A Genomic Approach to Investigate the Evolution of Exophagy in *Anopheles* mosquitoes

Danica Joy Fabrigar

Jesus College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2015

Abstract

The development of behavioural resistance in *Anopheles* mosquitoes is an important threat to the sustainability of current vector interventions, which selectively target mosquitoes that are found indoors. Changes in the following behaviour have been reported in several *Anopheles* vectors: biting time, host preference, biting and resting behaviour. However, a critical review of the published literature reveals that behavioural resistance has not been demonstrated clearly in *Anopheles*, highlighting important limitations in using entomological data to investigate the evolution of behaviour.

The availability of next generation sequencing and low-cost custom genotyping technologies makes it possible to investigate behavioural resistance using genetic information as an alternative approach. Using such technologies, this thesis aims to investigate whether *Anopheles* vectors are evolving towards outdoor-biting behaviour ('exophagy') by examining the genetic difference between the indoor- and -outdoor population of the main African malaria vectors, *An. coluzzii* and *An. gambiae* from The Gambia and Uganda, respectively. We also investigated the role of olfactory-related genes in driving the potential genetic difference between the two populations.

We found limited divergence between the two populations overall, however close inspection identified SNPs potentially associated with outdoor-biting behaviour as well as genomic regions displaying elevated levels of differentiation between the two populations. Most notably, gustatory receptors, *Gr11* and *Gr12* and the odour receptor, *OR1* were identified as well as, genes related to transcription, signal transduction and catalysis. We discuss the potential role of these genes in

determining biting behaviour and suggest that further investigation of these genomic regions is required to confirm the early-stage divergence of indoor- and outdoor-biting *Anopheles* populations.

Acknowledgements

I would like to express my thanks to my supervisor, Prof. Charles Godfray, for his support, guidance and insights. I am tremendously grateful for his time and effort spent in reading my thesis and providing meticulous feedback.

I am grateful to Jesus College for granting me the Sir Richard Southwood Scholarship, which enabled me to undertake a DPhil.

I would like to acknowledge my collaborators from the MalariaGEN, especially Prof. Dominic Kwiatkowski, Kirk Rockett, Alistair Miles, Tiago Antão and Christina Hubbart – liberally sharing their scripts, laboratory resources and knowledge. Their enthusiasm for the project and guidance helped me in all the time of research and without them, I think I would still be buried under many lines of codes and tubes of unsequenced mosquito samples.

I would also like to thank Prof. Steve Lindsay for providing valuable insights during the initial planning stages of the fieldwork experiments and for putting us in touch with our African collaborators. My sincerest gratitude is dedicated to the team of researchers and volunteers who helped me execute the fieldwork in The Gambia and Uganda, especially Musa Jawara from the Medical Research Council (MRC) unit, The Gambia and Kilama Maxwell from the Infectious Diseases Research Collaboration (IDRC), Uganda. My time in The Gambia was a truly memorable experience.

I'd like to thank the current and past members of the Godfray lab for making me feel welcome and for sharing their knowledge of both the academic and non-academic world.

I dedicate my thanks to my mum and my sister for being my cheerleaders. My mum was a source of inspiration throughout my life, always demonstrating by example that hard work and dedication will get you far in life. My sister has also been extremely supportive. She demonstrated great kindness when she welcomed Drew and I into her home, taking care of us when we needed it most. Drew's family also deserves a big thank you – Katie and Ash have been a second family to me, patiently indulging Drew when he talks about my work for me.

I also thank the Darwins girls, Klara, Berti, Vikki, Izzy, Ellie B., Ellie W. and Catherine for the many hours of fun and listening to my PhD woes during lunch and tea breaks. To my friends, Ana, Richard, Niall, Harsita, Sanjay and Kyros thank you for always keeping me grounded and for helping me find a useful outlet for my PhD frustrations, by encouraging me to drink, eat, laugh and weightlift. They say you meet lifelong friends at university and I am fortunate enough to count them as my closest friends.

Finally, to my partner, Drew Hammond, whose patience, pep talks and humour have been sources of strength during the difficult periods of this journey. Drew listened eagerly to my ideas - even when it is, sometimes, in the early hours of the night. He also dutifully read through my earlier drafts, offering valuable insights and suggestions. More importantly, Drew had the tremendous grace to overlook my frustrations and carefully masterminded distractions by organising fun trips and activities when I needed it the most. Thank you Drew for making this journey beautiful.

Contents

List of tables

List of figures

1. Introduction

Malaria burden

Mosquitoes are vectors of both viral and parasitic diseases of significant medical importance, from malaria to dengue fever, West Nile virus and other arboviruses (Eldridge, 2005). Of these, malaria is the most devastating, representing one of the leading causes of death and morbidity. The disease is caused by *Plasmodium* parasites, which are transmitted through the bite of an infected female *Anopheles* mosquito. Over 100 species of *Plasmodium* have been described but only five of these infect humans and are thus, relevant to human public health: *P. vivax*, *P. falciparum*, *P. ovale*, *P. malariae* and *P. knowlesi* (CDC, 2010).

Over three billion people are estimated to be at risk of contracting malaria, with approximately one million deaths, according to one estimate, occurring annually (Murray *et al.*, 2012). An overwhelming majority of these deaths occur in Africa, predominantly in pregnant women and children under the age of five (WHO, 2011). The economic costs of the disease are also great, accounting for the loss of 44.7 million disability adjusted life years (DALYs) and 40% of healthcare spending in endemic countries (DFID, 2010). Such consequences have led to a strong association of malaria with poverty so that the countries with the highest risk of malaria infection tend to be among the poorest (DFID, 2010, Keiser *et al.*, 2004).

Malaria vectors

Mosquitoes of the genus *Anopheles* exclusively transmit mammalian malaria. Of approximately 500 recognised species, only 41 are of major public health

importance (Hay *et al.*, 2010) while another 30 or so have a minor role as vectors (Collins and Paskewitz, 1995). The importance of each species as vectors varies geographically. In sub-Saharan Africa, where the malaria burden is the highest, malaria is predominantly transmitted by the *An. gambiae* and *An. funestus* complexes (Sinka *et al.*, 2010).

Anopheles gambiae complex

The *An. gambiae* complex consists of at least seven morphologically indistinguishable sibling species that vary in ecology, behaviour and propensity as vectors (White *et al.*, 2011). Members include *An. gambiae s.s., An. arabiensis*, *An. bwambae*, *An. merus*, *An. melas*, *An. quadrinnulatus* A and *An. quadrinnulatus* B (Coluzzi *et al.*, 2002, Hunt *et al.*, 1998, Coluzzi *et al.*, 1985). Although closely related, members exhibit almost complete reproductive isolation, with only low levels of hybridization occurring naturally (Takken and Lindsay, 2003). Of the *An. gambiae* complex, *An. gambiae s.s.* and *An. arabiensis* are considered the most important vectors while other members are considered less or of no importance due to their restricted range or tendency to feed on animals (zoophagic) instead (White *et al.*, 2011).

An. gambiae s.s. and *An. arabiensis* are distributed across Africa, overlapping in much of their range (Sinka *et al.*, 2010). Despite their common association, the larval stage of *An. gambiae s.s.* predominates in humid environments (Coetzee *et al.*, 2000) whilst *An. arabiensis* show greater tolerance for drier conditions (Coluzzi *et al.*, 1979). Such differences in habitat preference have resulted in an asynchronous seasonal prevalence of these species, such that *An. arabiensis* are at higher proportions during the dry season whilst *An. gambiae s.s.* is more

abundant during the rainy season. Furthermore, the species segregate in their feeding and resting behaviour with *An. gambiae s.s.* displaying greater tendency to bite humans (anthropophagy), and bite and rest indoors (endophagy and endophily) (Coetzee *et al.*, 2000) (Ng'habi *et al.*, 2011).

An. gambiae s.s. can be distinguished into five chromosomal forms: Mopti, Bamako, Savannah, Bissau and Forest. Although morphologically indistinguishable, there is suggestion of reproductive isolation among these forms due to differences in ecology and behaviour (Coluzzi *et al.*, 1985, della Torre *et al.*, 2001). Independent of these chromosomal forms, two incipient species have also been described and recently, have been elevated to full species status: *An. coluzzii* Coetzee & Wilkerson (hereafter *An. coluzzii*) and *An. gambiae* Giles (hereafter *An. gambiae*), formerly referred to as “M” and “S” molecular forms, respectively (Coetzee *et al.*, 2013). For simplicity, *An. coluzzii* and *An. gambiae* are hereafter referred to as “*An. gambiae s.s.*”.

An. coluzzii and *An. gambiae* are sympatric in west and central Africa, but only *An. gambiae* exists in east Africa and is therefore, presumed ancestral (della Torre *et al.*, 2005b). Furthermore, the two species differ in their ability to exploit alternative larval habitats, contributing to their divergence. While *An. coluzzii* is associated with temporally stable larval habitats, *An. gambiae* is associated with rain-dependent habitats that are largely unpolluted and predator-free (Caputo *et al.*, 2008, della Torre *et al.*, 2005). It is therefore thought that *An. coluzzii* arose following the birth and invasion of anthropogenic habitats (della Torre *et al.*, 2002, Coluzzi *et al.*, 2002) although some suggest an earlier split (Crawford and Lazzaro, 2010).

It is also believed that pre-mating isolation plays a significant role in the divergence of these incipient species, as evidenced by the low hybridisation rates in the wild (~1.2%) (Tripet *et al.*, 2001). Indeed, in areas where they are sympatric, mating occurs in segregated swarms even when they are less than 200 feet apart (Diabaté *et al.*, 2009). Although the evolutionary forces responsible for the divergence between these incipient species are not yet fully understood, many authors suggest that different fitness trade-offs imposed by their different habitats are drivers of their speciation (White *et al.*, 2011).

Vector control

Control interventions aimed at reducing the risk of malaria transmission broadly fall into two categories: 1) parasite control which uses chemotherapy-based approaches to prevent disease development and 2) vector control which limits exposure to the malaria parasite. Historically, control measures aimed at targeting the vector population have achieved considerable success in eradicating malaria in many parts of the world (Harrison, 1978). However, malaria continues to persist in some parts of the world, particularly in Africa because of the challenges met in programme management and the emergence of insecticide resistance.

There are several anti-vector control methods that are available such as insecticide-treated bednets (ITNs), indoor residual spraying (IRS), predatory mosquitoes (*Toxorhynchites* species) (Pates and Curtis, 2005), biolarvicides (Howard *et al.*, 2007) and odour-baited traps. Of these, of ITNs and IRS are the frontline methods of malaria vector control.

Insecticide-treated nets (ITNs)

First introduced 20 years ago (Lines *et al.*, 1987), ITNs have been a long-standing tool in vector control because it affords protection both at the individual- and community-level. At the individual level, bed nets act as physical barriers against endophagic (indoor-biting) vectors. When impregnated with insecticides, ITNs have an added chemical barrier by killing vectors that are attempting to feed or diverting vectors to search for a bloodmeal elsewhere (Lindsay *et al.*, 1989, Lindsay *et al.*, 1991). When used by the majority of the population, ITNs reduce the density, survival and feeding frequency of vector populations and therefore, protects ITN and non-ITN users alike from malaria transmission at the community-level (Phillips-Howard *et al.*, 2003, Killeen and Smith, 2007).

ITNs are one of the most cost-effective interventions against malaria. Long-lasting impregnated bednets (LLINs), which are a type of longer lasting ITNs, have an annual cost of $2.10 (US dollars; range: $1.48-2.64) per LLIN. This is equivalent to $1.05 per person protected each year (Yukich *et al.*, 2007). Because they are relatively cheap, ITNs are a prominent preventive measure for large-scale deployment in malaria-endemic areas.

The mass distribution and prolonged use of ITNs have achieved considerable reductions in malaria-related morbidity and mortality. Under full coverage, a Cochrane review of 22 randomised controlled trials concluded that ITNs reduce child mortality by 18% (range: 14-29%) and reduce the clinical incidence of malaria caused by *P. falciparum* and *P.* vivax infections by half (range: 39-62%) (Lengeler, 2004). Trials in pregnant women also show that ITNs are highly effective at reducing malaria parasite loads and the incidence of miscarriages (Gamble *et al.*, 2007).

Indoor residual spraying (IRS)

Similar to ITNs, IRS have been used in vector control since the early 1940s (Hays, 2000). It involves spraying of insecticides on the interior surfaces of the house such as walls and roofs, so that mosquitoes resting on these surfaces are killed. IRS therefore reduces malaria transmission by reducing vector density and the lifespan of female mosquitoes so that they are no longer able to infect new hosts (WHO, 2006).

Historically, IRS was central to the success of the Malaria Eradication Programme (1955-1969), which culminated in significant reductions of malaria transmission in Asia, Latin America and Southern Africa. Combined with other measures, the use of IRS also led to the eradication of malaria from Europe, the former USSR as well as, several countries in Asia and the Caribbean (WHO, 2006). Today, IRS and ITNs account for almost 60% of global investment in malaria control. These investments have led to remarkable declines in the malaria burden, reducing the global malaria mortality rate by 25% and by more than 33% in WHO African regions (WHO, 2013a).

The effectiveness of IRS and ITNs largely depend on the continued efficacy of insecticides against local vector populations. In this context, the spread of the knock-down resistance (*kdr*) gene in *An. gambiae s.s.*, which confers resistance to pyrethroids and cross-resistance to DDT, in west Africa is concerning (Martinez-Torres *et al.*, 1998, Chandre *et al.*, 1999, Diabaté *et al.*, 2002). Pyrethroid resistance has also been reported in *An. funestus* (Okoye *et al.*, 2008). ITNs are additionally limited by factors such as their proper use and maintenance (e.g. checking for holes, re-impregnation of older nets with insecticides) as well as, factors related to vector behaviour. Particularly, ITNs are ineffective in areas

where the predominant vectors bite outdoors or early in the evening when people are not sleeping under a bed net.

Alternative methods

Efforts are underway to develop novel vector control strategies which aim to counteract the challenges posed by the emergence of drug and insecticide resistance. For example, genetically modified mosquitoes designed to drive vector populations to significant reductions, extinction or to render them refractory to *Plasmodium* are garnering increasing attention as alternative tools (Alphey *et al.*, 2008, Windbichler *et al.*, 2012). These approaches are intended to be highly species-specific, limiting damage to the target vector species only (Wilke *et al.*, 2009) and if successful, can be used to complement traditional vector control tools. Much work will need to be done however before such strategies can be deployed.

Vector characteristics

An overwhelming majority of the global malaria burden is found in Africa where the most efficient and effective dominant vectors of human malaria are found: *An. gambiae s.s.*, *An. arabiensis* and *An. funestus*; although *An. funestus* is often considered less important when compared to the former two (White *et al.*, 2011, Takken and Lindsay, 2003, Sinka *et al.*, 2010).

The significance of the aforementioned species as important vectors lays within their overall vectorial capacity- the daily rate at which new malaria cases arise from a previous malaria case. Vectorial capacity is in turn determined by a combination of traits such as host preference, biting behaviour, population density relative to the host, longevity and vectorial competence. Of vectorial competence,

with the exception of *An. quadriannulatus,* all members of the complex are naturally susceptible to *P. falciparum* and are thus potential vectors (White *et al.*, 2011). Yet*, An. gambiae s.s.* and *An. arabiensis* appear to have higher competence for the uptake, development and transmission of *P. falciparum* (Takken and Lindsay, 2003).

Feeding behaviour

Clearly, *Anopheles* mosquitoes that exploit resources and habitats that are associated with humans are chief vector candidates of malaria. For instance, *An. gambiae s.s., An. arabiensis* and *An. funestus* display a marked preference for human bloodmeals (Fornadel *et al.*, 2010, Lefèvre *et al.*, 2009), while *An. quadriannulatus* prefer to feed on animal blood. Furthermore, *An. gambiae s.s.* and *An. funestus* have higher tendencies to bite and rest indoors, whereas *An. arabiensis* is generally associated with outdoor-biting and -resting behaviour (White *et al.*, 2011). Beyond their feeding behaviour, the three primary vectors also display close associations with human environments so that they are often found ovipositing in water bodies created through anthropogenic environmental modifications. Together, their close association with humans make them highly efficient malaria vectors (Collins and Besansky, 1994).

Substantial variation in feeding behaviour also exists within species. *An. arabiensis* exhibits an east-west behavioural cline where it is associated with outdoor and zoophagic behaviours in eastern Africa while it displays the reverse behaviour in the west (Tirados *et al.*, 2006). Lefèvre *et al.* (2009) also cite variations in the blood-feeding habit of *An. gambiae s.s.,* becoming more zoophagic towards western Africa (Takken and Lindsay, 2003). In Burkina Faso, a field study identified a new subgroup of *An. gambiae s.s.,* named Goundry which

appears to be non-endophillic and non-anthropophilic (Riehle *et al.*, 2011). Such variation in their feeding tendencies, even within the same species, therefore emphasises the effects of the interaction between genetic and environmental components.

Factors influencing host choice

The final outcome of host choice is determined by vector host preference, their time of biting and tendency to feed indoors or outdoors (Takken and Verhulst, 2013). In turn, extrinsic and intrinsic factors determine these behavioural characteristics. Extrinsic factors include host odour (Lindsay *et al.*, 1993a), blood quality (Takken *et al.*, 2002), infection status (Mukabana *et al.*, 2007, Lacroix *et al.*, 2005) and climate (Balenghien *et al.*, 2006, Laarman, 1958). In contrast, intrinsic factors include the mosquito's physiological state and genes that confer an adaptive advantage when feeding on a specific host or location (i.e. indoor or outdoor). The following provides a brief overview of both extrinsic and intrinsic determinants of host choice with emphasis on the role of odour and genes[*].

Intrinsic factors

Physiology

Shortly after mating, females express a strong behavioural response to blood meal sources, marking the onset of host-seeking behaviour (Foster and Takken, 2004). During this time, females express their inherent preference for a specific host. However, if suitable hosts are unavailable, limited (Killeen *et al.*, 2001) or inaccessible (Bøgh *et al.*, 1998), vectors may feed on other hosts because they have a greater need to safeguard their reproduction and fitness.

[*]A review by Takken and Verhulst (2013) provides further details of the factors influencing host choice.

Genetics

Differences in host preferences and propensity to rest/bite indoors or outdoors between the two closely related species *An. gambiae and An. arabiensis,* provide clear evidence for a genetic basis for mosquito feeding behaviour. The heritability of host preference has been demonstrated experimentally through backcrossing and selection experiments. When presented with a choice between a man and a cow in the same experimental hut, *An. gambiae s.s.* females and their progeny expressed the same host preference. Additionally, *An. gambiae s.s.* and *An. vestipennis* strains have shown altered host preferences within a just few generations of artificial selection (Ulloa *et al.*, 2004, Gillies, 1964). Similarly, hybrids between anthrophophillic and zoophagic strains of *Aedes aegypti* expressed host preferences that were intermediate to their parental strains (Mukwaya, 1977). These results demonstrate the influence of genes in shaping behavioural evolution.

To date little is known about the genetic determinants governing host preference, except that polymorphic chromosomal inversions may play a role. In Nigeria, a correlation between the site of collection (i.e. indoors/outdoors), and different structural arrangements (so called “karyotypes”) of the 2R inversion in *An. gambiae s.s.* and *An. arabiensis* was found. This also appeared to be related to climactic and ecological factors so that the arrangements commonly found indoors were associated with more arid zones while those found outdoors occurred at higher frequency in humid areas (Coluzzi *et al.*, 1977). Similarly, surveys conducted on *An melas* in The Gambia showed a significant variation in the frequency of 2Rn arrangements across villages and resting sites (i.e. indoor/outdoor sites) (Bryan *et al.*, 1987). In a more recent research, McBride *et*

al. (2014) showed that the divergence in host preference between anthrophophillic and zoophagic forms of *Ae. aegypti* is linked to seven major allelic variations which affect the expression and ligand-sensitivity of the odourant receptor, *OR4.* Differences in the protein sequences of 16 *Anopheles* species have also been demonstrated recently and provide an insight into how dominant vectors of human malaria came to specialise on humans (Neafsey *et al.*, 2015). Together, this suggests that mosquito behaviour is determined by a combination of genetic variants and physiological responses to environmental heterogeneities.

Extrinsic factors

External factors affect host choice, particularly in situations where the preferred host is not available and/or when their distribution is sparse (Killeen *et al.*, 2001). Adverse weather conditions (Laarman, 1958) can also prevent long-distance migration and cause vectors to feed on other hosts. Other factors such as gender-specific odours, host defensive behaviour (Charlwood *et al.*, 1995), infection status (Koella *et al.*, 1998, Koella *et al.*, 2002) and body mass (Torr *et al.*, 2006) of the host also contribute to the final outcome of host choice. For instance, host body size appears correlated with the size of the olfactory cue so that larger hosts presumably produce greater levels of carbon dioxide (Torr *et al.*, 2006). Similarly, *Plasmodium* parasites may exert an influence on vector feeding behaviour such that vectors harbouring the infective stages of the parasite preferentially feed on malaria-infected children rather than those that were not infected (Lacroix *et al.*, 2005).

<u>Odour</u>

Olfactory cues play a primary role in determining behaviours related to the search for hosts, sugar and oviposition sites (Takken and Knols, 1999). Much of the

current literature concerning olfaction in *Anopheles* mosquitoes focuses on the attractiveness of compounds such as carbon dioxide (Costantini *et al.*, 1996) (Gillies, 1980), carboxylic fatty acids (Bosch *et al.*, 2000, Knols *et al.*, 1997) and ammonia (Geier *et al.*, 1999, Smallegange *et al.*, 2002), which have now been confirmed as mosquito attractants.

Differences in odour sensitivity determine the host preference and final outcome of host choice in different species of *Anopheles*. For example, carbon dioxide (CO_2) drives greater responses in *An. quadriannulatus* and *An. arabiensis* than *An. gambiae s.s.* (Costantini *et al.*, 1996) (Dekker and Takken, 1998). These differences in response reflect their feeding behaviour; for zoophagic and opportunistic feeders, CO_2 is a reliable cue for the presence of a vertebrate host while for *An. gambiae s.s.,* detecting CO_2 alone does not signify human presence (Takken and Knols, 1999). Instead, odour blends associated with human sweat, skin microflora and urine elicit greater excitation in *An. gambiae s.s* (van den Broek and den Otter, 1999, Cork and Park, 1996, Carey *et al.*, 2010).

Odour reception and its importance in host-seeking activity

Host-seeking is loosely defined as the expression of behaviour that increases the chance to locate a suitable host and therefore, obtain a blood meal. It can be divided into three phases (Lehane, 2005):

1. *Non-oriented search*- internal signals driven by the gonotrophic cycle in *Anopheles* females activate non-oriented behaviour to increase the likelihood of them coming into contact with stimuli derived from a potential host.

2. *Activation and orientation*- the long-distance reception of host stimuli marks the activation of oriented flight towards the stimuli source.
3. *Attraction*- once they are in the host's immediate vicinity, the final phase involves expression of behaviours which determine the suitability of the host.

Molecular events in odour reception

Over two decades of research on invertebrate olfaction have identified several molecular components that are involved in the signal transduction of odours. These include G-proteins, odour binding proteins (*OBPs*), odour-degrading enzymes (*ODEs*), arrestins and odour receptors (*ORs*) (Bohbot *et al.*, 2010b).

Olfactory signal transduction is initiated once odour molecules are adsorbed through the porous cuticles of the olfactory organs, the antenna and the maxillary palps (Dahanukar *et al.*, 2005). Once adsorbed, *OBPs* transport odour molecules to *ORs* located in the dendritic membranes of olfactory receptor neurones (ORNs) (Bohbot *et al.*, 2010a). The binding of *ORs* to odour molecules triggers a signalling cascade involving secondary messenger pathway, leading to the depolarisation of the ORN dendrite and the generation of action potential signals to the antennal lobe(s) (Zwiebel and Takken, 2004). Ultimately, information regarding the quality, quantity and spatio-temporal patterns of the odour is conveyed.

ORs belong to a large family of G-protein coupled receptors (*GPCRs*) comprising of 79 *OR* genes in *Anopheles.* In insects, ORNs express two types of *OR* genes: a highly conserved, broadly expressed olfactory co-receptor (*orco*) and a

conventional†, selectively expressed *OR* (Benton *et al.*, 2006, Vosshall and Hansson, 2011). Together, they form a heterodimeric complex with the conventional *OR* conferring the complex' binding specificity (Larsson *et al.*, 2004) while the *orco* ensures that the ligands are bound properly. *ORs* are also thought to play a role in the signalling of chemo- and photoreception (Hill *et al.*, 2002).

OBPs are water-soluble, globular proteins that are secreted into the sensillar lymph where they bind, solubilise and transport odour compounds to *ORs*. *OBPs* belong to a large multi-gene family that includes pheromone-binding proteins (*PBPs*) and antennal-binding proteins *(ABPs*) (Xu *et al.*, 2003). In addition to their role in transport, *OBPs* are also though to facilitate the interaction of *ORs* and odour molecules as well as, the degradation of odour compounds by *ODEs* for the desensitisation and/or protection of *ORNs* from toxic chemicals (Andronopoulou *et al.*, 2006). To date, 66 *OBP* genes have been identified in *Anopheles* (Vieira and Rozas, 2011).

From gene to behaviour

The ability to perceive odour blends is determined by the tuning of *ORs*, which can be narrowly or broadly tuned (Wang *et al.*, 2010b, Carey *et al.*, 2010). Narrowly tuned receptors are of particular interest in the context of their role in host-seeking activity. For instance, *ORs* 1, 2, 5, 8 and 65 respond strongly to odour chemicals associated with human sweat (Meijerink *et al.*, 2000), breath (Takken and Knols, 1999), skin microflora (Verhulst *et al.*, 2009) and urine (Sun *et al.*, 2008). The species-specific response to odours is therefore determined by different *OR* tuning

†The term "conventional" refers to a superfamily of *OR* genes that encode multi-transmembrane domain receptors with no homology to nematode or vertebrate *ORs,* unlike the unconventional, *orco* which has homologs in other insect species (Hill *et al.,* 2002).

profiles that together, create a unique functional profile that reflects the needs of each species (Carey *et al.*, 2010).

OR families of *An. gambiae s.s.*, *Aedes aegypti* and *Drosophila melanogaster* show strong functional and protein sequence divergence as well as, species-specific expansion of particular *OR* genes or genes clusters (Hill *et al.*, 2002, Neafsey *et al.*, 2015). Carey *et al.* (2010) compared the *OR* response profiles of *D. melanogaster* with *An. gambiae s.s.* and found that *An. gambiae s.s.* responds to aromatic compounds while *D. melanogaster's ORs* are tuned to detect esters. Such differences mirrors their ecological differences, with *An. gambiae s.s.'s ORs* tuned to key compounds associated with humans whereas *D. melanogaster* is tuned to respond to fruit volatiles (Carey *et al.*, 2010). Such patterns of evolution therefore reflect a role of *OR* genes in mediating adaptation and environmental responses in individual species (McBride, 2007, Smadja *et al.*, 2012).

Changes in the gene structure of olfactory-related genes such as *ORs* and *OBPs* can alter an organism's sensitivity to odour compounds and consequently, modify their response profile. For instance, single-base pair changes in the DNA sequence, known as single nucleotide polymorphisms (SNPs), were associated with variable responses to key odour molecules in *D. melanogaster OBP99* (Wang *et al.*, 2010a, Rollmann *et al.*, 2010). Similarly, epigenetic effects such as methylation and histone modification alter gene expression (Jaenisch and Bird, 2003). Changes in gene expression through these modifications are heritable changes that are associated with the chromosome but are not based on DNA sequence (commonly known as "transgenerational epigenetic inheritance'') (Bird, 2007, Youngson and Whitelaw, 2008). This illustrates that variations in the DNA

sequence as well as, epigenetic effects are potential avenues through which variations and changes in mosquito feeding behaviour may arise.

Modern methods in vector population studies

Population genetics can be broadly defined as the study of variation in natural populations with regards to their gene and genotype frequencies, and the factors that influence this variation. These factors are evolutionary processes such as mutation, natural selection, gene flow, and random genetic drift, which may have locus-specific or genome-wide effects. Locus-specific processes, such as mutation and selection, provide information on the genes that may be involved in adaptation whereas demography and phylogenetic history may only be inferred by studying genome-wide processes (such as gene flow and drift). The use of genotypic information therefore allows these two effects to be separated in order to investigate population and ecological genetics as well, as evolution (Luikart *et al.*, 2003).

Over the past decade, there has been a rise in the development of population genetic approaches to study population structure and evolution. Considerable progress has been made with the advent of advanced computing and genotyping technologies which have made genome-wide sampling feasible at increasingly low costs. Furthermore, the genome of *An. gambiae s.s.* has been sequenced and is available publicly as well as, 17 other *Anopheles* species which include several of its sibling species (Holt *et al.*, 2002, Sharakhova *et al.*, 2007, Neafsey *et al.*, 2015, Marinotti *et al.*, 2013). The availability of such tools and resources therefore provides an avenue to investigate the genetic variability within and between-species at a level that was not possible before.

An important application for genotyping technologies is its use in understanding the diversity and dynamics of *Anopheles* populations. Though malaria control efforts have experienced multiple successes in reducing transmission rates, the ability of *Anopheles* populations to develop resistance to insecticides has made it difficult to eradicate malaria. One way in which we can overcome this problem is through exploiting the use of sequencing technologies for high-throughput genotyping of natural populations to capture rare, divergent genomic regions (Pool *et al.*, 2010, Helyar *et al.*, 2011). In doing so, it is hoped that we can gain an understanding of the complex population structure of these malaria vectors as well as, develop early warning of vector resistance occurring.

Genotyping technologies

Next-generation sequencing (NGS) technologies are a new generation of modern sequencing technologies. At present, there are three main platforms for large-scale DNA sequencing: Illumina/Solexa Genome Analyzer (Bentley *et al.*, 2008), Applied Biosystems SOLiD (Smith *et al.*, 2008) and Roche/454 FLX sequencing (Margulies *et al.*, 2005). These technologies have the capacity to generate large volumes of sequencing data, with a single sequencing run producing approximately 1 terabase of data within a relatively short period of time and at cost that is becoming more affordable (Illumina, 2013). The development of these technologies has made it increasingly feasible to obtain whole-genome information for entire populations, leading to a flurry of large-scale sequencing projects being launched (e.g. 5,000 insect genome project; see http://arthropodgenomes.org/wiki/i5K). These projects have increased our understanding of the relationship between genomic variation and phenotype. NGS

technologies also has important applications in public health, particularly in the surveillance of infectious disease and pathogenic outbreak strains (Lipkin, 2013).

For many applications, sequencing of entire genomes may not be necessary or practical, and therefore, high-throughput genotyping technologies such as the Sequenom MassARRAY iPLEX platform have been developed. With the Sequenom technology, approximately 100,000 genotypes per day can be generated, facilitating large-scale genotyping analysis with a fast turnaround time (Gabriel *et al.*, 2009). Sequenom has particular applications in the fine mapping of disease-causing loci in several genome-wide association studies (Chang *et al.*, 2009, Barrett *et al.*, 2008) as well as in candidate-genes studies.

Genetic information obtained from these technologies is readily applied to existing clustering methods to infer population demography. In this way, the population ancestry of individuals may be inferred (Falush *et al.*, 2003, Falush *et al.*, 2007) and the genetic relationships between individuals/populations may be quantified (Li *et al.*, 2008, Novembre *et al.*, 2008). Already it has revealed distinct geographic substructure of human populations (Hinds *et al.*, 2005) and has been useful for inferring divergence and evolutionary relationships in both non-model and model organisms such as birds (McCormack *et al.*, 2012) and *Drosophila* species (Macpherson *et al.*, 2007), respectively. NGS has also been a valuable tool in malaria research especially in the study of the genetic variation of *P. falciparum* isolates (Bright *et al.*, 2012, Manske *et al.*, 2012) and *Anopheles* populations (Besansky, 2008, MalariaGEN, 2013, Weetman *et al.*, 2010, Clarkson *et al.*, 2014).

Single Nucleotide Polymorphisms (SNPs) as molecular markers

Genetic markers are useful for a wide variety of applications such as in diagnostics and phylogenetic and taxonomic analysis (Scott *et al.*, 1993, Favia *et al.*, 1997). Ideal markers should be abundant, widely dispersed, easy to score and reproducible across different laboratories. Traditional markers include mitrochondrial DNA variants (Besansky *et al.*, 1997), microsatellites (Braginets *et al.*, 2003), restriction fragment length polymorphism (RFLP) (Favia *et al.*, 1997) and amplified fragment length polymorphisms (AFLP) (Black and Lanzaro, 2001).

In the past, microsatellites were the preferred markers because they were relatively easy to score and fairly abundant throughout the genome. Microsatellites are a class of repetitive DNA sequences that are 2-8 nucleotides long and are highly polymorphic in length (Selkoe and Toonen, 2006). Indeed microsatellites have been used to investigate the level of divergence (Lanzaro *et al.*, 1998) and gene flow (Kamau *et al.*, 1999) between *Anopheles* populations.

Increasing demands for high density genetic markers combined with advances in polymorphism detection and genotyping technologies have made SNPs the markers of choice for most genetic studies. SNPs are simple bi-allelic co-dominant markers that vary at a single base pair position. Compared to microsatellites, SNPs are much more abundant, occurring at a rate of one SNP per kb in humans (Wang *et al.*, 1998) and one SNP per ~100 bp in *An. gambiae* (Lawniczak *et al.*, 2010). Furthermore, methods of SNP detection are easily automated, have much lower error rates and are reproducible. Their application in fine-resolution population-based studies, association analysis and in identifying genomic regions under selection have been demonstrated extensively in recent years (e.g.

Weetman *et al.* (2014), Weetman *et al.* (2010), Reidenbach *et al.* (2012), Neafsey *et al*. (2010)).

Methods for studying population genetics: an overview

Currently, there are two popular methods used to infer population structure: principal component analysis (PCA) and cluster analysis. The application of PCA in population genetics was introduced by Cavalli-Sforza *et al.* in 1978 and has since then been a powerful tool for identifying complex population substructures, especially in mapping the worldwide substructure of human populations (Paschou *et al.*, 2007). Similarly, model-based clustering approaches to infer population structure such as STRUCTURE (Pritchard *et al.*, 2000) and ADMIXTURE (Alexander *et al.*, 2009) are gaining increasing popularity. Unlike PCA which is typically a first-step tool, cluster analysis has greater sensitivity in detecting hidden genetic structures (Jakobsson *et al.*, 2008). Together, both tools have been essential for studying ancient and recent population histories as well as, in establishing population substructures for the unbiased, analysis of association studies (Marchini *et al.*, 2004, Ziv and Burchard, 2003).

A major application of high-throughput genotyping and NGS technologies is their use in characterising genetic variation to identify targets of positive selection. There are several statistical methods available to identify positive selection[‡], but all methods broadly detect the following signatures: 1) reduction in variability and an excess of rare alleles (Tajima, 1989, Fay and Wu, 2000), 2) linkage disequilibrium in selected loci (Przeworski, 2002), 3) highly differentiated loci

[‡] See Biswas & Akey (2006) for an overview of these methods.

between populations (Weir and Cockerham, 1984) and 4) deviations of nucleotide variation from neutral expectations (Hudson *et al.*, 1987).

Another important use of genome-wide information is to test for associations between variants and a given trait or condition in a genome-wide association study (GWAS). Using allele frequency information derived from thousands of SNP loci, an association between a SNP variant and trait can be inferred under a logistic or linear regression model for dichotomous (e.g. indoor and outdoor-biting behaviour) and continuous traits (e.g. height), respectively (Cardon and Bell, 2001). Alternative methods of associations testing include chi-square or contingency table-based tests as well as, Cochran-Mantel-Haentzel test for stratified analysis (Stranger *et al.*, 2011).

Research objectives

Reductions in malaria transmission have been achieved as a result of the mass-scale use of ITNs and IRS. However, changes in host preference, biting time and in the propensity to bite and rest indoors or outdoors have been reported in *Anopheles* vectors in several countries. Such changes pose a threat to the sustainability of current interventions, which are predominantly indoor-based. Thus, there is a need to improve our understanding of *Anopheles* populations and their capacity to respond to human intervention through behavioural changes. Such research will need to investigate the local behavioural and genetic variation of vector populations to monitor any undesirable changes. This thesis aims to investigate whether *Anopheles* vectors are evolving towards outdoor-biting behaviour ('exophagy'). We also examine the role of odour-mediated processes in

modulating changes in feeding behaviour. Specifically, we had the following research objectives:

1. Conduct a systematic review to identify studies demonstrating the development of behavioural resistance in response to the mass-scale use of ITNs and IRS on *Anopheles* populations.
2. Describe the host-seeking behaviour of the indoor- and outdoor-biting populations of the *An. gambiae* complex and compare it to previous observations.
3. Compare the genetic difference between *An. gambiae* and *An. arabiensis,* and demonstrate a genetic difference between the indoor and outdoor populations of *An. coluzzii* and *An. gambiae* at candidate olfactory-related genes.
4. Demonstrate the use of intact legs as alternative PCR templates to bypass nucleic acid extract and improve current high-throughput genotyping of *Anopheles* mosquitoes.
5. Identify candidate genes/regions associated with the outdoor-biting phenotype in *An. gambiae* mosquitoes using a combined approach of F_{st} analysis and genome-wide association study.
6. Validate and elucidate the possible functional relevance of the top candidate genes/regions in an independent cohort of *An. gambiae* using a similar strategy and perform a meta-analysis of the combined sample.

