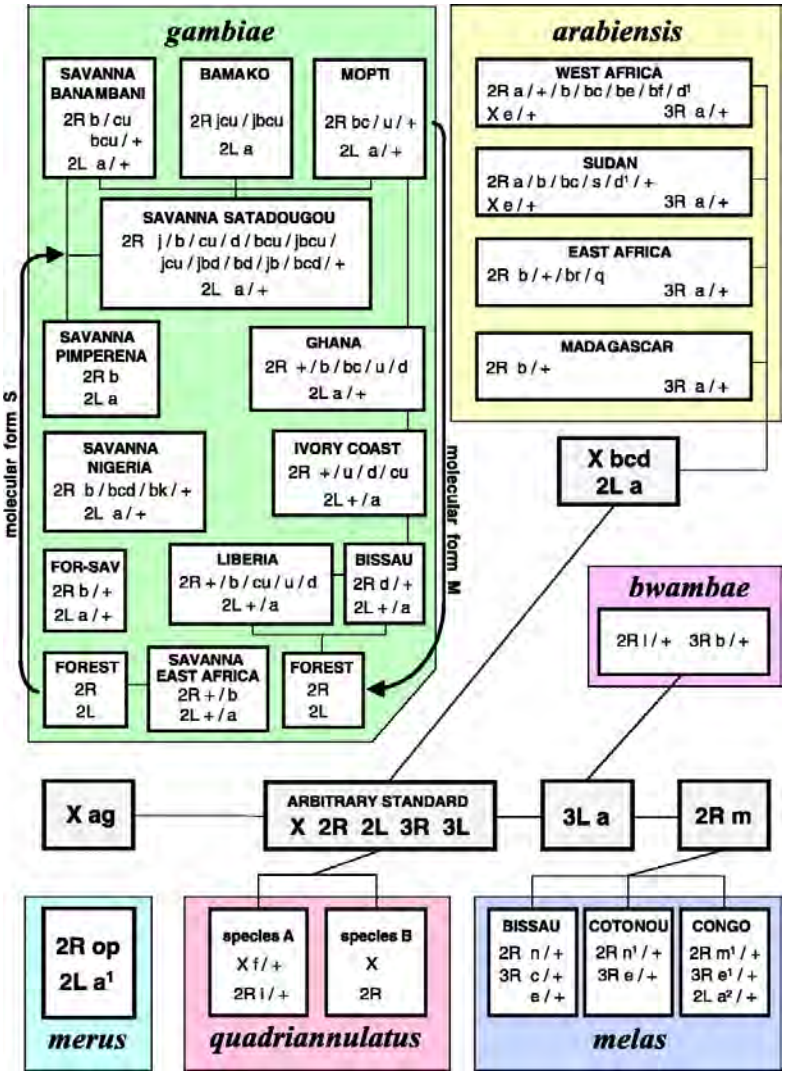
Polytene chromosomes of An. gambiae, demonstrating the characteristic heterokaryote loop for a) inversions 2Rb and 2Rc and b) inversion 2La. Lack of adjacency can be seen at each of the breakpoints providing some illustration of the physical barriers to recombination that underlie many of the actions of inversions.
Source: Arm-specific dynamics of chromosome evolution in malaria mosquitoes. (Sharakhova et al. 2011)

The inversion of chromosomal segments is assumed to be rare, such that all 2La chromosomes are believed to derive from a single inversion event. Due to their relative ease of identification, these features have been studied within dipterans for decades, in order to identify populations in which the distribution and relative proportions of these features indicates that there may be underlying population structure (Noor et al. 2001). In particular departures from Hardy-Weinberg equilibrium (HWE) that demonstrate an underrepresentation of predicted heterokaryotypes can indicate a non-panmictic population. Examinations of inversion frequencies within the *Anopheles gambiae* have frequently identified populations that are not in HWE, and yielded strong indications that *An. gambiae* was not panmictic; the contrasting frequencies of inversions in different environments led Coluzzi *et al.* to define five separate populations: Forest, Bissau, Mopti, Bamako and Savanna, each of which was associated with a different environment or locale and each of which had a different typical karyotype (Coluzzi et al. 1985): Forest had standard uninverted chromosomes; Bissau had high frequencies of 2Rd; Mopti was fixed for 2La and had high frequencies of 2Rbc and 2Ru; Bamako is fixed for the 2Rjc and u inversions and polymorphic for 2Rb; and Savanna is highly polymorphic, segregating c,u,j,d,k inversions and typically showing very high proportions of 2La and 2Rb.

Figure 3.3 : Chromosomal inversions of *Anopheles gambiae sensu lato*

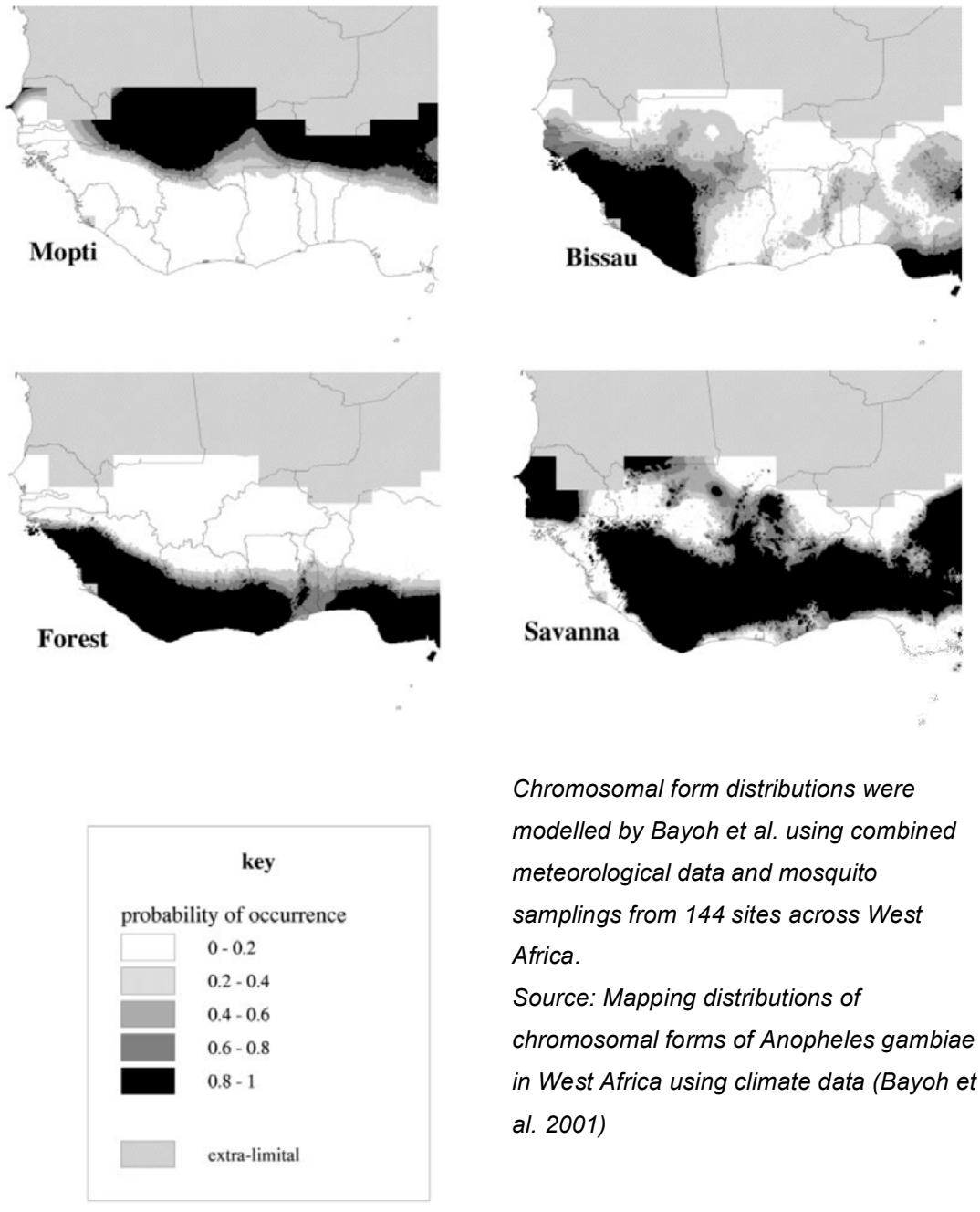
Chromosomal inversions are largely non deterministic for individual chromosomal forms, but their distribution shows recognisable differences between both sympatric and allopatric populations of An. gambiae, strongly supporting a model of low genetic homogenisation across large distances.

Source : A polytene chromosome analysis of the Anopheles gambiae species complex (Coluzzi et al. 2002)



These ‘chromosomal forms’ have very different patterns of distribution. At one extreme the Bamako form is restricted to southern Mali and Guinea Conakry, along the banks of the niger river. The other four – at least in west Africa where inversion polymorphisms are most prevalent and the most detailed studies have been performed – show varying degrees of dispersal. In a meta-analysis of samplings from Senegal in the west to Benin in the east Bayoh *et al.* (Bayoh et al. 2001) showed Bissau to be restricted to western wet coastal forests (largely within Guinea-Bissau, with some representation in the south of Nigeria), Forest occupied low lying and coastal forests, Mopti had occupied the arid sahel inland and Savanna the large savanna belts inbetween. However overlap between these zones, at the macrogeographic level at least, was extensive. All forms aside from Bamako were found to overlap somewhere within the humid tropical zone (Mopti in the north, Savanna across the zone, and Bissau and forest in the coastal parts), suggesting that the lack of panmixia was not related to geographic isolation.

Figure 3.4: Chromosomal forms distributions of *An. gambiae sensu stricto* in West Africa.



Chromosomal form distributions were modelled by Bayoh et al. using combined meteorological data and mosquito samplings from 144 sites across West Africa.
Source: Mapping distributions of chromosomal forms of Anopheles gambiae in West Africa using climate data (Bayoh et al. 2001)

The chromosomal forms also show varying degrees of reproductive isolation. Studies in Mali, where the Savanna, Mopti and Bamako forms exist in sympatry, found that hybrids between savannah and the two other forms were present, albeit at low frequencies, yet Mopti / Bamako hybrids were close to absent (one putative hybrid in over 17,000 samples) (Touré et al. 1998). It should be noted, however, that by karyotype alone and without any other genetic

markers outside of the inversions, it is difficult to distinguish low-frequency inversion forms within a population from true hybrids.

3.5: Molecular forms:

Although the chromosomal forms were a vital step in first dissecting the population structure of *Anopheles gambiae*, subsequent work has cast doubt on their reliability as markers of reproductively isolated populations.

In an attempt to devise molecular typing methods for the chromosomal forms, sequences of rapidly evolving intergenic spacer within ribosomal DNA were compared between Mopti, Bamako and Savanna individuals from Mali and 10 SNPs found that segregated the chromosomal forms. A PCR diagnostic was devised from these SNPs (Favia et al. 2001). However use of this test, in particular outside of Mali, soon demonstrated unexpected results. Although the geographically restricted Bamako form was predictable, Mopti and Savanna forms were not. Apparent hybrid chromosomal forms were found that had non-hybrid molecular test results, or mopti-like molecular results within savanna-like chromosomal forms – indicating that karyotypes were indicative but not definitive for the ribosomal DNA results. However, although they did not unambiguously type Mopti and Savanna chromosomal forms, the distribution of M/S markers strongly supported a hypothesis of incipient speciation in west Africa (della Torre et al. 2001).

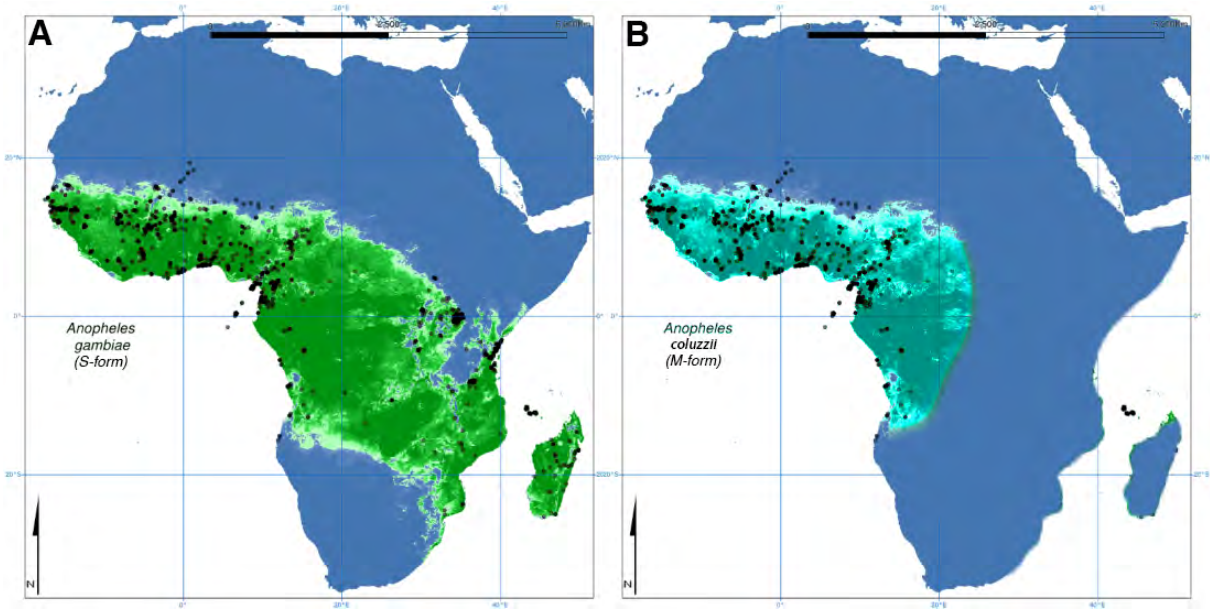
Some chromosomal forms sit well within this paradigm; Bissau is entirely comprised of M-form gambiae, and Bamako is entirely S-form; Mopti is predominantly, though not entirely M-form; however the chromosomal forms of Forest and Savanna contain both M and S form individuals. In the region studied by Della Torre *et al*, population segregation appears to be maintained between M and S. This is demonstrated by the lack of gene transfer that we might expect within unstructured populations: in Cote d'Ivoire the insecticide resistance allele *Kdr* is at high frequency in S-form mosquitoes, and would be expected to be strongly positively selected in a panmictic population, yet it is entirely absent in M-form individuals in this area.

Distribution

Since 2001 the M and S-form speciation event has been extensively studied. The availability of a molecular test has enabled the comparisons of the M/S-form ranges on a macrogeographic scale and investigations of the rates of hybridisation extending to thousands of individuals.

The range of the M-form centres roughly on the area in which it was first detected, being widely represented from the west coast of Africa as far east as Angola and the DRC (della Torre et al. 2005). For most of this range it is broadly sympatric with the S-form, though it extends both further north and south, being the only form found in the drier sahel of Senegal (the transition zone between southern savanna and the northern sahara desert), and in the borderline desert regions of southern Angola. Although isolated M-form individuals have been found in savanna regions of Zimbabwe (Masendu et al. 2004), the M-form is thought to be largely absent from East Africa (Lehmann & Diabate 2008).

Figure 3.5: Coarse-scale continent-wide distribution of molecular forms



Coarse-scale distributions of the M and S molecular forms. The two forms exist in sympatry in West and Central Africa, whilst only the S-form has regularly been found in East Africa.
Source: adapted from Anopheles mosquitoes - New insights into malaria vectors - chapter: Speciation in Anopheles gambiae (Lee & Lanzaro 2013), adapted after: On The Distribution and Genetic Differentiation of Anopheles gambiae s.s. molecular forms.(della Torre et al. 2005)

The S-form, in contrast, extends the width of Africa, being the only form found east of the great rift valley. Its latitudinal range is more restricted, being absent from the transitional desert regions, and from the drier habitats of Ethiopia, Somalia and Sudan (where Arabiensis predominates) (della Torre et al. 2005; Lee, Marsden, et al. 2013). Within the regions of sympatry, however, a different picture emerges; with one form or the other predominating in individual locations within the same range of distribution (i.e. < 50km) (Costantini et al. 2009). The question of whether the M-form and S-form are ‘good species’ has occupied vector biologists in recent years with a number of studies attempting to assess their relative levels of divergence and reproductive isolation. Much of this is intertwined with the distributions of

chromosomal inversions (see section 3.7), since where they do segregate they frequently do so along chromosomal lines, however the relationships are not deterministic.
The initial study implying incipient speciation in the M and S-forms highlighted a lack of hybrids as evidence of the reproductive isolation of the subspecies. Della torre *et al.* found that, despite the frequent coincidence of the M and S molecular forms with Mopti and Savanna chromosomal forms, individuals that were of 'hybrid' chromosomal form were not of hybrid M/S form, indicating that these were not hybrids, but a representation of low-frequency 'non-standard' chromosomes within the chromosomal forms. In fact, more than 1000 individual mosquitoes were tested, including in regions of M/S sympatry, yet this yielded only 3 hybrid genotypes. This is a level of hybridisation comparable to that between *An. gambiae* and *An. arabiensis* and strongly supports complete reproductive isolation.

Phenotypic differences

The M and S form mosquitoes are not merely differentiated by their ribosomal DNA. Both forms occupy an array of different niches and those niches differ from country to country, making it difficult to resolve form-specific traits; nevertheless some clear and robust phenotypic differences can be discerned.
The Mopti and Savanna chromosomal forms were originally identified due to their preferences for distinct habitats, in particular the greater tolerance of aridity of the Mopti form (Touré et al. 1994). Whilst the links between chromosomal and molecular form patterns are not perfectly maintained (although the M-form does broadly maintain the higher tolerance to aridity that was seen in the Mopti chromosomal form), it is perhaps not surprising that the M/S molecular forms display some distinct habitat preferences.
The phenotypic segregation that can be seen between the forms is not within adult habitats but larval. S-form mosquitoes are found in what might be termed the classical mosquito breeding sites – temporary rainpools and puddles. The M-form, in contrast are typically found in permanent bodies of water, often those that have a human origin (Lehmann & Diabate 2008). In exploiting these anthropogenic habitats, such as rice paddies and water storage containers, the M-form is able to extend its reproductive activity throughout the year. This is also likely to impact upon its ability to exploit more arid regions that are not accessible to the S-form. As a result of the strongly anthropogenic habitat preferences of the M-form, it is assumed that this divergence has arisen since the emergence of permanent human habitats in Africa (della Torre et al. 2002).
Other, potentially linked phenotypes are also apparent (summarised in lehmann *et al.* (Lehmann & Diabate 2008)). The development times of the M-form are significantly slower than the S-form, as can be afforded in the permanent larval sites that are at little risk of

drying up. Causality would be impossible to determine, but slower development is associated with larger body and wing sizes, and perhaps as a consequence wing beat speeds are slower, beating at 492hz to the S-form's 493hz (ibid).
Anecdotal evidence has for many years suggested the M-form mosquito could be found in 'dirtier' habitats, something that would tally well with its presence in rice fields, and this form does appear to have a higher tolerance to larval habitat pollutants such as ammonia (Tene Fossog et al. 2013). What is certain is that the S-form has a far weaker capacity for avoiding predation – something that is likely to be less crucial in a temporary pool that has had little chance to develop predatory fauna.
Indeed, transplantation experiments have been performed in which the offspring of wild females were placed in M or S-like habitats; unexpectedly the S-form outcompeted the M-form in all habitats, however when a 50:50 mixture of M and S forms were seeded in the same habitat the emergence success of M-forms increased, while emergence of S-forms decreased, indicating competition between the forms (Diabate et al. 2005), however despite being hampered by competition, the overall emergence of the S-form remained higher. However, studies of mixed forms in larval habitats demonstrated that when predators (predominantly *Notonectidae* "backswimmers") were present the M-form individuals consistently outperformed the S-form in both types of habitat (Diabate et al. 2008). M-form mosquitoes have additionally been shown to have more pronounced avoidance responses to *Notonectidae* resulting in a lower overall level of predation (Gimonneau et al. 2012).

It should be remembered, of course, that the larval habitat is not selected by the larva, but by the oviposition preferences of the female mosquito, therefore any discontinuity between the parental preferences for oviposition site and the larval capabilities within that habitat should be selected against.
This should create a strong evolutionary imperative for mosquitoes - females in particular - to mate assortatively, and there have been a number of attempts to ascertain the method of pre-zygotic isolation. Whilst the theory of selection on wingbeat frequency might be somewhat fanciful (Tripet et al. 2004), evidence of form-specific selection has been seen from mixed swarms (Dabire et al. 2013). However tethered females released into swarms of either the same or different forms showed no difference in their rates of insemination – suggesting, at least, that assortative mating in mixed swarms is not the major reason for reproductive isolation.
Differences in swarm formation are a more probable method of distinguishing between M and S-form males. S-form mosquitoes form swarms over bare ground, whereas M-swarms form over 'high-contrast' areas, such as might be formed at the border of a grassy area and a footpath, or over a well (Manoukis et al. 2009).

Although no notable differences in infectivity to *Plasmodium* have been shown between the M and S forms, M-form mosquitoes have also been shown to have greater longevity. This may or may not be related to the larger body sizes, but more significantly it could have a drastic effect on the vector capacities of this molecular form.

3.6: Speciation and genomics

The reproductive isolation of these species has been further investigated with genomic techniques of increasing resolution.

First attempts were made using microsatellite maps which confirmed that a region of differentiation between the populations could be found in the centromeric region of the X chromosome – that is, surrounding the rDNA IGS region used for genotyping, and outside of the Xag inversion that segregates *gambiae* from *arabiensis* (Wang et al. 2001). This result was further corroborated using intronic sequence from genes located within or near to the pericentromeric region of the X chromosome, demonstrating again the lack of homogenisation in this region between M and S forms (extending the analysis further from the centromere showed no such differentiation). This effect contrasts with A.gambiae / arabiensis populations (that show a similar degree of hybridisation to M/S forms in this region) in which the X-linked divergence extends across the chromosome (Stump et al. 2005). The authors suggest that it is not merely mechanical effects that underlie this effect, but instead that natural selection maintains this divergence in the face of gene flow.

The existence of centromeric islands of divergence lends weight to the ‘speciation islands’ model. According to this model, reproductive isolation is incomplete, with some hybridisation between M/S form mosquitoes. Gene flow is therefore ongoing, and would be expected to homogenise much of the genome. Areas where homogenisation does not take place should therefore contain genes that underlie the reproductive isolation itself (whether these are genes related to niche separation or assortative mating) (Stevison et al. 2011). Assuming this is a polygenic trait this is most likely to occur where co-adapted alleles cannot be efficiently broken up by recombination: in the mosquito this is predominantly in the centromeric regions and within chromosomal inversions.

An alternative model is that the speciation event is already advanced, and that the divergent regions are the effect of this lack of panmixia. For instance, in two separate unmixed populations a selective pressure such as a novel insecticide could be applied to a heterogeneous gene important for resistance; the result might be two separate purifying selections that would appear to be a single region of divergent selection. In practical terms the difference would be twofold: firstly there would be less homogeneity in the position of diverged regions – assuming species occupy marginally different niches they would be

subjected to differing degrees of purifying selection, and different loci would be affected; secondly, there would be a far smaller likelihood that the genes underlying reproductive isolation would be present within these diverged regions (Lee, Marsden, et al. 2013).
The development of novel high throughput genotyping methods enabled a more fine grained analysis of divergence across the genome. The first attempt to assess which of these models fit best was performed using PCR-RFLPs for SNPs within the three proposed speciation islands (i.e. the three centromeres). White *et al*. (White et al. 2010) genotyped 517 individuals in Mali, Burkina Faso, Cameroon, and Kenya, the first three countries spanning the range of the M-form and Kenya providing a pure S-form sample.
They found a high degree of linkage between the centromeric SNPs in all cases, something that would not be expected if there was ongoing gene flow. Although some individuals did show patterns suggesting F1 hybrids, it appears that these individuals do not make a significant genetic contribution to the overall population.
The first genome-wide studies support this view. There have been two large-scale studies, both performed in the West African regions with the putative highest levels of reproductive isolation. Two colonies of mosquitoes from Mali, the M-form *Mali-NIH* colony and the S-form *Pimperena*, were sequenced using sanger sequencing to around 6-fold coverage (comparable to a median level of coverage using NGS sequence, but still enough to leave gaps in the de-novo assembly). The resultant comparison between the genomes showed divergence was not clustered within areas of reduced recombination. Although there were higher numbers of fixed-differences between colonies within the x-centromere, areas of high divergence were found genome-wide – which would not be consistent with significant ongoing gene-flow. In addition, the loci that were diverged were informative. The most notable peak of divergence was located at ~25mb on chromosome 2L, and centred on the *RDL* (‘resistance to dieldrin’) gene. As the name might suggest this gene has previously been associated with insecticide resistance and would have been placed under strong selective pressure within the past half century (Du et al. 2005). Importantly this gene has been shown to have separate mutations conferring resistance in the ‘good’ species *An. gambiae* and *An. arabiensis,* with an alanine to glyceine mutation in *gambiae* and alanine to serine in *arabiensis.* Similarly the S-form Pimperena sequence contains a glyceine mutation to the M-form’s Serine mutation, suggesting independent sweeps in each form (Lawniczak et al. 2010).
However the examination of colonies is limited; the Mali and Pimperena colonies were founded from 80 and 5 iso-female lines respectively, and inbreeding will reduce the level of heterogeneity in the colony even further. This is of particular concern for the S-form Pimperena colony that was founded from just ten parental chromosomes.

A companion paper was published by Neafsey *et al*. which took a similarly high-throughput genotyping approach over a larger number of samples from the same region (Neafsey et al. 2010). The authors used the raw M and S sequencing reads in order to call variants found both within and between the M and S colonies; from a subset of these SNPs a 400,000 SNP Affymetrix microarray was developed in order to assay both within and between form variation. The chip was used to sample 60 individual mosquitoes, twenty each of Mopti, Savanna, and Bamako chromosomal forms, from an area of sympatry (testing of the molecular form using the favia-fanello (Fanello et al. 2002) indicated these groups to be, as expected, M-form, S-form and S-form respectively).

The work of Neafsey *et al.* confirmed many of the findings from Lawniczak *et al*. that the divergence was concentrated in the centromeres, but that differentiated regions could be found across the genome (Lawniczak et al. 2010). Moreover, the ability to measure within-form diversity also enabled the authors to identify multiple regions of purifying selection that were unique to each form – again suggesting an advanced speciation process with little gene flow to disrupt these swept regions. Further comparisons of M-form Mali and M-form Cameroon samples showed that, despite significant differentiation between Malian and Cameroonian samples, the genome-wide patterns of divergence between M and S molecular forms were not confined to Mali.

Studies under controlled conditions further demonstrated that there were no intrinsic incompatibilities within the speciation islands. More than 2000 F2 individuals derived from an artificial M-form/S-form hybridisation were genotyped using markers that segregated each of the speciation islands; no evidence of biased co-transmission was found, suggesting either that the centromeric linkage is due to genetic drift under advanced speciation, or that the three alleles are under strong coordinated selective pressure (Hahn et al. 2012).

More recently, Reidenbach *et al.* (Reidenbach et al. 2012) used the same chip used to examine variation within and between M and S-form mosquitoes from populations across West and Central Africa. Sampling in Mali (from the same villages samples by Neafsey *et al.*) as well as locations in neighbouring Burkina Faso, and Cameroon, the authors again found heterogeneous islands of divergence across genome and not merely in regions of reduced recombination.

However when regions that were diverged or swept in these regions were discounted they discovered that the samples appeared to cluster based on geography, perhaps indicating that introgression on a local level is still a factor. This apparent contradiction may be a result of variation in the degree of postmating barriers. That is, if there is no intrinsic incompatibility maintaining the speciation islands (as found by Hahn *et al.*) then the level of reproductive isolation will depend upon *extrinsic* factors. If, as is widely believed, these extrinsic factors are due to niche preferences (see section 3.7), this would suggest that the strength of

reproductive isolation would vary depending upon the prevailing environmental conditions. Population sizes for molecular forms are known to vary drastically with seasonal changes (Caputo et al. 2011); it may well be the case that with the reduction in available intra-form mating partners, hybridisation levels increase.

Introgression

West Africa

Other areas of Africa are believed to have less stringent barriers to gene flow. The west coast of Africa between Guinea-Conakry and Senegal was previously thought to be a region where the level of reproductive isolation between Mopti and Savanna chromosomal forms was lower and is now viewed as a hotspot for M/S form introgression. Far higher frequencies of hybrids are found here than in Mali or Cameroon (where <1% is typical). Senegal – 3% (Ndiath et al. 2008), The Gambia - 7% (Caputo et al. 2008) and in particular Guinea-Bissau 24% (Oliveira et al. 2008) all appear to have fewer or less strict barriers to gene flow.
To discern whether this represented a breakdown of the speciation event, or merely the maintenance of ancestral hybrid genotypes, markers for each of the putative centromeric speciation islands were tested by Caputo *et al.* (Caputo et al. 2011) in an analogous manner to the earlier work by Hahn *et al.* (Hahn et al. 2012). Sampling at four sites, three along The Gambia and one coastal site in Guinea Bissau, found that there was persistent (though incomplete) linkage between the centromeric islands. This strongly suggests that the hybrids in these regions represent a recent collapse of a previously segregated population. However the breakdown of marker linkage within one of these loci does suggest that this introgression is advanced, and not heavily selected against in this region.
A further examination of this region was undertaken by Nwakanma *et al.* (Nwakanma et al. 2013). Sampling almost 3500 mosquitoes from twelve sites spanning Senegal, The Gambia, Guinea-Bissau and Guinea-Conakry the authors found varying degrees of M/S hybridisation. In almost all regions hybridisation was higher than in Mali, though most were still at levels indicating some reproductive isolation was present (5-34%). At one sampling site, Caio in Guinea-Bissau (a diverse coastal habitat including rice plantations and open woodland), the hybrids were in Hardy-Weinberg equilibrium indicating complete panmixia, although it should be noted this was at the lowest sampling density with only 12 mosquitoes sampled in the locality.
The genome-wide divergence of these samples was also assessed using the AgSNP01 chip developed by Neafsey *et al*.. DNA was extracted from 20 randomly selected mosquitoes of

pure and hybrid forms from three sites, one each from The Gambia, Senegal and Guinea-Bissau; relative allele frequencies were then deduced from these pools.
Signal strength differed between the three locales, but all showed a distinct increase in divergence around the X centromere when comparing heterozygote pools to the pure M-form pools. A similar rise in divergence was not seen, however, when comparing the pools of hybrids to S-form samples. The asymmetry of this hybridisation is intriguing, and has been suggested from other studies also (Lee, Marsden, et al. 2013). The authors propose that this relates to the speciation of the M-form itself; that is, that the pre-mating barriers that have enabled the reproductive isolation of the M-form in one region may not be present continent wide. In this case, upon expansion of the M-form into novel niches (in which ecological segregation is not enforced), the local S-form mosquitoes that they encounter may not have developed the same pre-mating mechanisms; F1 individuals would therefore more successfully backcross into the ancestral S-form than the M.
Care should be taken, however, in interpreting many of these results. Pooled sequence has significant weaknesses in its ability to accurately assay allele frequencies and validation of at least a subset of the genotypes by a secondary method is preferable. In addition the AgSNP01 chip has inherent ascertainment biases (being designed from M/S colonies in Mali) might not accurately assay the fine-grained divergences in other regions; this could in particular lead to over-estimations of contrast values in the X-centromeric regions that were both probe-poor and designed from colonies with a known segregating locus in this region. The use of relative probe intensities as a measurement of divergence can also cause erroneous results. In an analysis of five systems involving closely related species with varying degrees of reproductive isolation, Cruickshank and Hahn (Cruickshank & Hahn 2014) have shown that relative measures of diversity can be misleading. In particular that they are liable to over-estimate divergence in regions of reduced diversity. This is particularly likely to be the case in regions where environmental pressures have caused a reduction in diversity in one or both of a species pair. It is clear, therefore, that care should be taken when interpreting diversity against a background of heterogeneous selection, as will be found in the mosquito.

Adaptive introgression:

Further work will be required to elucidate the precise nature of the breakdown in speciation in West Africa, and indeed the degree of hybridisation afforded by prevailing environmental conditions in other parts of Africa. Yet even in regions in which reproductive (or indeed geographical) isolation is pronounced; where hybridisation is reduced to levels seen between *gambiae* and *arabiensis* and hybrids are expected to have reduced fitness, loci that are

sufficiently advantageous can be exchanged between forms and even driven to high frequencies within a matter of years (Slotman et al. 2005)..

The first indication that stable introgression might take place between these reproductively isolated populations emerged from further analyses of SNPs within the centromeric speciation islands. Samples from locations that displayed a range of degrees of hybridisation – from highly segregated populations with little or no current gene flow, to some that were believed to be in complete panmixia – were genotyped using fifteen SNPs located in each of the pericentromeric regions of the chromosomes (within 5mb of X centromere and within 1Mb of the autosomal) (Lee, Marsden, et al. 2013).

In contrast to the work of White *et al.* (White et al. 2010), the results indicated that introgression had occurred even in regions of supposed complete reproductive isolation, leading to a reduction in the linkage of the centromeric islands. Regions of higher introgression had more pronounced breakdown of the centromeric linkage, with the 2L centromere showing particularly high levels of hybrid and S-form alleles.

Within this context, however were indications of maintained reproductive isolation. The results supported the findings of nwakanma *et al.* (Nwakanma et al. 2013), that the introgression was asymmetric. In regions with the highest frequencies of hybrids, pure S-form genotypes were notably absent, whilst pure M-form were maintained. In addition, the breakdown of linkage between genotypes in median-hybrid loci suggested that a lack of fitness in further generations acted gradually to remove backcross individuals in most loci.

As well as performing individual sampling in villages, Lee *et al.* also undertook a longitudinal study in one site, Selinkenyi in Mali (previously assumed to be a region of strong reproductive isolation). Sampling in eight different years between 1991 and 2012, the authors found the expected high levels of segregation between forms in the first decade and a half of the study – only 24 hybrids being found out of 398 individuals sampled. However a striking change was seen in 2006, when close to ¼ of the samples were hybrids, many of them F1s. The six years following this event were marked by an almost total loss of F1s from the population, but a maintenance of hybrid or M-like alleles on the 2L loci, to the point where S-like genotypes in this locus were almost entirely eliminated (in 2012 only 2 individuals out of 83 were found with entirely S genotypes). This indicates not only a pronounced selection against hybrids in this region, but also strong positive selection on the 2L locus.

It is now almost certain that the positive selection on this locus relates to the presence of *Kdr* (knockdown resistance) alleles. Mutations in the ‘voltage gated sodium channel’ (*VGSC*) gene provide resistance to both pyrethroid and DDT insecticides, and are expected to see extremely strong positive selection in regions where control measures are in place. Although it was previously highlighted as evidence for the reproductive isolation of the molecular forms, since Kdr alleles had been confined to S-form mosquitoes, in recent years it

has been well documented that the Kdr allele has, in fact introgressed into M-form populations (Weetman et al. 2012; Diabaté et al. 2002).
The interpretation that this represents extensive gene flow, however, is perhaps false. Clarkson *et al.* (Clarkson et al. 2014) have studied this phenomenon using genome-wide genotyping by illumina sequencing. Sampling across a ten-year timeline (from 2003 to 2013) at six sites in a region of sympatry in southern Ghana, the authors demonstrated a marked increase in *Kdr* alleles after introgression had occurred. However the additional genomic resolution afforded by whole-genome sequencing enabled the degree of introgression to be quantified. The hybridisation event that caused transfer of the Kdr allele did not, in fact, lead to extensive gene transfer; despite the presence of Kdr at near-fixation in M-form samples in 2013, the introgressed locus amounted to only ~1.5% of the genome. The prior M-like alleles were maintained across the majority of the genome and reproductive isolation between the M and S-form mosquitoes appears to be extremely robust.
It is notable that both Clarkson *et al.* and Lee *et al.* recorded a sudden increase in the frequency of hybrids in 2006, and this marked the clear beginning of the 2L locus introgression. The cause of this increase in hybridisation remains an open question, although environmental and climatic factors could have an effect on the ability of (particularly S-form) mosquitoes to find within-form mating opportunities. Similarly, in both cases the sampling time that immediately followed 2006 showed a marked drop in total vector numbers. It is tempting to infer that the reduction in vector numbers following each hybridisation event was caused by the increase in hybrid mosquitoes; in regions of distinct ecotypic separation hybrids would suffer reduced fitness compared to the pure molecular forms. Further longitudinal studies would be necessary in order to establish if this was the case.

Anopheles gambiae / coluzzii

Despite the apparent breakdown of speciation in ‘far-west’ Africa, the advanced state of reproductive isolation indicated from the studies in Central and West Africa, and the robustness of the assortative mating to isolated introgression events, was sufficient to lead to their definition as separate species. In 2013 the term *Anopheles gambiae s.s.* was restricted to only the S-form individuals, whilst the M-form was named *Anopheles Coluzzii* (Coetzee et al. 2013).

3.7: Ecotypification

Chromosomal inversions

There is no single inversion, nor group of inversions, that is definitive for the *gambiae* / *coluzzii* speciation event. However the contrasting distributions of these features between the two species does suggest that they play a role in the isolation of the populations.
The distribution of chromosomal forms is heterogeneous, with contrasting inversion frequencies found between habitat types. Indeed it has already been shown that the distribution of chromosomal inversions can be predicted by the prevailing climatic and meteorological conditions. Using previously published studies from more than 171 sites across western and central Africa, Bayoh *et al.* were able to predict the dominant chromosomal forms at each site using only coarse climatic variables: mean monthly precipitation, potential evapotranspiration, and mean minimum and maximum temperatures (Bayoh et al. 2001). Climate suitability zones were predicted for the Mopti, Bissau, Forest and Savanna chromosomal forms, yet importantly, these zones were not entirely distinct, with significant overlap over much of their ranges in which some degree of hybridisation is certain.
If drift is therefore negated as a possible cause of these allele frequency contrasts, selection should instead underlie their distribution. Moreover, predictable chromosomal forms are found in their 'correct' environments in both *gambiae* and *coluzzii* species (such as the homokaryotypic standard 'forest' individuals, with both 'M' and 'S' molecular forms in forested regions of high precipitation) (Lee et al. 2009). This further reinforces their association with ecological or climatic conditions.
The concept of inversions as single 'superalleles' providing adapted complexes to a particular environment is now widely accepted (Lee, Collier, et al. 2013), even if precise associations for most of the inversions have not been established. Despite the complex nature of these interactions, a broad pattern emerges in both the *Anopheles gambiae / coluzzi* species pair, and in the broader species complex: those groups that have the greatest compliment of inversion polymorphisms will be those that are able to exploit the greatest range of habitats (Coluzzi et al. 2002). For instance the Bissau form, segregating only the 2Rd and 2La inversions, has a limited range centred around Guinea in West Africa (43 out of 144 sites tested by Bayoh *et al.*) whereas the savanna chromosomal form, segregating to some degree all chromosomal inversions, has a range that stretches continent-wide (90 / 144 of the bayoh *et al.* sites).
Clearly, therefore, the inversions that are segregated between two parapatric populations can have a major effect on the potential for reproductive isolation: with strong inversion

differences reinforcing any other mechanisms of reproductive isolation, and shared inversions predisposing the populations to eventual collapse.

A good example of this can be found in the Bamako chromosomal form. This form is characterised by inversions 2Rjc and u, and polymorphic for only one inversion, 2Rb. It is solely found in a geographically limited larval distribution in laterite rock pools in the banks of the river Niger (Coulibaly et al. 2007). These pools are formed by erosion of the mineral rich rock, forming deep, and often interconnected, rockpools that provide a protective environment for the developing mosquito larvae. However their high mineral content presents the larvae with a challenging chemical environment – laterite rocks themselves being rich in iron and aluminium – and in common with permanent anthropogenic habitats, they have a higher complement of potentially predatory biota (Manoukis et al. 2008).

Despite a moderate level of hybridisation between the Bamako and the sympatric Savanna form (0-6.25% hybrids (Taylor et al. 2001)) comparisons between the two forms demonstrate significant reproductive isolation. Using the AgSNP01 400k chip, Neafsey *et al.* found regions of high levels of differentiation between S-savanna and S-Bamako individuals, in particular novel selective sweeps within one form that were not represented in the other. These regions were particularly found within the inversions themselves and clustered around the inversion breakpoints. However, importantly, they were also found in other autosomal regions suggesting that the divergence was not merely a result of suppression of recombination in these loci, but that reproductive isolation was robust. Given the moderate levels of hybrids seen in these loci, gene flow is certain to be ongoing, but it could be presumed that F1 Bamako / Savanna crosses displayed significantly reduced fitness in this environment.

Further supporting the extensive reproductive isolation of these chromosomal forms, Lee *et al.* used a tiling microarray to look at levels of differentiation between samples of Savanna and Bamako individuals from the same region, uncovering further significantly differentiated regions on the X-chromosome (Lee, Collier, et al. 2013).

The presence of differentiated regions throughout the genome raises the questions of whether the inversions, and the classical model of ecotypic speciation can really be found to underlie this differentiation – or if the inversions are incidental to an alternative X-linked mechanism of reproductive isolation.

Using statistical modelling of the frequencies of inversions and the number of adaptive loci captured, Manoukis *et al.* have demonstrated not only that the ecotypic theory of speciation is robust, but that the Bamako population conforms well to this model and is likely to display significant reproductive isolation (Manoukis et al. 2008). Despite the presence of Savanna and Bamako larvae in mixed larval habitats very few hybrids karyotypes are found (of 680 larvae sampled by Manoukis *et al.* only four were heterokaryotypic) suggesting that non-random mating is a factor in this ‘ecotypification’ event. Given the prior implication of X-linked

loci with assortative mating between *gambiae* / *coluzzii* groups it would be interesting to further investigate the X-chromosome loci identified by Lee *et al.*

Other chromosomal forms have also been suggested to be at the point of incipient speciation. Slotman *et al.* investigated potential subdivisions between the chromosomal forms in West Africa. Individuals were sampled from 18 sites in Mali (11 locations) and Cameroon (7 locations); microsatellite based comparisons between forest and savanna chromosomal forms in both *gambiae* and *coluzzii* demonstrated that the chromosomal forms within *coluzzii* had an elevated degree of differentiation (F_{ST} = 0.0406) when compared to *gambiae* (F_{ST} = 0.0053).

The same relationship was tested again in Cameroon, using microsatellites along with karyotype and M/S markers. Again the forest M form (that is, homokaryotypic standard) was found to be more highly differentiated than the others (Lee et al. 2009). This form was also significantly positively correlated with precipitation.

Physical / genomic effects of inversions

Studies of inversions in *Drosophila* have shown them to have a repressive effect on recombination. That is, in a heterokaryote, with one forward and one reversed section of the chromosome, a single recombinant will result in an extended dual-centromere and a shortened no-centromere chromosome – neither of which are viable (Stevison et al. 2011). As a result genetic exchange between the two inversion forms is suppressed, with gene flow occurring only via double-recombination events and gene conversion. This effect is most pronounced in the breakpoint regions, with suppression being almost total within 1.5Mb of the inversion breakpoint. As a result of this lack of recombination, inversions are inherited largely intact, and so segregate as single ‘superalleles’ within a population (Stevison et al. 2011).

It is assumed that inversion events are rare, so that the presence of inversions can be used to devise phylogenetic relationships via parsimony. However this is complicated by their transference as super alleles, since inversions are a source of large-scale introgression from sister or occasionally hybridising species. Indeed 2Rb and 2La are both thought to have introgressed to *An. arabiensis* after the *gambiae* speciation event (N Besansky personal communication). The frequent introgression of inversions can confuse attempts to deduce phylogenetics via parsimony of inversion forms, and even via genetic methods.

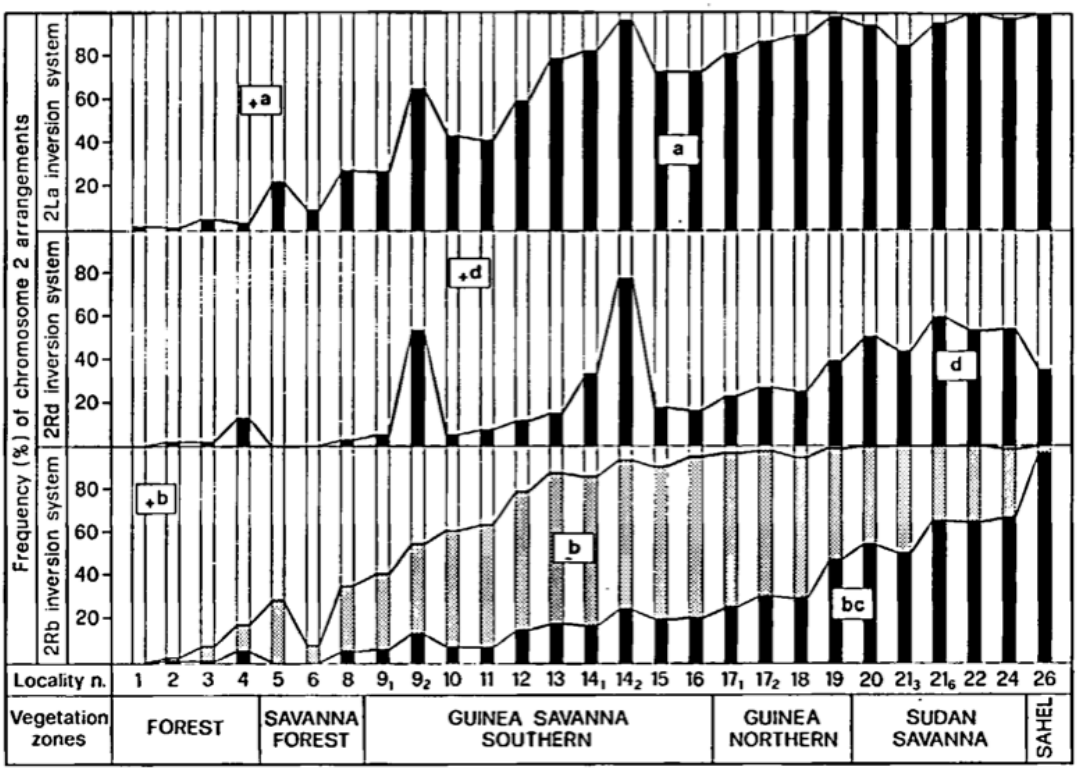
The reduction of recombination found in inversion breakpoints would, in the absence of any selection, lead to gradual divergence under drift only. However the reduction in recombination has also been hypothesised to facilitate the adaptation of linkage groups; that is, by preventing the disruption of groups of co-adapted alleles, chromosomal inversions can

facilitate adaptation to novel environments. In a similar manner, they can also lead to the maintenance of hybrid sterility between highly diverged populations (Noor et al. 2001).

Phenotypes and ecophenotypes

It is clear from the distributions of chromosomal forms that the presence of certain inversions carry with them advantageous factors for the exploitation of new habitats, however isolating which aspects the inversions are adapted to is problematic. Whilst broad geographical shifts are easy to discern, most of the ecotypes are not as specific as those seen for the Bamako form. Moreover, even within the Bamako niche, whilst it is clear that the 2Rj inversion provides adaptive advantages, the presence of $2R+^{j}$ individuals in the same larval habitats might indicate that the J-conferred advantages are either not drastic, or that it is not solely the larval habitat to which they are adapted.

Figure 3.6: Latitudinal clines of chromosomal inversion frequencies



Latitudinal clines shown for the three most frequent inversion systems, indicating a strong association for 2La with the drier regions of the sudan savanna and sahel.

Source: Coluzzi et al. - Chromosomal inversion intergradation and incipient speciation in Anopheles gambiae (Coluzzi et al. 1985)

The clearest example of association with such an 'ecophenotype' has been shown for the 2La inversion, which shows a clear gradient of distribution along latitudinal clines of aridity. The association with aridity was first identified by Coluzzi *et al.* (Coluzzi et al. 1979) when it was noted that the frequency of the 2La inversion differed greatly between ecological zones. In an exhaustive analysis of karyotypes in Nigeria, taking in twenty six sampling sites across forest / savannah and sahelian habitats, Coluzzi *et al.* demonstrated fixation or near-fixation of the standard and inverted forms in the forest and sahel respectively. The savanna sites In between these extremes demonstrated a gradation in frequencies, with the hotter northern zones showing high sahel-like 2La frequencies, and southern forest/savanna transition zones having low forest-like levels of 2La (see Figure 3.2).

More recent genetic dissections of this inversion, greatly assisted by the availability of a molecular test (White et al. 2007) have confirmed this association. Comparisons of individual mosquitoes from homokaryotypic $2L+^{a}$ / 2La colonies (each selected from a polymorphic 2L+/a colony) demonstrated a superior resistance to dessication in the 2La forms in both larval and adult life stages. Larvae were initially of similar robustness when exposed to heat, however after repeated exposure to heat stress, 2La-carrying individuals were found to be significantly more tolerant, perhaps suggesting a stronger heat-shock response (Rocca et al. 2009).

Adults were similarly tested and showed differences in aridity tolerance at eclosion and four days afterwards (Gray et al. 2009). Interestingly, the mechanisms by which the effects of dessication were negated appeared to be different at each adult timepoint: newly emerged 2La females showed lower rates of water loss than 2L+; whereas by four days rates of water loss were comparable, yet the higher initial water content of the 2La individuals provided them with superior dessication resistance. Whilst linkage between these phenotypes is entirely possible, the presence of two co-adapted features within the same inversion would be in keeping with the prevailing theories of ecotypic speciation.

Transcriptional responses were also evident between these inverted and standard colonies. Upon exposure to heat stress, both forms demonstrated massive upregulation of heat-shock proteins. The response from both karyotypes featured many of the same genes (principally the core families of heat-shock genes *hsp20, hsp70* and *hsp90*), however in 2La, the upregulation was of far greater magnitude than the 2L+ response and lasted longer (Cassone et al. 2011).

Interestingly, expression differences were seen both within and outside of the inversion itself, and the inversion was not significantly enriched for heat-shock proteins, pointing to a role for trans-acting regulatory genes in the heat-resistance machinery. Though the selection of homokaryotypic lines from a single heterokaryotypic colony that had been intercrossed for

many generations should ensure that the genetic background was relatively homogeneous outside of the inversion.

Resequencing of homokaryotypic samples from the wild confirmed this genetic profile. Comparisons of ecotypically segregated individuals along a latitudinal cline in Cameroon showed that divergence was principally inversion-based, with significant F_{ST} differences present only in the 2La inversion, and to a lesser degree in the 2Rb locus (Cheng et al. 2012). Within the 2La inversion, high-F_{ST} loci were found to contain a number of diverged cuticular proteins, as well as serine-threonine kinases, Ion transport proteins and G-protein coupled receptors. In comparison, the 2Rb inversion – that also demonstrates a similar aridity cline to 2La – did not exhibit high-F_{ST} loci containing genes that could be related to aridity tolerance, but instead had only one locus with an immunoglobulin-like cluster. This is perhaps evidence of trans-acting dessication-resistance mechanisms on 2Rb, or more likely, evidence of a confounding phenotype that is mechanistically unrelated to dessication resistance.

Inversion-associated copy number variations may also provide a feasible mechanism for the differences in the intensity of the 2La transcriptional response that may not show up from F_{ST} –based selection scans. Scanning for CNVs within resequencing data, Cheng *et al.* identified a number of regions with copy number polymorphisms, few if any of which were also in regions of high F_{ST}. It should be noted, however, that a lack of control for sequence divergence (significant on the 2La inversion) can lead to erroneous results with only small sample sizes. A scan of CNV differences between large numbers of individuals is yet to be performed.

The apparent lack of aridity-associated genes on the 2Rb inversion highlights the difficulty in dissecting inversion-phenotype relationships. The 2La inversion is rare in that its geographically-deduced phenotype has been successfully tested in the lab. Even so, conditions of high aridity are likely to have a knock-on effect on a range of other phenotypes, some of which are causally linked and some coincidentally.

In addition, 2La is a very large inversion; at 21Mb it is close to 10% of the entire genome, and naturally it possesses features that will be important for numerous phenotypes. In addition to its role in dessication resistance, 2La has also been previously associated with anti-plasmodial immunity (largely due to the presence of the PRI locus within the inverted region (Riehle et al. 2006)) and insecticide resistance (dieldrin and fipronil resistance being correlated with 2La karyotype (Brooke et al. 2000)), either of which could provide a strong selective pressure for its maintenance in the wild.

Phenotype-phenotype associations can also be difficult to disentangle. In behavioural studies, comparing internal sampling, external, and window-mounted ‘exit traps’, 2La karyotypes have been more frequently found in internal environs than outside or in exit traps,

both of which could indicate that the 2La karyotype is associated with the indoor-resting behaviour type (Coluzzi 1992). This may be related to the aridity tolerance, since in most regions of sub-saharan Africa the internal evening temperature will be boosted by cooking fires. However the relationship between aridity tolerance and indoor resting may be too subtle to disentangle with current methods.

It is important to note that both of the 2La associated behavioural phenotypes are important for malaria transmission and control. Increased aridity tolerance enables the vector to transmit malaria in more northerly regions of Africa and can be particularly problematic if the mosquito is able to colonise a region in which the standing human population lacks widespread resistance to malaria. Meanwhile indoor resting, as we have already seen, can indicate high degrees of anthropophily, but also leaves the mosquito susceptible to 'classical' control methods.

In addition, since these behaviours are both associated with an inversion polymorphism that was 'captured' within *Anopheles arabiensis*, it highlights the potential importance of hybridisation in extending the ranges and the vector competence of this species; introgression to *arabiensis* having extended the range of *Anopheles* and potentially increased its degree of anthropophily at a fell swoop. It is a curiosity of this system, however, that whilst the 2La inversion can introgress freely in both directions, the $2L+^{a}$ form cannot be established in an *An. arabiensis* population. The cause of this genetic incompatibility is not known.

Once established within a population, these polymorphisms can also give rise to precisely the type of behavioural heterogeneity that can militate against the actions of insecticide control (see section 1.2).

3.8: Chromosomal inversions and speciation

Although the link between chromosomal forms and molecular forms has been broken in recent years, nevertheless there remains a question over the relationship between chromosomal inversions and speciation.

The degree of divergence between some chromosomal forms, as measured by F_{ST} (see above), is comparable to that between the molecular forms (White et al. 2009), and the linkage of chromosomal and molecular forms in some regions is unlikely to be coincidental. If drift on this scale is combined with effective reproductive isolation, it is highly likely to lead to incompatibilities in collapsing populations. Such incompatibilities do not necessarily need to be established within the chromosomal inversion itself, and it is not entirely surprising that the resequencing experiments performed by Cheng *et al.* (Cheng et al. 2012) uncovered

numerous regions of high F_{ST} on the X chromosome when comparing ecotypically diverged samples.

Therefore even if the inversions are not the definitive cause of speciation, the presence of inversions might encourage and underlie ecological isolation, enabling sympatric or parapatric speciation events to take place that were not possible in a karyotypically homogeneous population. That is, by enabling 'niche expansions' into uncolonised ecotypes the presence of inversions might raise the level of reproductive isolation sufficiently so that alternative speciation loci can take effect.

Even without the issue of speciation, the process of niche expansion is in itself important for malaria transmission. At any scale we chose, the populations with the greatest complement of inversions is also that with the largest habitat. *An. gambiae, An coluzzii* and *An. arabiensis* are the only truly pan-african species in the *gambiae* species complex, and each segregates seven different inversions at high frequencies, with further fixed inversions delineating the species boundaries between the species. In contrast, the species within the complex that have restricted ranges – in particular merus, quadriannulatus and amharicus – segregate only 1-2 inversions each (*melas* is something of an intermediate case - segregating five inversions, restricted to the western African coast, but within this restriction, it is able to colonise a geographically diverse range of the coastal regions from northern Mauritania to Angola (Lee & Lanzaro 2013)).

At finer scale the savanna chromosomal form, segregating to some degree every major inversions except 2Rk, is highly cosmopolitan, showing frequent crossover with the habitats of the more restricted Forest and Bamako forms (1-2 inversions each).

Exhaustive surveys in both Cameroon and Burkina Faso have underlined this relationship between niche expansions and speciation. Examining *An. gambiae, An. coluzzii* and *An. arabiensis* individuals from hundreds of separate locations spanning both countries, Simard *et al.* (Simard et al. 2009) and Constantini *et al.* (Costantini et al. 2009) both showed inversions to be shared across all species at the macrogeographic level, but to differentiate between species within each location: thereby facilitating isolation at the local level.

4: Sequence-based karyotyping of the 2Rb Inversion

4.1: Introduction:

Why do we need a test?

As we have seen in the previous chapter, inversions are both indicative and frequently causative of population structure, however the kinds of exhaustive study that allowed Coluzzi *et al.* to deduce their ecotypic associations are seldom performed these days; the manpower required to karyotype thousands of specimens is not often available and can represent poor value for money when compared with molecular tests and genotyping (Krzywinski & Besansky 2003). The development of molecular and sequence based tests (such as the molecular test for the ribosomal M/S marker (Fanello et al. 2002)) for can therefore facilitate the study of population structure by reducing the workload as compared to polytene chromosome analysis, and by permitting retrospective studies of previously collected samples where nurse cells or salivary glands are not available.

Ultimately, a better understanding of the distribution of inversions and the degree of gene flow between parapatric populations will not only allow us to investigate some of the fundamental evolutionary biology questions about speciation, but may have significant public health benefits, enabling the superior application of current control methods.

However understanding population structure such as this is also essential for the successful use of genome mapping techniques. Given the broad distribution of the major chromosomal inversions, it is inconceivable that any sampling that took place in all but the most restricted locations would avoid capturing inversion polymorphisms. An understanding of this structure is vital if a representative sample is to be taken. However, even more so than other features of structured populations, an uneven distribution of chromosomal inversions is likely to show up as a strong signal in any association study: in a low-LD background, any increase in this signal is more likely to appear significant. Methods of identification of these inversions are therefore of wide utility both for assessments of ecological associations in the field, and for the removal of structured populations in association studies.

The extensive work on the genetics of 2La, that has confirmed its previously proposed role in aridity tolerance and enabled a genetic dissection of this phenomenon to take place, has largely been made possible by the development of a PCR test by White *et al.* (White et al. 2007). Yet tests for the other inversions are either absent, or suffer from lack of applicability across broad geographic regions (Lobo et al. 2010).

4.2: The 2Rb Inversion

The 2Rb inversion has a complex demographic history. Though it is unrepresented in the geographically restricted species (*quadriannulatus, bwambae, melas, merus* and *amharicus*) the two karyotypes are maintained at high frequencies in *An. gambiae, An. coluzzii* and *An. arabiensis* mosquitoes. Within *Anopheles gambiae s.s.* it segregates to some degree in all of the classic five chromosomal forms (albeit rare within the forest and bissau forms), and is present on both sides of the great rift valley (Coluzzi et al. 2002).

Like $2L+^{a}$, the 2Rb inversion in *Anopheles gambiae* is believed to share the same origin as that in *Anopheles arabiensis* (White et al. 2009) and is therefore directly comparable across the two species. The locus itself is confounded with other inversions: it is found as either as a single inversion or as part of the 2Rbc inversion with which it shares it's distal breakpoint and it may have undergone more than one chromosomal rearrangement (Sharakhov et al. 2006). Indeed the 2Rb and c inversions are sufficiently close to comprise a single system, with the further complication that the distal 2Rc breakpoint is either shared, or in close proximity to the less frequent (and mutually exclusive) 2Rd and 2Ru inversions.

Due to the prevalence and non-random distribution of the 2Rb inversion it has long been hypothesized that this is related to adaptation to ecotypes, although the ecotype to which it is associated has not been clearly defined. 2Rb has been shown to be non-randomly distributed with respect to both aridity and the availability of anthropogenic habitats (Coluzzi et al. 1979). The 2Rb inversion demonstrates a similar latitudinal cline to the 2La inversion, though this is complicated by the presence of the 2Rbc inversion. That is, the three karyotypes show a gradation from near-fixation of $2R+^{bc}$ in the humid forested regions, high levels of 2Rb in the savanna, with increasing proportions of 2Rbc in the northern savanna ($2R+^{bc}$ is almost eliminated in the northern savanna), and near fixation of 2Rbc in the Sahel (Coluzzi et al. 1985) (see Figure 3.2, Section 3.7).

Links have also been posited between human habitation and the 2Rb inversion. Samplings in Nigeria have shown that, in *arabiensis*, the 2Rb inversion is near fixed within urban populations and at lower frequencies outside (Coluzzi et al. 1977), though the dominance of *arabiensis* in urban settings within this region prevented an assessment in *gambiae s.s.* Moreover, in both species, 2Rb carrying mosquitoes have shown a disposition towards outdoor resting; Coluzzi *et al.* sampled mosquitoes by methods that distinguished between endophagic and endophilic behaviours, and gave further indications of anthropophily and zoophily. 2Rb and bc individuals were significantly more frequent when sampled by landing catches on animals and humans outdoors, and in window exit traps; $2R+^{bc}$ individuals were more frequently found within indoor human landing catches and by indoor spray-catch sampling (Coluzzi et al. 1979). In addition, Petrarca and Beier, sampling in the Kisumu region

of Kenya have shown 2Rb homokaryotes in *gambiae* to have a lower circumsporozoite prevalence than heterokaryotypes or standard homokaryotes, despite having similarly high frequencies (90-100%) of human-blood-positive samples (although 2Rb/b sample sizes were low for this experiment) (Petrarca & Beier 1992). Interestingly the presence of one copy of the 'b' inversion also negated a significant effect on parasite prevalence of the 2La inversion (ibid.).

Large scale data, along the aridity cline, shows a strikingly similar 2Rb/bc distribution in both *gambiae* and *arabiensis* (Coluzzi et al. 1985). However at finer scales, care should be taken when making inferences from *arabiensis*. Despite the inversion sharing the same origin in both species, it does so against different backgrounds and shows different linkages with other inversions. (Coluzzi et al. 1979).

Indeed, in contrast to 2La, where the macrogeographic association has been specifically tested and its genetic basis dissected, in the 2Rb inversions system neither the precise ecological adaptation, nor the genes conferring these adaptive traits have been identified. Attempts to replicate the aridity tests that demonstrated 2La association have generally failed. 2Rb association with aridity tolerance is weakly supported at best, and probably confounded with body size (Fouet et al. 2012).

The extra complexity involved in the 2Rb inversion, and in particular the confounding effect of other related inversions, will almost certainly make these relationships more difficult to discern. It is additionally possible that the ecophenotype has lower penetration than that seen for 2La - indeed the significantly lower F_{ST} that is seen in 2Rb, as compared to 2La (Cheng et al. 2012), is likely to indicate a less divergent phenotype. If, indeed, the correct phenotype is being tested. There are numerous linked aspects that change predictably along an aridity cline – reductions in the availability of natural (non-anthropogenic) habitats and a subsequent reduction in seasonality being the most obvious examples. Larval habitat choice, exophily and aestivation could all potentially be linked to aridity adaptations. Differences in antiplasmodial responses have also never been tested.

It is an open question whether novel hypotheses need to be deduced by large scale sampling or exploratory experiments, or if 2Rb merely needs to be isolated from the confounding 2Rbc and 2La inversions so that previously suggested phenotypic associations can be confirmed. What is certain is that, in either case, current karyotyping methods - being destructive to the mosquito, impossible to apply retrospectively, and laborious - will not enable us to gather the samples sizes that will be needed.