2. Changes in *Anopheles* biting behaviour in response to vector interventions: a systematic review

Abstract

Background

Emerging resistance of *Anopheles* vectors to insecticides threaten the current successes gained in reducing the malaria burden. Behavioural resistance in the form of changes in feeding behaviour has been reported in several countries following the introduction and scale-up of vector interventions. We aimed to examine the evidence of behavioural resistance in *Anopheles* vectors in response to indoor-based vector control.

Methods

We mined the Pubmed database using search terms related to behavioural resistance mechanisms in *Anopheles* mosquitoes. Studies were included if they investigated changes in one or more of the following behaviours: biting time, host preference, biting and resting behaviour. Only randomised controlled trials, non-randomised trials with pre-and post-intervention data, and pre-post studies were included. For host preference and resting behaviour, percentage reductions in vector density or human blood index (HBI) were calculated. To investigate biting behaviour, we calculated the difference in the proportion of indoor-biting mosquitoes caught between pre- and post-intervention periods.

Results

Fifteen studies were included in the analysis. Possible evidence of a genetic change towards early-biting behaviour was demonstrated for *An. farauti* and *An. funestus,* but not *An. gambiae,* by three studies. Similarly, potential evidence of a shift in the host preference and toward outdoor-resting behaviour was found for *An. funestus* by a total of three studies while six studies suggests that a change towards exophagy was observed in *An. farauti, An. funestus* and *An. gambiae.* Despite these results, no conclusion regarding behavioural resistance can be drawn from all fifteen studies owing to methodological weaknesses which limited data interpretation.

Conclusion

The scale-up of vector control may be causing a strong selection pressure for *Anopheles* vectors to develop resistant behaviours, however, due weaknesses in the study design there was limited evidence for behavioural resistance occurring in *Anopheles* vectors. Therefore, there is an urgent need to establish guidelines for measuring behavioural resistance. Here, we outline recommendations for measuring behavioural resistance in the field.

Introduction

Malaria remains a major public health problem causing, according to one estimate, over one million deaths annually (Murray *et al.*, 2012). While this burden has been reduced following the scale-up of effective interventions, reports of malaria resurgence threaten their sustainability. Scaling-up of intervention creates the risk that vectors will develop resistance and consequently, cause resurgence.

Anopheles mosquitoes, the exclusive vectors of human malaria, feed on mammalian blood. For the dominant vectors, the search for a blood meal host is characterised by their propensity to enter houses, bite humans and bite at times when most people are asleep (White *et al.*, 2011). Such behavioural traits are the cornerstones upon which current control methods are based. These include the mass distribution of insecticide-treated nets (ITN), long-lasting insecticide-treated nets (LLINs) and spraying of insecticides (IRS) (Lindsay *et al.*, 2002). However, the widespread use of these tools has had a substantial effect on mosquito survival and reproduction, creating a strong evolutionary pressure for the selection of resistant phenotypes (Haynes, 1988).

There are different mechanisms of resistance, often involving reduced sensitivity to insecticides through increased metabolic capacity or target site modifications (Hemingway *et al.*, 2004). Although less frequently reported, resistance can also occur in the form of behavioural changes. These include: 1) shifts in biting time - in an attempt to bite humans that are unprotected by bed nets, 2) change in host preference - to feed on more accessible species and 3) the propensity to bite and 4) rest outdoors (Russell *et al.*, 2011).

Confirmed examples of behavioural resistance in nature are rare, although it has been demonstrated by laboratory-based experiments in other insect species. A strain of German cockroaches, *Blattella germanica*, has evolved an aversion to glucose following exposure to gel baits containing lethal doses of pesticide ingredients (Silverman and Bieman, 1993). Linkage analysis using crosses between resistant and vulnerable strains showed that this behaviour was linked to the inheritance of a single, semi-dominant trait (Ross and Silverman, 1995). Similarly, leaf-cutter ants (*Atta cephalotes*) exhibited avoidance behaviour in the presence of the diurnal parasitoid fly, *Neodohrniphora curvinervis* by shifting from diurnal to nocturnal foraging activity (Orr, 1992).

In *Anopheles,* changes in vector feeding behaviour towards biting outdoors and early in the morning when humans are unprotected have been reported for several vector species (Charlwood and Graves, 1987, Russell *et al.*, 2011). Despite these observations, the evidence base for behavioural resistance in *Anopheles* mosquitoes has been a subject of contention. One reason is that many observations of altered vector behaviour can be attributed to changes in species composition, because the most behaviourally vulnerable taxa are selectively suppressed. Another reason is that apparent behavioural changes may also be attributed to the killing or repellent effect of the insecticide. In such cases, the indoor vector population is reduced (through the suppression of vector density or increased exit rates), increasing the proportion of biting that occurs outdoors. However this response is temporary and will disappear if the insecticide is removed unlike genuine behavioural resistance which is long-term, difficult to reverse and may potentially lead to the spread of resistance genes. Furthermore, studies investigating the impact of vector control are highly variable in their study

design and ecological setting making it difficult to generalise their findings. To our knowledge, no systematic review has been conducted to investigate behavioural resistance in *Anopheles* vectors.

We performed a systematic review to investigate the potential effect of vector control on *Anopheles* feeding behaviour. We aimed to assess the evidence of behavioural resistance in *Anopheles* populations so that true changes in vector behaviour can be disentangled from confounding effects such as changes in taxonomic composition.

Methods

Definition of behavioural resistance

Behavioural resistance is defined as the development of behaviours that result in reduced contact with insecticides, thereby increasing the probability of survival in an otherwise fatal environment (Sparks *et al.*, 1989). In the context of *Anopheles* vectors, this refers to changes in the following behaviours: i) time of biting, ii) host preference iii) propensity to rest indoors or outdoors (endophily and exophily, respectively) and iv) propensity to bite indoors or outdoors (endophagy and exophagy, respectively).

Review questions

Our aim was to provide a systematic review of the studies reporting behavioural resistance in *Anopheles* vectors following interventions. Specifically, we aimed to investigate if *Anopheles* vectors display behavioural resistance in response to the use of vector control interventions.

Literature Search

A systematic search of published literature in the PubMed database was performed in March 2013 and repeated in November 2014 using the following search strings: *Anopheles* and 1) outdoor-biting, 2) biting behaviour, 3) host preference, 4) biting time, 5) behaviour insecticide nets or bed nets. Potential articles were screened based on their abstracts and titles with the goal of finding any reference to behavioural resistance. Additionally, studies cited in the previously retrieved articles were subjected to the same screening procedure. Only articles published in the English language were included.

Eligibility criteria

Articles were critically appraised in full to assess their relevance to behavioural resistance in *Anopheles* vectors. Studies were included if they aimed to investigate the effects of untreated bed nets, insecticide treated nets (ITN or LLINs) and IRS on *Anopheles* behaviour. Additionally, articles were included if they investigated the following behavioural traits: biting time, host preference, resting and biting behaviour.

In terms of study designs, we included i) randomised controlled trial* with or without pre- and post-intervention data, ii) non-randomised control trials with or without pre- and post-intervention data and iii) pre-post study, where data was available before (pre-) and after (post-) the intervention was introduced. Unlike control trials where localities are assigned to an intervention or a control group, all localities receive the intervention after a period of pre-intervention surveillance in a pre-post study.

Data extraction and analysis

We extracted data from each study based on a pre-designed data extraction form (Table 2.1). The form required us to record key components of general study information (title, author, country of study, *Anopheles* species under investigation), study characteristics (vector control intervention, methodology used to sample vectors) and findings. We reviewed other articles referenced by the study for missing or incomplete details. This information was recorded in a spread sheet format.

*A control trial is considered randomised if the intervention (i.e. ITNs, LLINs, IRS) was randomly allocated to houses/villages. In contrast, the method of allocation is not random in a non-randomised control trial.

Studies investigating changes in the time of biting were extremely heterogeneous in the way that outcomes were reported. For example, some studies reported the mean biting density (bites/person/hour) while others reported the proportion of mosquitoes caught per hour. In contrast, resting collections were generally reported as the mean vector density collected indoors (i.e. average number of mosquitoes collected each night) while host preference was measured in terms of the proportion of bloodmeals derived from humans, termed human blood index (HBI). However for all three forms of behaviour resistance, we were unable to perform a meta-analysis because, either the outcome measures were heterogeneous or there was inadequate data provided. With regards to resting behaviour and host preference, no study reported the standard error for mean vector density and we could not calculate it from the data presented in the papers. Therefore we reported our findings in a structured, narrative format for all three types of behaviour resistance.

Almost all studies that investigated changes in biting behaviour reported the outcomes as the proportion of mosquitoes that were caught indoors, known as the "endophagy rate". From this we can measure the changes in endophagy by calculating the difference in proportion of (*D*) endophagic mosquitoes between 1) the pre-intervention (*p*) and post-intervention (*q*) periods for pre-post studies or 2) intervention and control areas (*q1-q0*) for control trials and controlled pre-post studies (i.e. intervention vs control studies but only post-intervention data is available). The proportion of endophagic mosquitoes was calculated using the following formula:

$$D = \frac{I}{N}$$

I = the total number of mosquitoes caught indoors
N= the total number of mosquitoes caught indoors and outdoors

Subsequently, the difference in proportion between the pre-intervention (*p*) and post-intervention (*q*) periods or between intervention and control areas was calculated (*q-p* or *q1-q0*), as well as the standard error (S.E.) of their difference:

$$\mathrm{S.E.}\,(\mathrm{q}-\mathrm{p}) = \sqrt{\left[\left(\frac{\mathrm{p}\,(1-\mathrm{p})}{\mathrm{N}_1}\right) + \left(\frac{\mathrm{q}(1-\mathrm{q})}{\mathrm{N}_2}\right)\right]}$$

Note: *q* and *p* were substituted for *q1* and *q0* if making a comparison between intervention and control areas

Negative *D* values represent an increase in outdoor-biting vectors relative to indoors, positive values suggests a decrease while a value of zero denotes no change. Where data is available for multiple sites, we calculated endophagy by taking the total number of mosquitoes collected across all sites. Furthermore, if data is available for multiple time points during the post-intervention period, for example yearly outcome, we used the latest available data (i.e. the last year of vector surveillance) for the post-intervention measure.

We calculated the percentage reduction in vector density or HBI for studies investigating resting behaviour, biting behaviour or host preference, using three different approaches. For pre-post studies (i.e. studies where only intervention areas were monitored during the pre- and post-intervention period) we calculated the percentage reduction using the following formula:

$$100 * \left(1 - \frac{\mathrm{q1}}{\mathrm{p1}}\right)$$

q1= outcome measure (e.g. mean number of indoor-resting mosquitoes) observed in the intervention area during the post-intervention period.
p1= outcome measure (e.g. mean number of indoor-resting mosquitoes) observed in the intervention area during the pre-intervention period.

Alternatively, for control trials with pre- and post-intervention data (i.e. randomised controlled pre-post studies) we calculated the reduction in vector density or HBI using the formula (Wilson *et al.*, 2014):

$$100 * \left(1 - \frac{\left(\frac{q1}{q0}\right)}{\left(\frac{p1}{p0}\right)} \right)$$

q1, q0= outcome measure (e.g. mean vector density) observed in the intervention and control areas during the post-intervention period, respectively.
p1, p0= outcome measure (e.g. mean vector density) observed in the intervention and control areas during the pre-intervention period, respectively.

Finally, for studies where only post-intervention data is available (i.e. control vs intervention study design), we used the following formula to calculate the percentage reduction in vector density:

$$100 * \left(1 - \frac{q1}{q0}\right)$$

q1, q0= outcome measure (e.g. mean vector density) observed in the intervention and control areas during the post-intervention period, respectively.

Where data is available for multiple intervention and control areas, we calculated the percentage reduction by calculating the mean value of the outcome measure (e.g. number of mosquitoes collected), applying equal weights to all sites. We did not calculate confidence intervals around percentage reductions because there were substantial variation in the study designs (e.g. the number of nights in which mosquitoes were collected differed between pre- and post-intervention periods) and the way outcomes were reported (e.g. total number of mosquitoes collected across all study sites rather than providing count data for each nightly collection and for each area).

Table 2.1. Data extraction form, adapted from Wilson *et al.* (2014).

Variable Name	Variable Description
PUBLICATION INFORAMTION	
Reference ID	Unique identifier assigned for a publication
Author	Authors of publication
Publication year	Year of publication
Journal	Journal where results are published
Volume	Volume of the journal
Issue	Issue number of the journal
Pages	Start page and end page of journal
Title	Title of publication or source
STUDY INFORMATION	
Study start year	Calendar year that the study was started. For studies using data from other publications (i.e. for comparing current surveillance data with historical information), the earliest year in which the data was collected have to be stated.
Study end year	Year that the study was completed
Country	Geographical location where the study was conducted
Controlled trial	Does the trial have any control arm or not?
Randomisation	Is the trial randomised? Yes/no/uncertain
Allocation	How were interventions allocated?
Blinding	Level of blinding - investigator, recipient of intervention, outcome assessor
Study design	i) randomised controlled trial, ii) randomised controlled trial, iii) pre-post studies (i.e. before and after intervention)
BACKGROUND INFORMATION	
Area	Exact location where study was conducted (e.g. name of the villages, longitude and latitude if given)
Intervention coverage	To what level are people protected by the vector intervention? (e.g. 75% have bednets)
Vector species 1	Name of the vector species under study
Vector species 2	Name of the vector species under study
Vector species 3	Name of the vector species under study
INTERVENTION DESCRIPTION	
Intervention	Type of intervention (e.g. untreated bednets, IRS)
Intervention: Number of zones that received the intervention	E.g. houses, villages
Intervention: Name of area	Name of the area that received treatment
Number of times intervention was administered	How many times were ITN re-impregnated with insecticides
Control	What is the control? (e.g. no bednets)
Number of control zones	E.g. houses, villages
Name of control area	Name of the area that received control
OUTCOME MEASUREMENT	

Mechanism of behavioural resistance?	Which mechanism of behavioural resistance was assessed? Time of biting, host preference, resting and biting tendencies
Outcome	E.g. mosquito abundance, bites/trap/hour, biting rate
Species composition	What was the species composition before and after intervention, or in control and intervention arms?
Length of baseline period	State total length
Length of post-intervention period	State total length
Time between intervention and collection of post-intervention data	Was it performed immediately?
Sampling method	Method used for sampling vectors (e.g. human landing catch, light trap, spray catch)
Detail on sampling method	Description of sampling method
How were the sampling sites chosen?	Describe how the sites were chosen for sampling. Does the paper say they were chosen randomly? How was the random selection performed? Is this described? Or was the sampling done in all houses?
Number of times outcome was measured during baseline period	
Number of times outcome was measured during post-intervention period	
Outcome information	Where in the article is the outcome described? (e.g. figure 1, etc.)
REPEAT IF MORE THAN ONE ENTOMOLOGICAL OUTCOME WAS INVESTIGATED	

Risk of bias and study quality

We assessed the risk of bias in the included studies using a risk of bias assessment form (Table 2.2). This form was developed for the purpose of this review and was adapted from Wilson *et al.* (2014). Briefly, we rated each article as having a high, low or unclear risk of bias for a number of parameters. We then gave each article an overall risk of bias assessment (low-high) based on the modal bias risk.

One parameter that is of particular interest to us is the bias introduced by species composition. Species composition and its changes during intervention, represent an important source of bias because vectors differ in their feeding behaviour, often exhibiting different peak biting activities and, indoor and outdoor biting/resting tendencies (White *et al.*, 2011). By failing to identify vectors to sibling-species level during both pre- and post- intervention periods, studies may falsely report an apparent shift towards outdoor-biting and zoophagic behaviour because vector control targets endophagic and anthropophilic vectors.

We assessed each study's quality using a study quality assessment form adapted from Wilson *et al.* (2014) (Table 2.3). Scores were allocated to a study depending on the study design, length of post-intervention surveillance, and were downgraded based on their risk of bias and the time between intervention and collection of post-intervention data. We consider the period between the intervention and collection of post-intervention data to be an important factor in determining the methodological quality of the study because we believe that selection-induced genetic changes in behaviour is unlikely to be captured if vector surveillance was conducted immediately after introducing interventions.

Table 2.2. Risk of bias assessment form, adapted from Wilson *et al.* (2014).

Criterion	Type of Bias	Risk Level Explanation		
		Low risk	**High Risk**	**Unclear**
GENERAL CONSIDERATIONS				
Sequence generation	Selection bias	Study is randomised and random nature of sequence generation is well described.	Study is not randomised.	No or unclear information reported (e.g. paper states that study is randomised but does not outline the method of sequence generation).
Blinding	Performance bias	Participants were not aware if they were allocated the intervention or, if more than one intervention was used, of which intervention they were allocated to during the study.	Participants were aware of which intervention was allocated to them.	No or unclear information reported.
Contamination	-	It is unlikely that the control group received the intervention.	Control group may have inadvertently received intervention (e.g. proximity of control and intervention areas).	No or unclear information reported.
Incorrect analysis	Over-dispersion	Analysis accounts for the variance associated with obtaining measurements from multiple areas, time-points	Incorrect analysis (e.g. data from intervention villages are pooled together despite differences in ecology or data from two yearly collections periods were combined into one).	No or unclear information reported.
OUTCOMES				
Baseline characteristics	Selection bias	Baseline entomological data available for at least 3 months prior to the introduction of intervention.	No baseline entomological data OR baseline entomological data available for short time period only.	No or unclear information reported.
Blinding (Detection)	Detection bias	Investigators or data collectors were not aware of intervention allocation OR objective measurement technique used for data collection (e.g. light traps).	Non-standardised measurement technique used for data collection (e.g. human landing catches, knock-down spray catches, manual aspiration) OR efficiency of sampling technique likely to vary between study arms (e.g. less vectors will be collected by a light	No or unclear information reported.

			trap in a bednet versus no bednet village).	
Sampling method	-	Method used for data collection is appropriate for type of behaviour being measured (e.g. human landing catches for biting behaviour) OR method used is appropriate for the comparison (e.g. HLC in intervention vs HLC in control arms).	Method used for data collection is inappropriate (e.g. known-down spray catches for biting behaviour). OR method used is inappropriate for the comparison (e.g. HLC in intervention vs light traps in control arms)	No or unclear information reported.
Human-bait rotation	Detection bias	Human-baits were rotated between indoor and outdoor avenues	Rotation was not performed	No or unclear information reported.
Incomplete outcome data	Attrition bias	No or low missing data (<20%), reason for missing data is unlikely to be related to the true outcome.	High missing data (>20%), missing data is likely to be related to the true outcome.	No or unclear information reported.
Monitoring of study site	Detection bias	Vector sampling was performed in all units	Vector sampling was not performed in all units OR sampling was not performed for both pre- and post-intervention period	No or unclear information reported.
Species composition	Sibling species confounding	Where appropriate, species were identified to sibling-species level OR the species composition is described for i) both pre-post periods or ii) both control and intervention arms.	Species were not identified to sibling-species level OR species identification was not performed for i) both pre-post periods or ii) both control and intervention arms.	No or unclear information reported.

Abbreviations: HLC= Human landing catches.

Table 2.3. Study Quality Assessment form, adapted from Wilson *et al.* (2014).

Study design (quality score)	Lower if:	Quality score
Randomised control trial (RCT) (+10)	Length of post-intervention surveillance: 0 >1 year or transmission season -0.5 >1 year or transmission season but limited repeat measures during this time -1 <1 year or transmission season	≥ 7 high quality ≥4 < 7 medium quality <4 low quality
Non- randomised control (nRCT) trial (+5)	Time between intervention and collection of post-intervention data: 0 >1 year -1 <Immediate	
Pre-post study (+4)	Risk of bias -0.5 medium -1 high	

Results

The initial search yielded 747 unique records. Of these, the full text of 39 articles was obtained. Nine studies were excluded because they lacked baseline data or had the wrong study design (Table 2.4). 10 out of 39 articles fulfilled the inclusion criteria while three others were found through screening of references. The updated search in November 2013 identified an additional 200 unique articles, of which two studies met the inclusion criteria. A total of 15 articles were therefore included in this review (Figure 2.1).

Table 2.5 summarises the study characteristics of the articles that were included in this review. Of the 15 studies that were included, 12 were rated as having low risk of bias, one with medium risk and a further two with high risk of bias (Table 2.6). Only one study was rated high quality, seven were medium quality and a further seven were low quality (Table 2.7).

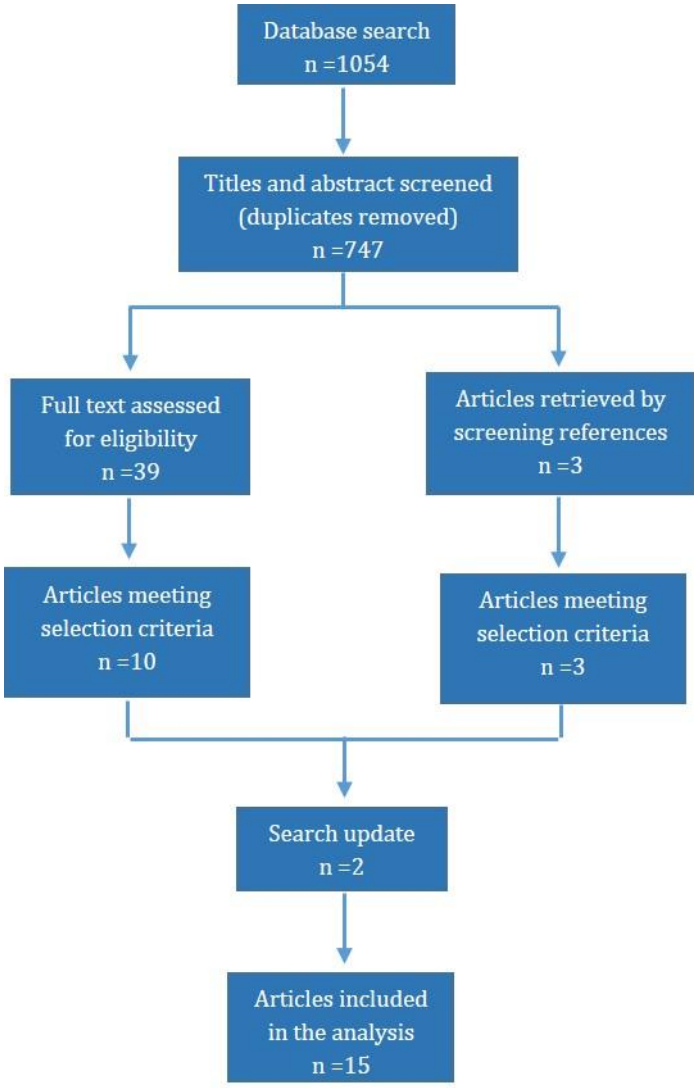

Figure 2.1. Flow chart of the study selection process.

Table 2.4. List of studies excluded and reasons for exclusion.

Author and date	Reason
Killeen *et al.* (2007)	Data from this study is already included in Russell *et al.'s* study (2011).
Reddy *et al.* (2011)	No baseline data reported.
Moiroux *et al.* (2012)	No baseline data reported: LLINs were selectively given to pregnant women and children aged <6 years during the first year of surveillance.
Fornadel *et al.* (2010)	No baseline data reported.
Quiňones *et al.* (1997)	Wrong comparison (ITN versus untreated bednets).
Lindsay *et al.* (1993b)	Wrong comparison (ITN versus untreated bednets).
Bayoh *et al.* (2010)	Programme with staged roll out of ITN. No control group.
Ilboudo-Sanogo *et al.* (2001)	A continuation of Cuzin-Ouattara *et al.*'s (1999) longitudinal study. Wrong comparison (all villages had ITCs; former control villages were given ITCs).
Lwetoijera *et al.* (2014)	No baseline data reported.

ITC=insecticide-treated curtains.

Table 2.5. Characteristics of studies included in the review.

Author (year)	Study Type	No. of areas	Country	Vector	Intervention	Length of study period	Time between intervention and collection of post-intervention data	1st outcome	2nd outcome	3rd outcome	4th outcome
Charlwood and Graves (1987)	Pre-post study	1 village with 12 houses	Papua New Guinea	*An. farauti* & *An. koliensis*	ITN- permethrin (0.4±0.2g/m^2)	46 nights	Immediate	Outdoor biting density (by species)	Indoor and outdoor resting density (by species). Performed manual aspirations inside houses and immediately outside the houses	Density of human-fed vectors (from resting collections)	-
Russell *et al.* (2011)	Pre-post study	1997 survey: 2 rural houses 2004 survey: 1 rural house; 2009 survey: 4 experimental huts	Tanzania	*An. gambiae s.l.* & *An. funestus*	ITN- permethrin & lambdacyhalothrin	12 years	7 & 12 years	Indoor and outdoor biting density (by species)	Mean biting density (mosquitoes/collector /hour) by species	-	-
Magesa *et al.* (1991)	Pre-post study	5 villages: sampled 3 houses/village	Tanzania	*An. gambiae s.l.* & *An. funestus*	ITN - permethrin (0.2g/m^2) and lambdacyhalothrin (10-30 mg/m^2); IRS-DDT	3 years	Immediate	Indoor and outdoor biting density (by species)	Mean biting time (by species)	Indoor and outdoor resting density (by species). Performed knock-down sprays and manual aspirations insides houses, and manual aspirations outside from pit traps	Density of human-fed vectors (from resting collections)
Braimah *et al.* (2005)	Non-randomised control trial	8 villages	Tanzania	*An. gambiae s.s.*, *An. arabiensis* & *An. funestus*	ITN- permethrin & lambdacyhalothrin	>1 year	6 years*	Peak biting time. Performed collections using light traps and mbita traps	-	-	-
Bøgh *et al.* (1998)	Non-randomised pre-post study	6 village pairs; 2 houses/village	Kenya	*An. gambiae* .& *An. funestus*	ITN- permethrin (0.5g/m^2)	> 1 year	<1 year	Indoor and outdoor resting density (by species). Performed knock-down sprays insides houses, and manual aspirations outside from pit traps	Density of human-fed vectors (from resting collections)	-	-
Mbogo *et al.* (1996)	Non-randomised pre-post study	72 zones: 38 intervention, 34 control	Kenya	*An. gambiae s.l.*	ITN- permethrin (0.5g/m^2)	3 years	1 year	Indoor and outdoor biting density (by species)	Indoor resting density (by species). Performed knock-down spray catches from one house/zone	Biting profile (using outdoor biting data)	Density of human-fed vectors (from resting collections)
Mathenge *et al.* (2001)	Non-randomised pre-post study	4 compounds: 2 intervention, 2 control	Kenya	*An. gambiae s.s., An. arabiensis* & *An. funestus*	ITN- permethrin (0.5g/m^2)	> 2 years	1 year	Biting profile (by species)	-	-	-
Mutuku *et al.*	Pre-post study	Pre-intervention: 30	Kenya	*An. gambiae*	ITN- permethrin	13 years	11 years	Indoor and outdoor	Density of human-fed	-	-

(2011)		villages: 20 from the north and 10 from the south coast; post-intervention: 8 villages		*s.l.*& *An. funestus*				resting density (by species). Performed knock-down spray catches inside houses and manual aspirations from clay pots outside	vectors (from resting collections)		
Cuzin-Ouattara *et al.* (1999)	Non-randomised pre-post study	4 villages: 2 intervention, 2 control	Burkina Faso	*An. gambiae s.l.*	ITC: permethrin (1g/m2); ITC was installed in every opening of the house	5 years	2 years	Indoor and outdoor biting density (by species)	-	-	-
Padonou *et al.* (2012)	Non-randomised pre-post study	5 districts: 4 intervention (IRS/LLIN), 1 control	Benin	*An. gambiae s.l.*	LLIN: bendiocarb (0.4g/m^2); IRS	< 2 years	Immediate	Indoor and outdoor biting density. Performed HLC in 4 villages/district	Indoor resting density. Performed manual aspirations and used window exit traps in 4 rooms/intervention area	-	-
Taylor (1975)	Non-randomised pre-post study	4 villages: 2 intervention, 2 control then became intervention (given intervention 1 year after the original intervention clusters)	Solomon Islands	*An. farauti, An. koliensis* & *An. punctulatus*	IRS: DDT (2g/m^2)	2 years	Immediate	Indoor and outdoor biting density	Biting profile (by species)	-	-
Bugoro *et al.* (2011)	Pre-post study	2 villages	Solomon Islands	*An. farauti*	LLIN: deltamethrin; IRS: lambdacyhalothrin	2 years	> 1 year	Indoor and outdoor biting density	mean biting density (mosquitoes/collector /hour) by species	-	-
Samarawickrema *et al.* (1992)	Pre-post study	1 village: 42 houses	Solomon Islands	*An. punctulatus & An. koliensis*	ITN- permethrin (0.5g/m^2)	> 2 years	Immediate	Indoor and outdoor biting density	Biting profile (by species)	-	-
Kabbale *et al.* (2013)	Randomised controlled trial	5 villages using ITN versus 5 villages without ITN	Uganda	*An. gambiae s.l.*& *An. funestus*	ITN	1 year	>5 years	Indoor and outdoor biting density (by species)	Biting profile	-	-
Ndiath *et al.* (2014)	Pre-post study	1 village	Senegal	*An. gambiae s.l.*	LLIN: deltamethrin	5 years	< 2 years	Indoor and outdoor biting density (by species)	Biting profile	Density of human-fed vectors (from resting collections)	

HLC= human landing catches, ITN= insecticide treated nets, LLIN= long-lasting insecticide treated nets, ITC= insecticide treated curtains. Control trials represent studies where outcomes from intervention areas were compared against areas without an intervention. Non-randomised pre-post study refers to control trial studies that performed vector surveillance during the pre- (before) and post-intervention (after) period.

Table 2.6. Study characteristics based on their risk of bias.

Author (year)	General considerations					Outcomes						
	Sequence Generation	Blinding	Contamination	Incorrect analysis	Baseline characteristics	Blinding (Detection)	Human-bait rotation	Sampling Method	Incomplete outcome data	Monitoring of study site	Species composition	**Overall rating**
Charlwood and Graves (1987)	NA	H	NA	L	L	H	NA	L	H	L	L	**L**
Russell *et al.* (2011)	NA	H	NA	L	L	H	L	L	L	H	L	**L**
Magesa *et al.* (1991)	U	H	U	H	L	H	U	L	H	H	H	**H**
Braimah *et al.* (2005)	U	U	U	H	H	L	NA	L	H	L	U	**H**
Bøgh *et al.* (1998)	U	H	L	L	L	H	NA	L	H	L	H	**M**
Mbogo *et al.* (1996)	U	H	L	H	L	H	L	L	H	L	L	**L**
Mathenge *et al.* (2001)	U	H	L	L	-	H	NA	L	L	L	L	**L**
Mutuku *et al.* (2011)	NA	H	NA	L	L	H	NA	L	L	L	L	**L**
Cuzin-Ouattara *et al.* (1999)	U	H	L	L	L	L	NA	L	L	L	U	**L**
Padonou *et al.* (2012)	U	H	U	L	L	H	L	L	L	L	H	**L**
Taylor (1975)	U	H	L	H	L	H	L	L	L	L	L	**L**
Bugoro *et al.* (2011)	NA	H	NA	L	L	H	L	L	L	H	L	**L**
Samarawickrema *et al.* (1992)	NA	H	L	L	L	H	L	L	L	L	L	**L**
Kabbale *et al.* (2013)	U	H	L	L	H	H	U	L	L	L	H	**M**
Ndiath *et al.* (2014)	NA	H	L	L	L	H	U	L	L	L	H	**L**

H= high risk of bias, M= medium risk of bias, L= low risk of bias, NA= not applicable, U= unclear.

Table 2.7. Assessment of study quality.

Study	Study design	Length of post-intervention surveillance	Time between intervention and collection of post-intervention data	Risk of bias	Overall score	Study quality
	RCT (+10), nRCT(+5), pre-post study (+4)	*-0.5 > 1 year/season but limited repeat measures < 1 year/ transmission season (-1)*	*Immediate (-1)*	*Medium (-0.5) High (-1)*		*≥ 7 high, ≥3 <7 medium, <3 low*
Charlwood and Graves (1987)	Pre-post study (+4)	46 nights (-1)	Immediate (-1)	Low (0)	2	**Low**
Russell *et al.* (2011)	Pre-post study (+4)	> 1 year (0)	7 & 12 years (0)	Low (0)	4	**Medium**
Magesa *et al.* (1991)	Pre-post study (+4)	2 years (0)	Immediate (-1)	High (-1)	2	**Low**
Braimah *et al.* (2005)	nRCT (+5)	3 nights (-1)	6 years (0)	High (-1)	3	**Low**
Bøgh *et al.* (1998)	Non-randomised pre-post study (+4)	Season (0)	<1 year (0)	Medium (-0.5)	3.5	**Medium**
Mbogo *et al.* (1996)	Non-randomised pre-post study (+4)	1 year (0)	1 year (0)	Low (0)	4	**Medium**
Mathenge *et al.* (2001)	Non-randomised pre-post study (+4)	Season (0)	1 year (0)	Low (0)	4	**Medium**
Mutuku *et al.* (2011)	Pre-post study (+4)	1 year (0)	11 years (0)	Low (0)	4	**Medium**
Cuzin-Ouattara *et al.* (1999)	Non-randomised pre-post study (+4)	2 years (0)	2 years (0)	Low (0)	4	**Medium**
Padonou *et al.* (2012)	Non-randomised pre-post study (+4)	1 year (0)	Immediate (-1)	Low (0)	3	**Medium**
Taylor (1975)	Non-randomised pre-post study (+4)	>2 years (0)	Immediate (-1)	Low (0)	3	**Medium**
Bugoro *et al.* (2011)	Pre-post study (+4)	3 months (-1)	> 1 year (0)	Low (0)	3	**Medium**
Samarawickrema *et al.* (1992)	Pre-post study (+4)	5 months (-1)	Immediate (-1)	Low (0)	2	**Low**
Kabbale *et al.* (2013)	RCT (+10)	1 year (0)	>5 years (0)	Medium (-0.5)	9.5	**High**
Ndiath *et al.* (2014)	Pre-post study (+4)	2 year (0)	<2 years (0)	Low (0)	4	**Medium**

RCT= Randomised control trial, nRCT= Non-randomised control trial.

Summary of study characteristics

Over half of the studies included in this review were control trials (i.e. control versus intervention; *n* = 8) and of these, all except two studies conducted vector surveillance before and after intervention, in both control and intervention areas. The remaining studies were pre-post studies (*n* = 7). Five studies investigated vector behaviour immediately after the intervention was introduced but only two sampled vectors again at a later date to observe any long-term behavioural changes (Taylor, 1975; Magesa *et al.*, 1991).

The length of the study period ranged from 46 nights (Charlwood and Graves, 1987) to 13 years (Mutuku *et al.*, 2011). Study locations were primarily located in Africa except for four studies that were in Papua New Guinea and the Solomon Islands (Charlwood and Graves, 1987; Taylor, 1975; Bugoro *et al.*, 2011; Samarawickrema *et al.*, 1992).

The majority of the studies (11/15) investigated behavioural resistance in the *An. gambiae* species complex. Eight studies investigated *An. funestus* while four studies examined members of the *An. punctulatus* group.

Summary of results

To improve the ease of interpretation, we have chosen to describe the studies according to the effect of vector control on the following behaviours: 1) biting pattern, 2) host preference, 3-4) propensity to rest and bite indoors or outdoors. In all cases, we lament that variations in the study designs and in the way outcomes were reported limited data interpretation, and point out that results were generally confounded by a difference in the species composition between intervention and control areas or a change in vector taxonomic composition following intervention.

Furthermore, we stress that the results of all studies except one (Cuzin-Ouattara *et al.*, 1999) should be interpreted with great caution because the studies performed vector surveillance in the presence of the intervention. In the presence of intervention, the mass killing effect of insecticides cannot be disentangled from the effects of selection because both lead to a reduction in the number of mosquitoes collected indoors. It is for this reason that we chose not to conduct a meta-analysis to summarise the evidence of biting behaviour because it will invalidate the results. Therefore, we urge readers to review the following evidence of behavioural resistance with this important caveat in mind.

Effects of vector control on time of biting

Changes in vector biting profile were measured by comparing the number of mosquitoes collected per hour in control and intervention villages, or before and after intervention. A total of 11 out of 15 articles investigated the impact of intervention on the biting time profile of vectors. Of these, nine reported a change towards early-biting behaviour following the introduction of vector control.