The development of a reliable molecular test for 2Rb will not only enable better assaying of population structure in the field – opening up possibilities for strata-control in the development of experimental samples – but it may allow us to answer long standing and

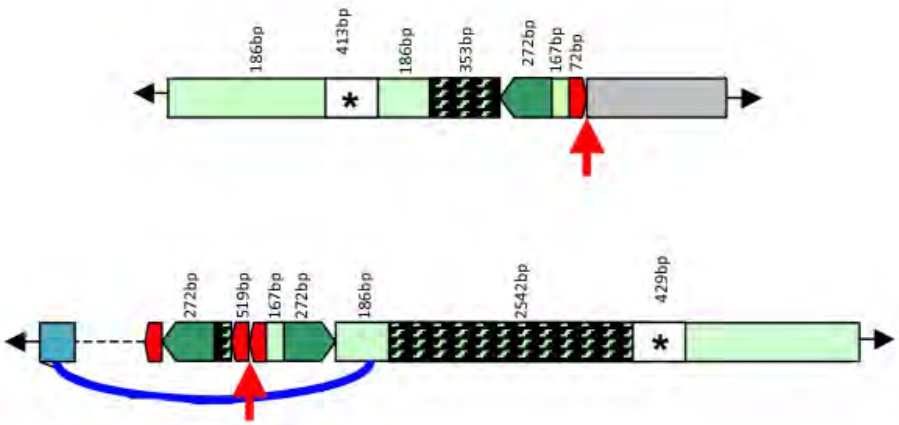
fundamental questions about the ecotypification of the *Anopheles gambiae* / *coluzzii* species pair.

4.3: Current methods of karyotyping

Molecular tests (*Anopheles gambiae*)

A previous attempt has been made to devise a molecular test for 2Rb in the same manner as one was developed for 2La. Lobo *et al.* devised a PCR test based upon a manual reassembly of the inversion breakpoints (Lobo et al. 2010).
A read close to the proximal 2Rb breakpoint was identified, and reads spanning the breakpoint were deduced by chromosome walking using traces identified as part of the *An. gambiae* M/S genome project (Lawniczak et al. 2010). Reassembly was therefore performed from whole genome sequence of an S-Pimperena colony, a colony that is known to be homokaryotypic for 2Rb. These manually assembled sequences were then used to identify scaffold sequences of the Pimperena (*An. gambiae s.s., 2Rb/b*) and Mali-NIH (*An. coluzzii* 2Rbc/bc) colonies, allowing those highly repetitive sequences that could not be manually assembled to be bridged.
The sequences of these two inverted colonies, should therefore describe two inverted forms, one with the breakpoint-sharing 2Rc inversion, while the PEST colony ("pink-eye standard" – referring to all inversion being in their 'standard' orientation) represents an uninverted $2R+^{bc}$ example.
The high degree of repetitive DNA present in the breakpoint region prevented a PCR test being devised that bridged the breakpoint itself, as had been possible for the 2La inversion (White et al. 2007). Instead the authors exploited a linked insertion around 1kb from the breakpoint to distinguish the forms.

Figure 4.1: breakpoint structure of the 2Rb inversion (distal)



Schematic overview of uninverted (top) and inverted (bottom) breakpoint of the 2Rb proximal inversion, as assembled by Lobo et al. Blue and gray boxes represent flanking sequence outside of the breakpoint; red boxes low-complexity, repetitive DNA. Dark green arrows represent the sequence that has been duplicated in the 2Rb inversion forms. Red arrows represent the putative breakpoints. Dashed lines represent gaps in the breakpoint assembly. Blue curved lines represent sequence linked by mate-pair information. The while box containing an asterisk indicates the region targeted by the PCR diagnostic assay. Segment sizes are given in above the image – it should be noted that these are the sizes in the Lobo et al. reassembly and not their size when re-aligned to the PEST genome.

Source: adapted from Lobo et al. Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion. (Lobo et al. 2010)

This PCR test, however, demonstrated poor accuracy when tested by Lobo *et al.* against field samples. Although this PCR test proved reliable in identifying b/b forms in southern Mali (the region that gave rise to the Mali-NIH and Pimperena colonies), giving a 100% accuracy rate from 42 samples, it showed extremely poor reliability in the malian +/+ samples at under 45% accuracy. Samples from Cameroon showed the opposite tendency, with 96% of +/+ samples being called correctly (n=70) yet only 85% and 81% of b/+ and b/b samples called correctly (n=65,37).

The miscall rate in Mali appears particularly high considering they sequences were derived from colonies founded in this region. However comparisons with flanking inversion karyotypes indicated that the high miscall rate in Mali corresponded perfectly with the presence of the ‘cu’ inversions on the $2R+^{b}$ backround. In these cases the $2R+^{b}$

chromosomes possessed the 2Rb-like insertion, indicating a potential re-inversion of the 2Rb back to $2R+^{b}$ in this region. The 11% failure rate within Cameroon was not explained.
The poor accuracy of the Lobo test, and in particular its variance over geographic distance limits its use in the field. The use of linked variation for the Lobo *et al.* test, rather than the breakpoint itself, means the test is susceptible to further genomic rearrangements or small indels – events that are likely to be more frequent in the highly repetitive sequence that characterises the breakpoints.
The development of a robust test for the 2Rb inversion is therefore of significant utility.

De-novo detection methods

Available methods

There are a number of methods of de-novo inversion calling, with varying degrees of accuracy and ease of use. Inversion detection methods spit into two broad categories: the first is those that attempt to identify breakpoints themselves – either by reassembly of the breakpoint region, or by identification of translocation of paired-end reads; the second category consists of algorithms that attempt to identify demographic signals associated with the inversion (arising from the suppression of recombination within the inversion).
The former approach is taken by a number of programs, such as VariationHunter (Hormozdiari et al. 2010), SVdetect (Zeitouni et al. 2010) and breakdancer (Chen et al. 2009). Although these algorithms perform creditably with simulated data, and in datasets where the paired-end strategy was designed with them in mind, they are less effective where reference errors occur, or where insert sizes are short. In work by Lledo *et al.* direct comparisons of paired-end mapping algorithms have indicated that datasets where template lengths are under 500bp are likely to miss up to 50% of simple inversions where the breakpoint sequence is accessible and unmodified. This false-negative rate rises to more than 80% when considering inversions associated with segmental duplications (Lucas Lledó & Cáceres 2013).
Algorithms that consider linkage patterns or demographic signals are not reliant upon insert sizes, and are therefore perhaps more likely to work on datasets that were not purpose-designed. The reduction in recombination in an inversion will lead to an increase in LD across the inversion, and in particular in the breakpoint region. Bansal *et al.* have developed and tested this theory in human HapMap data, identifying 176 candidate inversions ranging from 200 kb to several megabases in length. However the method is susceptible to natural variation in LD (common in wild *Anopheles* populations (Wang-Sattler et al. 2007)), and is

only designed to detect inversions that are inverted with respect to the reference in the majority of chromosomes in a population.

Perhaps the most promising approach is the comparison of distributions of alleles in populations that segregate inversions: as a consequence of drift between inversion karyotypes and a lack of recombination in heterokaryotes, populations that segregate an inversion – even if panmictic – will appear as admixtures of two diverged populations within the inverted region. This effect has been exploited by Sindi *et al.* who have developed a method to identify inversions by examining haplotype frequencies around inversion breakpoints (Sindi & Raphael 2010), this approach was implemented (and further generalised to genotype frequencies) in the inveRsion package (Cáceres et al. 2012). Testing in human HapMap (phase 2) data, however, indicates that despite an improvement over the method of Bansal *et al.* this approach still requires minor inversion frequencies over 20% in order to accurately predict variants.

PCA based approaches can also show inversion polymorphisms given sufficient sample sizes, and have been used to scan for novel inversions in (phase III) Hapmap data. As with the haplotype based methods, these exploit the behaviour of heterozygotes as 1:1 admixtures of each homokaryotype. Ma and Amos devised a parameter to measure the relative equidistance of the putative heterokaryotype cluster from the two homokaryotypes – calling an inversion if the clusters were sufficiently defined (within-cluster sum of squares < 0.08) and the heterokaryotype cluster deviated less than 8% from the centre of the two side clusters (Ma & Amos 2012).

All of these approaches have been developed and tested in human hapmap samples, where the frequency of miscalled SNPs is low and the reference genome is of extremely high quality (Frazer et al. 2007). Despite inversions having been a long-standing feature of dipteran genomic studies few studies have successfully applied inversion calling algorithms to these genomes. Dipteran genomics, and non-model genomes in general present a variety of different problems – largely relating to lower sequence coverage and frequent mis-assembly – that often preclude these methods being used without modification.

A manual combination approach was taken by Corbett-detig *et al.* in D.melanogaster (Corbett-Detig et al. 2012). Using paired-end mapping of reads with 300bp insert sizes to identify putative breakpoint regions, followed by examination of F_{ST} in the proposed breakpoint regions to confirm the validity of the breakpoints. This combination approach identified 12 putative loci that demonstrated both translocated reads and high F_{ST} in both breakpoints; these results were effectively confirmed by a survey in natural populations of *D. melanogaster* that identified the same set of inversions (Aulard et al. 2002).

Use of de-novo detection in *An. gambiae / coluzzii*

The specific challenges of *Anopheles* inversions can prevent most of these methods being used. *Anopheles* populations in the wild exhibit an excess of low-frequency SNPs, with very little linkage between chromosomal loci (Ag1000G Consortium n.d.). Even at nearby loci, linkage disequilibrium (LD) is negligible: r^2 is around 0.05 for variants within 1kb (Neafsey et al. 2010), a level that is significantly lower than in humans where SNPs within 1kb would be in near-perfect linkage (r^2 =~ 1); r^2 of 0.05 is a level more typically found in markers separated by half a megabase or more (Shifman 2003).

Against this background, detecting inversions via rises in LD or patterns of linked genotypes is problematic. For either of these detection algorithms to work, prior filtering of low-frequency variants would need to be performed, necessitating a large and geographically widely distributed dataset.

PCA based methods can be more robust at lower sample sizes, given that they inherently will identify the most informative SNPs in a given locus. This is generally assumed to be those that are linked to the inversion itself. Indeed PCA calling has been used with other high throughput genotyping studies, successfully identifying 2Rj, b, c, u and 2La inversions in 60 Malian samples of differing chromosomal forms (Neafsey et al. 2010). However attempts to apply the Ma and Amos method to *Anopheles* datasets were unsuccessful, with unlinked variation and non-inversion related drift preventing the ‘equidistance statistic’ (δ) from falling below the 0.08 limit even in regions where an inversion was known to segregate.

Although this statistic could be adjusted, far more grave problems were presented by the population structure of *Anopheles gambiae* itself. As shown above (section 3.5) reproductive isolation in these samples is frequent and often cryptic. This can cause frequent false-positives in PCA methods. PCA-based calling relies upon identifying heterokaryotes as a perfect admixture of the standard and inverted forms; structured populations will exhibit high degrees of natural admixture, which can give inversion-like signals where no inversion is present, any inversion identified using this method would therefore need to be confirmed using an independent detection method such as polytene chromosome analysis.

Identification of typing markers in *An. gambiae*

Whilst the use of de-novo methods to call novel inversions may be unattainable at the moment, the availability of assembled sequence for two of the major inversions enables us to identify reliable markers that are closely, or definitively linked to the inversion event itself. These can then themselves be used as a proxy for the presence of the entire inversion.

If the chromosomal reattachment takes place against a non-clean fracturing of the chromosome the inversion event can result in the duplication of sequence. Inversions can therefore be divided into those with 'cut-and-paste' and 'staggered' breakpoints (Corbett-Detig et al. 2012).

Copy number variants in *An. gambiae* inversions

Both the 2La and 2Rb inversions are characterised by such a tandem duplication at the breakpoint site. Being located proximal to the breakpoint in a region of almost complete suppression of recombination, these features should give reliable copy-number polymorphism that will be less susceptible to feature loss than the deletion used by Lobo *et al.*.

CNV calling in these regions can be challenging, however. Genome accessibility in inversion breakpoints is low, indeed inversion polymorphisms can cause these regions to be incorrectly assembled in the reference genome. Moreover, even where correctly assembled, the high F_{ST} differences between karyotypes can cause reductions in the alignment for one of the forms, a particular problem for copy-number assessments based upon sequence coverage.

4.4: Datasets

The choice of dataset that is available, or in which we hope to call the inversion, will define the inversion-calling methods that are able to be used. Prediction methods based on reassembly of breakpoints or mismatched read pairs will benefit from predictability – inbred colonies or lab strains will generate large amounts of predictable sequence with high frequencies of each karyotype if they are present at all. Alternatively, wild colonies are more likely to have rare inversions, and will have significant natural heterogeneity which will affect hybridisation efficiencies in an unpredictable manner; they are, however, the only datasets in which demographic measures (such as LD rises or principal components prediction) are likely to have any tractability.

in devising the karyotype calling method, we have used both a dataset with large numbers of predictable copy numbers for development, and a further highly heterogeneous dataset for testing and refinement of the method.

Colony crosses dataset

To develop a typing method for the karyotypes, a set of colony crosses were used. These consist of a set of F1 crosses of highly inbred colonies from geographically diverse locations, part of a wider study into genome accessibility in *Anopheles gambiae* / *coluzzii*.
The colonies that were crossed included the Mali-NIH and Pimperena colonies used for the M/S sequencing efforts (Lawniczak et al. 2010) along with colonies from West ('Akron', from Benin; 'G3' from The Gambia; 'Ghana' from Ghana – all *coluzzii*) and East Africa (Kisumu, from Kenya – both *gambiae s.s.*). Each cross contains parental sequences and individual sequences for each of the F1 offspring, along with additional colony individuals that were unsuccessful at mating.

The dataset was generated by the Wellcome Trust Centre for Human Genomics, Liverpool School of Tropical Medicine and Wellcome Trust Sanger Institute. Permission was given for its use in this thesis by Prof D Kwiatkowski. A manuscript describing the dataset is in preparation and expected to be submitted in 2015.

Anopheles 1000-genomes dataset (release 1)

SVC testing and refinement was performed within a large geographically and genetically diverse dataset. The *Anopheles* 1000-genome project (Ag1kG) is a global collaboration that aims to sample mosquitoes from across the continent; generating a broad-scale and high-resolution view of genetic variation in the vector species. As it is designed not as a single comprehensive dataset, but as a series of subprojects, it does not have a comprehensive representation of all of the major population subdivisions. It is also, in addition, currently in the first phase of a 3-phase data release (containing 765 samples, less than half of the expected final total). A subset of the Ag1kG set has been karyotyped by polytene chromosome analysis prior to sequencing, providing an additional validation step.

The dataset was developed as part of a large collaborative project including members from the Wellcome Trust Sanger Institute, Wellcome Trust Centre for Human Genomics, Liverpool School of Tropical Medicine, Pasteur Institute, University of Notre Dame, University of Rome, and Imperial College (Ag1000G 2014). A manuscript describing the dataset is currently in preparation and expected to be submitted by the end of 2014 (Ag1000G Consortium n.d.)

4.5 Rationale

It is the larger Ag1kG dataset that provides much of the necessity for the development of a novel typing method. The Ag1kG dataset is the largest, most comprehensive dataset of next-generation sequencing sets available for *Anopheles spp*. spanning both sides of the great rift valley and with representation of *Anopheles gambiae* and *coluzzii.* However whilst 26% of

the samples have been typed for 2La, little or no karyotyping has been performed on the other inversions; only 16% of samples have 2R karyotypes called and this applies only to samples from one region. The deduction of karyotypes is particularly important for the frequent and highly differentiated 2Rb inversion: reliable sequence-based karyotyping will allow us to examine the distribution of the 2Rb inversion across the continent, and is vital if we are to avoid the potential confounding effects of this inversion in population structure and association work.

The lack of a large number of samples that have been karyotyped by polytene chromosome analysis necessitates that any calling method is internally robust, and preferably that it is able to confirm the karyotypes using multiple genomic features – features that are as independent as possible. Following the work of Corbett-Detig *et al.*(ibid), who identified misaligned read pairs and confirmed with F_{ST} increases in the breakpoint regions, this would include identification of the breakpoints themselves along with confirmation using the demographic effects of those inversions.

Karyotyping of the other inversions is more challenging. Breakpoint reassembly has not been attempted for the 2Rc, 2Ru and 2Rd inversions, and as a result linked features or precise breakpoint coordinates are not available for analysis. Breakpoint identification would therefore require the use of de-novo methods of identification with little opportunity for verification. The 2Rj inversion, in contrast, has been well characterised and assembled breakpoint sequence is available (Coulibaly et al. 2007); however this inversion is present at high frequencies only within the Bamako chromosomal form, which exhibits a very narrow geographic distribution within Mali and is not represented within the Ag1kG dataset. Nevertheless, it is worth noting that the development of de-novo calling methods (or the adaptation of those described in section 4.3) will be greatly assisted by having a large number of accurate 2Rb calls. This type of development is currently not possible with the 2La breakpoint due to the extremely high F_{ST} differences between 2La and 2L+ karyotypes which causes a drastic drop in read coverage from one karyotype to the other (White et al. 2009), this degree of misalignment is not seen for any of the other inversions and precludes 2La being used for method development. The 2Rb inversion, in contrast, shows more moderate F_{ST} differences (ibid), and does not present the same issues with regard to read alignment. Whilst generating more widely applicable algorithms is sadly beyond the scope of this analysis, it is hoped that this work could provide the basis for their development in the near future.

The investigation proceeds in three parts: development of the karyotype calling method, use of this method in the Ag1kG dataset, and an assessment of the method using alternative karyotyping methods relying upon demographic signals. Each of these stages depends upon the results of the prior steps.

4.6: Methods I: SVC karyotyping of the 2Rb/2La inversions

Due to the short insert sizes chosen for both the colony crosses and Ag1kG datasets and the high frequency of misaligned reads throughout the *Anopheles* genome, breakpoint identification via the identification of misaligned reads was ruled out as a viable approach. Similarly, the extreme variation in *Anopheles gambiae* and frequent natural admixture was also considered a likely source of false-positive results for admixture based approaches – due to the lack of a broad-geography confirmation set this was also ruled out as a viable approach. As a result, identification of the 2Rb and 2La breakpoints was made via features that are inherently linked to the breakpoint itself, the tandem duplications identified by Sharakhov *et al.* (2006) and Lobo *et al.* (2010)

In order to devise a reliable dataset for the development of the method, multiple evidence streams were used to call karyotypes in the parents of the colony crosses, including previously identified karyotypes (MR4 n.d.) and novel variation data. Since the alignments in breakpoint regions were of too low quality for standard CNV calling methods to work, the samples with unambiguous karyotypes (parents and predictable F1s) were used to train a machine-learning algorithm to detect 2,3 and 4-copy samples, with validation performed according to Mendelian inheritance rules.

Finally, the algorithm was applied to the *Anopheles* 1000-genomes dataset to determine it's reliability across broad geographic areas. This also enabled the identification of reliable typing SNPs that could be used for karyotyping where genome sequencing was unavailable.

Karyotyping of colony crosses

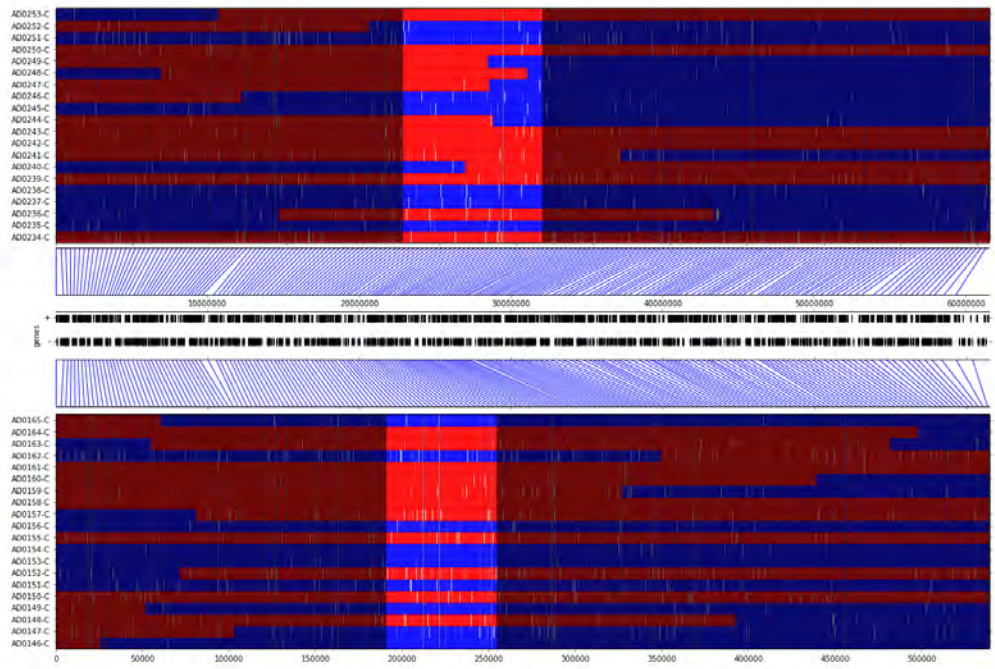
No single method of karyotyping was suitable for all individuals within the colony cross dataset, however using evidence from heterozygosity rates, prior karyotyping, and recombination patterns the majority of karyotypes for the cross parents were able to be determined. Where crosses were between two homokaryotes, this would also result in the confident prediction of all F1 offspring, providing further validation of the parental karyotype calls.

A subset of the MR4 colonies have been previously karyotyped as being fixed for individual inversions: Mali-NIH are known to be 2La/a, 2Rbc/bc; Pimperena 2La/+ 2Rb/b, and G3 2La/+ 2R+ (MR4 n.d.).

The availability of parental and F1 sequence has enabled phasing to be performed, demonstrating precise recombination boundaries for each of the offspring (Alistair Miles, WTCHG – private communication). This has allowed some of the parents to be definitively marked as homokaryotes, where a recombination boundary is found within a known

inversion. Furthermore, inverted homokaryotes are easily distinguishable from standards, since the presence of a recombination within an inversion will lead to artificial 'recombinations' at the inversion boundaries when mapped back to a homokaryotypic standard PEST reference (see figure 4.2).

Figure 4.2: phasing comparison between hetero and homokaryotypic F1 sequences :



Phasing graphs for F1 sequences of colony crosses showing the suppression of recombination within heterokaryotypic parents. The top set is homokaryotypic standard (2R+/+), while the bottom is heterokaryotypic (2Rb/+). This plot shows the distributions of maternal chromosomes within the F1 offspring only - each horizontal line represents one chromosome of the F1 offspring. Maternal chromosomes are painted red or blue, therefore colour changes represent recombination sites. The highlighted block is the location of the 2Rb inversion.

Source: Alistair Miles, WTCHG, Private communication

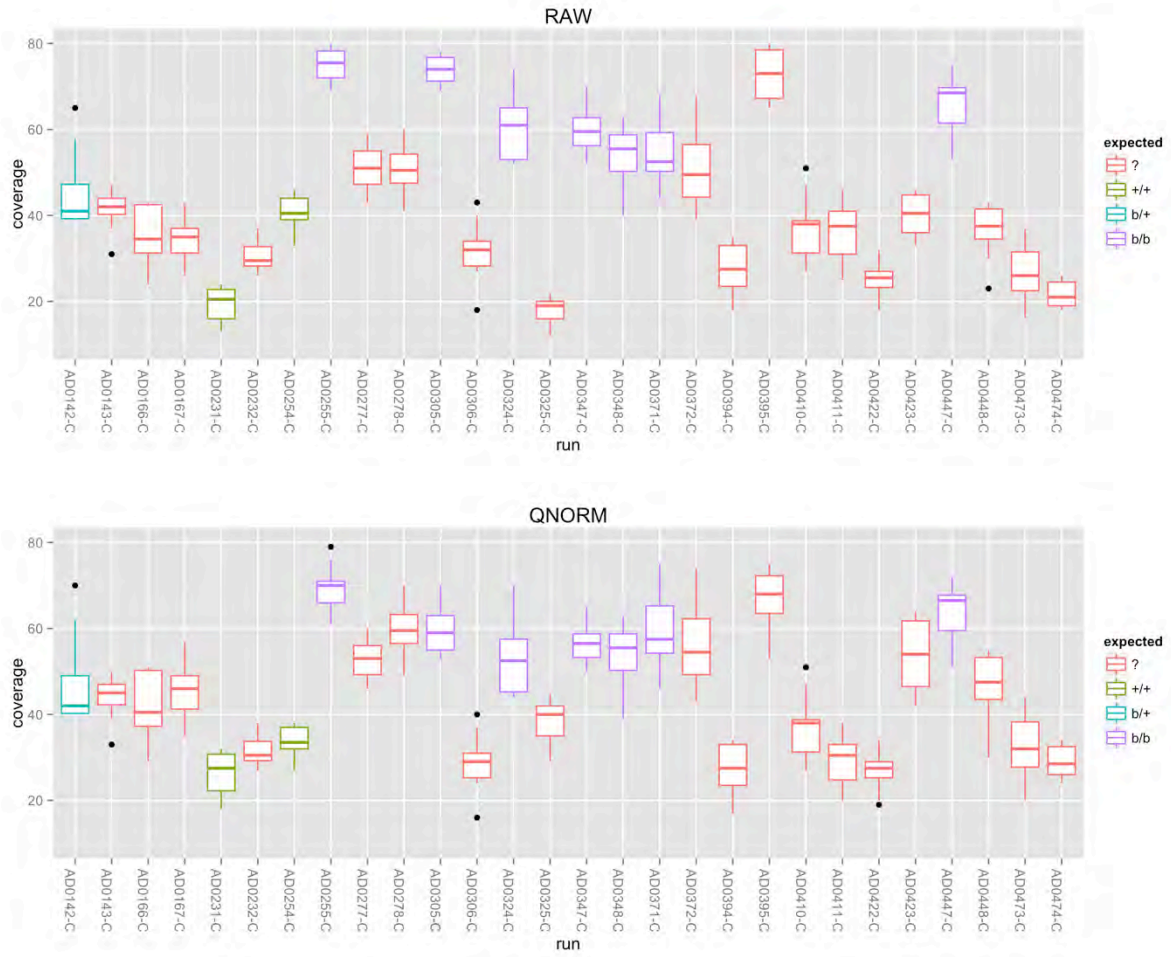
Finally, even where a recombination event does not take place within an inversion, the frequency of certain genotypes can indicate which orientation is present in the parent. A rise in 'homozygous alternative' (i.e. non-PEST) genotypes, if coincident with a known inversion, can strongly indicate an inverted karyotype, and a strong rise in heterozygotes can indicate a heterokaryote. This effect is particularly pronounced for the inversions with higher F_{ST} differences between forms, such as 2La and, to a lesser extent, 2Rb.

Using these three evidence streams we were able to identify known parental karyotypes for a subset of the colony crosses datasets, and as a result predicted karyotype distributions for the F1s.

Tandem duplication identification and typing

The method was developed using the 2Rb inversion; like the 2La inversion this inversion has assembled sequence for the breakpoint region and a broad distribution, however it does not have the same high degree of F_{ST} between karyotypes that might affect read coverage.
The locus in the 2Rb breakpoint that was duplicated in the Lobo *et al.* manual assembly was re-identified in the PEST chromosome by a BLAT alignment of the 2Rb sequence back to the reference, this was split into 50bp fragments allowing the identification of all misaligned or misplaced contigs. The 2Rb duplication in the distal breakpoint was localised to the region 2R:19025132-19025693 (561bp), the proximal duplication to 2R:26748285-26748653 (368bp).
Comparisons of the read coverage for both duplications in the colony crosses dataset showed characteristic rises in read coverage for the putative duplicated regions, providing support for this as a typable CNV linked to the inversion karyotype. In all cases the larger distal breakpoint duplications demonstrated a stronger relationship between karyotype and mean depth, so this breakpoint was chosen for the typing assay. Since depth calls were affected by the overall read coverage of the samples, all of the samples were quantile normalised prior to comparison (Bolstad et al. 2003), ensuring that mean depth and the overall distributions were identical between sets (see figure 4.3).

Figure 4.3: distribution of breakpoint depths across the distal 2Rb locus



Boxplot showing the distribution of coverage values within the 2Rb distal breakpoint for each of the parents of the colony crosses dataset (for 100bp bins across a 1kb region spanning the region 2R:19025000-19026000bp). Separate boxplots are given for the raw (top) and quantile normalised (bottom) values, showing the increased reliability of the quantile normalised samples. Expected karyotypes are based on manual assessments of the recombination patterns shown by phased data, and the relative numbers of heterokaryotes in the breakpoint regions.
It should be noted that a linear classification on this basis (e.g. based on mean or median depth across the entire 1kb locus) will not cleanly separate the three copy numbers.

Support Vector Classification of breakpoint copy number polymorphisms

Although there were strong indications of a valid CNV in this region, and despite the normalization, the coverage around this locus was variable; this is almost certainly related to the poor genome accessibility in this region. This stochastic variation in depth, and the small size of the locus, prevented an obvious depth cut-off being set that would cleanly separate the two, three and four copy-numbers. A machine-learning approach was therefore used.

SVC training

Mean read depth was calculated for a 100bp fragments tiling across the duplicated locus from 19025-19026kb including some of the flanking region on either side. A support vector classifier (SVC) was then used to classify the depth profiles into specific karyotypes. That is, given the ten 100bp depth figures, and a sufficiently large training set, the SVC will describe a ten-dimensional space that best defines each of the three karyotypes. Once defined, each novel example will be assessed for its probability to fit into each class and a final classification given. Using the known parents and predictable F1s as a training set, the SVC was therefore trained to classify depth values into 2, 3 and 4-copy samples. The specific SVC implementation used is available in the scikit-learn package (version 0.14.1), classifier code is available in appendix 3.

4.7 Methods II: PCA calling / karyotype validation

As was seen for the Lobo assay, results can be variable across even relatively close geographic regions. Therefore testing of the SVC by its application to a broad dataset was vital. In addition, this enables the identification of robust SNP-based markers that do not require high-coverage sequence.
A panel of 20 SNPs were selected according to the method of Chakraborty and Weiss (Chakraborty & Weiss 1988); that is, those with the highest allele frequency differences between homokaryotypic standard and inverted groups were chosen as karyotyping markers, with a further filter added to remove any markers where heterozygotes did not exhibit an allele frequency at or near to 0.5 (see figure 4.6).

PCA calling of karyotypes

In order to derive an independent assessment of karyotype, a PCA-based approach was taken. It should be noted that previous papers have used PCA with manual identification of karyotype pools as corroboratory evidence (Neafsey et al. 2010), however to derive a formal classification for our dataset, k-means clustering was applied to the PCA and a cutoff of the within-cluster sum-of-squares was applied in order to separate groups that segregate the inversion and do not (see table 4.3).
Karyotypes were assigned to the dataset on the basis of variance from the homokaryotypic standard PEST reference (see figure 4.8). Concordance between the SVC, Tag-SNP and PCA datasets was calculated only using those samples where the WSS was under 500.

2La Karyotyping via CNV-typing of breakpoint duplications

As a further test of the calling algorithm, and to extend the utility of the Ag1kG dataset, the same methods were applied to identifying the 2La inversion. Like the 2Rb inversion, 2La derives from a staggered breakpoint, that has lead to the duplication of the terminal exon in two genes, one at each end of the inversion: zinc-finger protein (*ZNF-183* / AGAP007068) at 2L:20521765-20523605 and an iduronate-2-sulfatase precursor (*IDS* / AGAP005778) at 2L:42163507-42164602. (Sharakhov et al. 2006).
As for 2Rb, predictable karyotypes were identified by analysis of heterozygosity rates and recombination breakpoints in the colony crosses dataset (see table 4.1). Quantile normalisation was applied and mean depth was then calculated for 100bp bins across a 1kb fragment encompassing each duplicated exon. Again, the distal breakpoint demonstrated higher accessibility than the proximal and was therefore used for the training region for the SVC caller (2L:42164-42165kb).

4.8 Methods III: linkage and genotype frequency calculations

To provide a preliminary overview of the distributions of the inversions and their segregation in the Ag1kG dataset, karyotype frequencies were assayed for Hardy-Weinberg equilibrium by chi-squared test. LD was also calculated between tag SNPs and the reset of the genome, in order to both demonstrate the higher physical linkage across the inversion, and to demonstrate any statistical linkage between other regions of the genome. The bioconductor suite was used to calculate LD (SNPstats) and Hardy-Weinberg equilibrium (genetics) (Gentleman et al. 2004).

4.9 Results I: karyotyping & method validation

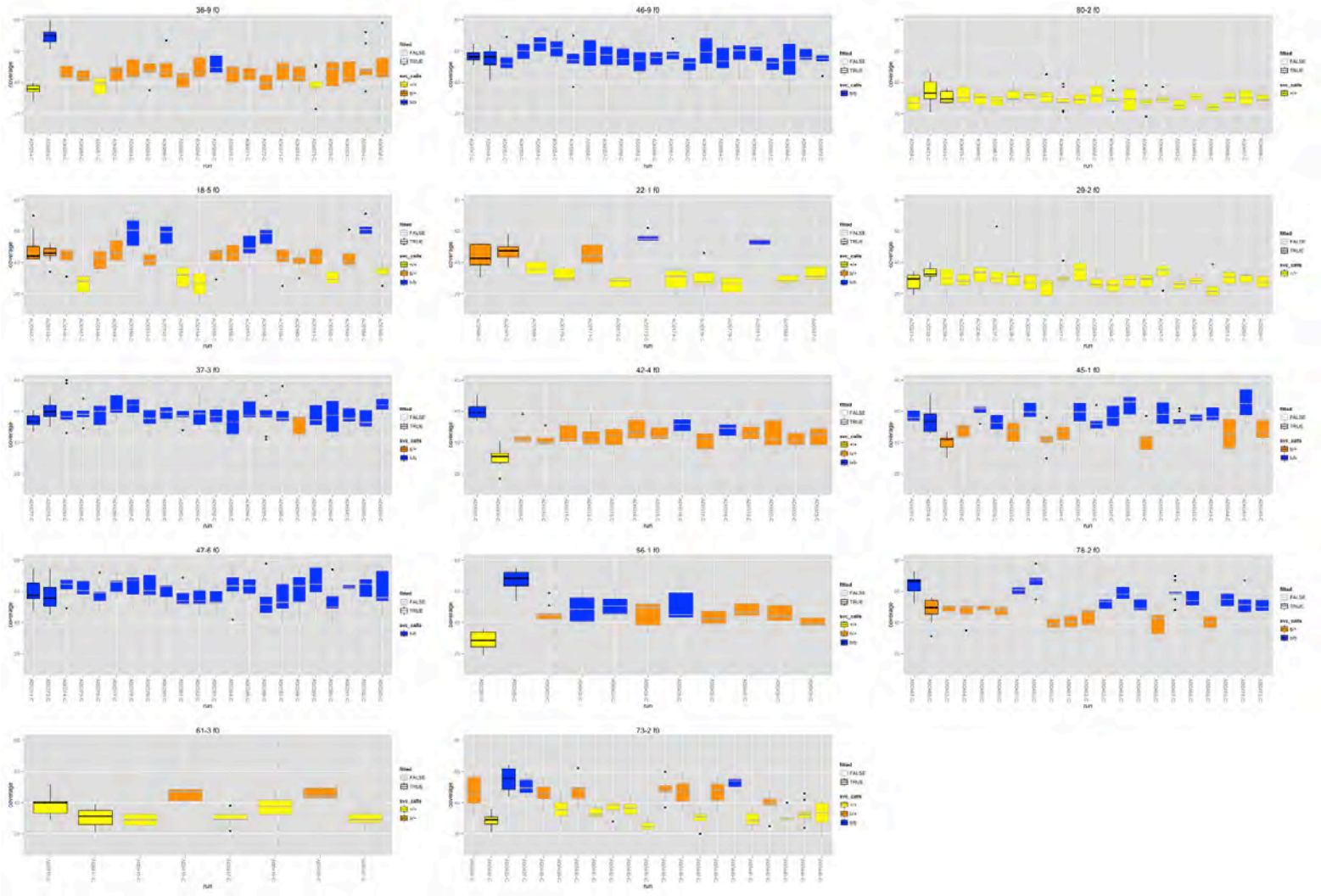
Examination of recombination patterns and heterozygosity in the colony cross parents allowed the identification of all heterokaryotes unambiguously and inverted forms where a recombination occurred within the breakpoint. The training set consisted of 71 individuals for the 2La and 2Rb inversions respectively.
Validation was then performed using the rest of the (unknown karyotype) F1 sequences. The availability of both parental and offspring sequences allowing us to define any calls which defied Mendelian inheritance rules as miscalls. Using these criteria, of the 238 samples tested, only four were found to be discordant. Whilst this may be an underestimate, ignoring miscalls that did not break Mendelian inheritance rules, an error rate of 2.05% compares

favourably to the other available typing methods (PCR assay 11-26%, polytene chromosome analysis: 2-5%). See Figure 4.4 for all SVC call results.

Table 4.1 Manual karyotype calls in colony-crosses data

				2Rb			2La		
indiv	sex	ms	cross	het-alt call	Phasing call	estimated copy number	het-alt call	Phasing call	estimated copy number
AD0142-C	F	m	18-5	b/+		3	a+		3
AD0143-C	M	ms	18-5			3			
AD0166-C	F	ms	22-1				a+		3
AD0167-C	M	m	22-1				a+		3
AD0231-C	F	m	29-2		std	2	aa	aa	2
AD0232-C	M	m	29-2				a+		3
AD0254-C	F	m	36-9		std	2	a+		3
AD0255-C	M	m	36-9	b/b	b/b	4	aa	aa	2
AD0277-C	F	s	37-3				a+		3
AD0278-C	M	s	37-3				a+		3
AD0305-C	F	m	42-4	b/b	b/b	4	aa	aa	2
AD0306-C	M	ms	42-4						
AD0324-C	F	m	45-1	b/b		4	aa	aa	2
AD0325-C	M	s	45-1				a+		3
AD0347-C	F	s	46-9	b/b		4	a+		3
AD0348-C	M	m	46-9	b/b	b/b	4	aa	aa	2
AD0371-C	F	m	47-6	b/b		4	aa	aa	2
AD0372-C	M	s	47-6				a+		3
AD0394-C	F	s	56-1						
AD0395-C	M	m	56-1						
AD0410-C	F	s	61-3				a+		3
AD0411-C	M	m	61-3				a+		3
AD0422-C	F	m	73-2				a+		3
AD0423-C	M	m	73-2				a+		3
AD0447-C	F	m	78-2	b/b		4	aa	aa	2
AD0448-C	M	ms	78-2				a+		3
AD0473-C	F	s	80-2				a+		3
AD0474-C	M	m	80-2				aa	aa	2

Figure 4.4: SVC calling results for colony cross samples : 2Rb



2Rb SVC calls for the entire colony crosses set. Standard forms are shown in yellow, heterokaryotes orange and inverted homokaryotes blue. Colony parents are the leftmost pair in each plot.
Mendelian errors were called if the offspring could not have emerged from the parents, or if the parent itself was miscalled.
In each case the most parsimonious explanation was used

Karyotype calls in the Ag1kG dataset

The SVC that was derived from the colony crosses was applied to all samples from the Ag1kG dataset, resulting in 765 calls with a calculated probability for each class. Calls were initially validated by comparison to a principal components analysis of SNPs within the inversion region. Principal components analysis was performed using the 'prcomp' method in R (package 'stats', R version 3.0.3 - components are calculated by singular value decomposition). K-means clustering was used to identify distinct clusters within the PCA, and karyotypes were assigned to these clusters manually, using the locations of the manually-karyotyped samples as a guide.

It is believed that most of the miscalls are related to poor low read coverage in the breakpoint regions of some of the Ag1kG samples, reducing our ability to accurately gauge copy number. However SNP calls – not being as susceptible to low read coverage – should be more robust to poor quality sequence sets.

Training the SVC for the 2La inversion was more challenging. Higher divergence between 2La / 2L+ karyotypes, and a lack of predictable F1 genotypes in the colony crosses dataset reduced the efficacy of the SVC calling, as well as reducing our ability to call definitive Mendelian errors in the testing set – only four samples out of 238 were clear as contravening Mendelian inheritance, however this almost certainly underestimates the miscall rate. Application of the 2La SVC to the Ag1kG set demonstrated similar concordance with PCA-derived karyotypes as was seen in 2Rb, but far lower median probabilities from the SVC caller.

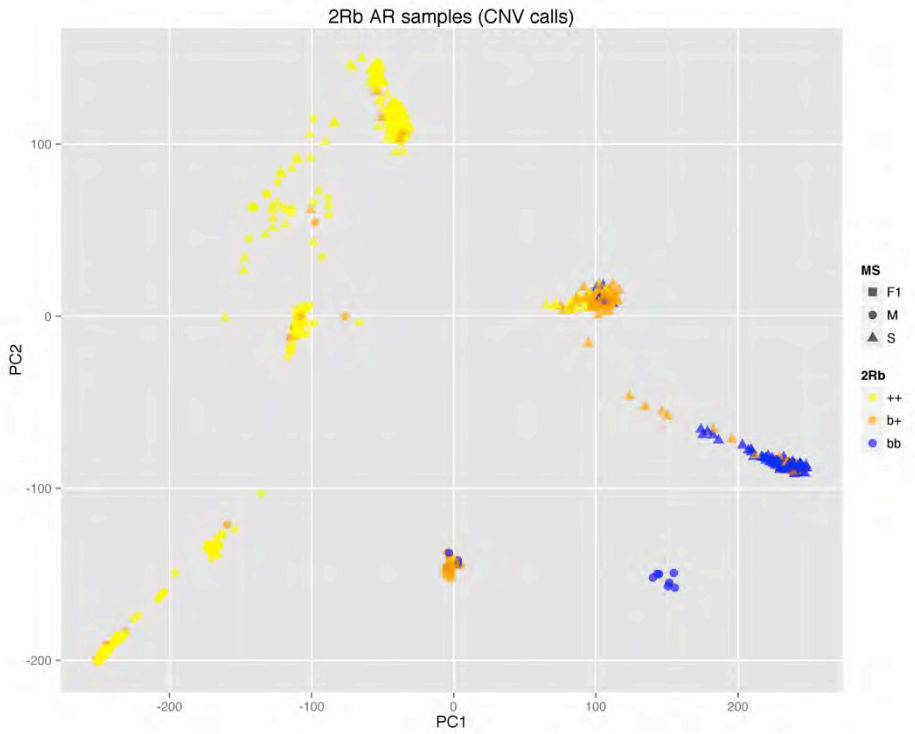
2Rb Tag-SNP identification

Further refinement of the SVC calls was made by selection SNPs that typed for the breakpoint duplication – i.e. the majority of correct calls in each cluster. Due to the large sample sizes, despite the small percentage of miscalls, this resulted in an improvement over the SVC alone, and an apparently viable method of accurately typing SNPs (see following section for validation). The use of the 'most different' SNPs in fact removed the noise inherent in the SVC classification, and typing using the 20 SNPs with the highest MAF-delta (by selecting the modal genotype from the panel) gave SNP calls of 100% accuracy according to the PCA comparison (see figure 4.5-4.7).

Indeed the panel of 20 SNPs was excessive for calling this inversion; assessment of subsets of these SNPs identified an individual SNP (2R:26371581) that was able to identify 95% of all candidates with under 1% of samples miscalled. Increasing this to a 3 SNP panel resulted in

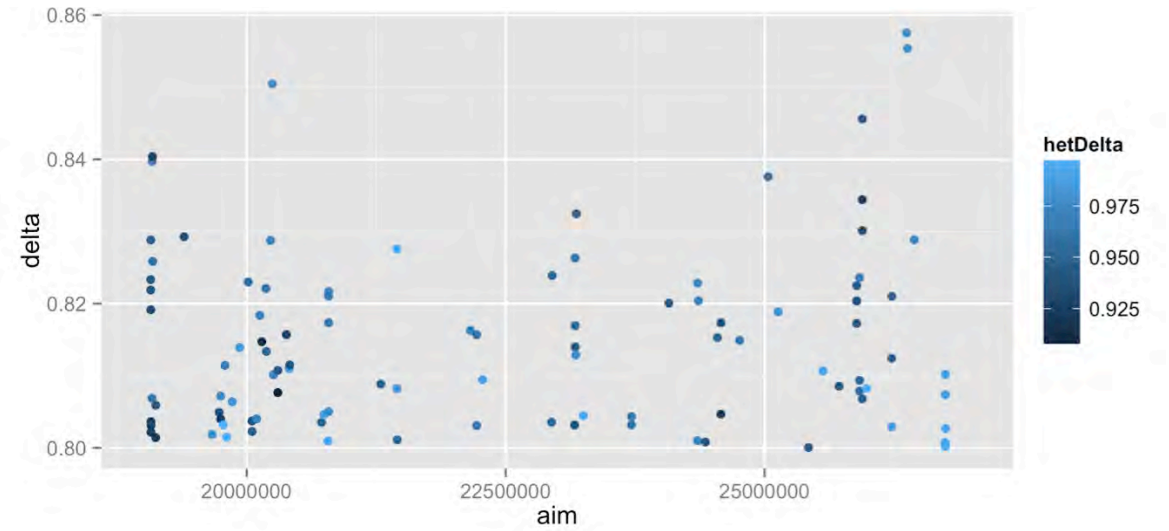
calls for 98% of samples with perfect accuracy (see table 4.2a); samples that we were unable to call were later filtered from the dataset due to low overall read coverage.

Figure 4.5 : SVC calls for the Ag1kG (release 1) set



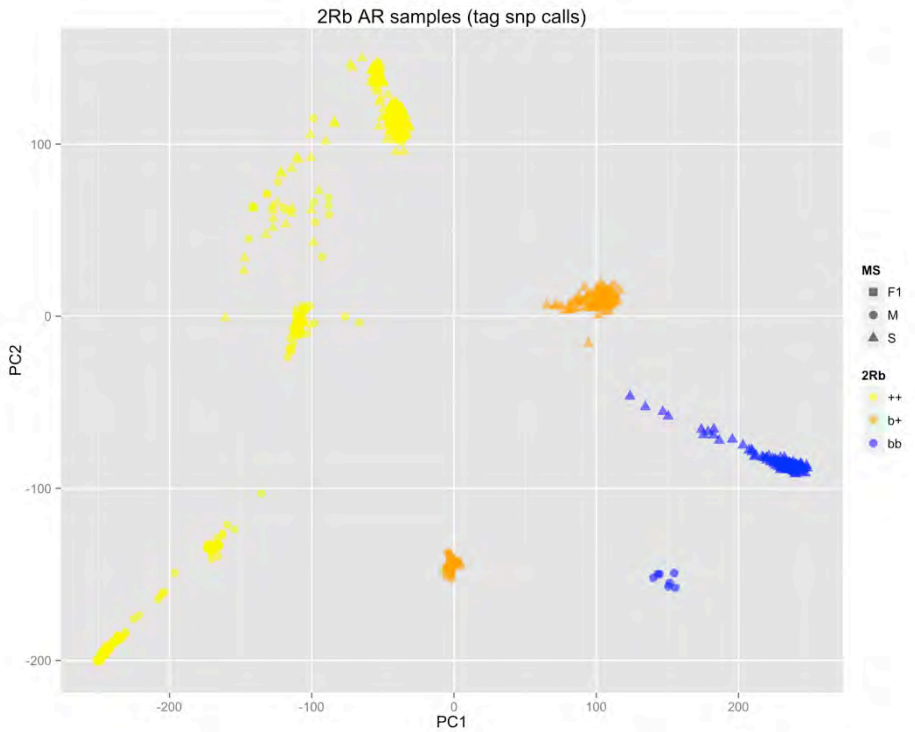
The PCA graph was created by singular value decomposition (using the 'prcomp' function in R). Each point on the graph represents one sample, and is coloured according to SVC-called inversion karyotype and shaped according to M/S molecular marker Low concentration samples have been removed from this set.

Figure 4.6: Tag SNP selection by Chakraborty and Weiss' method:



Marker selection by the Chakraborty and Weiss method. Markers are separated by their minor allele frequency differences between the two putative 2Rb / 2R+ homokaryotype sets. The delta statistic is the minor allele frequency difference between the two sets (only samples with a MAF difference greater than 0.8 were shown.
Marker colour is the hetDelta statistic, defining deviation of the putative heterokaryote sample from the MAF=0.5 expectation (blue / 1 = perfect admixture, black / 0 = no admixture signal).
The 20 markers with the highest delta statistic, and a hetDelta of > 0.8 were chosen as tag SNPs.

Figure 4.7a: 2Rb tag-SNP calls for the Ag1kG dataset (release 1)

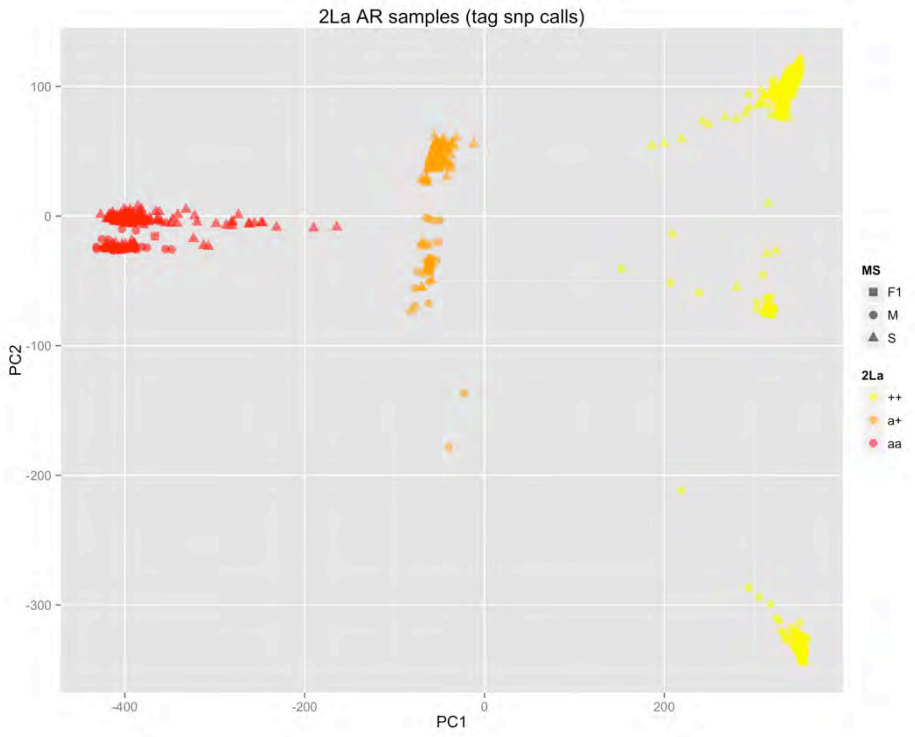


The PCA graph is identical to that shown in Fig 4.5. However points here are coloured by the tag-SNP-called inversion karyotype and shaped according to M/S molecular marker. Low concentration samples have been removed from this set.

2La tag-SNP selection

Despite the lower efficicacy of the 2La SVC karyotyping, refinement of these calls by the method of Chakraborty and Weiss was far more successful. The higher F_{ST} between 2L+ and 2La forms (that is the vastly reduced proportion of variants that segregate in both forms), meant that a larger number of markers were available to type this inversion.

Figure 4.7b: 2La tag-SNP calls for the Ag1kG dataset (release 1)



The PCA graph was created by singular value decomposition (using the 'prcomp' function in R). Each point on the graph represents one sample, and is coloured according to SVC-called inversion karyotype for 2La and shaped according to M/S molecular marker. Low concentration samples have been removed from this set.

Selection of a 20 SNP panel, and smaller subsets of these, were assessed as before. A far greater number of no-calls were present; 2La has a lower alignment efficiency, and of the SNPs that perfectly typed the inversion, none was able to call more than 82% of the samples (though each of them could do this without any erroneous calls). However panels of 3 markers were able to call 96.5% of samples with 100% accuracy according to PCA confirmation (see table 4.2b). Again, samples without calls were almost exclusively those with low read coverage.

Table 4.2: minimal typing barcodes for 2Rb and 2La inversions :

Table 4.2a : Typing barcodes for 2Rb

Tag SNP bp	Plex Size	Miscall rate	No call rate	Failure rate	Min Failure
19090656	1	0.0023	0.1247	0.1270	0.1143
25945079	1	0.0046	0.0831	0.0878	0.0751
26371581	1	0.0069	0.0485	0.0554	0.0427
19090656, 25945079	2	0.0069	0.0473	0.0543	0.0416
25945038, 25945079	2	0.0069	0.0716	0.0785	0.0658
19090656, 26371581	2	0.0092	0.0335	0.0427	0.0300
19074736, 25945079, 26371581	3	0.0000	0.0185	0.0185	0.0058
19074736, 25945079, 26380901	3	0.0000	0.0185	0.0185	0.0058
19074736, 21449883, 26371581	3	0.0000	0.0196	0.0196	0.0069
19074736, 25945079, 26371581, 26380901	4	0.0000	0.0150	0.0150	0.0023
19074736, 21449883, 25945079, 26371581	4	0.0000	0.0162	0.0162	0.0035
19074736, 25917705, 25945079, 26380901	4	0.0000	0.0162	0.0162	0.0035
19074736, 25033805, 25945079, 26371581, 26380901	5	0.0000	0.0139	0.0139	0.0012
19074736, 25917705, 25945079, 26371581, 26380901	5	0.0000	0.0139	0.0139	0.0012
19073606, 19074736, 19088508, 25945079, 26371581	5	0.0000	0.0150	0.0150	0.0023

Table 4.2b : Typing barcodes for 2La

Tag SNP bp	Plex Size	Miscall rate	No call rate	Failure rate	Min Failure
21368262	1	0.0000	0.1894	0.1894	0.1663
20816458	1	0.0000	0.3037	0.3037	0.2806
21257391	1	0.0000	0.5023	0.5023	0.4792
20816458, 21368262	2	0.0000	0.1155	0.1155	0.0924
21257391, 21368262	2	0.0000	0.1409	0.1409	0.1178
20816458, 21257391	2	0.0000	0.2113	0.2113	0.1882
21368262, 39965217, 41133006	3	0.0000	0.0346	0.0346	0.0115
21750923, 39965217, 41133006	3	0.0000	0.0358	0.0358	0.0127
21055567, 41133006, 41788771	3	0.0000	0.0404	0.0404	0.0173
21055567, 39965217, 41133006, 41524088	4	0.0000	0.0300	0.0300	0.0069
20816458, 39631457, 39965217, 41133006	4	0.0000	0.0312	0.0312	0.0081
21055567, 21368262, 39965217, 41133006	4	0.0000	0.0312	0.0312	0.0081
20816458, 39631457, 39965217, 41133006, 41524088	5	0.0000	0.0277	0.0277	0.0046
20727760, 39631457, 39965217, 41133006, 41524088	5	0.0000	0.0289	0.0289	0.0058
20816458, 21055567, 39965217, 41133006, 41524088	5	0.0000	0.0289	0.0289	0.0058

SVC validation with PCA-derived karyotypes

Concordance of tag-snp calls with larger demographic patterns was calculated by comparison to PCA-derived karyotypes. As detailed above, these are susceptible to false-positive results due to admixture, however we would not expect this admixture to generate the characteristic duplication in the breakpoint; nor conversely would be expect the breakpoint duplication to generate a broader signal of admixture in the dataset.
A high degree of correlation between these results is therefore strong evidence for the validity of the SVC/tag-SNP approach.
PCA-derived calls are calculated for those populations showing evidence of segregation of the inversion; it should be noted that this is not all of the populations that are indicated to segregate the inversion by SVC/TagSNP analysis – however it does include all of those populations seen to segregate the inversion at high frequencies.

Despite the difficulty of resolving copy numbers in the breakpoint region, SVC calls in general demonstrated strong correlation with the PCA-derived karyotypes: 2Rb SVC samples showed a (Pearson) correlation of 0.881, whilst 2La demonstrated a correlation of 0.975. Tag SNP comparisons, on the other hand, demonstrated perfect correlation (i.e. Pearson cor=1 in both datasets) with the PCA derived samples across all samples (n=455 / 559 for 2La / 2Rb respectively).

Figure 4.6 a/b: PCA-based karyotyping of the 2Rb / 2La inversions (following page)

PCA-based calls of 2La and 2Rb inversions. PCA plots are calculated by single value decomposition using the prcomp function in R. Classification of the karyotypes was calculated by examining the total variation from the PEST reference; inversions should have a surfeit of fixed differences from the homokaryotypic standard reference sequence. Countries in which the inversions do not segregate at high frequencies are not identified by this method.

Table 4.3 a/b: within-cluster sum-of-squares of PCA clusters, 2Rb (following page)

Within-cluster sum of squares results for all kmeans clusters within the country-specific PCA graphs (Figure 4.6 and 4.7). sum of squares are calculated for each of the coloured clusters and scaled to the total number of samples. Putative karyotypes are then assigned due to the number of non-pest variants (pest is homokaryotypic, hence heterokaryote and homokaryote inverted groups should be more highly diverged from the reference). Populations marked with an asterisk were chosen for correlation comparisons.

Figure 4.6 PCA-based karyotyping of the 2Rb inversion in the AG1kG :AR1 dataset

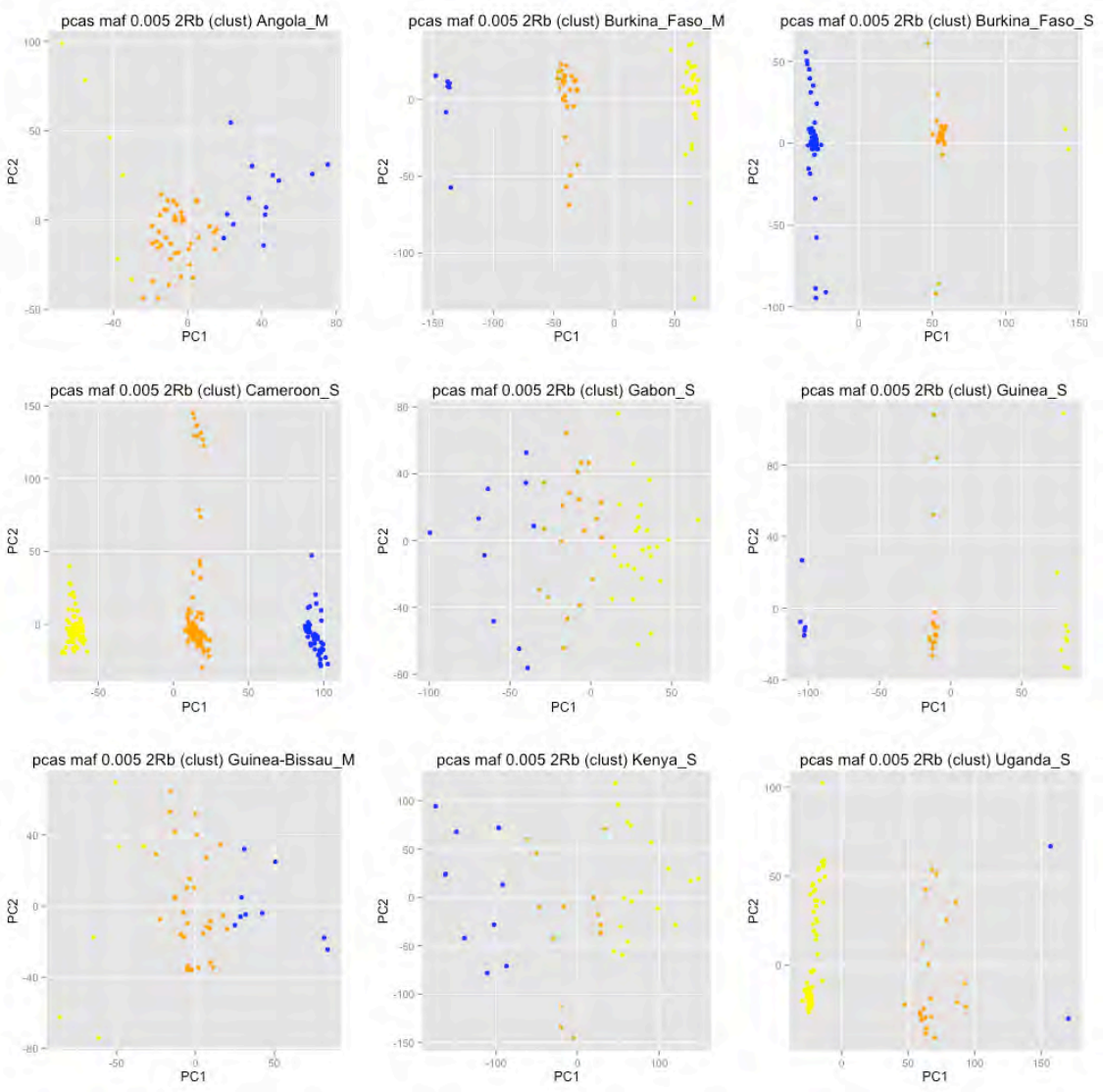


Table 4.3a : within-cluster sum-of-squares of PCA clusters, 2Rb

	++	b+	bb	sum wss	delta
Angola_M	95.56	268.24	168.69	532.49	-3.67
Burkina_Faso_M *	15.3	18.03	22.28	55.6	-1.91
Burkina_Faso_S *	6.78	5.51	1.12	13.4	-0.2
Cameroon_S *	11.18	16.76	11.46	39.4	0.82
Gabon_S	132.16	360.04	133.05	625.25	1.32
Guinea_S *	2.38	5.55	1.67	9.59	-0.04
Guinea-Bissau_M	272.05	125.14	463.64	860.83	4.24
Kenya_S	946.4	1062.55	977.72	2986.67	12.66
Uganda_S *	13.65	44.75	136.11	194.51	-1.87

Figure 4.8 PCA-based karyotyping of the 2Rb inversion in the AG1kG :AR1 dataset

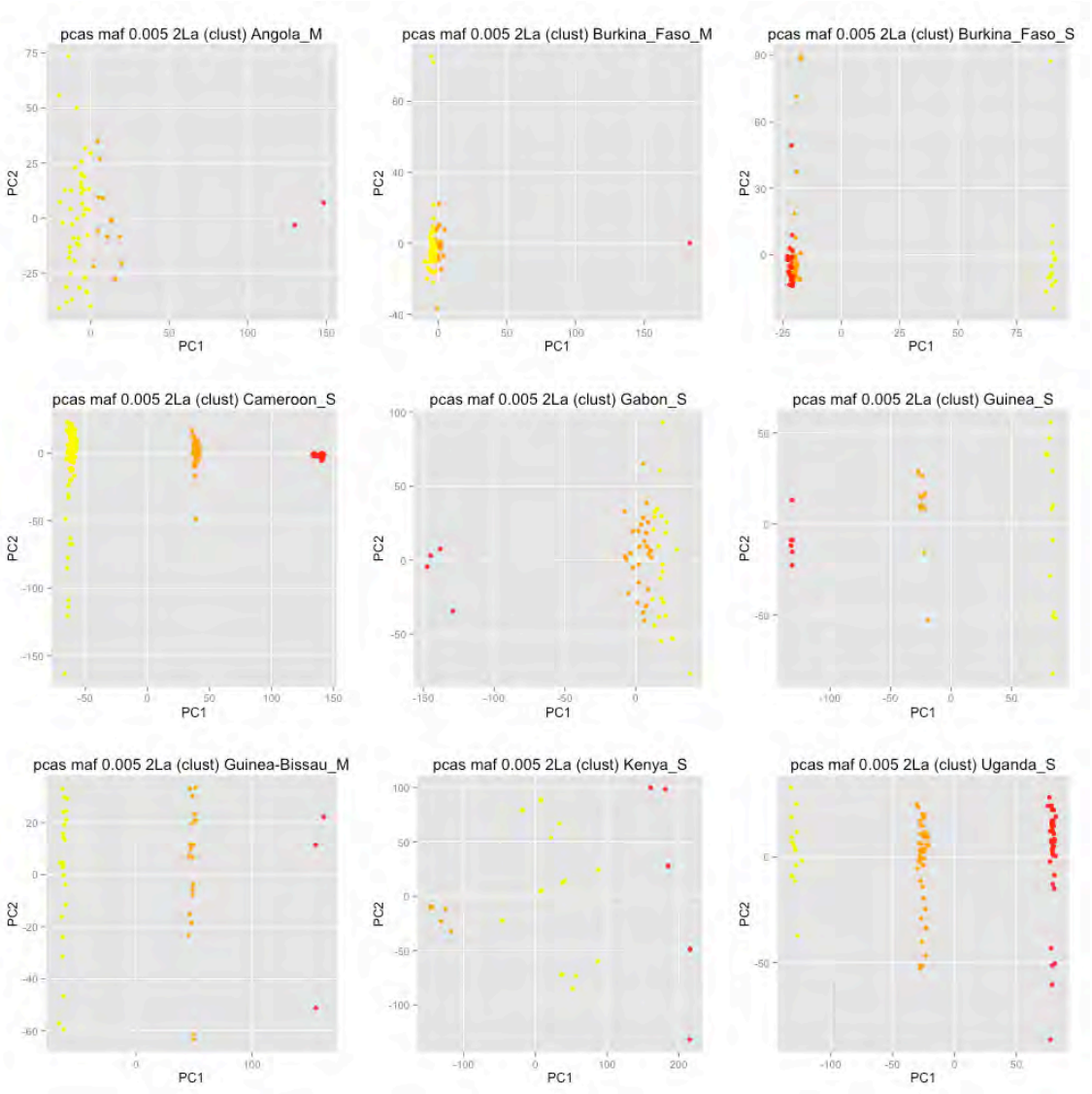


Table 4.3b : within-cluster sum-of-squares of PCA clusters, 2La

	++	a+	aa	sum wss	delta
Angola_M	32.78	30.52	85.52	148.82	-55.27
Burkina_Faso_M	0.00	2.54	3.01	5.55	-88.82
Burkina_Faso_S	0.52	1.07	0.77	2.36	-53.91
Cameroon_S *	3.66	2.01	3.97	9.65	-0.64
Gabon_S	31.58	49.37	33.85	114.79	63.84
Guinea_S *	0.24	5.34	4.50	10.08	0.44
Guinea-Bissau_M *	11.07	3.05	2.47	16.60	1.04
Kenya_S	84.85	399.74	1157.91	1642.50	-6.96
Uganda_S *	4.51	1.69	5.70	11.90	-0.84

Confirmation with polytene chromosome samples

Further validation was provided by comparison of the SVC calls to the subset of karyotyped samples available from Cameroon. Of 123 samples with determined karyotypes, 113 were correctly called for 2Rb (an error rate of 8.1%). There was no indication of any association between these ten errors and adjacent karyotypes, which might indicate the kind of location-specific errors found by Lobo *et al.* (Lobo et al. 2010). All ten were called as standard for the 2Rj, c and u inversions and only one possessing a 2Rd inversion.