Five studies investigating changes in the *An. gambiae* complex were possibly confounded by the fact that the vectors were not determined to sibling-species level (Magesa *et al.*, 1991; Kabbale *et al.*, 2013; Russell *et al.*, 2011; Braimah *et al.*, 2005; Mbogo *et al.*, 1996). Magesa *et al.* (1991) reported that there was a change towards early-biting activity in one out of the two villages tested after intervention (p = 0.001). An unknown proportion of the vectors collected were identified to species level and the majority were *An. gambiae s.s.* however, it was unclear whether the test was performed during both the pre- and post-intervention periods. Given the lack of details provided, it was not possible to determine

whether the observed change towards early-biting behaviour was a consequence of changes in sibling species composition rather than behavioural resistance. It should also be noted that this study had a high risk of bias and was rated low quality.

Similarly, Kabbale *et al.* (2013) did not determine the membership of *An. gambiae s.l.* to sibling-species level. The study reported no significant difference in the biting time profile of *An. gambiae s.l.* and *An. funestus* between intervention and control areas (p = 0.05). This lack of difference suggests little evidence for behavioural resistance occurring in *An. funestus.* However, the results are more difficult to interpret for *An. gambiae s.l.* because the observed outcome may be explained by a change in species composition.

Russell *et al.* (2011) investigated the impact of ITNs on *An. gambiae s.l.* and *An. funestus* in Tanzania. When compared to the data reported by Kileen *et al.* (2006) during the pre-intervention period they found a significant shift from *An. gambiae s.s.* towards *An. arabiensis* during the post-intervention years while *An. funestus* remained the dominant vector throughout the study period. They also reported that significantly fewer *An. gambiae s.l.* and *An. funestus* were caught during late hours (21.00-05.00) compared with what was observed before the mass distribution of ITNs. Taking into consideration the change in the species composition of the *An. gambiae* complex, these results indicate that there was a change in the biting cycle of *An. funestus* but not *An. gambiae*. As vector surveillance was carried out in houses containing ITNs it was not possible to determine, however, whether this change represents a genuine evolutionary change in *An. funestus* rather than an artefact of the killing effect of the insecticide.

Mbogo *et al.* (1996) reported that the *An. gambiae* complex had a higher tendency towards early-biting activity in the intervention zones compared to the control zones. However, the proportion of *An. gambiae s.s.* differed significantly between the two zones (p = 0.0001). Most likely, the difference in the biting profile between the two zones is attributable to the difference in species composition between the two zones, and not selection.

Braimah *et al.* (2005) performed a non-randomised control trial to investigate the impact of ITN on the time of biting. The study compared the number of mosquitoes caught during times that people are unprotected by bednets (i.e. 20:00-22:00 and 5:00-07:00) between control and intervention villages. A significantly higher proportion of host-seeking events occurred during these hours in the intervention (8%) villages compared to control villages (17%) (p = 0.005). However, the authors pooled the results of *An. gambiae s.s.* and *An. funestus* together during data analysis. Although both vectors typically exhibit similar biting patterns, it is likely that there is substantial behavioural heterogeneity between these vectors. As this study failed to take into account the behavioural differences between species, we cannot conclude if there is evidence of behavioural resistance occurring in this population. We also considered this study as having a high risk of bias and of low quality, mainly due to the short period of the post-intervention surveillance.

Mathenge *et al.* (2001) reported a significant difference in the frequency of *An. gambiae s.s.* over hours of collection between intervention and control areas (D = 0.038, p = 0.001), but not in *An. funestus*. The proportion of *An. gambiae s.s.* and *An. arabiensis* did not differ significantly during the course of the night (19:00-

07:00) in either intervention and control areas (p >0.05) [*], allowing the authors to pool the results of the two vectors. Given that only a small difference in the biting cycle was observed between control and intervention villages, the authors argue that the large sample size of the study likely accounted for the finding of a statistically significant difference (10,341 and 5,853 *An. gambiae s.l.* in intervention and control villages, respectively). These results therefore give limited evidence that behavioural resistance is occurring in the *An. gambiae s.s.* population in Kenya.

Charlwood and Graves (1987) measured the outdoor-biting activity of *An. farauti* and *An. koliensis* and found that both species exhibited a shift in their peak biting time from post-midnight to pre-midnight. This study was rated low quality however, because of a lack of longer term follow-up data (post-intervention = 46 nights).

Regarding vectors in the Solomon Islands, Taylor (1975) and Bugoro *et al.* (2011) reported a significant change of *An. faurauti* towards early-biting behaviour (p = 0.01 and 0.002, respectively). Both studies were rated as having low risk of bias and of medium quality; however both studies were limited by the fact that vector surveillance was performed in the presence of an intervention and therefore, we cannot conclude whether we would observe similar results if vectors were collected when the intervention is absent (e.g. collection of vectors from houses that have no ITNs but are within the intervention village zone).

Samarawickrema *et al.* (1992) investigated the effect of ITN introduction in one village and found that there was no change in the biting pattern of *An. punctulatus* and *An. koliensis*. However, due to inadequate reporting we were unable to

[**]*An. gambiae s.s.* was consistently the dominant species in both intervention and control villages (74% and 80%, respectively).

extract the appropriate data, making the results challenging to interpret. This study was also rated low quality due to limitations in the study design.

Overall, while it may appear that *An. farauti* and *An. funestus* have undergone genetic changes towards early-biting behaviour, it is more likely that the observed changes were a direct consequence of the killing effect of insecticides. Similarly, the evidence for *An. gambiae s.l.* is weak and often explained by a change in species composition following intervention.

Effects of vector control on host preference

Host preference was measured using HBI, which represents the proportion of bloodmeals obtained from humans. All six studies that evaluated changes in host preferences found a reduction in the proportion of mosquitoes that had fed on humans (Table 2.8). With the exception of Bøgh *et al.*'s (1998), Charlwood and Graves' (1987) and Mutuku *et al.*'s study (2011) however, all six studies failed to investigate the true impact of vector control on host preference because vectors were not sampled at both indoor and outdoor locations.

Charlwood and Graves (1987) investigated changes in the host preference of the *An. punctulatus* complex. The study reported a significant reduction in the HBI of *An. farauti* following ITN introduction in Papua New Guinea (44% reduction; p <0.001) but a similar analysis could not be performed for *An. koliensis*. Prior to vector control, *An. koliensis* was highly anthropophillic with an HBI of approximately 89%. However after ITNs were introduced, too few bloodfed *An. koliensis* were caught, leading to a sample size of two specimens. Although both specimens had fed on humans, representing a 21% increase in HBI, the effect of vector control on the host preference of *An. koliensis* could not be determined due

to this small sample the limited sample size. This study was rated low quality because surveillance was conducted immediately after the intervention was introduced and the surveillance period was too short (46 nights).

Bøgh *et al.* (1998) found a change in the host preference of *An. funestus* in Kenya. Following the introduction of ITNs, they reported a 47% and 14% reduction in the HBI of *An. funestus* and *An. gambiae* in intervention villages, respectively. By comparison, there was no change in the host preference of vectors in control villages. Regarding changes in the HBI of outdoor vectors, *An. funestus* exhibited an 86% reduction in HBI in intervention villages while again, no change was observed in the control villages. Due to the low number of mosquitoes collected outdoors (n = 5), it was not possible to analyse the HBI changes of *An. gambiae's* outdoor population. The observed reduction in the HBI of *An. funestus* may be explained by a genetic change towards zoophilic behaviour, representing a true occurrence of behavioural resistance. It is more likely however, that it was a result of the killing effect of insecticides; given that vectors were collected in the presence of an intervention we were unable to distinguish between the two explanations.

Similarly, Mutuku *et al.* (2011) sampled both indoor and outdoor *An. gambiae s.l.* and *An. funestus* in Kenya. They found a greater reduction in the HBI of *An. gambiae s.l.* than *An. funestus* (15% versus 5% reduction, respectively). However, the study also found that the major vector changed towards *An. arabiensis* (85%) following the introduction of ITN. While these results suggest the possibility of behavioural resistance occurring in *An. funestus*, the reduced HBI observed in *An. gambiae s.l.* is likely explained by the change in vector species composition towards the zoophilic vector, *An. arabiensis,* rather than a true change in host

preference. Like Bøgh *et al.*'s study (1998) however, the observed reduction in the HBI of *An. funestus* may be an artefact of the killing effect of insecticides.

Mbogo *et al.* (1996) compared the HBI of *An. gambiae s.l.* caught indoors between control and intervention villages after ITN introduction. They found that the HBI was slightly lower (~7%) in the intervention group compared to the control group but this difference was not significant (p <0.05). However, only 20 mosquitoes were tested for their blood-meal origins in the intervention group as opposed to 280 mosquitoes in the control group. Therefore, there may not be sufficient power to detect a difference between the two groups. We also rated this study as having low confidence in the outcome reported because the control and intervention villages had substantially different species composition (83% versus 49% *An. gambiae* in control and intervention villages, respectively), which may have confounded the results.

Magesa *et al.* (1991) investigated the host preference of *Anopheles* by collecting outdoor-resting mosquitoes. They observed modest to great reductions in the HBI of all *Anopheles* vectors; however, we identified several limitations. Firstly, complete data for both pre- and post-intervention periods was available for one out of five villages only. Secondly, the data for *An. rivulorum* and *An. funestus* were pooled because the authors were, at times, unable to distinguish between the two sibling species. *An. funestus* is widely cited in the literature as being highly anthorphophilic while *An. rivulorum's* feeding behaviour appears opportunistic, feeding on the host species that is most abundant (Kawada *et al.*, 2012). Finally, the study lacked sufficient power (n ≤10 in *An. rivulorum/An.* funestus) and data on the indoor-resting population. Taken together, this study was limited in its ability to determine changes in host preference.

Ndiath *et al.* (2014) investigated the impact of LLINs on the host preference of *An. gambiae* complex, captured indoors by knock-down spray catches. All three vectors exhibited a reduction in HBI with the greatest reduction observed in *An. arabiensis* (65% reduction) while *An. gambiae s.s.* exhibited low to modest reductions (4% and 24% reduction in *An. coluzzii* and *An. gambiae,* respectively). However these outcomes are potentially under-representative of the overall vector behaviour as no data regarding outdoor-biting vectors was collected. Furthermore, surveillance was conducted in the presence of LLINs, limiting the study's ability to distinguish behavioural resistance from other explanations of behavioural changes.

All except three studies (Charlwood and Graves, 1987, Mutuku *et al.*, 2011, Bøgh *et al.*, 1998) failed to investigate the true impact of vector control on host preference because data was available either for indoor or outdoor vectors only. As mentioned previously, sampling of both indoor and outdoor populations is an essential prerequisite to adequately measure any changes in vector behaviour. As Charlwood and Graves's study (1987) was rated low quality, only two out of six studies† had possible evidence of behavioural resistance occurring through increased zoophilic behaviour, although their findings are more likely explained simply as an effect of the insecticides on reducing vector density.

†It should be noted that while we mention Bøgh *et al.*'s and Mutuku *et al.*'s study (2011) as potentially having evidence of behavioural resistance in the form of altered host preference, this applies to *An. funestus* only.

Table 2.8. Effects of ITN and LLINs on *Anopheles'* host preference

Study	Vector species	Surveillance method	Intervention	Measure	Control (HBI)		Intervention (HBI)		% reduction in HBI
					Pre	Post	Pre	Post	
Charlwood and Graves (1987)	*An. farauti*	Aspiration	ITN	No. of human-fed vectors (Indoors and outdoors)	-	-	0.72	0.4	44.35
	An. koliensis				-	-	0.83	1	-20.69[≠]
Magesa *et al.* (1991)	*An. gambiae s.l.*	Pit trap collections	ITN	No. of human-fed vectors (Mlingano only; outdoors)*	-	-	0.89	0.85	4.167
	An. funestus/An. rivulorum[¥]				-	-	0.43	0.13	70.83
Bøgh *et al.* (1998)	*An. gambiae s.s.*	PSC & pit trap collections	ITN	No. of human-fed vectors (indoors and outdoors)	0.98	1	0.99	0.86	14.18
	An. funestus				0.99	0.99	0.99	0.34	65.17
Mbogo *et al.* (1996)	*An. gambiae s.l.*	PSC	ITN	No. of human-fed vectors (Indoors)	-	0.86	-	0.8	7.44
Mutuku *et al.* (2011)	*An. gambiae s.l.*	PSC, claypot sampling & HLC	ITN	No. of human-fed vectors (indoors and outdoors)		-	0.98	0.84	14.65
	An. funestus					-	0.99	0.94	5.47
Ndiath *et al.* (2014)	*An. coluzzi*	PSC	LLINs	No. of human-fed vectors (Indoors)	-	-	0.81	0.62	24.1
	An. gambiae				-	-	0.88	0.85	3.51
	An. arabiensis				-	-	0.74	0.26	64.52

*Data represents data from Mlingano village only because inadequate data was provided for the remaining village. [≠]Sample size was too low (n = 2) to allow accurate measurement of percentage reduction. [¥]Results for *An. funestus* and *An. rivluorum* were pooled together because the authors couldn't distinguish between the two vectors. PSC= pyrethrum spray catches, HLC= human landing catches, ITN= insecticide-treated nets, LLINs= long-lasting insecticide bednets, HBI= Human blood index.

Effects of vector control on resting behaviour

All of the six studies that investigated resting behaviour found a reduction in the number of indoor-resting vectors following intervention (Table 2.9). With the exception of Bøgh *et al.*'s study (1998) however, all six studies failed to investigate the true impact of vector control on resting behaviour because, in most cases, only the indoor–resting vectors were sampled.

Magesa *et al.* (1991) investigated the impact of ITNs on the resting behaviour of *An. gambiae* complex. They reported a reduction in the number of indoor- and outdoor-resting mosquitoes that correlated with a reduction in the overall population density following intervention. The results however, were not described well enough to allow data extraction, making the results challenging to interpret. This study was also considered as having a high risk of bias and of low quality due to limitations in their study design.

Mutuku *et al.* (2011) reported a 75% and 92% reduction in the indoor-resting population of *An. gambiae s.l.* and *An. funestus,* respectively in Kenya. However, their results were likely confounded by the fact that the major vector changed from *An. gambiae* to *An. arabiensis* following ITN introduction. Similarly, reductions in the indoor-resting population of *An. gambiae s.l.* from Benin were reported by Padonou *et al.* (2012) (99% and 14% reduction in IRS and LIIN villages, respectively), but they did not determine the *An. gambiae* complex to sibling-species level. Both studies also lacked data on the outdoor-resting vectors during the pre-intervention phase.

Similarly, Mbogo *et al.* (1996) reported a significant difference in the number of indoor-resting *An. gambiae s.l.* and *An. funestus* in intervention zones compared

to control zones (88.5% and 95.6% reduction for *An.gambiae s.l.* and *An.funestus*, respectively in intervention zones; p <0.0001). As stated previously however, the proportion of *An. gambiae s.s.* was significantly different between the two zones (p = 0.0001). They also did not perform outdoor collections and therefore, have limited evidence that behavioural resistance has occurred.

Charlwood and Graves (1987) investigated the impact of ITNs in *An. farauti* and *An. koliensis* in Papua New Guinea, and found a 78% and 85.7% reduction in the overall number of resting vectors, respectively during the post-intervention period. Although both indoor and outdoor collections were performed, these results were pooled during data analysis. Therefore, it was not possible to determine if there were changes in the indoor- and outdoor-resting tendencies of either vector.

Only Bøgh *et al.* (1998) adequately measured changes in resting behaviour in response to vector control. They found that ITNs had a greater impact on the resting behaviour of *An. gambiae s.s.* than *An. funestus.* They reported a reduction in both indoor- (96% reduction) and outdoor-resting (77% reduction) densities of *An. gambiae s.s.* while only the indoor-resting population (94% reduction) of *An. funestus* appears to be affected. Interestingly, *An. funestus* exhibited a 9% increase in outdoor-resting densities. These results suggest that *An. funestus* may be exhibiting behavioural resistance but not *An. gambiae s.s..* However, it is more likely that the proportional decline in indoor-resting vectors was a direct result of the mass killing effect of the insecticides.

In summary, all except one study (Bøgh *et al.*, 1998) failed to investigate the true impact of vector control on resting behaviour. To obtain a more reliable estimate of the behavioural response of vectors to control intervention, it is important that both

the indoor- and outdoor-resting population are sampled. The results of all studies were additionally confounded by the fact that vector surveillance was conducted in the presence of the intervention and therefore, it was not possible to determine whether any changes in vector density were a result of the insecticide causing increased mortality and/or higher exit rates.

Table 2.9. Effects of ITN, LLINs and IRS on *Anopheles* resting behaviour.

Study	Vector species	Surveillance method	Intervention	Measure	Control (Mean)		Intervention (Mean)		% reduction in vector density
					Pre	Post	Pre	Post	
Charlwood and Graves (1987)	*An. farauti*	Aspirator	ITNs	No. of indoor- and outdoor-resting (mean number per collector)	-	-	6.45	1.41	78.14
	An. koliensis				-	-	1.4	0.2	85.71
*Magesa *et al.* (1991)	*An. gambiae s.l.*	PSC & pit trap collections	ITNs	Mean number of indoor- and outdoor-resting mosquitoes collected per pit trap/house	-	-	-	-	-
Bøgh *et al.* (1998)	*An. gambiae*	PSC	ITNs	No. of indoor-resting female mosquitoes per house	307.17	256.5	215.17	6.83	96.20
		Pit trap collections		No. of outdoor-resting female mosquitoes per house/pit trap	3.67	1.5	3.5	0.33	76.72
	An. funestus	PSC		No. of indoor-resting female mosquitoes per house	252.17	766.33	211	37.33	94.18
		Pit trap collections		No. of outdoor-resting female mosquitoes per house/pit trap	3.833	26.5	5.5	41.5	-9.15
Mbogo *et al.* (1996)	*An. gambiae s.l.*	PSC	ITNs	No. of indoor-resting mosquitoes	-	1.79	-	0.21	88.53
	An. funestus				-	0.23	-	0.01	95.64
Mutuku *et al.* (2011)	*An. gambiae s.l.*	PSC	ITNs	Mean number of indoor-resting mosquitoes per house	-	-	1.83	0.46	74.86
	An. funestus				-	-	4.27	0.36	91.57
Padonou *et al.* (2012)	*An. gambiae s.l.*	PSC	IRS	No. of indoor-resting females	49.33	51.67	124.6	0.8	99.39
			LLINs		53.33	52.08	53.33	44.67	14.24

*Data could not be extracted successfully due to ambiguous way of reporting. PSC= pyrethrum spray catches, ITN= insecticide-treated nets, IRS= indoor residual spraying, LLINs= long-lasting insecticide bednets.

Effects of vector control on biting behaviour

A total of eleven studies investigated changes in biting behaviour and of these, all except one study (Charlwood and Graves, 1987) provided adequate information to allow us to calculate the difference in proportion of endophagy (*D*) between pre- and post-intervention periods (Table 2.10).

Two studies provided evidence that behavioural resistance is occurring in *An. farauti* in the Solomon Islands (Taylor, 1975, Bugoro *et al.*, 2011). Taylor (1975) investigated the biting behaviour of *An. farauti* in San Cristobel Island and found a change towards exophagy (*D* = -0.54; 95% CI: -0.55, -0.52) following DDT spraying. In this study, vectors were surveyed immediately after the introduction of vector control for a period of approximately two years. Similarly, Bugoro *et al.* (2011) found a small increase in the exophagic behaviour of *An. farauti* (*D* = -0.08; 95% CI: -0.11, -0.04) following the distribution of LLINs and IRS in the Temotu Province. Surveillance was conducted a year after LLIN and IRS were introduced for three months. Given that a change towards early-night biting activity was also observed in both studies[*], there is a higher likelihood that the observed changes in behaviour represent a true occurrence of behavioural resistance. We cannot exclude however, the possibility that these results were a mere consequence of the direct killing effect of the insecticides.

Cuzin-Ouattara *et al.* (1991) measured the effect of insecticide-treated curtains (ITCs) on the biting behaviour of *An. gambiae s.l.* in Burkina Faso. The study was a randomised controlled trial involving a total of 158 villages but long-term surveillance of indoor- and outdoor-biting vectors (~ 3 years) was performed in

[*]It is thought that biting behaviour is related to vector biting profile with exophagic vectors expressing a greater tendency towards early-biting behaviour because outdoor hosts are only accessible during early hours of the night.

four villages only (two intervention and two control villages). For intervention villages, indoor-biting vectors were collected in houses with and without ITCs. Using the data collected from unprotected houses, we calculated a *D* estimate of -0.04 (95% CI: -0.06. -0.02), indicating a change towards outdoor-biting behaviour following intervention (Table 2.9). In the presence of insecticides, it is not possible to determine whether the reduction of indoor-biting vectors is due to the effect of insecticides or selection. Because indoor collections were performed in houses where the intervention is not present, these results provide the clearest evidence of behavioural resistance out of the studies included in this review. However, the vectors were not determined to the sibling species level. Therefore, we were unable to reject the possibility that the results were confounded by a difference in the species composition between intervention and control villages.

Charlwood and Graves (1987) reported a decrease in the average number of outdoor-biting mosquitoes collected for both *An. farauti* and *An. koliensis* (30% and 54% reduction in vector density, respectively). Lack of data on their indoor-biting counterparts precluded our calculation of *D*; however it is likely that these reductions are caused by an overall decline in the vector population, suggesting that vector behaviour remained unchanged following ITN introduction. This study was rated low quality because it has a limited ability to detect any long-term behavioural changes (post-intervention follow-up = 46 nights).

There is no hard evidence that *An. punctulatus* exhibited a shift towards outdoor-biting behaviour in Samarawickrema *et al.*'s study (1992) (*D* = -0.13; 95% CI: -0.13, 0.07). During the pre-intervention phase, a cyclone caused damage to the study area and reduced the mosquito population, contributing to the wide fluctuation in the monthly vector density reported by the authors. As this study was

confounded by the mass killing effect caused by the cyclone, it is difficult to determine whether a behavioural change has occurred.

Mbogo *et al*.'s (1996) study was likely confounded by the fact that the vector taxonomic composition differed between the intervention and control areas. The study found significantly higher rates of exophagy in intervention areas compared to control areas (p <0.0001), after adjusting for the difference in the pre-intervention biting rates between the two groups. However, post-intervention surveillance showed that control areas consisted mainly of *An. gambiae* (83%) while *An. gambiae* (49%) *and An. merus* (44%) were the dominant vectors in the intervention areas. Without pre-intervention measures of species composition, it is difficult to determine if these are genuine observations of behavioural resistance occurring rather than a mere effect of a change in vector taxonomic composition.

Ndiath *et al.* (2014) reported a significant increase in exophagy in both *An. arabiensis* and *An. gambiae* after the introduction of LLINs (p <0.001). However we were unable to calculate D at the sibling species level due to inadequate reporting. At best, the results of our analysis showed weak evidence in support of Ndiath ***et al.'s*** (2014) findings (D = -0.03; 95% CI: -0.13, 0.07). Once again however, vector surveillance was conducted in the presence of the LLINs and thus, it is more likely that the relative increase in outdoor-biting vectors were a consequence of the direct killing effect of the insecticides rather than an effect of selection.

There is some evidence of a change towards exophagic behaviour in *An. funestus* but not *An. gambiae s.l.* in two studies. In Russell *et al.*'s (2011) study, both *An. funestus* and *An. gambiae s.l.* exhibited an increase in exophagic behaviour;

however as the taxonomic composition of the *An. gambiae* complex shifted towards *An. arabiensis* following vector control (p = 0.0001) we consider the results of low value with regards to the *An. gambiae* complex. By comparison, as *An. funestus* remained the dominant vector throughout the study period, there is evidence that *An. funestus* is exhibiting behavioural resistance. In Kabbale *et al.*'s (2013) study, *An. funestus* exhibited a slight increase in exophagy (D = -0.02; 95% CI: -0.14, 0.09) but no change was observed in the *An. gambiae* complex (D = 0.02; 95% CI: -0.02, 0.07). This study was also rated as having a medium risk of bias because the membership of the *An. gambiae* complex was not determined to sibling-species level and there was no data available before ITNs were introduced. While the observed proportional decline in indoor-biting *An. funestus* may be indicative of behavioural resistance in both studies, it may also be explained by the mass killing effect of the insecticide as both studies performed vector surveillance while the intervention was present.

Magesa *et al.* (1991) investigated the effect of ITNs on the biting behaviour of *An. gambiae s.l.* in Tanzania. Although, a significantly higher number of outdoor-biting was reported in one out of two intervention villages (p <0.05), the overall effect showed that there was no change towards exophagic behaviour (D = 0.04; 95% CI: 0.01, 0.06). As discussed previously, this study was rated low quality, had a high risk of bias and was potentially confounded by the fact that the taxonomic composition of the vectors may have changed. Taken together, there is limited evidence that behavioural resistance has been observed in this study.

Both IRS and LLIN villages in Padonou *et al.*'s (2012) study exhibited an increase in the exophagic behaviour of *An. gambiae s.l.* in Benin (D = -0.24 and -0.14, respectively). However, we were unable to calculate the confidence interval of the

D estimate as they did not provide the raw number of vectors collected. Furthermore, they did not determine the sibling species identity of the *An. gambiae* complex and therefore, it has limited evidence that behavioural resistance has occurred.

Overall, while the results indicate that *An. farauti* and *An. funestus* might have undergone a genetic change towards exophagic behaviour, the evidence for *An. gambiae s.l.* is weak due to limitations associated with the study design (i.e. inadequate reporting, lack of pre-intervention data) and because the results were confounded by a change in species composition following intervention.

Table 2.10. Effect of vector control on *Anopheles* biting behaviour.

Study	Vector species	Intervention	Comparison	*D*	95% CI (Lower, Upper)
Russell *et al.* (2011)	*An. gambiae s.l.*	ITN	Intervention group pre- vs post-intervention	-0.01	-0.05, 0.03
	An. funestus			-0.31	0.38, -0.24
Magesa *et al.* (1991)	*An. gambiae s.l.*	ITN		0.04	0.01, 0.06
Taylor *et al.* (1975)	*An. farauti*	IRS		-0.54	-0.55, -0.52
Bugoro *et al.* (2011)	*An. farauti*	LLIN		-0.08	-0.11, -0.04
Samarawickrema *et al.* (1992)	*An. punctulatus*			-0.03	-0.13, 0.07
Ndiath *et al.* (2014)	*An. gambiae s.l.*	LLIN		-0.03	-0.07. 0.01
Padonou *et al.* (2012)*	*An. gambiae s.l.*	IRS		-0.24	-
		LLIN		-0.14	-
Kababale *et al.* (2013)	*An. gambiae s.l.*	LLIN	Intervention vs control group post-intervention	0.02	-0.02, 0.07
	An. funestus			-0.02	-0.14, 0.09
Mbogo *et al.* (1991)	*An. gambiae s.l.*	ITN		-0.1	-0.12, -0.08
Cuzin-Ouattara *et al.* (1991)†	*An. gambiae s.l.*	ITC		-0.04	-0.06. -0.02

Negative values represent a change towards outdoor-biting behaviour. Vectors were sampled by human landing catches except for Cuzin-Ouattara *et al.*'s (1991) study where light traps positioned over a volunteer sleeping under a bednet were used to collect biting mosquitoes. *D*= difference in endophagy, ITN= insecticide-treated nets, IRS= indoor residual spraying, LLIN= long-lasting insecticide bednets, ITC=insecticide-treated curtains. *Confidence intervals could not be calculated because the raw number of vectors was not reported. †Post-intervention vector surveillance was performed in houses without ITCs, thereby representing a more accurate measure of vector population density. Charlwood and Graves (1987) were excluded from the analysis as they performed outdoor-biting collections only.

Discussion

This review aimed to investigate whether *Anopheles* vectors display evidence of behavioural resistance in response to the use of vector control interventions. We reviewed a total of 15 studies investigating one or more of the following behavioural resistance mechanisms: altered biting time activity, host preference, feeding and resting habits. There was substantial heterogeneity between all 15 studies, displaying variations in vector species, ecological settings, study designs and outcome measures. These variations have made it difficult to generalise the findings of the studies however, no conclusion can be drawn regarding behavioural resistance owing to methodological weaknesses.

Most studies investigating the impact of vector control interventions are understandably primarily designed to see if the intervention reduces transmission, not whether it results in changes in vector behaviour, which is usually a secondary consideration. This will inevitably result in some measurement bias and residual confounding. Studies investigating changes in the *An. gambiae* complex were confounded (or potentially confounded) by a difference in the species composition between intervention and control areas or a change in vector taxonomic composition following intervention. A shift towards zoophagic, exophagic vectors is a predictable effect of vector control as it selectively targets endophagic and anthropophilic vectors. However it is important to distinguish this effect from genuine behavioural resistance because it has long-term effects that are difficult to reverse and may potentially lead to the spread of resistance genes. Another major problem that we encountered was the fact that all studies, except one (Cuzin-Ouattara *et al.*, 1999), performed mosquito collections in the presence of the intervention or while the intervention was still ongoing. In the presence of

intervention, it is not possible to determine whether the reduction in the number of mosquitoes collected indoors is due to behavioural resistance occurring rather than the mass killing effect of insecticides. At best, this review was only able to identify studies that might have evidence of behavioural resistance.

Most studies summarised changes in vector behaviour by combining the data from all villages. This would almost certainly underestimate the variation in vector behaviour between villages and lead to an over-dispersion of the results. This is best exemplified by Mbogo *et al.*'s (1996) study, where the results were summarised from 38 intervention and 34 control villages. In this case, lack of information prevented us from determining the extent of variability and level of uncertainty in the results they provided.

It was particularly difficult to make strong conclusions regarding changes in resting behaviour and host preference because we did not contact the authors to request further information. Doing so may have allowed us to calculate the confidence intervals for the percentage reduction of vector density and HBI, as confidence intervals and standard deviations were omitted in the articles. The evidence for resting behaviour and host preference is therefore weak.

With regard to resting behaviour and host preference, vector surveillance was focused on sampling indoor mosquitoes with few studies performing outdoor collections. For resting behaviour, only one out of six studies had available data on outdoor-resting behaviour, while only half of the studies measured the host preference of outdoor mosquitoes. As we have mentioned frequently during this review, characterisation of vector behaviour from both indoors and outdoors is crucial for the proper evaluation of behavioural resistance in response to vector

control. Therefore, any reported behavioural changes in almost all of these studies were accorded with less weight because they were liable to sampling bias.

Though it is feared that behavioural shifts will reduce control efficacy, it is also possible that they will not impact malaria control. Indeed, reductions in malaria prevalence have been reported in bed net areas despite the predominant vector species displaying exophagic behaviour (Charlwood and Graves, 1987). Increased control efforts may cause reduced vectorial capacity because of a shift towards early reproduction at the expense of longevity. Shifts to zoophagic behaviour may also prove beneficial for the prospect of elimination if livestock feeding persists long enough to drive human-human/livestock-human transmission to critically low levels (Ferguson *et al.*, 2012).

Our review has several limitations that should be noted. We did not contact the authors to request further information. Doing so may have allowed us to calculate the confidence intervals for the percentage reduction of vector density and HBI, as confidence intervals and standard deviations were omitted in the articles. Without knowing the uncertainty around percentage reductions it was not possible to make any conclusions regarding changes in resting behaviour and host preference. Furthermore, we did not do a full search of the grey literature (i.e. unpublished trials, trials published in non-English language journals) which may mean that publication and language bias was introduced, resulting in over-reporting of studies demonstrating behavioural changes. The majority of the studies were at low risk of bias but were of mixed quality. This was mostly attributable to the fact that the study designs yielded low scores (e.g. Pre-post study = 4 points). We also relaxed the scoring threshold somewhat in terms of the length of the post-intervention surveillance so that studies with surveillance data for one year or one

transmission season were not downgraded. However, evolutionary forces work on a slow time-scale and studies would likely need to be of a longer duration before behavioural resistance can be observed.

Owing to methodological weaknesses in the study design there is limited evidence that behavioural resistance is occurring in *Anopheles* vectors. This review highlights the need to improve reporting and establish guidelines to improve the assessment of intervention impact on vector behaviour. Distinguishing between temporary, adaptive traits and true behavioural resistance is essential because the spread of resistance genes conferring undesirable behavioural changes, such as outdoor-biting behaviour, threatens the sustainability of malaria control since current interventions are poorly equipped to handle such changes. To demonstrate evidence of behavioural resistance, vector behaviour must be surveyed at both indoor and outdoor locations. Entomological data prior to the introduction of vector control must be available for a minimum period of one year and following vector control, for a period spanning several years to detect any long-term behavioural changes. It is also important to exclude the direct killing effect of insecticides as an explanation for any proportional changes in indoor-associated behaviour by conducting vector surveillance in the absence of an intervention, either through halting the intervention while vector surveillance is conducted or by performing collections in houses that do not have the intervention. To prevent the results from being confounded by changes in the taxonomic composition, proper identification of vectors to species and sibling species level during both the pre- and post-intervention period must be carried out. Finally, there should be some considerations of the study design (i.e. RCT) and the sample size required to minimise risk of bias and achieve biologically

meaningful results. Alternatively, molecular and genetic tools may be exploited to investigate the genetic differentiation between the different types of behaviour (e.g. endophagy vs exophagy).

3. A survey of host-seeking behaviour of the *Anopheles gambiae* complex in the Gambia

Abstract

The recent and sustained decline of malaria in The Gambia, attributed in part to increasing insecticide-treated bednet (ITNs) coverage, has prompted a scale-up of vector interventions with the goal of local elimination. This study aimed to investigate the host-seeking behaviour of the major vectors, *An. gambiae s.s.* and *An. arabiensis* in Wali Kunda, Gambia. We surveyed the hourly biting rates of these vectors during the dry season in 2012 and compared it to rates obtained in 1992 during the wet season. Both vectors displayed a tendency to bite indoors and during early mornings (peak biting time 4:00-5:00 am and, 1:00-3:00 and 4:00-6:00 am, respectively). There were no significant difference in the proportion of mosquitoes biting indoors and outdoors, both at the within and between species level (p >0.05). Compared to the nearby village of Saruja in 1992, the *An. gambiae* complex displayed an earlier mean biting and lower endophagy rate (p >0.01), suggesting that *An. gambiae* complex is exhibiting a behavioural change in Wali Kunda. However, due to differences in the species composition between Saruja and Wali Kunda and issues associated with comparing behaviour from different locations, our findings provide limited evidence of behavioural resistance.

Introduction

Anopheles gambiae sensu stricto and *An. arabiensis* are major vectors of malaria in The Gambia. A third vector, *An. melas,* is restricted to the western parts of the country where it makes a minor contribution towards disease burden (Bryan *et al.*, 1987). Vector control against *Anopheles* vectors remains an indispensable tool in the current effort to reduce malaria transmission. Indeed, anti-malaria campaigns in the Gambia have prioritised the use of insecticide treated nets (ITNs) and indoor residual spraying (IRS). In 2006, it was estimated that bednet coverage in The Gambia was approximately 63% (Ceesay *et al.*, 2008). However, effective use of these interventions requires adequate and continued surveillance of the distribution and abundance of local vectors to monitor changes in vector populations and, consequently, malaria transmission dynamics.

Studies of the distribution of the *An. gambiae* complex in The Gambia are limited. Bryan *et al.* (1982) surveyed the *An. gambiae* complex across The Gambia, representing one of the earliest surveys describing its distribution and behaviour. Since then, relatively few studies have described the bionomics of this complex (Caputo *et al.*, 2008, Quiňones *et al.*, 1997). Although the distribution of the three *An. gambiae s.l.* species (*An. gambiae s.s., An. arabiensis* and *An. melas*) have remained fairly constant since Bryan *et al.*'s (1982) first description, there is still a paucity of information regarding the abundance and composition of these vectors at a local scale. Information regarding vector biting behaviour has also been scarce with major studies conducted only during the start of vector control trials (Lindsay *et al.*, 1993a). The importance of such information has been highlighted over recent years in light of potential evolved changes in vector behaviour (also referred to as "behavioural resistance") and species composition in response to

vector control (Reddy *et al.*, 2011, Gillies and Smith, 1960, Bayoh *et al.*, 2010). Monitoring such a change is important for understanding the epidemiology of malaria transmission and for appropriate control interventions.

This chapter presents the results of adult *An. gambiae s.l.* collections in the small village of Wali Kunda, Gambia. We describe the species composition, biting cycle and proportion of indoor and outdoor biting of the members of the complex found in the area. Next, we compared how these results differ from data collected in 1992 to look for initial indications of behavioural changes. It is hoped that this information will contribute towards the surveillance of vectors in response to vector control and the development of more effective control measures.

Methods

Ethics statement

The study protocol was reviewed and approved by The Gambia Government/Medical Research Council (MRC) Joint Ethics Committee. Verbal and written consent was obtained from volunteer mosquito collectors prior to the start of the study (SCC reference: 1300).

Study site

Collections were carried out at Wali Kunda in the rural, central river region of the Gambia (13^{o} 34'N, 14^{o} 55'W) (Figure 3.1). The area is composed of a small fishing village with approximately 50 inhabitants and a small field station belonging to the MRC. Hereafter I use the name Wali Kunda to include the MRC field site and the village. The area is described as flat Sudan savannah; it is approximately 180 km from the coast and is located by the south bank of the River Gambia. *Anopheles* breeding sites include rice fields and swamplands located to the south of the village. Previous entomological surveys showed that *An. gambiae s.s.* and *An. arabiensis* are common within the area, while *An. melas* is present elsewhere in The Gambia. Furthermore, the incipient species *An. coluzzii* (formerly known as the M molecular form) is abundant within the study area while *An. gambiae* (formerly known as the S molecular form) are found elsewhere (Jawara *et al.*, 2009). For simplicity, *An. coluzzii* and *An. gambiae* will be collectively referred to as *"An. gambiae s.s.*".