A direct comparison of these microscopically determined karyotypes with the PCA graph for 2Rb indicated that all of them were misplaced on the graph – a strong indication of errors in the polytene chromosome analysis (see figure NN). In all cases the SVC-derived calls placed the sample in the correct PCA cluster see (figure 4.9a).

Comparison of the 2La calls to the previously karyotyped samples was similar to the 2Rb calls. Of 123 samples with 2La karyotypes, 120 were in concordance with the SVC-derived call (error rate 2.4%). Comparisons of the discordant calls to the PCA results suggested the SVC-derived calls to be the correct ones (see figure 4.9b).

Figure 4.9 distribution of discordant cytotyped samples within Ag1kG dataset (Cameroon samples only):

Figure 4.9a :

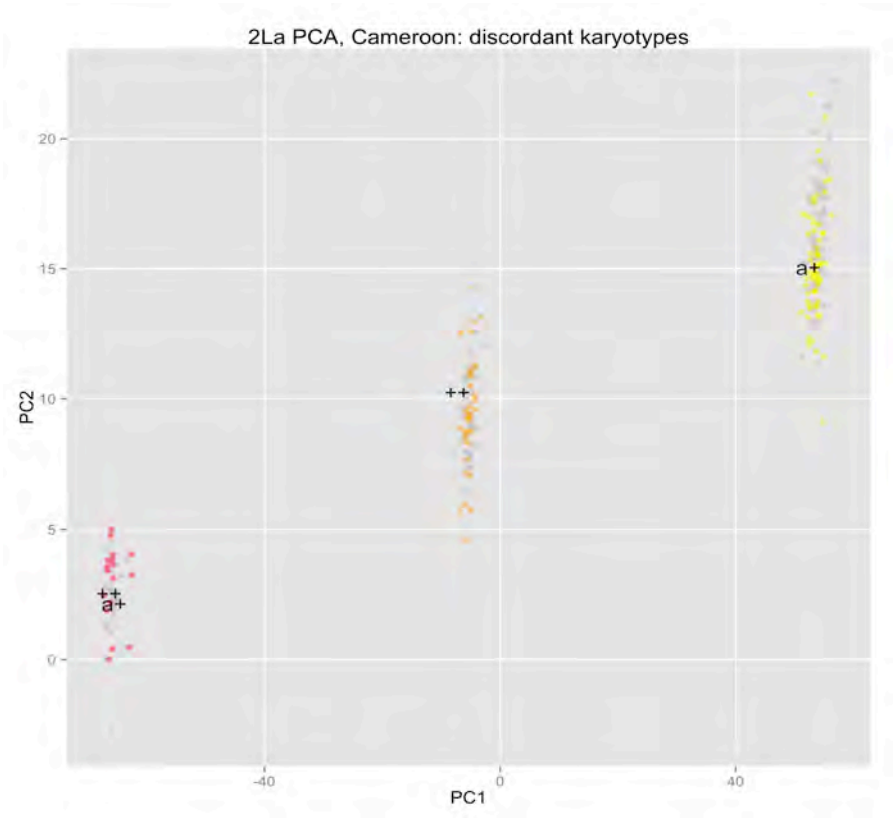
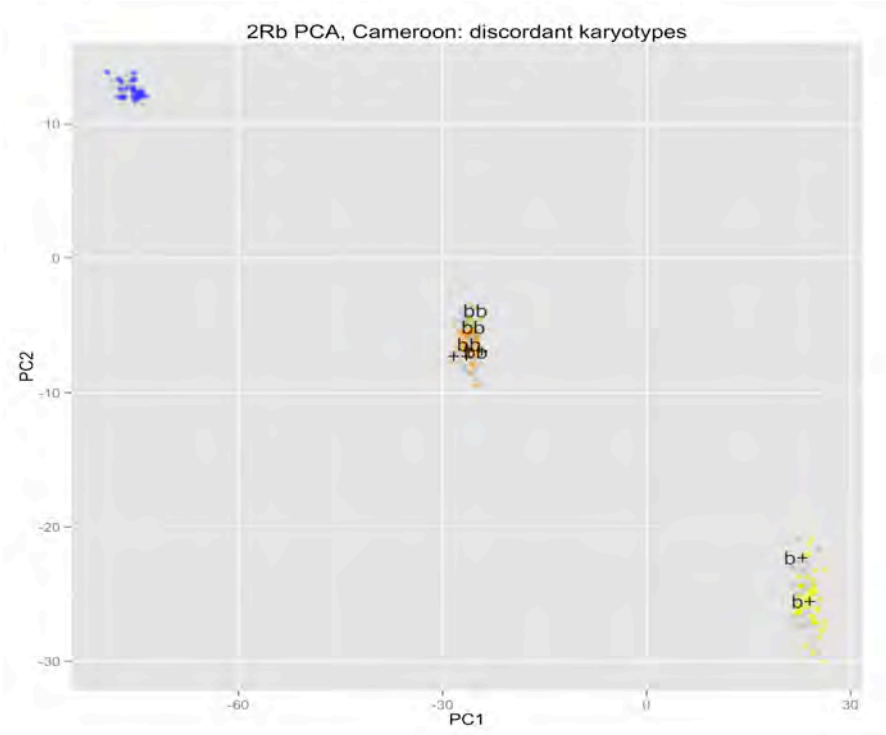


Figure 4.9b:



Discordant samples, those in which the karyotypes called via polytene chromosome analysis did not agree with those called via tag-SNP, are plotted on a whole-dataset PCA of the karyotype regions for the a) 2La and b) 2Rb inversions. Concordant samples are shown as points, coloured by their tag-SNP derived karyotype, Discordant samples are shown as text, labelled by their karyotype as determined by polytene chromosome analysis.
The siting of these discordant samples amongst the 'wrong' karyotype group may suggest miscalling during polytene chromosome analysis or during collation of sample metadata.

Table 4.4: concordance table of samples typed by all methods :

Table 4.4a: 2Rb concordance table:

	cyto	svc	tag	pca
cyto	1	0.842	0.952	0.952
svc	0.842	1	0.895	0.895
tag	0.952	0.895	1	1
pca	0.952	0.895	1	1

Table 4.4b: 2La concordance table:

	cyto	svc	tag	pca
cyto	1	0.924	0.947	0.947
svc	0.924	1	0.978	0.978
tag	0.947	0.978	1	1
pca	0.947	0.978	1	1

4.10 Results II: Inversion distribution

The results presented in this section relate to the distribution of 2La and 2Rb karyotypes within the Ag1kG dataset. These demonstrate, in different subsets of the dataset, reproductive isolation between sympatric populations of different species (Burkina Faso) and apparent parapatric segregation of the same species (Cameroon).
Both of these issues will be investigated more fully in following publications. As a result this section should be seen as a preliminary overview demonstrating the potential utililty of the karyotyped dataset, rather than an in-depth investigation of these factors.

Inversion Distributions

Distributions of inversion karyotypes were heterogeneous, higher frequencies of segregating inversions were seen in Central and West Africa than in East or Far-West Africa (see figure 4.8). Calculations of Hardy-Weinberg proportions indicated, for the majority of countries where panmictic populations were expected, that inversions were in equilibrium.
Of the samples where *gambiae* / *coluzzii* are in sympatry, the sample from central Africa (Burkina Faso) shows inversion karyotypes that are in Hardy-Weinberg equilibrium when *gambiae* / *coluzzii* are considered separately, yet with very different distributions in each species – indicating strong reproductive isolation of *gambiae* / *coluzzii* in this region. In contrast, the 'far west' samples, from Guinea-Bissau are marked by a lack of similar separation between M and S forms, indicating high levels of hybridisation in this region. Results from Cameroon, consisting solely of *gambiae s.s.* (S-form) samples were the only samples not in HWE, showing a clear lack of heterokaryotes and apparent within-form population structure in this region.

Monophyly of 2Rb / 2La

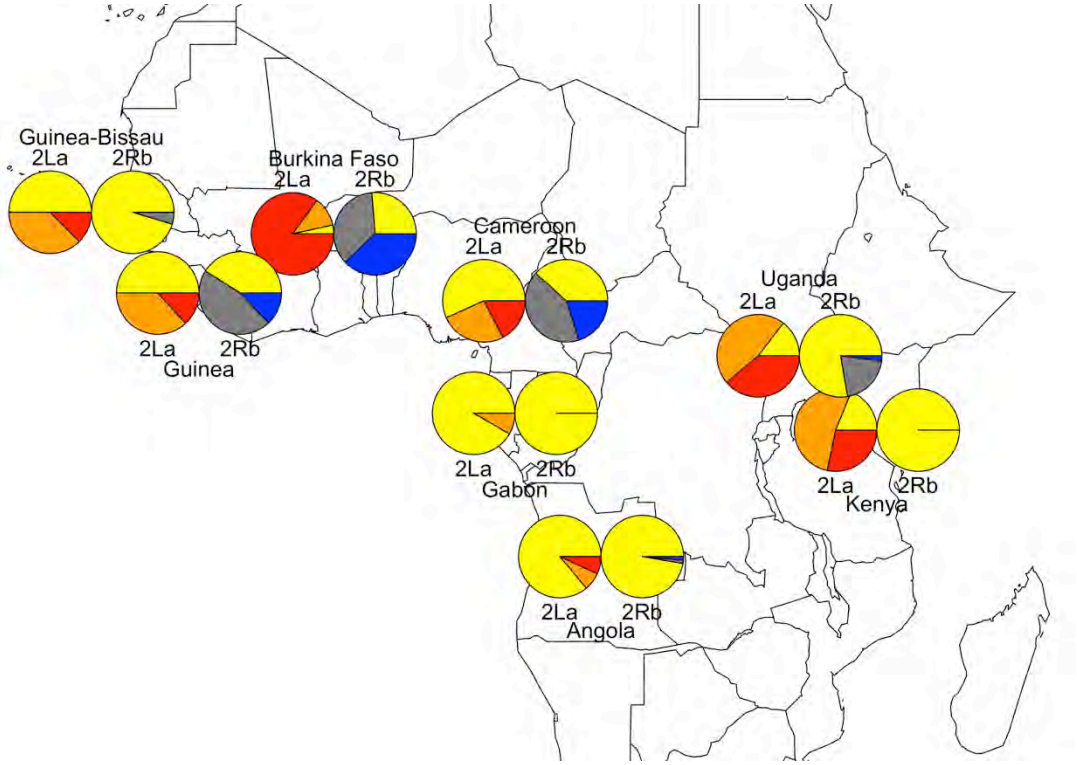
It is important to note that although the CNV loci showed better fidelity in the distal rather than proximal breakpoints, the tag SNPs show no such bias. An ancestral haplotype can be seen that extends across both breakpoints, providing strong support for the monophyly of both these inversions. This supports the results of Sharakhov *et al.* (Sharakhov et al. 2006) in which they established the single origin of the 2La inversion by comparison of inversion breakpoints between *An. gambiae* and *An. arabiensis.* The monophyly of 2Rb has not previously been established.

Table 4.5: Deviation of inversion frequencies from Hardy-Weinberg equilibrium

	2La			HWE	2Rb			HWE
country	+/+	+/a	a/a	p-val	+/+	+/b	b/b	p-val
Angola	62	5	5	1.00E-04	71	1	0	1.00
Burkina Faso	6	19	148	1.30E-03	47	59	67	2.00E-04
Cameroon	160	73	49	1.00E-04	109	116	57	1.22E-02
Gabon	55	5	0	1.00E+00	60	0	0	1
Guinea	19	14	5	4.63E-01	16	18	4	1.00
Guinea-Bissau	32	22	8	2.34E-01	59	0	3	1.00E-04
Kenya	10	28	15	7.89E-01	53	0	0	1
Uganda	15	48	40	1.00E+00	80	21	2	6.34E-01

Significant deviation from Hardy-Weinberg equilibrium (HWE) are seen within a number of populations, in particular the Burkina-Faso and Cameroon samples segregate both inversions at high frequencies deviate from HWE. Further dissection of these two datasets indicates that reproductive isolation within Burkina-Faso is subdivided along gambiae / coluzzii lines, whereas Cameroonian samples demonstrate population structure within Anopheles gambiae alone.

Figure 4.10: Geographical distributions of 2Rb and 2La inversions in the Ag1kG (release 1) dataset



Genome-wide linkage to inversions

Long-range LD was examined by calculating r^2 between the 20 marker SNPs for 2Rb and the rest of the genome using the snpStats package (version 1.10.0), part of the Bioconductor suite, and custom R scripts.

These long-range LD scans were performed in the four countries in which 2Rb segregated to a high degree; Burkina Faso, Cameroon, Guinea and Uganda. Linkage patterns differed between countries. Strong linkage was seen between 2Rb and 2La within Cameroon and Guinea, whereas 2Rb was linked to centromeric regions (with particularly strong signals for chromosome 3 in Burkina Faso. Samples from Uganda showed no significant LD to regions outside of the 2Rb inversion itself.

Figure 4.11: long range LD to 2Rb markers across the genome

Long range LD was calculated for all countries in which the 2Rb locus segregated (Burkina Faso, Cameroon, Guinea and Uganda. The graphs show linkage to chromosomes 2,3 and X respectively (centromeres are marked by a dashed line in chromosomes 2 and 3, and absent in the telocentric X-chromosome).

Different patterns of linkage are seen to genetically distant locations depending upon the situation:

Cameroon (green), in which population segregates between ecotypes, shows strong linkage to the 2La locus; Burkina Faso (red), where the segregation is along gambiae / coluzzii lines shows strong linkage to the centromeric speciation islands but none to the 2La locus (even though it does segregate in this dataset).

4.11 Discussion

Discussion I: validity and transferrability of SVC / tag-SNP typing

Following the approach of Corbett-Detig *et al.* we have karyotyped the samples using a two-fold approach with one method identifying the physical breakpoint itself, and the other detecting the demographic effects of those inversions. Whilst we have used demographic effects as independent corroboration of the breakpoint identification (as did Corbett-Detig *et al.*) it is worth considering the degree of independence of these measures.

The use of PCA data to corroborate the breakpoint relies on the independence of these two measures – that the presence of the breakpoint will not inherently lead to the generation of an admixture signal within the PCA. Here the great size of these inversions, ~21Mb for 2La and 7mb for 2Rb (Sharakhova et al. 2011) greatly reduce the likelihood that a 0.5kb locus could dominate a PCA that is derived from 347,000 and 434,000 SNP markers respectively. Similarly, the previously identified location of the CNV duplications within the area of suppressed recombination within each breakpoint reduces greatly the likelihood of loss of this feature as a marker for the inversion.

Indeed, perhaps the only potential source of loss of the CNV marker for the inversion is either a very specific (and highly improbable) deletion within the breakpoint, or the re-use of this breakpoint for a subsequent inversion – which would leave the duplication intact, but would revert the inversion to it's standard form. In their attempts to devise a molecular karyotyping test for the 2Rb inversion, Lobo *et al.* encountered an increased number of failures for 2R+ karyotypes where they also demonstrated a 2Rcu karyotype – perhaps indicating a re-inversion of the 2Rb karyotype to 2R+'. Sadly a lack of known 2Rcu variants does not allow us to test this theory extensively, however the development of typing methods for the 2Rc and 2Ru inversions in coming years may uncover additional 2Rcu samples within this dataset.

The transferability of this method to other datasets is extremely poor. SVC calls are only viable in a quantile normalised dataset that is directly comparable to the training set (as was

the case for the colony crosses and Ag1kG dataset), consequently this method cannot easily be applied to other data without normalisation of that set to the same profile.
The approach was worth taking only due to the size and importance of the Ag1kG dataset, and should therefore be seen as something of a one-off. However the minimal set of typing SNPs shown in table 4.4 are viable candidates for the development of molecular typing assays. Potential choices for this approach would be to devise a RFLP from one or more of the loci, which would provide a low-cost method of assaying the inversion that could be used in challenging environments, such as in the field. However whilst this approach would be viable for the 2La inversion (that already has a reliable molecular karyotyping method), the lack of one clear candidate for the 2Rb inversion in fact means that a multi-SNP approach would be preferable.
The ability to make entirely unambiguous calls with only three SNPs opens up the possibility of developing cheaper genotyping methods to be used in the field. Should typing panels of these sizes be maintained for the other five inversions we could envisage ascertaining each chromosomal form with 2-25 markers – sufficiently small to fit into a single sequenom 'plex' – enabling extremely cost-effective karyotyping. One could envisage the development of such an approach in the near future, following the identification of typing SNPS for the other inversions. Desirable as this would be, it is currently beyond the scope of this project, particularly as it would require a dataset of independent karyotyped samples including extensive representation of 2Rcu samples from Mali, the source of Lobo *et al.*'s misdiagnoses. This is, at the moment, a problematic area to sample from.
PCA methods themselves are somewhat unreliable. Any measure that relies on the detection of admixture between karyotype calls will be susceptible to false-positives in the presence of genuinely admixed populations. Perhaps more problematically, such methods will not be viable unless all three karyotypes are represented at high frequencies.
This would appear to have been the case in at least three of our populations, in Gabon, Angola and Burkina Faso – where 2La frequencies were too low to dominate the first principal component, as well as in Kenya, where the diversity of this dataset increased the within-sample sum of squares in each karyotype cluster.
Nevertheless, even with these omissions, there are 455 and 559 samples respectively in which the Tag SNPs and PCA are in perfect concordance, providing strong support for the validity of the breakpoint identifications.

Discussion II: geographical distribution of the 2La and 2Rb inversions

The availability of robust inversion karyotypes for these two inversions over a dataset of this geographic spread is unprecedented, and allows us to examine some of the features of this

distribution. Whilst a detailed investigation of these effects is ongoing and is beyond the scope of this thesis, broad patterns can be discerned within the data.

Running somewhat contrary to previous assertions of minimal LD in these species, linkage was clearly detectable between the inversions and distant genomic loci, although the linkage pattern varied significantly depending on the context.

Examinations of Hardy-Weinberg equilibrium are illustrative of some of the known factors of population structure on macrogeographic scales. The existence of differing proportions of 2Rb and 2La frequencies in *gambiae* / *coluzzii* samples in Burkina Faso supports strong reproductive isolation in this region, in contrast, the 'far west' samples, from Guinea-Bissau are marked by a lack of similar separation between the two species, indicating high levels of hybridisation in this region, consistent with the results of Caputo *et al.* (Caputo et al. 2011). Results from Cameroon indicate reproductive isolation between chromosomal forms related to sampling site: further investigations will be required to establish the ecological differences between these sites that may underlie this parapatric isolation.

These results are entirely consistent with the results of the long-range LD scans: regions such as Cameroon, where inversion frequencies indicates an ecotypic split, demonstrate linkage between potential markers of ecotypification; whereas in regions such as Burkina Faso, where frequencies suggest a division along species boundaries, linkage is strongest with putative speciation islands.

Although firm conclusions are difficult to drawn from this uneven sampling set, it is nevertheless interesting to note that, of the eight country datasets represented in the samples, it is the two regions with the highest proportions of segregating inversions that show the strongest signals of reproductive isolation. Further dissection of this dataset, and the development of markers for the other major chromosomal inversions, will surely elucidate more about the relationships between structural variants and reproductive isolation in this species pair.

Acknowledgements:

The work represented in this chapter has been principally carried out by the author, but is reliant upon the work of numerous others, in particular in the use of the data from the Anopheles 1000 genomes project and the Anopheles colony-crosses project.
The colony crosses dataset was conceived in collaboration between the Wellcome Trust Sanger Institute (WTSI), Wellcome Trust Center for Human Genomics (WTCHG) and Liverpool School of Tropical Medicine (LSTM). Crossing of mosquito lines and rearing of the F1 offspring was carried out at LSTM; Sequencing of mosquito DNA was carried out at WTSI; genotyping and basic analysis was carried out at WTCHG. Particular thanks must go to Tiago Antão and Alistair Miles for the genotype calling and phasing of F1 samples upon which the initial karyotype assessments rely.
Manual calling of karyotypes within the crosses dataset, sample normalisation, development and validation of the SVC calling algorithm were carried out by the author. SVC calling of AG1kG data, Tag SNP identification, Validation of tag-SNP calls, and analysis of inversion distribution and panmixia were also carried out by the author.
This work received financial support from the Wellcome Trust, and the Pasteur / Paris University PhD Program.

5 : Genome Wide Association of Anti-Plasmodium Immune Factors in *An Coluzzii*:

5.1 Introduction

The techniques of genome wide association studies are now widely used in a variety of species. In malaria work they have already been successfully used to map immune factors within the human genome (Dunstan et al. 2012; Jallow et al. 2009) and drug resistance factors for *Plasmodium* genes (Van Tyne et al. 2011). However, although the selection of refractory lines from a phenotypically heterogeneous colony indicates that this variation has a strong genetic component (Collins et al. 1986), until now the mosquito has proven resistant to these approaches.

Much of the reason for this involves the inherent intractability of the mosquito as a system; put simply, it is more difficult to infect a suitably large cohort of mosquitoes with *Plasmodium* than it is to grow a colony of the parasite in vitro, or to find enough malaria-infected humans – particularly when passive case detection is used, as was the case for both Dunston *et al.* and Jallow *et al.* Indeed *Plasmodium* virulence factors in the mosquito host are comparably difficult to work with and have so far gone unexamined. However, in addition to the usual complexities of a two-genome system, the mosquito genomics is also unfavourable to genomic mapping, much of this is connected to the highly polygenic nature of the immune response (as seen in chapter 2) and the complexity of the population structure (as seen in chapter 3).

Requirements for GWAS experiments

In any organism, successful GWAS is reliant on a combination of linkage, diversity, and allele penetrance. Linkage disequilibrium (LD) is particularly important as a characteristic. A locus in high LD with it's neighbouring variants (i.e. non-random co-incidence) will present a greater number of potential 'tag' SNPs, increasing our ability to detect the locus; at the same time, since haplotypes are longer, the number of independent tests is reduced and lower *P*-values are required to establish association. Contrasting extremes of LD are also problematic. While high LD ensures the ability to detect is maximised, LD that is too high will hinder the ability to isolate the causative locus within a long haplotype. Few populations in any species will present an ideal range of LD to maximise both detectability and resolution.

High levels of diversity will similarly decrease our ability to detect loci. This is partly as a result of a reduction in LD (in low-diversity *spp*. a recombination between two largely identical chromosomes will not reduce haplotype length) but also by maintaining a broader range of allele frequencies. An excess of low-frequency alleles in particular will act to reduce LD (since they cannot be in strong linkage with more common variants) and will increase the number of individual tests being carried out.
The final criterion, penetrance, relates to the likelihood that an individual carrying the allele of interest has the phenotype being mapped – an ideal relationship would be deterministic, such that an allele always and only occurred in the presence of the phenotype. Optimum statistical power is achieved when LD is highest (i.e. $r^2 = 1$), the frequency difference between alleles is near 0, and the condition is related to a single allele of strong effect.

5.2: Challenges of genomic mapping in *An. gambiae*

I: High Diversity / Low LD

It is increasingly apparent that *Plasmodium falciparum* infection in *Anopheles gambiae* is a particularly intractable subject for association studies; due both to the vector's genome and the dynamics of the infection. Despite recombination rates that are not significantly higher than those found in other eukaryotes (Pombi et al. 2006), levels of LD in *A. gambiae* are negligible, with r^2 dropping to around 0.05 for variants within 1kb (Neafsey et al. 2010). This is similar to levels of linkage found in the *P. falciparum* genome (where r^2 at 1kb is also under 0.1) (Manske et al. 2012) and a good deal less than found in humans, where mean r^2 is maintained well above this level for half a megabase or more and SNPs within 1kb are frequently in perfect linkage (Shifman 2003). Diversity in the mosquito is also at least an order of magnitude greater than that found in human (Wilding et al. 2009); preliminary studies in the *Ag* 1000 genomes project indicate that 72% of variants have an allele frequency of under 0.5% (Ag1000G Consortium n.d.). Gathering sufficient samples to assay alleles at 0.5% frequency would be next to impossible, however even for the more frequent alleles, these effects combine to raise the statistical bar to proof and make the development of effective tag-SNP panels close to impossible.

II: Low penetrance / multigenicity:

The diversity of the immune response, and the apparent ability of multiple pathways to show anti-plasmodial actions also complicates mapping efforts. Far from the case where a phenotype is linked to a single allele of strong effect, the mosquito apparently has a variety of intricately balanced recognition and effect mechanisms (see chapter 2). *Plasmodium*

immunity is a highly multigenic trait. Moreover it seems that, save for a few isolated cases such as TEP1R/S, most alleles will be of moderate to low effect, due to the countervailing selection pressures acting on the mosquito immune system. For instance, activation of immune pathways can incur a significant metabolic cost on the mosquito (Hurd et al. 2005; Garver et al. 2009) imposing a pressure for moderation in the immune response. Balancing selection is therefore likely to predominate in many loci. Whilst a small number of strong-effect alleles can and do segregate in nature (White et al. 2011), allele penetrance of the majority of weaker-effect alleles is likely to be low. Identifying these genes – particularly in populations where a strong effect allele segregates – is likely to be challenging.
Indeed, the broad spectrum of genes that have been associated so far – including signal transduction (Blandin et al. 2009; Harris et al. 2010), recognition (Harris et al. 2010; Li et al. 2013), and effector (Harris et al. 2010) molecules – also points to a highly multifactorial phenotype, and suggests that a high number of rare variants actually contribute to the variation in phenotype. This is in apparent disagreement with the common-disease/common-variant hypothesis (Iyengar & Elston 2007), which suggests, for a disease with up to 40% prevalence in the wild, a small number of common polymorphisms will underlie the variation in resistance.
However, for a multiplicity of weak alleles, particularly when combined with a lack of significant haplotype structure, we would expect to make successful associations only where a large number of polymorphic immune genes were clustered in a single genomic locus, such as the PRI (Riehle et al. 2006).

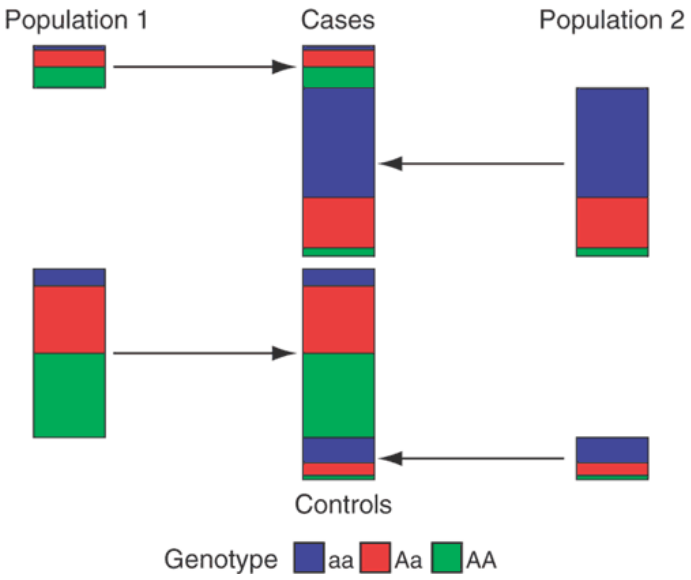
III: Phenotype tractability

Even where alleles of strong effect are present, the dynamics of *P.falciparum* infections themselves are unhelpful. In the wild, both infection prevalence and intensity are low and *Plasmodium* oocyst numbers in wild populations typically show a binomial distribution with a mode at or near zero (Tripet et al. 2008). Many laboratory colonies give more statistically amenable levels of infection, enabling statistically robust comparisons to be made. Yet, even in inbred colonies, infection rates vary greatly both at the individual and population level, based on genetic factors as well as stochastic factors such as bloodmeal size. As a result nominally ‘susceptible’ mosquitos will frequently be parasite free.
More problematically for a genetic association study, these colonies are likely to have lost many alleles important for immune function and may not be an accurate model for resistance in the wild. It is for these reasons that work in the insect vector has previously been restricted to either lower resolution or targeted methods.

IV: Population structure

It is not just the excess of low frequency variants that is problematic. The highly structured populations that typify this mosquito are also challenging. The lack of allele sharing means that attempts to replicate results using variants on one region (even if they are locally common) will frequently fail even in nearby locales. However, it is also well known from human studies that population structure can cause both type 1 and type 2 errors if divergent sub-populations are sampled unevenly (Marchini et al. 2004) (see fig 5.1)

Figure 5.1: The effects of population structure at a SNP locus



If two cryptic and reproductively isolated subpopulations differ in their frequencies of an allele, and if the proportions of cases and controls also differs across these subpopulations, then, despite there being no difference in the proportion of alleles within each population, a comparison across both will show a false association with the phenotype.
Source: Marchini et al.: The effects of human population structure on large genetic association studies. (Marchini et al. 2004)

This is a particular issue for *Anopheles gambiae* / *coluzzii* not just because of the major speciation event between the two species, but the apparent presence of reproductive isolation within those species at all scales of sampling. Indeed, cryptic populations have been shown even within single sampling sites; Riehle *et al.* have shown shared larval habitats that support two or more distinct populations, one of which, 'Goundry', exhibits reduced immunity to *Plasmodium* (Riehle et al. 2011). Clearly any phenotypic study that captured a mixture of these populations would suffer disproportionately from erroneous results due to the higher representation of markers linked to the Goundry form in the 'intense-infection' category (even though these markers may not actually be phenotypically linked). The presence of reproductively and phenotypically divergent populations even in the same sampling locations required either post-sampling statistical control when conducting an association study, or the

separation of the complex population into true panmictic groups as much as this can possibly be achieved. As we have seen in chapters 3 and 4, whilst the relationships between chromosomal forms are somewhat fluid, chromosomal inversions remain the most tractable markers of population structure in *Anopheles gambie s.l.*.

5.3 Previous mapping attempts:

In *Anopheles gambiae*

Due to these difficulties, previous mapping attempts in *Anopheles* have sought to work with a controlled sampling population and a limited set of markers – either by dealing with controlled crosses so that a QTL approach can be used, or by pre-selecting potential immune targets in an association study.

Riehle *et al.* used a QTL approach to identify the *Plasmodium resistance island* (PRI) (Riehle et al. 2006) – a region of co-localised immune genes containing a number of then-novel immune genes (including a large cluster of LRR genes containing, amongst others, the APL1A-C family). The PRI region was found to explain 89% of the variation in resistance to *Plasmodium* in this cross, however at 15MB it contains over 900 genes and mapping causative genes within this locus remains a challenge, not least because the PRI spans the breakpoint of the 2La inversion reducing mapping resolution due to the inversion-related increase in LD.

Limited-target approaches have been applied to the mosquito in two separate studies. Both Harris *et al.* (Harris et al. 2010) and Horton *et al.* (Horton et al. 2010) began association studies from a curated set of genes identified as having an immune role.

Finally Li *et al*. attempted to refine the PRI; using blocks of co-expressed genes to identify genes within the locus with immune effect (Li et al. 2013). Whilst all of these studies have successfully identified variants that altered the vector competence of the mosquito, none of the approaches would have either the generality or the resolution to detect a completely novel individual gene

These two methods have contrasting advantages and disadvantages. QTL mapping is unbiased, able to identify any phenotype-linked region of importance. However its ability to localise regions is poor: typically performed by crosses of inbred lines, it relies on recombination to break up those markers that do segregate in order to resolve smaller loci. In addition, since it normally requires genetically restricted lines it cannot sample a large amount of natural variation.

Target-limited steps, in comparison, are often able to identify individual genes; indeed markers will typically be chosen to code for specific genes. However the reduction in targets

is typically by manual curation of (known or putative immune) genes, meaning it has no power to detect novel associations. Whilst it may have some of the advantages of an association study, it could not be considered 'genome-wide'

Controlled-diversity mapping in other species

Whilst the diversity of *Anopheles* is somewhat extreme, other organisms have encountered comparable problems; often at the opposite end of the scale (requiring an increase in diversity rather than a reduction). Controlled-diversity populations have proven to be a valuable resource for mapping in model organisms, where resources such as the Jackson laboratory's diversity outbred (DO) line (Valdar et al. 2006) and the *Drosophila* Synthetic Population Resource (King, Macdonald, et al. 2012) (DSPR) have enabled fine-resolution mapping to be performed on a background of median diversity (King, Merkes, et al. 2012; Svenson et al. 2012). Both the DO and DSPR are derived from advanced inter-crosses of inbred lines (many with prior associated phenotypic traits) with breeding controlled to prevent excessive genetic drift.

5.4 Founder colony mapping

The challenge in *Anopheles* is similar to the diversity management seen in the DO / DSPR projects, yet the context is very different. The challenge in model organisms is to map a wide range of phenotypes within a series of highly controlled lines, where we can assume the required phenotypic diversity is present. In comparison, malariologists and vector biologists instead wish to investigate a smaller range of phenotypes (principally immune and behavioural), where it is desirable to capture natural variation.

To this end, we have developed a series of diversity-controlled *Anopheles* colonies to be used for association studies. These are intended to provide a workable intermediate between wild mosquitoes and lab colonies; possessing sufficient linkage that associations can be detected, and a representative degree of wild variation, but with sufficient control of population strata so that type 2 errors due to population structure are avoided. Rather than a cross of inbred lines these colonies have been generated using a limited number of founder females, from geographically limited collections with control for known markers of population structure.

This colonisation has two major effects. By founding the colonies from 10-20 individuals, we reduce the level of diversity to a manageable level, whilst also raising the minimum minor allele frequency to 0.05-0.025. This allows us to use a lower sequencing coverage than

would be necessary in wild colonies and still achieve an acceptable level of accuracy in our variant detection.

The second effect of this colonisation is to increase linkage between alleles on the same parental chromosomes – to $R^2 = 1$ in many cases. Effectively we create a mapping population with a large number of SNPs that perfectly type their haplotypes, boosting our ability to detect linked haplotypes and vastly reducing the number of individual tests that are performed. The combined effect of these two factors enables us to query low-frequency variants in a way that would be impossible in the wild.

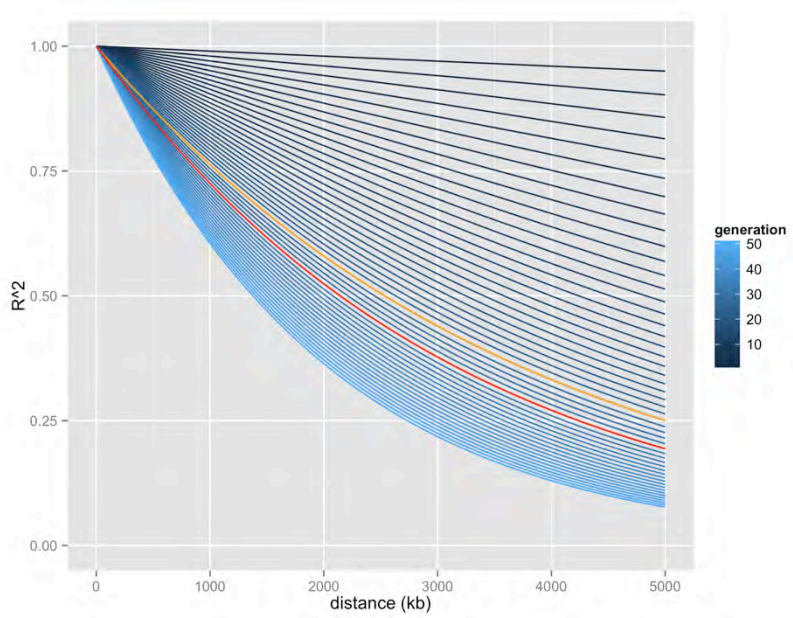
These colonies are then freely intercrossed (mated within the colonies without any specific mating scheme) resulting in a predictable loss of LD with each generation as founding haplotypes are broken up. Experiments are performed when an intermediate level of LD is predicted to be present, around 30 generations (see figure 5.2).

At the samples generations the theoretical maximum coverage of each marker is predicted to be 821kb for generation 27 and 693kb for generation 30. This is calculated for markers that were in perfect linkage at the time of founding and using $r^2 = 0.8$ as a determinant of marker coverage - as is typically used for human GWAS studies (Hoffmann et al. 2011). It should be noted that the true coverage is impossible to determine using this dataset and the applicability of human criteria to anopheline research could be a matter for some debate.

We used these ‘founder’ colonies in a two-stage GWAS study of infection with *Plasmodium*. In stage one, phenotypically divergent individuals are selected against a haplotypic background resulting from founding effects. Broad regions containing the causative variants are identifiable in phenotype pools due to an overall reduction in diversity and a shift in allele/haplotype frequencies across pools in those regions of the genome in association with the phenotype. For regions of the genome not associated with the phenotype, segregation of haplotypes would be random across pools. In stage two, individual genotypes are used for fine mapping and for replication of the association. As a distinct advantage over GWAS in wild populations, we are able to functionally characterise the loci against the same genetic background before attempting to type the associated variants in a wild population.

Figure 5.2: Decay of LD from the founding event

Fig 5.2a: theoretical decay of perfect linkage (r^2=1) from founding



Theoretical decay of linkage from the founding event for a pair of markers in perfect linkage. As markers are broken up by recombination Linkage decay is shown from the top (black) line for those in the first generation after founding, to the light blue 50th generation. Linkage decays rapidly for those that are separated by more than 2-3mb, but is maintained at shorter distances. Yellow and red lines mark the generations at which founder 3 (Mali – generation 27) and founder 9 (Burkina Faso – generation 30) were infected. It should be noted that the magnitude of the reductions in LD decrease with successive generations.

Fig 5.2b: comparison of theoretical decay to actual LD at generations 27 and 30

Theoretical decay is shown here for markers between D'=1 and D'=0.5, along with comparisions to D' as calculated from actual microsatellite markers in the two founder lines (D' is used since r^2 .cannot be calculated for multiallelic markers such as microsatellites). While D' decays with distance as expected there is a clear excess of high-linkage samples. This is probably indicative if significant inbreeding within the colonies.

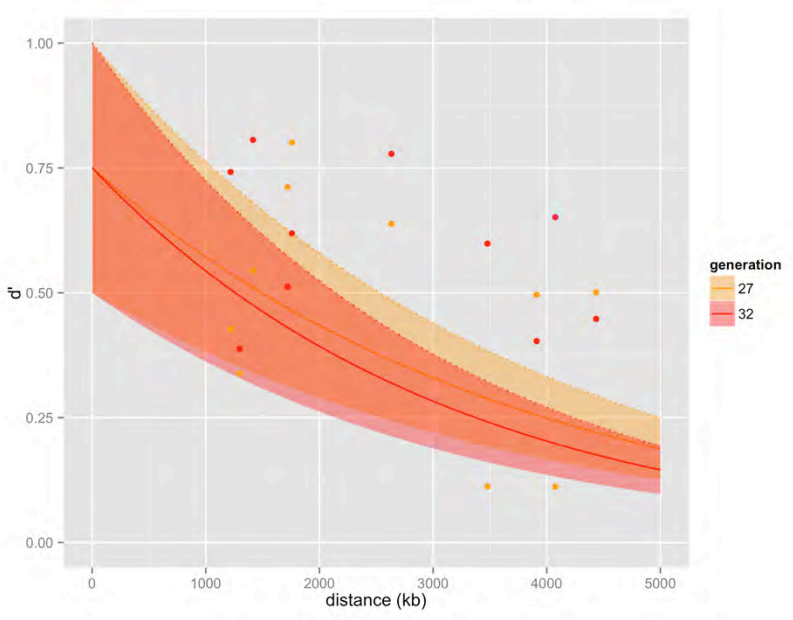
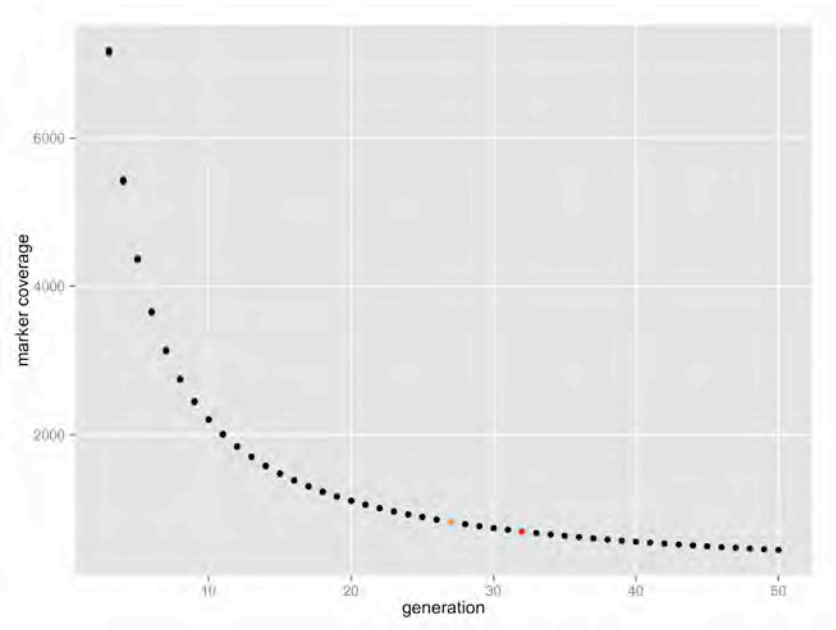


Fig 5.2c: Theoretical decay of highest marker coverage from founding :



Theoretical decay of marker coverage from the founding event based on markers in perfect linkage. Coverage decay proceeds rapidly from the first generation but slows as overall LD in the colony is reduced. Yellow and red points mark the generations at which founder 3 (Mali – generation 27) and founder 9 (Burkina Faso – generation 30) were infected. Theoretical maximum marker coverage at these generations was 821 and 693kb respectively.

5.5: Methods:

Founder populations

The populations sampled were from Mali (Goundry region) and Burkina Faso (Bancoumana) – two regions with high degrees of chromosomal polymorphisms and both of which host *An. gambiae* and *coluzzii* in sympatry.

Gravid females were captured by aspiration indoors in order to ensure that bloodfeeding and any assortative mating had happened under natural conditions. They were then placed in individual oviposition tubes and any resulting eggs collected. Females that laid eggs were collected and stored in ethanol for DNA extraction. F1 eggs from these presumed panmictic groups were placed in a pan of water with Tetramin fish food. Emerged adults were reared under standard conditions at 26°C and 80% humidity, 12 h light/dark cycle with access to cotton soaked in 10% sucrose solution.

DNA of the founding females was screened to ensure they were identical in terms of the IGS marker that defines *An. gambiae / coluzzii* (at the time of experimentation, these were still defined as a single species), and the 2La inversion. Only F1 adults from females identified as *A. gambiae* M molecular form with the karyotype 2La/a, were used to initiate two distinct colonies: Founder population 03 (hereafter Fd03) was started with the F1 offspring from six

mated isofemales originating from Mali and founder population 9 (Fd09) was created with the offspring of eleven mated isofemales from Burkina Faso.
The 2Rb inversion was not controlled for, though it segregates at high frequencies in both countries. The 2Rb markers identified in chapter 4 were not available at the time of founding - leaving only the unreliable Lobo assay for genotyping. It was further considered that, due to its lower F_{ST} as compared to 2La, it was less likely to prove a confounding influence on the association study.
The resultant colonies: founder 3 (Mali) and founder 9 (Burkina Faso) were maintained in the lab for up to 30 generations in order to break up founding haplotypes.
The diversity of the resultant colonies was determined by microsatellite genotyping with comparisions made to wild populations, clearly demonstrating both the maintenance of representative levels of diversity, and the increase in the levels of some low-frequency wild alleles.
At 27 (Mali) and 30 (Burkina Faso) generations, the colonies were infected with *Plasmodium falciparum*. Mosquitoes were dissected at 7-8 days post-infection in order to determine the oocyst load in the midgut. DNA was extracted from all samples. DNA samples were classified into zero, low and high infection pools based upon the oocyst counts (with the low/high separation determined separately for each founder). DNA was then pooled and sequenced on an illumina hi-seq.

Plasmodium falciparum gametocyte culture and mosquito infection

Stocks of *P. falciparum* (isolate NF54) were cultured using the tipper-table system developed by Ponnudurai *et al.* (Ponnudurai et al. 1982) as implemented in the CEPIA mosquito infection facility of the Pasteur Institute (Mitri et al. 2009).
For infection gametocytes are mixed with fresh erythrocytes in AB serum before being transferred to a membrane feeder at 37°C. Mosquitoes were allowed to feed for 15 minutes. Only females that fed well were used for further analysis.

Analysis of mosquito infection phenotypes

Infection phenotypes were oocyst infection prevalence and intensity. Oocyst prevalence is the fraction of mosquitoes carrying at least one oocyst, while intensity is the number of oocysts per mosquito determined only in mosquitoes with ≥1 oocyst. Midguts of bloodfed females were dissected 7-8 days post-infection, stained in 1xPBS buffer with 0.4% mercury dibromofluorescein (Sigma) and the number of oocysts per midgut was determined by light microscopy. Carcasses of the dissected mosquitoes were stored at -20□C until DNA extraction. Genomic DNA was extracted from individual female mosquitoes by

homogenization in 100ul DNAZol (Invitrogen, CA, USA) using a disposable pestle, following the manufacturer's protocol.

Sequencing of mosquito phenotype pools

Based on the observed number of oocysts detected in the females of a *P. falciparum* infection experiment, mosquitoes were assigned to one of three phenotype categories: mosquitoes with zero oocysts (i.e. uninfected), mosquitoes with low infection intensity and mosquitoes with high infection intensity, as follows: Fd03 phenotype pools were constituted from 20 mosquitoes for the "Zero" pool, 17 mosquitoes with 1-5 oocysts for the "Low" pool and 14 mosquitoes with ≥10 oocysts for the "High" pool. Each of the Fd09 phenotype pools was constituted with genomic DNA of 20 mosquitoes, mosquitoes with 1-6 oocysts contribute to the "Low" pool and mosquitoes with more than 29 oocysts for the "High" pool. DNA concentrations were determined by picogreen and DNAs of individual mosquitoes were combined at equal molarity to obtain a total of 700 ng per phenotype pool. The pooled DNAs were subjected to Illumina sequencing.

Mapping method

Genomic mapping was performed in a novel two-stage analysis. An initial coarse mapping performed from the pooled sequence to identify candidate loci, and a second fine mapping to confirm and resolve these loci more finely.

Coarse Mapping

Coarse mapping was initially attempted by seeking regions of divergence between pools; effectively looking for fixation between uninfected and infected samples, or between infected and highly infected. However as a result of high remaining diversity in the colonies, and a low sequencing coverage in the pools (around 0.5X per chromosome) our ability to detect regions of strong F_{ST} was extremely limited.
After the failure of this approach, an alternative method was devised. As the phenotype pools had been selected from a single pool, it could be seen as an application of artificial selection, therefore broad selective sweeps should be visible as they would be in nature. The three phenotype pools were therefore compared in terms of heterozygosity, and their relative heterozygosity reported as a proportion of the total. This approach indicated regions in which the heterozygosity had been significantly reduced in one or two of the pools: candidate regions in which a causative haplotype had been enriched.

Illumina sequences were aligned to the AgamP3 genome (Holt et al. 2002) using Bowtie version 0.12.7 (Langmead et al. 2009). Reads with low mapping quality (MQ < 40) were removed and allele frequencies called using samtools mpileup (Li et al. 2009). No attempt was made to distinguish between low frequency alleles and sequencing errors. Pooled heterozygosity was calculated across sliding windows (10kb windows, 1kb steps) for each of the phenotype pools individually, as well as for the whole founder colony combined, using the *Hp* metric proposed by Qanbari *et al.* (Qanbari et al. 2012). Relative diversity (*HpR*) was calculated as the proportion of pool heterozygosity relative to total heterozygosity across each founder after normalising for overall read-depth in each pool. Standard deviation of *HpR* values (*SHpR*) was used to identify regions with over-represented haplotypes. Random resampling was performed for 1000 permutations for each window to establish significance cutoff values. Allele frequencies were selected randomly from each pool (though with locus positions unchanged), the S*HpR* values recalculated for each permutation. S*HpR* values were then segmented using the *fastseg* Bioconductor (Reimers & Carey 2006) package to identify clearly differing regions. Regions below $1e^{-4}$ probability according to the permutation analysis were removed. Broad regions of association were then selected for subsequent fine mapping.

Ni = read depth at locus *i*

ni = major allele depth at locus *i*

$$Hp = \frac{2 \sum_{i=1}^{1} \mathrm{ni} \ \sum_{i=1}^{1} (Ni - ni)}{\left(\sum_{i=1}^{1} \mathrm{ni} \ + \ \sum_{i=1}^{1} (Ni - ni)\right)^2}$$

HpP = pool *Hp*

HpT = total *Hp*

$$HpR = \sum_{i=1}^{l} \frac{(HpPi - \overline{HpP})}{HpTi}$$

Fine mapping by Sequenom genotyping

Fine mapping within these loci was performed via individual genotyping by Sequenom MassArray™. Probes were selected and designed using the pooled sequence from the coarse mapping, and then individual genotyping was performed on the full infection sample

from which the pools were comprised (the 'deconvoluted' pools). An attempt to use alleles shared between the two pools (i.e. variants that predated the segregation of the two populations) failed –frequently showing monomorphic alleles in one or the other population. Subsequently separate typing plexes were designed that were specific to each founder. Those samples where a positive association was found were then typed in an alternative infection from the same founder line (one that had not been previously pool-sequenced). Individual were categorised into zero, low and high infection phenotypes using the same rules applied previously and assessed via logistic regression using the PLINK software (Purcell et al. 2007).

Loci identified from pooled sequence during the coarse mapping phase were filtered on the basis of differences in the proportion of reads showing the alternate allele (used here as a proxy for minor allele frequency). SNPs with the greatest differences in read-counts between phenotype pools were used to design SNP plexes for genotyping using the Sequenom MassARRAY platform. It should be noted that the SNPs were selected only on the basis read-count differences, without any enrichment for immune or other genes. A single plex (20-25 individual SNP assays) was designed for each locus.

DNA from individual mosquitoes, from the same experimental infection that was pool-sequenced, were typed with SNPs specific to that founder. This included individuals used to generate the pools and additional samples that did not contribute to the phenotype pools. A second completely independent experimental infection of the same founder colony, one that had not been subjected to pooled sequencing, was genotyped in the same way. The former shows that the pooled analysis results are recapitulated in individual genotyping (technical replication) while the latter demonstrates experimental replication with entirely different biological material (biological replication).

Individual mosquitos were categorized into binary phenotypes with respect to infection prevalence (uninfected/infected) and infection intensity (low infected/high infected) using the same oocyst cutoffs employed for pooling. Logistic regression was used to test for significant association with phenotype using PLINK (Purcell et al. 2007) and all statistics controlled for multiple testing. Replicate infections were tested for significance both individually and across replicates.

Locus characterisation and 2Rb inversion typing

Putative variants were filtered for sequencing quality, and consequences of variants were called for both colonies using the Ensembl Variant Effect Predictor (v2.3) (McLaren et al. 2010) against VectorBase genebuild AgamP3.5 (Megy et al. 2012) and using Ensembl API 65.3 (Dec 2011). Enrichment for gene ontology terms was calculated by Fisher's exact test

using custom R scripts and topGO, from the Bioconductor suite (Gentleman et al. 2004). Due to an inability to ensure representation of all individuals within the sequencing pools, Watterson's theta could not be calculated, therefore *d*N:*d*S ratios were assessed by locus counting using custom R scripts; due to the lack of available codon substitution data for this species, no attempt was made to account for multiple substitutions or codon bias in *d*N:*d*S results. Molecular karyotyping of the 2Rb inversion for Fd09 was carried out by a published method (Lobo et al. 2010). Molecular karyotyping results were confirmed against a panel of individuals previously karyotyped by polytene chromosome analysis.

RNAi knockdown of candidate genes:

RNA-mediated silencing was performed on the extant founder 3 colony, ensuring, as far as possible, that the phenotype was tested in the same genetic background as the association study. Founder-3 mosquitoes were injected with *ds*RNA for each of the *TOLL* genes and a *P. falciparum*-infected bloodmeal presented. Double-stranded RNAs were synthesized from PCR amplicons using the T7 Megascript Kit (Life Technologies) from the primers listed in Table 5.5.
For each gene, 500ng of dsRNA (but not more than 207nl volume) were injected into the thorax of cold-anesthetized *A. gambiae* females one day after emergence, using a nanoinjector (Nanoject II; Drummond). A subset of the mosquitoes were assayed via rtPCR in order to determine the efficiency of the knockdown. Primers used in PCR for gene knockdown verification are listed in Table 5.5.

Midguts were dissected and counted as described above. Oocyst count at 8 days post infection was used to derive phenotypes, and ≥30 mosquitoes were dissected per infection replicate. Two to three independent replicate infections were performed.
Differences in the prevalence of infection (i.e. proportion infected vs proportion uninfected) were tested using the Chi-Square test. For infection intensities the Wilcoxon non-parametric test was used, comparing the number of oocysts only amongst those with one or more oocysts. However it should be noted that despite the removal of uninfected samples, ties were still present in the Wilcoxon analysis. P-values for all replicates were combined by the method of Fisher.

5.6: Results

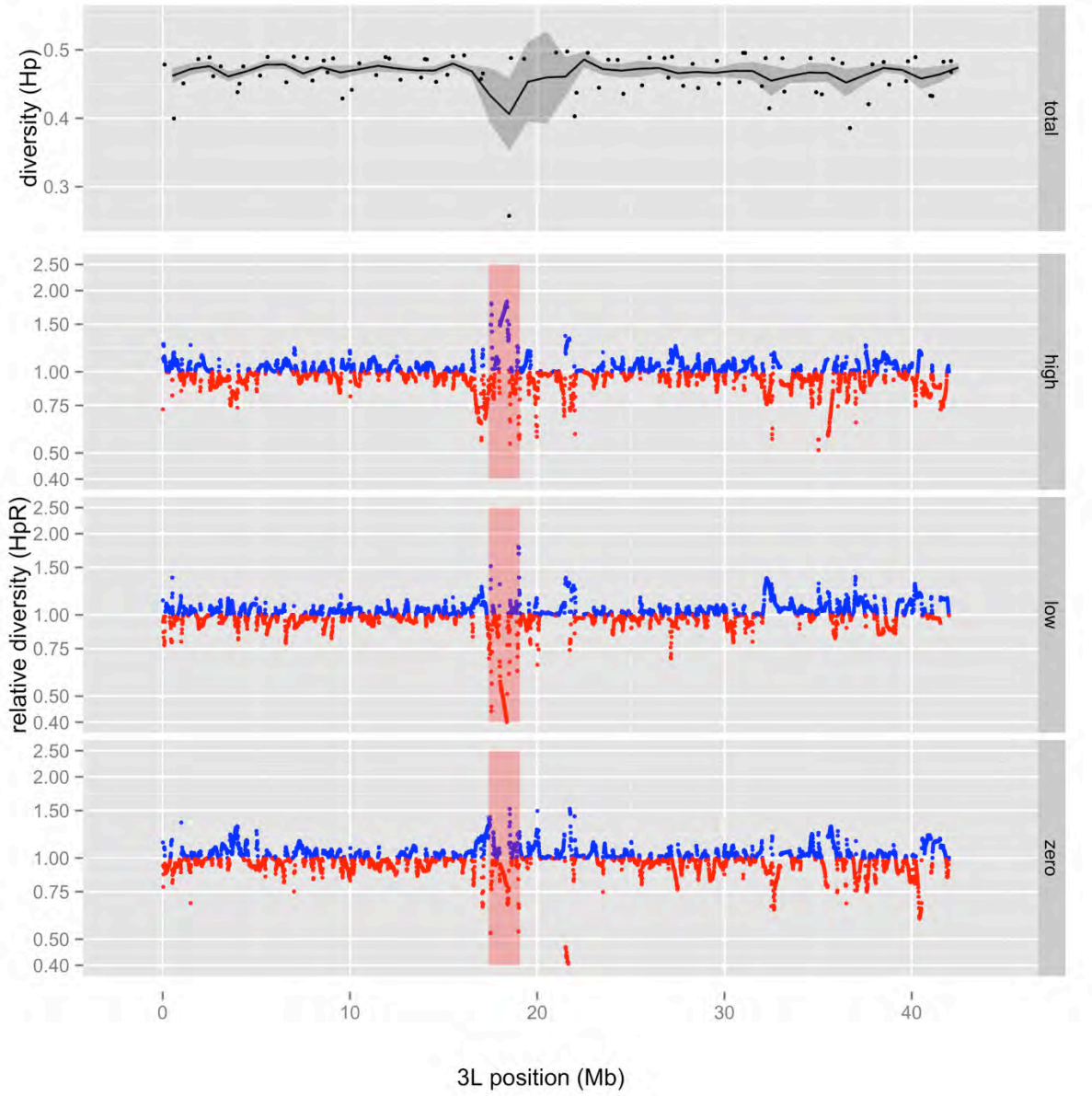
Coarse mapping identifies candidate loci

Coarse mapping identified three candidate loci, two in founder 9 and one in founder 3. The founder 9 loci were both large and both located on chromosome 2R; the first from ~17.5-26.5mb and the second from ~47.5-60mb.

Genetic intervals that displayed reduced nucleotide diversity in a phenotype pool as compared to the diversity of the total 'population' at the same interval (*SHpR*) were detected by sliding window analysis of the phenotype pool sequences. The analysis showed 26 high-*SHpR* regions on chromosome 2R in Fd09, and ten on 3L in Fd03. Individual loci varied in size from 13kb to 511kb. Collapsing adjacent loci on the basis of proximity (≤5Mb) yielded three candidate loci: 3L:17409-19071kb (Fd03), 2R:17385-26524kb (Fd09), 2R:47490-60531kb (Fd09). These candidate loci are named 3.1, 9.1, and 9.2, respectively (Figure 5.1, appendix 4,5). The distal 2R locus (17-26Mb) is coincident with the 2Rb paracentric chromosomal inversion.

A permutation analysis was carried out on the *SHpR* values, by reselecting allele frequencies randomly from each of the phenotype pools. After 1000 tests, the 99.9th percentile of the permuted S*HpR* values was found to be 0.208. At a median S*HpR* of 0.509 and a maximum of 1.42 the selected ten *SHpR* regions from which the locus is derived are well within the 99.9th percentile of those selected randomly (equivalent to a P-value of 0.001), the combined locus 3.1 has a median S*HpR* of 0.215. However, this analysis does not permute LD within the sample, and consequently does not control for the potential positional effects of centromeric or inverted regions.

Figure 5.3a: coarse mapping results illustrating the founder 3 locus:



Coarse mapping plots show heterozygosity measures across chromosome 3L.
At the top is heterozygosity across the total founder 3 sample, the three graphs below show heterozygosity in phenotype pools relative to total (blue lines = increased heterozygosity, red lines = reduced). The coarse-mapped founder 3 locus is marked with a red vertical shaded band and is situated on chromosome 3L 17.4-19.1mb. Pool identity is given on the far right margin of the graph.

Functional description of candidate loci

Coding sequences within candidate loci were analysed for enrichment of Gene Ontology (GO) predicted functional categories. The two candidate loci from Fd09 contain 609 genes for candidate locus 9.1 and 708 genes for candidate locus 9.2. While the large number of genes in the two Fd09 loci might reduce the probability of detecting significant enrichments, both Fd09 loci demonstrated enrichment for genes with potential immune functions, with highly significant enrichment ($P=7.8^{e-6}$) for monoxygenase function in locus 9.1, and the presence of multiple peroxidases in locus 9.2, consistent with either a detoxification or ROS-based immune response. Analysis of genes with significantly enriched GO terms on locus 9.1, however, indicated that most of these genes belong to a single cytochrome P450 cluster between 17.4 and 21.1Mb.

Due to the coincidence of candidate locus 9.1 with the 2Rb chromosomal inversion, molecular karyotyping was carried out on all Fd09 samples. There was only one 2Rb/b individual in the high pool, with heterozygotes occurring randomly between all three pools (giving 2/3/3 copies of the inversion in zero/low/high pools respectively). Thus, there was no association of the frequency of 2Rb inversion genotypes or alleles with membership in phenotypic pools. The Fd03 candidate locus 3.1 contains only 74 genes, the majority of them with no functional information. Of those with characterized function, two encode Toll-family proteins, *TOLL 10* and *TOLL 11* (AGAP001187,AGAP001186) and, largely due to the presence of these two genes, this locus shows significant enrichment for receptor binding ($p=3.50e^{-07}$) and signal transduction ($p=4.45e^{-03}$).

Fine mapping of candidate loci

SNPs within the candidate loci displaying the greatest difference in minor allele frequencies between any two phenotype pools were selected for genotyping of individual mosquitoes. Due to the high diversity within the two colonies, and the expected high Fst between them, SNPs were selected and tested only within each founder colony. A total of 44 SNPs were chosen from Fd09 across candidate loci 9.1 and 9.2, and 23 SNPs from Fd03 for candidate locus 3.1. It should be noted that whilst markers were deliberately biased towards those with a high probability of association within the fine-mapped samples, there was no pre-selection of markers with relation to specific immune-related genes.

For each founder colony, fine-mapping was performed by genotyping DNAs from all of the individual mosquitoes from the original infections; that is, the infections from which the phenotype pools were comprised. For Fd03, a second replicate infection from the same colony (that had not been previously pool-sequenced) was also genotyped; these two

replicate infections were assessed by logistic regression separately, and – where the odds ratio indicated the same effect – as one experiment.

Fine-mapping, Founder 9 :

Candidate locus 9.1 contained no significant (or even close to significant) SNPs in the deconvoluted pools, for either infection intensity or prevalence. Given the coincidence of this locus with the 2Rb inversion and the confirmed presence of the 2Rb locus within this colony this seems to strongly imply a type 2 error due to the presence of the 2Rb inversion.
Candidate locus 9.2 also contained no significant markers using a significance cutoff of p<=0.01. The lowest P-value in either experiment was p=3.18e-3 at 2R:55095410 within locus 9.2. Despite the presence of a number of markers with reduced, though non-significant, p-values for infection prevalence (2R: 52949796, 53067336, 53475245, demonstrating p-values of 6.55e-2, 6.43e-2, 4.44e-2 respectively) none of the p-values were significant after multiple testing correction (an adaptive monte-carlo permutation analysis as implemented in PLINK) and replicate infections were not genotyped in order to look for association.
All founder 9 fine-mapping results are shown in appendix 5.

Fine mapping : Founder 3

Candidate locus 3.1 contained two SNPs with significant association after permutation for both oocyst prevalence and intensity. The variant at 3L:18559884 is a C:A mutation in the intergenic region between *TOLL 11* (AGAP011186) and *TOLL 10* (AGAP011187) with an odds ratio of 7.79 in samples of high infection intensity (p=0.00148, calculated across both replicates). The variant at 3L:18552220 is a T:C mutation located immediately downstream of *TOLL 10*, with an odds ratio of 3.15 in relation to infection prevalence (p=0.002594, calculated across both replicates). Both of these markers were individually significant in one replicate each, with significance values in each case passing multiple testing correction (implemented as above), as well as being jointly significant and passing multiple testing correction when assessed as a combined experiment, or when results were combined using Fishers' method. Combining replicates in all cases emerged as the more conservative approach and was used in order to call significance and to generate all Manhattan plots.

Table 5.1 Association results for Fd03 individuals (following page)

Association is tested with respect to infection prevalence, i.e. zero *cf* low + high (ST1), and intensity, i.e. low *cf* high (ST2). Logistic regression p-values are given for rep1, rep2, both reps combined. P-values for replicates combined by Fisher's method, and after multiple testing correction are also given. Markers with significant signals after Fisher combination and Bonferroni correction are highlighted

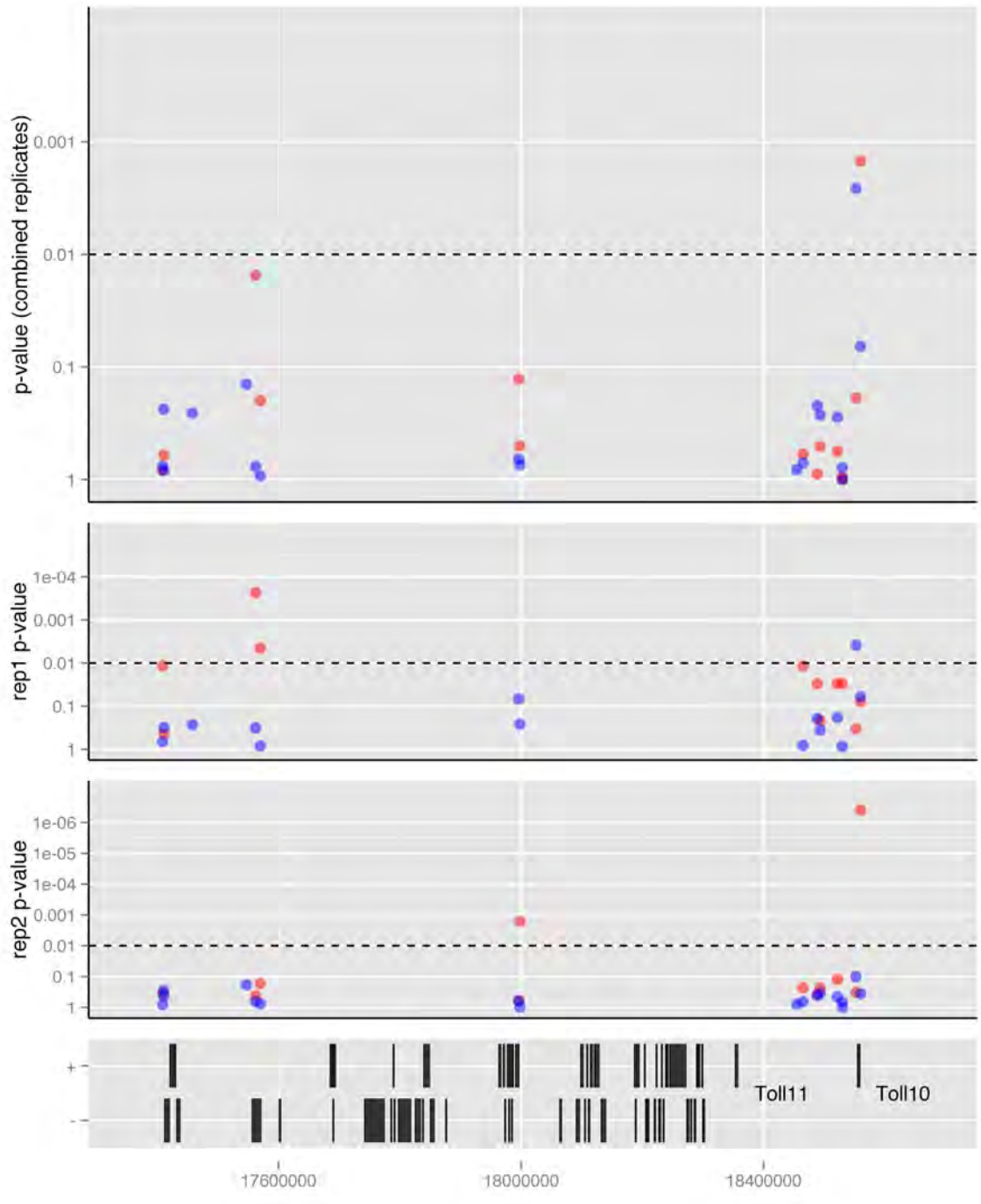
Table 5.1a: Fd03 prevalence assoc

		Replicate 1			Replicate 2			Combined		Combined (Fisher)	
SNP_ID	BP	Odds ratio	P value	P adj	Odds ratio	P value	P.adj	Odds ratio	P value	Fisher	Bonferroni
3L_17409051_G-A	17409051	1.345	6.66E-01	1.00E+00	0.9052	8.13E-01	7.78E-01	1.112	7.62E-01	8.73E-01	1.00E+00
3L_17410678-94_3-2	17410678	-	-	1.00E+00	0	2.86E-01	8.57E-01	0	2.38E-01	-	-
3L_17410825_G-A	17410825	-	3.12E-01	3.57E-01	0	4.20E-01	8.57E-01	1.351	8.32E-01	3.97E-01	1.00E+00
3L_17458216_C-A	17458216	0	2.69E-01	7.78E-01	-	-	1.00E+00	0	2.57E-01	-	-
3L_17546985_C-G	17546985	-	-	1.00E+00	0	1.85E-01	3.41E-01	0	1.42E-01	-	-
3L_17557010_T-C	17557010	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_17562156_A-T	17562156	1.677	3.21E-01	6.67E-01	0.818	6.41E-01	7.78E-01	1.103	7.66E-01	5.30E-01	1.00E+00
3L_17569921-2_2	17569921	1.132	8.38E-01	1.00E+00	0.9011	7.74E-01	7.78E-01	0.973	9.28E-01	9.29E-01	1.00E+00
3L_17995944_T-A	17995944	0	6.78E-02	1.28E-01	1.29	6.10E-01	7.27E-01	0.8202	6.61E-01	1.73E-01	1.00E+00
3L_17998025_A-G	17998025	0	2.61E-01	5.88E-01	1.01	9.86E-01	1.00E+00	0.8419	7.40E-01	6.06E-01	1.00E+00
3L_18454541_A-C	18454541	-	-	1.00E+00	1.25	7.93E-01	1.00E+00	0.8333	8.20E-01	-	-
3L_18464918_G-A	18464918	0.8571	8.11E-01	5.56E-01	0.8571	6.61E-01	8.57E-01	0.8957	7.11E-01	8.70E-01	1.00E+00
3L_18468533_A-C	18468533	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_18468614_G-A	18468614	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_18485024_T-A	18485024	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_18488581_G-A	18488581	0.4274	1.93E-01	2.39E-01	0.7534	4.15E-01	5.88E-01	0.6977	2.21E-01	2.82E-01	1.00E+00
3L_18493248-9_2	18493248	0.3696	3.59E-01	5.56E-01	0.6963	3.43E-01	5.88E-01	0.6827	2.67E-01	3.81E-01	1.00E+00
3L_18521673-4_2	18521673	0.4193	1.83E-01	2.00E-01	0.7714	4.54E-01	7.27E-01	0.7287	2.79E-01	2.90E-01	1.00E+00
3L_18529641_G-T	18529641	1.142	8.65E-01	1.00E+00	0.8411	6.79E-01	7.27E-01	0.908	7.88E-01	9.00E-01	1.00E+00
3L_18530001-2_2	18530001	-	-	1.00E+00	1	1.00E+00	8.57E-01	1	1.00E+00	-	-
3L_18552220_T-C	18552220	8.5	3.86E-03	5.19E-02	2.154	1.00E-01	2.86E-01	3.149	2.59E-03	3.43E-03	4.46E-02
3L_18559884_C-A	18559884	0.2549	6.08E-02	1.27E-01	0.6425	3.58E-01	5.88E-01	0.4842	6.55E-02	1.05E-01	1.00E+00
3L_18569272_A-G	18569272	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-

Table 5.1b: Fd03 intensity assoc

		Replicate 1			Replicate 2			Combined		Combined (Fisher)	
SNP_ID	BP	Odds ratio	P value	P.adj	Odds ratio	P value	P.adj	Odds ratio	P value	Fisher	Bonferroni
3L_17409051_G-A	17409051	-	1.17E-02	1.70E-02	0	3.49E-01	1.00E+00	1.143	8.28E-01	2.64E-02	2.38E-01
3L_17410678-94_3-2	17410678	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_17410825_G-A	17410825	0	4.39E-01	8.57E-01	-	-	1.00E+00	0	6.06E-01	-	-
3L_17458216_C-A	17458216	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_17546985_C-G	17546985	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_17557010_T-C	17557010	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_17562156_A-T	17562156	-	**2.33E-04**	**1.70E-02**	0	4.16E-01	1.00E+00	3.4	**1.53E-02**	**9.91E-04**	**8.92E-03**
3L_17569921-2_2	17569921	-	**4.48E-03**	**1.70E-02**	3.857	1.69E-01	1.14E-01	1.92	2.00E-01	**6.20E-03**	5.58E-02
3L_17995944_T-A	17995944	-	-	1.00E+00	0	6.06E-01	1.00E+00	0	1.28E-01	-	-
3L_17998025_A-G	17998025	-	-	1.00E+00	16.5	1.60E-03	3.64E-02	1.8	5.04E-01	-	-
3L_18454541_A-C	18454541	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_18464918_G-A	18464918	-	1.17E-02	1.88E-01	0	2.35E-01	6.25E-01	0.7302	5.95E-01	1.89E-02	1.70E-01
3L_18468533_A-C	18468533	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_18468614_G-A	18468614	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_18485024_T-A	18485024	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-
3L_18488581_G-A	18488581	-	3.01E-02	5.05E-02	3.375	3.73E-01	5.24E-01	0.9231	8.94E-01	6.17E-02	5.55E-01
3L_18493248-9_2	18493248	-	2.17E-01	3.13E-01	4.833	2.34E-01	7.78E-01	0.5952	5.09E-01	2.02E-01	1.00E+00
3L_18521673-4_2	18521673	-	3.01E-02	5.05E-02	3.533	1.26E-01	1.42E-01	1.357	5.63E-01	2.50E-02	2.25E-01
3L_18529641_G-T	18529641	-	3.01E-02	5.05E-02	-	-	1.00E+00	1.036	9.59E-01	-	-
3L_18530001-2_2	18530001	-	-	1.00E+00	-	-	1.00E+00	1	1.00E+00	-	-
3L_18552220_T-C	18552220	0	3.34E-01	6.43E-01	0	3.24E-01	6.25E-01	2.286	1.89E-01	3.49E-01	1.00E+00
3L_18559884_C-A	**18559884**	-	7.89E-02	3.13E-01	44.67	**3.95E-07**	**1.27E-02**	7.786	**1.48E-03**	**5.70E-07**	**5.13E-06**
3L_18569272_A-G	18569272	-	-	1.00E+00	-	-	1.00E+00	-	-	-	-

Figure 5.4: Fine mapping via Sequenom SNP typing:



This is a Manhattan plot of the region within the coarse-mapped founder 3 locus. Association is calculated by logistic regression of founder 3 SNPs using individual genotypes called by Sequenom genotyping. Association is calculated (a) across both biological replicates and (b) separately for individual reps 1 and 2. Red points show p values for association with intensity (high vs low pools) blue points show p values for SNP association with prevalence (zero vs low + high pools). The dashed line shows the significant p-value (p=0.01). Note the different y-axes. Genes within the locus are shown beneath the w plot, *TOLL 11* and *TOLL10* are labelled.

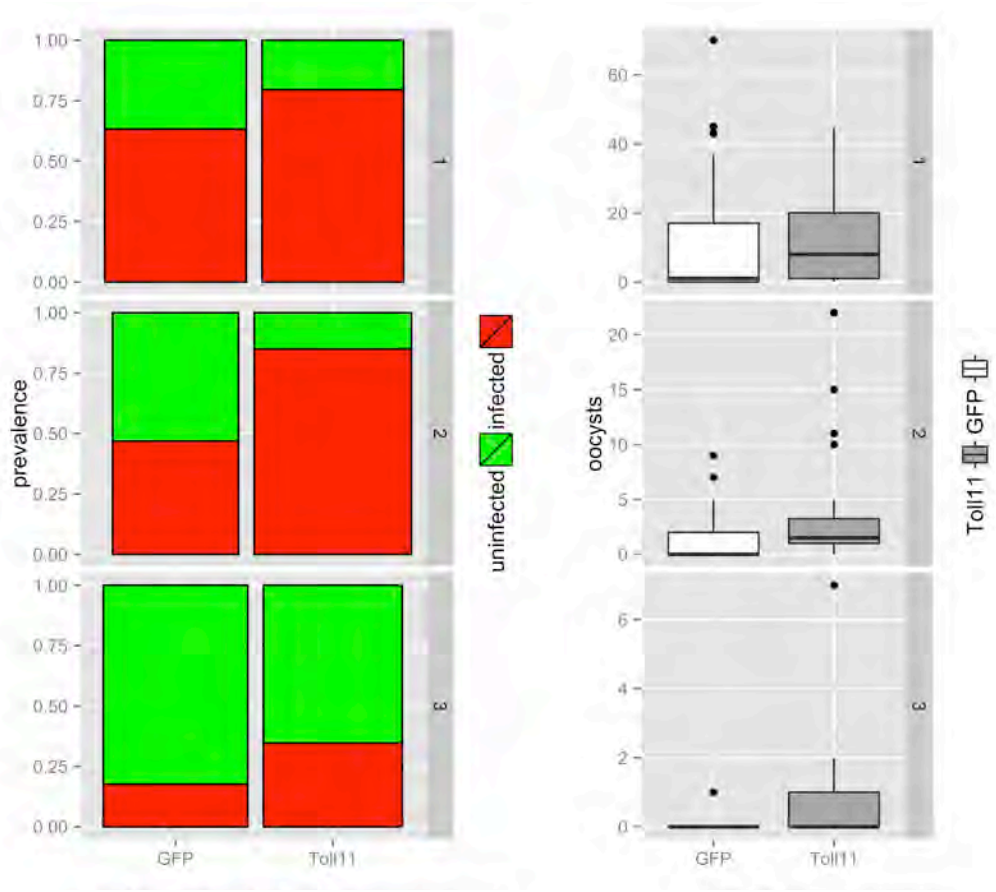
TOLL10 / 11

The two associated SNPs are both found in close proximity to two genes in the *TOLL* family: 18559884 is located downstream of *TOLL10* and 18552220, is found in the intergenic spacer between *TOLL11* and *TOLL10*.
Combined analysis across both replicates demonstrated a consistent response in both sets (albeit of differing magnitude) and this response was robust to multiple testing correction (Bonferroni). No other genes were within the expected coverage range of these two markers, and no other genes in the locus were found to have a statistically significant locus across both replicates. We therefore consider this locus to have been refined to this two-gene region.

RNAi knockdowns : *TOLL 11* displays protective function against *P. falciparum*

The small size of the locus enabled us to test both (i.e. all) genes for their effect on antiplasmodial immunity. A functional test of *TOLL 11* by RNAi-mediated gene silencing followed by challenge with *P. falciparum* reveals significant protection by *TOLL 11* against oocyst infection prevalence. Three replicates were performed for *TOLL11*. Knockdowns of this gene caused an increase in oocyst prevalence of 16-38% across three replicates (mean 24.5%) with a mean risk ratio for infected as compared to uninfected categories of 1.71 (individual=1.26, 1.81, 2.08 respectively). Probabilities for oocyst prevalence differed between replicates (p=0.1175, 0.0014, 0.0864), with only one replicate showing individual significance, however the consistent increase in infection prevalence across three replicates was significant (p=0.0011, p values combined by the method of Fisher). *TOLL 11* had no effect upon oocyst intensity (p-values=0.85, 0.85, 0.32; combined=0.82).
TOLL10 knockdowns were less predictable: Uninfected/infected risk ratios (0.56, 0.69, mean=0.63) did not indicate a phenotype for infection prevalence, and Wilcoxon rank-sum results were non-significant (p=0.12, 0.19, combined=0.11), despite showing a similar tendency across both replicates.
One result was mildly significant for infection intensity (p=0.038) with the knockdown individuals showing a decrease in mean oocyst count, however despite a similar tendency for a reduction in oocyst count in the second replicate this was also non-significant across the two replicates (p=0.038, 0.54; combined p=0.10). Due to this non-significant second replicate, a third replicate was not pursued.
In each case the efficacy of the knockdown was assessed by rtPCR in a subset of the infected samples (see figure 5.7).

Figure 5.5: dsRNA-mediated knockdowns of TOLL11

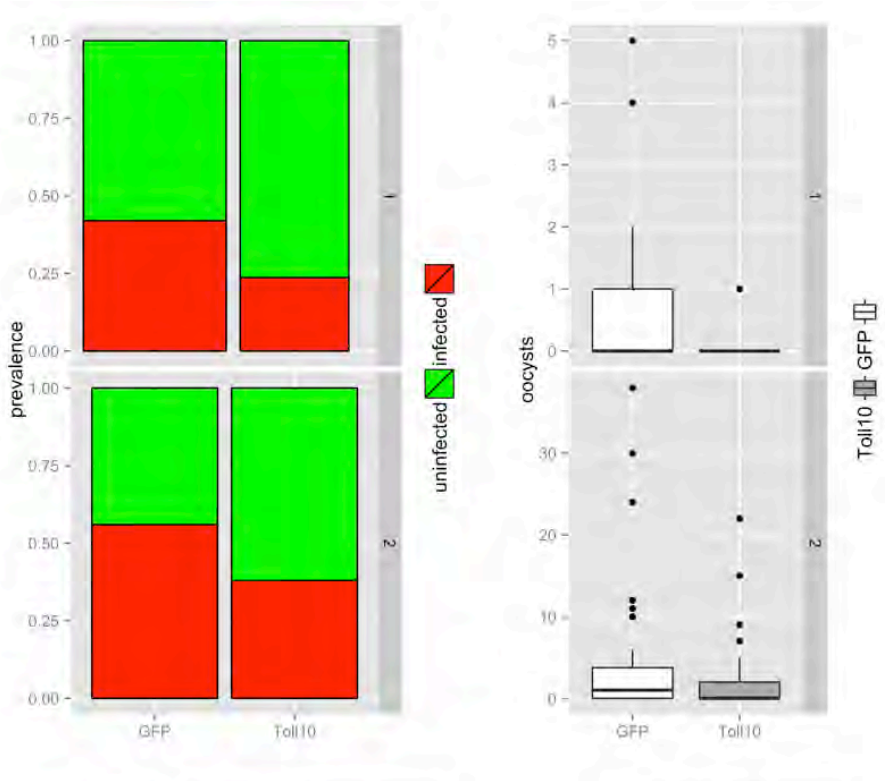


Toll11 RNAi mediated gene silencing results. (a) Prevalence of infection, green = 0 oocysts, red >1 oocyst for 3 replicate knockdowns of founder 3 individuals (replicate number in grey box). P-values are calculated by X^2 test of infected/uninfected individuals. (b) Intensity of infection measured by the number of oocysts in midguts. Note the difference in y axes across replicates. Significance is calculated by Wilcoxon rank-sum test on infected individuals only. P-values across replicates for intensity and prevalence were combined by the method of Fisher.

Table 5.3: TOLL11 RNAi knockdown prevalence / intensity results

	rep.1	rep.2	rep.3	combined
GFP uninfected	18	17	42	
GFP infected	31	15	9	
	63.27%	46.88%	17.65%	
TOLL 11 uninfected	10	6	32	
TOLL 11 infected	39	34	17	
	79.59%	85.00%	34.69%	
Risk ratio	1.258	1.813	1.966	1.679
p	0.1175	0.0014	0.0864	0.00106
GFP mean oocysts	11.24	1.34	0.18	
TOLL 11 mean oocysts	12.14	3.23	0.49	
Risk ratio	0.742	1.024	inf	0.833
p	0.8495	0.8479	0.3217	0.8182

Figure 5.6: dsRNA-mediated knockdowns of TOLL10

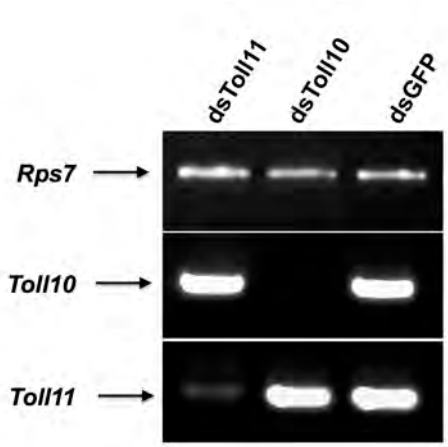


Toll10 RNAi mediated gene silencing results. As with Fig 5.5 (a) Prevalence of infection, green = 0 oocysts, red >1 oocyst for 3 replicate knockdowns of founder 3 individuals (replicate number in grey box). P-values are calculated by X^2 test of infected/uninfected individuals. (b) Intensity of infection measured by the number of oocysts in midguts. Significance is calculated by Wilcoxon rank-sum test on infected individuals only.
Despite having a similar tendency over the two replicates, the lack of any significance in the second replicate was considered to be definitive, and a third round of replication was therefore not pursued.

Table 5.4: TOLL10 RNAi knockdown prevalence / intensity results

	rep.1	rep.2	combined
GFP uninfected	29	24	
GFP infected	21	28	
	42.00%	53.85%	
TOLL 10 uninfected	29	30	
TOLL 10 infected	9	19	
	23.68%	38.78%	
Risk ratio	0.564	0.692	
p	0.1168	0.1875	0.10558
GFP mean oocysts	0.72	3.96	
TOLL 10 mean oocysts	0.24	1.8	
Risk ratio	inf	1.188	1.429
p	0.0383	0.5381	0.1006

Figure 5.7 : Knockdown confirmation of *Toll10* / *Toll11*



The efficiency of Toll10 / Toll11 silencing is demonstrated by rtPCR in a subset of the injected mosquitoes.
Knockdown efficiency is shown by comparison to the invariant S7 loading control.
It should be noted that, due to the need to allow oocyst development before assaying, the rtPCR illustrates the reduced transcript levels at the time of infection, but not necessarily at the time of assay.

Table 5.5: RNAi knockdown / verification primers

Sequence of the primers used for the synthesis of the double-stranded RNA and for knockdown verification. The primers called V are used for the knockdown verification

T7-GFP-F	taatacgactcactatagggcatggtgagcaagggcgag
T7-GFP-R	taatacgactcactatagggcttacttgtacagctcgtc
T7-TOLL10-F	taatacgactcactataggaggaccgggtgccggccaagc
T7-TOLL10-R	taatacgactcactataggtccggtatcgtggagtacacg
T7- TOLL11-F	taatacgactcactataggtacggtgtgcggctgtgcaagg
T7- TOLL11-R	taatacgactcactataggtcgggcatggcaaaccgcagc
TOLL10-VF	tacgcgctgcccgacgtgccg
TOLL10-VR	tccggtatcgtggagtacacg
TOLL11-VF	tgcctgtgggcccactcccgg
TOLL11-VR	tcgggcatggcaaaccgcagc
Rps7-F	aggcgatcatcatctacgtgc
Rps7-R	gtagctgctgcaaacttcgg

5.7: Discussion

We have devised a novel method of genomic mapping via pooled sequence and recently collected 'founder' colonies and demonstrated its applicability to a complex non-model genome. This novel method has been used to identify a locus containing two genes that were orthologous to known immune receptors. We have further confirmed the antiplasmodial action of one of these by dsRNA-mediated knockdown. This may also have implications for the future study of immunity in wild samples. Both the success of the method and the potential immune consequences will be discussed below.

I: loss-of-heterozygosity mapping:

This strata-controlled mapping method has successfully identified a novel locus containing two genes, one of which we have confirmed to have a significant phenotype via dsRNA-mediated knockdown. Whilst these genes were both in the cadre of 'potential immune genes' that were identified in chapter 2 / appendix 3 (and indeed they have both been used as candidates in target-gene studies) neither has been previously assigned an immune role by any other method. Given the considerable pleiotropy shown by other members of the TOLL family it was far from certain that these would have an immune function at all, and more specifically their implication in the anti-plasmodial immune system is entirely novel.
It is important to note, also, that at no point was the locus refined on the basis of gene annotation or homology, therefore no potential 'annotation bias' was introduced into the method. Indeed given the unbiased nature of the coarse mapping, the identification of genes with known immune homologues reinforces the validity of the locus.
The functional testing, too, was not prejudiced by any immune annotations (i.e. 100% of the candidate genes within the locus were tested) although we were, of course, aided significantly in this by the small nature of the fine-mapped locus. Given the low conversion rate of the Sequenom assays, we cannot entirely rule out the possibility that genes outside the locus contributed to the phenotypic signal, yet were not associated due to marker failure in one or both replicates.
This method represents, therefore, the first truly whole-genome assay for *Plasmodium* immune function in this species. And moreover, a method that provides a tractable and unbiased method of investigating phenotypic diversity that can lead to highly resolved loci. Assuming that resolution of QTL studies is likely to remain low (in the absence of a 'collaborative-cross' style resource), this method is the only current means of identifying entirely novel individual genes of strong effect.

Mapping method assessment:

This method has a number of significant advantages over previous approaches and provides a solid basis for future association studies. As compared to single-mosquito mapping, this method requires only moderate expenditure on both sequencing and genotyping, however it's benefits are not limited to cost effectiveness. It also shows particular pluses when applied to *Anopheles gambiae* – a genome characterized by high levels of diversity and markedly low LD. The mosquito husbandry techniques used in this study alleviate both of these problems permitting statistically robust associations to be made.
The use of distinct colonies from different locations has two major advantages. Primarily this allows us to reduce the risk of type 2 errors due to population structure. However it also has the effect of isolating individual genomic effects, enabling us to detect lower effect or modulatory variants without the influence of strong-effect genes such as Tep1 R/S dominating the signal. Finally the availability of these strata-controlled founder colonies assists in subsequent studies, since we are able to carry out functional assessment of mapped genes against the same genetic background.

Coarse mapping

Loss-of-diversity scanning is also well suited to highly diverse species such as *A.gambiae*. Classical association of a causative mutation using tagging SNPs requires that those tag SNPs repeatedly occur in a sample set. That is, that the exact haplotype that links causative locus to tag locus occurs at high frequency. In low LD - high diversity species, where haplotype length is short, the likelihood of this is greatly reduced. By instead performing our coarse mapping using loss of diversity we can view a reduction in the overall number of haplotypes without relying on a single haplotype linked to a single tag SNP.
Nevertheless some major caveats do remain. The sequencing dataset examined here contains neither technical nor biological replication. This lack of replication drastically reduces our ability to make basic assessments of the method as we do not have comparable pools from which to derive an estimate of the null model. Phenotype pools are not comparable across colonies (e.g. FD09 low vs Fd03 low), and within each colony we have only phenotypically linked pools. It follows that a true, unbiased, power analysis is currently impossible.
A biological replicate from the same colony – that is, a separate infection of the homogeneous colony – would have allowed us calculate *Hp* and *HpR for* all permutations of of pools across both infections. Doing this we could build up a fair model of the distribution of the Hp statistic under the null model; significance could then be fairly tested by assessing the deviation of the real HpR values from the null model.

Indeed, the problems caused by lack of replication are not limited to defining genome-wide significance. *Anopheles* is known to show major deviations in both variance and linkage across the genome, particularly in pericentromeric regions, or regions of large structural variation such as chromosomal inversions, with many of these regions linked to processes of speciation (Lawniczak et al. 2010) and reproductive isolation (Coluzzi et al. 2002). Regions such as these, where LD is increased and diversity reduced, will serve to amplify relative reductions in heterogeneity and will therefore greatly increase the likelihood of false positive results in the coarse-mapping stage. The type 2 error associated with the 2Rb inversion serves to illustrate the susceptibility of this method to genomic structure.

Since our permutation destroyed haplotype structure, we were unable to assess how the haplotypes in a particular region would behave under a null model; that is, since the within-pool variation cannot be assessed fairly for different parts of the genome, no control is possible for broader genomic effects such as a general reduction in theta within the locus. Such a reduction would have a greater effect on the variance in the HpR score in a linked set than in the unlinked permutation set. Indeed, similar relative-diversity approaches have been shown to be biased as a result of such fluctuations in diversity and LD within *Anopheles* (Cruickshank & Hahn 2014) and we could reasonably expect the HpR statistic to be susceptible to the same effect.

Indeed simple improvements could have been made simply by sequencing subsets of the same phenotype pools; whilst this would reduce the depth of sampling for each phenotype pool it would have allowed for an unbiased measurement of the distribution of within-pool diversity without requiring additional infections.

It remains to be seen whether further replication from the same location (that is, founding multiple colonies from the same location) would be worthwhile, or even desirable. Though it might be considered that multiple colonies would enable us to attempt replication of the signal, as well as performing the permutation outlined above, this is not necessarily the case. Given the uneven sampling of wild haplotypes that is revealed by the microsatellite analysis, where some alleles are present at greatly increased frequencies and others not at all, it is somewhat doubtful that the two colonies would be directly comparable, and type 1 errors are highly likely. Similarly, stochastic variations in captured haplotypes may generate very different distributions of HpR from one colony to the other; therefore this does not obviate the need for replicate infections per colony and may provide no greater benefits over sampling in a different location as was performed here.

Finally, pooled sequencing, although cost effective, also comes with distinct disadvantages. We are not able to ascertain the degree of concordance between read coverage and allele frequencies with any real confidence, and this has a drastic effect on our ability to select

effective assays for fine mapping. Although experiments in *Drosophila* have indicated a high correlation between these two measures (i.e. concordance >0.9 for GATK-validated SNPs) (Zhu et al. 2012), these results may not be directly transferrable to *Anopheles*. In comparisons between pooled and individual Sanger sequences in the mosquito, Weetman *et al.* (Wilding et al. 2009) found a higher degree of noise and a lower correlation (R^2 = 0.61). This problem was particularly acute for the *Anopheles coluzzi* that comprises both of our founder populations. This doubtless contributed the lack of marker conversion we saw during fine mapping (see below) and the poor fidelity of the minor-read count when used as a proxy for minor allele frequency.
It is almost certain that these differences are due to the use of the PEST genome as a reference, and this method should improve once the *An. gambiae s.s.* (Molecular form S) and *An. coluzzi* genomes are fully assembled.

Fine mapping:

The fine mapping approach was based on long-established methods of association and a proven genotyping technology, and was performed at the individual level, removing a large degree of the uncertainty from this stage. However marker design was based upon the prior coarse mapping, and could therefore be significantly biased by marker selection.
In all three of the fine-mapped loci the genotyping plex was not evenly spread across the assayed region. Selection of markers was biased towards individual regions within the loci, and, particularly in the case of locus 9.2, the chosen markers do not always cover the entire locus.
Instead of being chosen for even coverage, given the difficulty in distinguishing real variants from sequencing errors (indeed the impossibility at this level of coverage and with no replication) markers were prioritized if they had a high read-shift between phenotype pools – i.e. if they were most likely to be real SNPs and by implication to give us a positive result in the fine-mapping association. At no point, however, were they chosen on the basis of the functional predictions of the genes within the locus. The enrichment for markers in the *Toll10/11* region is - from the point of view of marker selection - entirely coincidental.
In terms of the experimental design, this has the severe disadvantage of preventing us considering the fine mapping to be a validation of the coarse-mapped locus. That is, had markers been chosen independently from the coarse mapping, their association in the fine mapping would be an independent corroboration of the reduction in haplotype diversity, however as they were chosen to replicate this signal this is not the case. Though a complete lack of association in the fine mapping could reasonably be taken as a negation of the coarse-mapping result, a positive association is reinforcement rather than true validation.

However, this does mean that the identification of immune-family genes in this case is not due to manual selection. Not only might we anecdotally consider the presence of two immune-family genes to be something of an anecdotal validation of the method in itself, but more importantly it means that this method would be equally adept at identifying entirely novel genes, an important criterion for genome wide association studies.
An improved version of this experiment, with valid biological replication in the coarse mapping stage, would, of course, remove many of the statistical weaknesses of the method whilst maintaining this all-important neutrality.

Gene identification

The identification of *Toll10* and *Toll11* as being the genes underlying the locus is based on the association of two SNP markers on chromosome 3L at 18,552,220 and 18,559,884bp. The two genes are at 18,353,919-18,358,304 (*Toll11*) and 18,555,488-18,559,531 (*Toll10*). Both markers are adjacent to *Toll10*, one upstream and one downstream, with the upstream marker also 200kb downstream of *Toll11*. It is important to note the distinction between a potentially causative variant and a SNP marker. A putative causative variant would have to be within the coding or regulatory region of a gene (a somewhat arbitrary region, but typically defined as between 1-5kb upstream of a gene and 0.5-1kb downstream (McLaren et al. 2010)), whereas a marker need only be in strong LD with the causative variant. The coverage of these markers will hence be determined by the expected LD within the dataset, and as we calculated this to be up to 821kb (for markers in perfect linkage at founding and at $r^2 = 0.8$ at the time of infection) this could potentially give our two markers a maximum coverage of 17,731,220 to 19,380,884bp - more than enough to cover both *Toll10* and *Toll11* (along with another 71 genes). Sadly, since the only individual genotyping undertaken during this study was actively looking for markers linked to phenotype, we are not able to gauge true LD within the locus, however the proximity of these two markers to both *Toll10* and *Toll11* means that even a three-fold drop in r^2 would still include both of these genes within the locus.

Confirmation by functional assay

Since the use of the fine-mapping as statistical validation for the coarse mapping is problematic, confirming the validity of the locus is left to the functional assay. This is an approach that is widely used in *Anopheles* research (Stathopoulos et al. 2014; Harris et al. 2010) due to the difficulty of replication studies, yet as a validation it is not without weaknesses.

It remains to be determined what proportion of immune genes give an immune phenotype, and since publication and ascertainment bias will prevent us from ascertaining the true proportion of genes that have an immune effect from the previous literature, only a large-scale (genome wide?) knockdown study would allow us to determine the 'false discovery rate' of RNAi. This is a major fault in those studies where a large locus is identified and then a gene manually chosen from within this locus on the basis of its functional prediction. This is notably *not* the assay we have performed here: the locus was defined as two genes based on the positions of the fine-mapped SNPs, and both genes within this locus were tested – with 50% giving a statistically significant immune-related phenotype. We were, of course, assisted greatly in this by the small size of the locus, which reduced the number of testable genes to a very manageable two. It could be argued that the locus, with a potential coverage of more than 1.6Mb, should be extended further. It is certainly possible that genes outside of this locus for which we did not detect associable SNPs contributed to the coarse-mapped region, or indeed that further replicates might have confirmed the significance of the loci at 17562156 and 17569221 that were weakly significant for infection prevalence in replicate one. However, in the absence of such confirmatory replicates, the use of the two closest genes to the associated markers seems an eminently sensible and defensible approach and it is unlikely that the false positive rate for RNAi knowndowns approaches the 50% positive rate seen here.

Replication in wild samples:

If we truly aim to discover the genes important for disease transmission, it is desirable to map, replicate and test in mosquitos that are as close as possible to those that will be found in the wild. Founder colonies such as these (and the newly-founded 'ngousso' strain used by Harris *et al*) represent an effective compromise between tractability and applicability, however their applicability as a model – and as a mapping resource - will inevitably diminish with time.

Moreover, whilst gene knockdowns and other functional studies demonstrate the efficacy of GWAS techniques in discovering the immune function of novel genes, this does not necessarily mean that the variants themselves are important for malaria transmission anywhere outside of the region and time that the founder line was established. Replication in the wild would be required in order to define any of these alleles as markers of susceptibility or refractoriness.

Demonstrating the effect of this locus in the wild remains a challenge. Due to a history of rapid expansion followed by decline (O'Loughlin et al. 2014) the majority of variants in *A.gambiae* are rare and there is little allele sharing between divergent populations. The

failure to apply the same markers to two founders also points to the distinct lack of allele sharing in the *gambiae* populations; given the proximity of the two countries from which the founders were taken (the countries are adjacent and Burkina Faso shares a similar climate to southern Mali), there is little reason to expect greater success in marker replication in any more distant populations. Of the five SNPs showing association in either replicate, only one has been demonstrated to segregate in a previously sequenced field colony (3L:17569921:GC/AA in the MR4 'Kisumu' colony) and it is unlikely that a panel of the same SNPs would be found at high allele frequencies in any given field sample.
Coupled with the low infection intensities that will be encountered in field settings, it is doubtful that attempts to replicate association of this locus in the wild with the same SNP panel would be successful.
Mapping directly from wild samples – without either the intermediate step of founder colonies or pre-selection of targets is even more unlikely. Until a reliable, broad-geography, haplotype map is developed for this species that would allow imputation against SNPs known to type for ancestrally informative haplotypes, the prospects for performing successful genome-wide association studies directly from wild individuals will remain slim.

II Mapped locus:

Without resorting to pre-selection of immune genes, or 'cherry picking' from within a large region, we have identified a locus containing two genes that are members of a large family of immune receptors. We have further demonstrated one of these genes to have a measurable effect on the development of *Plasmodium falciparum* in the Mosquito midgut. However it is not possible to draw conclusions about the mechanism or function of this anti-*Plasmodium* action from this experiment.
Nevertheless, with a view to identifying the work that would be necessary in order to functionally dissect this association signal, it is worth considering potential mechanisms and functions of these genes, including its relative importance in the wild.

Does TOLL11 stimulate the TOLL pathway?

The TOLL10 and TOLL11 genes identified in this locus form an orthologous group in the mosquito and are the only two toll genes representing a novel gene expansion in the *culicidae* - frequently an indication of species or clade specific functions, and quite possibly associated with a highly prevalent parasite that infects this clade but not other dipterans. Although the TOLL11 gene is paralogous to TOLL1, and now has been implicated in the same metabolic function, it is by no means certain that this receptor stimulates the same pathway. However the involvement of this locus in a TOLL-like response would have major

implications for immunity in the field. It is worth considering, therefore, the possibility that the TOLL11 gene identified here either stimulates or modulates the canonical TOLL pathway. It is notable that there has been there are no similar duplications within the rest of the TOLL pathway. Despite the proliferation of TOLL receptor homologues (10 paralogues are currently annotated within the *An. gambiae* genome), there is no apparent duplication of the downstream components: *MYD88, TUBE, PELLE, TRAF6* and *CACTUS* are *all* maintained as 1:1:1 orthologues between the three sequenced mosquitoes (Waterhouse et al. 2007). The combination of a highly conserved intracellular C-terminal domain and the lack of duplication of the rest of the pathway certainly lends itself to the interpretation that the downstream path for many TOLL receptors is similar to that seen in *TOLL1*. Indeed expression motifs of *dmTOLL9* indicate downstream Toll-pathway activation (Bettencourt et al. 2004) and the expression of chimeric genes (with a Toll1 extracellular domain and Toll5 intracellular) has demonstrated activation of the canonical TOLL pathway via the TOLL5 intracellular domain (Tauszig et al. 2000). However attempts to discern a similar effect using chimeric genes for other TOLL genes were unsuccessful, and similar experiments in the mosquito have yet to be performed. The coordinated decrease in *Cecropin* and *Defensin* expression in TOLL11 knockdowns may also indicate TOLL-pathway stimulation, although further replicates would need to be performed in order to confirm this result.

Anti-*P. falciparum* effect

Despite the incontrovertible effect on the development on *P.falciparum* we have demonstrated in the RNAi-mediated knockdown, is it not possible to say whether the pathway activated here either reacts to, or is directed specifically against *P.falciparum.* Immune responses to bacteria have been previously shown to have an effect on *Plasmodium* development (Gendrin & Christophides 2013), and it is consistent with the work of Frolet *et al.* (Frolet et al. 2006) that this reduction in oocyst prevalence could be achieved by, for example, an increase in the basal expression of TEP1 in reaction to other pathogens encountered by the mosquito prior to the immune challenge with *falciparum.*

Importance of the *TOLL11* response in the wild

Given their homology, it is not entirely surprising that *TOLL11* and *10* might be immune genes, yet they are perhaps not the genes we would have most expected to be identified by a GWAS study. If this locus stimulates or modulates the canonical TOLL pathway, or if it has another non-IMD immune function, in either case this would run counter to prior expectations which have suggested an IMD-pathway dominated immune response to *P. falciparum* (Garver et al. 2009). Though the stimulation of the TOLL pathway should be considered

conjecture, there are good reasons to consider this a viable possibility (see above) and as such it is worth considering why this has not been identified within laboratory colonies.
This is particularly pertinent in light of the prior associations of variants in *TOLL*6 (Harris et al. 2010) and *TOLL*5B (Horton et al. 2010) with *P.falciparum* resistance in the target-gene studies. This is therefore the third time 'semi-wild' colonies have been assayed and the third time a TOLL receptor that has been identified; a fact that may indicate a more significant role for TOLL family in *P.falciparum* infection.
It is a severe weakness of this method that the founder colonies cannot necessarily be said to represent an even sampling of the wild populations, but merely to provide and amalgam to it. As a result we cannot be sure that the haplotypes captured in this colony are of major or minor importance in the real world – that is, whilst alleles can be shown to have statistical significance in the colony, does not necessarily translate to high significance in the wild.

Did we miss an IMD signal?

An IMD-dominated immune response to *P.falciparum* challenge has been widely accepted since investigations by Garver *et al.* (Garver et al. 2009) that involved a direct comparison of IMD/TOLL function in '*Keele'* strain mosquitoes first indicated a dominant role for the IMD pathway in this immune response. Yet neither founder showed significant variation in IMD pathway genes.
It is possible that absence of IMD-related signal in this case is due a lack of significant phenotypic variation in the IMD response; that is, that it remains a strong response, but one that does not vary between individuals. Or that the requisite differences in phenotype were present, but suffered from a lack of associable variation in the IMD immune genes, or from insufficient LD to identify a locus. As the coarse-mapped loci are neither in pericentromeric regions or associated with chromosomal inversions this too would have to be linked to an overall drop in diversity.
However it is also conceivable that the IMD-dominated response seen in inbred lab colonies is of lesser importance in the wild. Differences in the immune response have been noted previously between inbred lab colonies and wild-caught mosqitoes (Cohuet et al. 2006) and it is entirely possible that the repeated detection of *TOLL* receptors is indicative of a greater role for these receptors than has been previously considered.
The majority of laboratory studies use highly inbred mosquitos along with an exceptionally virulent parasite strain or murine malaria to which the vector has poor immunity; as a result they demonstrate high infection intensities. The oocyst counts seen here are far closer to those that would be seen in the wild, where the typical mosquito is uninfected, and our results point to a non-IMD pathway response as underlying the phenotypic variation. The

reasons for the differences in infection intensity between wild and lab mosquitoes are little investigated, but it is certain that mosquito immunity plays a large role.
Comparisons of TOLL/IMD pathways by Garver *et al.* – the study upon which the dominance of the IMD pathway in anti-*Plasmodium* immunity has been based – showed oocyst counts seen in GFP-injected controls (mean: 21, median: 9) that were an order of magnitude greater than our infections (mean 2.2, median 0). Infection intensity may affect the dominance of one immune pathway over any other. Differential transcriptional responses to high and low intensity infections have been shown to occur under both P.falciparum and P.berghei challenge (Mendes et al. 2011) identifying a series of GPRKs with potential for downstream MAPK activation; moreover the IMD response itself has also been demonstrated to be dependent on infection intensity, with the most significant effects recorded at moderate infection intensities (median oocyst no: 7). It is notable in this context that, when performing our knockdowns of *TOLL11*, the replicate with the highest median oocyst load was also the least significant.
Unlike earlier studies combining a highly infectious parasite strain with a highly susceptible mosquito strain, oocyst counts in this experiment approach those found in natural settings. Dominance of a non-*IMD* response at these infection intensities would have significant implications for the study of malaria transmission. Further work will be needed to assess the relative importance of these loci in *Plasmodium* infection

Potential implications of the TOLL11 immune response

Although the TOLL11 / 10 locus cannot be shown to have the same magnitude of effect on oocyst number as IMD, the association of this locus with the immune response to *P.falciparum* may still have significant implications for current practices in malaria parasitology. This is true whether or not it is subsequently shown that the TOLL11 receptor activates the downstream TOLL pathway, or if it is activated by co-infected pathogens. Investigations into the *Anopheles* immune system have typically concentrated on a small number of opsonins and highly specific responses. Conformation of this locus's importance in the wild would suggest that, in fact, a higher diversity of immune receptors is of importance, or that an 'off-target' stimulation of the immune system is of comparable importance to the direct anti-*plasmodium* response.
The IMD pathway portrays a nuanced response to different *Plasmodium* species (Mitri et al. 2009) and to geographically diverse isolates of *Plasmodium* (Molina-Cruz et al. 2012). This response dominates at the levels typically seen in lab colonies, and is more amenable for use in transgenic control methods and has been the subject of numerous functional and genomic investigations. Previous RNAi studies of IMD have shown significant effects on

oocyst count at both high and low intensities (Meister et al. 2005; Garver et al. 2012), indicating a continuing effect in high-intensity infections.
However this does not necessarily mean it represents the major factor determining infection prevalence. In a comprehensive set of knockdowns of IMD pathway components, only the Rel2 inhibitor caspar has been thus far demonstrated to have a high-magnitude effect on prevalence (Garver et al. 2012), exhibiting a mean risk ratio of 0.70 under low and medium-intensity infections. Notably this effect is of smaller magnitude than we have demonstrated for *TOLL11* silencing.
Even if *TOLL11*-mediated immunity is only able to counteract small numbers of gametocytes, and even if it is typically overcome by high-intensity infections, this would not necessarily render it unimportant in the wild. Any immune response effective at low levels will mark the difference between the one-and-zero oocyst counts that are crucial for the effective blocking of transmission.

Acknowledgements:

Although the analysis, interpretation and writing contained within this chapter is entirely that of the author, the work was carried out as part of a large collaborative project. This chapter should therefore also be considered the work of: Karin Eiglmeier[1], Christian Mitri[1], Michelle M. Riehle[2], Wamdaogo M. Guelbeogo[3], Awa Gneme[3], Alison T. Isaacs[1], Boubacar Coulibaly[4], Emma Brito-Fravallo[1], Gareth Maslen[5], Daniel Mead[5], Oumou Niare[4], Sekou F. Traore[4], N'Fale Sagnon[3], Dominic Kwiatkowski[5] and Kenneth D. Vernick[1,2,]*
Mosquito sampling, the establishment of founder lines, and maintenance of these lines was carried out by KE, WMG, AG, ATI, BC and EBF. Founder line infection and phenotyping was carried out by CM and KE. Sequencing of founder lines was carried out by DM and GM. Statistical methods development, data analysis and interpretation was carried out by SNR. Fine-mapping marker design was carried out by SNR and MMR. RNAi knockdowns were carried out by CM. Analysis of knockdown assays was carried out by SNR. Study design was conceived by KV, MMR and SNR

This work received financial support from the European Commission FP7 InfraVec program, the European Research Council project AnoPath, the National Institutes of Health USA, and the Pasteur / Paris University PhD Program.

1 Institut Pasteur, Unit of Insect Vector Genetics and Genomics
2 University of Minnesota, USA.
3 Centre National de Recherche et de Formation sur le Paludisme, Ouagadougou, BURKINA FASO.
4 Malaria Research and Training Centre, University of Mali , Bamako, MALI.
5 Wellcome Trust Sanger Institute and Wellcome Trust Centre for Human Genetics

6: Wider Importance and Implications

I would hope that over the course of this thesis I have managed to elucidate some of the major challenges in malaria genomics and how it relates to malaria control, in particular the necessity of understanding population structure as a prerequisite to either studying immunity or interrupting zoonotic transmission. During the course of this work we have added to the store of knowledge in ways that will have implications for future research, both in techniques developed, and in the linkage of genetic loci to both phenotype and population structure.
We have deduced new and reliable SNP markers for population subdivisions of Anopheline mosquitoes, enabling major structural variants to be typed at a higher degree of accuracy than is available by any alternative method, and in a dataset of unparalleled geographic spread. There are already strong indications within the Ag1kG data that these inversions segregate population and ecotypic subdivisions, and further elucidation of these phenomena will no doubt shed more light on the varying subdivisions of these mosquito populations.
We have also performed the first genome-wide investigation into malarial vector competence in *Anopheles gambiae* / *coluzzii*; assaying the degree of association to infection prevalence and intensity without the pre-selection of markers or the use of known-phenotype crosses. The method exploits known markers of population structure and natural assortative mating to derive colonies that are panmictic and as far as possible a fair representation of the wild poulations. It has resulted in the identification of a two-gene locus with SNPs associated with infection prevalence and intensity, and containing two genes. This has been validated in two separate infections. The two TOLL receptors that are found in this locus, TOLL10 and TOLL11 are part of a known immune family and part of a mosquito-specific gene expansion, yet they have never, until now, been assigned a clear immune function, nor have they been implicated in the anti-*Plasmodium* response. Finally we have functionally tested all of the genes within this locus by dsRNA mediated knockdown, confirming the phenotype of one gene; TOLL11 has been shown to engender a reduction in infection prevalence in wild mosquitoes and for its action to be subject to natural variation in at least one population of mosquitoes from Mali.

6.1: Potential uses of the results:

Inversion karyotype markers

The availability of robust SNP-barcodes for 2Rb and 2La will enable the construction of far larger genomic datasets than has previously been possible, removing the laborious and

restrictive process of polytene chromosome analysis from the workflow required to construct them. Indeed the availability of 765, fully-sequenced, continent-wide samples that are now associated with the two most common karyotypes has already allowed us to examine reproductive isolation between *Anopheles gambiae* and *coluzzii* and between ecotypes within *Anopheles gambiae*. Subsequent investigations into these phenomena, combined with further development of markers for the other five major inversions will no doubt shed further light on the genomic and ecological effects of these inversions.

By utilising the breakpoint duplication as a inversion marker, we have also provided a confirmation that even the smallest copy number polymorphisms can be identified; the normalisation techniques that were used to enable depth comparisons within the breakpoints will also allow CNV detection across the genome – currently a sorely under-examined aspect of variation. Through analysis of the Ag1kG dataset, including karyotype-specific signals of purifying selection and CNVs (that are frequently linked to expression differences), it is hoped that some of the specific mechanisms underpinning the maintenance of these inversions could be discovered.

The 2Rb / 2La karyotyping should also provide a basis for the identification of the remaining five common chromosomal inversions, either by a recapitulation of the same methods, or by providing a large sample set of confirmed karyotypes against which inversion-calling algorithms such as inveRsion (Cáceres et al. 2012) and breakdancer (Chen et al. 2009) can be tested and validated (see chapter 4).

The availability of linked markers allows us to more easily investigate any potential novel phenotypic associations with these inversions, as has been done for the 2La inversion and aridity tolerance. In particular, if the other inversions can be successfully typed, we may be able to examine phenotypic associations of the inversions in isolation from linked inversions such as the 2Rbcu cluster or of epistatic interactions between inversions that are not physically linked, such as the frequently co-occurring 2La and 2Rb.

TOLL11 upstream/downstream pathway dissection

The results of the association study have more specific implications. The implication of TOLL10 and TOLL11 in the antiplasmodial immune response is novel and adds to our understanding of the innate immune response to this parasite. It is also another piece of evidence that, in combination with previous association studies in this species (Harris et al. 2010; Horton et al. 2010), point to a greater role for the family of TOLL receptors in anti-plasmodial immunity. This result suggests and enables a number of potential follow-up investigations in future years.

The stimulation of these TOLL receptors is unknown (as indeed are the ligands for all *AgTOLL* receptors): whilst identifying the particular Spätzle ligand that activates the receptor may be achievable, perhaps the more interesting question might be whether this is a specific response to the *Plasmodium falciparum* parasite, or if the actions of TOLL11 are in response to, for example, bacterial infections and have a corollary effect on the *Plasmodium* parasite. Such effects have been previously seen for microbiota in Anopheline mosquitoes, in almost all cases with gram-negative bacteria (Cirimotich et al. 2011). Particularly interesting in this regard are *Wolbachia spp*. which are now known to be present in natural populations of *Anopheles gambiae* (Baldini et al. 2014) and have been shown to induce refractoriness to *Plasmodium* in *Anopheles stephensi* (Bian et al. 2013).
The downstream actions of TOLL11 are also unknown. Further work will be required to establish if the anti-plasmodial action is effected through stimulation or modification of the canonical TOLL pathway. The TOLL pathway has already been shown to have an effect on *Plasmodium* infections in laboratory colonies – albeit one smaller in magnitude than IMD. In addition, coordinated activation of the *TOLL*-mediated Cec1 and Def1 defensins (both of which were seen to increase in expression under *TOLL11* knockdown) has been shown to interrupt malaria transmission in *Aedes* mosquitos (Kokoza et al. 2010).
Whilst the TOLL10 / 11 locus has been functionally tested, we have not uncovered any specific causative variants as is the case for the *TEP1* R/S alleles (Blandin et al. 2009), or the various pyrethroid-resistance-generating *KDR* mutants (Mathias et al. 2011; Alout et al. 2013). Neither of the two significant SNPs identified during fine mapping is sufficiently close to *TOLL11* to influence its behaviour. Targeted resequencing of this locus may uncover specific coding sequences that are linked to the loss or gain of function in TOLL11 or 10, most likely either by affecting the binding efficiency of the LRR domains, or the downstream signal transduction via the TLR domain. The identification of causative variants, rather than neutral linked loci, would provide simple methods to assess wild function, since the presence of an allele could be defined as definitive of phenotype rather than merely indicative.

Replication of founder-colony-mapped loci

Even if causative variants cannot be identified, replication of these results in wild samples is desirable. Although we have demonstrated that a genetic effect can be detected within the founder colony, we have not shown that this effect is detectable or important in the wild. This is, of course, a far higher standard of proof. Replication of wild effect is much more difficult; the low degree of allele sharing between populations presents a major barrier to replication studies in the wild. But this is important if we are to discover the most important immune reactions in natural conditions.

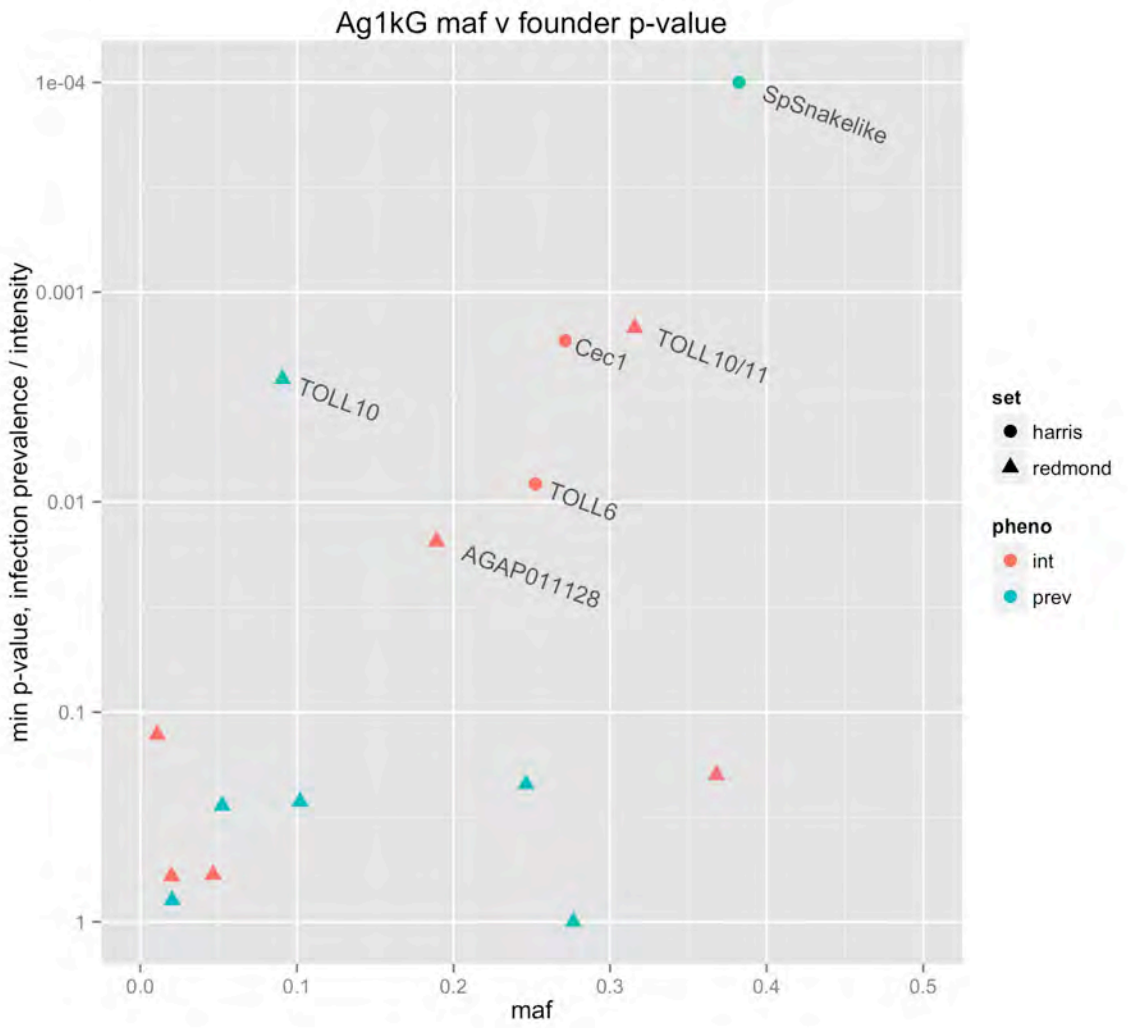
Although replication or direct mapping in the wild is beyond our current capabilities, this is likely to be possible in the near future. Failure of fine-mapping assays in our dataset is likely to derive from sequencing errors, or those markers where undetected variants in flanking sequence impaired binding of Sequenom probes. In addition many primers demonstrated variation, but at a lower than expected allele frequency. Not only are these minor allele frequencies too low to enable statistically significant association themselves, but the presence of the variant still affects the degree of multiple testing correction that is required. The colonisation process upon which this mapping method relies destroys our ability to ascertain true allele frequencies in the wild, and therefore hampers any efforts to chose high quality markers for replication studies.

Yet using data that is currently available, replication of colony-derived associations should be a viable prospect. The availability of a broad-geography sample of variation, the Ag1kG dataset, would allow pre-filtering of markers to select those with the highest probability of being found in novel populations – either as a function of allele frequencies, or their segregation in geographically distinct locations. In effect it is capable of replacing the information lost during colonisation, allowing us to derive a set of unbiased markers that are at a sufficiently high minor allele frequency so that phenotype associations are viable. It is notable in this regard that significant assays in both chapter 5 and the prior association study of Harris *et al.* are also invariably of high MAF (see figure 6.1)

Moreover, the coarse-mapped locus can be searched for linked variation from within the Ag1kG set, providing a set of markers to assay natural variation in this locus in as many populations as possible, whilst reducing the degree of multiple testing introduced by low-frequency variants.

Should variants of sufficiently strong effect be present, the availability of a curated set of markers that are typable, and represented across broad geographic regions may even enable direct mapping in wild colonies via SMFA (Bousema et al. 2012) or other phenotypic assays (see below).

Figure 6.1 : MAF / P-values for Ag1kG markers within the Harris / Redmond sets



6.2 Direct mapping of wild alleles : A viable prospect ?

At the outset of the mapping experiment it was stated that the inherent difficulties encountered within the *Anopheles* / *Plasmodium* infection system, both in terms of their genomics and the inherent noise in a two-system phenotype made wild mapping a distinct improbability.

During the course of this study, and in the course of the *Anopheles* 1000-genomes project, we have not succeeded in dispelling these fears, and in fact the degree of diversity and the near-absence of LD has in fact been shown to be even more severe than was previously thought (Ag1000G Consortium n.d.). Nevertheless the data generated by this project, and the lessons learnt during the founder-mapping study, raise the possibility of performing genomic mapping of malaria transmission directly in the wild.

Genomic mapping from direct wild sampling has, in fact, already been achieved or is underway for other phenotypes that are important factors for vector control (see chapter 1), such as endo/exophily (Fabrigar 2014) and insecticide resistance (Weetman et al. 2010). A crucial difference between both of these studies, however, is the relative simplicity of the phenotypes. Whilst mapping these phenotypes is a significant challenge (in particular the highly complex phenotype of endo/exophily) they are nevertheless mappings of only one genome. Insecticide resistance in particular, due to its recent application and strong selective pressure is frequently associated with clear signals of purifying selection within resistant populations.

The innate immune system, as we saw in chapter 2, has multiple modes of detection, complex pathways and a wide variety of effectors - any one of which might affect the rather blunt measurement of oocyst count. This will inevitably reduce the strength of the association signal, and in systems where a major change has not recently occurred (such as the introduction of a novel strain of the parasite) the genomic signals are not likely to be clear. However attempts to map host-pathogen interactions have the additional complexity of the pathogen genome to consider. Comparative infections of multiple isolates of *Plasmodium falciparum* with multiple isolates of *Anopheles gambiae* have indicated genotype-genotype interactions that drastically change the degree of resistance shown by mosquito isolates (Lambrechts et al. 2005) and infection of local and non-local strains of *Plasmodium* with wild mosquitoes have indicated strong local adaptation of the parasite to the vector – most likely involving specific modes of evasion of the mosquito immune system (Harris et al. 2012). Both of these effects would be near-impossible to predict *a priori*.

All of this indicates that an extremely large number of samples will be required to ensure a successful result. Fabrigar *et al.* have sampled 403 mosquitoes (171 in the GWAS, 232 in the replication) in order to derive associations with endo/exophily (D Fabrigar private communication), and Weetman *et al.* typed more than 1500 mosquitoes in order to re-identify the strongly swept *Kdr* locus (Weetman et al. 2010). Given the lower odds ratios found in our mapping results (Redmond et al. n.d.), and those of Harris *et al.* (Harris et al. 2010) we can expect the required sample sizes to be considerably larger than both of these.

Moreover, given the high degree of population structure demonstrated by the mosquito (see chapter 3) and the expected effect this would have on their relative infectivity to a single strain of *P. falciparum,* it will be vital in any wild mapping attempt that both the phenotype and the population structure are monitored, and where possible, closely controlled.