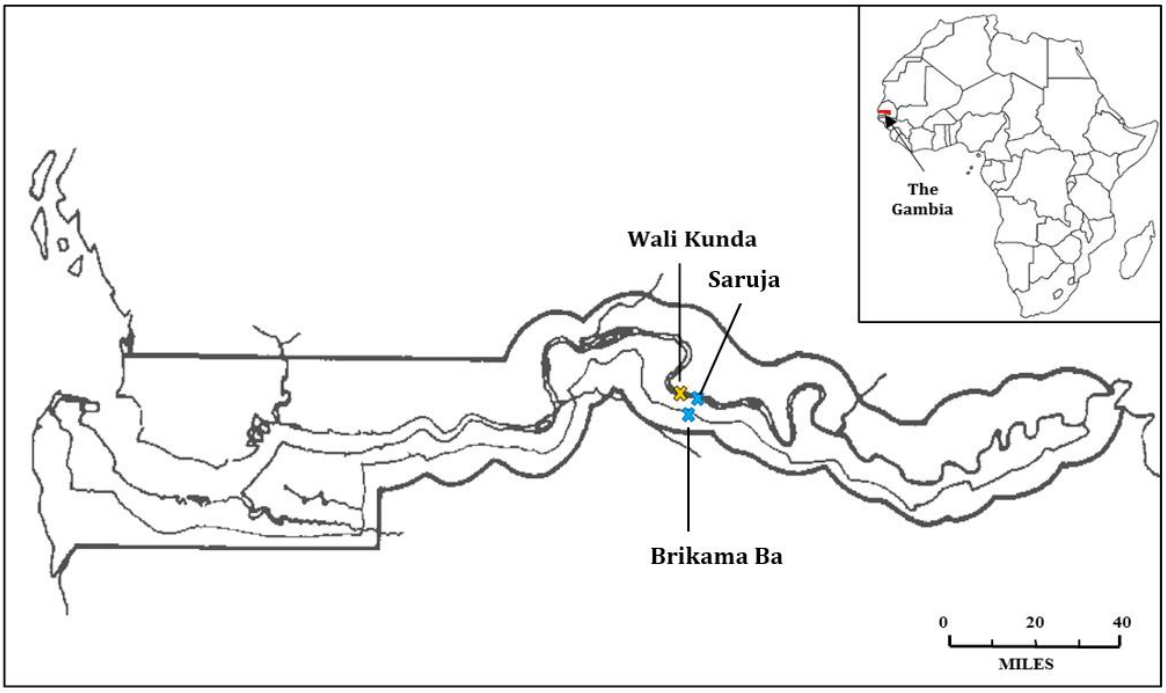

Figure 3.1. A map of The Gambia. Wali Kunda is a small field site owned by the MRC where various entomological studies take place. Neighbour villages, Saruja and Brikama Ba are also indicated. Adapted from Caputo *et al.* (2008).

HLC Volunteers

A total of 20 men aged between 16 and 60 years were recruited to act both as mosquito baits and collectors, referred to as human landing collections (HLC). Volunteers were from Wali Kunda or other nearby villages such as Brikama Ba and Wellingara. Collectors were divided into pairs and each pair was allocated a position (i.e. outdoor/ indoor position). Different pairs were used every 2 hours. Pairs also changed positions and time periods each night to reduce fatigue and any bias from differences in individual odours and collecting abilities. Each volunteer gave their consent to participate in the trial and was paid a nightly rate of £3.50 (200 dalasis). Snacks and refreshments were also provided each night.

Indoor location

Indoor collections were performed inside an experimental hut in Wali Kunda. The hut has a simple square design (2m x 2m) with plywood ceiling, thatched roof, walls made of mud and cement, open eaves, a window in each wall and a door

facing north-east towards the River Gambia (Figure 3.2). The hut was situated in an area of open grassland with scattered trees. Host-seeking mosquitoes can enter via the windows and eaves. The door was held slightly ajar (approximately 20mm gap) to simulate village conditions.

Figure 3.2. Experimental hut in Wali Kunda. The verandas were left open in order to allow mosquitoes to enter through the eaves and windows.

Outdoor location

Outdoor collections took place in three locations: 1) south of the hut, by the gate of the MRC field site and just before the surrounding rice field, 2) north of the hut and further into the site's compound and 3) north-west of the huts and inside Wali Kunda village. All outdoor locations were at least 60 metres away from the indoor collectors (Figure 3.3).

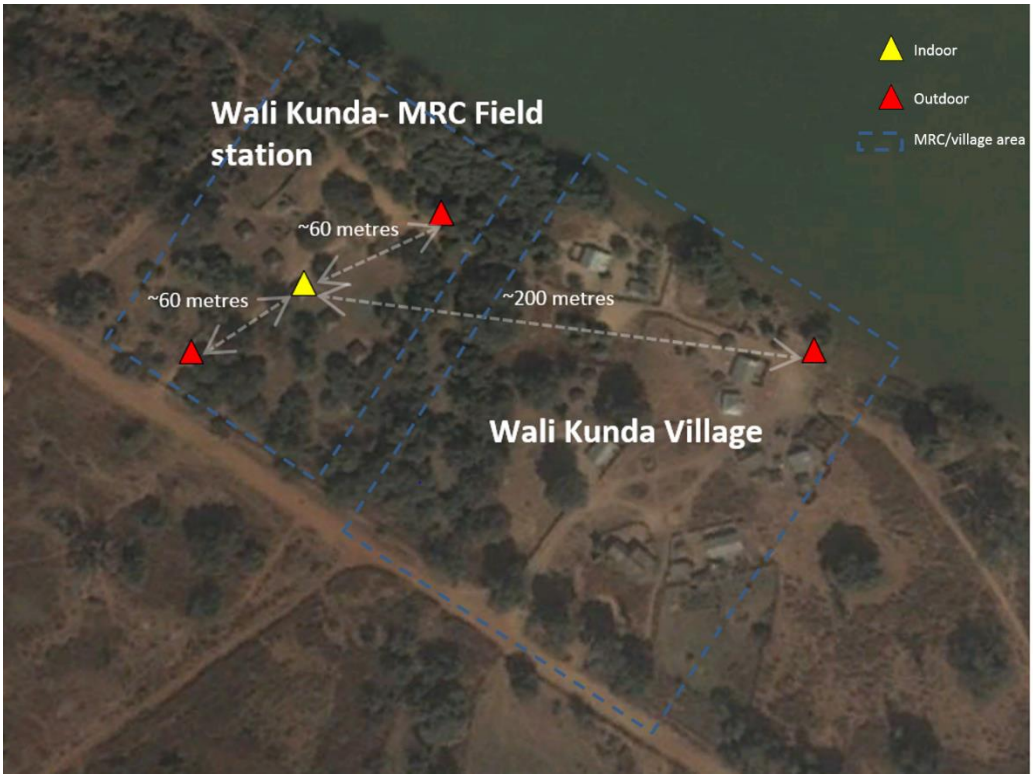

Figure 3.3. Schematic of mosquito collection sites in Wali Kunda. Adapted from Google Earth (Google Inc., 2013).

Mosquito collection

Mosquitoes were collected over 19 days during the early dry season (late October-early November) in 2012. The collection period was divided into four phases with the end of each phase marked by a one-day break from collections.

The first phase served as the pilot study and lasted for four consecutive nights. During this phase, one indoor and two outdoor positions within the MRC site were chosen to maximise outdoor collections. Four pairs of collectors were used and, each team member helped the other during the collection by shining a torch while the other collected landing mosquitoes in test tubes sealed with cotton wool. Indoor collectors were positioned inside an experimental hut with windows and eaves as ports of entries for host-seeking mosquitoes. Outdoor collectors were positioned approximately 60-70 metres from the indoor location and at least 100 metres away from each other. Indoor collections were performed from 19:00-6:00

hours while partial night catches (19:00-22:00 hours) were performed outdoors; the time of night people would normally be exposed to mosquitoes.

Following poor outdoor collections during the first phase, the collection period was extended to 9:00-7:00 hours, for both indoor and outdoor collections for the remainder of the experiment. Additional collectors were therefore recruited. Caught mosquitoes were stored in individual tubes and collections were retrieved from each team every hour.

Species identification

An. gambiae s.l. mosquitoes were identified based on morphology (Gillet, 1972), using the following three diagnostic characters: 1) irregularly speckled legs, 2) three pale rings including one at the tip of the palps and 3) third dark area on vein 1 with a pale interruption of the wings. *An. gambiae s.l.* mosquitoes were stored in RNAlater or dried with silica gel at -20^{o}C prior to transport to the laboratory. *Culex* and *Mansonia* mosquitoes were not identified to species but were counted at the genus level.

Preserved *An. gambiae s.l.* were thawed and refrozen to allow 1-3 legs to be removed from each individual. The DNA from the legs were used as template DNA for species identification following the PCR procedure described by Scott *et al.* (1993). Legs from each mosquito were placed in 10µl of master mix (5 µl REDTaq ReadyMix master mix, 3.4 µl H_20 and 0.4 µl of each primer) containing species-specific primers for *An. gambiae s.s., An. arabiensis* and *An. merus*.

An. gambiae s.s. mosquitoes were identified further to *An. coluzzii* and *An. gambiae* following the PCR methods of Wilkins (2006). Briefly, 1µl of DNA was added to the following master mix: 5 µl REDTaq ReadyMix Reaction Mix, 0.4µl of

M and S primer (25 pmol/µl each), and 2.4µl of water. PCR products were run on 2% agarose gel at 100 volts for 2 hours and visualised under UV light.

Data analysis

Using the capture data, we calculated the proportion of mosquitoes displaying the following behaviours: endophagy, exophagy and “nocturnality”. Endophagy and exophagy were calculated based on the proportion of mosquitoes caught indoors and outdoors, respectively relative to the total number of mosquitoes caught in each location. Nocturnality is defined as the proportion of mosquitoes caught during peak sleeping hours (21:00-5:00) (Govella *et al.*, 2010).

The standard error of proportions (S.E.) were calculated using the following formula:

$$\mathrm{S.E.} = \sqrt{\left(\frac{\mathrm{p}(1-\mathrm{p})}{\mathrm{n}}\right)}$$

p= proportion of mosquitoes displaying endophagic, exophagic or nocturnal behaviour
n= total number of mosquitoes

A fisher exact test was performed to test if the proportion of the mosquitoes collected indoors and outdoors differed within and between species.

Data comparison

Current endophagy rates and mean time of biting were compared with estimates obtained in 1992, during a control trial of ITNs (Quiňones *et al.*, 1997). The trial consisted of ten pairs of villages and each pair received either a treated or an untreated bednet. Although the authors did not collect mosquitoes in Wali Kunda, two of the villages were close enough to allow comparison. Saruja (untreated) is 1.85 km away while Brikama Ba (treated) is located 3.71 km from Wali Kunda

(Figure 3.1). To investigate changes in endophagy, we calculated the difference in proportion of mosquitoes caught biting indoors (i.e. endophagy) between 1992 and 2012†. Comparisons were made based on the pooled number of *An. gambiae s.l.* because their results were presented at the species level only. 95% confidence intervals (CI) were calculated from the standard error (S.E.) of their difference using the following formula:

$$95\%\ \mathrm{CI} = \mathrm{D} \pm (1.96 * \mathrm{S.E.})$$

D= difference in proportion of endophagy
SE= standard error of *D*

And *p*-values were approximated from the *z*-score:

$$\mathrm{z} = \frac{\mathrm{D}}{\mathrm{S.E.}}$$

We calculated the mean time of biting, indoors and outdoors, as outlined by Quiňones *et al.*'s (1997). For each collection time, we assumed that mosquitoes were collected in the middle of the period. For example, mosquitoes collected during 19:00-20:00 hours were assumed to have been collected at 19:30 hours. Next, we weighted each collection time by the number of mosquitoes caught in each period and calculated the mean biting time. For both endophagy and mean time of biting, we compared our data with the following: 1) The overall rate in all control villages, 2) rates in Saruja and 3) rates in Brikama Ba.

†See Chapter 2 for the formulas to calculate the difference in proportion of endophagy and the standard error of their difference.

Results

Of the mosquitoes that were collected, 0.7% were *An. gambiae s.l.* while 0.33% and 98% were *Culex* and *Mansonia* species, respectively. The remaining 0.97% were made up of other *Anopheles* species such as *An. pharoensis* and *An. zeimanni.* Out of the 499 morphologically identified *An. gambiae s.l., An. arabiensis* was the dominant species (62%) while the remainder were composed of *An. coluzzii* (37%) and *An. gambiae* (1%), broadly similar to previous observations during the early dry season (Caputo *et al.*, 2008). An additional sample could not be resolved due to poor DNA amplification. Seven *An. gambiae s.s.* and one *An. arabiensis* were excluded from further analysis because their hour of collection was not recorded.

Figure 3.4 shows the number of *An. gambiae s.s.* and *An. arabiensis* collected indoors and outdoors, per hour. The majority of the host-seeking activity took place during sleeping hours in both species (0.76 and 0.69 nocturnality in *An. gambiae s.s.* and *An. arabiensis,* respectively). Both species also displayed low outdoor activity before 10:00 pm (8% and 4% for *An. gambiae s.s.* and *An. arabiensis,* respectively).

Indoor-biting activity by *A. gambiae s.s.* increased steadily throughout the night with peak activity occurring early in the morning between 04:00-05:00 am, and rapidly declining thereafter. By comparison, the highest, sustained activity of outdoor-biting was observed during the early morning (03:00-06:00 am). In *An. arabiensis,* indoor biting occurred earlier in the evening, with peak activity occurring between 01:00-02:00 am and again at, 04:00-05:00 am. Similar levels of

activities were observed for their outdoor biting counterparts with the highest level of activity occurring during the later hours/early morning.

A slightly higher proportion of mosquitoes were collected indoors (0.58 ±0.04 and 0.56 ±0.03, respectively) than outdoors in both species (0.42 ±0.04 and 0.44 ±0.03 respectively) (Table 3.1; Figure S1). However, the proportion of mosquitoes biting indoors and outdoors were not significantly different within and between species (p >0.05).

The taxonomic composition of the *An. gambiae* complex differed between 1992 and 2012, with *An. gambiae s.s.* being the prevailing taxon (96.1%) in 1992 compared to 2012 where *An. arabiensis* dominated (62%). We attributed this to the seasonal variation of the two vectors, reflecting how *An. arabiensis* is better adapted to drier conditions. As there was no significant difference in the endophagy rates between species, we proceeded to pool all samples as one group (i.e. *An. gambiae s.l.*) and investigated the difference in endophagy between our data and those observed in 1992. We found a significant reduction in endophagy in Saruja only (D= -0.21; p <0.01). We observed a slight, but non-significant reduction in endophagy in control villages while an increase in endophagy was observed in Brikama Ba (D= -0.015 and 0.13, respectively; p >0.05) (Table 3.2). Regarding changes in mean biting time, we found that the mean biting time of *An. gambiae s.l.*, indoors and outdoors, was broadly similar to those observed in the control villages and Brikama Ba in 1992 (Table 3.3). However, there is some evidence to suggest that *An. gambiae s.l.* has changed towards early biting behaviour when our data was compared to Saruja, although this was not confirmed statistically.

Table 3.1. The proportion of mosquitoes caught indoors and during sleeping hours in Wali Kunda, Gambia.

Species	Behaviour	Proportion	± S.E.	N
An. gambiae s.s.	Endophagy	0.58	0.037	104/178
	Nocturnality	0.76	0.032	136/178
An. arabiensis	Endophagy	0.56	0.028	172/309
	Nocturnality	0.69	0.026	223/309

Nocturnality: 21:00-5:00. S.E. = standard error.

Table 3.2. Difference in the proportion of endophagic, *An. gambiae s.l.* between 1992 and 2012.

Comparison	Diff. in endophagy (95% CI)	*p*-value
Control villages (*n* = 1293) vs Wali Kunda	-0.015 (-0.07,0.04)	0.572
Saruja (control; n = 31) vs Wali kunda	-0.208 (-0.36,-0.05)	0.008
Brikama Ba (intervention; *n* = 44) vs Wali Kunda	0.13 (-0.02,0.29)	0.084

n= number of mosquitoes collected. Abbreviation: Diff. = Difference.

Table 3.3. Mean biting time of indoor and outdoor-collected, *An. gambiae s.l.* in The Gambia in 1992 and 2012.

	Mean biting time	
Location/Year	Indoor (*n*)	Outdoor (*n*)
Control villages/1992	2:53 (749)	2:58 (544)
Saruja (control)/1992	1:47 (24)	1:47 (24)
Brikama Ba (intervention)/1992	2:30 (19)	3:30 (25)
Wali Kunda/2012	2:54 (276)	2:48 (211)

n= number of mosquitoes collected.

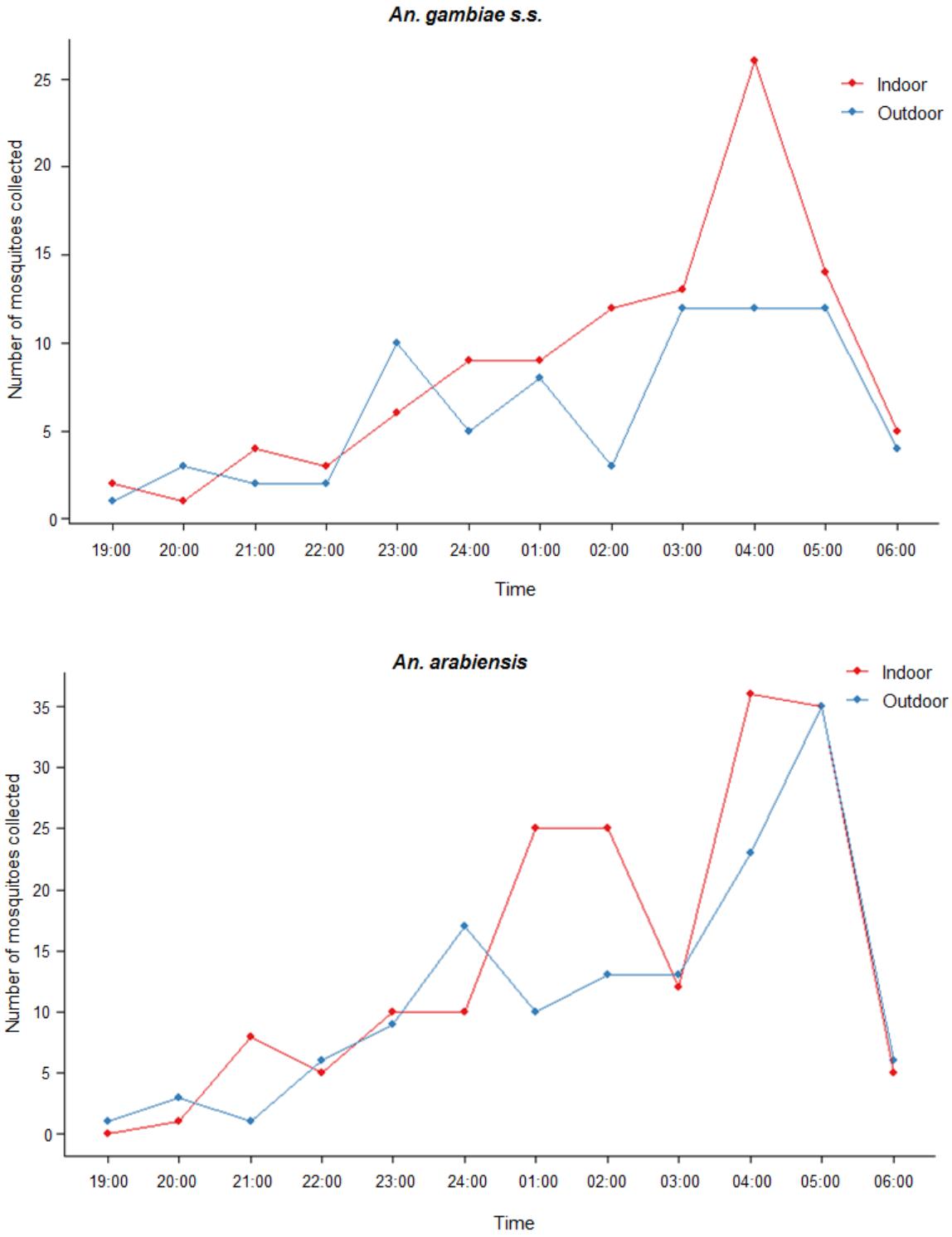

Figure 3.4. Time segregated biting profile of *An. gambiae s.s.* and *An. arabiensis* in Wali Kunda, Gambia in 2012.

Discussion

Here, we aimed to describe the host-seeking behaviour of the *An. gambiae* complex in Wali Kunda, Gambia. We measured the endophagy rate and biting pattern of both *An. gambiae s.s.* and *An. arabiensis* and, compared our observations with those obtained 20 years ago during an ITN trial in villages located across The Gambia which included neighbour villages, Saruja and Brikama.

We did not find a significant difference between the proportion of mosquitoes biting indoors and outdoors in both species, suggesting that the *An. gambiae s.l.* population in Wali Kunda are opportunistic, feeding depending on the availability of accessible human hosts. Based on similar observations reported elsewhere (Lemasson *et al.*, 1997, Oyewole and Awolola, 2006, Antonio-Nkondjio *et al.*, 2002, Bockarie *et al.*, 1993), it seems that both species exhibits greater phenotypic plasticity in their feeding habits beyond what was previously thought (Sinka *et al.*, 2010). The lack of indoor-outdoor preference exhibited by both species in Wali Kunda is therefore not atypical. Relating these results to the endophagy rates observed in 1992, *An. gambiae* complex were less endophagic when compared to the nearby village, Saruja (1.85km away) (D= -0.21; p <0.01).

Evidence for shifts in biting time in response to vector control has been found in several populations (Charlwood and Graves, 1987, Moiroux *et al.*, 2012, Magesa *et al.*, 1991). Early-biting behaviour can reduce the efficiency of ITNs to protect people against infectious bites because the peak activity of *Anopheles* vectors coincides with the time that most people are out of bed and accessible outside. In Wali Kunda, we observed that the majority of host-seeking activity occurred after

midnight in both species. This observation is broadly similar to previous observations where peak biting occurred between 2:00-5:00 am in villages in mid-Gambia (Quiňones *et al.*, 1997). However, there is some indication that the biting behaviour of the *An. gambiae* complex has changed towards biting earlier in the morning, as suggested by the mean biting times observed in Wali Kunda compared to Saruja in 1992 (2:50 am versus 1:47 am, respectively).

Nevertheless, it is wise to be cautious about drawing conclusions because there were several limitations associated with comparing our data against Quiñones *et al.*'s (1997) study. First, vector behaviour may simply differ between sampling locations as a result of differences in selection pressure and ecological settings, and not as a consequence of selection driving behavioural changes. Second, the differences in the taxonomic composition of the two datasets have almost certainly confounded our results. As Quiñones *et al.*'s (1997) study lacked a real control group (i.e. no bednets), it was also not possible to compare how vector behaviour in 2012 differed from the true behaviour of vectors in 1992. Finally, given the small sample sizes in Saruja (n = 31) and Brikama Ba (n = 44) the behaviour sampled may not be representative of the populations' true behaviour.

Although this study did not measure the number of mosquitoes that fed on humans, we assumed that every mosquito caught landing on our collectors represents a blood meal derived from humans. However Caputo *et al.* (2008) found that in Wali Kunda, only 10% of *An. gambiae s.s.* caught indoors fed on humans, suggesting that our method is a crude measurement. Therefore, our study could be improved by measuring the level of blood feeding that takes place on humans in order to provide a comprehensive overview of vector feeding behaviour in Wali Kunda, particularly in cases of multiple blood-feeding. It is

possible that the availability of cattle within the area has sustained the *An. gambiae s.l.* population, which may explain the opportunistic feeding behaviour observed in this study.

After 20 years of vector control, there is some indication that behavioural resistance, in the form of changes in their biting cycle and indoor-outdoor preference, is occuring in *An. gambiae s.l.*. However, the evidence is weak due to a difference in the species composition between the two datasets and a lack of a more appropriate local historical data for comparison.

4. Genetic differentiation in olfactory-related genes between indoor- and outdoor-collected *Anopheles coluzzii* and *Anopheles gambiae*

Abstract

Anopheles coluzzii and *An. gambiae* (hereafter "*An. gambiae s.s."* refers to both species) are major vectors of malaria in sub-Saharan Africa. They are characterised by their strong propensity for biting humans and to enter houses during sleeping hours. Such behaviours are driven by olfactory cues that are perceived by odour receptors (*ORs*) and odour-binding proteins (*OBPs*). Using the Sequenom iPLEX platform, we genotyped single nucleotide polymorphisms (SNPs) at *ORs* and *OBPs* to determine whether there is a genetic difference between indoor- and outdoor-collected *An. coluzzii* and *An. gambiae.* We also compared these vectors with the zoophagic, *An. arabiensis.* Our analysis found little genetic difference between indoor and outdoor-populations, except at the narrowly tuned receptor, *OR1* where markers displayed moderate to high levels of linkage in the outdoor population of *An. gambiae.* We also found substantial gene flow and interbreeding between *An. gambiae s.s* and *An. arabiensis,* as expected from such closely related species. However, we found that *An. coluzzii* and *An. gambiae* were highly diverged from *An. arabiensis* at *OR1* and *OR7,* suggesting a potential role of these genes as modulators of host-seeking preference.

Introduction

Olfactory cues direct a number of behaviours in mosquitoes. This includes food-source foraging (for blood and sugar), oviposition site selection, predator avoidance and mating (Takken and Knols, 1999). Of these, the search for a blood meal host is arguably the most important behaviour because it contributes to the ability of these mosquitoes to transmit diseases such as malaria. Indeed, the main malaria vectors, *Anopheles coluzzii* and *An. gambiae* (hereafter "*An. gambiae s.s."* refers to both species while "*An. gambiae"* refers specifically to the S form) have high propensities to choose human hosts. While significant progress has been made in characterising the molecular components of mosquito olfactory transduction (Hill *et al.*, 2002, Wang *et al.*, 2010a), it remains poorly understood how these components modulate host-seeking preference.

In insects, the process of odour detection begins in the olfactory receptor neurons (ORNs) which are housed within sensory hair structures, sensilla. ORNs are abundantly distributed throughout the mosquito body although the majority are located on the olfactory organs, antennae and maxillary palps (Dahanukar *et al.*, 2005). Odour compounds perceived through these organs are detected by odour receptor (*ORs*) complexes which are expressed by ORNs (Bohbot *et al.*, 2010a). Typically, ORNs express a distinct *OR* gene together with a common, odorant co-receptor (*orco*). The *orco,* formerly named *OR83b* in *Drosophila* and *OR7* in *An. gambiae s.s.,* functions as a chaperone (Vosshall and Hansson, 2011, Sato *et al.*, 2008) while a ligand-selective *OR* confers the complex's binding specificity (Benton *et al.*, 2006).

Odour binding proteins (*OBP*s) are additional components that are involved in olfactory transduction. *OBP*s are present in high concentrations in the olfactory tissues (Pelosi *et al.*, 2006) and act as transporters of odour molecules towards ORs (Pelosi, 1998, Pevsner and Snyder, 1990). Despite their abundance, *OBP*s do not appear essential to olfactory perception (Bohbot *et al.*, 2010b). Experiments expressing *OR* proteins in *Drosophila* (known as heterologous expression systems) have shown that *ORs* can be activated in the absence of *OBP*s (Hallem *et al.*, 2004, Wetzel *et al.*, 2001). Yet, in the presence of *OBPs, ORs* exhibited increased sensitivity to semiochemicals suggesting an additional role of *OBP*s in odour recognition (Matsuo *et al.*, 2007, Wang *et al.*, 2007). Further evidence from electrophysiological experiments, *in vitro* binding assays and behavioural observations in several insect species lends support to the important, albeit unclear role of *OBP*s in the olfactory signalling pathway (Kim *et al.*, 1998, Zhou *et al.*, 2004, Bette *et al.*, 2002, Pophof, 2004).

Given their principal role in the olfactory signalling pathway, *ORs* and *OBPs* are thought to play a role in determining an insect's odour sensitivity and in turn, affect olfactory-driven behaviours[*]. Indeed, *An. gambiae s.s.*, *Aedes aegypti* and *Drosophila melanogaster* show strong divergence as well as, species-specific expansion of particular *OR* and *OBP* genes (Hill *et al.*, 2002, Zhou *et al.*, 2008). Their role in facilitating adaptation to different ecological habitats is also exemplified by the *An. gambiae* complex, where members differ in their status as malaria vectors because of differences in their human-biting behaviour. Human-associated odours drive greater response in *An. gambiae s.s.* whereas its non-vector sibling species, *An. quadriannulatus* is primarily attracted to cow-derived odours (Pates *et al.*, 2001,

[*]See Chapter 1 for information regarding odour reception and, *ORs* and *OBPs.*

Gillies and Coetzee, 1987). Furthermore, olfactory-related genes including *OBP*s are differentially expressed in incipient species*, An. coluzzii* and *An. gambiae* (Cassone *et al.*, 2008). Taken together, differences in odour sensitivities and/or odour ligand-receptor specificities appear to play a central role in mediating host-seeking behaviour and thus, their adaptation to local environments.

The genetic basis of individual- and species-specific odour recognition remains unresolved, although significant progress has been made recently. One emerging hypothesis is that polymorphisms in olfactory-related genes can alter sensitivity to odours and consequently, modify olfactory-driven behaviour (Keller *et al.*, 2007, Wang *et al.*, 2007). In *Drosophila,* varied responses to key odour molecules were associated with single nucleotide polymorphisms (SNPs) and insertion/deletions (indels) in *OBP* and *OR* genes (Rollmann *et al.*, 2010, Wang *et al.*, 2007). Similarly, differences in host preferences between anthrophophilic and zoophagic strains of *Ae. aegypti* were linked to major allelic variations in *OR4* (McBride *et al.*, 2014). The effect of such polymorphisms include modified receptor-ligand interaction, impaired function, altered expression and mRNA structure (Keller *et al.*, 2007, Wang-Sattler *et al.*, 2007, Wang *et al.*, 2010b, Kimchi-Sarfaty *et al.*, 2007). Any of these changes can alter gene expression and/or ligand sensitivity and give rise to alternative behavioural phenotypes.

Since the publication of the *Anopheles* genome, bioinformatics-based strategies have identified a total of 79 and 66 *OR* and *OBP* genes, respectively (Vieira and Rozas, 2011, Holt *et al.*, 2002). Furthermore, recent developments in sequencing technologies have made SNP genotyping relatively easy and affordable. The Sequenom MassARRAY platform is one such technology that allows the genotyping of several thousand SNPs in several hundred samples, with fast turnaround (Gabriel

et al., 2009). Such technology is suitable for targeted, exploratory analysis. Thus, the convenience offered by the Sequenom technology coupled with the abundant distribution of SNPs within the *Anopheles* genome (approximately one SNP per 125 bp) provide an attractive system for population structure studies and identifying SNPs linked to phenotypic traits.

Using the Sequenom technology, this study aimed to investigate the role of olfactory-related genes in mediating the indoor-outdoor host seeking preferences of *An. gambiae s.s..* We also investigated how the primarily zoophagic*, An. arabiensis* differs from *An. gambiae s.s.* at these genes*.* We used a combined approach of literature search and bioinformatics analysis, to identify candidate *OR* and *OBP* genes with a putative role in *Anopheles* host-seeking behaviour. Our primary aim was therefore to identify genetic differences between indoor- and outdoor-collected *An. gambiae s.s.* to determine whether there is evidence of a genetic change towards outdoor-biting behaviour*.*

Methods

Identification of candidate genes

To identify candidate *OR* and *OBP* genes with a putative role in host-seeking activity, we used both literature and gene expression analysis data. Given their primary role in odour reception, *ORs* were our primary candidates while *OBP*s were our secondary candidates. First, we mined the literature for *OR* and *OBP* genes with known or implicated role in host-seeking in *Anopheles.* Potential candidates were selected based on their temporal and tissue expression using the VectorBase and FlyBase databases. We aimed to shortlist ORs and *OBP*s based on the following characteristics:

- Head-specific/ head-biased gene expression
- Female-specific/ female-biased gene expression
- Reduced expression after a blood meal
- Circadian rhythm expression- higher expression during the time of host-seeking
- Experimental evidence demonstrating strong response to human-associated odours

We hypothesised that *OR* and *OBP* genes involved in host-seeking are principally expressed in the mosquito head where the main olfactory organ, the antennae, is located. As blood-feeding is restricted to female behaviour, we prioritised genes which had female-specific/ female-biased expression. Based on *Anopheles'* nocturnal activity, we also posit that *OR* and *OBP* expression display temporal and conditional expression. Indeed, behavioural bioassays have shown that *An. coluzzii*

and *An. gambiae* females are refractory to odour stimulation following a blood meal (Takken *et al.*, 2001). Diel variation in odour responsiveness have also been demonstrated, both at the transcript (Rund *et al.*, 2011) and electrophysiological level (Krishnan *et al.*, 1999).

Genotyping

A list of 2,456 SNPs, of west- and east-African origin, found in each candidate gene were provided by collaborators from the MalariaGEN project. These SNPs were identified as part of MalariaGEN's effort to capture the genetic diversity of *Anopheles* populations.

To speed up SNP assay design, we excluded SNPs that were proximal to another SNP; proximal SNPs tend to introduce non-specific binding and decrease the efficiency of primer binding. We defined proximal SNPs as a SNP that was within 15 bases (both upstream and downstream) of our target SNP. 291 candidate SNPs passed this criteria and were submitted to the Sequenom SpectroDESIGNER® software for assay design of primers. Of those submitted, 256 SNP assays were successfully designed, corresponding to four multiplexes while the remaining SNPs failed because extension primers could not be designed. Each multiplex contained a human (PfTRAP) and *P. falciparum* (G6PD) SNP which served as the positive and negative control, respectively. Due to cost restrictions, we ran 2 out of 4 multiplexes comprising a total of 78 SNPs (Table S3).

An. gambiae s.l. from Uganda were genotyped whereas only *An. coluzzii* were genotyped for the Gambian dataset. All genotyping plates also included *P. falciparum* laboratory isolates and pooled human DNA (HapMap CEPH reference panel) which allowed us to assess the validity of Sequenom's genotype calls.

Experimental design

To investigate the intra- and inter-species population structure of *An. gambiae s.l.* from the Gambia and Uganda, we created three datasets to allow the following comparisons: 1) indoor versus outdoor *An. coluzzii* from The Gambia, 2) indoor versus outdoor *An. gambiae* from Uganda and 3) inter-species comparison of *An. gambiae s.l.* from Uganda.

Sample collection and preparation*

Briefly, indoor- and outdoor-biting mosquitoes were collected by human landing catches (HLC) in both The Gambia and Uganda. In The Gambia, collections were performed for the whole night (19:00-7:00 hours) while collections were conducted for part of the night only in Uganda (19:00-24:00). *An. gambiae s.l.* mosquitoes were morphologically identified and stored at -20°C in individual tubes containing 80% ethanol or RNAlater.

An. gambiae s.l. mosquitoes were prepared for whole-genome amplification by: 1) removing legs from each mosquito and placing them into 96-well plates containing 5µl of water per well or by 2) adding 5µl of genomic DNA (1ng/ µl) to each well of a 96-well plate. For gDNA extractions, mosquitoes were homogenised using the AllPrep DNA/RNA Micro kit (Qiagen, Manchester, UK) according to the manufacturers' protocols, except that DNA was eluted in 100µl of RNase-free water. Hereafter, these samples are referred to as gDNA samples. We also selected 26 laboratory, *An. gambiae s.s.* mosquitoes to do a full gDNA extraction for concordance comparisons as outlined under "Assay performance and validation". Hereafter, these samples are referred to as control samples.

*See Chapters 3 and 5 for details regarding sample collection and preparation of Gambian and Ugandan samples, respectively.

Sequenom procedure

SNP genotyping using the Sequenom technology consists of several steps outlined below. An overview of the entire workflow is shown in (Figure 4.1).

Primer-Extension Preamplification (PEP)

Leg and gDNA samples were whole-genome amplified using the primer-extension preamplification (PEP) protocol (Zhang *et al.*, 1992). The mastermix consisted of: 1.10µl of N15 primers (Sigma, UK), 2.5µl of 10x PCR buffer (160 mM $(NH_4)_2SO_4$, 670 mM Tris-HCl, pH 8.8), 1.25µl of $MgCl_2$ (50mM), 0.63µl mixture of dNTPs (8 mM each) and 0.25µl Taq polymerase (Biotaq, 5U/µl), made up to 20µl with water. Fifty primer-extension cycles were carried out in a MJ Research thermocycler. The PCR cycling conditions consisted of a 4 minute denaturation step at 94°C, a 2 minutes annealing step at 37°C, a programmed ramping step of 1sec /0.1°Cto 55°C, 4 minutes incubation at 55°C for extension, followed by 5 minutes final extension at 72°C. PEP DNA were stored at -20°C until used.