Wild phenotype measurement:

Mapping insecticide resistance is greatly aided by the existence of long-established methods of phenotyping: standardised assays were developed more than fifty years ago, and (WHO Expert Committee on Insecticides 1963) and the use of the CDC bottle assay and WHO paper assay for testing have been standardised and used in their current form for the past thirty years (WHO 2013; Brogdon & McAllister 1998) giving results that are highly comparable between samples. These methods have differing advantages in the field. For instance, the WHO paper assay requires the purchasing of identical insecticide-impregnated papers from a centralised source - removing a great deal of variation due to operator error, but greatly increasing the cost - whilst the CDC bottle assay is more amenable to implementation at local levels and with novel insecticides, at the cost of some of the replicability of the results.

Attempts to derive a similar standardised test for resistance to *Plasmodium* are in their infancy, though in recent years two tests in particular have emerged for field-based analyses: the direct skin-feeding assay and the direct membrane feeding assay (Bousema et al. 2012). During a meta-analysis of these two techniques (using only experiments where direct comparisons were made using the same infected blood), direct skin feeding was found to show the highest degree of transmission, infecting more than twice the proportion of mosquitoes as membrane feeding (ibid). Clearly, however, this assay would allow little control of gametocyte densities, and no control of *Plasmodium* genotypes. Whilst direct skin feeding is applicable to studies of transmissibility using a single, homogeneous, mosquito colony this technique would not be applicable to a genomic mapping of *Anopheles* immune responses.

The direct membrane feeding assay (DMFA), in which infected blood is extracted and mosquitoes are exposed on a membrane feeder does allow control of both density and genotype. The technique can be implemented using blood directly extracted from infected patients; blood re-suspended in its own serum, allowing control of gametocyte density; or blood re-suspended in heterologous serum, allowing for control of serological effects. Although there is some loss of infectivity after blood extraction (potentially related to premature exflaggelation of the parasite due to the drop in temperature after extraction), and significant variation in infection prevalence is introduced by serological effects (with heterologous serum indicating increased infection prevalences to homologous). Nevertheless DMFA assays demonstrate reduced variability and high correlation to direct feeding assays. DMFA have, in fact, already been used in an association study demonstrating that insecticide resistance alleles affect vector competence (Alout et al. 2013). However it should be noted in this case that the vector populations in this study were closely controlled – with one colony of

susceptible and one colony each that was positive for the *Kdr*+ve and *ace1* +ve insecticide resistance alleles. The technique has yet to be implemented successfully directly for genomic mapping using a wild population.

Similar techniques enable use of a controlled clone (or set of clones) of *P. falciparum:* the so-called 'standard membrane feeding assay' (SMFA) further reduces the effects of genotype-genotype interactions by requiring only the mixing of a cultured stock of *Plasmodium* with fresh human serum (in the same manner as was performed for our functional assays - see section 5). The higher oocyst counts achieved using SMFA, as compared to DMFA, enable the use of oocyst count as a correlate for infection intensity (typically the DMFA will only count prevalence). However comparisons of SMFA using empirical data have shown less-than-optimal reproducibility (van der Kolk et al. 2005); individuals are comparable within replicates, but high variance between replicates means it is difficult to compare quantitative results across experiments. Despite this comparability issue, and following optimisation of the method, this technique is increasingly widely used in assays of transmission-blocking elements and is likely to remain a mainstay of transmission assaying for the foreseeable future (Miura et al. 2013). Nevertheless, regional variability in susceptibility may mean that the effects of local adaptation (Harris et al. 2012) are confounding to any association studies that seek to use a single parasite isolate, and direct comparisons of DMFA with SMFA suggest that there remains further optimisation to be done before the SMFA can be used as an amalgam for the more biologically relevant DFA or DMFA (Nunes et al. 2014). It may be the case that, despite the additional variation introduced when dealing with multiple parasite genotypes, and the resultant need to both genotype and statistically control for parasite genotypes (see below), the DMFA is still the most appropriate assay for surveying in the field. Large scale comparative field trials are yet to be carried out.

Population structure control

While performing these follow-up studies in wild mosquitoes it will not be possible to apply the same degree of prior control of population structure as we were able to during the founder mapping. The founder colonies presented us with the ability to ensure that each colony consisted of a single inversion karyotype and extensive intercrossing ensured that the assayed population was panmictic. Wild populations are highly unlikely to be homokaryotypic in all locations and, given the degree of population structure encountered in the Ag1kG project, it is certain that some degree of population structure will be encountered in wild populations.

In the best case scenario this would be restricted to sympatric *gambiae* / *coluzzii* populations that could be distinguished using the favia-fanello test (Fanello et al. 2002), however

encountering cryptic population structure is also a distinct possibility. Some degree of post-assay control for population structure will be crucial.
At the simplest this could involve the identification of M/S markers and karyotypes after assaying, and only seeking association in populations where these did not vary. Yet even if we simply aimed to identify, or sub-select, homokaryotypic populations, it is difficult to envisage how this can be achieved using the destructive assays of polytene chromosomes. The ability to perform karyotyping post-hoc using SNP genotypes will be crucial for any wild mapping attempt. This is currently limited to the the 2La and 2Rj inversions, but if we are to prevent the kind of type 2 errors seen in our founder line from Burkina Faso (where the 2Rb inversion was present but obscure) this will need to be extended to the other 2R inversions. The identification of typing markers for all seven inversions as illustrated above should be seen as a necessary pre-requisite for wild mapping.
Of course, it is not only the vector genome that will show unpredictable heterogeneity. If the DMFA or DFA assays are used, and if we aim to isolate *Anopheles* immune effects from any confounding genotype-genotype interactions, this will also require some surveying for population structure differences in the parasite (beyond that expected under the divergence-by-distance model). Techniques are already available for this barcoding work that should allow detection of cryptic population structure in any parasite genotypes, and can be implemented using the low levels of parasite genotypes that we would expect within the vector at the time of oocyst assaying (Daniels et al. 2008). Similar techniques to detect population structure in the mosquito are notably absent.

Vector population barcoding

However the development of parasite barcoding can suggest a pathway for the development of similar vector barcoding to allow post-hoc identification of population structure.
The designers of the *P. falciparum* barcode used a set of high-frequency markers identified from a worldwide collection of parasites, with markers exhibiting mean minor allele frequencies of 0.35. Using this panel of just 24 markers, each population that was considered to be unique by other genotyping methods – including 114 samples from a worldwide dataset genotyped by genome-wide microarrays – also was able to generate a unique SNP barcode (Daniels et al. 2008). In the six years since its publication, this simple tool has facilitated numerous further analyses of complex infections, and mapping studies in the parasite (Van Tyne et al. 2011; Park et al. 2012).
Due to the much larger genome and greater positional heterogeneity of the *Anopheles* genome, it will not be sufficient to take a random selection of high-frequency markers and hope they illustrate all cases of reproductive isolation, instead we would need to identify

frequent markers that can identify large linked blocks that are expected to segregate differently in sympatric, parapatric and allopatric populations.
Inversions are an obvious starting point for the development of such a genotyping barcode in *Anopheles*. As it stands, they are the only markers known to be in high LD and to segregate commonly in populations. In the Ag1kG dataset just two inversions were sufficient to identify which sympatric and parapatric populations were panmictic and which were not (that is, in all cases in which we saw sympatric reproductive isolation, we also saw inversion karyotypes that were not in Hardy-Weinberg equilibrium). We might therefore expect that the identification of typing SNPs for the remaining five common inversions would identify the majority of population-level segregations. A panel of these markers, along with a subset of centromeric markers to identify species level segregations (White et al. 2010; Caputo et al. 2011), and a set of high-frequency markers in non-inverted regions (i.e. chromosome 3) for detecting fine-scale segregations, would represent a panel of SNPs with sufficient resolution that they could act as a barcode for detecting population segregation at all levels. Using this barcode it may then be possible to perform phenotype assays only from within populations represented by a single barcode, avoiding the deleterious effects of inversions on association.

Identification of heterokaryotypic populations

Though the use of homogenous colonies would ameliorate some of the statistical difficulties in a wild association study, it is possible that when typing a population which shows a large degree of karyotypic hetereogeneity the selection of homokaryotes will have the effect of drastically reducing the effective population size – not only reducing power to resolve loci, but reducing the number of wild alleles that are sampled. However typing of all seven inversions may enable us to make a more nuanced identification of panmictic populations. Instead of keeping the assayed populations karyotypically homogeneous, as we did in the Fd03 and Fd09 lines, in cases where a population was known to be panmictic and segregated an inversion, we could chose assay populations which were predictably heterogeneous.
It is near certain that – for most all situations – there will be no single definitive inversion that segregates populations, and population boundaries are likely to be porous. However, if the patterns seen in the reproductively isolated populations in Burkina Faso and Cameroon in the Ag1kG data are typical, it seems almost certain that particular proportions or

combinations of inversion karyotypes will be able to classify mosquitoes into true panmictic populations with a high degree of accuracy.
Effectively we would aim to recapitulate the work of Coluzzi *et al.* (Coluzzi et al. 1985; Coluzzi et al. 1979) in defining isolated populations by their inversion karyotype. Whilst some of these populations would be homokaryotypic, we should also be able to identify and assay important heterokaryotypic panmictic populations such as the highly polymorphic and cosmopolitan 'Savanna' form (see section 3). The machine learning approaches that were able to discern the duplication within the breakpoints would be highly applicable to such a non-linear classification. For instance, after training on a set of known chromosomal forms, or another population from a single location and in Hardy-Weinberg equilibrium, an SVC should be able to classify a smaller sampling of similar karyotypes as being of one of those forms, or an admixture of more than one.

Genotyping

The selection of typing markers can be fraught with problems. Selecting only frequent markers, or those in the highest linkage is likely to identify only those that represent the oldest population subdivisions, such as *gambiae* / *coluzzii* speciation, or the oldest of the inversions. However the inclusion of all of the markers can lead to impractically high p-values being required as a result of multiple testing. Where association studies are concerned less is frequently more, however it is important to avoid ascertainment bias in the selection of markers.
The genotyping method itself will govern which markers are available, with technologies such as the AgSNP01 chip (Neafsey et al. 2010), a 400k affymetrix genotyping chip, providing genome-wide coverage of frequent alleles, but with a significant enrichment for alleles that distinguish *gambiae* and *coluzzii* forms (Ag1000G Consortium n.d.). Alternatively individual resequencing using, eg, illumina sequencing will ensure that all alleles are assayed, but at exhorbitant cost – greatly reducing the sample sizes that can be genotyped.
Two alternative approaches can reduce the cost of sequence-based genotyping. Positional coverage of the genome can be reduced, enabling larger numbers of samples to be multiplexed in a single run. This is primarily either exome sequencing, in which probes baited with coding sequence are used in order to capture exons specifically from the non-coding background (Bamshad et al. 2011). As a result of this 'exome capture' only open reading frames would be assayed.
Although this technique is more likely to discover the causative variants behind loss-of-function mutants (at the significant cost of any regulatory variants of importance), and is cost-effective on an individual basis, the development of exome-capture probes is expensive and

time consuming. It is likely to be for this reason that this technique has not been implemented in *Anopheles*.

A second option for restricted-coverage genotyping is RAD-tag sequencing (Davey et al. 2011), in which genomic DNA is digested with a small number of restriction enzymes and their flanking sequences sequenced. This technique relies upon markers being in linkage with the putative causative marker, something that might present a challenge in the extremely low LD environment of the *Anopheles* genome, although it may be achievable at a local level. Caveats are also suggested as regards polymorphisms in restriction target sites, which would bias the representation of some of the RAD markers (the importance of which would scale both linearly with the number of restriction sites and the expected LD). In general both technologies enable higher coverage to be achieved per sample than is possible with genomic sequence and would permit individual sequence to be used at an achievable cost.

The second alternative is via pooling of samples as we performed for the founder lines. This ensures that all regions of the genome are assayed, and can reduce costs significantly. However it does not allow the same use of individual level associations as would be possible with RAD-tag, exome, or whole-genome individual sequencing. It should also be noted that the coarse mapping technique applied in section 5 would be wholly inappropriate for wild populations, in which the panmixia of the wild population could not be ensured.

A recent review by Schlotterer *et al.* (Schlötterer et al. 2014) has elucidated some of the relative advantages of these methods, however care should be taken with their conclusions regarding pooled sequencing. What testing has been carried out with pooled sequencing has typically been performed using inbred drosophila lines such as the Drosophila Genetic Reference Panel (Mackay et al. 2012; Zhu et al. 2012). Comparison of these resources to our own founder lines demonstrated a far lower degree of correlation between pooled sequence genotypes in *Anopheles* as compared to *Drosphila*, and neither of these are comparable to wild mosquitoes. Given the uneven representation of low-frequency alleles, the high numbers of sequencing errors and the inaccessibility of many regions of the genome, the conclusions of Schlotterer *et al.*, that "*pool-seq provides more accurate allele frequency estimation at a lower cost than sequencing of individuals*" (Schlötterer et al. 2014), seem ambitious in the extreme – (particularly where effective population sizes are sufficiently low that individuals could feasibly be sequenced, as was the case in founder colonies). Perhaps more importantly, even if pools were sequenced to high frequency (i.e. significantly greater than 100X coverage) – which would enable sequencing errors to be discerned from low-frequency alleles (Nielsen et al. 2011), it is not necessarily an advantage to identify all variants in the sample. Heterozygosity in *Anopheles gambiae* / *coluzzii* is extreme, and the vast majority of alleles are of extremely low frequency. The lack of power seen in both the Harris *et al.* and Redmond *et al.* datasets when low-MAF alleles were genotyped may

instead suggest the use of panels of high frequency, accessible SNPs that could be typed using any of the above methods. The publication of the Ag1kG dataset should enable an optimal panel of markers to be devised.

Summary

There are clearly a large number of challenges still to be met before association studies can be performed in wild populations. A true power analysis is difficult to envisage until a larger amount of DMFA sampling has been performed; typing panels within the accessible genome have not been identified; and replication of colony-derived associations (a far less challenging experiment) has not been performed.
But with increase in knowledge of both the genomic and population structure of the mosquito, the refinement of the membrane-feeding assays, and the continuing falling cost of genotyping, this may be possible in coming years. We would hope that the work undertaken in this thesis could have contributed some small part of the knowledge that would make this possible.

Bibliography

Introduction

Dondorp, A.M. et al., 2010. Artemisinin resistance: current status and scenarios for containment. Nature reviews. Microbiology, 8(4), pp.272–80.

Gallup, J. & Sachs, J., 2001. The economic burden of malaria. Am J Trop Med Hyg, 64(1_suppl), pp.85–96.

WHO, 2013. WHO | Factsheet on the World Malaria Report 2013. Factsheet on the World Malaria Report 2013.

Worrall, E., Basu, S. & Hanson, K., 2005. Is malaria a disease of poverty? A review of the literature. Tropical medicine & international health : TM & IH, 10(10), pp.1047–59.

Chapter 1

Ahoua Alou, L.P. et al., 2012. Status of pyrethroid resistance in Anopheles gambiae s. s. M form prior to the scaling up of Long Lasting Insecticidal Nets (LLINs) in Adzopé, Eastern Côte d'Ivoire. *Parasites & vectors*, 5(1), p.289.

Alonso, P.L. et al., 2011. A research agenda to underpin malaria eradication. *PLoS medicine*, 8(1), p.e1000406.

Bayoh, M.N. et al., 2010. Anopheles gambiae: historical population decline associated with regional distribution of insecticide-treated bed nets in western Nyanza Province, Kenya. *Malaria journal*, 9, p.62.

Beeman, R.W., Friesen, K.S. & Denell, R.E., 1992. Maternal-effect selfish genes in flour beetles. *Science (New York, N.Y.)*, 256(5053), pp.89–92.

Beier, J.C. et al., 2008. Integrated vector management for malaria control. *Malaria journal*, 7 Suppl 1, p.S4.

Bian, G. et al., 2013. Wolbachia invades Anopheles stephensi populations and induces refractoriness to Plasmodium infection. *Science (New York, N.Y.)*, 340(6133), pp.748–51.

Brogdon, W.G., McAllister, J.C. & Vulule, J., 1997. Heme peroxidase activity measured in single mosquitoes identifies individuals expressing an elevated oxidase for insecticide resistance. *Journal of the American Mosquito Control Association*, 13(3), pp.233–7.

Brown, A., 2002. Personal experiences in the malaria eradication campaign 1955-1962. *Journal of the Royal Society of Medicine*, 95(3), pp.154–6.

Brown, A.W., 1986. Insecticide resistance in mosquitoes: a pragmatic review. *Journal of the American Mosquito Control Association*, 2(2), pp.123–40.

Chandre, F. et al., 1999. Status of pyrethroid resistance in Anopheles gambiae sensu lato. *Bulletin of the World Health Organization*, 77(3), pp.230–4.

Charlwood, J.D. & Dagoro, H., 1989. Collateral effects of bednets impregnated with permethrin against bedbugs (Cimicidae) in Papua New Guinea. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 83(2), p.261.

Chen, C.-H. et al., 2007. A Synthetic Maternal-Effect Selfish Genetic Element Drives Population Replacement in Drosophila. *Science*, 316(5824), pp.597–600.

Cohen, J.M. et al., 2012. Malaria resurgence: a systematic review and assessment of its causes. *Malaria journal*, 11(1), p.122.

contracostamosquito.com, Contra Costa Mosquito & Vector Control District. Available at: http://www.contracostamosquito.com/ [Accessed May 26, 2014].

Coosemans, M. & Carnevale, P., 1995. Malaria vector control: a critical review of chemical methods and insecticides. *Annales de la Société Belge de Médecine Tropicale*, 75(3), pp.13–31.

Coulibaly, M.B. et al., 2007. PCR-based karyotyping of Anopheles gambiae inversion 2Rj identifies the BAMAKO chromosomal form. *Malar J*, 6, p.133.

Curtis, C.F., 1990. Appropriate technology in vector control.

Diabate, A. et al., 2005. Larval development of the molecular forms of Anopheles gambiae (Diptera: Culicidae) in different habitats: a transplantation experiment. *J Med Entomol*, 42(4), pp.548–553.

Dialynas, E. et al., 2009. MIRO and IRbase: IT tools for the epidemiological monitoring of insecticide resistance in mosquito disease vectors. *PLoS Negl Trop Dis*, 3(6), p.e465.

Dieter, K.L., Huestis, D.L. & Lehmann, T., 2012. The effects of oviposition-site deprivation on Anopheles gambiae reproduction. *Parasites & vectors*, 5, p.235.

Eckhoff, P., 2013. Mathematical models of within-host and transmission dynamics to determine effects of malaria interventions in a variety of transmission settings. *The American journal of tropical medicine and hygiene*, 88(5), pp.817–27.

Fanello, C. et al., 2003. The pyrethroid knock-down resistance gene in the Anopheles gambiae complex in Mali and further indication of incipient speciation within An. gambiae s.s. *Insect Mol Biol*, 12(3), pp.241–245.

Fletcher, M., Teklehaimanot, A. & Yemane, G., 1992. Control of mosquito larvae in the port city of Assab by an indigenous larvivorous fish, Aphanius dispar. *Acta Tropica*, 52(2-3), pp.155–166.

Frentiu, F.D. et al., 2014. Limited dengue virus replication in field-collected Aedes aegypti mosquitoes infected with Wolbachia. M. J. Turell, ed. *PLoS neglected tropical diseases*, 8(2), p.e2688.

Gerold, J.L., 1977. Evaluation of some parameters of house-leaving behaviour of Anopheles gambiae s.l. *Acta Leidensia*, 45, pp.79–90.

Gimnig, J.E. et al., 2001. Characteristics of Larval Anopheline (Diptera: Culicidae) Habitats in Western Kenya. *Journal of Medical Entomology*, 38(2), pp.282–288.

Gorrochotegui-Escalante, N., Fernandez-Salas, I. & Gomez-Dantes, H., 1998. Field evaluation of Mesocyclops longisetus (Copepoda: Cyclopoidea) for the control of larval Aedes aegypti (Diptera Culicidae) in northeastern Mexico. *Journal of medical entomology*, 35(5), pp.699–703.

Griffin, J.T. et al., 2010. Reducing Plasmodium falciparum malaria transmission in Africa: a model-based evaluation of intervention strategies. S. Krishna, ed. *PLoS medicine*, 7(8), p.17.

Gu, W. & Novak, R.J., 2005. Habitat-based modeling of impacts of mosquito larval interventions on entomological inoculation rates, incidence, and prevalence of malaria. *The American journal of tropical medicine and hygiene*, 73(3), pp.546–52.

Hargreaves, K. et al., 2000. Anopheles funestus resistant to pyrethroid insecticides in South Africa. *Medical and Veterinary Entomology*, 14(2), pp.181–189.

Harris, A.F. et al., 2012. Successful suppression of a field mosquito population by sustained release of engineered male mosquitoes. *Nature biotechnology*, 30(9), pp.828–30.

Hay, S.I. et al., 2010. Developing global maps of the dominant anopheles vectors of human malaria. *PLoS medicine*, 7(2), p.e1000209.

Helinski, M.E.H., Parker, A.G. & Knols, B.G.J., 2009. Radiation biology of mosquitoes. *Malaria journal*, 8 Suppl 2(Suppl 2), p.S6.

Hemingway, J., 1983a. Biochemical studies on malathion resistance in Anopheles arabiensis from Sudan. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 77(4), pp.477–480.

Hemingway, J., 1982. The biochemical nature of malathion resistance in Anopheles stephensi from Pakistan. *Pesticide Biochemistry and Physiology*, 17(2), pp.149–155.

Hemingway, J., 1983b. The genetics of malathion resistance in Anopheles stephensi from Pakistan. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 77(1), pp.106–108.

Hemingway, J., Miyamoto, J. & Herath, P.R.J., 1991. A possible novel link between organophosphorus and DDT insecticide resistance genes in Anopheles: Supporting evidence from fenitrothion metabolism studies. *Pesticide Biochemistry and Physiology*, 39(1), pp.49–56.

Hemingway, J. & Ranson, H., 2000. Insecticide resistance in insect vectors of human disease. *Annual review of entomology*, 45, pp.371–91.

Hoffmann, A.A. et al., 2011. Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission. *Nature*, 476(7361), pp.454–7.

Holm, I. et al., 2012. Diverged alleles of the Anopheles gambiae leucine-rich repeat gene APL1A display distinct protective profiles against Plasmodium falciparum. K. Michel, ed. *PloS one*, 7(12), p.e52684.

Van den Hurk, A.F. et al., 2012. Impact of Wolbachia on infection with chikungunya and yellow fever viruses in the mosquito vector Aedes aegypti. *PLoS neglected tropical diseases*, 6(11), p.e1892.

Karch, S. et al., 1992. Efficacy of Bacillus sphaericus against the malaria vector Anopheles gambiae and other mosquitoes in swamps and rice fields in Zaire. *Journal of the American Mosquito Control Association*, 8(4), pp.376–80.

Kawada, H. et al., 2011. Multimodal pyrethroid resistance in malaria vectors, Anopheles gambiae s.s., Anopheles arabiensis, and Anopheles funestus s.s. in western Kenya. P. L. Oliveira, ed. *PloS one*, 6(8), p.e22574.

Killeen, G.F., 2013. A second chance to tackle African malaria vector mosquitoes that avoid houses and don't take drugs. *The American journal of tropical medicine and hygiene*, 88(5), pp.809–16.

Kirby, M.J. & Lindsay, S.W., 2009. Effect of temperature and inter-specific competition on the development and survival of Anopheles gambiae sensu stricto and An. arabiensis larvae. *Acta tropica*, 109(2), pp.118–23.

Klassen, W., 2009. Introduction: development of the sterile insect technique for African malaria vectors. *Malaria journal*, 8 Suppl 2, p.I1.

Klassen, W. & Curtis, C.F., 2005. *History of the Sterile Insect Technique* V. A. Dyck, J. Hendrichs, & A. S. Robinson, eds., Berlin/Heidelberg: Springer-Verlag.

Kweka, E.J. et al., 2012. Anopheline larval habitats seasonality and species distribution: a prerequisite for effective targeted larval habitats control programmes. J. F. Turrens, ed. *PloS one*, 7(12), p.e52084.

Lehmann, T. & Diabate, A., 2008. The molecular forms of Anopheles gambiae: a phenotypic perspective. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 8(5), pp.737–46.

Livadas, G.A., 1952. is it necessary to continue indefinitely DDT residual spraying programmes? Available at: https://extranet.who.int/iris/restricted/handle/10665/64195 [Accessed April 20, 2014].

Lockwood, J.A., 2008. *Six-Legged Soldiers*,

Lyimo, E.O., Takken, W. & Koella, J.C., 1992. Effect of rearing temperature and larval density on larval survival, age at pupation and adult size of Anopheles gambiae. *Entomologia Experimentalis et Applicata*, 63(3), pp.265–271.

Majori, G., Ali, A. & Sabatinelli, G., 1987. Laboratory and field efficacy of Bacillus thuringiensis var. Israelensis and Bacillus sphaericus against Anopheles gambiae s.l. and Culex quinquefasciatus in Ouagadougou, Burkina Faso. *Journal of the American Mosquito Control Association*, 3(1), pp.20–5.

Marinotti, O. et al., 2013. Development of a population suppression strain of the human malaria vector mosquito, Anopheles stephensi. *Malaria journal*, 12(1), p.142.

Marten, G.G. et al., 1989. Natural control of larval Anopheles albimanus (Diptera: Culicidae) by the predator Mesocyclops (Copepoda: Cyclopoida). *Journal of medical entomology*, 26(6), pp.624–7.

Marten, G.G., Bordes, E.S. & Nguyen, M., 1994. Use of cyclopoid copepods for mosquito control. , pp.491–496.

Mattingly, P.F., 1962. Mosquito behaviour in relation to disease eradication programmes. *Annual review of entomology*, 7, pp.419–36.

Maull, E.A. et al., 2012. Evaluation of the association between arsenic and diabetes: a National Toxicology Program workshop review. *Environmental health perspectives*, 120(12), pp.1658–70.

Mbogo, C.N. et al., 1996. The impact of permethrin-impregnated bednets on malaria vectors of the Kenyan coast. *Medical and veterinary entomology*, 10(3), pp.251–9.

Melander, A.L. & Experiment, A., 1914. Can insects become resistant to sprays? *Journal of Economic Entomology*, 7, pp.167–173.

Mereta, S.T. et al., 2013. Physico-chemical and biological characterization of anopheline mosquito larval habitats (Diptera: Culicidae): implications for malaria control. *Parasites & vectors*, 6(1), p.320.

Midega, J.T. et al., 2007. Estimating dispersal and survival of Anopheles gambiae and Anopheles funestus along the Kenyan coast by using mark-release-recapture methods. *Journal of medical entomology*, 44(6), pp.923–9.

Mitri, C. et al., 2009. Fine pathogen discrimination within the APL1 gene family protects Anopheles gambiae against human and rodent malaria species. D. S. Schneider, ed. *PLoS pathogens*, 5(9), p.e1000576.

Mittal, P.K., Biolarvicides in vector control: challenges and prospects. *Journal of vector borne diseases*, 40(1-2), pp.20–32.

Mohamed, A.A., 2003. Study of larvivorous fish for malaria vector control in Somalia, 2002. *Eastern Mediterranean health journal = La revue de santé de la Méditerranée orientale = al-Majallah al-ṣiḥḥīyah li-sharq al-mutawassiṭ*, 9(4), pp.618–26.

Moiroux, N. et al., 2012. Changes in Anopheles funestus biting behavior following universal coverage of long-lasting insecticidal nets in Benin. *The Journal of infectious diseases*, 206(10), pp.1622–9.

mosquitoes.org, 2014. www.mosquitoes.org -- The Alameda County Mosquito Abatement District. Available at: http://www.mosquitoes.org/ [Accessed May 26, 2014].

Mwangangi, J.M. et al., 2013. The role of Anopheles arabiensis and Anopheles coustani in indoor and outdoor malaria transmission in Taveta District, Kenya. *Parasites & vectors*, 6(1), p.114.

Nájera, J.A., González-Silva, M. & Alonso, P.L., 2011. Some lessons for the future from the Global Malaria Eradication Programme (1955-1969). *PLoS medicine*, 8(1), p.e1000412.

Okie, S., 2008. A New Attack on Malaria. *New England Journal of Medicine*. Available at: http://www.nejm.org/doi/full/10.1056/NEJMp0803483 [Accessed June 15, 2014].

Omlin, F.X. et al., 2007. Anopheles Gambiae Exploits the Treehole Ecosystem in Western Kenya: A New Urban Malaria Risk? *American Journal of Tropical Medicine and Hygiene*, 77(6).

Pates, H. & Curtis, C., 2005. Mosquito behavior and vector control. *Annual review of entomology*, 50, pp.53–70.

Prapanthadara, L.A. & Ketterman, A.J., 1993. Qualitative and quantitative changes in glutathione S-transferases in the mosquito Anopheles gambiae confer DDT-resistance. *Biochemical Society transactions*, 21 (Pt 3)(3), p.304S.

Protopopoff, N. et al., 2013. High level of resistance in the mosquito Anopheles gambiae to pyrethroid insecticides and reduced susceptibility to bendiocarb in north-western Tanzania. *Malaria journal*, 12, p.149.

Reddy, M.R. et al., 2011. Outdoor host seeking behaviour of Anopheles gambiae mosquitoes following initiation of malaria vector control on Bioko Island, Equatorial Guinea. *Malaria journal*, 10(1), p.184.

Riehle, M.M. et al., 2011. A cryptic subgroup of Anopheles gambiae is highly susceptible to human malaria parasites. *Science*, 331(6017), pp.596–598.

Roberts, L. & Enserink, M., 2007. Malaria. Did they really say ... eradication? *Science (New York, N.Y.)*, 318(5856), pp.1544–5.

RollBackMalaria, 2014. Global Malaria Action Plan: Elimination and Eradication: Achieving Zero Transmission. Available at: http://www.rollbackmalaria.org/gmap/2-3.html [Accessed May 30, 2014].

Rose, W.H., 1981. An Evaluation of Entomological Warfare as a Potential Danger to the United States and European NATO Nations. *U.S. Army Test and Evaluation Command*. Available at: http://www.thesmokinggun.com/file/attack-killer-mosquitoes-0?page=1 [Accessed May 28, 2014].

Rozendaal, J.A., 1989. Biting and resting behavior of Anopheles darlingi in the Suriname rainforest. *Journal of the American Mosquito Control Association*, 5(3), pp.351–8.

Russell, T.L. et al., 2013. Successful malaria elimination strategies require interventions that target changing vector behaviours. *Malaria journal*, 12(1), p.56.

Sabatinelli, G. et al., 1991. [Impact of the use of larvivorous fish Poecilia reticulata on the transmission of malaria in FIR of Comoros]. *Annales de parasitologie humaine et comparée*, 66(2), pp.84–8.

Sanford, M.R. et al., 2013. A preliminary investigation of the relationship between water quality and Anopheles gambiae larval habitats in western Cameroon. *Malaria journal*, 12(1), p.225.

scc.gov.org, 2014. Mosquitofish - Vector Control District - County of Santa Clara. Available at: http://www.sccgov.org/sites/vector/Pages/mosquitofish.aspx [Accessed May 26, 2014].

Scholte, E.-J. et al., 2005. An entomopathogenic fungus for control of adult African malaria mosquitoes. *Science (New York, N.Y.)*, 308(5728), pp.1641–2.

Scholte, E.-J., Knols, B.G.J. & Takken, W., 2004. Autodissemination of the entomopathogenic fungus Metarhizium anisopliae amongst adults of the malaria vector Anopheles gambiae s.s. *Malaria journal*, 3(1), p.45.

Schuler, M.A., 2011. P450s in plant-insect interactions. *Biochimica et biophysica acta*, 1814(1), pp.36–45.

Seligman, L.M. et al., 2002. Mutations altering the cleavage specificity of a homing endonuclease. *Nucleic acids research*, 30(17), pp.3870–9.

Service, U.S.A.M. & Office, U.S.S.-G., 1963. *Preventive medicine in World War II.*, Office of the Surgeon General, Dept. of the Army.

Shiff, C., 2002. Integrated Approach to Malaria Control Integrated Approach to Malaria Control. , 15(2).

Shousha, A.T., 1948. Species-eradication: The Eradication of Anopheles gambioe from Upper Egypt, 1942-1945. *Bulletin of the World Health Organization*, 1(2), pp.309–52.

Sinh Nam, V. et al., 2012. Community-based control of Aedes aegypti by using Mesocyclops in southern Vietnam. *The American journal of tropical medicine and hygiene*, 86(5), pp.850–9.

Sinkins, S.P. & Gould, F., 2006. Gene drive systems for insect disease vectors. *Nature reviews. Genetics*, 7(6), pp.427–35.

Soper, F.L. & Wilson, D.B., 1943. Anopheles gambiae in Brazil 1930 to 1940. , (91/4 x61/4 ins).

Sreenivasamurthy, S.K. et al., 2013. A compendium of molecules involved in vector-pathogen interactions pertaining to malaria. *Malaria journal*, 12(1), p.216.

Sundararaman, S., 1958. The behaviour of A. sundaicus Rodenwaldt in relation to the application of residual insecticides in Tjilatjap, Indonesia. *Indian journal of malariology*, 12(2), pp.129–56.

Tanner, M. & Savigny, D. de, Malaria eradication back on the table. *Bulletin of the World Health Organization*, 86(2), p.82.

Taylor, P. & Mutambu, S.., 1986. A review of the malaria situation in Zimbabwe with special reference to the period 1972–1981. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 80(1), pp.12–19.

"The malERA Consultative Group on Vector Control," 2011. A research agenda for malaria eradication: vector control. *PLoS medicine*, 8(1), p.e1000401.

Trapido, H., 1954. Recent experiments on possible resistance to DDT by Anopheles albimanus in Panama. *Bulletin of the World Health Organization*, 11(4-5), pp.885–9.

Vulule, J.M. et al., 1994. Reduced susceptibility of Anopheles gambiae to permethrin associated with the use of permethrin-impregnated bednets and curtains in Kenya. *Medical and veterinary entomology*, 8(1), pp.71–5.

Walker, K. & Lynch, M., 2007. Contributions of Anopheles larval control to malaria suppression in tropical Africa: review of achievements and potential. *Medical and veterinary entomology*, 21(1), pp.2–21.

Werren, J.H., Baldo, L. & Clark, M.E., 2008. Wolbachia: master manipulators of invertebrate biology. *Nature reviews. Microbiology*, 6(10), pp.741–51.

White, M.T. et al., 2011. Modelling the impact of vector control interventions on Anopheles gambiae population dynamics. *Parasites & vectors*, 4(1), p.153.

White, N. et al., 1999. Averting a malaria disaster. *The Lancet*, 353(9168), pp.1965–1967.

WHO, 2014a. The technical basis for coordinated action against insecticide resistance: preserving the effectiveness of modern malaria vector control: meeting report. *2011*. Available at: http://apps.who.int/iris/handle/10665/44526 [Accessed April 20, 2014].

WHO, 2014b. WHO | Malaria control: the power of integrated action.

WHO, 1976. WHO-supported collaborative research projects in India: the facts. *WHO chronicle*, 30(4), pp.131–9.

Windbichler, N. et al., 2011. A synthetic homing endonuclease-based gene drive system in the human malaria mosquito. *Nature*, 473(7346), pp.212–5.

Wu, K.C. et al., 1993. [Studies on distribution and behavior of Anopheles minimus and its role of malaria transmission in Hainan Province at present]. *Zhongguo ji sheng chong xue yu ji sheng chong bing za zhi = Chinese journal of parasitology & parasitic diseases*, 11(2), pp.120–3.

Yakob, L., 2011. Epidemiological consequences of a newly discovered cryptic subgroup of Anopheles gambiae. *Biology letters*, 7(6), pp.947–9.

Yen, J.H. & Barr, A.R., 1971. New hypothesis of the cause of cytoplasmic incompatibility in Culex pipiens L. *Nature*, 232(5313), pp.657–8.

Zhen, T.M., Jennings, C.D. & Kay, B.H., 1994. Laboratory studies of desiccation resistance in Mesocyclops (Copepoda: Cyclopoida). *Journal of the American Mosquito Control Association*, 10(3), pp.443–6.

Zieske, F., 1995. An Investigation of Paul Cézanne's Watercolors With Emphasis on Emerald Green.

De Zulueta, J., 1959. Insecticide resistance in Anopheles sacharovi. *Bulletin of the World Health Organization*, 20, pp.797–822.

Chapter 2

Agaisse, H. et al., 2003. Signaling Role of Hemocytes in Drosophila JAK/STAT-Dependent Response to Septic Injury. *Developmental Cell*, 5(3), pp.441–450.

Agaisse, H. & Perrimon, N., 2004. The roles of JAK/STAT signaling in Drosophila immune responses. *Immunological Reviews*, 198(1), pp.72–82.

Barillas-Mury, C., 2007. CLIP proteases and Plasmodium melanization in Anopheles gambiae. *Trends in parasitology*, 23(7), pp.297–9.

Bian, G. et al., 2005. Transgenic alteration of Toll immune pathway in the female mosquito Aedes aegypti. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38), pp.13568–73.

Blandin, S. et al., 2004. Complement-Like Protein TEP1 Is a Determinant of Vectorial Capacity in the Malaria Vector Anopheles gambiae. *Cell*, 116(5), pp.661–670.

Blandin, S. a, Marois, E. & Levashina, E. a, 2008. Antimalarial responses in Anopheles gambiae: from a complement-like protein to a complement-like pathway. *Cell host & microbe*, 3(6), pp.364–74.

Blandin, S.A. et al., 2009. Dissecting the genetic basis of resistance to malaria parasites in Anopheles gambiae. *Science*, 326(5949), pp.147–150.

Blandin, S.A., Marois, E. & Levashina, E.A., 2008. Antimalarial responses in Anopheles gambiae: from a complement-like protein to a complement-like pathway. *Cell host & microbe*, 3(6), pp.364–74.

Boutros, M., Agaisse, H. & Perrimon, N., 2002. Sequential activation of signaling pathways during innate immune responses in Drosophila. *Developmental cell*, 3(5), pp.711–22.

Brennan, C.A. & Anderson, K. V, 2004. Drosophila: the genetics of innate immune recognition and response. *Annual review of immunology*, 22, pp.457–83.

Castillo, J.C., Robertson, A.E. & Strand, M.R., 2006. Characterization of hemocytes from the mosquitoes Anopheles gambiae and Aedes aegypti. *Insect biochemistry and molecular biology*, 36(12), pp.891–903.

Cerenius, L. & Söderhäll, K., 2004. The prophenoloxidase-activating system in invertebrates. *Immunological reviews*, 198, pp.116–26.

Christophides, G.G.K. et al., 2002. Immunity-related genes and gene families in Anopheles gambiae. *Science (New York, N.Y.)*, 159(2002), pp.159–65.

Cohuet, A. et al., 2006. Anopheles and Plasmodium: from laboratory models to natural systems in the field. *EMBO reports*, 7(12), pp.1285–9.

Das, S. et al., 2009. Specificity of the innate immune system: a closer look at the mosquito pattern-recognition receptor repertoire. In *Insect Infection and Immunity: Evolution, Ecology, and Mechanisms (Google eBook)*. Oxford University Press, p. 266.

Dong, Y., Aguilar, R., et al., 2006. Anopheles gambiae immune responses to human and rodent Plasmodium parasite species. *PLoS pathogens*, 2(6), p.e52.

Dong, Y., Taylor, H.E. & Dimopoulos, G., 2006. AgDscam, a hypervariable immunoglobulin domain-containing receptor of the Anopheles gambiae innate immune system. D. Schneider, ed. *PLoS biology*, 4(7), p.e229.

Frolet, C. et al., 2006. Boosting NF-kappaB-dependent basal immunity of Anopheles gambiae aborts development of Plasmodium berghei. *Immunity*, 25(4), pp.677–85.

Garver, L.S. et al., 2012. Anopheles Imd pathway factors and effectors in infection intensity-dependent anti-Plasmodium action. D. S. Schneider, ed. *PLoS pathogens*, 8(6), p.e1002737.

Garver, L.S., de Almeida Oliveira, G. & Barillas-Mury, C., 2013. The JNK pathway is a key mediator of Anopheles gambiae antiplasmodial immunity. K. Deitsch, ed. *PLoS pathogens*, 9(9), p.e1003622.

Garver, L.S., Dong, Y. & Dimopoulos, G., 2009. Caspar controls resistance to Plasmodium falciparum in diverse anopheline species. *PLoS pathogens*, 5(3), p.e1000335.

Garver, L.S., Xi, Z. & Dimopoulos, G., 2008. Immunoglobulin superfamily members play an important role in the mosquito immune system. *Developmental and comparative immunology*, 32(5), pp.519–31.

Gobert, V. et al., 2003. Dual activation of the Drosophila toll pathway by two pattern recognition receptors. *Science (New York, N.Y.)*, 302(5653), pp.2126–30.

Gouagna, L.C. et al., 1998. The early sporogonic cycle of Plasmodium falciparum in laboratory-infected Anopheles gambiae: an estimation of parasite efficacy. *Tropical Medicine and International Health*, 3(1), pp.21–28.

Gupta, L. et al., 2009. The STAT pathway mediates late-phase immunity against Plasmodium in the mosquito Anopheles gambiae. *Cell host & microbe*, 5(5), pp.498–507.

Hillyer, J.F.J.F., 2010. Chapter 12: Mosquito Immunity. In K. Söderhäll, ed. *Invertebrate Immunity*. Advances in Experimental Medicine and Biology. Boston, MA: Springer US, pp. 218–238.

Hoffmann, J.A., 2003. The immune response of Drosophila. *Nature*, 426(6962), pp.33–8.

Holt, R.A. et al., 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science*, 298(5591), pp.129–149.

Kim, W. et al., 2004. Ectopic expression of a cecropin transgene in the human malaria vector mosquito Anopheles gambiae (Diptera: Culicidae): effects on susceptibility to Plasmodium. *Journal of medical entomology*, 41(3), pp.447–55.

Kimbrell, D.A. & Beutler, B., 2001. The evolution and genetics of innate immunity. *Nature reviews. Genetics*, 2(4), pp.256–67.

Kumar, S. et al., 2010. A peroxidase/dual oxidase system modulates midgut epithelial immunity in Anopheles gambiae. *Science (New York, N.Y.)*, 327(5973), pp.1644–8.

Kumar, S. et al., 2003. The role of reactive oxygen species on Plasmodium melanotic encapsulation in Anopheles gambiae. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24), pp.14139–44.

Lavine, M.D. & Strand, M.R., 2002. Insect hemocytes and their role in immunity. *Insect Biochemistry and Molecular Biology*, 32(10), pp.1295–1309.

Lemaitre, B. et al., 1995. A recessive mutation, immune deficiency (imd), defines two distinct control pathways in the Drosophila host defense. *Proceedings of the National Academy of Sciences of the United States of America*, 92(21), pp.9465–9.

Lemaitre, B. et al., 1996. The Dorsoventral Regulatory Gene Cassette Controls the Potent Antifungal Response in Drosophila Adults. *Cell*, 86(6), pp.973–983.

Levashina, E.A. et al., 2001. Conserved Role of a Complement-like Protein in Phagocytosis Revealed by dsRNA Knockout in Cultured Cells of the Mosquito , Anopheles gambiae. *Cell*, 104, pp.709–718.

Li, Z., Zhang, S. & Liu, Q., 2008. Vitellogenin functions as a multivalent pattern recognition receptor with an opsonic activity. *PloS one*, 3(4), p.e1940.

Medzhitov, R., Preston-Hurlburt, P. & Janeway, C.A., 1997. A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. *Nature*, 388(6640), pp.394–7.

Meister, S. et al., 2009. Anopheles gambiae PGRPLC-mediated defense against bacteria modulates infections with malaria parasites. D. S. Schneider, ed. *PLoS pathogens*, 5(8), p.e1000542.

Meister, S. et al., 2005. Immune signaling pathways regulating bacterial and malaria parasite infection of the mosquito Anopheles gambiae. *Proceedings of the National Academy of Sciences of the United States of America*, 102(32), pp.11420–5.

Meister, S., Koutsos, A.C. & Christophides, G.K., 2004. The Plasmodium parasite--a “new” challenge for insect innate immunity. *International journal for parasitology*, 34(13-14), pp.1473–82.

Mendes, A.M. et al., 2008. Conserved mosquito/parasite interactions affect development of Plasmodium falciparum in Africa. D. S. Schneider, ed. *PLoS pathogens*, 4(5), p.e1000069.

Minakhina, S., Tan, W. & Steward, R., 2011. JAK/STAT and the GATA factor Pannier control hemocyte maturation and differentiation in Drosophila. *Developmental biology*, 352(2), pp.308–16.

Mitri, C. et al., 2009. Fine pathogen discrimination within the APL1 gene family protects Anopheles gambiae against human and rodent malaria species. D. S. Schneider, ed. *PLoS pathogens*, 5(9), p.e1000576.

Myllymäki, H., Valanne, S. & Rämet, M., 2014. The Drosophila imd signaling pathway. *Journal of immunology (Baltimore, Md. : 1950)*, 192(8), pp.3455–62.

Nüsslein-Volhard, C. & Wieschaus, E., 1980. Mutations affecting segment number and polarity in Drosophila. *Nature*, 287(5785), pp.795–801.

Obbard, D.J. et al., 2008. The evolution of TEP1, an exceptionally polymorphic immunity gene in Anopheles gambiae. *BMC evolutionary biology*, 8(1), p.274.

Oliveira, G. d. A., Lieberman, J. & Barillas-Mury, C., 2012. Epithelial Nitration by a Peroxidase/NOX5 System Mediates Mosquito Antiplasmodial Immunity. *Science*, 335(6070), pp.856–859.

Osta, M. a et al., 2004. Innate immunity in the malaria vector Anopheles gambiae: comparative and functional genomics. *The Journal of experimental biology*, 207(Pt 15), pp.2551–63.

Osta, M.A., Christophides, G.K. & Kafatos, F.C., 2004. Effects of mosquito genes on Plasmodium development. *Science (New York, N.Y.)*, 303(5666), pp.2030–2.

Park, J.-W. et al., 2007. Clustering of peptidoglycan recognition protein-SA is required for sensing lysine-type peptidoglycan in insects. *Proceedings of the National Academy of Sciences of the United States of America*, 104(16), pp.6602–7.

Pham, L.N. et al., 2007. A specific primed immune response in Drosophila is dependent on phagocytes. K. Vernick, ed. *PLoS pathogens*, 3(3), p.e26.

Povelones, M. et al., 2009. Leucine-rich repeat protein complex activates mosquito complement in defense against Plasmodium parasites. *Science*, 324(5924), pp.258–261.

Povelones, M. et al., 2011. Structure-function analysis of the Anopheles gambiae LRIM1/APL1C complex and its interaction with complement C3-like protein TEP1. *PLoS pathogens*, 7(4), p.e1002023.

Povelones, M. et al., 2013. The CLIP-domain serine protease homolog SPCLIP1 regulates complement recruitment to microbial surfaces in the malaria mosquito Anopheles gambiae. *PLoS pathogens*, 9(9), p.e1003623.

Riehle, M.M. et al., 2007. A major genetic locus controlling natural Plasmodium falciparum infection is shared by East and West African Anopheles gambiae. *Malaria journal*, 6(1), p.87.

Rodrigues, J. et al., 2010. Hemocyte differentiation mediates innate immune memory in Anopheles gambiae mosquitoes. *Science (New York, N.Y.)*, 329(5997), pp.1353–5.

Rono, M.K. et al., 2010. The major yolk protein vitellogenin interferes with the anti-plasmodium response in the malaria mosquito Anopheles gambiae. D. S. Schneider, ed. *PLoS biology*, 8(7), p.e1000434.

Rosetto, M. et al., 1995. Signals from the IL-1 receptor homolog, Toll, can activate an immune response in a Drosophila hemocyte cell line. *Biochemical and biophysical research communications*, 209(1), pp.111–6.

Roth, O. et al., 2009. Strain-specific priming of resistance in the red flour beetle, Tribolium castaneum. *Proceedings. Biological sciences / The Royal Society*, 276(1654), pp.145–51.

Royet, J., Gupta, D. & Dziarski, R., 2011. Peptidoglycan recognition proteins: modulators of the microbiome and inflammation. *Nature reviews. Immunology*, 11(12), pp.837–51.

Schnitger, A.K.D. et al., 2009. Two C-type lectins cooperate to defend Anopheles gambiae against Gram-negative bacteria. *The Journal of biological chemistry*, 284(26), pp.17616–24.

Sinden, R.E., 1999. Plasmodium differentiation in the mosquito. *Parassitologia*, 41(1-3), pp.139–48.

Smith, R.C., Vega-Rodríguez, J. & Jacobs-Lorena, M., 2014. The Plasmodium bottleneck: malaria parasite losses in the mosquito vector. *Memórias do Instituto Oswaldo Cruz*, (ahead), pp.1–18.

Takehana, A. et al., 2002. Overexpression of a pattern-recognition receptor, peptidoglycan-recognition protein-LE, activates imd/relish-mediated antibacterial defense and the prophenoloxidase cascade in Drosophila larvae. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21), pp.13705–10.

Valanne, S., Wang, J.-H. & Rämet, M., 2011. The Drosophila Toll signaling pathway. *Journal of immunology (Baltimore, Md. : 1950)*, 186(2), pp.649–56.

Vlachou, D. et al., 2005. Functional genomic analysis of midgut epithelial responses in Anopheles during Plasmodium invasion. *Current biology : CB*, 15(13), pp.1185–95.

Volz, J. et al., 2006. A genetic module regulates the melanization response of Anopheles to Plasmodium. *Cellular microbiology*, 8(9), pp.1392–405.

Warr, E. et al., 2006. Anopheles gambiae immune responses to Sephadex beads: involvement of anti-Plasmodium factors in regulating melanization. *Insect biochemistry and molecular biology*, 36(10), pp.769–78.

Warr, E. et al., 2008. The Gram-negative bacteria-binding protein gene family: its role in the innate immune system of anopheles gambiae and in anti-Plasmodium defence. *Insect molecular biology*, 17(1), pp.39–51.

Waterhouse, R.M. et al., 2007. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, 316(5832), pp.1738–1743.

Waterhouse, R.M. et al., 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic acids research*, 41(Database issue), pp.D358–65.

Waterhouse, R.M., Povelones, M. & Christophides, G.K., 2010. Sequence-structure-function relations of the mosquito leucine-rich repeat immune proteins. *BMC genomics*, 11, p.531.

White, B.J. et al., 2011. Adaptive divergence between incipient species of Anopheles gambiae increases resistance to Plasmodium. *Proc Natl Acad Sci U S A*, 108(1), pp.244–249.

Wright, V.M. et al., 2011. Differential activities of the Drosophila JAK/STAT pathway ligands Upd, Upd2 and Upd3. *Cellular signalling*, 23(5), pp.920–7.

Yan, R. et al., 1996. Identification of a Stat Gene That Functions in Drosophila Development. *Cell*, 84(3), pp.421–430.

Zambon, R.A. et al., 2005. The Toll pathway is important for an antiviral response in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 102(20), pp.7257–62.

Chapter 3

Bayoh, M.N., Thomas, C.J. & Lindsay, S.W., 2001. Mapping distributions of chromosomal forms of Anopheles gambiae in West Africa using climate data. *Medical and Veterinary Entomology*, 15(3), pp.267–274.

Besansky, N.J. et al., 1997. Patterns of Mitochondrial Variation Within and Between African Malaria Vectors, Anopheles gambiae and An. arabiensis, Suggest Extensive Gene Flow. *Genetics*, 147(4), pp.1817–1828.

Besansky, N.J. et al., 2003. Semipermeable species boundaries between Anopheles gambiae and Anopheles arabiensis: evidence from multilocus DNA sequence variation. *Proceedings of the National Academy of Sciences of the United States of America*, 100(19), pp.10818–23.

Brooke, B.D., Hunt, R.H. & Coetzee, M., 2000. Resistance to dieldrin + fipronil assorts with chromosome inversion 2La in the malaria vector Anopheles gambiae. *Medical and veterinary entomology*, 14(2), pp.190–194.

Caputo, B. et al., 2008. Anopheles gambiae complex along The Gambia river, with particular reference to the molecular forms of An. gambiae s.s. *Malaria journal*, 7(1), p.182.

Caputo, B. et al., 2011. The “far-west” of Anopheles gambiae molecular forms. D. Ortiz-Barrientos, ed. *PloS one*, 6(2), p.e16415.

Cassone, B.J. et al., 2011. Divergent transcriptional response to thermal stress by Anopheles gambiae larvae carrying alternative arrangements of inversion 2La. *Molecular ecology*, 20(12), pp.2567–80.

Cheng, C. et al., 2012. Ecological genomics of Anopheles gambiae along a latitudinal cline: a population-resequencing approach. *Genetics*, 190(4), pp.1417–32.

Clarkson, C.S. et al., 2014. Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. *Nature communications*, 5, p.4248.

Coetzee, M. et al., 2013. Anopheles coluzzii and Anopheles amharicus, new members of the Anopheles gambiae complex. *Zootaxa*, 3619(3), pp.246–274.

Coluzzi, M. et al., 2002. A polytene chromosome analysis of the Anopheles gambiae species complex. *Science (New York, N.Y.)*, 298(5597), pp.1415–8.

Coluzzi, M. et al., 1979. Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 73(5), pp.483–97.

Coluzzi, M. et al., 1985. Chromosomal inversion intergradation and incipient speciation in Anopheles gambiae. *Bolletino di zoologia*, 52(1-2), pp.45–63.

Coluzzi, M., 1992. Malaria vector analysis and control. *Parasitology Today*, 8(4), pp.113–118.

Costantini, C. et al., 2009. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in Anopheles gambiae. *BMC Ecol*, 9(1), p.16.

Coulibaly, M.B. et al., 2007. PCR-based karyotyping of Anopheles gambiae inversion 2Rj identifies the BAMAKO chromosomal form. *Malar J*, 6, p.133.

Cruickshank, T.E. & Hahn, M.W., 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular ecology*, 23(13), pp.3133–57.

Dabire, K.R. et al., 2013. Assortative mating in mixed swarms of the mosquito Anopheles gambiae s.s. M and S molecular forms, in Burkina Faso, West Africa. *Medical and veterinary entomology*, 27(3), pp.298–312.

Davidson, G., 1956. Insecticide resistance in Anopheles gambiae giles. *Nature*, 178(4535), pp.705–6.

Diabate, A. et al., 2008. Evidence for divergent selection between the molecular forms of Anopheles gambiae: role of predation. *BMC Evol Biol*, 8, p.5.

Diabate, A. et al., 2005. Larval development of the molecular forms of Anopheles gambiae (Diptera: Culicidae) in different habitats: a transplantation experiment. *J Med Entomol*, 42(4), pp.548–553.

Diabaté, A. et al., 2002. First report of the kdr mutation in Anopheles gambiae M form from Burkina Faso, west Africa. *Parassitologia*, 44(3-4), pp.157–8.

Du, W. et al., 2005. Independent mutations in the Rdl locus confer dieldrin resistance to Anopheles gambiae and An. arabiensis. *Insect molecular biology*, 14(2), pp.179–83.

Fanello, C., Santolamazza, F. & Della Torre, A., 2002. Simultaneous identification of species and molecular forms of the Anopheles gambiae complex by PCR-RFLP. *Medical and Veterinary Entomology*, 16(4), pp.461–464.

Favia, G. et al., 2001. Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of Anopheles gambiae s.s. *Insect Molecular Biology*, 10(1), pp.19–23.

Gimonneau, G. et al., 2012. Behavioural responses of Anopheles gambiae sensu stricto M and S molecular form larvae to an aquatic predator in Burkina Faso. *Parasites & vectors*, 5(1), p.65.

Gray, E.M. et al., 2009. Inversion 2La is associated with enhanced desiccation resistance in Anopheles gambiae. *Malar J*, 8(1), p.215.

Hahn, M.W. et al., 2012. No evidence for biased co-transmission of speciation islands in Anopheles gambiae. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1587), pp.374–84.

Lawniczak, M.K.N. et al., 2010. Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. *Science*, 330(6003), pp.512–514.

Lee, Y., Collier, T.C., et al., 2013. Chromosome Inversions, Genomic Differentiation and Speciation in the African Malaria Mosquito Anopheles gambiae B. Brooke, ed. *PLoS ONE*, 8(3), p.e57887.

Lee, Y. et al., 2009. Ecological and genetic relationships of the Forest-M form among chromosomal and molecular forms of the malaria vector Anopheles gambiae sensu stricto. *Malaria journal*, 8(1), p.75.

Lee, Y., Marsden, C.D., et al., 2013. Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, Anopheles gambiae. *Proceedings of the National Academy of Sciences of the United States of America*.

Lee, Y. & Lanzaro, G., 2013. Speciation in Anopheles gambiae — The Distribution of Genetic Polymorphism and Patterns of Reproductive Isolation Among Natural Populations. In S. Manguin, ed. *Anopheles mosquitoes - New insights into malaria vectors*. InTech.

Lehmann, T. et al., 1996. Genetic differentiation of Anopheles gambiae populations from East and west Africa: comparison of microsatellite and allozyme loci. *Heredity*, 77 (Pt 2)(November 1995), pp.192–200.

Lehmann, T., 2003. Population Structure of Anopheles gambiae in Africa. *Journal of Heredity*, 94(2), pp.133–147.

Lehmann, T. & Diabate, A., 2008. The molecular forms of Anopheles gambiae: a phenotypic perspective. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 8(5), pp.737–46.

Manoukis, N.C. et al., 2008. A test of the chromosomal theory of ecotypic speciation in Anopheles gambiae. *Proc Natl Acad Sci U S A*, 105(8), pp.2940–2945.

Manoukis, N.C. et al., 2009. Structure and dynamics of male swarms of Anopheles gambiae. *Journal of medical entomology*, 46(2), pp.227–35.

Masendu, H.T. et al., 2004. The sympatric occurrence of two molecular forms of the malaria vector Anopheles gambiae Giles sensu stricto in Kanyemba, in the Zambezi Valley, Zimbabwe. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 98(7), pp.393–6.

McLain, D.K. et al., 1989. Microgeographic variation in rDNA intergenic spacers of Anopheles gambiae in western Kenya. *Heredity*, 62 (Pt 2), pp.257–64.

Ndiath, M.O. et al., 2008. Dynamics of transmission of Plasmodium falciparum by Anopheles arabiensis and the molecular forms M and S of Anopheles gambiae in Dielmo, Senegal. *Malaria journal*, 7, p.136.

Neafsey, D.E. et al., 2010. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science (New York, N.Y.)*, 330(6003), pp.514–7.

Noor, M.A. et al., 2001. Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences of the United States of America*, 98(21), pp.12084–8.

Nwakanma, D.C. et al., 2013. Breakdown in the Process of Incipient Speciation in Anopheles gambiae. *Genetics*, 193(4), pp.1221–31.

Oliveira, E. et al., 2008. High Levels of Hybridization between Molecular Forms of <I>Anopheles gambiae</I> from Guinea Bissau. *Journal of Medical Entomology*, 45(6), pp.1057–1063.

Reidenbach, K.R. et al., 2012. Patterns of genomic differentiation between ecologically differentiated M and S forms of Anopheles gambiae in West and Central Africa. *Genome biology and evolution*.

Riehle, M.M. et al., 2011. A cryptic subgroup of Anopheles gambiae is highly susceptible to human malaria parasites. *Science*, 331(6017), pp.596–598.

Riehle, M.M. et al., 2006. Natural malaria infection in Anopheles gambiae is regulated by a single genomic control region. *Science*, 312(5773), pp.577–579.

Rocca, K.A.C. et al., 2009. 2La chromosomal inversion enhances thermal tolerance of Anopheles gambiae larvae. *Malar J*, 8(1), p.147.

Sharakhova, M. V et al., 2011. Arm-specific dynamics of chromosome evolution in malaria mosquitoes. *BMC evolutionary biology*, 11(1), p.91.

Simard, F. et al., 2009. Ecological niche partitioning between Anopheles gambiae molecular forms in Cameroon: the ecological side of speciation. *BMC Ecol*, 9(1), p.17.

Slotman, M.A. et al., 2005. Differential Introgression of Chromosomal Regions Between Anopheles gambiae and Anopheles Arabiensis. *Am J Trop Med Hyg*, 73(2), pp.326–335.

Stevison, L.S., Hoehn, K.B. & Noor, M.A.F., 2011. Effects of inversions on within- and between-species recombination and divergence. *Genome biology and evolution*, 3(0), pp.830–41.

Stump, A.D. et al., 2005. Centromere-proximal differentiation and speciation in Anopheles gambiae. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44), pp.15930–5.

Taylor, C. et al., 2001. Gene Flow Among Populations of the Malaria Vector, Anopheles gambiae, in Mali, West Africa. *Genetics*, 157(2), pp.743–750.

Tene Fossog, B. et al., 2013. Physiological correlates of ecological divergence along an urbanization gradient: differential tolerance to ammonia among molecular forms of the malaria mosquito Anopheles gambiae. *BMC ecology*, 13(1), p.1.

Della Torre, A. et al., 2001. Molecular evidence of incipient speciation within Anopheles gambiae s.s. in West Africa. *Insect Mol Biol*, 10(1), pp.9–18.

Della Torre, A., 1996. Polytene chromosome preparation from anopheline mosquitoes. In J. M. Crampton, C. Ben Beard, & C. Louis, eds. *The Molecular Biology of Insect Disease Vectors*. Dordrecht: Springer Netherlands.

Della Torre, A. et al., 1997. Selective Introgression of Paracentric Inversions Between Two Sibling Species of the Anopheles gambiae Complex. *Genetics*, 146(1), pp.239–244.

Della Torre, A. et al., 2002. Speciation within Anopheles gambiae--the glass is half full. *Science*, 298(5591), pp.115–117.

Della Torre, A., Tu, Z. & Petrarca, V., 2005. On the distribution and genetic differentiation of Anopheles gambiae s.s. molecular forms. *Insect biochemistry and molecular biology*, 35(7), pp.755–69.

Touré, Y.T. et al., 1994. Ecological genetic studies in the chromosomal form Mopti ofAnopheles gambiae s.str. in Mali, West Africa. *Genetica*, 94(2-3), pp.213–223.

Touré, Y.T. et al., 1998. The distribution and inversion polymorphism of chromosomally recognized taxa of the Anopheles gambiae complex in Mali, West Africa. *Parassitologia*, 40(4), pp.477–511.

Tripet, F. et al., 2004. The “wingbeat hypothesis” of reproductive isolation between members of the Anopheles gambiae complex (Diptera: Culicidae) does not fly. *Journal of medical entomology*, 41(3), pp.375–84.

Wang, R. et al., 2001. When genetic distance matters: measuring genetic differentiation at microsatellite loci in whole-genome scans of recent and incipient mosquito species. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19), pp.10769–74.

Wang-Sattler, R. et al., 2007. Mosaic genome architecture of the Anopheles gambiae species complex. *PloS one*, 2(11), p.e1249.

Weetman, D. et al., 2012. Gene flow-dependent genomic divergence between Anopheles gambiae M and S forms. *Molecular biology and evolution*, 29(1), pp.279–91.

White, B.J. et al., 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of Anopheles gambiae. *Molecular ecology*, 19(5), pp.925–39.

White, B.J. et al., 2007. Molecular karyotyping of the 2La inversion in Anopheles gambiae. *Am J Trop Med Hyg*, 76(2), pp.334–339.

White, B.J. et al., 2009. The population genomics of trans-specific inversion polymorphisms in Anopheles gambiae. *Genetics*, 183(1), pp.275–88.

White, G.B., 1985. Anopheles bwambae sp.n., a malaria vector in the Semliki Valley, Uganda, and its relationships with other sibling species of the An.gambiae complex (Diptera: Culicidae). *Systematic Entomology*, 10(4), pp.501–522.

Chapter 4

Ag1000G, C., 2014. Ag1000G: Anopheles gambiae 1000 Genomes | www.malariagen.net. Available at: http://www.malariagen.net/projects/vector/ag1000g [Accessed August 20, 2014].

Ag1000G Consortium, The Anopheles 1000 genome project (manuscript in preparation).

Aulard, S., David, J.R. & Lemeunier, F., 2002. Chromosomal inversion polymorphism in Afrotropical populations of Drosophila melanogaster. *Genetics Research*, 79(01), pp.49–63.

Bolstad, B.M. et al., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2), pp.185–193.

Cáceres, A. et al., 2012. Identification of polymorphic inversions from genotypes. *BMC bioinformatics*, 13(1), p.28.

Caputo, B. et al., 2011. The “far-west” of Anopheles gambiae molecular forms. D. Ortiz-Barrientos, ed. *PloS one*, 6(2), p.e16415.

Chakraborty, R. & Weiss, K.M., 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences*, 85(23), pp.9119–9123.

Chen, K. et al., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9), pp.677–81.

Cheng, C. et al., 2012. Ecological genomics of Anopheles gambiae along a latitudinal cline: a population-resequencing approach. *Genetics*, 190(4), pp.1417–32.

Coluzzi, M. et al., 2002. A polytene chromosome analysis of the Anopheles gambiae species complex. *Science (New York, N.Y.)*, 298(5597), pp.1415–8.

Coluzzi, M. et al., 1977. Behavioural divergences between mosquitoes with different inversion karyotypes in polymorphic populations of the Anopheles gambiae complex. *Nature*, 266(5605), pp.832–833.

Coluzzi, M. et al., 1979. Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 73(5), pp.483–97.

Coluzzi, M. et al., 1985. Chromosomal inversion intergradation and incipient speciation in Anopheles gambiae. *Bolletino di zoologia*, 52(1-2), pp.45–63.

Corbett-Detig, R.B., Cardeno, C. & Langley, C.H., 2012. Sequence-based detection and breakpoint assembly of polymorphic inversions. *Genetics*, 192(1), pp.131–7.

Cortes, C. & Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20(3), pp.273–297.

Coulibaly, M.B. et al., 2007. PCR-based karyotyping of Anopheles gambiae inversion 2Rj identifies the BAMAKO chromosomal form. *Malar J*, 6, p.133.

Fanello, C., Santolamazza, F. & Della Torre, A., 2002. Simultaneous identification of species and molecular forms of the Anopheles gambiae complex by PCR-RFLP. *Medical and Veterinary Entomology*, 16(4), pp.461–464.

Fouet, C. et al., 2012. Adaptation to aridity in the malaria mosquito Anopheles gambiae: chromosomal inversion polymorphism and body size influence resistance to desiccation. J. Pinto, ed. *PloS one*, 7(4), p.e34841.

Frazer, K.A. et al., 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), pp.851–61.

Gentleman, R.C. et al., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), p.R80.

Hormozdiari, F. et al., 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, 26(12), pp.i350–7.

Krzywinski, J. & Besansky, N.J., 2003. Molecular systematics of Anopheles: from subgenera to subpopulations. *Annual review of entomology*, 48, pp.111–39.

Lawniczak, M.K.N. et al., 2010. Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. *Science*, 330(6003), pp.512–514.

Lobo, N.F. et al., 2010. Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion. *Malaria journal*, 9(1), p.293.

Lucas Lledó, J.I. & Cáceres, M., 2013. On the Power and the Systematic Biases of the Detection of Chromosomal Inversions by Paired-End Genome Sequencing Z. Liu, ed. *PLoS ONE*, 8(4), p.12.

Ma, J. & Amos, C.I., 2012. Investigation of inversion polymorphisms in the human genome using principal components analysis. F. Emmert-Streib, ed. *PloS one*, 7(7), p.e40224.

MR4, Malaria Research and Reference Reagent Resource Center (MR4). Available at: http://www.mr4.org/ [Accessed July 30, 2014].

Neafsey, D.E. et al., 2010. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science (New York, N.Y.)*, 330(6003), pp.514–7.

Petrarca, V. & Beier, J.C., 1992. Intraspecific chromosomal polymorphism in the Anopheles gambiae complex as a factor affecting malaria transmission in the Kisumu area of Kenya. *The American journal of tropical medicine and hygiene*, 46(2), pp.229–37.

Sharakhov, I. V et al., 2006. Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the Anopheles gambiae complex. *Proc Natl Acad Sci U S A*, 103(16), pp.6258–6262.

Sharakhova, M. V et al., 2011. Arm-specific dynamics of chromosome evolution in malaria mosquitoes. *BMC evolutionary biology*, 11(1), p.91.

Shifman, S., 2003. Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics*, 12(7), pp.771–776.

Sindi, S.S. & Raphael, B.J., 2010. Identification and frequency estimation of inversion polymorphisms from haplotype data. *Journal of computational biology : a journal of computational molecular cell biology*, 17(3), pp.517–31.

Wang-Sattler, R. et al., 2007. Mosaic genome architecture of the Anopheles gambiae species complex. *PloS one*, 2(11), p.e1249.

White, B.J. et al., 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of Anopheles gambiae. *Molecular ecology*, 19(5), pp.925–39.

White, B.J. et al., 2007. Molecular karyotyping of the 2La inversion in Anopheles gambiae. *Am J Trop Med Hyg*, 76(2), pp.334–339.

White, B.J. et al., 2009. The population genomics of trans-specific inversion polymorphisms in Anopheles gambiae. *Genetics*, 183(1), pp.275–88.

Zeitouni, B. et al., 2010. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics (Oxford, England)*, 26(15), pp.1895–6.

Chapter 5

Ag1000G Consortium, The Anopheles 1000 genome project (manuscript in preparation).

Bettencourt, R. et al., 2004. Toll and Toll-9 in Drosophila innate immune response. *Journal of endotoxin research*, 10(4), pp.261–8.

Blandin, S.A. et al., 2009. Dissecting the genetic basis of resistance to malaria parasites in Anopheles gambiae. *Science*, 326(5949), pp.147–150.

Cohuet, A. et al., 2006. Anopheles and Plasmodium: from laboratory models to natural systems in the field. *EMBO reports*, 7(12), pp.1285–9.

Collins, F.H. et al., 1986. Genetic selection of a Plasmodium-refractory strain of the malaria vector Anopheles gambiae. *Science (New York, N.Y.)*, 234(4776), pp.607–10.

Coluzzi, M. et al., 2002. A polytene chromosome analysis of the Anopheles gambiae species complex. *Science (New York, N.Y.)*, 298(5597), pp.1415–8.

Cruickshank, T.E. & Hahn, M.W., 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular ecology*, 23(13), pp.3133–57.

Dunstan, S.J. et al., 2012. Variation in human genes encoding adhesion and proinflammatory molecules are associated with severe malaria in the Vietnamese. *Genes and immunity*, 13(6), pp.503–8.

Frolet, C. et al., 2006. Boosting NF-kappaB-dependent basal immunity of Anopheles gambiae aborts development of Plasmodium berghei. *Immunity*, 25(4), pp.677–85.

Garver, L.S. et al., 2012. Anopheles Imd pathway factors and effectors in infection intensity-dependent anti-Plasmodium action. D. S. Schneider, ed. *PLoS pathogens*, 8(6), p.e1002737.

Garver, L.S., Dong, Y. & Dimopoulos, G., 2009. Caspar controls resistance to Plasmodium falciparum in diverse anopheline species. *PLoS pathogens*, 5(3), p.e1000335.

Gendrin, M. & Christophides, G.K., 2013. The Anopheles Mosquito Microbiota and Their Impact on Pathogen Transmission. *Anopheles mosquitoes - New insights into malaria vectors*, p.book chapter 2.

Gentleman, R.C. et al., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), p.R80.

Harris, C. et al., 2010. Polymorphisms in Anopheles gambiae immune genes associated with natural resistance to Plasmodium falciparum. *PLoS Pathog*, 6(9).

Hoffmann, T.J. et al., 2011. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics*, 98(6), pp.422–30.

Holt, R.A. et al., 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science*, 298(5591), pp.129–149.

Horton, A.A. et al., 2010. Identification of three single nucleotide polymorphisms in Anopheles gambiae immune signaling genes that are associated with natural Plasmodium falciparum infection. *Malaria journal*, 9(1), p.160.

Hurd, H. et al., 2005. EVALUATING THE COSTS OF MOSQUITO RESISTANCE TO MALARIA PARASITES. *Evolution*, 59(12), pp.2560–2572.

Iyengar, S.K. & Elston, R.C., 2007. The genetic basis of complex traits: rare variants or “common gene, common disease”? *Methods in molecular biology (Clifton, N.J.)*, 376, pp.71–84.

Jallow, M. et al., 2009. Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature genetics*, 41(6), pp.657–65.

King, E.G., Merkes, C.M., et al., 2012. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. *Genome research*, 22(8), pp.1558–66.

King, E.G., Macdonald, S.J. & Long, A.D., 2012. Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. *Genetics*, 191(3), pp.935–49.

Langmead, B. et al., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), p.R25.

Lawniczak, M.K.N. et al., 2010. Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. *Science*, 330(6003), pp.512–514.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–9.

Li, J. et al., 2013. Genome-block expression-assisted association studies discover malaria resistance genes in Anopheles gambiae. , 110(51).

Lobo, N.F. et al., 2010. Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion. *Malaria journal*, 9(1), p.293.

Manske, M. et al., 2012. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature*, 487(7407), pp.375–9.

Marchini, J. et al., 2004. The effects of human population structure on large genetic association studies. *Nature genetics*, 36(5), pp.512–7.

McLaren, W. et al., 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)*, 26(16), pp.2069–70.

Megy, K. et al., 2012. VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic acids research*, 40(Database issue), pp.D729–34.

Meister, S. et al., 2005. Immune signaling pathways regulating bacterial and malaria parasite infection of the mosquito Anopheles gambiae. *Proceedings of the National Academy of Sciences of the United States of America*, 102(32), pp.11420–5.

Mendes, A.M. et al., 2011. Infection intensity-dependent responses of Anopheles gambiae to the African malaria parasite Plasmodium falciparum. *Infection and immunity*, 79(11), pp.4708–15.

Mitri, C. et al., 2009. Fine pathogen discrimination within the APL1 gene family protects Anopheles gambiae against human and rodent malaria species. D. S. Schneider, ed. *PLoS pathogens*, 5(9), p.e1000576.

Molina-Cruz, A. et al., 2012. Some strains of Plasmodium falciparum, a human malaria parasite, evade the complement-like system of Anopheles gambiae mosquitoes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(28), pp.E1957–62.

Neafsey, D.E. et al., 2010. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science (New York, N.Y.)*, 330(6003), pp.514–7.

Neale, B.M. et al., 2011. Testing for an unusual distribution of rare variants. *PLoS genetics*, 7(3), p.e1001322.

O’Loughlin, S.M. et al., 2014. Genomic Analyses of Three Malaria Vectors Reveals Extensive Shared Polymorphism but Contrasting Population Histories. *Molecular biology and evolution*, p.msu040–.

Pombi, M. et al., 2006. Variation in recombination rate across the X chromosome of Anopheles gambiae. *The American journal of tropical medicine and hygiene*, 75(5), pp.901–3.

Ponnudurai, T. et al., 1982. Cultivation of fertile Plasmodium falciparum gametocytes in semi-automated systems. 1. Static cultures. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 76(6), pp.812–8.

Purcell, S. et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3), pp.559–75.

Qanbari, S. et al., 2012. A High Resolution Genome-Wide Scan for Significant Selective Sweeps: An Application to Pooled Sequence Data in Laying Chickens N. Singh, ed. *PLoS ONE*, 7(11), p.e49525.

Reimers, M. & Carey, V.J., 2006. Bioconductor: an open source framework for bioinformatics and computational biology. *Methods Enzymol*, 411, pp.119–134.

Riehle, M.M. et al., 2011. A cryptic subgroup of Anopheles gambiae is highly susceptible to human malaria parasites. *Science*, 331(6017), pp.596–598.

Riehle, M.M. et al., 2006. Natural malaria infection in Anopheles gambiae is regulated by a single genomic control region. *Science*, 312(5773), pp.577–579.

Shifman, S., 2003. Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics*, 12(7), pp.771–776.

Stathopoulos, S. et al., 2014. Genetic dissection of Anopheles gambiae gut epithelial responses to Serratia marcescens. D. S. Schneider, ed. *PLoS pathogens*, 10(3), p.e1003897.

Svenson, K.L. et al., 2012. High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics*, 190(2), pp.437–47.

Tauszig, S. et al., 2000. Toll-related receptors and the control of antimicrobial peptide expression in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 97(19), pp.10520–5.

Tripet, F., Aboagye-Antwi, F. & Hurd, H., 2008. Ecological immunology of mosquito-malaria interactions. *Trends in parasitology*, 24(5), pp.219–27.

Turissini, D.A., Gamez, S. & White, B.J., 2014. Genome-Wide Patterns of Polymorphism in an Inbred Line of the African Malaria Mosquito Anopheles gambiae. *Genome biology and evolution*, 6(11), pp.3094–104.

Van Tyne, D. et al., 2011. Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in Plasmodium falciparum. *PLoS genetics*, 7(4), p.e1001383.

Valdar, W. et al., 2006. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics*, 38(8), pp.879–87.

Waterhouse, R.M. et al., 2007. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, 316(5832), pp.1738–1743.

White, B.J. et al., 2011. Adaptive divergence between incipient species of Anopheles gambiae increases resistance to Plasmodium. *Proc Natl Acad Sci U S A*, 108(1), pp.244–249.

Wilding, C.S. et al., 2009. High, clustered, nucleotide diversity in the genome of Anopheles gambiae revealed through pooled-template sequencing: implications for high-throughput genotyping protocols. *BMC genomics*, 10(1), p.320.

Zhu, Y. et al., 2012. Empirical validation of pooled whole genome population re-sequencing in Drosophila melanogaster. J.-S. Seo, ed. *PloS one*, 7(7), p.e41901.

Chapter 6:

Ag1000G Consortium, The Anopheles 1000 genome project (manuscript in preparation).

Alout, H. et al., 2013. Insecticide resistance alleles affect vector competence of Anopheles gambiae s.s. for Plasmodium falciparum field isolates. J. Vontas, ed. *PloS one*, 8(5), p.e63849.

Baldini, F. et al., 2014. Evidence of natural Wolbachia infections in field populations of Anopheles gambiae. *Nature communications*, 5, p.3985.

Bamshad, M.J. et al., 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics*, 12(11), pp.745–55.

Bian, G. et al., 2013. Wolbachia invades Anopheles stephensi populations and induces refractoriness to Plasmodium infection. *Science (New York, N.Y.)*, 340(6133), pp.748–51.

Blandin, S.A. et al., 2009. Dissecting the genetic basis of resistance to malaria parasites in Anopheles gambiae. *Science*, 326(5949), pp.147–150.

Bousema, T. et al., 2012. Mosquito feeding assays to determine the infectiousness of naturally infected Plasmodium falciparum gametocyte carriers. N. Kumar, ed. *PloS one*, 7(8), p.e42821.

Brogdon, W.G. & McAllister, J.C., 1998. Simplification of adult mosquito bioassays through use of time-mortality determinations in glass bottles. *Journal of the American Mosquito Control Association*, 14(2), pp.159–64.

Cáceres, A. et al., 2012. Identification of polymorphic inversions from genotypes. *BMC bioinformatics*, 13(1), p.28.

Caputo, B. et al., 2011. The “far-west” of Anopheles gambiae molecular forms. D. Ortiz-Barrientos, ed. *PloS one*, 6(2), p.e16415.

Chen, K. et al., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9), pp.677–81.

Cirimotich, C.M. et al., 2011. Natural microbe-mediated refractoriness to Plasmodium infection in Anopheles gambiae. *Science (New York, N.Y.)*, 332(6031), pp.855–8.

Coluzzi, M. et al., 1979. Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 73(5), pp.483–97.

Coluzzi, M. et al., 1985. Chromosomal inversion intergradation and incipient speciation in Anopheles gambiae. *Bolletino di zoologia*, 52(1-2), pp.45–63.

Daniels, R. et al., 2008. A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. *Malaria journal*, 7(1), p.223.

Davey, J.W. et al., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature reviews. Genetics*, 12(7), pp.499–510.

Fabrigar, D. et al., 2014. Genome-wide analysis of indoor and outdoor-biting populations of An. gambiae s.s. in the Gambia. In *Challenges in Malaria Research 2014*.

Fanello, C., Santolamazza, F. & Della Torre, A., 2002. Simultaneous identification of species and molecular forms of the Anopheles gambiae complex by PCR-RFLP. *Medical and Veterinary Entomology*, 16(4), pp.461–464.

Harris, C. et al., 2012. Plasmodium falciparum produce lower infection intensities in local versus foreign Anopheles gambiae populations. *PloS one*, 7(1), p.e30849.

Harris, C. et al., 2010. Polymorphisms in Anopheles gambiae immune genes associated with natural resistance to Plasmodium falciparum. *PLoS Pathog*, 6(9).

Horton, A.A. et al., 2010. Identification of three single nucleotide polymorphisms in Anopheles gambiae immune signaling genes that are associated with natural Plasmodium falciparum infection. *Malaria journal*, 9(1), p.160.

Kokoza, V. et al., 2010. Blocking of Plasmodium transmission by cooperative action of Cecropin A and Defensin A in transgenic Aedes aegypti mosquitoes. *Proceedings of the National Academy of Sciences of the United States of America*, 107(18), pp.8111–6.

Van der Kolk, M. et al., 2005. Evaluation of the standard membrane feeding assay (SMFA) for the determination of malaria transmission-reducing activity using empirical data. *Parasitology*, 130(Pt 1), pp.13–22.

Lambrechts, L. et al., 2005. Host genotype by parasite genotype interactions underlying the resistance of anopheline mosquitoes to Plasmodium falciparum. *Malaria journal*, 4, p.3.

Mackay, T.F.C. et al., 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384), pp.173–8.

Mathias, D.K. et al., 2011. Spatial and temporal variation in the kdr allele L1014S in Anopheles gambiae s.s. and phenotypic variability in susceptibility to insecticides in Western Kenya. *Malaria journal*, 10(1), p.10.

Miura, K. et al., 2013. Qualification of standard membrane-feeding assay with Plasmodium falciparum malaria and potential improvements for future assays. D. J. Diemert, ed. *PloS one*, 8(3), p.e57909.

Neafsey, D.E. et al., 2010. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science (New York, N.Y.)*, 330(6003), pp.514–7.

Nielsen, R. et al., 2011. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, 12(6), pp.443–51.

Nunes, J.K. et al., 2014. Development of a transmission-blocking malaria vaccine: Progress, challenges, and the path forward. *Vaccine*, 32(43), pp.5531–9.

Park, D.J. et al., 2012. Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite. , 109(32), pp.13052–13057.

Redmond, S. et al., Association Mapping by Pooled Sequence Identifies TOLL 11 as a Protective Factor against Plasmodium falciparum in Anopheles gambiae. *in preparation*.

Schlötterer, C. et al., 2014. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, 15(11), pp.749–763.

Van Tyne, D. et al., 2011. Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in Plasmodium falciparum. *PLoS genetics*, 7(4), p.e1001383.

Weetman, D. et al., 2010. Association mapping of insecticide resistance in wild Anopheles gambiae populations: major variants identified in a low-linkage disequilbrium genome. K. Y. K. Chan, ed. *PloS one*, 5(10), p.e13140.

White, B.J. et al., 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of Anopheles gambiae. *Molecular ecology*, 19(5), pp.925–39.

WHO, 2013. *WHO | Test procedures for insecticide resistance monitoring in malaria vector mosquitoes*, World Health Organization.

WHO Expert Committee on Insecticides, 1963. *Insecticide Resistance and Vector Control*,

Zhu, Y. et al., 2012. Empirical validation of pooled whole genome population re-sequencing in Drosophila melanogaster. J.-S. Seo, ed. *PloS one*, 7(7), p.e41901.

Appendices

A1: The *Plasmodium* lifecycle

The plasmodium parasite has a complex life cycle. Obligately cycling through both an invertebrate and vertebrate host, passing through two distinct cell types, and undergoing multiple rounds of sexual and asexual proliferation.

From the point of infection in the mammalian host, sporozoites invade the liver cells where it forms a parasitophorous vacuole. Within this vacuole they undergo exoerythrocytic / liver-stage development into schizonts. Proliferation into thousands of merozoites takes place within schizonts over a period of 5-16 days. Upon rupture these merozoites are released into the bloodstream. Invasion of red blood cells by the liver-derived merozoites begins the erythrocytic cycle: a repeated cycle of proliferation-by-schizogony within erythrocytes. The merozoite will develop first into an immature (ring stage) trophozoite, then a mature trophozoite and finally a schizont. Ruptured schizonts again release more merozoites that in turn invade more erythrocytes. Finally, in response to cues that are as-yet-unknown (but likely to be stress related, such as erythrocyte age, hypoxia and schizonticidal drug pressure (Sinden et al. 2012)), the erythrocytic cycle is interrupted as ring-stage trophozoites develop into sexually differentiated gametocytes.

The uptake of an infected bloodmeal by the mosquito vector ingests both male and female gametocytes. Triggered by the drop in pH and temperature in the mosquito midgut, the gametocytes develop further into motile exflagellated microgametes (male) and macrogametes (female); these in turn fuse and develop into the motile ookinete stage that traverses the mosquito midgut wall. Successful penetration of the midgut will be followed by the transformation of the ookinete into an oocyst, within which further proliferation takes place into a large number of sporozoites. This process takes between 8 and 15 days depending on the *Plasmodium* species. After rupture of the oocyst, infective sporozoites migrate to the mosquito salivary gland. Should the mosquito take another bloodmeal, these infective sporozoites will be transferred to the mammalian host.

Sinden, R.E. et al., 2012. The biology of sexual development of Plasmodium: the design and implementation of transmission-blocking strategies. *Malaria journal*, 11(1), p.70.

A2 : Immune family genes – potential targets for association mapping ?

In developing hypothesis-free methods of gene identification we do, of course, aim to provide unbiased assessments of immune function, however the presence of paralogous gene families indicates that the targets of genomic mapping projects are likely to be found within them, and in particular, novel expansions of gene families can indicate adaptation to clade-specific function, such as might be expected in infection of *Anopheles* with *P.falciparum*. With this in mind, it is helpful to elucidate the gene familes that can be readily identified in the mosquito, and which of those have and have not been ascribed a function.

LRIMs

The Leucine-rich immune protein (LRIM) family includes both partners in the TEP1-stabilising complex (LRIM1/APL1C) and the other gene implicated in the *P.falciparum* immune response, APL1A. It is logical therefore to expect others to have immune function, and perhaps also to mediate TEP1 opsonisation to other parasites:
The LRIMs are a subset of the LRR (leucine-rich-repeat containing) superfamily that also contains the TOLL receptors and other immune proteins. They represent a mosquito-specific family of proteins, with orthologues in *Aedes* and *Culex spp*. but none found in other diptera. They are characterised by a leucine rich repeat structure, followed by a pattern of cysteine residues and, in most cases, a coiled-coil domain – both of which are thought to be implicated in dimerization, They can be divided into broad families depending on how many leucine-rich-repeats (LRRs) they exhibit and which other structures are present. The short and long families containing 6-7 and 10+ repeats respectively, along with transmembrane types (featuring a C-terminal transmembrane domain) and the coil-less types. LRIM1 and the APL1s are all in the ‘long’ class (Waterhouse et al. 2010), and any putative alternate binding partners for APL1A or other convertase complexes are most likely to come from this subfamily. The long-form LRIM4 has been shown to be upregulated in the midgut upon *P.falciparum* ookinete invasion, while the coil-less LRIM17 has been shown by gene knocndown to have an immune role against both *P.berghei* and *P. falciparum* (Dong et al. 2006).

TEPs

TEP1 is one of seven genes in species-specific expansion in gambiae, many of which still possess a functioning thioester domain.

Of the three other TEPs identified in the *gambiae* genome, TEP2 appears to be unique to Anopheles, and TEP15 forms an orthologous cluster with the other culicidae. Both genes are upregulated upon bloodmeal, peaking at 3 hours after feeding (Marinotti et al. 2006).
Both TEP1 and TEP15 are within a larger phylogenetic cluster along with DmTEP1 and 2 (DmTEP1 is unrelated to AgTEP1) (Waterhouse et al. 2007) and also seen to be upregulated upon infection with the arbovirus sindbis, although knockdowns do not identify any of them as being essential for immunity (Mudiganti 2006).
Only one gene forms an orthologous cluster with drosophila; AgTEP13, AaTEP13 and dmTEP6 are 1:1:1 orthologues, however they have all lost their thioester domains and are highly unlikely to bind pathogens in the same manner as TEP1 (Waterhouse et al. 2007).

TOLLs:

The TOLL family are transmembrane proteins, featuring a single TM domain, a leucine rich repeat region of variable size, highly divergent extracellular N-terminal and a conserved C-terminal intracellular TOLL-ILK1 receptor (Christophides et al. 2002).
As well as the canonical *TOLL 1* (detailed above), immune and developmental functions have been attributed to other TOLL receptors. Anti-fungal and anti-viral functions have been described for the *dmTOLL5* (Tauszig et al. 2000) and *dmTOLL7* (Nakamoto et al. 2012) respectively.
Anopheles TOLL genes have shown notable conservation of these immune roles, with *TOLL 1* and *TOLL 5* activity against gram positive bacteria (Christophides et al. 2002), and an antifungal role for *TOLL 5* (Shin et al. 2006). Specific activation mechanisms of the TOLL receptors other than *TOLL 1* are not yet characterised, though the diversity of the extracellular domains may indicate their activation with proteins other than spätzle-1 (the roles of other members of the 6-gene *A.gambiae* spätzle family are yet to be defined). Similarly the highly conserved intracellular C-terminal domain, and the lack of duplication in the TOLL signal transduction cascade, lends itself to the interpretation that the downstream path for many TOLL receptors is similar or identical to that seen in *TOLL 1*. Indeed expression motifs of *dmTOLL9* indicate downstream TOLL-pathway activation (Bettencourt et al. 2004) and the expression of chimeric genes (with a *TOLL 1* extracellular domain and *TOLL 5* intracellular) has demonstrated activation of the *TOLL 1* pathway via the *TOLL 5* intracellular domain (Tauszig et al. 2000). However attempts to discern a similar effect using chimeric genes for other TOLL genes were unsuccessful, and similar experiments in the mosquito have yet to be performed.

Two separate gene expansions are seen in this family in the culicidae when compared to other dipterans - frequently an indication of species or clade specific functions. A duplication of *TOLL 1* and *TOLL 5* in the mosquito has led to four genes in the mosquito: *TOLL 1a, TOLL 1b, TOLL 5a* and *TOLL 5b*. It is also assumed from this that these genes were once physically linked in the ancestral genome (with a later translocation having separated them in *Drosophila*) (Christophides et al. 2002). Precise functions of these genes, and indeed the identity of the canonical *TOLL* receptor is not confirmed (the most direct orthologue is *TOLL 1a*), however mapping approaches have implicated both *TOLL 5a* and *TOLL 5b* in antifungal immunity (Horton et al. 2010).
The other expansion comprises genes *TOLL10* and *TOLL11.* These genes form an orthologous group in the mosquito and represent an entirely novel gene expansion in the *Culicidae.* Most closely related to *agTOLL7*, they have no reciprocal orthologue in *Drosophila*. More distantly related to the dmTOLLs than the TOLL1/5 cluster, they have not previously been implicated in mosquito immunity via either mapping or gene knockdown.
As with other TOLL genes, *TOLL10* and *TOLL11* show significant differential regulation during embryonic development (Goltsev et al. 2009) and both genes are upregulated in response to bloodfeeding (Marinotti et al. 2006). *TOLL11* (though not *TOLL10*) has further been shown to be significantly differentially regulated in response to infection with *Plasmodium spp.* (Mendes et al. 2011), and is upregulated in response to avian malaria challenge in *Aedes aegypti* (Zou et al. 2011).

Spätzles

There are 6 spätzle genes in *Anopheles*, all of which show a high degree of conservation with both Aedes and Anopheles (Waterhouse et al. 2007). AgSPZ1 is assumed to be the ligand for AgTOLL1, due to both genes' orthology with drosophila. The other five are therefore strong candidates as potential ligands for the other TOLL receptors, and potential substrates for an upstream CLIPB.

CLIPs

The 54 clip-domain serine protease (CLIPs) within the A.gambiae genome cluster into five broad clip families, named A-E.
Some of these clips carry mutations predicted to interrupt their CLIP domain, meaning they will have no protease activity; however these serine-protease homologues (SPHs) may be able to act as either co-factors or inhibitors to other proteases. Almost all of the CLIP-A genes encode SPHs, and three of the CLIP-As (1,6 and 7) are known to be upregulated upon bacterial challenge (Christophides et al. 2004).

The CLIP-Bs represent the largest family within *gambiae*, and contains genes similar to the TOLL-activating *dmSPE* and *dmEaster* genes. *dmEaster* is an orthologue of CLIPB5, although there is no direct orthologue for *dmSPE*, and which gene or genes within the CLIP-Bs activates this pathway in the immune context remains a mystery. In addition to the previously described roles of CLIP-B14 and 15 in the PPO cascade, CLIPBs 1,4 and 9 are upregulated on bacterial and malarial challenge (Christophides et al. 2004).

CLIP-Cs include orthologues of the Drosophila genes *Snake* and *Persephone*, activators of the TOLL pathway in response to developmental cues and fungal infection respectively, though few other members of the family have been functionally annotated.
Subfamilies D and E are largely uncharacterised. CLIP-Es are a recently described and highly divergent family consisting almost entirely of SPHs. The SPCLIP1 gene that has beenidentified as a TEP1-convertase groups within the CLIP-Es; this group would be a logical place to seek alternative convertases for different pathogens (Povelones et al. 2013).

SRPNs

There are a total of 18 SRPNs that have been identified in the *Anopheles gambiae* genome, one of which, AgSRPN13, is not present in the reference sequence but has been confirmed by EST analysis (Waterhouse et al. 2007).
Out of those 18, analysis of the protease domain indicates that 12 are likely to be active protease inhibitors, and 6 (SRPNs 11-14 and 18-19) have sufficiently severe mutations to have lost their protease capacity.
Highly conserved between mosquitoes, with a plethora of 1:1 orthologues between Aedes and Anopheles, they are less well conserved between mosquitoes and Drosophila. In addition to the two SRPNs that have been implicated in *P.berghei* immunity (see above), SRPN10 has also been implicated in apoptosis in response to *P.berghei* infection (Danielli et al. 2005).
It is fairly assumed that many SRPNs will serve as inhibitors of the CLIPs, as has been shown for AgSRPN3 and the moth (*Manduca sexta*) protein PAP3 in vitro (Michel et al. 2006) (SRPN3's binding partner in *A.gambiae* is currently unknown). Expression mapping of both CLIP and SRPN families show regular co-clustering of SRPNs and CLIPs, indicating frequent co-regulation of both families (Maccallum et al. 2011).

Bettencourt, R. et al., 2004. Toll and Toll-9 in Drosophila innate immune response. *Journal of endotoxin research*, 10(4), pp.261–8.

Christophides, G.G.K. et al., 2002. Immunity-related genes and gene families in Anopheles gambiae. *Science (New York, N.Y.)*, 159(2002), pp.159–65.

Christophides, G.K., Vlachou, D. & Kafatos, F.C., 2004. Comparative and functional genomics of the innate immune system in the malaria vector Anopheles gambiae. *Immunological reviews*, 198, pp.127–48.

Danielli, A. et al., 2005. Overexpression and altered nucleocytoplasmic distribution of Anopheles ovalbumin-like SRPN10 serpins in Plasmodium-infected midgut cells. *Cellular microbiology*, 7(2), pp.181–90.

Dong, Y. et al., 2006. Anopheles gambiae immune responses to human and rodent Plasmodium parasite species. *PLoS pathogens*, 2(6), p.e52.

Goltsev, Y. et al., 2009. Developmental and evolutionary basis for drought tolerance of the Anopheles gambiae embryo. *Developmental biology*, 330(2), pp.462–70.

Horton, A.A. et al., 2010. Identification of three single nucleotide polymorphisms in Anopheles gambiae immune signaling genes that are associated with natural Plasmodium falciparum infection. *Malaria journal*, 9(1), p.160.

Maccallum, R.M., Redmond, S.N. & Christophides, G.K., 2011. An expression map for Anopheles gambiae. *BMC genomics*, 12(1), p.620.

Marinotti, O. et al., 2006. Genome-wide analysis of gene expression in adult Anopheles gambiae. *Insect molecular biology*, 15(1), pp.1–12.

Mendes, A.M. et al., 2011. Infection intensity-dependent responses of Anopheles gambiae to the African malaria parasite Plasmodium falciparum. *Infection and immunity*, 79(11), pp.4708–15.

Michel, K. et al., 2006. Increased melanizing activity in Anopheles gambiae does not affect development of Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America*, 103(45), pp.16858–63.

Mudiganti, U., 2006. *Insect Response to Alphavirus Infection*, ProQuest.

Nakamoto, M. et al., 2012. Virus recognition by Toll-7 activates antiviral autophagy in Drosophila. *Immunity*, 36(4), pp.658–67.

Povelones, M. et al., 2013. The CLIP-domain serine protease homolog SPCLIP1 regulates complement recruitment to microbial surfaces in the malaria mosquito Anopheles gambiae. *PLoS pathogens*, 9(9), p.e1003623.

Shin, S.W., Bian, G. & Raikhel, A.S., 2006. A toll receptor and a cytokine, Toll5A and Spz1C, are involved in toll antifungal immune signaling in the mosquito Aedes aegypti. *The Journal of biological chemistry*, 281(51), pp.39388–95.

Sinden, R.E. et al., 2012. The biology of sexual development of Plasmodium: the design and implementation of transmission-blocking strategies. *Malaria journal*, 11(1), p.70.

Tauszig, S. et al., 2000. Toll-related receptors and the control of antimicrobial peptide expression in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 97(19), pp.10520–5.

Waterhouse, R.M. et al., 2007. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, 316(5832), pp.1738–1743.

Waterhouse, R.M., Povelones, M. & Christophides, G.K., 2010. Sequence-structure-function relations of the mosquito leucine-rich repeat immune proteins. *BMC genomics*, 11, p.531.

Zou, Z. et al., 2011. Transcriptome analysis of Aedes aegypti transgenic mosquitoes with altered immunity. *PLoS pathogens*, 7(11), p.e1002394.

A3: Support Vector Classification code

All code is written in Python and requires numpy (v1.6.2), scipy (v0.11.0) and the scikit-learn package (v0.14.1). Input files are tab-delimited text containing mean depth for 100bp fragments across the duplicated locus.

File format :

metadata			locus	normalization	0-100bp	1-200bp	2-300bp	3-400bp	4-500bp	5-600bp	6-700bp	7-800bp	8-900bp	9-1000bp
AD0142_C	F	m	2R:19025000-19026000	QNORM	65	52	73	42	43	47	42	43	39	45
AD0143_C	M	ms	2R:19025000-19026000	QNORM	47	44	33	49	52	52	41	47	46	49
AD0146_C	F	No	2R:19025000-19026000	QNORM	49	41	46	48	49	39	43	30	46	50
AD0147_C	F	No	2R:19025000-19026000	QNORM	30	31	26	32	19	7	18	19	28	27
AD0148_C	F	No	2R:19025000-19026000	QNORM	48	53	45	43	43	34	31	35	37	48
AD0149_C	F	No	2R:19025000-19026000	QNORM	40	42	40	56	42	55	41	67	55	51

…

SVC training code

```
import os
import numpy
from sklearn import svm
from sklearn import cross_validation
from sklearn import grid_search
import getopt

# get options:
def usage():
    print "python ./classifier.py -t|target target_col_index  -v|values vals_start-vals_end \
"
    print "\t [-c|check check_col_index -s|save file_name_for_SVC_object ]  \ "
    print "\t filename_of_value_table (TSV)  \ "
    sys.exit(1)


targetCol = 7
vals_st=10
vals_en=19


opts_s = 't:v:c:s:'
opts_l = ['target=','values=','check=','save=']

try:
    optlist, args = getopt.getopt(sys.argv[1:], opts_s, opts_l)
except getopt.GetoptError as err:
    print str(err)
    usage()
```

```
for opt, val in optlist:
    if opt in ('-t', '--target'):
        targetCol=int(val)
    elif opt in ('-v','--values'):
        vals_st,vals_en = string.split(val,'-')
        vals_st = int(vals_st)
        vals_en = int(vals_en)
    elif opt in ('-c','--check'):
        checkCol=int(val)
    elif opt in ('-s','--save'):
        SVC_FN=val

COV_TABLE_FN = args[0]

# Function Definitions

#GET SVC WITH CROSS VALIDATION
def fitSVCWithCrossVal(values, target, param_grid=None):
    if param_grid is None:
        param_grid = [
          {'C': [1, 10, 100, 1000], 'kernel': ['linear']},
          {'C': [1, 10, 100, 1000], 'gamma': [0.01, 0.001, 0.0001, 0.00001], 'kernel':
['rbf']},
         ]
    svr = svm.SVC(probability=True)
    clsf = grid_search.GridSearchCV(svr, param_grid)
    clsf.fit(values, target)
    clsf_best = clsf.best_estimator_
    sys.stderr.write(str(clsf.best_estimator_)+"\n")
    return clsf

covTable = np.genfromtxt(COV_TABLE_FN,skip_header=0,
                         delimiter="\t",dtype='S20')



values = covTable[:,vals_st-1:vals_en]
values = values.astype('float',copy=True)
targets = covTable[:,targetCol-1]

targets_t = targets[targets != '']
values_t = values[targets != '',:]
sys.stderr.write(str(len(targets_t))+ " lines found with target\n")

# make SVC from values in table
svc_o=fitSVCWithCrossVal(values_t, targets_t)
svc = svc_o.best_estimator_

for i in range(0,len(targets)):
    value = values[i,:]
    svc_out = svc.predict([value])
    out_line = list(covTable[i]) + list(svc_out)
    print "\t".join(out_line)
    #    print line
if SVC_FN is not None:
    import pickle
    pickle.dump(svc, open( SVC_FN, "wb" ) )
```

A2.2: SVC calling code

```
import os
```

```
import numpy
import petl.interactive as etl
import prettytable

import getopt
import sys


# get options:
def usage():
    print "python ./run_SVC_classifier.py -f|svc|file file_name_for_SVC_object  -v|values
vals_start-vals_end \ "
    print "\t values_table_filename "


opts_s = 'f:v:'
opts_l = ['svc=','file=','values=']

try:
    optlist, args = getopt.getopt(sys.argv[1:], opts_s, opts_l)
except getopt.GetoptError as err:
    print str(err)
    usage()

for opt, val in optlist:
    if opt in ('-f', '--svc','--file'):
        SVC_FN=val
    elif opt in ('-v','--values'):
        vals_st,vals_en = string.split(val,'-')
        vals_st = int(vals_st)
        vals_en = int(vals_en)

COV_TABLE_FN = args[0]

covTable = np.genfromtxt(COV_TABLE_FN, delimiter="\t",dtype='S20')

values = covTable[:,vals_st-1:vals_en]
values = values.astype('float',copy=True)

# load SVC and fit to values
import pickle
if SVC_FN is not None:
    svc = pickle.load(open(SVC_FN, "rb"))

#if probability, print prob values
#if not just print call
if svc.probability:
    for i in range(0,shape(values)[0]):
        value = values[i,:]
        svc_out = svc.predict([value])
        svc_out_p = svc.predict_proba([value])
        out_line = np.concatenate((covTable[i], svc_out, svc_out_p[0]))
        print "\t".join(out_line)
else:
    for i in range(0,shape(values)[0]):
        value = values[i,:]
        svc_out = svc.predict([value])
        out_line = list(covTable[i]) + list(svc_out)
        print "\t".join(out_line)
```

A3 : Sample metadata Ag1kG-AR1 set

ind	country	region	contributor	year	m_s	sex
AB0085-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0087-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0088-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0089-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0090-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0091-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0092-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0094-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0095-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0097-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0098-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0099-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0100-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0101-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0103-C	Burkina Faso	Bana	Austin Burt	2012	S	F
AB0104-C	Burkina Faso	Bana	Austin Burt	2012	S	F
AB0109-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0110-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0111-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0112-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0113-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0114-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0117-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0119-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0122-C	Burkina Faso	Bana	Austin Burt	2012	M	M
AB0123-C	Burkina Faso	Bana	Austin Burt	2012	M	M
AB0124-C	Burkina Faso	Bana	Austin Burt	2012	M	M
AB0126-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0127-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0128-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0129-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0130-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0133-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0134-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0135-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0136-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0137-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0138-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0139-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0140-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0142-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0143-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0145-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0146-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0147-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0148-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0151-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0153-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0155-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0157-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0158-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0159-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0160-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0161-C	Burkina Faso	Bana	Austin Burt	2012	S	M
AB0164-C	Burkina Faso	Pala	Austin Burt	2012	S	M
AB0166-C	Burkina Faso	Pala	Austin Burt	2012	S	M
AB0169-C	Burkina Faso	Pala	Austin Burt	2012	S	M
AB0170-C	Burkina Faso	Pala	Austin Burt	2012	S	M

AB0171-C	Burkina Faso	Pala	Austin Burt	2012	S	M
AB0172-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0173-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0174-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0175-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0176-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0177-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0178-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0179-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0181-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0182-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0183-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0184-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0185-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0186-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0187-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0188-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0189-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0190-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0191-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0192-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0197-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0198-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0199-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0201-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0202-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0203-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0204-C	Burkina Faso	Pala	Austin Burt	2012	M	F
AB0205-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0206-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0207-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0208-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0209-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0210-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0211-C	Burkina Faso	Bana	Austin Burt	2012	S	F
AB0212-C	Burkina Faso	Bana	Austin Burt	2012	M	F
AB0213-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0217-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0219-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0221-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0222-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0223-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0224-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0226-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0227-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0228-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0229-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0231-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0233-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0234-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0235-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0236-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0237-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0238-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0239-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0240-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0241-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0242-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0243-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0244-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0246-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0249-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0250-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	M	F
AB0251-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F

AB0252-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0253-C	Burkina Faso	Sourukoudinga	Austin Burt	2012	S	F
AB0256-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0257-C	Burkina Faso	Pala	Austin Burt	2012	M	F
AB0258-C	Burkina Faso	Pala	Austin Burt	2012	M	F
AB0260-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0261-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0262-C	Burkina Faso	Pala	Austin Burt	2012	M	F
AB0263-C	Burkina Faso	Pala	Austin Burt	2012	M	F
AB0264-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0265-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0266-C	Burkina Faso	Pala	Austin Burt	2012	M	F
AB0267-C	Burkina Faso	Pala	Austin Burt	2012	M	F
AB0268-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0270-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0271-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0272-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0273-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0274-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0276-C	Burkina Faso	Pala	Austin Burt	2012	M	F
AB0277-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0278-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0279-C	Burkina Faso	Pala	Austin Burt	2012	M	F
AB0280-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0281-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0282-C	Burkina Faso	Pala	Austin Burt	2012	M	F
AB0283-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AB0284-C	Burkina Faso	Pala	Austin Burt	2012	S	F
AC0090-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0091-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0092-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0093-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0094-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0095-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0096-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0097-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0098-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0099-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0100-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0101-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0102-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0103-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0104-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0106-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0107-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0108-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0109-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0110-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0111-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0112-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0113-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0114-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0115-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0116-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0117-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0118-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0119-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0120-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0121-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0122-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0123-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0124-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0125-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0126-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F

AC0127-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0128-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0129-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0130-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0131-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0132-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0133-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0135-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0136-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0137-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0138-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0139-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0140-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0142-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0143-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0144-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0145-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0147-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0148-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0149-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0150-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0151-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0152-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0153-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0154-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0156-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0158-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0159-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0160-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0161-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0162-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0163-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0164-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0166-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0167-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0168-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0169-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0170-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0171-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0172-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0173-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0174-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0176-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0178-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0179-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0180-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0181-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0182-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0183-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0184-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0186-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0187-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0188-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0189-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0190-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0191-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0192-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0193-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0194-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0195-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0196-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0197-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0199-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0200-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F

AC0201-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0202-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AC0203-C	Uganda	Nagongera, Tororo	Martin Donnelly	2012	S	F
AJ0023-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0024-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0032-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0035-C	Guinea-Bissau	Antula	Joao Pinto	2010	M	F
AJ0036-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0039-C	Guinea-Bissau	Antula	Joao Pinto	2010	M	F
AJ0043-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0044-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0045-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0047-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0051-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0052-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0056-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0061-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0063-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0064-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0066-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0070-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0071-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0072-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0074-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0075-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0076-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0077-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0078-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0081-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0084-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0085-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0086-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0088-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0090-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0092-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0093-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0096-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0097-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0098-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0100-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0101-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0102-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0103-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0105-C	Guinea-Bissau	Antula	Joao Pinto	2010	M/S	F
AJ0107-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0109-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0113-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AJ0115-C	Guinea-Bissau	Antula	Joao Pinto	2010	M	F
AJ0116-C	Guinea-Bissau	Antula	Joao Pinto	2010	S	F
AK0065-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0066-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0067-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0068-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0069-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0070-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0072-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0073-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0074-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0075-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0076-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0077-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0078-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0079-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0080-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F

AK0081-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0082-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0085-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0086-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0087-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0088-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0089-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0090-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0091-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0092-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0093-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0094-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0095-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0096-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0098-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0099-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0100-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0101-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0102-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0103-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0104-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0105-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0106-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0108-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0109-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0110-C	Kenya	Kilifi-Junju	Janet Midega	2012	S	F
AK0116-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0119-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AK0127-C	Kenya	Kilifi-Mbogolo	Janet Midega	2012	S	F
AN0007-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0008-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0009-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0010-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0011-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0012-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0014-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0016-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0017-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0018-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0019-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0020-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0022-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0023-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0024-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0025-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0026-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0027-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0028-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0029-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0030-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0031-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0032-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0033-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0034-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0035-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0036-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	M
AN0037-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0038-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0039-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0040-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0041-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0042-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0043-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0045-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F

AN0046-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0047-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0048-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0049-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0050-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0051-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0053-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0054-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0055-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0056-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0057-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0058-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0059-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0060-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0063-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0064-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0065-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0066-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0067-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0068-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0069-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0070-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0071-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0072-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0073-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0074-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0075-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0076-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0077-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0079-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0080-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0081-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0082-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0083-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0084-C	Cameroon	Zembe-Borongo	Nora Besansky	2009	S	F
AN0085-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0086-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0087-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0088-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0089-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0090-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0091-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0092-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0093-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0094-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0095-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0096-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0097-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0098-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0099-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0100-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0101-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0102-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0103-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0104-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0105-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0106-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0107-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0108-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0109-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0111-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0112-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0113-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0114-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F

AN0115-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0117-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0120-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0121-C	Cameroon	Gado-Badzere	Nora Besansky	2009	S	F
AN0122-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0123-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0124-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0125-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0126-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0127-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0128-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0129-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0130-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0131-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0132-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0134-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0135-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0136-C	Cameroon	Mayos	Nora Besansky	2009	S	M
AN0137-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0138-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0139-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0140-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0141-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0143-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0147-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0149-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0151-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0152-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0153-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0154-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0155-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0156-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0157-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0158-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0159-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0160-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0162-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0163-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0164-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0165-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0166-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0167-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0168-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0169-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0170-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0171-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0172-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0173-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0174-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0175-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0176-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0177-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0178-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0179-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0180-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0181-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0182-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0183-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0184-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0185-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0186-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0187-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0188-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0189-C	Cameroon	Mayos	Nora Besansky	2009	S	F

AN0190-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0191-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0192-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0193-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0194-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0196-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0197-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0198-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0199-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0200-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0201-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0202-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0203-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0204-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0205-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0206-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0207-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0208-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0209-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0210-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0212-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0213-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0214-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0215-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0217-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0218-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0219-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0220-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0221-C	Cameroon	Mayos	Nora Besansky	2009	S	F
AN0222-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0223-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0224-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0225-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0226-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0227-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0228-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0229-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0230-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0231-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0233-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0234-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0235-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0236-C	Cameroon	Daiguene	Nora Besansky	2009	S	M
AN0237-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0238-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0239-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0240-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0241-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0242-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0243-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0244-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0245-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0246-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0247-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0248-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0250-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0251-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0252-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0253-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0254-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0255-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0256-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0258-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0259-C	Cameroon	Daiguene	Nora Besansky	2009	S	F

AN0260-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0261-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0262-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0263-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0264-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0266-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0267-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0268-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0269-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0270-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0272-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0275-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0276-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0277-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0280-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0282-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0283-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0284-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0285-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0286-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0287-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0288-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0290-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0291-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0292-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0294-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0295-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0296-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0297-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0298-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0299-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0300-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0301-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0303-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0304-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0305-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0307-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0308-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0309-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0310-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0312-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0313-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0314-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0315-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0317-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0318-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0319-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AN0321-C	Cameroon	Daiguene	Nora Besansky	2009	S	F
AR0007-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0008-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0009-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0010-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0011-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0012-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0014-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0015-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0017-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0019-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0020-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0021-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0022-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0023-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0024-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0026-C	Angola	Luanda	Joao Pinto	2009	M	F

AR0027-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0034-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0035-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0042-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0043-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0045-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0047-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0049-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0050-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0051-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0053-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0054-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0057-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0059-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0061-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0062-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0063-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0065-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0066-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0069-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0070-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0071-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0072-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0073-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0074-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0075-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0076-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0078-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0079-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0080-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0081-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0083-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0084-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0086-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0087-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0089-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0090-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0092-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0093-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0095-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0096-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0098-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0099-C	Angola	Luanda	Joao Pinto	2009	M	F
AR0100-C	Angola	Luanda	Joao Pinto	2009	M	F
AS0001-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0002-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0003-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0004-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0006-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0007-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0008-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0009-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0010-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0011-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0012-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0013-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0014-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0015-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0016-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0017-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0018-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0019-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0020-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0021-C	Gabon	Libreville	Joao Pinto	2000	S	F

AS0022-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0024-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0026-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0028-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0030-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0032-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0033-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0034-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0035-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0036-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0037-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0039-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0042-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0044-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0045-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0047-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0049-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0052-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0053-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0054-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0055-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0056-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0058-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0059-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0064-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0065-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0066-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0068-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0069-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0070-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0071-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0072-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0073-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0074-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0076-C	Gabon	Libreville	Joao Pinto	2000	S	F
AS0077-C	Gabon	Libreville	Joao Pinto	2000	S	F
AV0001-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0002-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0003-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0004-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0005-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0007-C	Guinea	Koraboh	Kenneth Vernick	2012	M/S	F
AV0008-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0009-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0010-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0011-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0012-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0013-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0014-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0015-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0018-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0024-C	Guinea	Koraboh	Kenneth Vernick	2012	S	F
AV0026-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0027-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0029-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0030-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0031-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0032-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0033-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0034-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0035-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0036-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0039-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0041-C	Guinea	Koundara	Kenneth Vernick	2012	S	F

AV0044-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0045-C	Guinea	Koundara	Kenneth Vernick	2012	S	F
AV0047-C	Guinea	Koundara	Kenneth Vernick	2012	S	F

A4 : 2Rb karyotype calls, Ag1kG-AR1 set

ind	country	P.std.a	P.het.a	P.inv.a	cyto	svc	tag	pca
AB0085-C	Burkina Faso	0.04	0.06	0.90	NA	bb	bb	bb
AB0087-C	Burkina Faso	0.04	0.14	0.81	NA	b+	b+	b+
AB0088-C	Burkina Faso	0.03	0.07	0.90	NA	++	++	++
AB0089-C	Burkina Faso	0.04	0.08	0.88	NA	b+	b+	b+
AB0090-C	Burkina Faso	0.06	0.08	0.86	NA	++	++	++
AB0091-C	Burkina Faso	0.04	0.06	0.90	NA	++	++	++
AB0092-C	Burkina Faso	0.04	0.08	0.88	NA	b+	b+	b+
AB0094-C	Burkina Faso	0.04	0.05	0.91	NA	b+	b+	b+
AB0095-C	Burkina Faso	0.04	0.13	0.83	NA	++	++	++
AB0097-C	Burkina Faso	0.04	0.07	0.89	NA	++	++	++
AB0098-C	Burkina Faso	0.04	0.06	0.89	NA	b+	b+	b+
AB0099-C	Burkina Faso	0.04	0.07	0.89	NA	++	++	++
AB0100-C	Burkina Faso	0.04	0.06	0.90	NA	++	++	++
AB0101-C	Burkina Faso	0.04	0.07	0.89	NA	++	++	++
AB0103-C	Burkina Faso	0.06	0.16	0.78	NA	bb	bb	bb
AB0104-C	Burkina Faso	0.04	0.05	0.91	NA	bb	bb	bb
AB0109-C	Burkina Faso	0.04	0.07	0.88	NA	++	++	++
AB0110-C	Burkina Faso	0.05	0.19	0.76	NA	b+	b+	b+
AB0111-C	Burkina Faso	0.07	0.14	0.80	NA	++	++	++
AB0112-C	Burkina Faso	0.08	0.27	0.64	NA	++	++	++
AB0113-C	Burkina Faso	0.04	0.08	0.88	NA	bb	bb	bb
AB0114-C	Burkina Faso	0.04	0.08	0.88	NA	b+	b+	b+
AB0117-C	Burkina Faso	0.05	0.93	0.02	NA	bb	bb	bb
AB0119-C	Burkina Faso	0.04	0.06	0.90	NA	bb	bb	bb
AB0122-C	Burkina Faso	0.04	0.11	0.85	NA	++	++	++
AB0123-C	Burkina Faso	0.05	0.26	0.69	NA	++	++	++
AB0124-C	Burkina Faso	0.04	0.05	0.91	NA	b+	b+	b+
AB0126-C	Burkina Faso	0.14	0.84	0.02	NA	bb	bb	bb
AB0127-C	Burkina Faso	0.04	0.11	0.85	NA	bb	bb	bb
AB0128-C	Burkina Faso	0.04	0.09	0.88	NA	bb	bb	bb
AB0129-C	Burkina Faso	0.04	0.06	0.90	NA	b+	b+	b+
AB0130-C	Burkina Faso	0.07	0.17	0.76	NA	++	b+	b+
AB0133-C	Burkina Faso	0.04	0.06	0.90	NA	bb	bb	bb
AB0134-C	Burkina Faso	0.05	0.22	0.72	NA	bb	b+	b+
AB0135-C	Burkina Faso	0.04	0.04	0.92	NA	bb	bb	bb
AB0136-C	Burkina Faso	0.04	0.07	0.89	NA	bb	bb	bb
AB0137-C	Burkina Faso	0.04	0.09	0.87	NA	bb	b+	b+
AB0138-C	Burkina Faso	0.06	0.32	0.62	NA	b+	b+	b+
AB0139-C	Burkina Faso	0.05	0.06	0.89	NA	++	++	++
AB0140-C	Burkina Faso	0.05	0.13	0.82	NA	++	++	++
AB0142-C	Burkina Faso	0.04	0.06	0.90	NA	++	++	++
AB0143-C	Burkina Faso	0.04	0.06	0.89	NA	b+	bb	bb
AB0145-C	Burkina Faso	0.03	0.94	0.03	NA	bb	bb	bb
AB0146-C	Burkina Faso	0.04	0.24	0.71	NA	bb	bb	bb
AB0147-C	Burkina Faso	0.04	0.06	0.90	NA	b+	b+	b+
AB0148-C	Burkina Faso	0.04	0.94	0.01	NA	bb	bb	bb
AB0151-C	Burkina Faso	0.08	0.89	0.02	NA	b+	b+	b+
AB0153-C	Burkina Faso	0.29	0.68	0.03	NA	b+	b+	b+
AB0155-C	Burkina Faso	0.04	0.06	0.90	NA	b+	b+	b+
AB0157-C	Burkina Faso	0.03	0.95	0.02	NA	b+	b+	b+
AB0158-C	Burkina Faso	0.05	0.12	0.83	NA	b+	b+	b+
AB0159-C	Burkina Faso	0.04	0.05	0.91	NA	bb	bb	bb
AB0160-C	Burkina Faso	0.05	0.07	0.88	NA	bb	bb	bb
AB0161-C	Burkina Faso	0.04	0.09	0.87	NA	bb	bb	bb
AB0164-C	Burkina Faso	0.05	0.17	0.78	NA	++	++	++
AB0166-C	Burkina Faso	0.05	0.12	0.84	NA	b+	b+	b+
AB0169-C	Burkina Faso	0.05	0.24	0.71	NA	b+	bb	bb
AB0170-C	Burkina Faso	0.34	0.63	0.03	NA	b+	b+	b+
AB0171-C	Burkina Faso	0.05	0.13	0.82	NA	b+	b+	b+
AB0172-C	Burkina Faso	0.04	0.04	0.92	NA	bb	bb	bb

AB0173-C	Burkina Faso	0.04	0.13	0.83	NA	bb	bb	bb
AB0174-C	Burkina Faso	0.04	0.06	0.90	NA	b+	b+	b+
AB0175-C	Burkina Faso	0.03	0.07	0.90	NA	bb	bb	bb
AB0176-C	Burkina Faso	0.06	0.18	0.76	NA	bb	bb	bb
AB0177-C	Burkina Faso	0.04	0.06	0.89	NA	bb	bb	bb
AB0178-C	Burkina Faso	0.04	0.06	0.90	NA	b+	b+	b+
AB0179-C	Burkina Faso	0.04	0.07	0.89	NA	b+	b+	b+
AB0181-C	Burkina Faso	0.05	0.10	0.85	NA	++	++	++
AB0182-C	Burkina Faso	0.09	0.25	0.66	NA	b+	b+	b+
AB0183-C	Burkina Faso	0.05	0.06	0.89	NA	b+	b+	b+
AB0184-C	Burkina Faso	0.05	0.06	0.89	NA	++	++	++
AB0185-C	Burkina Faso	0.04	0.05	0.91	NA	++	++	++
AB0186-C	Burkina Faso	0.04	0.06	0.89	NA	b+	b+	b+
AB0187-C	Burkina Faso	0.08	0.43	0.49	NA	++	++	++
AB0188-C	Burkina Faso	0.04	0.06	0.90	NA	bb	bb	bb
AB0189-C	Burkina Faso	0.04	0.04	0.92	NA	b+	b+	b+
AB0190-C	Burkina Faso	0.04	0.06	0.90	NA	bb	b+	b+
AB0191-C	Burkina Faso	0.04	0.05	0.91	NA	b+	b+	b+
AB0192-C	Burkina Faso	0.04	0.06	0.90	NA	bb	b+	b+
AB0197-C	Burkina Faso	0.04	0.10	0.85	NA	bb	bb	bb
AB0198-C	Burkina Faso	0.05	0.06	0.89	NA	b+	b+	b+
AB0199-C	Burkina Faso	0.04	0.05	0.91	NA	b+	b+	b+
AB0201-C	Burkina Faso	0.04	0.04	0.92	NA	bb	bb	bb
AB0202-C	Burkina Faso	0.04	0.05	0.90	NA	bb	bb	bb
AB0203-C	Burkina Faso	0.06	0.12	0.82	NA	bb	bb	bb
AB0204-C	Burkina Faso	0.05	0.09	0.86	NA	++	++	++
AB0205-C	Burkina Faso	0.25	0.72	0.03	NA	b+	bb	bb
AB0206-C	Burkina Faso	0.04	0.04	0.92	NA	b+	bb	bb
AB0207-C	Burkina Faso	0.05	0.04	0.91	NA	b+	bb	bb
AB0208-C	Burkina Faso	0.05	0.07	0.88	NA	b+	bb	bb
AB0209-C	Burkina Faso	0.04	0.19	0.76	NA	++	++	++
AB0210-C	Burkina Faso	0.04	0.06	0.90	NA	b+	b+	b+
AB0211-C	Burkina Faso	0.06	0.07	0.87	NA	b+	b+	b+
AB0212-C	Burkina Faso	0.03	0.06	0.90	NA	b+	b+	b+
AB0213-C	Burkina Faso	0.04	0.10	0.86	NA	++	++	++
AB0217-C	Burkina Faso	0.04	0.06	0.90	NA	b+	b+	b+
AB0219-C	Burkina Faso	0.04	0.10	0.86	NA	++	++	++
AB0221-C	Burkina Faso	0.04	0.08	0.89	NA	++	++	++
AB0222-C	Burkina Faso	0.08	0.22	0.71	NA	bb	bb	bb
AB0223-C	Burkina Faso	0.05	0.05	0.90	NA	b+	b+	b+
AB0224-C	Burkina Faso	0.13	0.86	0.02	NA	b+	b+	b+
AB0226-C	Burkina Faso	0.06	0.17	0.77	NA	++	b+	b+
AB0227-C	Burkina Faso	0.05	0.13	0.82	NA	b+	b+	b+
AB0228-C	Burkina Faso	0.04	0.07	0.89	NA	bb	bb	bb
AB0229-C	Burkina Faso	0.03	0.07	0.90	NA	b+	b+	b+
AB0231-C	Burkina Faso	0.75	0.21	0.04	NA	bb	bb	bb
AB0233-C	Burkina Faso	0.07	0.91	0.02	NA	bb	bb	bb
AB0234-C	Burkina Faso	0.07	0.43	0.51	NA	++	++	++
AB0235-C	Burkina Faso	0.07	0.17	0.76	NA	bb	bb	bb
AB0236-C	Burkina Faso	0.05	0.12	0.84	NA	b+	b+	b+
AB0237-C	Burkina Faso	0.06	0.13	0.81	NA	++	++	++
AB0238-C	Burkina Faso	0.05	0.11	0.84	NA	b+	bb	bb
AB0239-C	Burkina Faso	0.06	0.11	0.83	NA	bb	bb	bb
AB0240-C	Burkina Faso	0.04	0.07	0.89	NA	++	++	++
AB0241-C	Burkina Faso	0.04	0.06	0.91	NA	bb	bb	bb
AB0242-C	Burkina Faso	0.04	0.04	0.92	NA	++	++	++
AB0243-C	Burkina Faso	0.05	0.08	0.87	NA	++	++	++
AB0244-C	Burkina Faso	0.04	0.06	0.90	NA	bb	bb	bb
AB0246-C	Burkina Faso	0.04	0.12	0.84	NA	b+	b+	b+
AB0249-C	Burkina Faso	0.04	0.06	0.89	NA	++	++	++
AB0250-C	Burkina Faso	0.09	0.41	0.51	NA	++	++	++
AB0251-C	Burkina Faso	0.10	0.88	0.02	NA	bb	bb	bb
AB0252-C	Burkina Faso	0.04	0.08	0.88	NA	b+	b+	b+
AB0253-C	Burkina Faso	0.07	0.16	0.77	NA	bb	bb	bb