Species complex and molecular form identification†

An. gambiae s.l. were identified to sibling species using Scott ***et al.'s*** PCR procedures using primers specific to *An. gambiae s.s.*, *An. arabiensis* and *An. merus. An. gambiae s.s.* mosquitoes were then identified to incipient species, *An. coluzzii* and *An. gambiae* following the PCR methods of Wilkins *et al.* (2006). Briefly, 1µl aliquots of PEP DNA was added to the following master mix: 5µl REDTaq ReadyMix Reaction Mix, 0.4µl of *An. coluzzii* and *An. gambiae* primers (25 pmol/µl each), and 2.4µl of water. Primers create fragments of 427bp and 335bp corresponding to *An. coluzzii* and *An. gambiae* respectively.

†Details of the sibling species and incipient species (molecular forms) PCR identification are outlined in Chapter 3 and also, in Appendices 4 and 5.

PCR products were run on 2% agarose gel at 100 volts for 2 hours and were visualised under UV light. Neat, intense bands were indicative of a successful PEP run. Samples that failed the PCR reaction were re-run.

First-round PCR

Sequenom's MassARRAY Designer software designs PCR and extension primers for each SNP to be investigated. First-round primers were made up to 100µM with water. The PCR master mix consisted of: 0.69µl primer mix (containing equal amounts of sense and anti-sense primers), 0.56µl $MgCl_2$ (25mM), 0.17µl dNTPs (25mM each), 1.07µl 10x Taq buffer, 2.32µl MilliQ water and 0.34µl HotStart Taq (5U/µl) (Qiagen, UK). 4.5µl of PCR master mix was added to each well of a 384-well PCR plate (Thermo-Fisher Scientific, UK) containing 2µl of 1:10 PEP DNA. Plates were sealed with Microseal 'A' lids (Bio-Rad, UK) and centrifuged at 1000rmp for 1 minute. The PCR was run in a MJ thermocycler with the following conditions: 94^{o}C for 15 min, 44 cycles of (94^{o}C for 20 sec, 56^{o}C for 30 sec, 72^{o}C for 1 min), and 72^{o}C for 3 min and 15^{o}C for 15 min. 1µl aliquots of the product were run on 2% agarose gel to confirm the PCR had worked prior to further processing.

Post-PCR shrimp-alkaline phosphatase (SAP) clean-up

Unincorporated dNTPs were removed with iPLEX shrimp-alkaline phosphatase (SAP) mixture (0.17µl of 10x SAP buffer, 0.3µl of SAP (1.7 U/µl SAP) and 1.53µl of water). 2µl of SAP mixture were added to 5µl of first-round PCR products. Plates were re-sealed with Microseal 'A' lids and centrifuged at 3,000 rpm for 30 seconds. Treated plates were then incubated at 37^{o}C for 40 min, followed by heat inactivation at 85^{o}C for 5 min and then cooling to 15^{o}C for 15 min.

Primer extension

Extension primers were re-suspended in 300µM stock solutions. Primer extension was carried out using the following master mix: 0.27µl iPLEX termination mixture, 0.06µl extension Taq, 0.27µl extension buffer, 1.67µl primer mixture and 0.44µl water. 2µl of master mix were added per well. After a 30 second centrifugation at 3,000 rpm, the extension reaction was run using the following conditions: 94^{o}C for 30 sec, 40 cycles of (94^{o}C for 5 sec, 5 cycles of [52^{o}C for 5 sec, 80^{o}C for 5 sec]), then 72^{o}C for 3 min and 15^{o}C for 15 min. Note: the final concentration of the extension primers was determined by their mass: <3000Da 0.625 µM; 3000- 5800 Da 0.84 µM; 5800-7000 Da 1.04 µM; 7000- 10,000 Da 1.25 µM; >10,000 Da 1.5 µM.

Resin clean-up

To optimise mass spectrometry analysis, products from the primer extension are treated with resin to remove salts such as Na^{+}, K^{+} and Mg^{2+} ions. 6mg ion-exchange resin (Sequenom®) and 16µl MilliQ water were added to each well. Plates were rotated for 30 minutes and subsequently, centrifuged to pellet the resin. 15nl of samples were subsequently spotted onto SpectroCHIPS and run on the Mass-Spectrometer.

Assay performance and validation

Genotypes were called using the Sequenom® Typer v4.03 software which measures the mass of each primer extension product and compares it with the calculated mass of all possible extension products in the multiplex. All data were initially inspected to assess the assay quality. Any assays with weak signals, positive signals detected in water samples or ambiguous cluster plots were

discarded. Samples were also discarded if their genotype call were ambiguous between a homozygous or heterozygous call (Figure S2).

Next, we determined assay performance based on the pass rate across all samples. Pass rates were calculated as the proportion of samples in which a genotype was successfully called. We also evaluated genotyping accuracy in the control samples by performing a concordance analysis between Sequenom's genotype calls and pre-existing calls produced from the Illumina sequencing platform and GATK's Unified Genotyper (UG) calling algorithm, supplied by the MalariaGEN. Concordance was calculated as the percentage of calls where both methods produced identical genotypes, once SNPs where Sequenom could not produce a genotype were discarded.

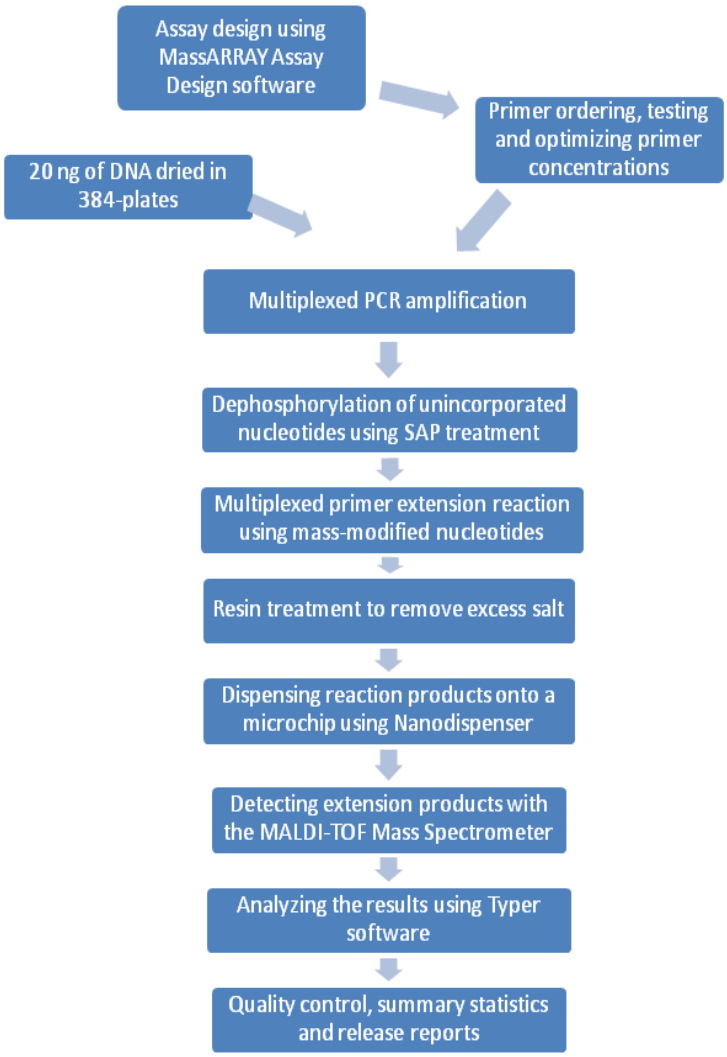

Figure 4.1. Workflow of Sequenom genotyping procedure. The procedure consists of a first round locus-specific PCR reaction, followed by a single base extension using mass-modified dideoxynucleotide terminators of a primer that anneals immediately upstream of the site of interest. Genotypes are detected using MALDI-TOF, Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight, mass spectrometry (FIMM, 2012).

Quality filtering

For each dataset, genotypes were filtered based on four criteria: 1) “missingness”, 2) deviations from Hardy-Weinberg Equilibrium (HWE), 3) concordance and 4) allele frequency. Quality assessment was performed using the PLINK v1.07 (Purcell *et al.*, 2007).

"Missingness"

SNPs with a high proportion of missing genotypes are uninformative markers and often add noise during analysis. First, we excluded SNPs with pass rates below 90%. Next, we split the data according to the four datasets and for each dataset, excluded SNPs that had more than 10% missing data. We also excluded samples with more than 10% missing data.

Hardy-Weinberg Equilibrium (HWE)

Extreme deviations from HWE can be indicative of genotyping error. Using an exact test implemented by PLINK, markers were tested for significant deviations from HWE. SNPs were marked for possible exclusion using a threshold of $p < 1.x10^{-6}$. We subsequently examined the cluster plots of these SNPs to evaluate the accuracy of their genotype calls. SNPs were excluded on the basis of having ambiguous cluster plots.

Concordance

We performed a concordance analysis using the genotype calls obtained from the control samples. We examined the concordance of Sequenom versus UG, and identified SNPs with concordance below 90%. As with HWE filtering, we verified Sequenom's calls by examining the cluster plots of these SNPs and excluded those with poor clustering.

Allele Frequency

As a final step, monomorphic SNPs were excluded.

Population genetic analysis

Genetic polymorphism

We calculated the observed heterozygosity and minor allele frequency (MAF) per locus using the R adegenet package and PLINK software, respectively (Purcell *et al.*, 2007, Jombart, 2008). Mean MAF of the two populations were compared using a paired t-test. Similarly, a t-test was performed to detect if the overall observed heterozygosity (H_o) was significantly different from expected (H_e).

H_e at each locus was calculated as:

$$\mathrm{H_e} = 1 - \sum \mathrm{p_i^2}$$

p_i = frequency of the allele *i*
$\sum$ = sum of all alleles

At the population level, H_e was calculated as:

$$\frac{1}{\mathrm{K}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \left(1 - \sum_{\mathrm{i}=1}^{\mathrm{m(k)}} \mathrm{f_i^2}\right)$$

m(k) = number of alleles of locus
K = total number of loci
f_i = frequency of allele *i* in a given population

Divergence and gene flow estimates

We calculated the genetic differentiation between indoor and outdoor populations using Nei's unbiased estimator of pairwise F_{st} (Nei, 1973), as implemented in the R adegenet package:

$$F_{st}(\mathrm{A,B}) = \frac{\left(\mathrm{H_t} - \frac{\mathrm{n_A H_e(A) + n_B H_e(B)}}{\mathrm{n_A + n_B}}\right)}{\mathrm{Ht}}$$

A and *B* = two populations of sample size n_A and n_B
He = expected heterozygosity of a given population (*He(A) OR He(B))*
H_t = expected heterozygosity in the whole dataset

F_{st} ranges from zero to one, corresponding to little to complete divergence between populations. We also calculated F_{st} for each locus using the R pegas package (Paradis, 2010), to identify locus-specific effects such as selection.

Gene flow is defined as the movement of alleles between populations; it can provide important insights into the ancestry of two populations. For each locus, we estimated gene flow (*Nm*) using locus-specific F_{st} estimates (Slatkin, 1985, Wright, 1951).

$$Nm = \frac{(1 - F_{st})}{4F_{st}}$$

Nm represents the number of migrants exchanged between populations per generation. Both estimates of *Nm* and F_{st} assume an infinite island model: 1) mutation and selection are absent, 2) there is no spatial structure between populations such that each population is equally accessible from every other and 3) each population has reached an equilibrium between migration and drift (Whitlock and McCauley, 1999). Although the assumption regarding selection is violated under our hypothesis, *Nm* estimates based on the F_{st} method are robust to selection and population structure (Slatkin and Barton, 1989). *Nm* is inversely related to F_{st} so that low *Nm* estimates results in relatively large values of F_{st} and vice versa.

Population structure

To investigate genetic relationships between samples, we performed a principal component analysis (PCA) using the R package, SNPRelate (Zheng *et al.*, 2012). SNPs with missing genotypes were replaced by the mean of the corresponding

allele frequency. We used plots of the first, second and third principal components to visualise variation within and between populations.

We also compared the results of the PCA with those of the clustering software, ADMIXTURE v1.23 (Alexander *et al.*, 2009). Using a maximum likelihood approach, ADMIXTURE estimates the genetic relationship between populations by assigning them to clusters ("*K*") that have distinctive allele frequencies. It is therefore useful for identifying population structure and mixed ancestry among populations. We ran ADMIXTURE using 10 cross-validation procedures and simulated 1 to 5 populations (K = 1-5) while all other parameters were set to default. We chose the model corresponding to an estimate of K with the lowest cross-validation error, giving the optimal number of populations.

Hardy-Weinberg Equilibrium (HWE)

SNP markers were tested for departures from HWE using the R adegenet package (Jombart, 2008). Under the null hypothesis, markers are in HW proportions in the absence of processes such as selection and non-random mating. Deviations from HWE were estimated using a chi-square test and one-tailed probability values were generated using 10,000 Monte Carlo permutations of the dataset.

Linkage disequilibrium

Pair-wise estimates of linkage disequilibrium (LD) were measured for all possible marker pairs using the correlation coefficient, r^2. We used the PLINK software to estimate r^2 and restricted the analysis between marker pairs that were no more than 1 Mb apart. We then used the R package LDheatmap to draw triangle plots of LD for each chromosome arm (Shin *et al.*, 2006). r^2 values range from 0 (no

linkage between two loci) to 1 (complete linkage between two loci), with r^2 defined as:

$$r^2 = \frac{(x_{AB} - p_A p_B)^2}{p_A p_a p_B p_b}$$

$p_A p_a$; $p_B p_b$ = frequency of alleles A and a at locus 1; frequency of alleles B and b at locus 2

X_{AB} = frequency of haplotypes having allele A at locus 1 and allele B at locus 2

Results

Identification of candidate genes

The initial literature search focused on olfactory-related genes with putative roles in host-seeking behaviour. We then compared these genes against published microarray data to identify their tissue, bloodmeal and circadian rhythm expression (Marinotti *et al.*, 2006, Baker *et al.*, 2011, Rund *et al.*, 2011). In this way we hoped to identify genes that met one or more of our search criteria. We also examined all other known *OR* genes available in VectorBase and performed similar analyses. Of the 79 *OR* genes, 71 had expression data thereby allowing us draw an expression heatmap using data available from VectorBase (Figure 4.2).

For many of the ORs and *OBP*s we found that gene expression was not restricted to the head. Similarly, the microarray data yielded no female-specific genes although female-bias has been demonstrated for some. Taken together, the overall ambiguity of the data made our selection of candidate genes challenging using our original filtering criteria. We therefore selected genes that have been studied extensively in the literature for their putative role in host-seeking behaviour using techniques such as behavioural bioassays and electrophysiological responses to odour compounds. Our search resulted in a list of 8 *OR* and 5 *OBP* genes (Table 4.1).

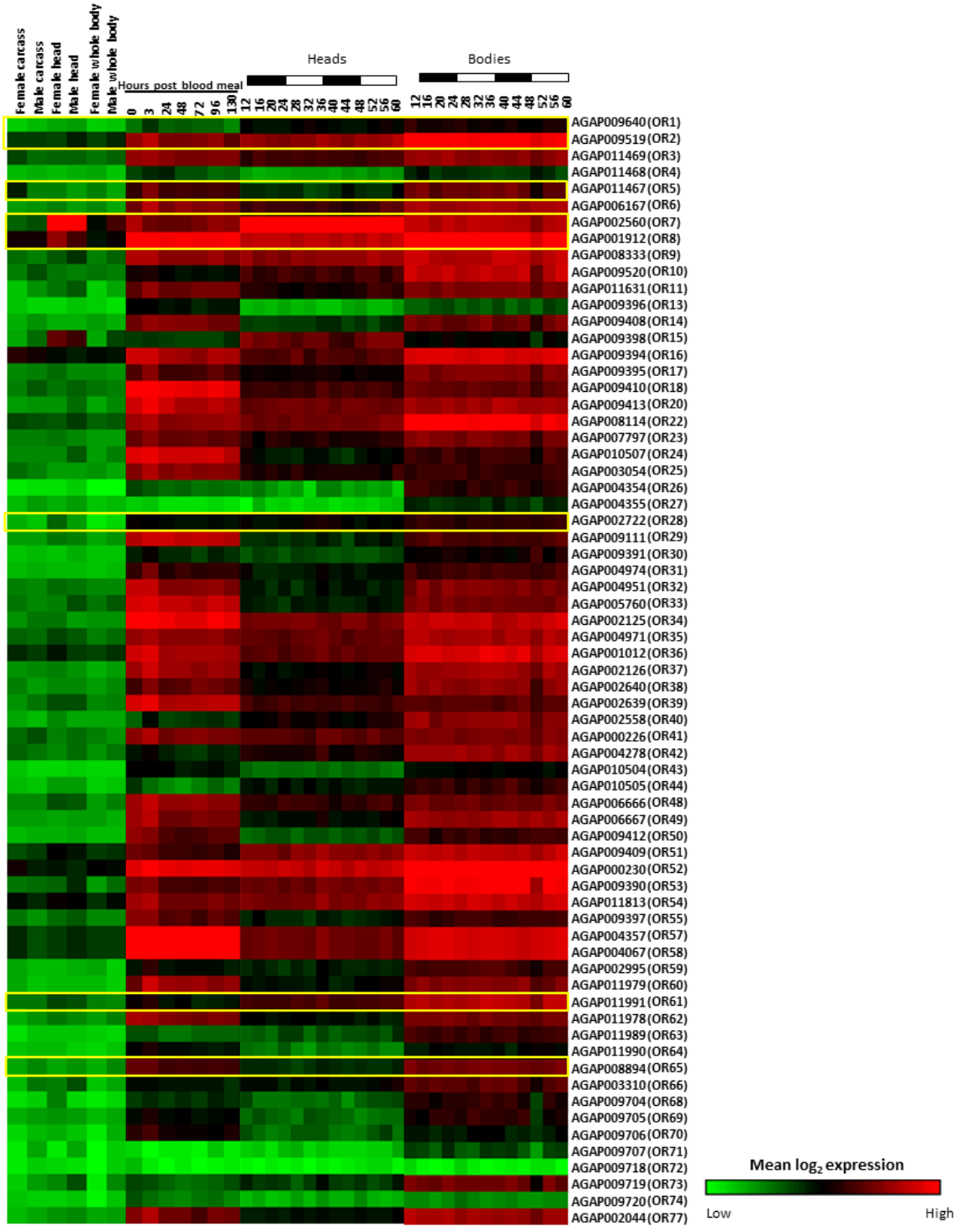

Figure 4.2. Tissue, blood meal and circadian rhythm expression of 71 *OR* genes in *An. gambiae s.s.*. Circadian rhythm expression is displayed separately for the mosquito head and body under light (▭) and dark (▬) conditions. Boxed in yellow (▭) are 8 *OR* genes which have been postulated to have a role in host-seeking behaviour. Bioinformatics analysis used microarray data generated by Baker *et al.* (2011), Marinotti *et al.* (2006) and Rund *et al.* (2011). Examination of the microarray data yielded no obvious candidates genes under our criteria of head-specific/biased expression, female-specific/biased expression, reduced expression post bloodmeal and higher expression during the dark phase.

Table 4.1. List of candidate genes and their relevance.

Vectorbase ID	Gene Name	Relevance	Reference
AGAP002560	OR7	A known *Drosophila orco* ortholog; responds to components of human sweat and skin emanation; rhythmically expressed with peak expression during the night/dark phase	Hallem *et al.* (2004); Kwon *et al.* (2006); Pitts *et al.* (2006); Rund *et al.* (2011)
AGAP009519	OR2	Narrowly tuned receptors; responds to components of human breath and sweat	Carey *et al.* (2010)
AGAP009640	OR1		
AGAP001912	OR8		
AGAP008894	OR65	Narrowly tuned receptor; responds strongly to 2-ethylphenol, a compound found in animal urine	
AGAP011467	OR5	Responds to 2,3-butanedione, a metabolic by product of human skin microflora	
AGAP002722	OR28	broadly tuned receptor	Lu *et al.* (2007)
AGAP0011991	OR61	Female-biased expression in the antennae	Baker *et al.* (2011); Iaotrou and Biessmann (2008)
AGAP007287	*OBP*47	Circadian rhythm expression with peak expression during the night/dark phase	Rund *et al.* (2011)
AGAP010409	*OBP*22	Lower level of blood-feeding following RNAi silencing; rhythmically expressed with peak expression during the night/dark phase	Das and Dimopoulos (2008); Rund *et al.* (2011, 2013)
AGAP012321	*OBP*26		
AGAP010489	*OBP*4	Circadian rhythm expression expressed with peak expression during the night/dark phase; Significant reduction in blood-feeding propensity following RNAi silencing	Das and Dimopoulos (2008)
AGAP001556	*OBP*7	10-fold higher expression in females	Biessmann (2005)

Assay performance and validation

Of the 76 SNP assays, we discarded 10 SNPs because of ambiguous clustering and positive signals detected in the water samples. We excluded an additional 13 SNPs because their pass rate was below the threshold cut-off (90%). The remaining 53 SNPs represented our core SNP set which produced an average pass rate of 96.66% (Figure 4.3). We performed concordance analysis by comparing the genotype calls of Sequenom with those generated from whole genome sequences via the Illumina platform. The analysis comprised of 1,378 genotype comparisons (53 SNPs x 26 control samples), producing an overall concordance rate of 99.33%. We found that nearly all 53 SNPs had a concordance rate of 100%, except for three (77-96%). To identify the source of disconcordance, we re-examined the cluster plots of these three SNPs. Surprisingly, none of the SNPs had poor clustering, suggesting that the majority of the disconcordance were due to differences in Sequenom's and UG's genotype calling algorithms.

Genotyping via the Sequenom platform is achieved by measuring the signal intensity for each allele. Weak signals due to poor PCR amplification or failed primer reaction lead to allele misclassification and missing data. Misclassification occurs because homozygotes typically show a strong signal for only one allele while heterozygotes have signals for both alleles (Conneely, 2008, Heid *et al.*, 2008). This can potentially bias genotype calling since heterozygotes with weaker signals may be misclassified as a homozygote. To identify such genotyping bias, we summarised the disconcordance in all three SNPs according to their genotype classes: a) homozygous reference, b) heterozygous and c) homozygous alternate

(Table 4.2). The disconcordance matrix showed no bias towards a particular genotype class, suggestive of random effects in genotyping error.

Missing data because of Sequenom's failure to successfully call a genotype is most likely to occur in heterozygotes as they often look like sequencing errors. In contrast, the Illumina platform has greater sensitivity because of the high number of independent sequence reads at each SNP position. Therefore, by comparing the occurrence of UG heterozygous calls against missingness in the Sequenom data, we were able to detect bias against heterozygotes. We found no correlation between UG heterozygous calls and missingness in the Sequenom data (r^2 = -0.24) (Figure S3), suggesting that genotype discrepancies were due to allele misclassifications rather than the failure to determine a genotype. As the overall genotype concordance was high (99%) we proceeded with the analysis using the genotype calls produced by Sequenom.

Table 4.2. Disconcordance matrix between Sequenom (SQ) and the GATK's Unified Genotyper (UG).

		UG		
		Homozygous Reference	Heterozygous	Homozygous Alternate
SQ	Homozygous Reference	57/60	-	-
	Heterozygous	-	10/16	-
	Homozygous Alternate	-	-	2/2

Values represent the number of calls that were concordant out of the total calls per genotype class (e.g. 57/60 were concordant for the homozygous reference allele). The matrix represent the results of 3/53 typed SNPs which had a concordance rate below 100%. Concordance was calculated based on the control samples of 26 laboratory, *An. gambiae s.s.*

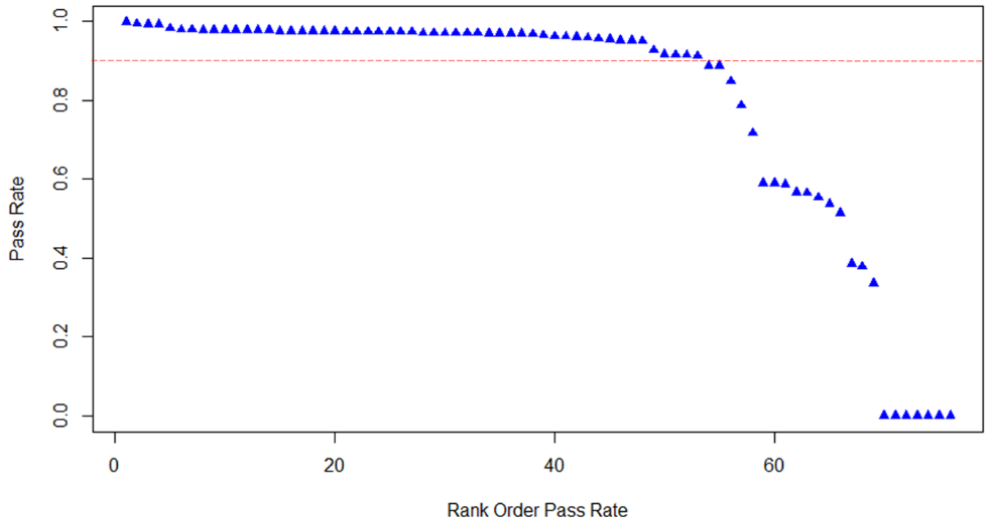

Figure 4.3. Assay pass rate of 76 SNPs. The pass rate for each SNP were calculated across all 477 samples from the Gambia and Uganda. The Y axis represents the pass rate and the X axis represents the rank order of the pass rates. The red horizontal line has been drawn at a pass rate of 0.9 (90%) for a nominal cut-off threshold.

Samples

A total of 185 *An. gambiae s.s.* mosquitoes were collected from the Gambia and of these, 183 and 2 corresponded to incipient species, *An. coluzzii* and *An. gambiae,* respectively. In contrast, samples from Uganda consisted of 47 *An. arabiensis*, 231 *An. gambiae* and 3 hybrids between *An. gambiae* and *An. arabiensis*. An additional 11 Ugandan samples failed to produce PCR amplicons despite repeated attempts to identify their species identity.

Quality filtering

Table 4.3 provides an outline of the number of SNPs and samples retained after quality filtering. As we aim to perform multiple comparisons, different analyses were based on different numbers of SNPs, depending on the number of samples included in the dataset. The variable number of SNPs and samples included were due to the filtering rules applied based on “missingness”. Furthermore, both the HWE test and concordance analysis identified markers to be excluded (p <1.x10^{-6} and concordance rate <90%, respectively). However, we decided to keep all

markers on the basis that their cluster plots showed acceptable levels of genotype clustering. It should be noted that not all genes were included in the datasets due to the allele frequency filter.

Table 4.3. Metrics of datasets generated based on the species or population included and a non-missing genotyping rate of 90%.

Data set	Country	Species	Comparison	No. of samples	No. of SNPs
1	Gambia and Uganda	*An. gambiae s.s.*	Indoor vs outdoor	412	34
2			Gambia vs Uganda	412	34
3	Uganda	*An. gambiae s.l.*	Between species	279	31

For two samples collected in Uganda, species identification was unsuccessful despite repeated attempts to identify them, thus we classified their species type based on the results of the PCA analysis.

Population genetic analysis

Differentiation between indoor and outdoor *An. gambiae s.s.* populations

An. coluzzii samples were clearly distinguishable from *An. gambiae* by PCA Figure S4). To avoid potential bias due to the genetic differences between the two species, we conducted the analysis of indoor and outdoor populations separately for each species.

Low frequency variants dominated the allele frequency spectrum in both species (Figure 4.4). In *An. coluzzii*, outdoor-collected samples had a slightly greater number of low frequency variants than indoor-collected samples. However, we found no significant difference between the population average MAF between the two populations in both *An. coluzzii* and *An. gambiae* (95% CI: -0.07-0.05 and -0.08-0.06, respectively; p <0.05).

Overall heterozygosity was low in both species and population, yielding an H_o mean range of 0.17-0.23 (Table 4.4). At the species level, significant departures from HW expectation were found in 12 out of 57 tests performed in *An. coluzzii* (*OR65*, *OR1*, *OR5, OR7* and *OR8*) and *An. gambiae* (*OR65*, *OR7* and *OR8* (*p*

<0.05) (Table S4). All deviations were associated with positive F_{is} values, suggestive of deficits in heterozygotes. However, overall H_o was significantly lower than H_e (95% CI: 0.002-0.028; p <0.05) in *An. coluzzii* only. To elucidate whether the deficit in heterozygotes was due to pooling of subpopulations of different genotype frequencies (Wahlund effect), we repeated the test separately for indoor- and outdoor-collected *An. coluzzii*. Mean H_o remained significantly lower than H_e in the outdoor population (95% CI: 0.001-0.031; p <0.05), excluding Wahlund effect explanation.

We found very little differentiation between the indoor and outdoor populations, yielding an overall F_{st} estimate of 0.003 for both *An. coluzzii* and *An. gambiae*. Only 10 (F_{st} = 2.4x10^{-6}-0.01) and 6 (F_{st} = 0.002-0.03) loci were divergent between the two populations in *An. coluzzii* and *An. gambiae,* respectively while the rest showed no differences (F_{st} = 0). As a result, gene flow estimates could only be calculated for these loci, yielding an *Nm* range suggestive of moderate gene flow at these divergent genes (*Nm* = 25-250 and 7-125 in *An. coluzzii* and *An. gambiae,* respectively) (Table S5). The results of both the PCA and cluster analysis also showed little differentiation between the two populations, with indoor samples genetically indistinguishable from outdoor samples, in both species (Figure 4.5, Figure 4.6 &Figure S5). Taken together, these results indicate that indoor- and outdoor-collected samples belong to a single panmictic population.

Pairwise LD tests were performed across all loci for each species and population (Figure 4.7, Figure S6 & Figure S7). Linkage between markers was generally low between most locus-pair combinations, in both species. Given that these markers were generally more than 1 Mb apart, this result was not surprising. The strongest linkage was between OR1_3R_37462048 and OR1_3R_37462009 (r^2 >0.7) in all

population-species combination, which can be attributed to the close physical linkage between the two loci (39bp apart). Interestingly, markers in the narrowly tuned receptor, *OR1* showed moderate to high levels of LD in the outdoor population of *An. gambiae* only, potentially indicative of population-specific selection (Figure 4.7).

Table 4.4. Average heterozygosity for each species and population.

	An. coluzzii (The Gambia)		*An. gambiae* (Uganda)	
	H_o	H_e	H_o	H_e
Indoor	0.182	0.197	0.171	0.178
Outdoor	0.207*	0.227	0.18	0.183
All	0.178*	0.194	0.175	0.181

Results are based on 31 and 26 markers in *An. coluzzii* and *An. gambiae,* respectively. "All" = indoor and outdoor populations are pooled into a single population. *Significant departures of H_o from H_e using a t-test (p <0.05). Abbreviations: H_o =observed heterozygosity; H_o =expected heterozygosity.

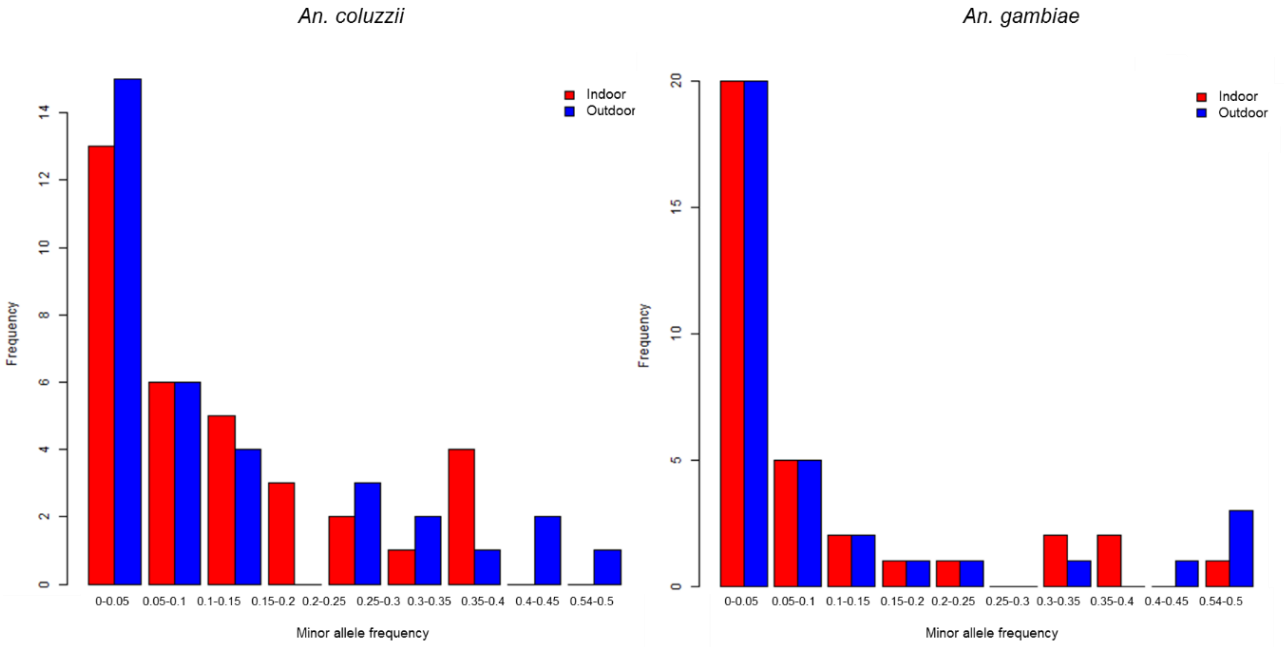

Figure 4.4. Minor allele frequency spectrum of SNPs in indoor- and outdoor-collected mosquitoes of *An. coluzzii* and *An. gambiae* from The Gambia and Uganda, respectively. This graph represents the results of 34 SNPs.

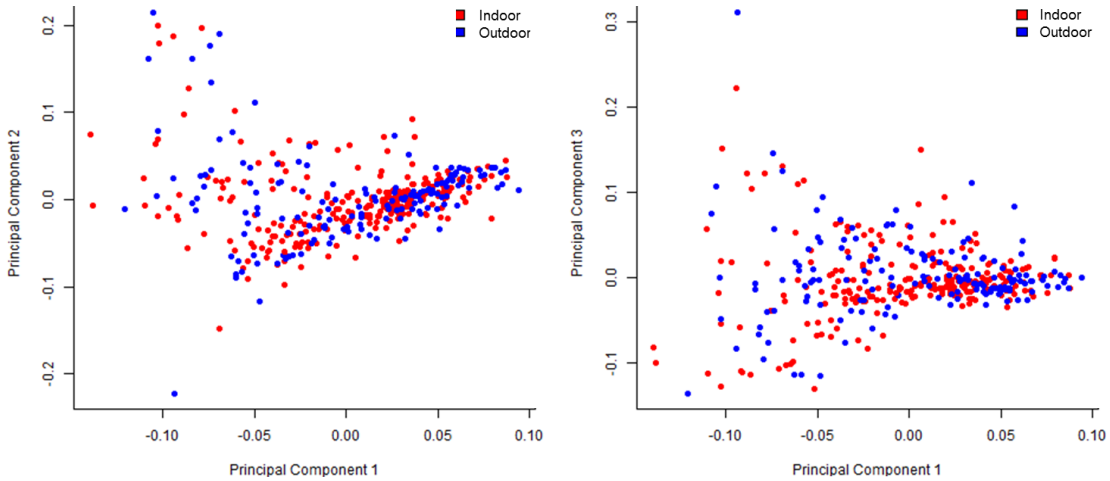

Figure 4.5. Principal component analysis of indoor and outdoor- collected *An. coluzzii* and *An. gambiae* from Uganda and Gambia using 31 and 26 SNPs, respectively. A single mosquito corresponds to a single point.

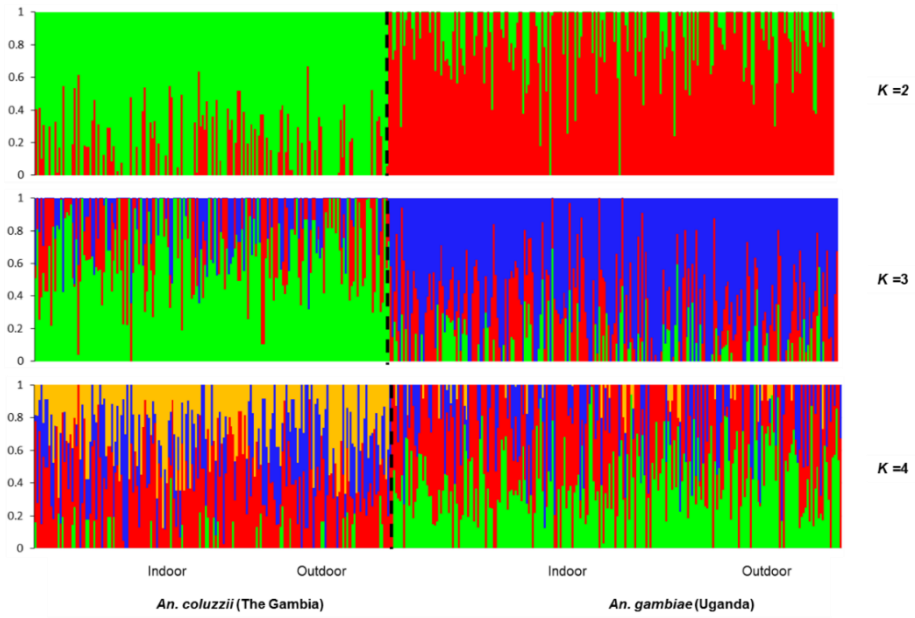

Figure 4.6. Results of the cluster analysis. Each sample is represented by a vertical stacked column. Under a model of two to five subpopulations (*K* = 2-5), each sample is partitioned into *K* coloured segments that represents an individual's estimated membership fractions into *K* subpopulation. The results of the cluster analysis suggested a single population as the optimal model (Figure S2). Indoor and outdoor samples could not be partitioned despite increasing number of subpopulations (*K*), however *An. coluzzii* could be distinguished from *An. gambiae* under a model of two subpopulations (*K* = 2).

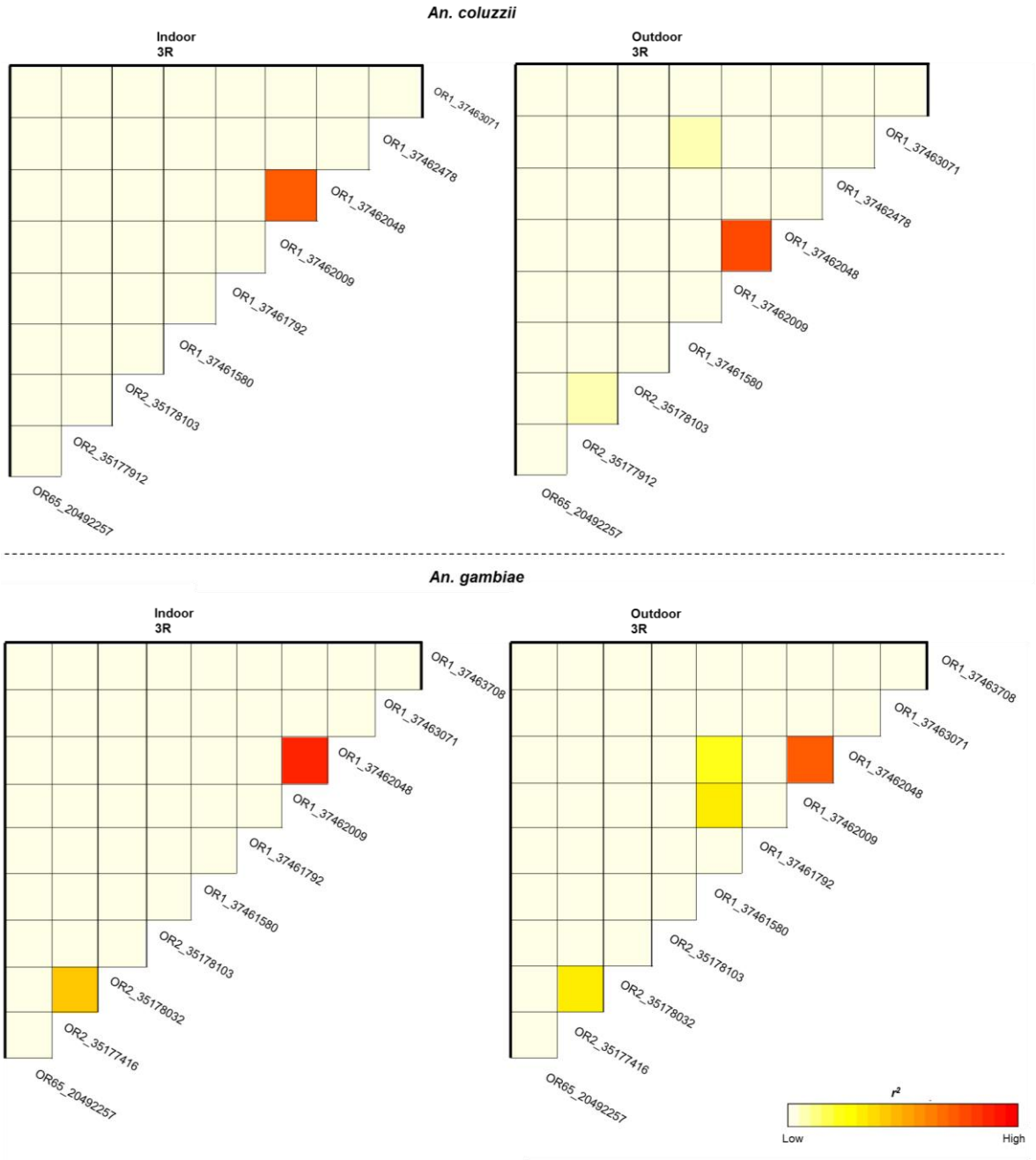

Figure 4.7. Linkage disequilibrium, r^2 estimate between SNP markers in chromosome arm 3R, displaying linkage disequilibrium in *An. coluzzii* and *An. gambiae* from the Gambia and Uganda, respectively. Linkage between markers are represented by r^2 values that range from 0 (white) to 1 (red). OR1 showed moderate levels of LD in outdoor-collected *An. gambiae* suggesting that this gene is under selection in this population.

Identification of species divergence

We investigated the divergence between species by performing a PCA and cluster analysis. Both analyses distinguished *An. gambiae* from *An. arabiensis*; however, the hybrid samples clustered with *An. gambiae*,suggestive of introgressive hybridisation in the direction *An. gambiae* (Figure 4.8A). The result of the cluster

analysis suggests that there is a substantial amount of gene flow between *An. gambiae* and *An. arabiensis.* Indeed, we found that up to over 65% of the *An. gambiae's* genetic variation has shared ancestry with *An. arabiensis* (Figure 4.8B).

These analyses were consistent with the results of the F_{st} analysis which showed that the divergence between the two species is low, yielding a pairwise F_{st} estimate of 0.02. Pronounced differences (F_{st} >0.1) between the two species were found in nine loci only, with the highest differentiation at *OR1* (position: 37462441; F_{st} = 0.93) and *OR7* (position: 22850829; F_{st} = 0.78) (Table S5). At the *OR1* loci, allele, T was rare in *An. gambiae* (MAF = 0.07) while it was at high frequencies in *An. arabiensis* (MAF = 0.47). Differences between allele frequencies were however not significant overall (95% *CI:* -0.04-0.04, *p* >0.05), with both species dominated by rare alleles (MAF <0.05) (Figure 4.9). Accordingly, gene flow estimates were high (*Nm* range: 0.02- 62) for loci with shared polymorphisms between *An. gambiae* and *An. arabiensis.* Taken together, these analyses suggest that there is substantial interbreeding between *An. gambiae* and *An. arabiensis* in Tororo district, Uganda.

LD patterns differ between populations, revealing important insights into a population's history such as effective population size and selection. We found that the pattern of LD differed substantially between species (Figure 4.10). In *An. gambiae,* estimates of r^2 suggested little LD between markers except in chromosome 3R where marker pairs, OR1_3R_37462048 and OR1_3R_37462009 showed strong linkage (r^2 = 0.84), presumably due to their close proximity (39bp apart). We also found modest linkage (r^2 = 0.41) in *OR2* between markers, OR2_3R_35178032 and OR2_3R_35177416 (600bp apart). In contrast, r^2 estimates were low in *OR1* and *OR2* in *An. arabiensis* (r^2 = 0-0.23) but

LD was substantial in chromosome 2R. In particular, we found modest linkage (r^2 >0.45) between two marker pairs in OR7 (OR7_2R_22853803 and OR7_2R_22850829; OR7_2R_22856191 and OR7_2R_22850829) that were ~3kb and 5.4kb apart, respectively. Given that we observed lower genetic diversity in *An. arabiensis* than *An. gambiae,* we consider these differences to be a result of differences in the effective population size between the two species and not selection.

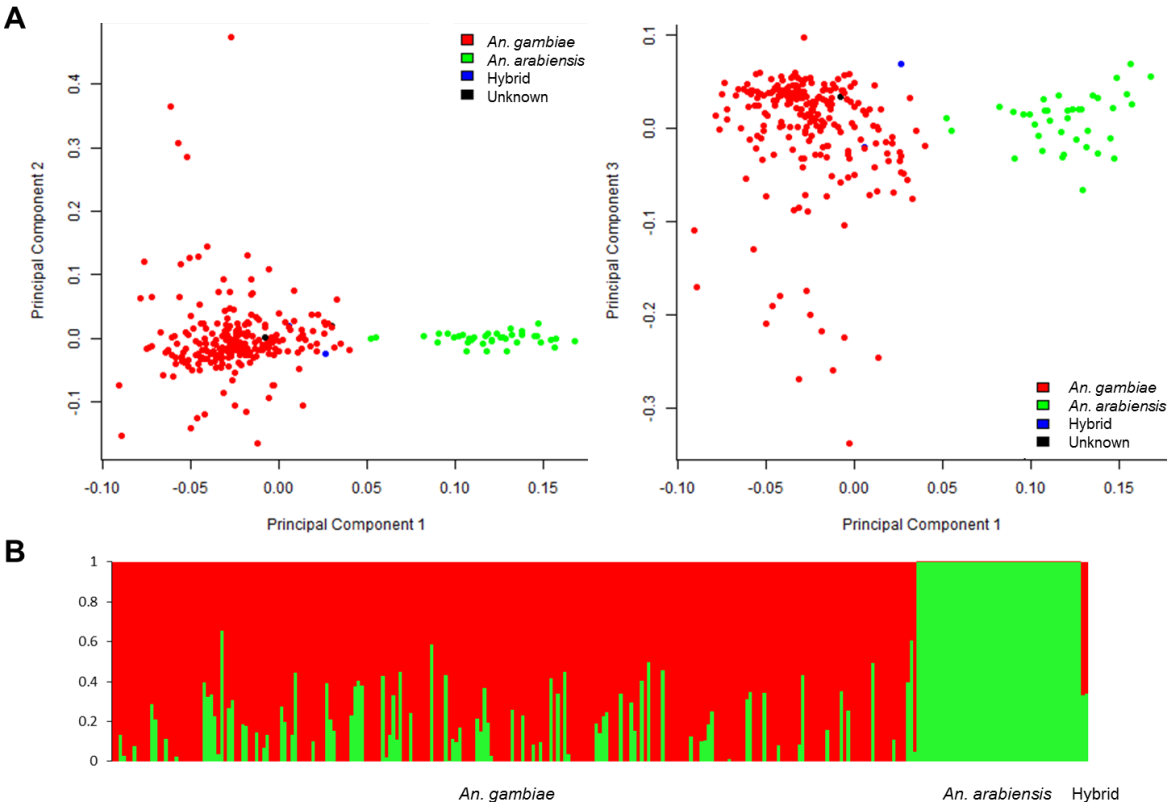

Figure 4.8. Inter-species relationship of mosquitoes from Tororo district, Uganda using 31 autosomal SNPs. A) Principal components analysis showing the first three principal components. *An. arabiensis* samples have lower genetic diversity compared to *An. gambiae* samples, as shown by the first two principal components. B) Results from the program, ADMIXTURE. Individuals were grouped according to their species identity as previously identified using Scott *et al.*'s (1993) PCR procedure. Two samples with unknown species identity were assigned to *An. gambiae* species based on the results of the PCA analysis. Colours represent the inferred ancestry from two ancestral populations (K = 2). Each individual is represented by a vertical bar.

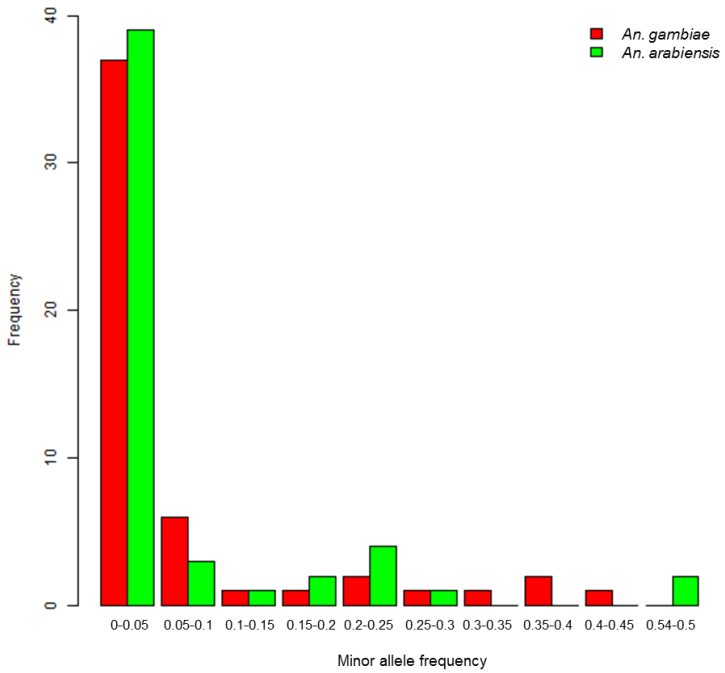

Figure 4.9. Minor allele frequency spectrum of *An. gambiae* and *An. arabiensis* from Uganda.

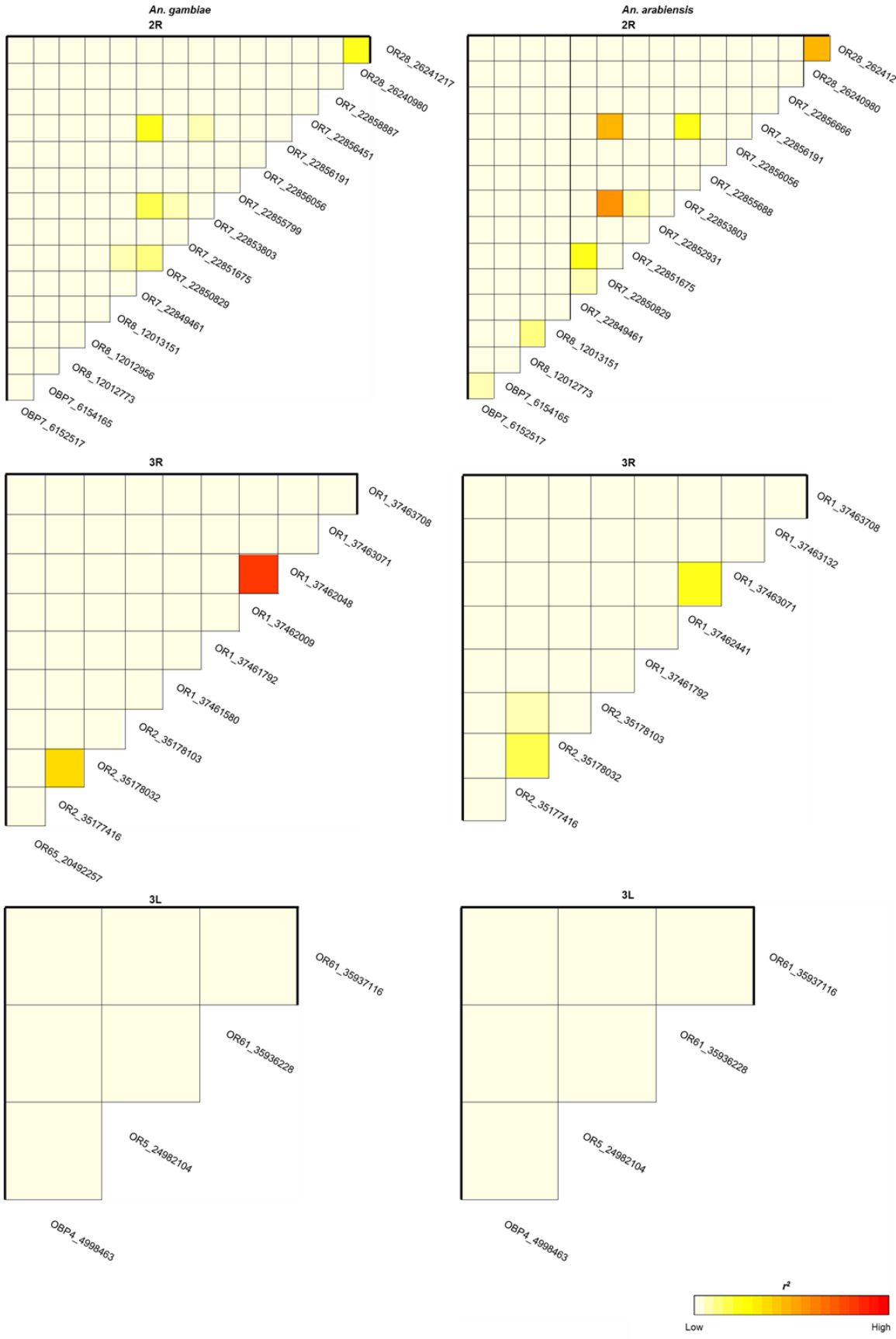

Figure 4.10. Linkage disequilibrium, r^2 estimates between SNP marker pairs in *An. gambiae* and *An. arabiensis* from Uganda. Each square represents pairwise r2 values between SNP markers; r^2 values range from 0 (white) to 1 (red).

Data Summary

Table 4.5 and Table S5 summarises the divergence observed in *An. gambiae s.l.* in the Gambia and Uganda. Divergence between populations (within and between species) appears to increase with phylogenetic distance. Therefore, indoor and outdoor populations of *An. gambiae s.s.* appear the least differentiated while phylogenetically distinct *An. gambiae s.s.* and *An. arabiensis* were the most divergent.

Table 4.5. Pairwise F_{st} estimates of inter- and intra-species comparisons.

Comparison	Pairwise *Fst*
An. coluzzii: indoor vs outdoor	0.001
An. gambiae: indoor vs outdoor	0.003
An. gambiae vs *An. arabiensis*	0.021

Discussion

Host-seeking behaviour is driven by olfactory cues, which are principally perceived by ORs and *OBP*s. Typically, host-biting behaviour in *A. gambiae s.s.* occurs indoors; however, the scale-up of vector interventions has increased the fitness cost associated with this behaviour and potentially, driving some populations to switch towards outdoor-biting behaviour (see Chapter 2). We aimed to describe the inter-species differences between *An. gambiae* and *An. arabiensis,* and investigate if there is a genetic difference between indoor- and outdoor-collected vectors at 13 *OR* and *OBP* genes.

An. arabiensis samples (pass rate = 97%) were reliably typed using *An. gambiae*-derived primers‡, suggesting that there is a substantial amount of gene flow and interbreeding between *An. gambiae* and *An. arabiensis.* Their low divergence (F_{st} = 0.02) coupled with their similar allele spectra further supports this notion. Nevertheless, the two species were distinguishable by PCA and ADMIXTURE analysis because they were highly differentiated at a few loci, particularly at the putative human-odour receptors, *OR1* and *OR7.* We also found modest linkage between *OR7* markers in *An. arabiensis.* Taken together, these results suggest that *OR1* and *OR7* warrant further investigations as candidate genes implicated for mediating the different host preferences of these closely related species.

There were little genetic differences between indoor- and outdoor-collected samples of *An. coluzzii* and *An. gambiae*, as evidenced by the following observations: 1) low F_{st} estimates between indoor and outdoor populations (F_{st} = 0.003), 2) similar allele distribution patterns and 3) no substructure from the PCA

‡See Chapter 5.

and ADMIXTURE analyses. The low genetic divergence observed here is consistent with another study where spatially close populations of *An. gambiae s.s.* in Tanzania (20-50 km apart) exhibited an F_{st} range of 0.003-0.01 (Ng'habi *et al.*, 2011). However in the presence of selection, we expected indoor and outdoor populations to exhibit divergence at candidate loci despite their close proximity (collected 10-50 km apart).

Significant departures from HWE were associated with heterozygote deficiency in *An. coluzzii*. Deficiency in heterozygotes can be attributed to factors such as population sub-structuring, inbreeding, the presence of null alleles and selection. The lack of sub-structuring and the maintained heterozygote deficiency despite pooling of the indoor and outdoor mosquitoes suggest that sub-structuring is an unlikely explanation.

Another possible cause of deficits in heterozygotes is the presence of null alleles, where mutations in one or both primer binding sites prevent amplification during PCR. Null alleles produce either non-amplified genotypes or false homozygotes due to the absence of a second amplicon in heterozygotes, leading to inaccurate alleles frequencies characterised by an excess of homozygotes (Crooks *et al.*, 2013). Although we cannot wholly exclude the presence of null alleles, we consider null alleles an unlikely explanation because of the stringent filters that we applied before data analysis.

In contrast, there is weak evidence for the inbreeding explanation. *An. coluzzii* samples were collected during the start of the dry season, when the likelihood of inbreeding is high due to the limited availability of breeding sites. Further

investigation of LD structure using an increased number of markers would allow us to confirm whether the deficit in heterozygotes was indeed due to inbreeding.

Homozygote excess may also arise because of the presence of segregating chromosomal inversions within a population. Individuals that are heterozygous for an inversion have reduced recombination compared to homozygotes. In turn, reduced recombination can result in adaptive alleles to accumulate, facilitating divergence (Lee *et al.*, 2013). Indeed, chromosomal inversions in *An. gambiae* have played a role in climatic adaptations at the geographic and microhabitat scale (i.e. indoor-and outdoor-resting) (Coluzzi *et al.*, 1977). Deficits in heterozygotes may therefore be partly attributed to the reduced recombination within inversions, as *OR7* is located within the known segregating 2Rb inversion in The Gambia (Bryan *et al.*, 1982).

Using a limited number of markers we were able to successfully identify the divergence between *An. gambiae* and *An. arabiensis.* However we were unable to find evidence of a genetic difference between indoor and outdoor populations of *An. gambiae s.s..* The lack of markers representing each gene may explain the lack of differentiation found in this study especially if the LD decay rate is high, as unlinked markers are poor predictors of the state of non-genotyped SNPs. Another potential explanation is that there may indeed be no genetic differentiation between indoor and outdoor populations.

Performing data analysis using a larger dataset would provide higher-resolution and allow us to conclusively determine if there is a genetic difference between indoor and outdoor populations. Yet, a targeted approach may present similar challenges because the search for divergence is limited to a few candidate genes.

In contrast, a genome-wide approach increases the potential of identifying genes under selection regardless of whether their function was known before (Amos *et al.*, 2010).

The lack of population structure and excess of homozygotes emphasise the need for further investigation using dense SNP markers. Specifically, examination of genes located both within and outside of inverted regions may provide insights into the source of heterozygote deficiency observed in *An. coluzzii*. Finally, adopting an unbiased, genome-wide approach may prove more useful in identifying indoor and outdoor divergence while the genes implicated in host-seeking behaviour remain unresolved.

5. High-Throughput genotyping of *Anopheles* mosquitoes using intact legs by Sequenom iPLEX

Abstract

Recent developments in genotyping technologies coupled with the growing desire to characterise genome variation in *Anopheles* populations opens the opportunity to develop more effective genotyping strategies for high-throughput screening. A major bottleneck of this goal is nucleic acid extraction. Here, we examined the feasibility of using intact portions of a mosquito's leg as sources of template DNA for whole genome amplification (WGA) by Primer-Extension Pre-amplification. We used the Sequenom MassARRAY platform to genotype 78 SNPs for 265 WGA leg samples. We performed nucleic acid extraction on 36 mosquito carcasses and compared the genotype call concordance with their corresponding legs, and observed full concordance. Using three legs instead of one improved genotyping success rates (96% versus 89%, respectively), although this difference was not significant. We provide a proof of concept that WGA reactions can be performed directly on mosquito legs, thereby eliminating the need to extract nucleic acid. This approach is straightforward, sensitive and allows both species determination and genotyping of *Anopheles* mosquitoes to be performed in a high-throughput manner. Our protocol also leaves the mosquito body intact facilitating other experimental analysis to be undertaken on the same sample. Based on our findings, this method would also be suitable for use with other insect species.

Introduction

Mosquitoes transmit diseases of significant medical importance, from malaria to dengue fever and the West Nile virus (Eldridge, 2005). Of these, malaria is the most devastating, causing the death, according to one estimate, of over one million people each year (Murray *et al.*, 2012). Understanding the genome variation of *Anopheles* mosquitoes, the exclusive vectors of human malaria, has important implications in all aspects of *Anopheles* research. This includes improvements on current vector control strategies, aiding taxonomic classifications and resolving cryptic species.

The development of PCR-based genotyping methods has made significant improvements in genetic analysis. Although still widely used, these methods have shown increasing limitations in cost-effectiveness and scope for large-scale analysis. The development of high-throughput genotyping technologies (e.g. Taqman and MALDI-TOF method) constitutes an important step in resolving these limitations (Livak *et al.*, 1995, Haff and Smirnov, 1997). These technologies not only allow the detection of hundreds to thousands of markers but also increases sample throughput by several orders of magnitude. Yet, isolation of genomic DNA from mosquitoes remains a rate-limiting step for high-throughput screening because it is elaborate and time-consuming with limited capacity for processing large numbers of samples. Furthermore, such practice poses additional constraints to researchers who wish to perform a multi-omics study, as it restricts analysis of the same biological sample. While several extraction kits are commercially available to try and bypass this problem, cost and time are yet again important limitations.

We have adapted the standard leg-PCR methodology (Scott *et al.*, 1993) used for species identification of *Anopheles* mosquitoes, to whole genome amplification (WGA) using the Primer-Extension Pre-amplification (PEP) protocol (Zhang *et al.*, 1992). The amplified DNA provides a plentiful resource for high-throughput genotyping (e.g. species identification, mass screening of resistance genes and population genetics) when used in conjunction with the Sequenom MassArray® platform. By bypassing DNA extraction, large numbers of samples can be processed with relative ease.

Material and methods

Ethics statement

Prior to the start of the study, verbal and written consent was obtained from volunteer mosquito collectors. The study protocol was approved by the Uganda National Council of Science and Technology (UNCST; reference HS789).

Sample collection

An. gambiae s.l. mosquitoes were identified morphologically, according to the key of Gillies & Coetzee (1987). Collections were performed in Tororo district, Uganda in April 2013. Mosquitoes were caught by human landing catches and were stored at -20°C in individual tubes containing 80% ethanol.

Sample preparation

Legs were removed from each mosquito and placed into 96-well plates containing 5µl of water per well. We also selected the carcass of 36 mosquitoes to do a full nucleic acid extraction for concordance comparisons. Bodies were homogenised using the DNeasy Blood and Tissue kit (Qiagen, Manchester, UK) according to the manufacturer's standard protocol. The DNA elution step was performed only once using 100µl of RNase-free water. Hereafter, these samples are referred to as carcass samples. DNA concentrations were measured using the PicoGreen assay (Invitrogen, Paisley, UK) and 5µl of gDNA (1ng/µl) were added to each well of a 96-well plate.

Whole-genome amplification

Leg and carcass samples were whole-genome amplified with the following PEP-PCR (Zhang *et al.*, 1992) mastermix: 1.1µl of N15 primers, 2.5µl of 10x PCR buffer (Biotaq, 5U/µl), 1.25µl of MgCl2 (50mM), 0.63µl mixture of dNTPs (2mM

each) and 0.25µl Taq polymerase (Biotaq, 5U/µl), and made up to 20µl with water. The samples were thermocycled as follows: 4 minutes denaturation at 94°C, followed by 50 cycles of 2 minutes at 37°C, 0.1°C/second ramp to 55°C, 4 minutes at 55°C and followed 5 minutes final extension at 72°C. PEP DNA was stored undiluted at -20°C until used.

Species identification

1µl aliquots of the PEP DNA were used in a 20µl PCR reaction for sibling species determination and quality assessment (Scott *et al.*, 1993). The reaction mixture contained primers which create fragments corresponding to *An. gambiae s.s.* (390bp), *An. arabiensis* (315bp) and *An. merus* (466bp). PCR products (10 µl) were run on 2% agarose gel at 100 volts for 2 hours and visualised under UV light. Neat, intense bands were indicative of a successful PCR. The PCR identification was re-run for any failed reactions or *An. gambiae-An. arabiensis* hybrids.

Genotyping

Candidate SNPs were selected based on genes that are currently under investigation by the authors in another research. Assays were designed using the Sequenom SpectroDESIGNER® v3.1.2.5 software according to the manufacturer's instructions and based on the *Anopheles gambiae* PEST reference genome version 3.7 (see URL). Two multiplexes consisting of 78 SNPs (Table S3), located in different chromosome regions, were selected. Genotyping was performed with the Sequenom iPLEX technology following standard protocols (Gabriel *et al.*, 2009), except that 2 µl of 1:10 dilution of PEP per typing reaction was used (MalariaGEN, 2014). Leg and carcass samples were genotyped regardless of their *An. gambiae s.l.* species type.

Assay performance and validation

All data were initially inspected on the Sequenom® Typer v4.2 software to assess assay quality. Any assays with weak signals, positive signals detected in water samples or ambiguous cluster plots were discarded. Next, we used the manual calling option to assign samples that were ambiguous between a homozygous or heterozygous call to a “no call”.

For each SNP, we determined assay performance based on the pass rate across all samples. Pass rates were calculated as the proportion of samples in which a genotype was successfully called. We also calculated the concordance between carcass and leg sample for each SNP. Next, we measured the pass rate for each sample across SNPs that have passed filtering stage. Additionally, we examined sample performance based on three comparisons: 1) carcass versus leg samples, 2) number of legs used and 3) leg samples of *An. gambiae s.s.* versus *An. arabiensis.* A Mann-Whitney U test for non-parametric distributions was performed to determine significant differences between all three comparisons using the software program, R (R Development Core Team, 2010).

Results

Samples

A total of 292 *An. gambiae s.l.* mosquitoes were collected. Of these, 47 were *An. arabiensis*; 231 were *An. gambiae s.s.*; 3 were hybrids between *An. gambiae s.s.* and *An. Arabiensis*; 11 failed to produce PCR amplicons despite repeated attempts to identify their species identity; 27 samples were not genotyped because we could not recover any legs from them. Table 5.1 shows the distribution of the number of legs collected across the samples. In total, 36 carcass and 265 leg samples were genotyped for 78 SNPs.

Table 5.1. Average pass rate according to sample type, number of legs used and species identity.

Group	Number of legs or species	Number of samples	Average pass rate	Pass rate SD	95% CI
carcass	-	36	0.964	0.163	0.911-1.017
Leg	0.5	2	0.969	0	0.969-0.969
	1	36	0.888	0.268	0.801-0.976
	1.5	8	0.977	0.014	0.967-0.987
	3	246	0.957	0.151	0.948-0.975
Species	*An. gambiae s.s.*	231	0.981	0.049	0.975-0.987
	An. arabiensis	47	0.973	0.016	0.968-0.977
	Hybrid	3	0.682	0.496	0.121-1.244
	Unknown	11	0.244	0.358	0.033-0.456

SD = standard deviation, CI = confidence interval.

Assay performance

Of the 78 SNP assays, we discarded 10 because of ambiguous clustering and positive signals detected in the water samples. We excluded an additional 4 SNPs because their pass rate was below the threshold cut-off (<80%). The remaining 64 assays represented our core SNP set which produced an average pass rate of 93.07% (Figure 5.1).

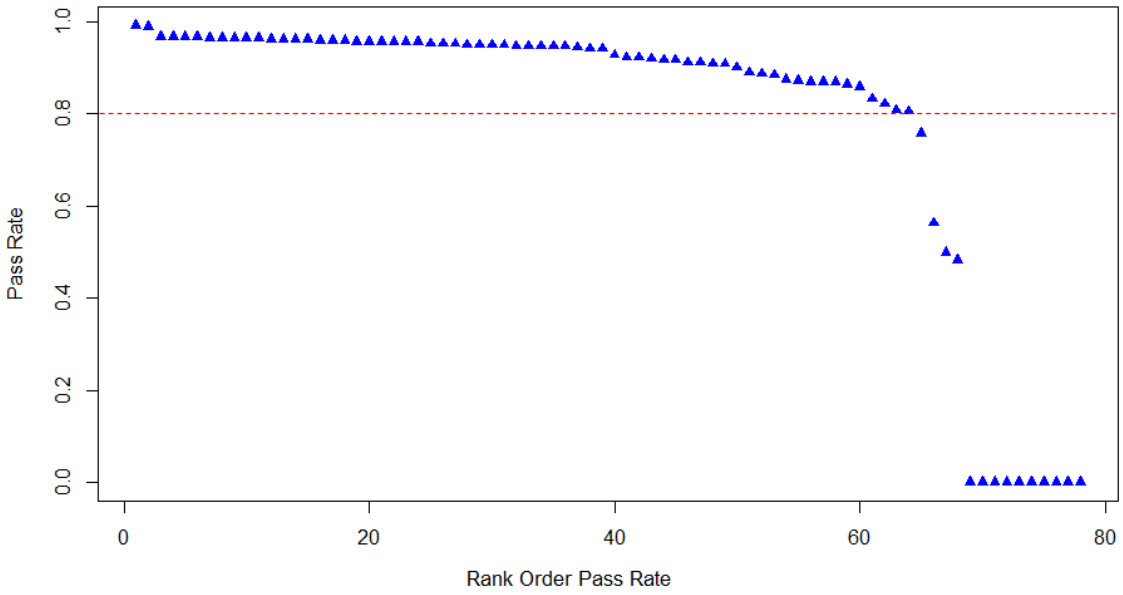

Figure 5.1. Pass rate for each SNP were calculated across all 265 leg and 36 carcass samples. The Y axis represents the pass rate and the X axis represents the rank order of the pass rates. The red horizontal line has been drawn at a pass rate of 0.8 (80%) for a nominal cut-off threshold.

Concordance

We evaluated the accuracy of the genotype calls in the core SNP set by calculating the concordance between 36 carcass samples and their corresponding legs. This gave a total of 2,304 genotypes to compare (64 SNPs x 36 carcass samples), producing an overall concordance of 99.8%. However we note that genotyping failed for one carcass sample, while one leg sample had a considerably lower pass rate than the others (69%). If we exclude both samples, we observe 100% concordance.

Carcass versus leg samples

To avoid confounding effects related to species divergence, we restricted the following comparisons to *An. gambiae s.s.* samples only. On average, leg samples had a higher pass rate (97.87%, 95% CI: 96.13-99.61) than their corresponding carcass samples (96.40%, 95% CI: 91.08-1.00) (Figure 5.2A). We observed that the lower average in carcass samples was caused by one sample that failed to call any genotypes (1.56% pass rate). Once it was excluded, the carcass group

had an improved average pass rate of 99.11% (95% CI: 98.75-99.47). However, we found no significant difference between the two groups (p = 0.81). The remaining leg samples yielded an average pass rate of 98.12% (95% CI: 97.43-98.81).

Number of legs used

Due to sample degradation during storage, it was not possible to obtain the same number of legs for all samples. This resulted in a large variation in the sample size across the legs group, ranging from 2 to 246 samples (Table 5.1; Figure 5.2B). To limit the effects of variance and sample size on pass rate distributions, we restricted comparisons to the 1 and 3 legs groups (36 versus 246 legs, respectively) and found no significant difference (p = 0.18).

Species type

We also examined sample pass rate by species type to detect if there is any bias against *An. arabiensis* samples due to the divergence of *An. arabiensis* from the PEST reference genome. *An. gambiae s.s.* samples had significantly (p = 2.05 x 10^{-6}) higher pass rates with an average of 98.09% (95% CI: 97.45-98.72) compared to 97.31% in *An. arabiensis* (95% CI: 96.84-97.77). However, we observed that the difference between the two groups is small (1% difference in average pass rates), suggesting that *An. arabiensis* samples can be reliably typed using *An. gambiae*-derived primers.

Samples whose species type could not be determined yielded poor pass rates (~8-9%) except for two samples (≥ 95% pass rate) that increased the average pass rate to 24.43% (95% CI: 14.19-34.68). 1 out of 3 hybrid samples had a poor pass rate at 10.94% (Figure 5.3).