AB0256-C	Burkina Faso	0.07	0.91	0.02	NA	bb	bb	bb
AB0257-C	Burkina Faso	0.04	0.08	0.88	NA	bb	bb	bb
AB0258-C	Burkina Faso	0.05	0.19	0.76	NA	bb	bb	bb
AB0260-C	Burkina Faso	0.04	0.07	0.89	NA	b+	b+	b+
AB0261-C	Burkina Faso	0.36	0.57	0.07	NA	bb	bb	bb
AB0262-C	Burkina Faso	0.04	0.05	0.91	NA	bb	bb	bb
AB0263-C	Burkina Faso	0.05	0.06	0.89	NA	bb	b+	b+
AB0264-C	Burkina Faso	0.05	0.11	0.84	NA	bb	bb	bb
AB0265-C	Burkina Faso	0.09	0.16	0.75	NA	bb	bb	bb
AB0266-C	Burkina Faso	0.10	0.16	0.74	NA	bb	bb	bb
AB0267-C	Burkina Faso	0.04	0.06	0.91	NA	++	++	++
AB0268-C	Burkina Faso	0.06	0.10	0.84	NA	bb	bb	bb
AB0270-C	Burkina Faso	0.03	0.95	0.02	NA	b+	b+	b+
AB0271-C	Burkina Faso	0.05	0.09	0.86	NA	bb	bb	bb
AB0272-C	Burkina Faso	0.05	0.04	0.91	NA	bb	bb	bb
AB0273-C	Burkina Faso	0.04	0.08	0.88	NA	bb	bb	bb
AB0274-C	Burkina Faso	0.06	0.09	0.85	NA	bb	bb	bb
AB0276-C	Burkina Faso	0.10	0.13	0.77	NA	b+	b+	b+
AB0277-C	Burkina Faso	0.05	0.09	0.87	NA	b+	b+	b+
AB0278-C	Burkina Faso	0.04	0.07	0.89	NA	b+	b+	b+
AB0279-C	Burkina Faso	0.05	0.08	0.87	NA	b+	b+	b+
AB0280-C	Burkina Faso	0.03	0.08	0.88	NA	bb	bb	bb
AB0281-C	Burkina Faso	0.11	0.16	0.73	NA	b+	++	++
AB0282-C	Burkina Faso	0.04	0.09	0.86	NA	b+	b+	b+
AB0283-C	Burkina Faso	0.04	0.04	0.92	NA	b+	bb	bb
AB0284-C	Burkina Faso	0.04	0.05	0.91	NA	bb	bb	bb
AC0090-C	Uganda	0.06	0.82	0.12	NA	++	++	++
AC0091-C	Uganda	0.06	0.82	0.12	NA	b+	b+	b+
AC0092-C	Uganda	0.04	0.11	0.85	NA	bb	b+	b+
AC0093-C	Uganda	0.15	0.83	0.02	NA	++	++	++
AC0094-C	Uganda	0.14	0.79	0.07	NA	++	++	++
AC0095-C	Uganda	0.08	0.18	0.75	NA	b+	++	++
AC0096-C	Uganda	0.05	0.90	0.04	NA	++	++	++
AC0097-C	Uganda	0.06	0.10	0.85	NA	++	b+	b+
AC0098-C	Uganda	0.07	0.28	0.65	NA	++	++	++
AC0099-C	Uganda	0.81	0.15	0.04	NA	++	++	++
AC0100-C	Uganda	0.06	0.72	0.22	NA	++	++	++
AC0101-C	Uganda	0.28	0.69	0.03	NA	++	++	++
AC0102-C	Uganda	0.96	0.02	0.02	NA	++	++	++
AC0103-C	Uganda	0.04	0.08	0.88	NA	++	++	++
AC0104-C	Uganda	0.17	0.80	0.03	NA	++	++	++
AC0106-C	Uganda	0.04	0.08	0.88	NA	++	++	++
AC0107-C	Uganda	0.94	0.03	0.03	NA	++	++	++
AC0108-C	Uganda	0.04	0.05	0.91	NA	++	++	++
AC0109-C	Uganda	0.05	0.06	0.89	NA	b+	b+	b+
AC0110-C	Uganda	0.07	0.68	0.25	NA	++	++	++
AC0111-C	Uganda	0.06	0.38	0.56	NA	bb	b+	b+
AC0112-C	Uganda	0.05	0.75	0.20	NA	++	++	++
AC0113-C	Uganda	0.10	0.77	0.13	NA	++	++	++
AC0114-C	Uganda	0.06	0.92	0.02	NA	b+	b+	b+
AC0115-C	Uganda	0.05	0.11	0.84	NA	++	++	++
AC0116-C	Uganda	0.04	0.06	0.90	NA	++	++	++
AC0117-C	Uganda	0.21	0.70	0.08	NA	++	++	++
AC0118-C	Uganda	0.19	0.78	0.03	NA	bb	b+	b+
AC0119-C	Uganda	0.41	0.52	0.07	NA	++	++	++
AC0120-C	Uganda	0.95	0.03	0.02	NA	b+	b+	b+
AC0121-C	Uganda	0.04	0.04	0.92	NA	++	++	++
AC0122-C	Uganda	0.07	0.82	0.11	NA	++	++	++
AC0123-C	Uganda	0.08	0.81	0.11	NA	++	++	++
AC0124-C	Uganda	0.05	0.07	0.88	NA	++	++	++
AC0125-C	Uganda	0.04	0.79	0.17	NA	b+	b+	b+
AC0126-C	Uganda	0.83	0.12	0.05	NA	++	++	++
AC0127-C	Uganda	0.04	0.07	0.89	NA	++	++	++
AC0128-C	Uganda	0.03	0.95	0.03	NA	++	++	++

AC0129-C	Uganda	0.04	0.06	0.90	NA	bb	bb	bb
AC0130-C	Uganda	0.04	0.10	0.85	NA	++	++	++
AC0131-C	Uganda	0.05	0.85	0.10	NA	++	++	++
AC0132-C	Uganda	0.05	0.05	0.90	NA	++	++	++
AC0133-C	Uganda	0.21	0.71	0.08	NA	bb	b+	b+
AC0135-C	Uganda	0.09	0.90	0.01	NA	++	++	++
AC0136-C	Uganda	0.09	0.83	0.08	NA	++	++	++
AC0137-C	Uganda	0.04	0.05	0.91	NA	++	++	++
AC0138-C	Uganda	0.46	0.45	0.09	NA	++	++	++
AC0139-C	Uganda	0.90	0.07	0.03	NA	++	++	++
AC0140-C	Uganda	0.04	0.07	0.90	NA	++	++	++
AC0142-C	Uganda	0.07	0.81	0.13	NA	bb	b+	b+
AC0143-C	Uganda	0.04	0.08	0.88	NA	++	++	++
AC0144-C	Uganda	0.07	0.20	0.73	NA	++	++	++
AC0145-C	Uganda	0.06	0.93	0.01	NA	++	++	++
AC0147-C	Uganda	0.04	0.07	0.90	NA	++	++	++
AC0148-C	Uganda	0.05	0.13	0.83	NA	++	++	++
AC0149-C	Uganda	0.05	0.77	0.18	NA	++	++	++
AC0150-C	Uganda	0.59	0.35	0.06	NA	++	++	++
AC0151-C	Uganda	0.03	0.96	0.01	NA	++	++	++
AC0152-C	Uganda	0.06	0.93	0.02	NA	++	++	++
AC0153-C	Uganda	0.69	0.25	0.05	NA	++	++	++
AC0154-C	Uganda	0.07	0.67	0.26	NA	++	++	++
AC0156-C	Uganda	0.06	0.87	0.06	NA	bb	b+	b+
AC0158-C	Uganda	0.06	0.92	0.02	NA	b+	b+	b+
AC0159-C	Uganda	0.95	0.03	0.02	NA	++	++	++
AC0160-C	Uganda	0.06	0.87	0.07	NA	++	++	++
AC0161-C	Uganda	0.06	0.35	0.59	NA	++	++	++
AC0162-C	Uganda	0.10	0.81	0.10	NA	b+	b+	b+
AC0163-C	Uganda	0.07	0.84	0.09	NA	bb	b+	b+
AC0164-C	Uganda	0.03	0.96	0.01	NA	++	++	++
AC0166-C	Uganda	0.05	0.46	0.49	NA	++	++	++
AC0167-C	Uganda	0.07	0.78	0.16	NA	++	++	++
AC0168-C	Uganda	0.05	0.88	0.06	NA	b+	b+	b+
AC0169-C	Uganda	0.06	0.13	0.80	NA	++	++	++
AC0170-C	Uganda	0.08	0.81	0.11	NA	++	++	++
AC0171-C	Uganda	0.06	0.07	0.87	NA	b+	++	++
AC0172-C	Uganda	0.04	0.11	0.85	NA	++	++	++
AC0173-C	Uganda	0.08	0.77	0.15	NA	++	++	++
AC0174-C	Uganda	0.10	0.84	0.06	NA	bb	b+	b+
AC0176-C	Uganda	0.62	0.34	0.04	NA	++	++	++
AC0178-C	Uganda	0.26	0.62	0.12	NA	++	++	++
AC0179-C	Uganda	0.05	0.05	0.90	NA	++	++	++
AC0180-C	Uganda	0.04	0.08	0.88	NA	++	++	++
AC0181-C	Uganda	0.62	0.32	0.06	NA	b+	b+	b+
AC0182-C	Uganda	0.04	0.05	0.91	NA	++	++	++
AC0183-C	Uganda	0.18	0.79	0.04	NA	bb	bb	bb
AC0184-C	Uganda	0.05	0.10	0.85	NA	++	++	++
AC0186-C	Uganda	0.85	0.11	0.04	NA	++	++	++
AC0187-C	Uganda	0.07	0.68	0.25	NA	++	++	++
AC0188-C	Uganda	0.08	0.81	0.11	NA	++	++	++
AC0189-C	Uganda	0.04	0.05	0.91	NA	b+	b+	b+
AC0190-C	Uganda	0.04	0.05	0.91	NA	++	++	++
AC0191-C	Uganda	0.04	0.06	0.90	NA	++	++	++
AC0192-C	Uganda	0.05	0.16	0.79	NA	++	++	++
AC0193-C	Uganda	0.04	0.11	0.85	NA	++	++	++
AC0194-C	Uganda	0.04	0.05	0.91	NA	b+	b+	b+
AC0195-C	Uganda	0.07	0.81	0.12	NA	++	++	++
AC0196-C	Uganda	0.04	0.06	0.90	NA	++	++	++
AC0197-C	Uganda	0.03	0.07	0.89	NA	++	++	++
AC0199-C	Uganda	0.06	0.86	0.08	NA	++	++	++
AC0200-C	Uganda	0.07	0.71	0.23	NA	++	++	++
AC0201-C	Uganda	0.04	0.06	0.91	NA	++	++	++
AC0202-C	Uganda	0.13	0.86	0.02	NA	++	++	++

AC0203-C	Uganda	0.06	0.66	0.27	NA	++	b+	b+
AJ0023-C	Guinea-Bissau	0.86	0.11	0.03	NA	++	++	NA
AJ0024-C	Guinea-Bissau	0.95	0.03	0.02	NA	b+	++	NA
AJ0032-C	Guinea-Bissau	0.04	0.13	0.83	NA	++	++	NA
AJ0035-C	Guinea-Bissau	0.04	0.94	0.02	NA	++	++	NA
AJ0036-C	Guinea-Bissau	0.81	0.16	0.03	NA	++	++	NA
AJ0039-C	Guinea-Bissau	0.93	0.05	0.02	NA	b+	++	NA
AJ0043-C	Guinea-Bissau	0.07	0.90	0.04	NA	++	++	NA
AJ0044-C	Guinea-Bissau	0.18	0.79	0.03	NA	++	++	NA
AJ0045-C	Guinea-Bissau	0.04	0.07	0.89	NA	++	++	NA
AJ0047-C	Guinea-Bissau	0.06	0.91	0.03	NA	++	++	NA
AJ0051-C	Guinea-Bissau	0.04	0.91	0.05	NA	++	++	NA
AJ0052-C	Guinea-Bissau	0.91	0.06	0.03	NA	++	++	NA
AJ0056-C	Guinea-Bissau	0.92	0.05	0.03	NA	bb	++	NA
AJ0061-C	Guinea-Bissau	0.31	0.65	0.04	NA	++	++	NA
AJ0063-C	Guinea-Bissau	0.04	0.93	0.03	NA	++	++	NA
AJ0064-C	Guinea-Bissau	0.26	0.71	0.03	NA	++	++	NA
AJ0066-C	Guinea-Bissau	0.68	0.28	0.04	NA	++	++	NA
AJ0070-C	Guinea-Bissau	0.93	0.04	0.03	NA	++	++	NA
AJ0071-C	Guinea-Bissau	0.93	0.04	0.02	NA	++	++	NA
AJ0072-C	Guinea-Bissau	0.87	0.09	0.04	NA	b+	++	NA
AJ0074-C	Guinea-Bissau	0.04	0.07	0.89	NA	++	++	NA
AJ0075-C	Guinea-Bissau	0.05	0.93	0.02	NA	++	++	NA
AJ0076-C	Guinea-Bissau	0.06	0.60	0.34	NA	++	++	NA
AJ0077-C	Guinea-Bissau	0.07	0.87	0.06	NA	++	++	NA
AJ0078-C	Guinea-Bissau	0.04	0.94	0.02	NA	++	++	NA
AJ0081-C	Guinea-Bissau	0.08	0.90	0.02	NA	++	++	NA
AJ0084-C	Guinea-Bissau	0.04	0.95	0.02	NA	b+	++	NA
AJ0085-C	Guinea-Bissau	0.81	0.16	0.02	NA	++	++	NA
AJ0086-C	Guinea-Bissau	0.89	0.06	0.05	NA	++	++	NA
AJ0088-C	Guinea-Bissau	0.71	0.26	0.03	NA	++	++	NA
AJ0090-C	Guinea-Bissau	0.12	0.83	0.05	NA	++	++	NA
AJ0092-C	Guinea-Bissau	0.26	0.72	0.02	NA	++	++	NA
AJ0093-C	Guinea-Bissau	0.05	0.93	0.02	NA	b+	++	NA
AJ0096-C	Guinea-Bissau	0.91	0.06	0.03	NA	++	++	NA
AJ0097-C	Guinea-Bissau	0.04	0.95	0.01	NA	++	++	NA
AJ0098-C	Guinea-Bissau	0.95	0.03	0.02	NA	b+	++	NA
AJ0100-C	Guinea-Bissau	0.95	0.03	0.02	NA	++	++	NA
AJ0101-C	Guinea-Bissau	0.94	0.03	0.03	NA	++	++	NA
AJ0102-C	Guinea-Bissau	0.94	0.03	0.03	NA	++	++	NA
AJ0103-C	Guinea-Bissau	0.06	0.92	0.02	NA	++	++	NA
AJ0105-C	Guinea-Bissau	0.87	0.09	0.04	NA	++	++	NA
AJ0107-C	Guinea-Bissau	0.62	0.34	0.04	NA	++	++	NA
AJ0109-C	Guinea-Bissau	0.69	0.23	0.07	NA	++	++	NA
AJ0113-C	Guinea-Bissau	0.06	0.91	0.02	NA	++	++	NA
AJ0115-C	Guinea-Bissau	0.54	0.40	0.06	NA	b+	++	NA
AJ0116-C	Guinea-Bissau	0.06	0.92	0.03	NA	++	++	NA
AK0065-C	Kenya	0.05	0.06	0.89	NA	++	++	NA
AK0066-C	Kenya	0.04	0.05	0.92	NA	++	++	NA
AK0067-C	Kenya	0.96	0.02	0.02	NA	++	++	NA
AK0068-C	Kenya	0.04	0.90	0.06	NA	++	++	NA
AK0069-C	Kenya	0.51	0.46	0.03	NA	++	++	NA
AK0070-C	Kenya	0.04	0.05	0.91	NA	++	++	NA
AK0072-C	Kenya	0.04	0.78	0.17	NA	++	++	NA
AK0073-C	Kenya	0.13	0.84	0.02	NA	++	++	NA
AK0074-C	Kenya	0.28	0.69	0.03	NA	++	++	NA
AK0075-C	Kenya	0.04	0.86	0.09	NA	++	++	NA
AK0076-C	Kenya	0.81	0.16	0.03	NA	++	++	NA
AK0077-C	Kenya	0.03	0.95	0.02	NA	++	++	NA
AK0078-C	Kenya	0.95	0.03	0.02	NA	++	++	NA
AK0079-C	Kenya	0.38	0.59	0.03	NA	++	++	NA
AK0080-C	Kenya	0.11	0.88	0.02	NA	++	++	NA
AK0081-C	Kenya	0.96	0.02	0.02	NA	++	++	NA
AK0082-C	Kenya	0.05	0.93	0.01	NA	++	++	NA

AK0085-C	Kenya	0.09	0.68	0.23	NA	++	++	NA
AK0086-C	Kenya	0.06	0.81	0.13	NA	++	++	NA
AK0087-C	Kenya	0.10	0.11	0.79	NA	++	++	NA
AK0088-C	Kenya	0.05	0.59	0.36	NA	++	++	NA
AK0089-C	Kenya	0.04	0.09	0.87	NA	++	++	NA
AK0090-C	Kenya	0.06	0.92	0.02	NA	++	++	NA
AK0091-C	Kenya	0.12	0.85	0.03	NA	++	++	NA
AK0092-C	Kenya	0.11	0.21	0.69	NA	++	++	NA
AK0093-C	Kenya	0.10	0.11	0.79	NA	++	++	NA
AK0094-C	Kenya	0.07	0.12	0.81	NA	++	++	NA
AK0095-C	Kenya	0.04	0.93	0.03	NA	++	++	NA
AK0096-C	Kenya	0.05	0.06	0.88	NA	++	++	NA
AK0098-C	Kenya	0.05	0.10	0.86	NA	++	++	NA
AK0099-C	Kenya	0.06	0.07	0.87	NA	++	++	NA
AK0100-C	Kenya	0.03	0.95	0.01	NA	++	++	NA
AK0101-C	Kenya	0.05	0.07	0.89	NA	++	++	NA
AK0102-C	Kenya	0.06	0.10	0.84	NA	++	++	NA
AK0103-C	Kenya	0.95	0.03	0.02	NA	++	++	NA
AK0104-C	Kenya	0.03	0.96	0.01	NA	++	++	NA
AK0105-C	Kenya	0.49	0.48	0.03	NA	++	++	NA
AK0106-C	Kenya	0.96	0.02	0.02	NA	++	++	NA
AK0108-C	Kenya	0.05	0.94	0.02	NA	++	++	NA
AK0109-C	Kenya	0.05	0.05	0.90	NA	++	++	NA
AK0110-C	Kenya	0.05	0.06	0.89	NA	++	++	NA
AK0116-C	Kenya	0.06	0.33	0.61	NA	b+	++	NA
AK0119-C	Kenya	0.94	0.04	0.02	NA	b+	++	NA
AK0127-C	Kenya	0.08	0.10	0.82	NA	++	++	NA
AN0007-C	Cameroon	0.87	0.11	0.03	NA	++	++	++
AN0008-C	Cameroon	0.92	0.05	0.03	NA	b+	b+	b+
AN0009-C	Cameroon	0.04	0.95	0.01	NA	b+	b+	b+
AN0010-C	Cameroon	0.92	0.04	0.03	NA	++	++	++
AN0011-C	Cameroon	0.91	0.05	0.04	NA	b+	b+	b+
AN0012-C	Cameroon	0.93	0.04	0.04	NA	++	++	++
AN0014-C	Cameroon	0.95	0.03	0.03	NA	++	++	++
AN0016-C	Cameroon	0.47	0.49	0.04	NA	++	b+	b+
AN0017-C	Cameroon	0.05	0.94	0.01	NA	bb	bb	bb
AN0018-C	Cameroon	0.88	0.08	0.04	NA	++	++	++
AN0019-C	Cameroon	0.91	0.05	0.04	NA	++	++	++
AN0020-C	Cameroon	0.73	0.19	0.08	NA	b+	b+	b+
AN0022-C	Cameroon	0.08	0.90	0.02	NA	++	b+	b+
AN0023-C	Cameroon	0.07	0.54	0.39	NA	++	b+	b+
AN0024-C	Cameroon	0.06	0.37	0.57	NA	b+	bb	bb
AN0025-C	Cameroon	0.09	0.89	0.02	NA	++	++	++
AN0026-C	Cameroon	0.04	0.05	0.92	NA	b+	b+	b+
AN0027-C	Cameroon	0.02	0.96	0.02	NA	b+	bb	bb
AN0028-C	Cameroon	0.93	0.04	0.03	NA	b+	b+	b+
AN0029-C	Cameroon	0.51	0.45	0.04	NA	++	b+	b+
AN0030-C	Cameroon	0.06	0.07	0.87	NA	++	++	++
AN0031-C	Cameroon	0.11	0.87	0.02	NA	++	b+	b+
AN0032-C	Cameroon	0.04	0.93	0.02	NA	b+	bb	bb
AN0033-C	Cameroon	0.07	0.91	0.02	NA	++	b+	b+
AN0034-C	Cameroon	0.05	0.14	0.81	NA	++	b+	b+
AN0035-C	Cameroon	0.94	0.03	0.03	NA	++	b+	b+
AN0036-C	Cameroon	0.06	0.08	0.85	NA	bb	bb	bb
AN0037-C	Cameroon	0.04	0.93	0.03	NA	++	b+	b+
AN0038-C	Cameroon	0.05	0.93	0.02	NA	b+	b+	b+
AN0039-C	Cameroon	0.04	0.95	0.01	NA	b+	b+	b+
AN0040-C	Cameroon	0.07	0.18	0.75	NA	bb	bb	bb
AN0041-C	Cameroon	0.05	0.93	0.02	NA	b+	b+	b+
AN0042-C	Cameroon	0.07	0.91	0.02	NA	b+	bb	bb
AN0043-C	Cameroon	0.04	0.08	0.88	NA	bb	bb	bb
AN0045-C	Cameroon	0.09	0.88	0.03	NA	bb	b+	b+
AN0046-C	Cameroon	0.05	0.06	0.89	NA	b+	b+	b+
AN0047-C	Cameroon	0.04	0.08	0.88	NA	bb	bb	bb

AN0048-C	Cameroon	0.05	0.07	0.88	NA	bb	bb	bb
AN0049-C	Cameroon	0.04	0.06	0.90	NA	bb	bb	bb
AN0050-C	Cameroon	0.03	0.96	0.01	NA	bb	bb	bb
AN0051-C	Cameroon	0.28	0.70	0.03	NA	b+	b+	b+
AN0053-C	Cameroon	0.08	0.31	0.60	NA	b+	bb	bb
AN0054-C	Cameroon	0.04	0.05	0.91	NA	b+	bb	bb
AN0055-C	Cameroon	0.89	0.06	0.05	NA	b+	b+	b+
AN0056-C	Cameroon	0.09	0.89	0.02	NA	bb	b+	b+
AN0057-C	Cameroon	0.69	0.25	0.06	NA	bb	bb	bb
AN0058-C	Cameroon	0.04	0.04	0.92	NA	bb	bb	bb
AN0059-C	Cameroon	0.10	0.88	0.02	NA	bb	bb	bb
AN0060-C	Cameroon	0.04	0.08	0.88	NA	b+	bb	bb
AN0063-C	Cameroon	0.27	0.71	0.02	bb	bb	bb	bb
AN0064-C	Cameroon	0.05	0.12	0.83	b+	b+	b+	b+
AN0065-C	Cameroon	0.08	0.90	0.02	bb	bb	bb	bb
AN0066-C	Cameroon	0.72	0.24	0.04	bb	bb	bb	bb
AN0067-C	Cameroon	0.05	0.09	0.87	b+	bb	b+	b+
AN0068-C	Cameroon	0.95	0.03	0.02	b+	bb	b+	b+
AN0069-C	Cameroon	0.95	0.02	0.02	bb	bb	bb	bb
AN0070-C	Cameroon	0.04	0.07	0.89	bb	bb	bb	bb
AN0071-C	Cameroon	0.04	0.09	0.87	bb	bb	bb	bb
AN0072-C	Cameroon	0.09	0.88	0.03	b+	b+	b+	b+
AN0073-C	Cameroon	0.93	0.05	0.03	b+	b+	b+	b+
AN0074-C	Cameroon	0.05	0.94	0.01	b+	b+	b+	b+
AN0075-C	Cameroon	0.94	0.03	0.02	b+	b+	b+	b+
AN0076-C	Cameroon	0.06	0.06	0.88	b+	++	b+	b+
AN0077-C	Cameroon	0.94	0.04	0.03	++	++	++	++
AN0079-C	Cameroon	0.93	0.04	0.03	b+	++	b+	b+
AN0080-C	Cameroon	0.95	0.02	0.02	++	++	++	++
AN0081-C	Cameroon	0.17	0.81	0.03	bb	bb	bb	bb
AN0082-C	Cameroon	0.04	0.05	0.90	b+	++	b+	b+
AN0083-C	Cameroon	0.08	0.90	0.02	b+	++	++	++
AN0084-C	Cameroon	0.95	0.03	0.03	b+	++	++	++
AN0085-C	Cameroon	0.88	0.08	0.04	b+	b+	b+	b+
AN0086-C	Cameroon	0.11	0.87	0.02	b+	b+	b+	b+
AN0087-C	Cameroon	0.95	0.02	0.02	++	++	++	++
AN0088-C	Cameroon	0.07	0.90	0.02	bb	bb	bb	bb
AN0089-C	Cameroon	0.05	0.08	0.87	bb	bb	bb	bb
AN0090-C	Cameroon	0.04	0.11	0.85	bb	bb	bb	bb
AN0091-C	Cameroon	0.04	0.06	0.90	b+	b+	b+	b+
AN0092-C	Cameroon	0.05	0.10	0.85	++	++	++	++
AN0093-C	Cameroon	0.06	0.08	0.85	b+	b+	b+	b+
AN0094-C	Cameroon	0.16	0.82	0.02	bb	bb	bb	bb
AN0095-C	Cameroon	0.04	0.09	0.88	b+	b+	b+	b+
AN0096-C	Cameroon	0.06	0.93	0.01	b+	b+	b+	b+
AN0097-C	Cameroon	0.30	0.68	0.03	b+	b+	b+	b+
AN0098-C	Cameroon	0.04	0.95	0.01	bb	b+	bb	bb
AN0099-C	Cameroon	0.92	0.04	0.04	bb	b+	bb	bb
AN0100-C	Cameroon	0.05	0.93	0.02	bb	b+	bb	bb
AN0101-C	Cameroon	0.10	0.88	0.02	++	++	++	++
AN0102-C	Cameroon	0.06	0.93	0.02	bb	b+	bb	bb
AN0103-C	Cameroon	0.07	0.91	0.02	bb	bb	bb	bb
AN0104-C	Cameroon	0.05	0.72	0.23	bb	b+	b+	b+
AN0105-C	Cameroon	0.05	0.08	0.87	bb	bb	bb	bb
AN0106-C	Cameroon	0.95	0.03	0.02	++	++	++	++
AN0107-C	Cameroon	0.85	0.11	0.04	b+	b+	b+	b+
AN0108-C	Cameroon	0.06	0.06	0.88	bb	bb	bb	bb
AN0109-C	Cameroon	0.28	0.69	0.02	b+	b+	b+	b+
AN0111-C	Cameroon	0.14	0.81	0.05	b+	bb	b+	b+
AN0112-C	Cameroon	0.05	0.09	0.86	bb	bb	bb	bb
AN0113-C	Cameroon	0.96	0.02	0.02	b+	b+	b+	b+
AN0114-C	Cameroon	0.05	0.06	0.89	b+	b+	b+	b+
AN0115-C	Cameroon	0.96	0.02	0.02	++	++	++	++
AN0117-C	Cameroon	0.75	0.21	0.04	bb	b+	b+	b+

AN0120-C	Cameroon	0.04	0.95	0.01	bb	bb	bb	bb
AN0121-C	Cameroon	0.04	0.06	0.91	b+	b+	b+	b+
AN0122-C	Cameroon	0.91	0.06	0.02	NA	++	++	++
AN0123-C	Cameroon	0.95	0.02	0.02	NA	++	++	++
AN0124-C	Cameroon	0.87	0.07	0.06	NA	++	++	++
AN0125-C	Cameroon	0.89	0.06	0.05	NA	++	++	++
AN0126-C	Cameroon	0.95	0.03	0.02	NA	b+	b+	b+
AN0127-C	Cameroon	0.95	0.02	0.02	NA	b+	b+	b+
AN0128-C	Cameroon	0.11	0.86	0.03	NA	++	++	++
AN0129-C	Cameroon	0.93	0.03	0.03	NA	b+	b+	b+
AN0130-C	Cameroon	0.82	0.14	0.03	NA	++	++	++
AN0131-C	Cameroon	0.95	0.02	0.02	NA	++	b+	b+
AN0132-C	Cameroon	0.94	0.03	0.03	NA	++	++	++
AN0134-C	Cameroon	0.93	0.04	0.03	NA	b+	b+	b+
AN0135-C	Cameroon	0.08	0.17	0.75	NA	++	++	++
AN0136-C	Cameroon	0.89	0.06	0.05	NA	++	++	++
AN0137-C	Cameroon	0.96	0.02	0.02	NA	++	++	++
AN0138-C	Cameroon	0.93	0.03	0.03	NA	b+	b+	b+
AN0139-C	Cameroon	0.92	0.04	0.04	NA	b+	b+	b+
AN0140-C	Cameroon	0.82	0.10	0.08	NA	++	++	++
AN0141-C	Cameroon	0.89	0.06	0.05	NA	++	++	++
AN0143-C	Cameroon	0.06	0.92	0.02	NA	b+	b+	b+
AN0147-C	Cameroon	0.92	0.05	0.04	NA	++	++	++
AN0149-C	Cameroon	0.05	0.93	0.01	NA	bb	bb	bb
AN0151-C	Cameroon	0.91	0.06	0.03	NA	++	b+	b+
AN0152-C	Cameroon	0.95	0.03	0.02	NA	b+	b+	b+
AN0153-C	Cameroon	0.89	0.06	0.05	NA	bb	b+	b+
AN0154-C	Cameroon	0.96	0.02	0.02	NA	b+	b+	b+
AN0155-C	Cameroon	0.89	0.08	0.03	NA	++	++	++
AN0156-C	Cameroon	0.26	0.72	0.03	NA	b+	b+	b+
AN0157-C	Cameroon	0.94	0.03	0.03	NA	b+	b+	b+
AN0158-C	Cameroon	0.93	0.03	0.03	NA	++	b+	b+
AN0159-C	Cameroon	0.95	0.03	0.02	NA	++	++	++
AN0160-C	Cameroon	0.95	0.03	0.02	NA	++	++	++
AN0162-C	Cameroon	0.90	0.06	0.05	NA	++	++	++
AN0163-C	Cameroon	0.95	0.03	0.03	NA	++	++	++
AN0164-C	Cameroon	0.92	0.05	0.04	NA	++	++	++
AN0165-C	Cameroon	0.11	0.87	0.02	NA	++	++	++
AN0166-C	Cameroon	0.90	0.07	0.03	NA	++	++	++
AN0167-C	Cameroon	0.95	0.03	0.02	NA	b+	b+	b+
AN0168-C	Cameroon	0.94	0.04	0.03	NA	++	++	++
AN0169-C	Cameroon	0.96	0.02	0.02	NA	b+	bb	bb
AN0170-C	Cameroon	0.94	0.04	0.03	NA	++	++	++
AN0171-C	Cameroon	0.94	0.04	0.02	NA	++	b+	b+
AN0172-C	Cameroon	0.89	0.09	0.03	NA	++	++	++
AN0173-C	Cameroon	0.96	0.02	0.02	NA	++	++	++
AN0174-C	Cameroon	0.89	0.06	0.05	NA	b+	b+	b+
AN0175-C	Cameroon	0.88	0.07	0.05	NA	++	++	++
AN0176-C	Cameroon	0.13	0.84	0.02	NA	++	++	++
AN0177-C	Cameroon	0.91	0.05	0.04	NA	bb	bb	bb
AN0178-C	Cameroon	0.95	0.03	0.02	NA	b+	b+	b+
AN0179-C	Cameroon	0.87	0.07	0.06	NA	b+	b+	b+
AN0180-C	Cameroon	0.96	0.02	0.02	NA	b+	b+	b+
AN0181-C	Cameroon	0.87	0.07	0.06	NA	++	++	++
AN0182-C	Cameroon	0.11	0.87	0.02	NA	++	++	++
AN0183-C	Cameroon	0.95	0.03	0.03	NA	++	++	++
AN0184-C	Cameroon	0.78	0.19	0.03	++	++	++	++
AN0185-C	Cameroon	0.95	0.03	0.03	b+	++	b+	b+
AN0186-C	Cameroon	0.91	0.05	0.04	++	++	++	++
AN0187-C	Cameroon	0.95	0.03	0.02	NA	bb	bb	bb
AN0188-C	Cameroon	0.83	0.10	0.07	b+	b+	b+	b+
AN0189-C	Cameroon	0.93	0.04	0.03	++	++	++	++
AN0190-C	Cameroon	0.93	0.05	0.02	++	++	++	++
AN0191-C	Cameroon	0.92	0.05	0.03	bb	bb	bb	bb

AN0192-C	Cameroon	0.91	0.07	0.03	++	++	++	++
AN0193-C	Cameroon	0.87	0.09	0.04	++	++	++	++
AN0194-C	Cameroon	0.89	0.06	0.05	b+	b+	b+	b+
AN0196-C	Cameroon	0.29	0.68	0.02	++	++	++	++
AN0197-C	Cameroon	0.95	0.03	0.02	b+	b+	b+	b+
AN0198-C	Cameroon	0.04	0.95	0.01	b+	++	b+	b+
AN0199-C	Cameroon	0.85	0.12	0.02	++	++	++	++
AN0200-C	Cameroon	0.75	0.15	0.10	++	++	++	++
AN0201-C	Cameroon	0.64	0.31	0.05	b+	b+	b+	b+
AN0202-C	Cameroon	0.90	0.07	0.03	++	++	++	++
AN0203-C	Cameroon	0.94	0.03	0.03	++	++	++	++
AN0204-C	Cameroon	0.94	0.03	0.02	++	++	++	++
AN0205-C	Cameroon	0.95	0.03	0.03	b+	b+	b+	b+
AN0206-C	Cameroon	0.94	0.03	0.03	++	++	++	++
AN0207-C	Cameroon	0.74	0.19	0.07	b+	b+	b+	b+
AN0208-C	Cameroon	0.83	0.14	0.03	b+	b+	b+	b+
AN0209-C	Cameroon	0.24	0.73	0.03	b+	b+	b+	b+
AN0210-C	Cameroon	0.90	0.06	0.04	++	b+	b+	b+
AN0212-C	Cameroon	0.88	0.06	0.05	++	++	++	++
AN0213-C	Cameroon	0.86	0.08	0.06	++	++	++	++
AN0214-C	Cameroon	0.17	0.81	0.02	b+	b+	b+	b+
AN0215-C	Cameroon	0.93	0.04	0.03	b+	b+	b+	b+
AN0217-C	Cameroon	0.78	0.12	0.09	++	++	++	++
AN0218-C	Cameroon	0.94	0.03	0.03	++	++	++	++
AN0219-C	Cameroon	0.92	0.05	0.03	++	++	++	++
AN0220-C	Cameroon	0.92	0.05	0.04	++	++	++	++
AN0221-C	Cameroon	0.88	0.07	0.06	b+	b+	b+	b+
AN0222-C	Cameroon	0.92	0.05	0.03	NA	++	++	++
AN0223-C	Cameroon	0.93	0.04	0.03	NA	++	++	++
AN0224-C	Cameroon	0.07	0.92	0.02	NA	b+	b+	b+
AN0225-C	Cameroon	0.94	0.03	0.03	NA	b+	b+	b+
AN0226-C	Cameroon	0.92	0.05	0.03	NA	++	++	++
AN0227-C	Cameroon	0.07	0.91	0.02	NA	b+	b+	b+
AN0228-C	Cameroon	0.90	0.06	0.04	NA	++	++	++
AN0229-C	Cameroon	0.92	0.05	0.04	NA	bb	bb	bb
AN0230-C	Cameroon	0.04	0.13	0.83	NA	bb	bb	bb
AN0231-C	Cameroon	0.91	0.05	0.04	NA	++	++	++
AN0233-C	Cameroon	0.41	0.54	0.05	NA	++	++	++
AN0234-C	Cameroon	0.03	0.94	0.03	NA	++	++	++
AN0235-C	Cameroon	0.41	0.55	0.05	NA	++	++	++
AN0236-C	Cameroon	0.04	0.16	0.80	NA	bb	bb	bb
AN0237-C	Cameroon	0.66	0.31	0.03	NA	b+	b+	b+
AN0238-C	Cameroon	0.04	0.07	0.89	NA	b+	b+	b+
AN0239-C	Cameroon	0.63	0.34	0.03	NA	++	++	++
AN0240-C	Cameroon	0.95	0.02	0.02	NA	++	++	++
AN0241-C	Cameroon	0.04	0.27	0.69	NA	b+	b+	b+
AN0242-C	Cameroon	0.94	0.03	0.03	NA	++	++	++
AN0243-C	Cameroon	0.04	0.05	0.92	NA	b+	bb	bb
AN0244-C	Cameroon	0.91	0.04	0.04	NA	++	++	++
AN0245-C	Cameroon	0.92	0.06	0.02	NA	++	++	++
AN0246-C	Cameroon	0.95	0.03	0.02	NA	b+	b+	b+
AN0247-C	Cameroon	0.81	0.15	0.04	NA	b+	b+	b+
AN0248-C	Cameroon	0.05	0.12	0.83	NA	b+	++	++
AN0250-C	Cameroon	0.04	0.05	0.92	NA	b+	bb	bb
AN0251-C	Cameroon	0.84	0.13	0.03	NA	b+	b+	b+
AN0252-C	Cameroon	0.80	0.15	0.05	NA	++	++	++
AN0253-C	Cameroon	0.10	0.87	0.02	NA	bb	bb	bb
AN0254-C	Cameroon	0.96	0.03	0.02	NA	++	b+	b+
AN0255-C	Cameroon	0.92	0.05	0.03	NA	++	++	++
AN0256-C	Cameroon	0.91	0.05	0.04	NA	b+	b+	b+
AN0258-C	Cameroon	0.04	0.07	0.89	NA	bb	bb	bb
AN0259-C	Cameroon	0.88	0.06	0.05	NA	++	++	++
AN0260-C	Cameroon	0.06	0.93	0.01	NA	b+	b+	b+
AN0261-C	Cameroon	0.90	0.06	0.04	NA	++	++	++

AN0262-C	Cameroon	0.11	0.88	0.02	NA	++	++	++
AN0263-C	Cameroon	0.85	0.11	0.05	NA	b+	b+	b+
AN0264-C	Cameroon	0.89	0.08	0.03	NA	++	++	++
AN0266-C	Cameroon	0.04	0.09	0.87	NA	bb	bb	bb
AN0267-C	Cameroon	0.66	0.30	0.04	NA	b+	b+	b+
AN0268-C	Cameroon	0.82	0.14	0.04	NA	b+	b+	b+
AN0269-C	Cameroon	0.04	0.12	0.84	NA	bb	bb	bb
AN0270-C	Cameroon	0.88	0.08	0.04	NA	++	++	++
AN0272-C	Cameroon	0.04	0.08	0.88	NA	bb	b+	b+
AN0275-C	Cameroon	0.08	0.89	0.03	NA	++	b+	b+
AN0276-C	Cameroon	0.04	0.05	0.91	NA	bb	bb	bb
AN0277-C	Cameroon	0.69	0.28	0.03	NA	b+	b+	b+
AN0280-C	Cameroon	0.30	0.67	0.03	bb	b+	bb	bb
AN0282-C	Cameroon	0.92	0.04	0.03	b+	b+	b+	b+
AN0283-C	Cameroon	0.89	0.06	0.05	b+	b+	b+	b+
AN0284-C	Cameroon	0.95	0.02	0.02	++	++	++	++
AN0285-C	Cameroon	0.93	0.04	0.03	++	++	++	++
AN0286-C	Cameroon	0.11	0.88	0.02	b+	b+	b+	b+
AN0287-C	Cameroon	0.18	0.80	0.02	b+	b+	b+	b+
AN0288-C	Cameroon	0.95	0.02	0.02	++	++	++	++
AN0290-C	Cameroon	0.18	0.80	0.02	bb	bb	bb	bb
AN0291-C	Cameroon	0.92	0.04	0.04	++	++	++	++
AN0292-C	Cameroon	0.88	0.06	0.05	++	++	++	++
AN0294-C	Cameroon	0.92	0.05	0.03	++	++	++	++
AN0295-C	Cameroon	0.03	0.06	0.90	bb	bb	bb	bb
AN0296-C	Cameroon	0.95	0.03	0.02	++	++	++	++
AN0297-C	Cameroon	0.91	0.05	0.04	bb	b+	b+	b+
AN0298-C	Cameroon	0.93	0.04	0.03	++	b+	b+	b+
AN0299-C	Cameroon	0.92	0.04	0.03	++	++	++	++
AN0300-C	Cameroon	0.95	0.03	0.02	++	++	++	++
AN0301-C	Cameroon	0.05	0.25	0.70	bb	bb	bb	bb
AN0303-C	Cameroon	0.04	0.95	0.02	++	++	++	++
AN0304-C	Cameroon	0.96	0.02	0.02	b+	++	b+	b+
AN0305-C	Cameroon	0.94	0.04	0.02	bb	++	b+	b+
AN0307-C	Cameroon	0.89	0.06	0.05	b+	b+	b+	b+
AN0308-C	Cameroon	0.94	0.04	0.03	++	++	++	++
AN0309-C	Cameroon	0.93	0.04	0.03	b+	b+	b+	b+
AN0310-C	Cameroon	0.03	0.96	0.01	++	++	++	++
AN0312-C	Cameroon	0.90	0.05	0.05	++	++	++	++
AN0313-C	Cameroon	0.35	0.62	0.03	++	++	++	++
AN0314-C	Cameroon	0.95	0.03	0.02	++	++	++	++
AN0315-C	Cameroon	0.96	0.03	0.02	++	++	++	++
AN0317-C	Cameroon	0.87	0.07	0.06	++	++	++	++
AN0318-C	Cameroon	0.92	0.04	0.04	++	++	++	++
AN0319-C	Cameroon	0.95	0.03	0.02	++	++	++	++
AN0321-C	Cameroon	0.84	0.12	0.04	++	++	++	++
AR0007-C	Angola	0.81	0.16	0.03	NA	b+	++	NA
AR0008-C	Angola	0.90	0.07	0.02	NA	++	++	NA
AR0009-C	Angola	0.93	0.04	0.03	NA	b+	++	NA
AR0010-C	Angola	0.85	0.13	0.02	NA	++	++	NA
AR0011-C	Angola	0.96	0.02	0.02	NA	b+	++	NA
AR0012-C	Angola	0.91	0.06	0.03	NA	++	++	NA
AR0014-C	Angola	0.10	0.88	0.02	NA	b+	++	NA
AR0015-C	Angola	0.94	0.03	0.02	NA	b+	++	NA
AR0017-C	Angola	0.96	0.03	0.02	NA	b+	++	NA
AR0019-C	Angola	0.95	0.03	0.02	NA	++	++	NA
AR0020-C	Angola	0.95	0.03	0.02	NA	b+	++	NA
AR0021-C	Angola	0.96	0.02	0.02	NA	++	++	NA
AR0022-C	Angola	0.95	0.03	0.02	NA	++	++	NA
AR0023-C	Angola	0.94	0.03	0.02	NA	++	++	NA
AR0024-C	Angola	0.95	0.03	0.02	NA	b+	++	NA
AR0026-C	Angola	0.05	0.93	0.02	NA	++	++	NA
AR0027-C	Angola	0.95	0.03	0.02	NA	b+	++	NA
AR0034-C	Angola	0.90	0.07	0.03	NA	++	++	NA

AR0035-C	Angola	0.95	0.03	0.02	NA	++	++	NA
AR0042-C	Angola	0.93	0.04	0.03	NA	++	++	NA
AR0043-C	Angola	0.95	0.03	0.02	NA	++	++	NA
AR0045-C	Angola	0.95	0.03	0.02	NA	++	++	NA
AR0047-C	Angola	0.78	0.19	0.03	NA	b+	++	NA
AR0049-C	Angola	0.54	0.43	0.03	NA	++	++	NA
AR0050-C	Angola	0.77	0.20	0.03	NA	++	++	NA
AR0051-C	Angola	0.34	0.61	0.05	NA	++	++	NA
AR0053-C	Angola	0.91	0.05	0.04	NA	b+	++	NA
AR0054-C	Angola	0.88	0.07	0.05	NA	++	++	NA
AR0057-C	Angola	0.95	0.03	0.03	NA	++	++	NA
AR0059-C	Angola	0.93	0.04	0.03	NA	++	++	NA
AR0061-C	Angola	0.96	0.02	0.02	NA	++	++	NA
AR0062-C	Angola	0.91	0.06	0.03	NA	++	++	NA
AR0063-C	Angola	0.94	0.04	0.02	NA	++	++	NA
AR0065-C	Angola	0.96	0.02	0.02	NA	++	++	NA
AR0066-C	Angola	0.96	0.02	0.02	NA	++	++	NA
AR0069-C	Angola	0.94	0.03	0.03	NA	++	++	NA
AR0070-C	Angola	0.96	0.02	0.02	NA	++	++	NA
AR0071-C	Angola	0.94	0.04	0.02	NA	++	++	NA
AR0072-C	Angola	0.87	0.10	0.03	NA	++	++	NA
AR0073-C	Angola	0.86	0.10	0.03	NA	++	++	NA
AR0074-C	Angola	0.96	0.02	0.02	NA	++	++	NA
AR0075-C	Angola	0.94	0.04	0.02	NA	++	++	NA
AR0076-C	Angola	0.92	0.06	0.02	NA	++	++	NA
AR0078-C	Angola	0.94	0.04	0.03	NA	++	++	NA
AR0079-C	Angola	0.93	0.05	0.02	NA	++	++	NA
AR0080-C	Angola	0.95	0.03	0.02	NA	++	++	NA
AR0081-C	Angola	0.94	0.03	0.03	NA	++	++	NA
AR0083-C	Angola	0.95	0.03	0.02	NA	++	++	NA
AR0084-C	Angola	0.68	0.29	0.03	NA	b+	++	NA
AR0086-C	Angola	0.95	0.03	0.03	NA	++	++	NA
AR0087-C	Angola	0.95	0.03	0.02	NA	++	++	NA
AR0089-C	Angola	0.94	0.04	0.02	NA	++	++	NA
AR0090-C	Angola	0.93	0.03	0.03	NA	++	++	NA
AR0092-C	Angola	0.95	0.03	0.03	NA	++	++	NA
AR0093-C	Angola	0.88	0.07	0.05	NA	++	++	NA
AR0095-C	Angola	0.95	0.03	0.02	NA	++	++	NA
AR0096-C	Angola	0.93	0.04	0.03	NA	++	++	NA
AR0098-C	Angola	0.93	0.04	0.03	NA	++	++	NA
AR0099-C	Angola	0.93	0.04	0.02	NA	++	++	NA
AR0100-C	Angola	0.91	0.06	0.03	NA	++	++	NA
AS0001-C	Gabon	0.95	0.03	0.02	NA	++	++	NA
AS0002-C	Gabon	0.96	0.02	0.02	NA	++	++	NA
AS0003-C	Gabon	0.95	0.03	0.03	NA	++	++	NA
AS0004-C	Gabon	0.92	0.05	0.02	NA	++	++	NA
AS0006-C	Gabon	0.89	0.06	0.05	NA	++	++	NA
AS0007-C	Gabon	0.96	0.02	0.02	NA	++	++	NA
AS0008-C	Gabon	0.95	0.03	0.02	NA	++	++	NA
AS0009-C	Gabon	0.78	0.12	0.09	NA	++	++	NA
AS0010-C	Gabon	0.92	0.06	0.02	NA	++	++	NA
AS0011-C	Gabon	0.93	0.04	0.03	NA	++	++	NA
AS0012-C	Gabon	0.95	0.03	0.02	NA	++	++	NA
AS0013-C	Gabon	0.94	0.04	0.02	NA	++	++	NA
AS0014-C	Gabon	0.84	0.13	0.04	NA	++	++	NA
AS0015-C	Gabon	0.87	0.09	0.03	NA	++	++	NA
AS0016-C	Gabon	0.85	0.10	0.05	NA	b+	++	NA
AS0017-C	Gabon	0.94	0.04	0.02	NA	b+	++	NA
AS0018-C	Gabon	0.92	0.06	0.02	NA	++	++	NA
AS0019-C	Gabon	0.82	0.14	0.04	NA	++	++	NA
AS0020-C	Gabon	0.20	0.78	0.03	NA	++	++	NA
AS0021-C	Gabon	0.95	0.03	0.02	NA	++	++	NA
AS0022-C	Gabon	0.94	0.04	0.02	NA	++	++	NA
AS0024-C	Gabon	0.96	0.02	0.02	NA	++	++	NA

AS0026-C	Gabon	0.95	0.03	0.02	NA	++	++	NA
AS0028-C	Gabon	0.10	0.88	0.02	NA	++	++	NA
AS0030-C	Gabon	0.94	0.03	0.02	NA	++	++	NA
AS0032-C	Gabon	0.95	0.03	0.02	NA	++	++	NA
AS0033-C	Gabon	0.92	0.06	0.02	NA	++	++	NA
AS0034-C	Gabon	0.83	0.14	0.03	NA	++	++	NA
AS0035-C	Gabon	0.92	0.05	0.04	NA	++	++	NA
AS0036-C	Gabon	0.94	0.03	0.03	NA	++	++	NA
AS0037-C	Gabon	0.77	0.20	0.03	NA	++	++	NA
AS0039-C	Gabon	0.95	0.03	0.02	NA	++	++	NA
AS0042-C	Gabon	0.89	0.08	0.03	NA	++	++	NA
AS0044-C	Gabon	0.89	0.07	0.04	NA	++	++	NA
AS0045-C	Gabon	0.63	0.34	0.03	NA	++	++	NA
AS0047-C	Gabon	0.80	0.12	0.08	NA	++	++	NA
AS0049-C	Gabon	0.82	0.14	0.04	NA	b+	++	NA
AS0052-C	Gabon	0.96	0.02	0.02	NA	b+	++	NA
AS0053-C	Gabon	0.90	0.06	0.05	NA	b+	++	NA
AS0054-C	Gabon	0.06	0.93	0.01	NA	++	++	NA
AS0055-C	Gabon	0.76	0.22	0.03	NA	++	++	NA
AS0056-C	Gabon	0.88	0.09	0.04	NA	++	++	NA
AS0058-C	Gabon	0.95	0.02	0.03	NA	++	++	NA
AS0059-C	Gabon	0.93	0.05	0.02	NA	++	++	NA
AS0064-C	Gabon	0.74	0.23	0.04	NA	++	++	NA
AS0065-C	Gabon	0.82	0.14	0.04	NA	++	++	NA
AS0066-C	Gabon	0.92	0.05	0.03	NA	++	++	NA
AS0068-C	Gabon	0.93	0.05	0.02	NA	++	++	NA
AS0069-C	Gabon	0.94	0.03	0.03	NA	++	++	NA
AS0070-C	Gabon	0.95	0.03	0.02	NA	++	++	NA
AS0071-C	Gabon	0.05	0.74	0.22	NA	++	++	NA
AS0072-C	Gabon	0.95	0.03	0.02	NA	++	++	NA
AS0073-C	Gabon	0.90	0.07	0.03	NA	++	++	NA
AS0074-C	Gabon	0.92	0.05	0.03	NA	++	++	NA
AS0076-C	Gabon	0.91	0.06	0.03	NA	++	++	NA
AS0077-C	Gabon	0.23	0.71	0.06	NA	++	++	NA
AV0001-C	Guinea	0.04	0.05	0.91	NA	bb	b+	b+
AV0002-C	Guinea	0.05	0.06	0.89	NA	++	++	++
AV0003-C	Guinea	0.13	0.85	0.02	NA	++	++	++
AV0004-C	Guinea	0.91	0.05	0.04	NA	bb	b+	b+
AV0005-C	Guinea	0.54	0.42	0.04	NA	++	++	++
AV0007-C	Guinea	0.84	0.09	0.07	NA	++	++	++
AV0008-C	Guinea	0.11	0.87	0.02	NA	bb	b+	b+
AV0009-C	Guinea	0.54	0.43	0.03	NA	++	b+	b+
AV0010-C	Guinea	0.09	0.89	0.02	NA	b+	b+	b+
AV0011-C	Guinea	0.05	0.12	0.82	NA	b+	b+	b+
AV0012-C	Guinea	0.04	0.94	0.02	NA	b+	b+	b+
AV0013-C	Guinea	0.46	0.50	0.04	NA	b+	b+	b+
AV0014-C	Guinea	0.89	0.06	0.05	NA	++	++	++
AV0015-C	Guinea	0.04	0.94	0.03	NA	b+	++	++
AV0018-C	Guinea	0.15	0.83	0.02	NA	bb	bb	bb
AV0024-C	Guinea	0.91	0.05	0.04	NA	b+	++	++
AV0026-C	Guinea	0.94	0.04	0.03	NA	++	b+	b+
AV0027-C	Guinea	0.96	0.02	0.02	NA	bb	b+	b+
AV0029-C	Guinea	0.27	0.70	0.03	NA	b+	b+	b+
AV0030-C	Guinea	0.49	0.48	0.03	NA	b+	b+	b+
AV0031-C	Guinea	0.24	0.72	0.04	NA	bb	bb	bb
AV0032-C	Guinea	0.95	0.02	0.02	NA	b+	b+	b+
AV0033-C	Guinea	0.92	0.04	0.03	NA	b+	b+	b+
AV0034-C	Guinea	0.03	0.08	0.89	NA	bb	bb	bb
AV0035-C	Guinea	0.06	0.11	0.83	NA	bb	bb	bb
AV0036-C	Guinea	0.04	0.15	0.80	NA	b+	bb	bb
AV0039-C	Guinea	0.93	0.05	0.02	NA	b+	b+	b+
AV0041-C	Guinea	0.92	0.05	0.03	NA	b+	b+	b+
AV0044-C	Guinea	0.94	0.04	0.03	NA	++	++	++
AV0045-C	Guinea	0.66	0.30	0.04	NA	b+	b+	b+

AV0047-C	Guinea	0.93	0.04	0.02	NA	++	++	++

A5 : 2La karyotype calls, Ag1kG-AR1 set

ind	country	P.std.a	P.het.a	P.inv.a	c2La	s2La	2La	p2La
AB0085-C	Burkina Faso	0.041	0.061	0.898	NA	aa	aa	NA
AB0087-C	Burkina Faso	0.045	0.143	0.813	NA	aa	aa	NA
AB0088-C	Burkina Faso	0.035	0.066	0.899	NA	aa	aa	NA
AB0089-C	Burkina Faso	0.044	0.081	0.876	NA	aa	aa	NA
AB0090-C	Burkina Faso	0.065	0.078	0.857	NA	aa	aa	NA
AB0091-C	Burkina Faso	0.038	0.062	0.901	NA	aa	aa	NA
AB0092-C	Burkina Faso	0.038	0.078	0.884	NA	aa	aa	NA
AB0094-C	Burkina Faso	0.039	0.053	0.908	NA	aa	aa	NA
AB0095-C	Burkina Faso	0.040	0.128	0.832	NA	aa	aa	NA
AB0097-C	Burkina Faso	0.038	0.074	0.888	NA	aa	aa	NA
AB0098-C	Burkina Faso	0.043	0.064	0.893	NA	aa	aa	NA
AB0099-C	Burkina Faso	0.041	0.066	0.893	NA	aa	aa	NA
AB0100-C	Burkina Faso	0.039	0.057	0.905	NA	aa	aa	NA
AB0101-C	Burkina Faso	0.043	0.069	0.888	NA	aa	aa	NA
AB0103-C	Burkina Faso	0.061	0.161	0.778	NA	aa	aa	NA
AB0104-C	Burkina Faso	0.040	0.052	0.908	NA	aa	aa	NA
AB0109-C	Burkina Faso	0.040	0.075	0.885	NA	aa	aa	NA
AB0110-C	Burkina Faso	0.051	0.188	0.762	NA	aa	aa	NA
AB0111-C	Burkina Faso	0.068	0.137	0.795	NA	aa	aa	NA
AB0112-C	Burkina Faso	0.084	0.272	0.644	NA	aa	aa	NA
AB0113-C	Burkina Faso	0.042	0.077	0.881	NA	aa	aa	NA
AB0114-C	Burkina Faso	0.036	0.081	0.883	NA	aa	aa	NA
AB0117-C	Burkina Faso	0.051	0.934	0.015	NA	a+	a+	NA
AB0119-C	Burkina Faso	0.039	0.056	0.905	NA	aa	aa	NA
AB0122-C	Burkina Faso	0.042	0.109	0.849	NA	aa	aa	NA
AB0123-C	Burkina Faso	0.054	0.259	0.687	NA	aa	aa	NA
AB0124-C	Burkina Faso	0.041	0.049	0.910	NA	aa	aa	NA
AB0126-C	Burkina Faso	0.142	0.839	0.018	NA	a+	a+	NA
AB0127-C	Burkina Faso	0.038	0.109	0.853	NA	aa	aa	NA
AB0128-C	Burkina Faso	0.038	0.086	0.876	NA	aa	aa	NA
AB0129-C	Burkina Faso	0.040	0.059	0.901	NA	aa	aa	NA
AB0130-C	Burkina Faso	0.067	0.173	0.760	NA	aa	aa	NA
AB0133-C	Burkina Faso	0.037	0.064	0.899	NA	aa	aa	NA
AB0134-C	Burkina Faso	0.053	0.225	0.722	NA	aa	aa	NA
AB0135-C	Burkina Faso	0.039	0.043	0.919	NA	aa	aa	NA
AB0136-C	Burkina Faso	0.040	0.071	0.889	NA	aa	aa	NA
AB0137-C	Burkina Faso	0.040	0.089	0.871	NA	aa	aa	NA
AB0138-C	Burkina Faso	0.061	0.322	0.616	NA	aa	aa	NA
AB0139-C	Burkina Faso	0.050	0.055	0.895	NA	aa	aa	NA
AB0140-C	Burkina Faso	0.049	0.131	0.820	NA	aa	aa	NA
AB0142-C	Burkina Faso	0.042	0.061	0.898	NA	aa	aa	NA
AB0143-C	Burkina Faso	0.044	0.063	0.893	NA	aa	aa	NA
AB0145-C	Burkina Faso	0.028	0.939	0.033	NA	a+	a+	NA
AB0146-C	Burkina Faso	0.044	0.241	0.715	NA	aa	aa	NA
AB0147-C	Burkina Faso	0.038	0.063	0.899	NA	aa	aa	NA
AB0148-C	Burkina Faso	0.044	0.942	0.015	NA	a+	a+	NA
AB0151-C	Burkina Faso	0.082	0.894	0.025	NA	a+	a+	NA
AB0153-C	Burkina Faso	0.294	0.681	0.025	NA	a+	a+	NA
AB0155-C	Burkina Faso	0.043	0.058	0.900	NA	aa	aa	NA
AB0157-C	Burkina Faso	0.034	0.951	0.016	NA	a+	a+	NA
AB0158-C	Burkina Faso	0.050	0.117	0.834	NA	aa	aa	NA
AB0159-C	Burkina Faso	0.041	0.052	0.907	NA	aa	aa	NA
AB0160-C	Burkina Faso	0.049	0.073	0.878	NA	aa	aa	NA
AB0161-C	Burkina Faso	0.041	0.094	0.865	NA	aa	aa	NA
AB0164-C	Burkina Faso	0.051	0.168	0.781	NA	aa	aa	NA
AB0166-C	Burkina Faso	0.047	0.115	0.838	NA	aa	aa	NA
AB0169-C	Burkina Faso	0.052	0.240	0.708	NA	aa	aa	NA
AB0170-C	Burkina Faso	0.344	0.631	0.026	NA	a+	a+	NA
AB0171-C	Burkina Faso	0.052	0.130	0.818	NA	aa	aa	NA
AB0172-C	Burkina Faso	0.039	0.044	0.917	NA	aa	aa	NA
AB0173-C	Burkina Faso	0.038	0.127	0.835	NA	aa	aa	NA
AB0174-C	Burkina Faso	0.040	0.060	0.900	NA	aa	aa	NA
AB0175-C	Burkina Faso	0.035	0.065	0.900	NA	aa	aa	NA

AB0176-C	Burkina Faso	0.057	0.179	0.764	NA	aa	aa	NA
AB0177-C	Burkina Faso	0.042	0.065	0.893	NA	aa	aa	NA
AB0178-C	Burkina Faso	0.039	0.061	0.900	NA	aa	aa	NA
AB0179-C	Burkina Faso	0.043	0.072	0.885	NA	aa	aa	NA
AB0181-C	Burkina Faso	0.046	0.104	0.849	NA	aa	aa	NA
AB0182-C	Burkina Faso	0.094	0.249	0.657	NA	aa	aa	NA
AB0183-C	Burkina Faso	0.048	0.064	0.889	NA	aa	aa	NA
AB0184-C	Burkina Faso	0.045	0.064	0.890	NA	aa	aa	NA
AB0185-C	Burkina Faso	0.039	0.049	0.912	NA	aa	aa	NA
AB0186-C	Burkina Faso	0.041	0.064	0.895	NA	aa	aa	NA
AB0187-C	Burkina Faso	0.084	0.428	0.488	NA	aa	aa	NA
AB0188-C	Burkina Faso	0.040	0.063	0.897	NA	aa	aa	NA
AB0189-C	Burkina Faso	0.043	0.040	0.916	NA	aa	aa	NA
AB0190-C	Burkina Faso	0.041	0.055	0.903	NA	aa	aa	NA
AB0191-C	Burkina Faso	0.039	0.053	0.908	NA	aa	aa	NA
AB0192-C	Burkina Faso	0.039	0.061	0.901	NA	aa	aa	NA
AB0197-C	Burkina Faso	0.042	0.104	0.855	NA	aa	aa	NA
AB0198-C	Burkina Faso	0.054	0.058	0.888	NA	aa	aa	NA
AB0199-C	Burkina Faso	0.038	0.054	0.907	NA	aa	aa	NA
AB0201-C	Burkina Faso	0.043	0.040	0.917	NA	aa	aa	NA
AB0202-C	Burkina Faso	0.045	0.052	0.903	NA	aa	aa	NA
AB0203-C	Burkina Faso	0.064	0.120	0.816	NA	aa	aa	NA
AB0204-C	Burkina Faso	0.053	0.088	0.859	NA	aa	aa	NA
AB0205-C	Burkina Faso	0.249	0.724	0.027	NA	a+	a+	NA
AB0206-C	Burkina Faso	0.038	0.045	0.917	NA	aa	aa	NA
AB0207-C	Burkina Faso	0.047	0.042	0.912	NA	aa	aa	NA
AB0208-C	Burkina Faso	0.052	0.071	0.877	NA	aa	aa	NA
AB0209-C	Burkina Faso	0.044	0.194	0.762	NA	aa	aa	NA
AB0210-C	Burkina Faso	0.045	0.058	0.897	NA	aa	aa	NA
AB0211-C	Burkina Faso	0.059	0.069	0.872	NA	aa	aa	NA
AB0212-C	Burkina Faso	0.035	0.062	0.903	NA	aa	aa	NA
AB0213-C	Burkina Faso	0.039	0.097	0.864	NA	aa	aa	NA
AB0217-C	Burkina Faso	0.040	0.064	0.896	NA	aa	aa	NA
AB0219-C	Burkina Faso	0.043	0.102	0.856	NA	aa	aa	NA
AB0221-C	Burkina Faso	0.038	0.076	0.886	NA	aa	aa	NA
AB0222-C	Burkina Faso	0.076	0.217	0.707	NA	aa	aa	NA
AB0223-C	Burkina Faso	0.045	0.055	0.900	NA	aa	aa	NA
AB0224-C	Burkina Faso	0.126	0.856	0.017	NA	a+	a+	NA
AB0226-C	Burkina Faso	0.061	0.169	0.770	NA	aa	aa	NA
AB0227-C	Burkina Faso	0.049	0.130	0.821	NA	aa	aa	NA
AB0228-C	Burkina Faso	0.040	0.066	0.894	NA	aa	aa	NA
AB0229-C	Burkina Faso	0.034	0.065	0.900	NA	aa	aa	NA
AB0231-C	Burkina Faso	0.754	0.210	0.036	NA	++	a+	NA
AB0233-C	Burkina Faso	0.068	0.912	0.020	NA	a+	a+	NA
AB0234-C	Burkina Faso	0.067	0.428	0.506	NA	aa	aa	NA
AB0235-C	Burkina Faso	0.067	0.173	0.760	NA	aa	aa	NA
AB0236-C	Burkina Faso	0.047	0.116	0.836	NA	aa	aa	NA
AB0237-C	Burkina Faso	0.058	0.130	0.812	NA	aa	aa	NA
AB0238-C	Burkina Faso	0.047	0.109	0.844	NA	aa	aa	NA
AB0239-C	Burkina Faso	0.063	0.106	0.831	NA	aa	aa	NA
AB0240-C	Burkina Faso	0.043	0.065	0.892	NA	aa	aa	NA
AB0241-C	Burkina Faso	0.035	0.059	0.905	NA	aa	aa	NA
AB0242-C	Burkina Faso	0.037	0.044	0.920	NA	aa	aa	NA
AB0243-C	Burkina Faso	0.047	0.084	0.869	NA	aa	aa	NA
AB0244-C	Burkina Faso	0.041	0.059	0.901	NA	aa	aa	NA
AB0246-C	Burkina Faso	0.040	0.123	0.837	NA	aa	aa	NA
AB0249-C	Burkina Faso	0.043	0.062	0.894	NA	aa	aa	NA
AB0250-C	Burkina Faso	0.087	0.407	0.506	NA	aa	aa	NA
AB0251-C	Burkina Faso	0.100	0.883	0.017	NA	a+	a+	NA
AB0252-C	Burkina Faso	0.044	0.079	0.877	NA	aa	aa	NA
AB0253-C	Burkina Faso	0.066	0.165	0.769	NA	aa	aa	NA
AB0256-C	Burkina Faso	0.072	0.905	0.022	NA	a+	a+	NA
AB0257-C	Burkina Faso	0.043	0.079	0.878	NA	aa	aa	NA
AB0258-C	Burkina Faso	0.053	0.191	0.757	NA	aa	aa	NA
AB0260-C	Burkina Faso	0.041	0.068	0.890	NA	aa	aa	NA
AB0261-C	Burkina Faso	0.360	0.573	0.067	NA	a+	a+	NA
AB0262-C	Burkina Faso	0.042	0.051	0.907	NA	aa	aa	NA
AB0263-C	Burkina Faso	0.051	0.059	0.889	NA	aa	aa	NA