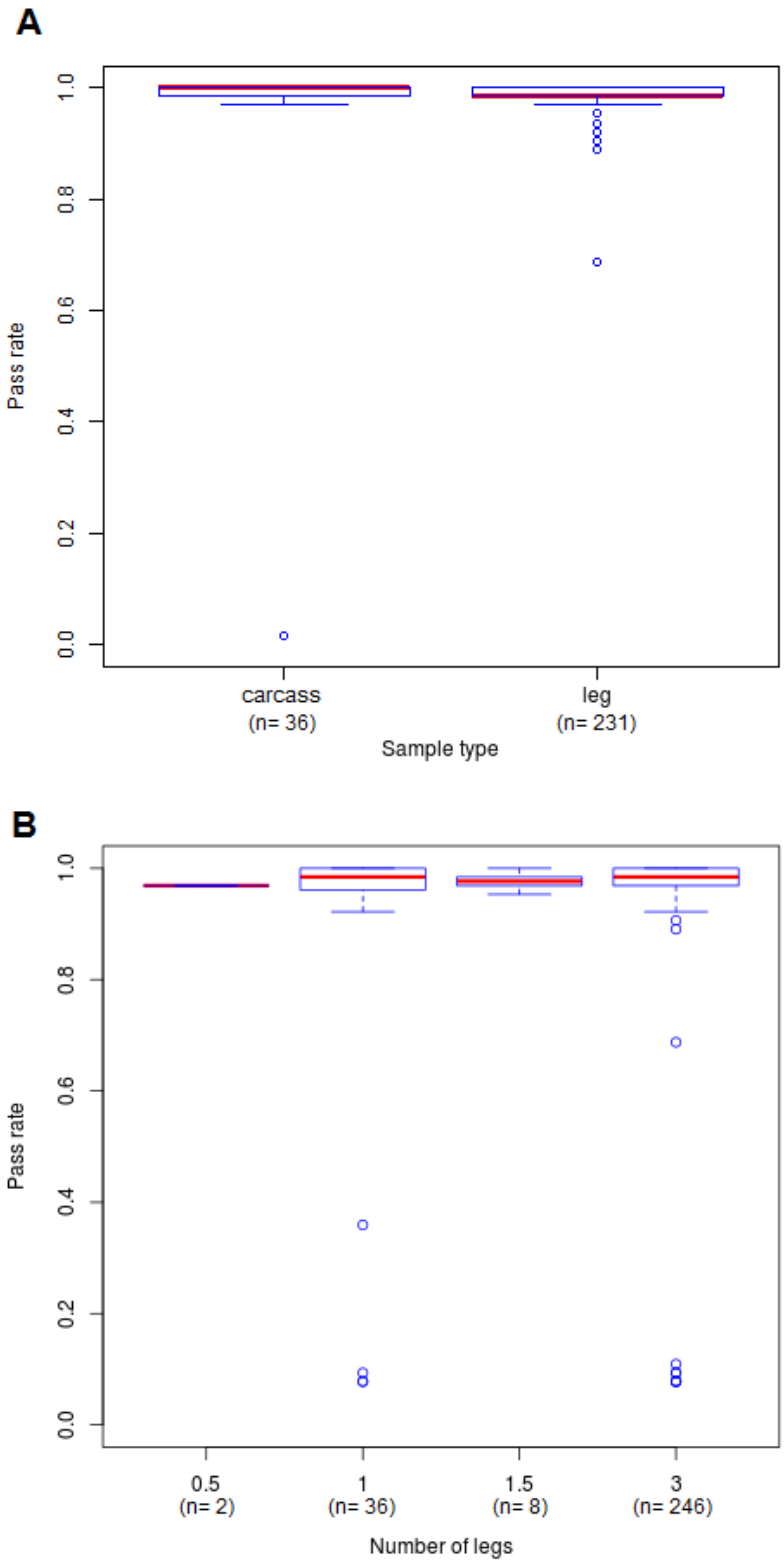

Figure 5.2. Assay performance according to sample type. Legs were used as direct PCR templates whereas DNA extraction was performed on carcass samples. A) 36 carcass versus 231 leg samples. Comparisons were restricted to *An. gambiae s.s.* B) Leg samples (all species were included) were further divided into the number of legs used for WGA. The red line represents the median while the box itself represents the lower and upper quartiles. The pass rate range is represented by the whiskers. Outlier data points are represented by blue circles. The number of samples for each sample type is indicated by "n" in brackets.

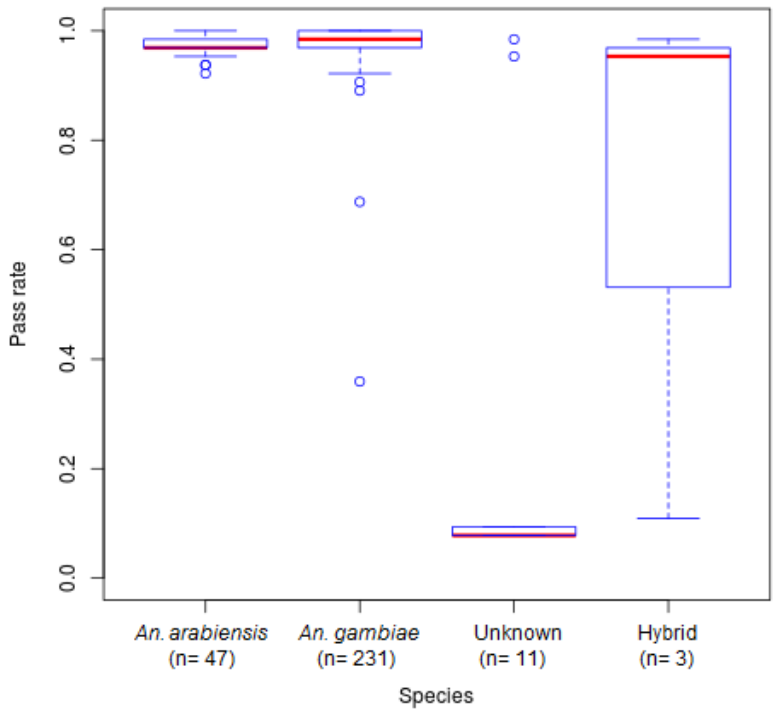

Figure 5.3. Assay performance according to species type. Hybrid= hybrid between *An. gambiae s.s*. and *An. arabiensis*. Samples whose species identity could not be resolved by the PCR procedure were classified as "unknown". The red line represents the median pass rate while the box itself represents the lower and upper quartiles. The whiskers show the pass rate range and outlier data points are represented by blue circles. The number of samples for each sample type is indicated by "n" in brackets.

Discussion

As proof of principle, we extended Scott *et* al.'s (1993) approach by using mosquito legs as sources of DNA for WGA by PEP-PCR and high-throughput genotyping. We found that the PEP DNA from a single leg produced genotype calls suitable for downstream analysis. Indeed, we found full concordance between the genotypes from carcass and their corresponding leg samples where both produced a result. We also found no significant difference between the pass rates of carcass and leg samples.

There was a large variation in the sample sizes, however we were able to compare the 1-leg group (89% pass rate, ±0.27 SD) against the 3-leg group (96% pass rate, ±0.15 SD) and found that using 3 legs achieved greater pass rates. Therefore, we advocate using a minimum of 3 legs to allow sufficient availability of DNA material and highlight the importance of proper storage of samples. We also found that *An. gambiae*-derived primers can reliably type *An. arabiensis* (97.31% pass rate) with great success despite their divergence from the PEST reference genome. However we acknowledge that pass rates may vary substantially according to the area of the genome being targeted. Therefore we encourage that this factor should be taken into account during the assay design stage and when trouble-shooting poor assay performance or assay failure. This also highlights the importance of performing the species identification step which can provide important insights when assessing assay performance.

Usually genotyping of *Anopheles* mosquitoes include the time-consuming process of genomic DNA isolation, followed by a species diagnostic PCR. In contrast, our method bypasses DNA extraction and combines the species diagnostic test with

the PEP quality check in a straightforward workflow. This allows researchers to exclude samples that performed poorly during WGA, as well as those with species identity that are not of interest, before proceeding to the genotyping stage. The mosquito also remains intact thereby allowing other experiments to be performed. This may prove invaluable for preserving limited samples of precious stock material, particularly for research involving vector-parasite interactions where the coordinated analysis of the same biological sample is desired.

A single 25µl PEP reaction generates sufficient material for typing ~3500 SNPs on the Sequenom iPLEX platform, enabling high-throughput genotyping of samples. Indeed, with a sample size of ~300 mosquitoes the turnaround time was approximately 72 hours: 24 hours to prepare the legs and perform the PEP reaction overnight, followed by the species diagnostic test and first-round PCR amplification of PEP DNA (48 hours). Finally, the last 24 hours constitute the last stages of the Sequenom procedure which include reaction clean-ups, iPLEX reaction, chip spotting and measurement. This can easily be scaled-up to larger sample sizes with a processing time that is relatively faster than if DNA extraction was performed. More loci (>2 multiplexes) can also be interrogated with little change to the overall processing time.

As Sequenom assay designs are short PCR amplicons (~100bp), the fragments produced by PEP (≤2kb fragments) (Dean *et al.*, 2001) are more than sufficient to allow amplification of SNPs/regions of interest. With this in mind, other genotyping platforms equivalent to Sequenom (e.g. Illumina GoldenGate platform) may also be considered. However, certain regions of the genome may be difficult to genotype using PEP (Paunio *et al.*, 1996). Thus, our method is not suitable for whole-genome sequencing or downstream applications that require longer DNA

fragments such as Southern blot analysis. In such cases, researchers may attempt WGA using the multiple displacement amplification (MDA) protocol (Dean *et al.*, 2001) instead of PEP, although we have not tested the feasibility of this ourselves. Nevertheless the workflow presented here holds promise for a range of applications, from genotyping small number of SNPs to genome-wide SNP analysis in thousands of samples.

Here we present an easily implemented workflow that allows mass identification of *Anopheles* species and genotyping thousands of SNPs with fast turnaround. Based on our results, it is possible to perform WGA and subsequent genotyping using other fragments of a specimen and other insect species. Our findings have important implications for malaria control where mass screening of field-collected mosquitoes will enable quick surveillance for insecticide resistance, especially in a rapidly evolving environment of vectors and parasites.

6. Genome-wide analysis of indoor and outdoor-biting populations of *Anopheles gambiae s.s.* identifies candidate genes for indoor-outdoor differentiation

Abstract

The development of *Anopheles* vectors towards outdoor-biting behaviour represents an important threat to malaria vector control, which are predominantly indoor-based. Using whole genome deep sequencing, we investigated whether there is a genetic difference between indoor and outdoor-biting populations of *An. coluzzii* collected in The Gambia. Several genes with a predicted role in learning, memory and circadian rhythm were identified as potentially under positive selection on the basis of estimates corresponding to the upper 0.1% of the F_{st} distribution. A number of these genes were proximal to SNPs that were significantly associated with outdoor-biting behaviour (threshold cut-off $p < 10^{-5}$) by a genome-wide association study (GWAS). A follow-up study and meta-analysis involving 177 genotyped SNPs in *An. gambiae* from Uganda identified genes with predicted roles in transcription, catalysis and notably, gustatory receptors, *Gr11* and *Gr12* and odour receptor, *OR1.* Further investigation of these genes is required to confirm their role in biting behaviour.

Introduction

Anopheles gambiae sensu stricto mosquitoes are the major vectors of human malaria in sub-Saharan Africa. The close association of this mosquito species with humans has played a significant role in determining its efficiency in transmitting human malaria parasites. Besides exhibiting a strong preference for human blood, *An. gambiae s.s.* tends to bite inside houses and at times when most people are in bed. These behaviours have been the basis of current vector control methods such as insecticide-treated nets (ITNs), which have substantially relieved the burden of malaria in recent years (WHO, 2013b). However, there are increasing frequent observations that vectors are biting outdoors - underlying limitations in vector control which are predominantly indoor-based (Reddy *et al.*, 2011, Mbogo *et al.*, 1996).

Behavioural traits such as host preference, time of biting and the tendency to feed indoors or outdoors underlie blood-feeding behaviour. These traits are expressed as a consequence of a complex interaction between physiological, environmental and genetic factors (Takken and Verhulst, 2013). However, little is known about the genetic basis of these behaviours. Early research involving selection experiments have shown that host choice can be altered after a few generations, demonstrating that host preference is a heritable trait (Gillies, 1964). Since then, it has been shown that sequence and inversion polymorphisms also play a role in determining resting behaviour (indoor- and outdoor-resting) (Coluzzi *et al.*, 1979, Riehle *et al.*, 2011). Despite these important insights, most of the genes involved in determining behaviour remain unknown. Yet the identification of variants, genes and pathways involved provides an avenue for early-warning surveillance of control measures failing as well as, the development of additional control methods.

Olfactory-related genes such as odour receptors (*ORs*) and odour-binding proteins (*OBPs*) are candidate genes for mediating olfactory-driven biting behaviour. In Chapter 4, we discussed the function of such genes, and identified *ORs* and *OBPs* as genes of interest (GOI) based on their putative role in host seeking behaviour. Our initial analysis found little genetic differences between indoor- and outdoor-biting populations at these genes, although the number of markers available limited the analysis. The development of high-throughput sequencing and genotyping technologies makes it possible to investigate these GOI further using dense SNP markers, as well as perform genome-wide analyses. Unlike the candidate gene approach, genome-wide analyses are unbiased by previous hypotheses concerning candidate genes and pathways, increasing the potential to identify genes under positive selection regardless of whether their function was known before.

Using genome-wide single-nucleotide polymorphisms (SNPs) ascertained from whole-genome sequencing, we investigated whether there is a genetic difference between indoor and outdoor-biting populations of *An. gambiae s.s.* We divided the genome into windows of 10kb to determine the extent of differentiation between the two populations and identified genomic regions of interest based on windows of elevated differentiation. Next, we performed a genome-wide association study (GWAS) to identify SNPs associated with outdoor-biting behaviour. We also described the results at GOI because of the major role that olfaction plays in determining host-seeking behaviour. Finally, we performed a follow-up association analysis using an independent dataset consisting of *An. gambiae s.s.* from Uganda, genotyped at SNPs of interest (SOI) using the Sequenom iPlex technology. We then combined the two datasets in a meta-analysis to identify

SNPs that exhibit significant associations with the outdoor-biting phenotype. Identifying genetic differences between indoor and outdoor populations is crucial for malaria control, which currently relies on *An. gambiae's* indoor-biting behaviour.

Methods

Study design and datasets

During the first stage of genomic analysis, we investigated the populating structure of *An. gambiae s.s.*, and conducted an F_{st} and association analysis using genome-wide SNPs. We described the results at 13 GOI, as well as the genome-wide signals at other loci. We then performed a follow-up study using a dataset that consisted of SNPs of interest (SOI), genotyped in *An. gambiae* from Uganda. Finally, we performed a meta-analysis to combine the results with the Gambian dataset.

Sample collection and preservation

Human landing catches (HLC) from locations 60-200m apart in Wali Kunda, The Gambia, were performed to collect indoor- and outdoor-biting mosquitoes. We performed outdoor collections from 19:00–22:00 hours, the time of night people would normally be exposed to mosquitoes, while indoor collections continued until dawn (19:00-6:00 hours). Following poor outdoor collections, collections were extended to 19:00-7:00 hours for both indoor and outdoor locations.

Mosquito collections from Uganda were conducted in a similar manner as described above, except that HLC was conducted for part of the night only (19:00-24:00). Collections were made in Tororo district, Uganda from locations at least 10 metres apart.

An. gambiae s.l. mosquitoes were identified based on morphology (Gillies and De Meillon, 1968, Gillies and Coetzee, 1987) and preserved in RNAlater, 80% ethanol or dried with silica gel at -20°C prior to transport to the laboratory. Mosquitoes that

had taken a blood meal were kept alive in the laboratory for two days to allow blood meals to digest before preservation*.

Sample preparation†

Briefly, 1-3 legs were removed from each mosquito and placed into 96-well plates. Legs were used as direct DNA templates for sibling species PCR determination using Scott *et al.*'s (1993) standard leg-PCR methodology (outlined below). Once *An. gambiae s.s.* samples were identified, DNA extraction was performed as described below.

DNA extraction

We extracted DNA from *An. gambiae s.s.* mosquitoes using the AllPrep DNA/RNA Micro kit (Qiagen, Manchester, UK), according to the manufacturers' protocols except that DNA was eluted in 100µl of RNase-free water. DNA was stored at -20°C until processing.

Species complex and molecular form identification

Legs were amplified in a 10µl PCR reaction for sibling species determination. The mixture consisted of: 5 µl of REDTaq ReadyMix Reaction Mix (Sigma-Aldrich, USA), 3.4 µl H_20 and 0.4 µl of each primer. Primers create fragments corresponding to *An. gambiae s.s.* (390bp)*, An. arabiensis* (315bp) and *An. merus* (466bp).

Once *An. gambiae s.s.* mosquitoes were identified, they were characterised further to *An. coluzzii* Coetzee (formerly "M" form) and *An. gambiae* (formerly "S" form) species following the PCR methods of Wilkins (2006). Briefly, 1µl of DNA

*Consult Chapter 3 and 5 for further details on the collection site, collection procedure and species identification of samples from the Gambia and Uganda, respectively.
† Chapter 3 provides a more detailed description of the sample preparation procedure.

was added to the following mastermix: 5 µl REDTaq ReadyMix Reaction Mix, 0.4µl of M and S primer (25 pmol/µl each), and 2.4µl of water. Primers create fragments of 427bp and 335bp corresponding to *An. coluzzii* and *An. gambiae* respectively.

10µl of PCR products were run on 2% agarose gel at 100 volts for 2 hours and were visualised under UV light. Samples that failed the PCR reaction were re-run. All *An. gambiae s.s.* samples were submitted to the MalariaGEN (MalariaGEN, 2013) for whole-genome sequencing and genotyping.

DNA sequencing and genotyping

Paired-end sequence reads (~100bp fragments) were generated using the Illumina Genome Analyzer II and HiSeq platforms to a read depth of approximately 30× in genotyped loci. Reads were aligned to the *An. gambiae* PEST reference genome version 3.7 using the BWA algorithm with the default parameters. After sequencing, collaborators from the MalariaGEN performed various steps to identify polymorphic sites (i.e. SNPs and insertions/deletions [indels]), otherwise known as “variants”, and samples were genotyped for these variants. We were therefore provided with a variant call format (VCF) file containing raw, unfiltered variants.

Quality control

Each genotype in the VCF file had meta-information describing the site’s properties such as its mapping quality (MQ) and depth of coverage (DP) across the samples (see section on variant filtering below). Using such information, we filtered the catalogue of variants supplied by the MalariaGEN to remove false positive and produce high quality SNPs. Figure 6.2 provides an outline of the different stages involved in this filtering process. Filters were applied using the

GATK toolkit, PLINK v1.07 and VCFTools v0.1.12 software (Danecek *et al.*, 2011, Purcell *et al.*, 2007, McKenna *et al.*, 2010).

Variant filtering

Variants were filtered in two steps. First, using the GATK's *SelectVariants tool,* we excluded indels and multi-allelic SNPs, keeping bi-allelic SNPs only (McKenna *et al.*, 2010). Next, we excluded low quality SNPs using the GATK's *VariantFiltration tool* with the parameters below. These parameters were based on the GATK's best practice guideline for filtering variants, although some additions were made following the suggestions of the MalariaGEN team. For example, we additionally excluded SNPs located in regions of repetitive DNA as they are likely artefacts from inaccurate read alignment. SNPs matching any of the following conditions were excluded‡:

- Quality by Depth (QD) < 5.0
- Mapping Quality (MQ) < 40.0
- Fisher Strand (FS) > 60.0
- Read Position Rank Sum Test (ReadPosRankSum) < -8.0
- Homopolymer Run (HRun) > 4
- Coverage (DP) < 4000 & > 7000
- Variants within repetitive regions

"Missingness" filter

Potential biases introduced by missing data were accounted for by removing SNPs with high levels of "missingness" (i.e. samples for which many SNPs cannot be genotyped and SNPs for which many samples do not have a genotype) using

‡See Appendix 8: Supplementary notes for more information regarding these parameters.

VCFtools. Similarly, we calculated the proportion of missing data for each sample once SNPs were filtered “for missingness”. Next, we excluded SNPs with data that was missing for more than 5% of the samples, as well as samples with more than 10% of missing data (Figure 6.1). We retained 28,721,881 SNPs and 175/176 samples after this step.

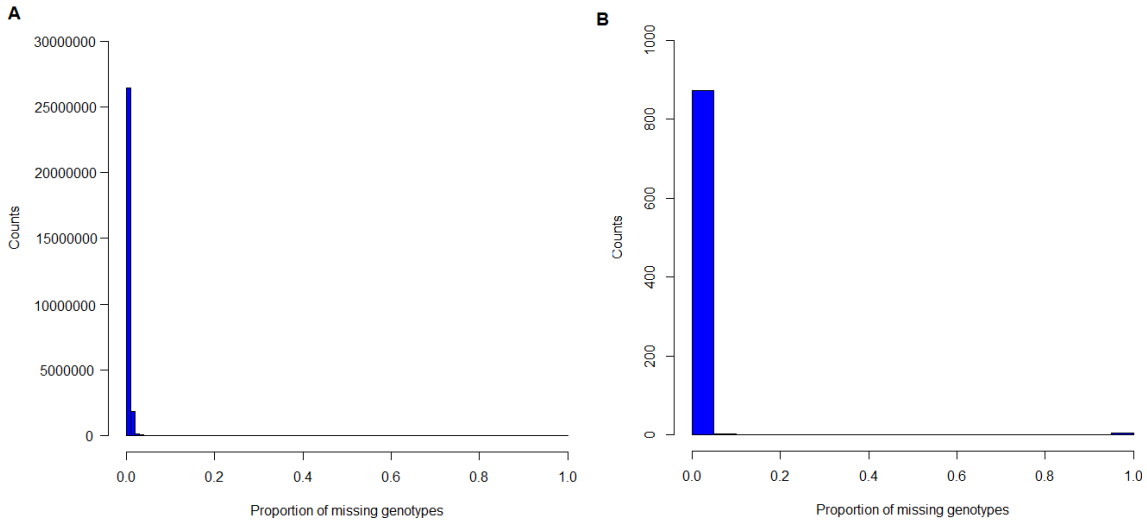

Figure 6.1. Proportion of missing data per SNP (A) and per sample (B) across all chromosomes. The proportion of missing data per sample (B) was calculated after a “missingness” filter was applied to SNPs.

Sample filtering

Additional statistics generated by collaborators from MalariaGEN provided the opportunity to filter samples according to their sequencing quality. Samples that failed any of the following conditions were considered low quality and were therefore excluded from further analysis:

- ≥12X coverage
- ≥ 85% of genome positions covered by at least one read
- >40% mapping rate
- <20% duplicate rate
- <3% mismatch rate

2/175 samples were excluded after this step.

Rare allele filtering

SNPs that have passed the "missingness" and sample filtering step were additionally filtered to exclude rare SNPs using PLINK. Besides having a higher probability than common SNPs of being false positives due to sequencing errors, rare SNPs typically have lower power, requiring a larger sample size or effect size to reflect the effect of the variant. Therefore we excluded SNPs with a minor allele frequency (MAF) of less than 10%, resulting in a final dataset of 3,470,851 SNPs. Hereinafter we will refer to the final SNP dataset simply as the SNP set

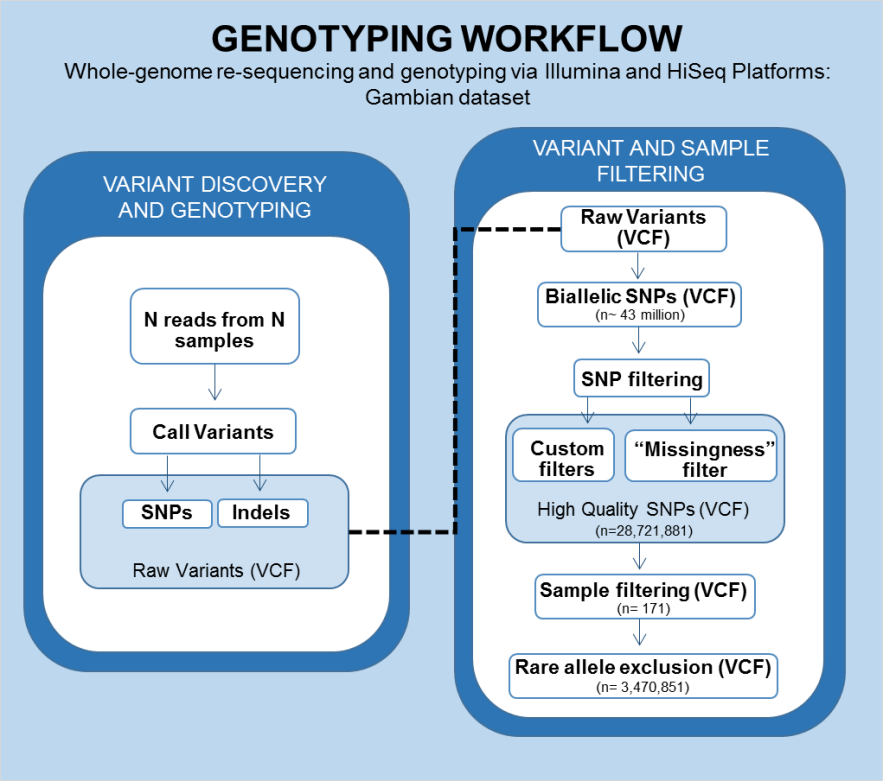

Figure 6.2. Basic workflow of the genotyping and filtering process to produce analysis-ready SNPs. Samples were submitted to the MalariaGEN for sequencing and raw, unfiltered variants such as SNPs and indels were provided in an initial variant call format file (VCF). We then filtered variants and samples in several steps, each generating a separate VCF, to produce high-quality data that is ready for analysis. During the custom filter stage, variants were filtered according to GATK's and MalariaGEN's guidelines, as detailed in the "variant filtering" section. SNPs with a MAF of less than 10% were excluded in the final filter step. The number (n) of SNPs or samples that were retained after each filtering step is also provided.

Annotation

SNPs in the SNPs set were annotated with information regarding their location within the transcript (e.g. coding, intronic) and impact (e.g. synonymous, non-synonymous). Annotation was performed using the program, SnpEff (Cingolani *et al.*, 2012) with the default parameters.

Pruning

Linkage between markers can confound results when performing a principal component analysis (PCA) and estimating population ancestry using the ADMIXTURE program. We used PLINK to produce a dataset containing a set of markers that are likely to be in linkage equilibrium. We used a window size of 500 SNPs, with the window shifted every 100 SNPs. SNP pairs with a correlation coefficient (r^2) greater than 0.1 were removed. 675,613 SNPs from the Gambian dataset were retained from this step; we refer to this set of SNPs as the pruned SNP set.

Genotyping using the Sequenom iPLEX

This section outlines the methods used to genotype samples collected from Uganda, for SOI, using the Sequenom MassArray platform.*

Identification of SNPs of interest (SOI)

SOI were identified based on the Gambian dataset using three approaches: 1) by choosing SNPs found within regions that correspond to the upper 0.1%-5% of the F_{st} distribution, 2) by choosing SNPs with *p-values* less than the threshold cut-off ($<10^{-5}$) from the GWAS and 3) by choosing SNPs within GOI, consisting of 8 odour

*Consult Chapter 3 and 5 for further details of the Sequenom genotyping procedure.

receptors (*ORs*) and 3 odour binding proteins (*OBPs*); these genes were identified previously as genes with a potential role in host-seeking activity†.

Whole-genome amplification and species identification

Briefly, legs were removed from each mosquito and were placed in 96-well plates (Thermo Scientific, UK) for whole-genome amplification using the primer-extension pre-amplication PCR (PEP-PCR) protocol (Zhang *et al.*, 1992). The quality of the PEP reaction was assessed for all samples using the species identification PCR reaction outlined by Scott *et al.* (1993). 1µl of PEP was loaded as a DNA template in a final reaction volume of 10µl. Samples were run on 2% agarose gel to check band intensity and fidelity. Neat, intense bands were indicative of a successful PCR reaction.

Genotyping

Assays were designed using the Sequenom SpectroDESIGNER® v3.1.2.5 software according to the manufacturer's instructions and based on the *An. gambiae* PEST reference genome version 3.7. The initial design stage consisted of 219 SNP assays that were split into seven multiplexes. Of these, we selected five multiplexes consisting of 177 SNPs as a compromise between costs and maximizing the number of SNPs we could type (i.e. >30 SNPs/assay) (Table S6). Genotyping was performed with the Sequenom iPLEX technology following standard protocols (Gabriel *et al.*, 2009) except that we used 2 µl of 1:10 PEP per reaction (Malaria Genomic Epidemiology Network, 2014).

†Details outlining the selection process and function of these GOI are available in Chapter 4.

Assay performance and filtering

The quality of the genotype calls was assessed using the Sequenom® Typer v4.2 software. Assays were discarded based on the following reasons: weak signals, positive signals detected in water samples or ambiguous cluster plots. Next, samples with ambiguous genotype calls (i.e. between a homozygous or heterozygous call) were manually assigned a “no call” genotype.

We included biallelic SNPs only in the analysis. We also excluded SNPs and samples based on “missingness”, however we relaxed the threshold to 80% “missingness” to retain as much data as possible (Figure 6.3). Furthermore, we examined the indoor samples for departures from Hardy-Weinberg equilibrium (HWE) and marked SNPs with *p-values* of less than 1×10^{-7} for potential exclusion. While departures from HWE can be indicative of potential genotyping error, it can also result from true associations. To distinguish between the two factors, we considered HWE for indoor samples only because as a proxy control group, there should be no association with the outdoor-biting phenotype in these samples. SNPs were excluded on the basis that their cluster plots did not show clear clustering of different genotypes (i.e. homozygote AA away from heterozygote AB). Of the 77 SNPs that were retained after filtering for “missingness”, 13 showed large deviations from Hardy-Weinberg equilibrium; however only 8 SNPs were removed from further analysis after inspection of their cluster plots. The final dataset therefore comprised of 69 SNPs.

Figure 6.4 provides an outline of the different stages involved during the filtering procedure of the Ugandan dataset.

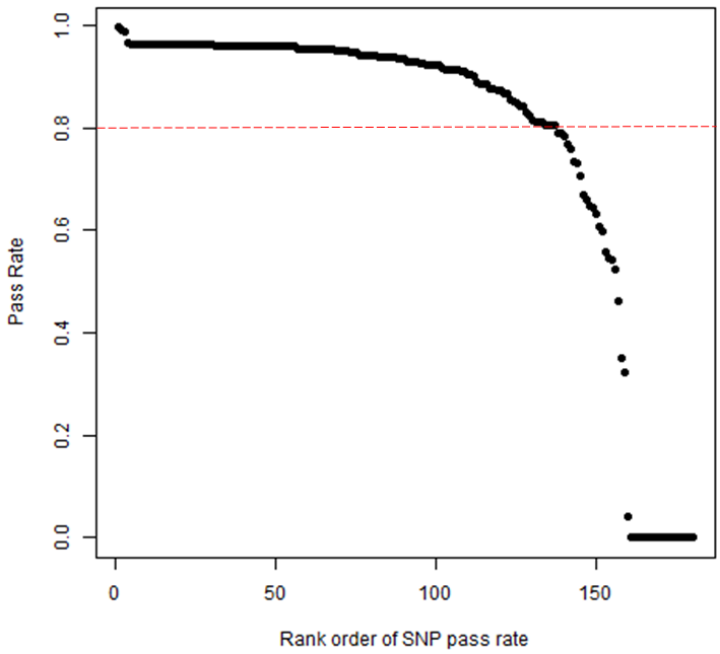

Figure 6.3. Pass rate for the 177 SNPs that were assayed. Pass rate for each SNP were calculated across 242 *An. gambiae* samples from Uganda. The Y axis the pass rate and the X axis represents the rank order of the pass rates. The red horizontal line has been drawn at a pass rate of 0.8 (80%) for a nominal cut-off threshold. red horizontal line has been drawn at a pass rate of 0.8 (80%) for a nominal cut-off threshold.

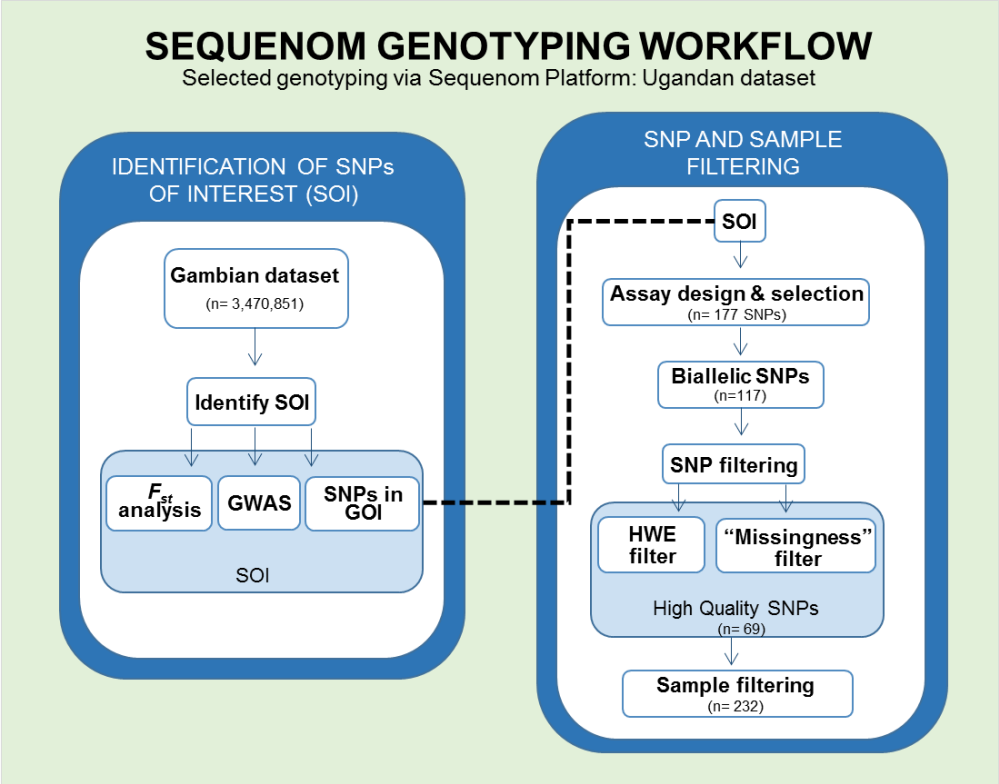

Figure 6.4. Basic workflow of the genotyping and filtering process to produce analysis-ready SNPs in the Ugandan dataset. Samples were genotyped for 177 SNPs of interest (SOI) using the Sequenom iPlex platform. The dataset was then filtered using several criteria such as excluding SNPs and samples with too much missing data, and excluding SNPs that deviate too much from Hardy-Weinberg Equilibrium (HWE) proportions. The number (*n*) of SNPs or samples that were retained after each filtering step is also provided.

Genome analysis

Population structure

To investigate population structure, we performed a principal component analysis (PCA) using the R package, SNPRelate (Zheng *et al.*, 2012). For a given pair of samples, their genetic distance was calculated as the number of SNPs at which the pair of samples had different alleles. The total variance between samples is then partitioned into principal components (PCs), which can be plotted to visualize the genetic variation between the samples. Typically, the first three PCs sufficiently explain most of the variation and relationship between samples because each succeeding PC accounts for less variability in the data.

To confirm the results of the PCA, we investigated population structure using a maximum likelihood clustering approach implemented by ADMIXTURE v1.23 program (Alexander *et al.*, 2009). ADMIXTURE, divides an individual's genome into groups corresponding to potential subpopulations (*K*) with different allele frequency profiles. Using this approach, the best model estimate of the number of ancestral populations can be estimated by choosing the *K* with the lowest cross-validation error. To run the ADMIXTURE program, we simulated 1 to 5 subpopulations (*K* = 1-5), *K* using 10 cross-validation procedures for each *K.* We used the pruned SNP set* for both PCA and ADMIXTURE analysis.

Population differentiation

To estimate the extent of genetic differentiation between indoor and outdoor samples, we calculated the unbiased estimator of population differentiation, F_{st} (Weir and Cockerham, 1984). F_{st} compares the variance of an allele frequency

*See section titled "Pruning".

over subpopulations to its variance in the total population. An F_{st} value of 1 suggests complete differentiation between two populations while zero estimates suggest panmixia - one interbreeding population.

We calculated F_{st} in non-overlapping windows of 10kb for each chromosome arm, using VCFtools. To identify highly differentiated regions, we first excluded windows with less than 50 SNPs to limit bias introduced by low SNP density. We then identified windows with F_{st} estimates that corresponded to the upper 0.1%, 1%, 2.5%, 5% and 10% of the distribution. We considered these windows as regions of interest, representing regions that are differentiating between the two populations and thus, potentially, under positive selection.

Inversions

Segregating chromosomal inversions were identified by visualising linkage between markers. If our samples were segregating for a specific inversion, we expect to see high linkage between markers within the inversion site. To decrease the computation time, we reduced the dataset to 1% of the total SNPs per chromosome arm and calculated the correlation coefficients, r^2, between the retained markers. Both the r^2 estimates and triangle plots of linkage disequilibrium were generated using R packages, LDheatmap (Shin *et al.*, 2006) and SNPRelate.

We characterised the inversion state of each sample by performing a PCA using SNPs located within the inverted regions. Each sample can have one of three possible inversion states: 1) “standard” arrangement, corresponding to a state without any inverted segments and denoted by a “+” sign (e.g. $2L^{+a}$ for inversion 2La), 2) “alternative” arrangement, corresponding to a state where both

chromosomal segments are in the reverse orientation and 3) “heterozygous” arrangement which corresponds to a sample that has both the alternative and standard arrangement (e.g. $2L^{+a}/a$).

Limited recombination between inverted and non-inverted segments cause these segments to evolve independently, creating an effect that is similar to population substructure. The three different inversion states are therefore easily identifiable from a PCA because it projects the samples into three equidistant clusters, with the middle cluster corresponding to heterozygous samples while the flanking clusters correspond to the two homokaryotypic states. To verify our findings, we conducted an F_{st} analysis to compare the level of differentiation between the three inversion states. We expected the highest differentiation between samples carrying the standard and alternative arrangements upon correct assignment.