AB0264-C	Burkina Faso	0.053	0.107	0.840	NA	aa	aa	NA
AB0265-C	Burkina Faso	0.090	0.157	0.754	NA	aa	aa	NA
AB0266-C	Burkina Faso	0.100	0.157	0.743	NA	aa	aa	NA
AB0267-C	Burkina Faso	0.038	0.056	0.905	NA	aa	aa	NA
AB0268-C	Burkina Faso	0.060	0.097	0.842	NA	aa	aa	NA
AB0270-C	Burkina Faso	0.032	0.951	0.017	NA	a+	a+	NA
AB0271-C	Burkina Faso	0.052	0.092	0.857	NA	aa	aa	NA
AB0272-C	Burkina Faso	0.046	0.040	0.914	NA	aa	aa	NA
AB0273-C	Burkina Faso	0.038	0.083	0.880	NA	aa	aa	NA
AB0274-C	Burkina Faso	0.062	0.090	0.848	NA	aa	aa	NA
AB0276-C	Burkina Faso	0.098	0.128	0.774	NA	aa	aa	NA
AB0277-C	Burkina Faso	0.046	0.086	0.868	NA	aa	aa	NA
AB0278-C	Burkina Faso	0.036	0.074	0.890	NA	aa	aa	NA
AB0279-C	Burkina Faso	0.048	0.079	0.873	NA	aa	aa	NA
AB0280-C	Burkina Faso	0.035	0.080	0.885	NA	aa	aa	NA
AB0281-C	Burkina Faso	0.111	0.156	0.733	NA	aa	aa	NA
AB0282-C	Burkina Faso	0.044	0.095	0.861	NA	aa	aa	NA
AB0283-C	Burkina Faso	0.038	0.042	0.921	NA	aa	aa	NA
AB0284-C	Burkina Faso	0.041	0.048	0.911	NA	aa	aa	NA
AC0090-C	Uganda	0.063	0.820	0.117	NA	a+	a+	a+
AC0091-C	Uganda	0.058	0.821	0.121	NA	a+	a+	a+
AC0092-C	Uganda	0.038	0.107	0.854	NA	aa	aa	aa
AC0093-C	Uganda	0.146	0.829	0.025	NA	a+	a+	a+
AC0094-C	Uganda	0.141	0.785	0.073	NA	a+	a+	a+
AC0095-C	Uganda	0.075	0.177	0.748	NA	aa	aa	aa
AC0096-C	Uganda	0.052	0.904	0.045	NA	a+	a+	a+
AC0097-C	Uganda	0.057	0.096	0.847	NA	aa	aa	aa
AC0098-C	Uganda	0.070	0.279	0.651	NA	aa	aa	aa
AC0099-C	Uganda	0.810	0.149	0.040	NA	++	++	++
AC0100-C	Uganda	0.064	0.720	0.216	NA	a+	a+	a+
AC0101-C	Uganda	0.283	0.689	0.028	NA	a+	a+	a+
AC0102-C	Uganda	0.955	0.023	0.021	NA	++	++	++
AC0103-C	Uganda	0.042	0.076	0.882	NA	aa	aa	aa
AC0104-C	Uganda	0.171	0.802	0.027	NA	a+	++	++
AC0106-C	Uganda	0.037	0.081	0.882	NA	aa	aa	aa
AC0107-C	Uganda	0.939	0.033	0.028	NA	++	++	++
AC0108-C	Uganda	0.042	0.052	0.907	NA	aa	aa	aa
AC0109-C	Uganda	0.045	0.062	0.893	NA	aa	aa	aa
AC0110-C	Uganda	0.074	0.679	0.247	NA	a+	a+	a+
AC0111-C	Uganda	0.058	0.379	0.564	NA	aa	aa	aa
AC0112-C	Uganda	0.047	0.748	0.205	NA	a+	a+	a+
AC0113-C	Uganda	0.099	0.773	0.128	NA	a+	a+	a+
AC0114-C	Uganda	0.061	0.921	0.017	NA	a+	a+	a+
AC0115-C	Uganda	0.051	0.107	0.842	NA	aa	aa	aa
AC0116-C	Uganda	0.040	0.060	0.899	NA	aa	aa	aa
AC0117-C	Uganda	0.211	0.704	0.085	NA	a+	a+	a+
AC0118-C	Uganda	0.191	0.783	0.025	NA	a+	a+	a+
AC0119-C	Uganda	0.409	0.517	0.074	NA	a+	++	++
AC0120-C	Uganda	0.948	0.028	0.024	NA	++	++	++
AC0121-C	Uganda	0.039	0.044	0.918	NA	aa	aa	aa
AC0122-C	Uganda	0.072	0.823	0.105	NA	a+	a+	a+
AC0123-C	Uganda	0.084	0.809	0.107	NA	a+	a+	a+
AC0124-C	Uganda	0.046	0.072	0.881	NA	aa	aa	aa
AC0125-C	Uganda	0.044	0.791	0.165	NA	a+	a+	a+
AC0126-C	Uganda	0.831	0.123	0.046	NA	++	++	++
AC0127-C	Uganda	0.041	0.069	0.890	NA	aa	aa	aa
AC0128-C	Uganda	0.028	0.946	0.026	NA	a+	a+	a+
AC0129-C	Uganda	0.037	0.060	0.903	NA	aa	aa	aa
AC0130-C	Uganda	0.045	0.101	0.854	NA	aa	aa	aa
AC0131-C	Uganda	0.054	0.848	0.099	NA	a+	a+	a+
AC0132-C	Uganda	0.046	0.052	0.903	NA	aa	aa	aa
AC0133-C	Uganda	0.207	0.709	0.085	NA	a+	a+	a+
AC0135-C	Uganda	0.086	0.900	0.014	NA	a+	a+	a+
AC0136-C	Uganda	0.092	0.826	0.082	NA	a+	a+	a+
AC0137-C	Uganda	0.037	0.052	0.911	NA	aa	aa	aa
AC0138-C	Uganda	0.457	0.452	0.090	NA	a+	++	++
AC0139-C	Uganda	0.903	0.066	0.031	NA	++	++	++
AC0140-C	Uganda	0.036	0.065	0.898	NA	aa	aa	aa

AC0142-C	Uganda	0.065	0.807	0.128	NA	a+	a+	a+
AC0143-C	Uganda	0.043	0.078	0.879	NA	aa	aa	aa
AC0144-C	Uganda	0.074	0.196	0.730	NA	aa	aa	aa
AC0145-C	Uganda	0.055	0.930	0.015	NA	a+	a+	a+
AC0147-C	Uganda	0.036	0.065	0.899	NA	aa	aa	aa
AC0148-C	Uganda	0.048	0.126	0.826	NA	aa	aa	aa
AC0149-C	Uganda	0.055	0.768	0.177	NA	a+	a+	a+
AC0150-C	Uganda	0.590	0.355	0.055	NA	++	++	++
AC0151-C	Uganda	0.030	0.957	0.013	NA	a+	a+	a+
AC0152-C	Uganda	0.056	0.927	0.016	NA	a+	a+	a+
AC0153-C	Uganda	0.695	0.253	0.052	NA	++	++	++
AC0154-C	Uganda	0.073	0.670	0.257	NA	a+	a+	a+
AC0156-C	Uganda	0.063	0.872	0.064	NA	a+	a+	a+
AC0158-C	Uganda	0.061	0.920	0.019	NA	a+	a+	a+
AC0159-C	Uganda	0.953	0.028	0.019	NA	++	++	++
AC0160-C	Uganda	0.059	0.872	0.068	NA	a+	a+	a+
AC0161-C	Uganda	0.061	0.354	0.585	NA	aa	aa	aa
AC0162-C	Uganda	0.097	0.807	0.096	NA	a+	a+	a+
AC0163-C	Uganda	0.069	0.839	0.092	NA	a+	a+	a+
AC0164-C	Uganda	0.027	0.960	0.013	NA	a+	a+	a+
AC0166-C	Uganda	0.050	0.461	0.490	NA	aa	a+	a+
AC0167-C	Uganda	0.065	0.775	0.159	NA	a+	a+	a+
AC0168-C	Uganda	0.055	0.885	0.061	NA	a+	a+	a+
AC0169-C	Uganda	0.062	0.134	0.804	NA	aa	aa	aa
AC0170-C	Uganda	0.076	0.812	0.111	NA	a+	a+	a+
AC0171-C	Uganda	0.063	0.070	0.868	NA	aa	aa	aa
AC0172-C	Uganda	0.042	0.108	0.850	NA	aa	aa	aa
AC0173-C	Uganda	0.079	0.775	0.147	NA	a+	a+	a+
AC0174-C	Uganda	0.099	0.845	0.056	NA	a+	a+	a+
AC0176-C	Uganda	0.618	0.338	0.044	NA	++	++	++
AC0178-C	Uganda	0.264	0.619	0.117	NA	a+	a+	a+
AC0179-C	Uganda	0.051	0.046	0.903	NA	aa	aa	aa
AC0180-C	Uganda	0.043	0.078	0.879	NA	aa	aa	aa
AC0181-C	Uganda	0.622	0.316	0.063	NA	++	++	++
AC0182-C	Uganda	0.037	0.051	0.912	NA	aa	aa	aa
AC0183-C	Uganda	0.175	0.789	0.036	NA	a+	a+	a+
AC0184-C	Uganda	0.050	0.096	0.854	NA	aa	aa	aa
AC0186-C	Uganda	0.851	0.110	0.039	NA	++	++	++
AC0187-C	Uganda	0.067	0.684	0.249	NA	a+	a+	a+
AC0188-C	Uganda	0.081	0.807	0.112	NA	a+	a+	a+
AC0189-C	Uganda	0.040	0.051	0.909	NA	aa	aa	aa
AC0190-C	Uganda	0.039	0.054	0.908	NA	aa	aa	aa
AC0191-C	Uganda	0.036	0.065	0.900	NA	aa	aa	aa
AC0192-C	Uganda	0.047	0.159	0.794	NA	aa	aa	aa
AC0193-C	Uganda	0.037	0.112	0.851	NA	aa	aa	aa
AC0194-C	Uganda	0.041	0.050	0.910	NA	aa	aa	aa
AC0195-C	Uganda	0.074	0.807	0.118	NA	a+	a+	a+
AC0196-C	Uganda	0.043	0.062	0.895	NA	aa	aa	aa
AC0197-C	Uganda	0.035	0.074	0.892	NA	aa	aa	aa
AC0199-C	Uganda	0.064	0.861	0.075	NA	a+	a+	a+
AC0200-C	Uganda	0.067	0.705	0.228	NA	a+	a+	a+
AC0201-C	Uganda	0.035	0.056	0.908	NA	aa	aa	aa
AC0202-C	Uganda	0.126	0.857	0.017	NA	a+	a+	a+
AC0203-C	Uganda	0.062	0.663	0.275	NA	a+	a+	a+
AJ0023-C	Guinea-Bissau	0.860	0.114	0.026	NA	++	++	++
AJ0024-C	Guinea-Bissau	0.951	0.029	0.020	NA	++	++	++
AJ0032-C	Guinea-Bissau	0.041	0.125	0.834	NA	aa	aa	aa
AJ0035-C	Guinea-Bissau	0.043	0.936	0.021	NA	a+	a+	a+
AJ0036-C	Guinea-Bissau	0.805	0.164	0.030	NA	++	++	++
AJ0039-C	Guinea-Bissau	0.930	0.048	0.022	NA	++	++	++
AJ0043-C	Guinea-Bissau	0.067	0.896	0.037	NA	a+	a+	a+
AJ0044-C	Guinea-Bissau	0.180	0.790	0.030	NA	a+	a+	a+
AJ0045-C	Guinea-Bissau	0.043	0.066	0.891	NA	aa	aa	aa
AJ0047-C	Guinea-Bissau	0.065	0.907	0.029	NA	a+	a+	a+
AJ0051-C	Guinea-Bissau	0.036	0.911	0.054	NA	a+	a+	a+
AJ0052-C	Guinea-Bissau	0.907	0.061	0.032	NA	++	++	++
AJ0056-C	Guinea-Bissau	0.919	0.051	0.030	NA	++	++	++
AJ0061-C	Guinea-Bissau	0.312	0.653	0.036	NA	a+	++	++

AJ0063-C	Guinea-Bissau	0.035	0.930	0.035	NA	a+	a+	a+
AJ0064-C	Guinea-Bissau	0.263	0.706	0.030	NA	a+	a+	a+
AJ0066-C	Guinea-Bissau	0.675	0.284	0.040	NA	++	++	++
AJ0070-C	Guinea-Bissau	0.927	0.040	0.033	NA	++	++	++
AJ0071-C	Guinea-Bissau	0.934	0.044	0.022	NA	++	++	++
AJ0072-C	Guinea-Bissau	0.872	0.092	0.036	NA	++	++	++
AJ0074-C	Guinea-Bissau	0.036	0.071	0.893	NA	aa	aa	aa
AJ0075-C	Guinea-Bissau	0.051	0.931	0.018	NA	a+	a+	a+
AJ0076-C	Guinea-Bissau	0.058	0.599	0.343	NA	a+	a+	a+
AJ0077-C	Guinea-Bissau	0.073	0.865	0.062	NA	a+	a+	a+
AJ0078-C	Guinea-Bissau	0.040	0.942	0.018	NA	a+	a+	a+
AJ0081-C	Guinea-Bissau	0.078	0.903	0.019	NA	a+	a+	a+
AJ0084-C	Guinea-Bissau	0.036	0.948	0.015	NA	a+	a+	a+
AJ0085-C	Guinea-Bissau	0.812	0.163	0.025	NA	++	++	++
AJ0086-C	Guinea-Bissau	0.891	0.057	0.053	NA	++	++	++
AJ0088-C	Guinea-Bissau	0.708	0.257	0.034	NA	++	++	++
AJ0090-C	Guinea-Bissau	0.118	0.835	0.048	NA	a+	a+	a+
AJ0092-C	Guinea-Bissau	0.262	0.717	0.022	NA	a+	a+	a+
AJ0093-C	Guinea-Bissau	0.052	0.925	0.023	NA	a+	a+	a+
AJ0096-C	Guinea-Bissau	0.912	0.059	0.029	NA	++	++	++
AJ0097-C	Guinea-Bissau	0.037	0.949	0.014	NA	a+	a+	a+
AJ0098-C	Guinea-Bissau	0.947	0.030	0.023	NA	++	++	++
AJ0100-C	Guinea-Bissau	0.946	0.033	0.021	NA	++	++	++
AJ0101-C	Guinea-Bissau	0.935	0.035	0.030	NA	++	++	++
AJ0102-C	Guinea-Bissau	0.944	0.028	0.028	NA	++	++	++
AJ0103-C	Guinea-Bissau	0.064	0.919	0.017	NA	a+	a+	a+
AJ0105-C	Guinea-Bissau	0.874	0.091	0.036	NA	++	++	++
AJ0107-C	Guinea-Bissau	0.625	0.336	0.040	NA	++	++	++
AJ0109-C	Guinea-Bissau	0.692	0.235	0.073	NA	++	++	++
AJ0113-C	Guinea-Bissau	0.064	0.911	0.025	NA	a+	a+	a+
AJ0115-C	Guinea-Bissau	0.536	0.402	0.062	NA	++	++	++
AJ0116-C	Guinea-Bissau	0.056	0.917	0.027	NA	a+	a+	a+
AK0065-C	Kenya	0.045	0.062	0.893	NA	aa	aa	NA
AK0066-C	Kenya	0.038	0.045	0.917	NA	aa	aa	NA
AK0067-C	Kenya	0.957	0.023	0.020	NA	++	++	NA
AK0068-C	Kenya	0.045	0.895	0.060	NA	a+	a+	NA
AK0069-C	Kenya	0.510	0.458	0.032	NA	a+	++	NA
AK0070-C	Kenya	0.037	0.053	0.910	NA	aa	aa	NA
AK0072-C	Kenya	0.044	0.782	0.174	NA	a+	a+	NA
AK0073-C	Kenya	0.132	0.844	0.024	NA	a+	a+	NA
AK0074-C	Kenya	0.280	0.690	0.030	NA	a+	a+	NA
AK0075-C	Kenya	0.042	0.863	0.095	NA	a+	a+	NA
AK0076-C	Kenya	0.814	0.160	0.026	NA	++	a+	NA
AK0077-C	Kenya	0.031	0.951	0.018	NA	a+	a+	NA
AK0078-C	Kenya	0.947	0.029	0.024	NA	++	++	NA
AK0079-C	Kenya	0.379	0.589	0.033	NA	a+	a+	NA
AK0080-C	Kenya	0.109	0.876	0.016	NA	a+	a+	NA
AK0081-C	Kenya	0.957	0.021	0.022	NA	++	++	NA
AK0082-C	Kenya	0.054	0.931	0.014	NA	a+	a+	NA
AK0085-C	Kenya	0.090	0.679	0.230	NA	a+	a+	NA
AK0086-C	Kenya	0.063	0.812	0.125	NA	a+	a+	NA
AK0087-C	Kenya	0.102	0.112	0.786	NA	aa	aa	NA
AK0088-C	Kenya	0.046	0.589	0.365	NA	a+	a+	NA
AK0089-C	Kenya	0.043	0.085	0.871	NA	aa	aa	NA
AK0090-C	Kenya	0.056	0.922	0.022	NA	a+	a+	NA
AK0091-C	Kenya	0.118	0.854	0.028	NA	a+	a+	NA
AK0092-C	Kenya	0.105	0.206	0.688	NA	aa	a+	NA
AK0093-C	Kenya	0.097	0.111	0.792	NA	aa	aa	NA
AK0094-C	Kenya	0.067	0.119	0.815	NA	aa	aa	NA
AK0095-C	Kenya	0.042	0.928	0.031	NA	a+	a+	NA
AK0096-C	Kenya	0.055	0.065	0.881	NA	aa	aa	NA
AK0098-C	Kenya	0.046	0.098	0.856	NA	aa	aa	NA
AK0099-C	Kenya	0.060	0.070	0.870	NA	aa	aa	NA
AK0100-C	Kenya	0.035	0.954	0.012	NA	a+	a+	NA
AK0101-C	Kenya	0.048	0.067	0.886	NA	aa	aa	NA
AK0102-C	Kenya	0.056	0.100	0.844	NA	aa	a+	NA
AK0103-C	Kenya	0.946	0.030	0.024	NA	++	++	NA
AK0104-C	Kenya	0.026	0.960	0.014	NA	a+	a+	NA

AK0105-C	Kenya	0.489	0.480	0.031	NA	a+	++	NA
AK0106-C	Kenya	0.959	0.023	0.018	NA	++	++	NA
AK0108-C	Kenya	0.046	0.936	0.018	NA	a+	a+	NA
AK0109-C	Kenya	0.049	0.055	0.897	NA	aa	aa	NA
AK0110-C	Kenya	0.045	0.063	0.892	NA	aa	aa	NA
AK0116-C	Kenya	0.060	0.328	0.612	NA	aa	aa	NA
AK0119-C	Kenya	0.939	0.043	0.019	NA	++	a+	NA
AK0127-C	Kenya	0.077	0.101	0.823	NA	aa	aa	NA
AN0007-C	Cameroon	0.865	0.109	0.025	NA	++	++	++
AN0008-C	Cameroon	0.916	0.050	0.034	NA	++	++	++
AN0009-C	Cameroon	0.041	0.947	0.011	NA	a+	a+	a+
AN0010-C	Cameroon	0.923	0.043	0.035	NA	++	++	++
AN0011-C	Cameroon	0.910	0.050	0.040	NA	++	++	++
AN0012-C	Cameroon	0.926	0.038	0.036	NA	++	++	++
AN0014-C	Cameroon	0.945	0.029	0.026	NA	++	++	++
AN0016-C	Cameroon	0.474	0.488	0.037	NA	a+	a+	a+
AN0017-C	Cameroon	0.046	0.941	0.014	NA	a+	a+	a+
AN0018-C	Cameroon	0.880	0.075	0.045	NA	++	++	++
AN0019-C	Cameroon	0.912	0.046	0.041	NA	++	++	++
AN0020-C	Cameroon	0.729	0.190	0.081	NA	++	++	++
AN0022-C	Cameroon	0.080	0.898	0.023	NA	a+	a+	a+
AN0023-C	Cameroon	0.073	0.539	0.388	NA	a+	aa	aa
AN0024-C	Cameroon	0.063	0.367	0.570	NA	aa	aa	aa
AN0025-C	Cameroon	0.092	0.888	0.020	NA	a+	a+	a+
AN0026-C	Cameroon	0.036	0.046	0.917	NA	aa	aa	aa
AN0027-C	Cameroon	0.025	0.959	0.016	NA	a+	a+	a+
AN0028-C	Cameroon	0.931	0.042	0.026	NA	++	++	++
AN0029-C	Cameroon	0.511	0.446	0.043	NA	a+	a+	a+
AN0030-C	Cameroon	0.056	0.072	0.872	NA	aa	aa	aa
AN0031-C	Cameroon	0.107	0.874	0.018	NA	a+	a+	a+
AN0032-C	Cameroon	0.040	0.935	0.025	NA	a+	a+	a+
AN0033-C	Cameroon	0.073	0.910	0.017	NA	a+	a+	a+
AN0034-C	Cameroon	0.046	0.143	0.811	NA	aa	aa	aa
AN0035-C	Cameroon	0.936	0.035	0.029	NA	++	++	++
AN0036-C	Cameroon	0.060	0.085	0.855	NA	aa	aa	aa
AN0037-C	Cameroon	0.037	0.929	0.033	NA	a+	a+	a+
AN0038-C	Cameroon	0.051	0.933	0.016	NA	a+	a+	a+
AN0039-C	Cameroon	0.039	0.946	0.015	NA	a+	a+	a+
AN0040-C	Cameroon	0.073	0.176	0.752	NA	aa	aa	aa
AN0041-C	Cameroon	0.052	0.929	0.019	NA	a+	a+	a+
AN0042-C	Cameroon	0.068	0.915	0.017	NA	a+	a+	a+
AN0043-C	Cameroon	0.038	0.083	0.879	NA	aa	aa	aa
AN0045-C	Cameroon	0.094	0.878	0.029	NA	a+	a+	a+
AN0046-C	Cameroon	0.048	0.059	0.893	NA	aa	aa	aa
AN0047-C	Cameroon	0.042	0.081	0.877	NA	aa	aa	aa
AN0048-C	Cameroon	0.046	0.074	0.879	NA	aa	aa	aa
AN0049-C	Cameroon	0.041	0.056	0.903	NA	aa	aa	aa
AN0050-C	Cameroon	0.030	0.957	0.013	NA	a+	a+	a+
AN0051-C	Cameroon	0.278	0.697	0.026	NA	a+	a+	a+
AN0053-C	Cameroon	0.083	0.312	0.605	NA	aa	aa	aa
AN0054-C	Cameroon	0.037	0.051	0.912	NA	aa	aa	aa
AN0055-C	Cameroon	0.890	0.060	0.050	NA	++	++	++
AN0056-C	Cameroon	0.092	0.890	0.018	NA	a+	a+	a+
AN0057-C	Cameroon	0.691	0.253	0.056	NA	++	++	++
AN0058-C	Cameroon	0.036	0.044	0.920	NA	aa	aa	aa
AN0059-C	Cameroon	0.101	0.876	0.023	NA	a+	a+	a+
AN0060-C	Cameroon	0.038	0.081	0.881	NA	aa	aa	aa
AN0063-C	Cameroon	0.268	0.708	0.025	a+	a+	a+	a+
AN0064-C	Cameroon	0.046	0.124	0.830	++	aa	aa	aa
AN0065-C	Cameroon	0.084	0.901	0.015	a+	a+	a+	a+
AN0066-C	Cameroon	0.723	0.239	0.038	a+	++	a+	a+
AN0067-C	Cameroon	0.047	0.086	0.867	aa	aa	aa	aa
AN0068-C	Cameroon	0.948	0.028	0.024	++	++	++	++
AN0069-C	Cameroon	0.954	0.024	0.022	++	++	++	++
AN0070-C	Cameroon	0.036	0.070	0.894	aa	aa	aa	aa
AN0071-C	Cameroon	0.045	0.086	0.870	aa	aa	aa	aa
AN0072-C	Cameroon	0.091	0.880	0.029	a+	a+	a+	a+
AN0073-C	Cameroon	0.926	0.046	0.028	++	++	++	++

AN0074-C	Cameroon	0.051	0.935	0.014	a+	a+	a+	a+
AN0075-C	Cameroon	0.945	0.032	0.024	++	++	++	++
AN0076-C	Cameroon	0.060	0.063	0.878	aa	aa	aa	aa
AN0077-C	Cameroon	0.937	0.036	0.027	++	++	++	++
AN0079-C	Cameroon	0.927	0.039	0.034	++	++	++	++
AN0080-C	Cameroon	0.955	0.025	0.020	++	++	++	++
AN0081-C	Cameroon	0.166	0.805	0.029	a+	a+	a+	a+
AN0082-C	Cameroon	0.043	0.055	0.902	aa	aa	aa	aa
AN0083-C	Cameroon	0.080	0.904	0.016	a+	a+	a+	a+
AN0084-C	Cameroon	0.946	0.028	0.026	++	++	++	++
AN0085-C	Cameroon	0.878	0.081	0.042	++	++	++	++
AN0086-C	Cameroon	0.105	0.874	0.021	a+	a+	a+	a+
AN0087-C	Cameroon	0.952	0.024	0.024	++	++	++	++
AN0088-C	Cameroon	0.075	0.904	0.021	a+	a+	a+	a+
AN0089-C	Cameroon	0.047	0.081	0.872	aa	aa	aa	aa
AN0090-C	Cameroon	0.038	0.108	0.854	aa	aa	aa	aa
AN0091-C	Cameroon	0.044	0.058	0.898	aa	aa	aa	aa
AN0092-C	Cameroon	0.046	0.102	0.852	aa	aa	aa	aa
AN0093-C	Cameroon	0.064	0.081	0.855	aa	aa	aa	aa
AN0094-C	Cameroon	0.156	0.824	0.020	a+	a+	a+	a+
AN0095-C	Cameroon	0.037	0.088	0.875	aa	aa	aa	aa
AN0096-C	Cameroon	0.059	0.926	0.015	a+	a+	a+	a+
AN0097-C	Cameroon	0.296	0.679	0.026	a+	a+	a+	a+
AN0098-C	Cameroon	0.038	0.948	0.014	a+	a+	a+	a+
AN0099-C	Cameroon	0.924	0.040	0.037	++	++	++	++
AN0100-C	Cameroon	0.051	0.933	0.016	a+	a+	a+	a+
AN0101-C	Cameroon	0.100	0.876	0.025	a+	a+	a+	a+
AN0102-C	Cameroon	0.056	0.927	0.017	a+	a+	a+	a+
AN0103-C	Cameroon	0.066	0.913	0.021	a+	a+	a+	a+
AN0104-C	Cameroon	0.051	0.718	0.231	a+	a+	a+	a+
AN0105-C	Cameroon	0.049	0.084	0.867	aa	aa	aa	aa
AN0106-C	Cameroon	0.947	0.028	0.025	++	++	++	++
AN0107-C	Cameroon	0.853	0.110	0.037	++	++	++	++
AN0108-C	Cameroon	0.063	0.062	0.876	a+	aa	aa	aa
AN0109-C	Cameroon	0.283	0.695	0.023	a+	a+	a+	a+
AN0111-C	Cameroon	0.142	0.812	0.047	a+	a+	a+	a+
AN0112-C	Cameroon	0.049	0.095	0.856	aa	aa	aa	aa
AN0113-C	Cameroon	0.960	0.021	0.019	++	++	++	++
AN0114-C	Cameroon	0.048	0.059	0.892	aa	aa	aa	aa
AN0115-C	Cameroon	0.956	0.025	0.019	++	++	++	++
AN0117-C	Cameroon	0.750	0.210	0.040	a+	++	a+	a+
AN0120-C	Cameroon	0.038	0.951	0.011	a+	a+	a+	a+
AN0121-C	Cameroon	0.038	0.057	0.905	aa	aa	aa	aa
AN0122-C	Cameroon	0.914	0.064	0.022	NA	++	++	++
AN0123-C	Cameroon	0.954	0.025	0.021	NA	++	++	++
AN0124-C	Cameroon	0.874	0.066	0.060	NA	++	++	++
AN0125-C	Cameroon	0.893	0.059	0.048	NA	++	++	++
AN0126-C	Cameroon	0.952	0.026	0.022	NA	++	++	++
AN0127-C	Cameroon	0.954	0.025	0.021	NA	++	++	++
AN0128-C	Cameroon	0.112	0.861	0.026	NA	a+	a+	a+
AN0129-C	Cameroon	0.931	0.035	0.035	NA	++	++	++
AN0130-C	Cameroon	0.822	0.144	0.034	NA	++	++	++
AN0131-C	Cameroon	0.952	0.024	0.024	NA	++	++	++
AN0132-C	Cameroon	0.938	0.033	0.029	NA	++	++	++
AN0134-C	Cameroon	0.935	0.039	0.027	NA	++	++	++
AN0135-C	Cameroon	0.077	0.175	0.748	NA	aa	aa	aa
AN0136-C	Cameroon	0.893	0.057	0.050	NA	++	++	++
AN0137-C	Cameroon	0.958	0.023	0.019	NA	++	++	++
AN0138-C	Cameroon	0.932	0.034	0.033	NA	++	++	++
AN0139-C	Cameroon	0.918	0.042	0.040	NA	++	++	++
AN0140-C	Cameroon	0.819	0.103	0.078	NA	++	++	++
AN0141-C	Cameroon	0.893	0.057	0.051	NA	++	++	++
AN0143-C	Cameroon	0.062	0.921	0.017	NA	a+	a+	a+
AN0147-C	Cameroon	0.915	0.047	0.038	NA	++	++	++
AN0149-C	Cameroon	0.054	0.934	0.013	NA	a+	a+	a+
AN0151-C	Cameroon	0.905	0.060	0.035	NA	++	++	++
AN0152-C	Cameroon	0.950	0.030	0.020	NA	++	++	++
AN0153-C	Cameroon	0.887	0.064	0.049	NA	++	++	++

AN0154-C	Cameroon	0.959	0.024	0.017	NA	++	++	++
AN0155-C	Cameroon	0.890	0.080	0.030	NA	++	++	++
AN0156-C	Cameroon	0.255	0.718	0.027	NA	a+	a+	a+
AN0157-C	Cameroon	0.942	0.032	0.025	NA	++	++	++
AN0158-C	Cameroon	0.934	0.035	0.032	NA	++	++	++
AN0159-C	Cameroon	0.947	0.031	0.022	NA	++	++	++
AN0160-C	Cameroon	0.950	0.027	0.023	NA	++	++	++
AN0162-C	Cameroon	0.896	0.057	0.048	NA	++	++	++
AN0163-C	Cameroon	0.946	0.029	0.025	NA	++	++	++
AN0164-C	Cameroon	0.915	0.047	0.037	NA	++	++	++
AN0165-C	Cameroon	0.108	0.869	0.023	NA	a+	a+	a+
AN0166-C	Cameroon	0.898	0.075	0.028	NA	++	++	++
AN0167-C	Cameroon	0.947	0.029	0.024	NA	++	++	++
AN0168-C	Cameroon	0.938	0.037	0.025	NA	++	++	++
AN0169-C	Cameroon	0.955	0.023	0.022	NA	++	++	++
AN0170-C	Cameroon	0.939	0.035	0.026	NA	++	++	++
AN0171-C	Cameroon	0.942	0.035	0.023	NA	++	++	++
AN0172-C	Cameroon	0.888	0.086	0.026	NA	++	++	++
AN0173-C	Cameroon	0.955	0.023	0.022	NA	++	++	++
AN0174-C	Cameroon	0.890	0.061	0.049	NA	++	++	++
AN0175-C	Cameroon	0.881	0.068	0.051	NA	++	++	++
AN0176-C	Cameroon	0.135	0.841	0.025	NA	a+	a+	a+
AN0177-C	Cameroon	0.908	0.052	0.040	NA	++	++	++
AN0178-C	Cameroon	0.949	0.030	0.021	NA	++	++	++
AN0179-C	Cameroon	0.867	0.073	0.060	NA	++	++	++
AN0180-C	Cameroon	0.956	0.024	0.020	NA	++	++	++
AN0181-C	Cameroon	0.867	0.075	0.058	NA	++	++	++
AN0182-C	Cameroon	0.112	0.867	0.021	NA	a+	a+	a+
AN0183-C	Cameroon	0.945	0.029	0.026	NA	++	++	++
AN0184-C	Cameroon	0.778	0.194	0.027	++	++	++	++
AN0185-C	Cameroon	0.945	0.026	0.028	++	++	++	++
AN0186-C	Cameroon	0.907	0.051	0.042	++	++	++	++
AN0187-C	Cameroon	0.945	0.030	0.025	++	++	++	++
AN0188-C	Cameroon	0.830	0.097	0.073	++	++	++	++
AN0189-C	Cameroon	0.933	0.038	0.029	NA	++	++	++
AN0190-C	Cameroon	0.929	0.049	0.022	++	++	++	++
AN0191-C	Cameroon	0.921	0.054	0.025	++	++	++	++
AN0192-C	Cameroon	0.906	0.067	0.026	++	++	++	++
AN0193-C	Cameroon	0.870	0.088	0.043	++	++	++	++
AN0194-C	Cameroon	0.892	0.060	0.047	++	++	++	++
AN0196-C	Cameroon	0.294	0.683	0.024	a+	a+	a+	a+
AN0197-C	Cameroon	0.952	0.026	0.022	++	++	++	++
AN0198-C	Cameroon	0.040	0.947	0.013	a+	a+	a+	a+
AN0199-C	Cameroon	0.855	0.123	0.022	++	++	++	++
AN0200-C	Cameroon	0.755	0.149	0.097	++	++	++	++
AN0201-C	Cameroon	0.643	0.311	0.046	++	++	++	++
AN0202-C	Cameroon	0.898	0.070	0.031	++	++	++	++
AN0203-C	Cameroon	0.943	0.028	0.029	++	++	++	++
AN0204-C	Cameroon	0.942	0.034	0.025	++	++	++	++
AN0205-C	Cameroon	0.948	0.026	0.026	++	++	++	++
AN0206-C	Cameroon	0.943	0.032	0.025	++	++	++	++
AN0207-C	Cameroon	0.737	0.195	0.068	++	++	++	++
AN0208-C	Cameroon	0.828	0.142	0.030	++	++	++	++
AN0209-C	Cameroon	0.245	0.729	0.026	NA	a+	a+	a+
AN0210-C	Cameroon	0.902	0.056	0.042	NA	++	++	++
AN0212-C	Cameroon	0.883	0.065	0.052	++	++	++	++
AN0213-C	Cameroon	0.860	0.085	0.055	++	++	++	++
AN0214-C	Cameroon	0.170	0.805	0.025	a+	a+	a+	a+
AN0215-C	Cameroon	0.933	0.036	0.031	++	++	++	++
AN0217-C	Cameroon	0.780	0.125	0.095	a+	++	++	++
AN0218-C	Cameroon	0.945	0.029	0.026	++	++	++	++
AN0219-C	Cameroon	0.921	0.045	0.034	++	++	++	++
AN0220-C	Cameroon	0.917	0.046	0.037	++	++	++	++
AN0221-C	Cameroon	0.875	0.069	0.055	++	++	++	++
AN0222-C	Cameroon	0.918	0.054	0.028	NA	++	++	++
AN0223-C	Cameroon	0.933	0.041	0.027	NA	++	++	++
AN0224-C	Cameroon	0.067	0.916	0.017	NA	a+	a+	a+
AN0225-C	Cameroon	0.938	0.033	0.029	NA	++	++	++

AN0226-C	Cameroon	0.920	0.045	0.035	NA	++	++	++
AN0227-C	Cameroon	0.074	0.910	0.017	NA	a+	a+	a+
AN0228-C	Cameroon	0.896	0.062	0.043	NA	++	++	++
AN0229-C	Cameroon	0.919	0.045	0.036	NA	++	++	++
AN0230-C	Cameroon	0.042	0.131	0.827	NA	aa	aa	aa
AN0231-C	Cameroon	0.913	0.048	0.039	NA	++	++	++
AN0233-C	Cameroon	0.414	0.537	0.050	NA	a+	++	++
AN0234-C	Cameroon	0.027	0.942	0.031	NA	a+	a+	a+
AN0235-C	Cameroon	0.407	0.545	0.047	NA	a+	++	++
AN0236-C	Cameroon	0.042	0.161	0.797	NA	aa	aa	aa
AN0237-C	Cameroon	0.656	0.311	0.033	NA	++	a+	a+
AN0238-C	Cameroon	0.036	0.073	0.891	NA	aa	aa	aa
AN0239-C	Cameroon	0.630	0.337	0.033	NA	++	a+	a+
AN0240-C	Cameroon	0.954	0.025	0.021	NA	++	++	++
AN0241-C	Cameroon	0.040	0.268	0.692	NA	aa	aa	aa
AN0242-C	Cameroon	0.939	0.032	0.029	NA	++	++	++
AN0243-C	Cameroon	0.036	0.046	0.918	NA	aa	aa	aa
AN0244-C	Cameroon	0.913	0.044	0.042	NA	++	++	++
AN0245-C	Cameroon	0.919	0.056	0.025	NA	++	++	++
AN0246-C	Cameroon	0.949	0.030	0.021	NA	++	++	++
AN0247-C	Cameroon	0.814	0.150	0.036	NA	++	++	++
AN0248-C	Cameroon	0.048	0.121	0.831	NA	aa	aa	aa
AN0250-C	Cameroon	0.037	0.047	0.916	NA	aa	aa	aa
AN0251-C	Cameroon	0.838	0.130	0.032	NA	++	++	++
AN0252-C	Cameroon	0.799	0.152	0.050	NA	++	++	++
AN0253-C	Cameroon	0.103	0.874	0.023	NA	a+	a+	a+
AN0254-C	Cameroon	0.958	0.026	0.016	NA	++	++	++
AN0255-C	Cameroon	0.923	0.051	0.025	NA	++	++	++
AN0256-C	Cameroon	0.913	0.048	0.039	NA	++	++	++
AN0258-C	Cameroon	0.044	0.065	0.890	NA	aa	aa	aa
AN0259-C	Cameroon	0.883	0.063	0.054	NA	++	++	++
AN0260-C	Cameroon	0.055	0.930	0.015	NA	a+	a+	a+
AN0261-C	Cameroon	0.896	0.060	0.044	NA	++	++	++
AN0262-C	Cameroon	0.105	0.877	0.018	NA	a+	a+	a+
AN0263-C	Cameroon	0.847	0.106	0.048	NA	++	++	++
AN0264-C	Cameroon	0.893	0.075	0.031	NA	++	++	++
AN0266-C	Cameroon	0.039	0.086	0.874	NA	aa	aa	aa
AN0267-C	Cameroon	0.663	0.298	0.039	NA	++	a+	a+
AN0268-C	Cameroon	0.819	0.144	0.037	NA	++	++	++
AN0269-C	Cameroon	0.039	0.125	0.837	NA	aa	aa	aa
AN0270-C	Cameroon	0.881	0.081	0.038	NA	++	++	++
AN0272-C	Cameroon	0.043	0.079	0.877	NA	aa	aa	aa
AN0275-C	Cameroon	0.083	0.888	0.029	NA	a+	a+	a+
AN0276-C	Cameroon	0.037	0.048	0.915	NA	aa	aa	aa
AN0277-C	Cameroon	0.691	0.276	0.033	NA	++	++	++
AN0280-C	Cameroon	0.299	0.675	0.026	a+	a+	a+	a+
AN0282-C	Cameroon	0.924	0.044	0.033	++	++	++	++
AN0283-C	Cameroon	0.893	0.060	0.047	++	++	++	++
AN0284-C	Cameroon	0.955	0.024	0.022	++	++	++	++
AN0285-C	Cameroon	0.928	0.039	0.033	++	++	++	++
AN0286-C	Cameroon	0.106	0.876	0.018	a+	a+	a+	a+
AN0287-C	Cameroon	0.180	0.797	0.023	++	a+	a+	a+
AN0288-C	Cameroon	0.954	0.023	0.023	++	++	++	++
AN0290-C	Cameroon	0.177	0.803	0.020	a+	a+	a+	a+
AN0291-C	Cameroon	0.921	0.041	0.038	++	++	++	++
AN0292-C	Cameroon	0.884	0.063	0.053	++	++	++	++
AN0294-C	Cameroon	0.920	0.046	0.035	++	++	++	++
AN0295-C	Cameroon	0.035	0.063	0.902	aa	aa	aa	aa
AN0296-C	Cameroon	0.949	0.029	0.022	++	++	++	++
AN0297-C	Cameroon	0.913	0.048	0.039	++	++	++	++
AN0298-C	Cameroon	0.932	0.037	0.031	++	++	++	++
AN0299-C	Cameroon	0.922	0.045	0.034	++	++	++	++
AN0300-C	Cameroon	0.951	0.026	0.023	++	++	++	++
AN0301-C	Cameroon	0.052	0.246	0.702	aa	aa	aa	aa
AN0303-C	Cameroon	0.036	0.949	0.015	a+	a+	a+	a+
AN0304-C	Cameroon	0.962	0.020	0.017	++	++	++	++
AN0305-C	Cameroon	0.941	0.040	0.019	++	++	++	++
AN0307-C	Cameroon	0.887	0.063	0.050	++	++	++	++

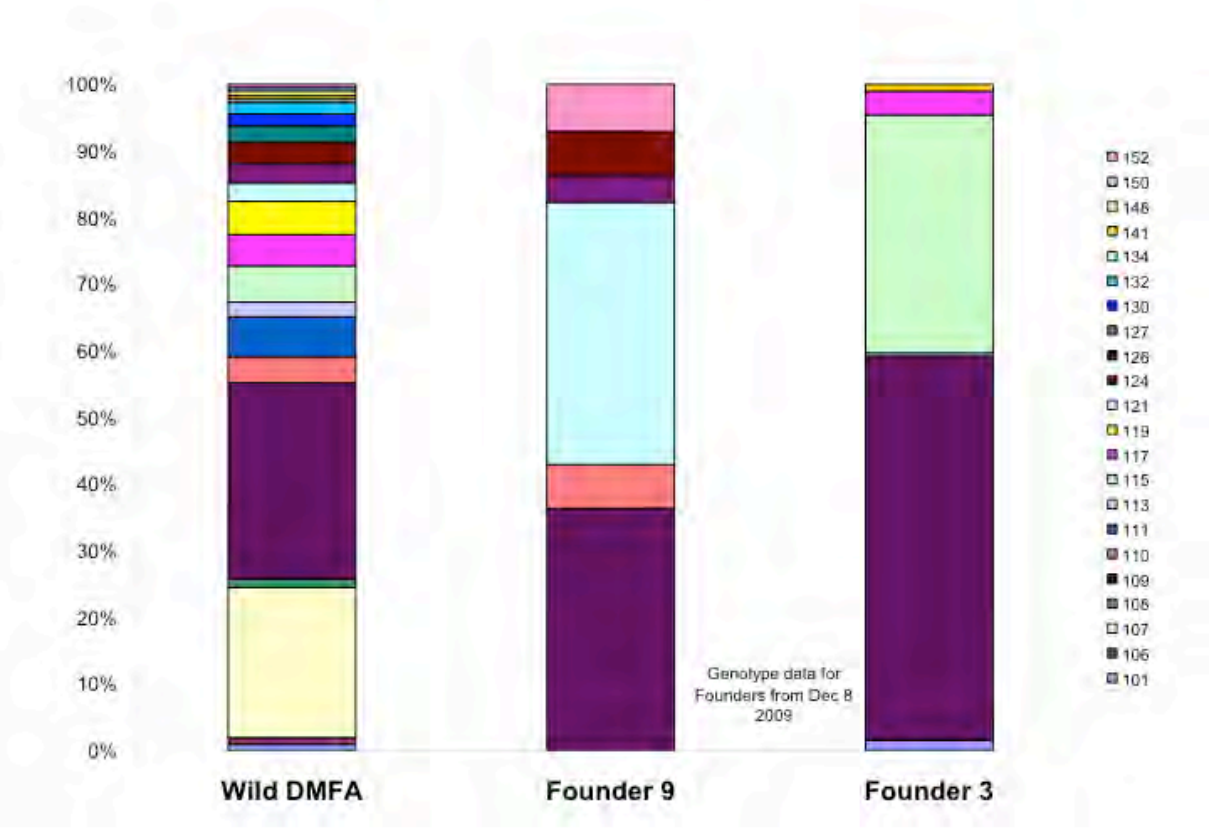
AN0308-C	Cameroon	0.938	0.036	0.026	++	++	++	++
AN0309-C	Cameroon	0.929	0.045	0.026	++	++	++	++
AN0310-C	Cameroon	0.030	0.959	0.012	a+	a+	a+	a+
AN0312-C	Cameroon	0.901	0.053	0.046	++	++	++	++
AN0313-C	Cameroon	0.351	0.621	0.027	a+	a+	a+	a+
AN0314-C	Cameroon	0.952	0.027	0.020	++	++	++	++
AN0315-C	Cameroon	0.956	0.026	0.018	++	++	++	++
AN0317-C	Cameroon	0.873	0.069	0.058	++	++	++	++
AN0318-C	Cameroon	0.918	0.043	0.040	++	++	++	++
AN0319-C	Cameroon	0.950	0.026	0.024	++	++	++	++
AN0321-C	Cameroon	0.841	0.122	0.037	a+	++	a+	a+
AR0007-C	Angola	0.811	0.155	0.034	NA	++	++	NA
AR0008-C	Angola	0.902	0.074	0.024	NA	++	++	NA
AR0009-C	Angola	0.935	0.037	0.028	NA	++	++	NA
AR0010-C	Angola	0.845	0.132	0.023	NA	++	++	NA
AR0011-C	Angola	0.955	0.024	0.021	NA	++	++	NA
AR0012-C	Angola	0.914	0.061	0.025	NA	++	++	NA
AR0014-C	Angola	0.097	0.879	0.024	NA	a+	a+	NA
AR0015-C	Angola	0.944	0.031	0.025	NA	++	++	NA
AR0017-C	Angola	0.956	0.026	0.018	NA	++	++	NA
AR0019-C	Angola	0.950	0.026	0.024	NA	++	++	NA
AR0020-C	Angola	0.953	0.026	0.021	NA	++	++	NA
AR0021-C	Angola	0.963	0.020	0.017	NA	++	++	NA
AR0022-C	Angola	0.951	0.027	0.021	NA	++	++	NA
AR0023-C	Angola	0.943	0.032	0.025	NA	++	++	NA
AR0024-C	Angola	0.950	0.026	0.023	NA	++	++	NA
AR0026-C	Angola	0.048	0.934	0.017	NA	a+	a+	NA
AR0027-C	Angola	0.953	0.025	0.022	NA	++	++	NA
AR0034-C	Angola	0.899	0.071	0.030	NA	++	++	NA
AR0035-C	Angola	0.953	0.027	0.020	NA	++	++	NA
AR0042-C	Angola	0.928	0.039	0.033	NA	++	++	NA
AR0043-C	Angola	0.954	0.026	0.020	NA	++	++	NA
AR0045-C	Angola	0.950	0.028	0.022	NA	++	++	NA
AR0047-C	Angola	0.781	0.192	0.027	NA	++	++	NA
AR0049-C	Angola	0.542	0.429	0.028	NA	a+	++	NA
AR0050-C	Angola	0.771	0.203	0.026	NA	++	++	NA
AR0051-C	Angola	0.335	0.615	0.050	NA	a+	++	NA
AR0053-C	Angola	0.911	0.049	0.040	NA	++	++	NA
AR0054-C	Angola	0.884	0.066	0.050	NA	++	++	NA
AR0057-C	Angola	0.947	0.025	0.028	NA	++	++	NA
AR0059-C	Angola	0.934	0.040	0.025	NA	++	++	NA
AR0061-C	Angola	0.958	0.025	0.018	NA	++	++	NA
AR0062-C	Angola	0.909	0.056	0.035	NA	++	++	NA
AR0063-C	Angola	0.939	0.040	0.021	NA	++	++	NA
AR0065-C	Angola	0.959	0.023	0.019	NA	++	++	NA
AR0066-C	Angola	0.960	0.021	0.019	NA	++	++	NA
AR0069-C	Angola	0.944	0.027	0.030	NA	++	++	NA
AR0070-C	Angola	0.955	0.023	0.021	NA	++	++	NA
AR0071-C	Angola	0.941	0.038	0.021	NA	++	++	NA
AR0072-C	Angola	0.874	0.098	0.028	NA	++	++	NA
AR0073-C	Angola	0.864	0.105	0.031	NA	++	++	NA
AR0074-C	Angola	0.956	0.022	0.022	NA	++	++	NA
AR0075-C	Angola	0.935	0.042	0.023	NA	++	++	NA
AR0076-C	Angola	0.918	0.060	0.022	NA	++	++	NA
AR0078-C	Angola	0.935	0.036	0.029	NA	++	++	NA
AR0079-C	Angola	0.934	0.046	0.020	NA	++	++	NA
AR0080-C	Angola	0.950	0.026	0.024	NA	++	++	NA
AR0081-C	Angola	0.937	0.035	0.029	NA	++	++	NA
AR0083-C	Angola	0.945	0.032	0.023	NA	++	++	NA
AR0084-C	Angola	0.676	0.290	0.034	NA	++	++	NA
AR0086-C	Angola	0.946	0.028	0.026	NA	++	++	NA
AR0087-C	Angola	0.945	0.030	0.024	NA	++	++	NA
AR0089-C	Angola	0.941	0.039	0.020	NA	++	++	NA
AR0090-C	Angola	0.934	0.034	0.033	NA	++	++	NA
AR0092-C	Angola	0.946	0.029	0.025	NA	++	++	NA
AR0093-C	Angola	0.883	0.069	0.047	NA	++	++	NA
AR0095-C	Angola	0.950	0.027	0.023	NA	++	++	NA
AR0096-C	Angola	0.926	0.040	0.033	NA	++	++	NA

AR0098-C	Angola	0.933	0.040	0.027	NA	++	++	NA
AR0099-C	Angola	0.935	0.044	0.021	NA	++	++	NA
AR0100-C	Angola	0.914	0.060	0.026	NA	++	++	NA
AS0001-C	Gabon	0.950	0.033	0.017	NA	++	++	NA
AS0002-C	Gabon	0.960	0.022	0.019	NA	++	++	NA
AS0003-C	Gabon	0.946	0.028	0.026	NA	++	++	NA
AS0004-C	Gabon	0.924	0.053	0.024	NA	++	++	NA
AS0006-C	Gabon	0.893	0.056	0.051	NA	++	++	NA
AS0007-C	Gabon	0.958	0.024	0.019	NA	++	++	NA
AS0008-C	Gabon	0.953	0.028	0.019	NA	++	++	NA
AS0009-C	Gabon	0.784	0.124	0.092	NA	++	++	NA
AS0010-C	Gabon	0.917	0.060	0.022	NA	++	++	NA
AS0011-C	Gabon	0.933	0.036	0.031	NA	++	++	NA
AS0012-C	Gabon	0.953	0.030	0.017	NA	++	++	NA
AS0013-C	Gabon	0.940	0.041	0.019	NA	++	++	NA
AS0014-C	Gabon	0.835	0.127	0.038	NA	++	++	NA
AS0015-C	Gabon	0.873	0.093	0.034	NA	++	++	NA
AS0016-C	Gabon	0.848	0.099	0.052	NA	++	++	NA
AS0017-C	Gabon	0.939	0.041	0.020	NA	++	++	NA
AS0018-C	Gabon	0.918	0.059	0.022	NA	++	++	NA
AS0019-C	Gabon	0.819	0.141	0.040	NA	++	++	NA
AS0020-C	Gabon	0.198	0.777	0.025	NA	a+	a+	NA
AS0021-C	Gabon	0.954	0.029	0.017	NA	++	++	NA
AS0022-C	Gabon	0.939	0.038	0.023	NA	++	++	NA
AS0024-C	Gabon	0.957	0.024	0.019	NA	++	++	NA
AS0026-C	Gabon	0.949	0.031	0.020	NA	++	++	NA
AS0028-C	Gabon	0.099	0.883	0.018	NA	a+	a+	NA
AS0030-C	Gabon	0.943	0.034	0.023	NA	++	++	NA
AS0032-C	Gabon	0.948	0.033	0.019	NA	++	++	NA
AS0033-C	Gabon	0.917	0.062	0.021	NA	++	++	NA
AS0034-C	Gabon	0.832	0.136	0.032	NA	++	++	NA
AS0035-C	Gabon	0.916	0.047	0.037	NA	++	++	NA
AS0036-C	Gabon	0.940	0.031	0.029	NA	++	++	NA
AS0037-C	Gabon	0.767	0.200	0.033	NA	++	++	NA
AS0039-C	Gabon	0.949	0.028	0.023	NA	++	++	NA
AS0042-C	Gabon	0.894	0.079	0.027	NA	++	++	NA
AS0044-C	Gabon	0.894	0.068	0.038	NA	++	++	NA
AS0045-C	Gabon	0.627	0.342	0.031	NA	++	++	NA
AS0047-C	Gabon	0.803	0.119	0.078	NA	++	++	NA
AS0049-C	Gabon	0.823	0.141	0.037	NA	++	++	NA
AS0052-C	Gabon	0.957	0.023	0.020	NA	++	++	NA
AS0053-C	Gabon	0.895	0.059	0.046	NA	++	++	NA
AS0054-C	Gabon	0.056	0.930	0.014	NA	a+	a+	NA
AS0055-C	Gabon	0.756	0.216	0.028	NA	++	++	NA
AS0056-C	Gabon	0.876	0.087	0.037	NA	++	++	NA
AS0058-C	Gabon	0.949	0.024	0.027	NA	++	++	NA
AS0059-C	Gabon	0.929	0.046	0.025	NA	++	++	NA
AS0064-C	Gabon	0.737	0.225	0.038	NA	++	++	NA
AS0065-C	Gabon	0.817	0.144	0.039	NA	++	++	NA
AS0066-C	Gabon	0.925	0.047	0.029	NA	++	++	NA
AS0068-C	Gabon	0.930	0.049	0.021	NA	++	++	NA
AS0069-C	Gabon	0.945	0.028	0.027	NA	++	++	NA
AS0070-C	Gabon	0.946	0.031	0.022	NA	++	++	NA
AS0071-C	Gabon	0.046	0.737	0.217	NA	a+	a+	NA
AS0072-C	Gabon	0.952	0.026	0.022	NA	++	++	NA
AS0073-C	Gabon	0.902	0.071	0.027	NA	++	++	NA
AS0074-C	Gabon	0.915	0.050	0.034	NA	++	++	NA
AS0076-C	Gabon	0.906	0.060	0.033	NA	++	++	NA
AS0077-C	Gabon	0.229	0.714	0.057	NA	a+	++	NA
AV0001-C	Guinea	0.041	0.052	0.907	NA	aa	aa	aa
AV0002-C	Guinea	0.052	0.061	0.887	NA	aa	aa	aa
AV0003-C	Guinea	0.126	0.854	0.020	NA	a+	a+	a+
AV0004-C	Guinea	0.906	0.051	0.043	NA	++	++	++
AV0005-C	Guinea	0.536	0.421	0.043	NA	a+	a+	a+
AV0007-C	Guinea	0.844	0.087	0.069	NA	++	++	++
AV0008-C	Guinea	0.114	0.868	0.018	NA	a+	a+	a+
AV0009-C	Guinea	0.544	0.425	0.031	NA	a+	a+	a+
AV0010-C	Guinea	0.089	0.887	0.025	NA	a+	a+	a+

AV0011-C	Guinea	0.054	0.124	0.822	NA	aa	aa	aa
AV0012-C	Guinea	0.037	0.944	0.019	NA	a+	a+	a+
AV0013-C	Guinea	0.457	0.503	0.040	NA	a+	a+	a+
AV0014-C	Guinea	0.889	0.060	0.051	NA	++	++	++
AV0015-C	Guinea	0.039	0.936	0.025	NA	a+	a+	a+
AV0018-C	Guinea	0.147	0.829	0.025	NA	a+	a+	a+
AV0024-C	Guinea	0.909	0.053	0.038	NA	++	++	++
AV0026-C	Guinea	0.937	0.036	0.027	NA	++	++	++
AV0027-C	Guinea	0.959	0.021	0.020	NA	++	++	++
AV0029-C	Guinea	0.272	0.701	0.027	NA	a+	a+	a+
AV0030-C	Guinea	0.489	0.479	0.032	NA	a+	a+	a+
AV0031-C	Guinea	0.238	0.725	0.037	NA	a+	a+	a+
AV0032-C	Guinea	0.955	0.024	0.021	NA	++	++	++
AV0033-C	Guinea	0.925	0.040	0.035	NA	++	++	++
AV0034-C	Guinea	0.034	0.076	0.889	NA	aa	aa	aa
AV0035-C	Guinea	0.057	0.114	0.829	NA	aa	aa	aa
AV0036-C	Guinea	0.044	0.155	0.802	NA	aa	NA	aa
AV0039-C	Guinea	0.934	0.046	0.019	NA	++	++	++
AV0041-C	Guinea	0.924	0.046	0.030	NA	++	++	++
AV0044-C	Guinea	0.938	0.037	0.025	NA	++	++	++
AV0045-C	Guinea	0.663	0.301	0.035	NA	++	++	++
AV0047-C	Guinea	0.932	0.044	0.024	NA	++	++	++

A6 : H603 microsatellite genotypes : founder colony representation of wild populations

Founder colonies were genotyped for a series of microsattelite markers and their diversity compared to prior samplings of the wild population. Results here are shown for the most polymorphic microsattelite H603 indicating the maintenance of a subset of the wild diversity at the time of genotyping. The increase in previously rare alleles enables the effective assaying of low-frequency alleles in the wild population, but reduces our ability to infer true importance of alleles in the wild.



Figures & Tables

Figures:

Tables

Résumé:

Population structure and genome wide association in the malaria vector Anopheles gambiae / coluzzii

Malgré le succès des insecticides pour le contrôle du paludisme, la transmission continue dans la plupart des pays d'Afrique sub-Saharienne. La recherche pour de nouveaux moyens de contrôle (plus spécialement la modification génétique des populations vecteurs), ou l'utilisation plus efficace des contrôles actuels, vont nécessiter une recherche sur les structures de populations de moustiques et les processus d'immunisation qui importent pour la transmission du Plasmodium chez les moustiques sauvages. Par ailleurs, l'utilisation des techniques d'association génomique 'GWAS' est basée sur un réelle compréhension des structures des populations.

Ma thèse inclura une description detaillée du système immunitaire du moustique, basée sur la recherche actuelle et des comparaisons génomiques; ainsi que des descriptions des principales voies immunitaires, et des gênes potentiels mal-caracterisés qui peuvent être trouvés dans une étude GWAS. Ainsi qu'une description des connaissances actuelles des structures des populations, dont la speciation du gambiae / coluzzii, et les effets des grandes variations structurelles.

Je présenterai le développement d'un nouveau moyen d'identification des variations structurelles ; utilisant les techniques d' "apprentissage automatique" permettant d'identifier les karyotypes directement à partir des séquences haut-débit, menant à des résultats d'une précision sans précédent.

Je présenterai également la première vraie cartographie génomique du 'tout-génome' du moustique. Les colonies sont fondées par des moustiques sauvages; les fondateurs sont controlées par strates, incluant également des sous-espèces et variations structurelles majeures. Avec ces colonies une méthode innovant de cartographie est utilisée: dans un premier temps, une identification des grandes régions au sein des groupes phenotypées par la perte de hétérogénéité; puis dans un second temps, le génotypage individuel 'Sequenom' sera utilisé pour une cartographie exacte. Cette méthode est utilisée pour l'identification d'une région avec un effet phenotypique sur la prévalence des infections dans la nature.

Enfin, je suggèrerai comment ces techniques peuvent être importantes à l'avenir pour l'application du contrôle génomique dans la nature.

Abstract:

Population structure and genome wide association in the malaria vector Anopheles gambiae / *coluzzii*

Despite successes in the use of insecticides in the control of malaria, malaria transmission continues in much of sub-saharan Africa. The search for novel methods of control (in particular genetic modification of vector populations), or of superior implementation of the currently available methods will require both greater knowledge of the population structure of the mosquito, and of the immune processes that are important in the wild. It is important to note that the mapping of novel immune genes, via genome wide association studies (GWAS) is predicated on a firm understanding of the population structure.

My thesis will include a detailed description of the mosquito innate immune system based on current research and comparative genomics; this will illustrate the major pathways that might be employed in the anti-malarial response, and some potential uncharacterised genes that might be implicated in any GWAS study. It will also include a summary of what is known about the mosquito's population structure, in particular the gambiae / coluzzii speciation event and the implication of chromosomal inversions in the speciation process.

I will present the development of a novel approach to the identification of chromosomal inversions; using machine-learning techniques in order to call inversion karyotypes directly from sequence, leading to calls of unprecedented accuracy.

I will also present the first truly genome-wide association study to have been performed in the mosquito. Strata-controlled populations of mosquitoes were derived from the wild, including restriction on the basis of subspecies and chromosomal inversion. A two-stage mapping design was then devised in which loss-of-heterozygosity is used to identify broad regions in phenotype pools, before fine-resolution mapping by Sequenom genotyping in individuals. This was used to identify a novel locus with a phenotypic effect on infection prevalence.

finally I will describe how these techniques and findings could be important in the future application of genetic control in the wild.