We used Brown’s (1970) formula, to determine if inversion frequencies deviated from HWE. *F* values greater than zero indicate a lack of heterozygotes while values less than zero indicate an excess of heterozygotes.

$$F = \frac{4a_1a_3 - a_2^2}{(2a_1 + a_2)(2a_3 + a_2)}$$

a_1 *and* a_3 = frequencies of the two homozygote classes
a_2 = frequency of heterozygotes
$F > 1.96/\sqrt{N}$ = significant departures from HWE; *N* = sample size

Association analysis

SNPs were tested for association with outdoor-biting behaviour using a logistic regression model under three different models of inheritance: additive, dominant and recessive. Under the model of additivity, having two copies of the minor allele (AA genotype) has twice the effect of having a single copy of the minor allele (Aa

genotype). In other words, the tendency to bite outdoors is greater for samples with two copies of the minor allele, while heterozygotes will exhibit a phenotype that lies between those of the two homozygotes. In contrast, the dominant model assumes that outdoor-biting behaviour is expressed as long as there is at least one copy of the minor allele. Samples are therefore grouped based on whether they have the minor allele (AA or Aa) or not (aa). Finally, the recessive model assumes that samples must have two copies of the minor allele (AA versus [Aa or aa]).

For the Gambian dataset, we performed a genome-wide association study (GWAS) using the SNP set and chose, a cut-off threshold of $p < 10^{-5}$. We also calculated the genomic inflation factor (λ) using the default parameters of the "estlambda" function from the R package, GenABEL (Aulchenko *et al.*, 2007). λ represents the level of inflation of the association test statistic due to differences in allele frequencies that can arise from population stratification, cryptic relatedness and genotyping errors. Generally, λ estimates of 1.1 are considered acceptable in a GWAS, while estimates above this indicate an excess of false positive results and thus, will require further correction (Yang *et al.*, 2011). The Ugandan dataset was subjected to the same association analysis as the Gambian dataset, except that the test was applied at SOI only.

Meta-analysis

The results of the Gambian and Ugandan dataset were combined using a custom R script to perform a weighted-linear combination of log odds ratios under a fixed-effect model. This method weights the effect size estimates (β coefficients) from each dataset, by their standard error estimates and calculates an overall effect size, standard error and *p-value.*

Results

Gambian dataset

Samples

A total of 496 *An. gambiae s.l.* were collected from The Gambia and of these, 311 were *An. arabiensis,* 183 were *An. coluzzii* and 2 were *An. gambiae.*

Out of the 185 *An. coluzzi* and *An. gambiae* that we submitted for sequencing, 12 were excluded for the following reasons: 3 had insufficient DNA available, 6 were sequenced poorly (i.e. poor reference alignment), 1 had too much missing data (>5% "missingness"), 2 had low quality sequence data (i.e. low coverage, mapping rate, etc.; see methods). To avoid additional noise introduced by the divergence between *An. coluzzii* and *An. gambiae*, we excluded *An. gambiae* samples following an initial principal component analysis which identified them as outliers (Figure 6.5). This left 171 samples for analysis, 104 of which were collected indoors while 67 were collected outdoors.

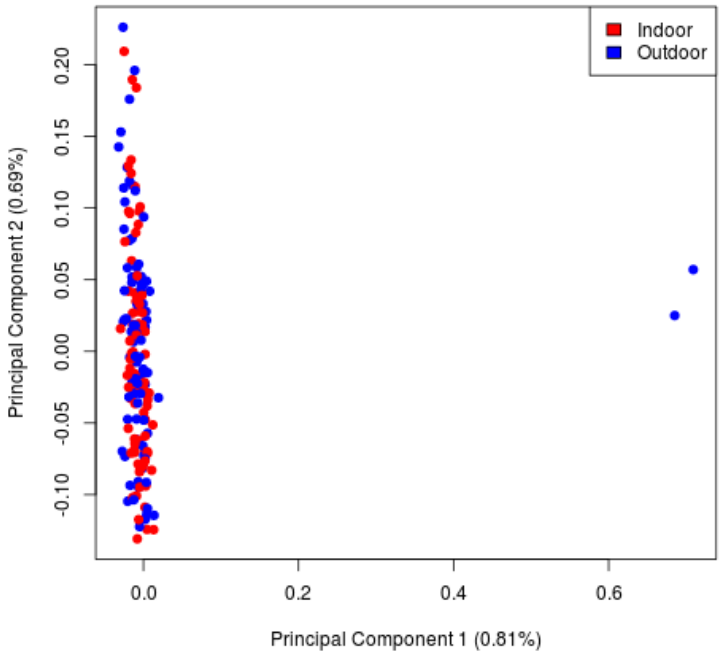

Figure 6.5. Principal component analysis (PCA) of 173 *An. coluzzii* and *An. gambiae* from The Gambia. The PCA distinguished *An. gambiae* as outliers from *An. coluzzii* along PC1 and were therefore excluded from the analysis. Each solid circle represents an individual.

Summary of genome data

After quality filtering, we obtained a final dataset of 3,470,851 SNPs. This SNP set represents approximately 8% of the original SNPs provided by the MalariaGEN (Table 6.1). The largest group of SNPs (38.72%) was found within intergenic regions while only 4.84% of SNPs were found within exonic regions (Figure 6.6). 23% of the exonic mutations were missense (change results in an altered amino acid sequence) while the remainder (~77%) were silent (the mutation does not result in a change of amino acid).

Table 6.1. **Summary of SNP data after filtering.**

Chr	# SNPs	# SNPs after filtering	Mean coverage/sample	Mean # of SNPs/kb
2L	9,810,426	733,487	29.04	14.86
2R	10,441,479	1,017,418	29.91	16.54
3L	7,998,621	647,119	29.50	15.44
3R	11,244,017	879,572	29.50	16.53
X	3,695,482	193,255	30.46	7.93
Total	43,190,025	3,470,851		

The data represents information across 171 *An. coluzzii* samples.

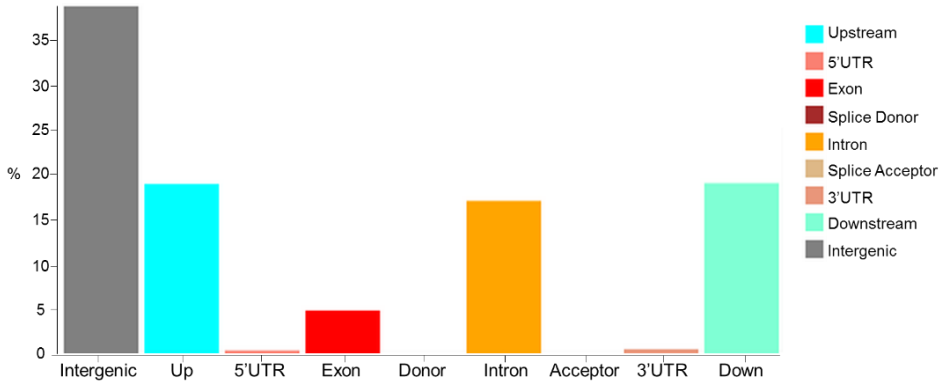

Figure 6.6. Functional annotation of the SNP set according to the location of the genome that they affect. Over 90% of the SNPs affect non-coding regions. Non-coding regions correspond to: intergenic regions, intronics, 5' and 3'UTR, and splice donor and acceptor sites. Upstream and downstream refers to a variant that is within 5 kb upstream or downstream of a gene. UTR = Untranslated region.

Ugandan dataset

A total of 177 SOI were genotyped in 242 *An. gambiae* from Uganda (231 *An. gambiae* plus 11 samples with unknown species identity but showed clustering with *An. gambiae* from the PCA analysis[†]), using the Sequenom iPlex platform. During filtering, 100 SNPs were excluded because 60 SNPs were monomorphic while 40 SNPs had too much missing data (>20% "missingness"). An additional 8 SNPs could not be included in the analysis because they were monomorphic in either indoor or outdoor samples. Furthermore, 10 out of 242 samples were excluded because they had too much missing data. After quality filtering, the Ugandan dataset comprised of 147 indoor and 85 outdoor *An. gambiae* that were genotyped for 59 SNPs.

Population structure

To investigate the relationship between population structure and site of collection, we carried out a PCA using the pruned SNP set. The analysis did not reveal any distinct sub-clusters, however we found that the first three principal components accounted for only ~2% of the total variance. This is likely explained by the fact that the genetic variation across all the samples was large, producing weak correlations (Figure 6.7). Consistent with the results of the PCA, the results of the ADMIXTURE program suggested that a single population was the optimal model of ancestral groups (Figure S11). Under a model of two sub-populations (K = 2), we found that the samples were assigned to subpopulations that were not associated with collection site (i.e. indoor/outdoor). Additionally, we found that higher values of K did not reveal any meaningful ancestries (Figure 6.8).

[†] See Chapter 4, Figure 4.8

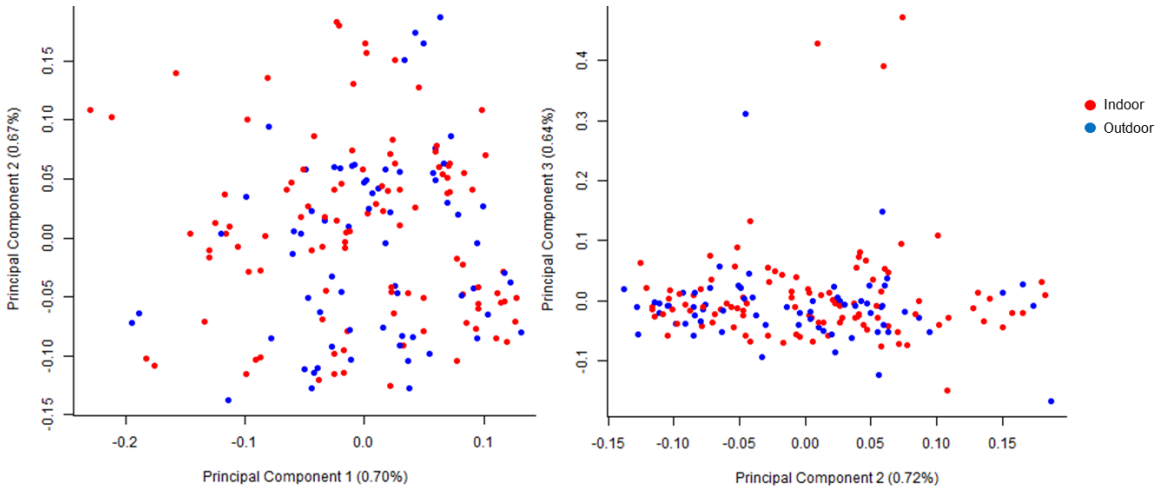

Figure 6.7. Principal component analysis (PCA) of indoor- and outdoor-collected *An. coluzzii* from The Gambia. Plots of the first three principal components (PC) and the percentage accounted for by each PC, using 675, 613 SNPs that were selected to minimise the confounding effects of linkage between markers. Each solid circle represents an individual and the colour is assigned according to their collection site.

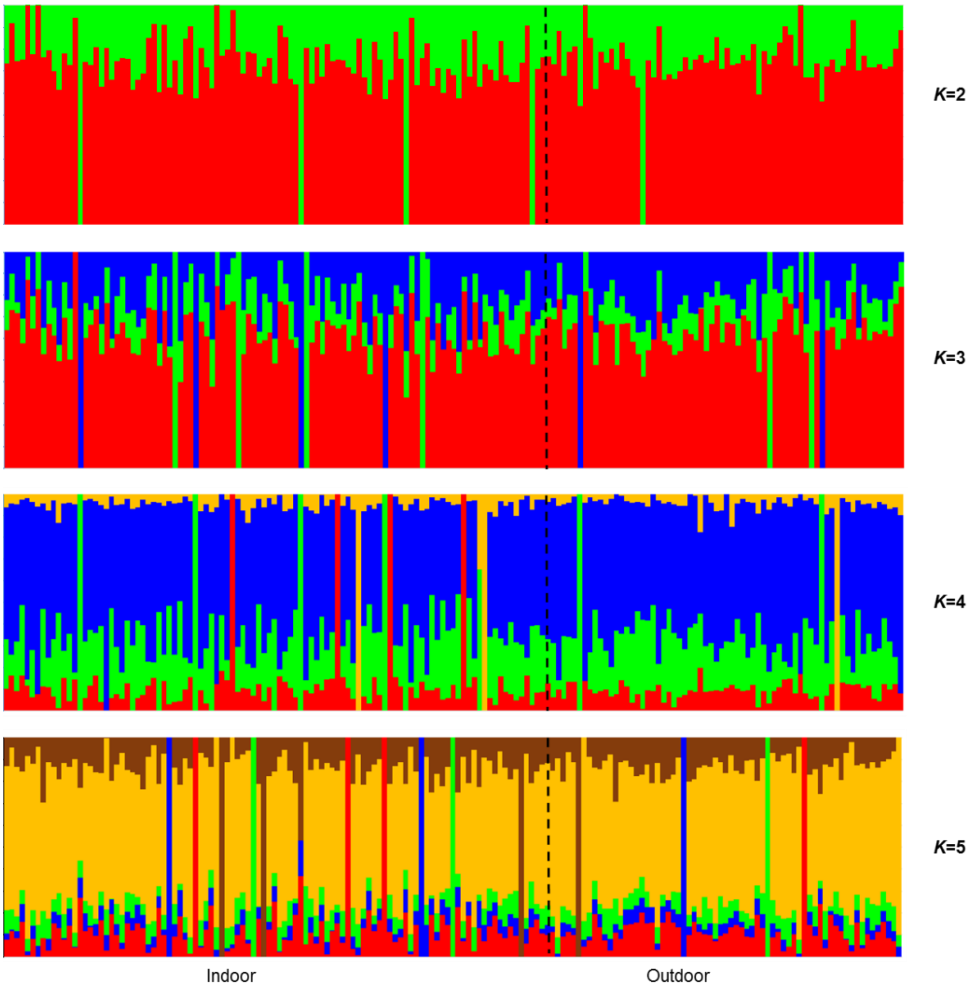

Figure 6.8. Population structure inferred from ADMIXTURE analysis. Each sample is represented by a vertical stacked column. Under a model of two to five subpopulations (*K* = 2-5), each sample is partitioned into *K* coloured segments that represents an individual's estimated membership fractions into *K* subpopulation.

Segregating inversions

High linkage between the markers, as shown in the LD plots, indicated that the inversions 2La, 2Rb and 2Rd were segregating in *An. coluzzii* (Figure 6.9). For each inversion, we used SNPs located within the inverted regions and performed a PCA and F_{st} analysis to compare the divergence between the three inversions states. The analyses showed clear clustering and high divergence between inversion states (Figure 6.10 & Figure 6.11). We found that divergence was highest when we compared the two homokaryotypic clusters for all three inversions (F_{st} = 0.54-0.78), as expected from two segments that are evolving independently because of limited recombination. Furthermore, divergence in the rearranged regions was higher than in collinear regions, as expected from a region with suppressed recombination.

Inversion polymorphisms on the second chromosome have been correlated with adaptation to arid environments, resting and feeding behaviour (Coluzzi *et al.*, 1979, Coluzzi *et al.*, 1985). Across our samples, we found that the inverted arrangement was rare in 2Rd but was at high frequency in 2Rb and 2La. However these frequencies did not deviate significantly from HWE (F <1.96/√N, p >0.05; Table 6.2), nor did we find a significant difference (x^2 = 0.58-3.62, df = 2, p >0.05) in the inversion frequencies of all three inversions between indoor and outdoor populations. Taken together, these findings suggest that that these inversions do not play a role in determining indoor- and outdoor-biting behaviour.

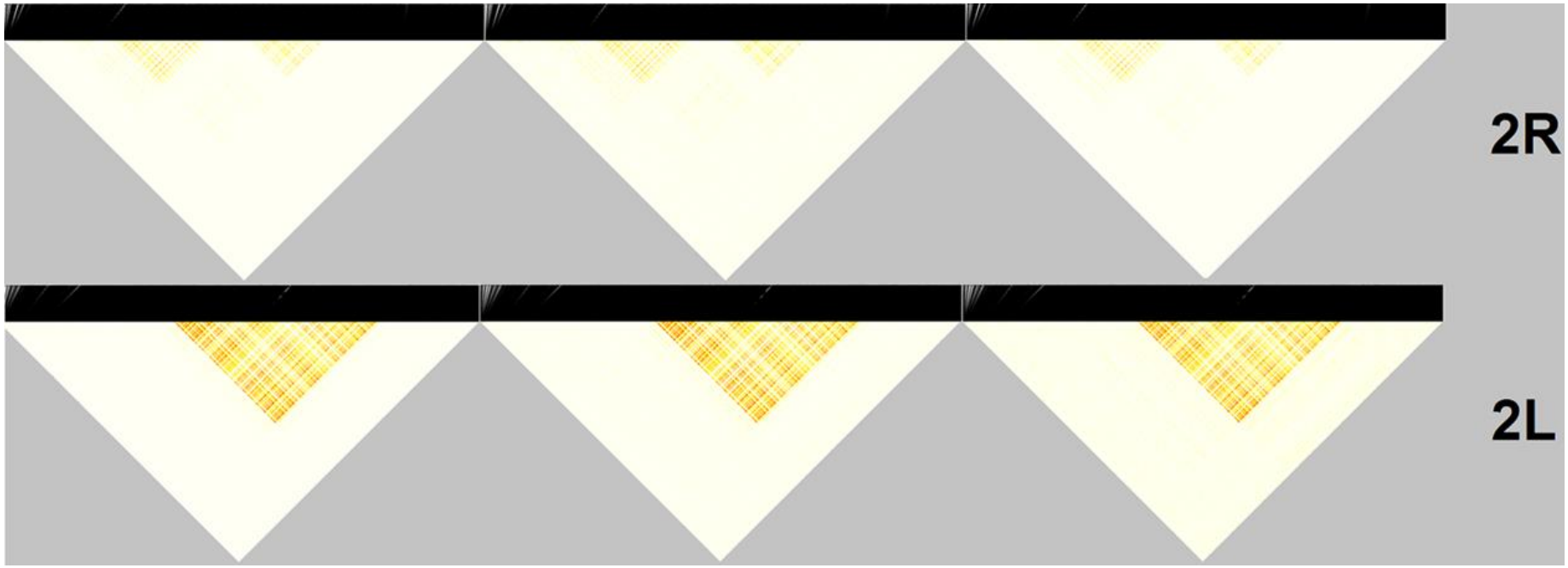

Figure 6.9. Triangle plot of genome-wide linkage disequilibrium (LD) in indoor (left), outdoor (centre) and combined (right) samples of *An. coluzzii* in chromosomes, 2L and 2R. SNPs were aligned along the bottom of the triangle plot and according to their chromosomal position. To measure LD, the squared correlation coefficient (r^2) estimates are calculated for a given pair of SNPs and shown as a diamond, at the intersection of the diagonals from each SNP. r^2 estimates are represented by a colour scale, ranging from yellow (no linkage, $r^2 = 0$) to red (complete linkage, $r^2 = 1$). Strong linkage between SNPs produces a patch of high intensity yellow/red at regions where inversions 2La, 2Rb and 2Rd occur. The plots were produced using 1% of the SNP dataset.

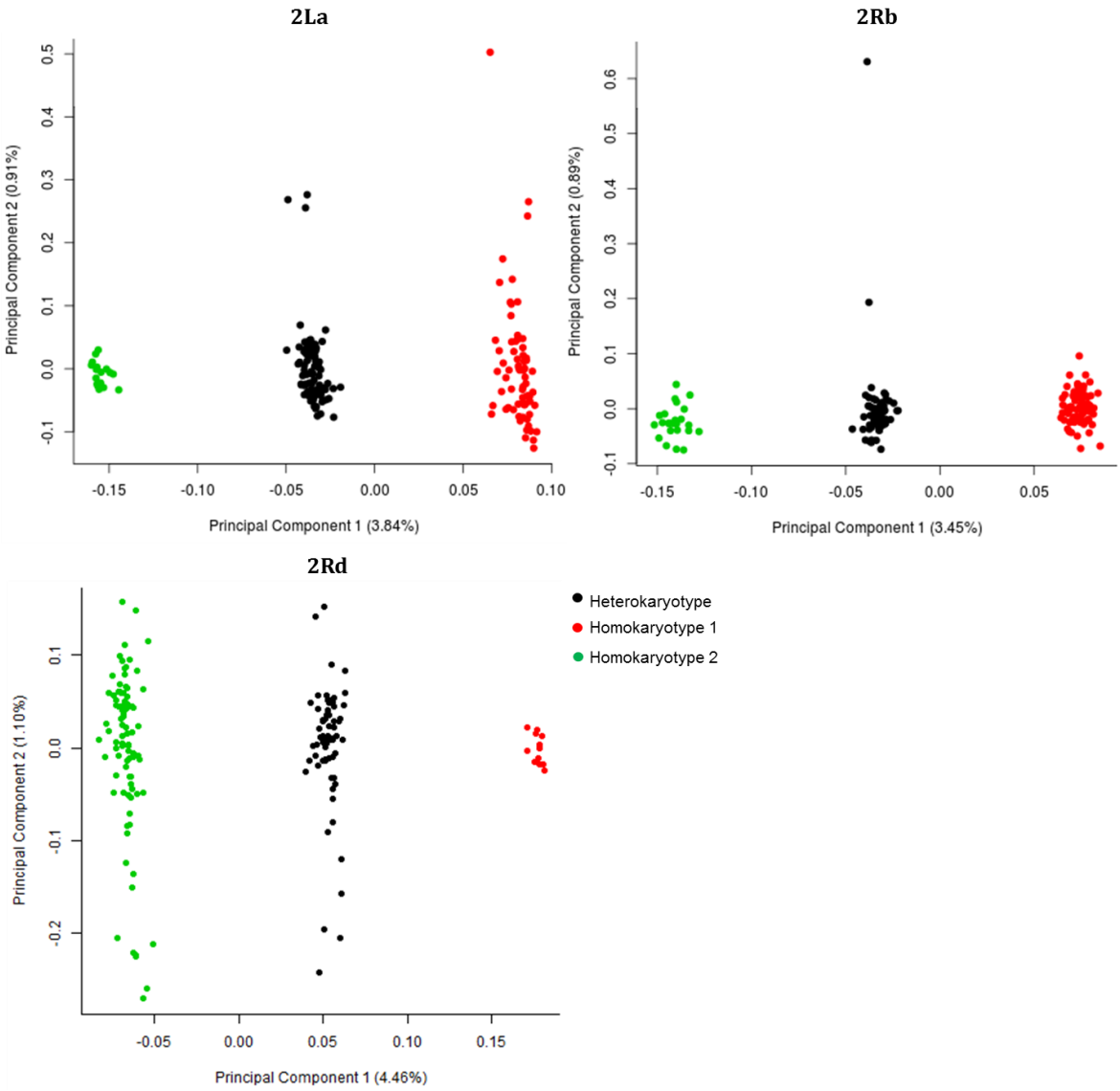

Figure 6.10. PCA plot using SNPs located within 2La, 2Rb and 2Rd inversions. Approximately 14,000 SNPs were used for the PCA analysis of 2La and 2Rb while ~6,000 SNPs were used for 2Rd. Homokaryotypes 1 and 2 correspond to samples that are homozygous for either the standard or alternative arrangements while heterokaryotypic samples carry both the alternative and standard arrangement.

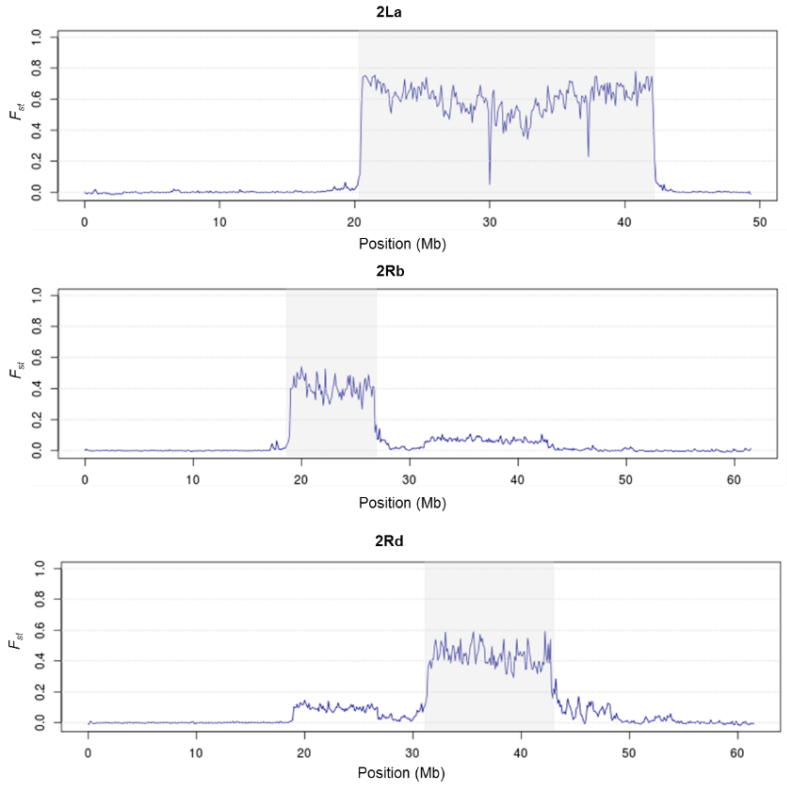

Figure 6.11. Non-overlapping sliding window analysis (10kb window) of F_{st} statistics between homokaryotypic samples in 2La (top), 2Rb (centre) and 2Rd (bottom). Grey shaded areas correspond to regions of inversion. For all three inversions, high divergence between the homokaryotypic samples was restricted to the site of inversions.

Table 6.2. Proportion of samples carrying the inverted arrangement for inversions 2La, 2Rb and 2Rd *in An. gambiae s.s.*

Group	N	1.96/ √N	2Rb (*F*)	2Rd (*F*)	2La (*F*)
Indoor	104	0.19	0.47 (0.17)	0.11 (0.10)	0.38 (-0.07)
Outdoor	67	0.24	0.43 (0.05)	0.03 (-0.19)	0.46 (-0.09)
All	171	0.15	0.46 (0.12)	0.08 (-0.01)	0.41 (-0.07)

F is a measure of HWE deviation and is enclosed in brackets. Value is significant (*) if *F*> 1.96/√N.

Estimates of genetic differentiation

Overall, we found that divergence was limited between the two populations (F_{st} = 0.00004) (Figure 6.12), with the highest level of differentiation at chromosome 2L. Differentiation was also limited at windows containing GOI (F_{st} = -8.87 x 10^{-4}-0.003) (Table 6.3). The highest differentiation was at *OBP47* and *OR65* with F_{st} estimates of 0.003 and 0.001, respectively. For both genes, the estimates were on the borderline of the upper 10% of the F_{st} distribution. With the

exception of *OR7*, these GOI were represented by a small number of SNPs (*n* = 1-77) relative to the size of the window therefore, it is not surprising that the F_{st} estimates were low.

We also identified regions that corresponded to the upper 0.1%-10% of the F_{st} distribution as regions under possible positive selection. However, we identified too many regions (~1800 regions) and so, we focused on regions that corresponded to the upper 0.1% of the distribution instead. The analysis identified a total of 22 regions containing 18 genes, with F_{st} estimates ranging from 0.012 to 0.021 (Table 6.4). For all except three genes that have no functions assigned (AGAP009801, AGAP028088, AGAP006526), it was possible to identify a biological relevance to biting behaviour. For example, *Gr11* and *Gr12* are gustatory receptors (*GR*s) that enable mosquitoes to detect chemical signals important for feeding and avoidance behaviour (Sparks *et al.*, 2013). AGAP003820 is a *Lin-10* homolog which has a predicted role in localising glutamate receptors (*GLRs*) (Rongo *et al.*, 1998). In the fruit fly, *Drosophila melanogaster, GLRs* function as chemosensory receptors (Benton *et al.*, 2009). If this function is conserved in *Anopheles* then it is possible that polymorphisms in *Lin-10* alter chemosensory-dependent behaviour such as biting behaviour.

The analysis also identified three genes (AGAP007717, *Beat IIb* and *Beat*) with predicted roles in mediating behavioural changes driven by learning and memory. AGAP007717 and *Beat IIb* (AGAP004369) are involved in cell adhesion and actin filament organisation. Research in *Drosophila* showed that *fasII*, a cell adhesion molecule (*CAM*), partly underlies olfactory learning and memory. Short-term memory defects as well as lost sensitivity to ethanol, driven by altered synapse plasticity and neural circuits, were among the observed effects of mutations in *fasII*

(Cheng *et al.*, 2001). Since then, several other *CAMs* have been implicated as mediators of synaptic rearrangements that can lead to altered behaviour (Murase and Schuman, 1999). In contrast, *Beat* proteins promote axon separation by acting as a competitive inhibitor of *CAMs,* thereby regulating neural development (Pipes *et al.*, 2001). Taken together, there is evidence to suggest that synaptic changes mediated by these genes might facilitate indoor- and outdoor-biting tendency.

We also identified genes involved in circadian rhythm. AGAP000627 is a casein kinase I whose function involves phosphorylating proteins. In *Drosophila, doubletime*, a casein kinase I, is a regulator of circadian locomotor activity (Williams and Sehgal, 2001), it is therefore possible that AGAP000627 plays a similar role as a regulator of circadian rhythm in *Anopheles*. Similarly, the *TubulinB* gene, a cytoskeletal component in *Anopheles* has exhibited up-regulated expression upon light stimulation (Das and Dimopoulos, 2008). Genes exhibiting rhythmic expression are interesting candidates given that the host-seeking activity of *Anopheles* is nocturnal. Other genes identified were AGAP004795, AGAP004797, AGAP004798, AGAP001370, AGAP003821 and AGAP004644 which play a role in transcription regulation, protein trafficking, signal transduction and proteolysis. Such genes might be involved in the downstream signalling that determines biting behaviour.

Considering that these 18 genes were found within the upper 0.1% of the F_{st} distribution, they represent interesting candidates that warrant further investigation to help identify the mechanisms underlying biting behaviour.

Table 6.3. F_{st} estimates of regions containing genes of interest (GOI).

Gene	Vectorbase ID	Chr	Gene Size (bp)	Region (Mb)	# SNPs in window	F_{st}
OBP47	AGAP007287	2L	820	45.01-45.02	188	3.03×10^{-3}
OR8	AGAP001912	2R	1,374	12.01-12.02	87	-2.39×10^{-3}
OR7	AGAP002560	2R	9,398	22.84-22.85	283	-3.09×10^{-3}
OR28	AGAP002722	2R	1,820	26.23-26.25	473	-3.27×10^{-3}
OBP22	AGAP010409	3L	558	2.85-2.86	96	-1.78×10^{-3}
OBP4	AGAP010489	3L	1,052	4.99-4.5	24	-3.09×10^{-3}
OR5	AGAP011467	3L	1,492	24.98-24.99	303	-3.08×10^{-3}
OR61	AGAP011991	3L	1,402	35.93-35.94	243	5.54×10^{-5}
OBP26	AGAP012321	3L	661	40.21-40.22	224	-8.87×10^{-4}
OR65	AGAP008894	3R	2,545	20.49-20.5	177	1.46×10^{-3}
OR2	AGAP009519	3R	3,372	35.17-35.18	189	3.7×10^{-4}
OR1	AGAP009640	3R	1,706	37.46-37.47	65	-1.76×10^{-3}

Table 6.4. Regions of the genome corresponding to the upper 0.1% of F_{st} distribution.

Chr	Region (Mb)	# SNPs	F_{st}	Genes
2L	3.63-3.64	101	0.020	AGAP004795, AGAP004797, AGAP004798
2L	8.21-8.22	128	0.013	-
2L	33.89-34.9	148	0.025	AGAP006526
2L	49.22-49.23	162	0.021	AGAP007717
2L	49.23-49.24	78	0.013	-
2R	0.25-0.26	133	0.020	-
2R	38.6-38.7	218	0.049	AGAP001370
2R	43.76-43.77	88	0.019	AGAP003820, AGAP003821
2R	47.63-47.64	152	0.018	-
2R	55.23-55.24	209	0.023	AGAP004369 (*Beat IIb*)
2R	58.95-58.96	77	0.024	AGAP004644
3L	5.26-5.27	140	0.017	AGAP010510 (*Tubulin β*)
3L	5.35-5.36	69	0.017	-
3L	7.07-7.08	52	0.017	-
3L	10.31-10.32	57	0.016	AGAP010783
3R	15.53-15.54	52	0.024	AGAP028088
3R	43.37-43.38	234	0.024	-
3R	43.66-43.67	195	0.025	AGAP009801, AGAP009802 (*Gr12*), AGAP009803 (*Gr11*)
3R	45.9-45.91	131	0.023	AGAP009947 (*Beat*)
3R	46.25-46.26	217	0.027	-
X	11.28-11.29	62	0.032	AGAP000627
X	18.66-18.67	91	0.021	-

Upper 0.1% F_{st} estimates were calculated separately for each chromosome arm and once SNPs with less than 50 SNPs were excluded. Genes (Vectorbase ID) were only listed if they were found within the specified genomic region.

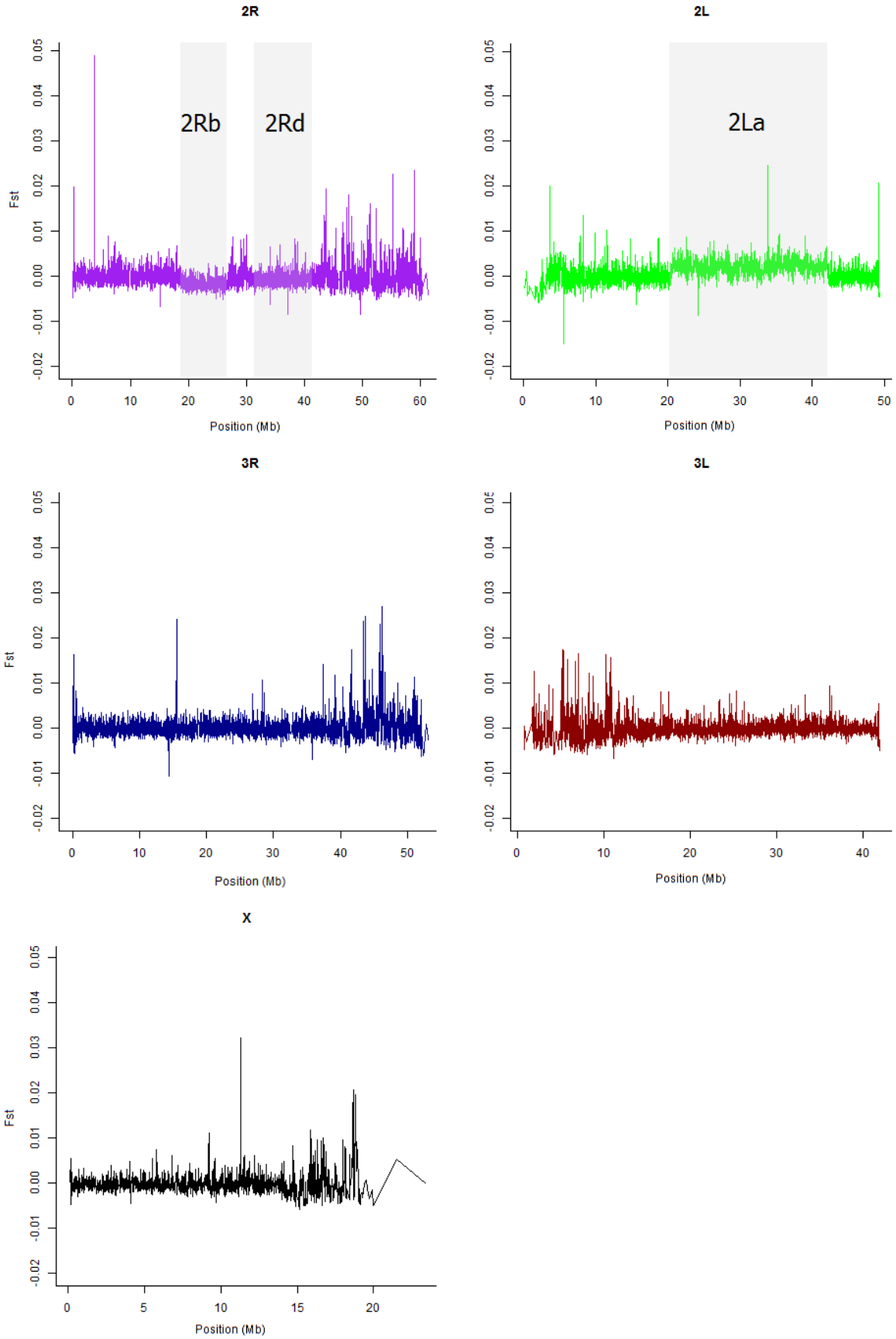

Figure 6.12. Differentiation between indoor- and outdoor-collected *An. coluzzii* from The Gambia. Weighted F_{st} was calculated using a non- overlapping window of 10 kb. Grey shaded areas represent inversions 2La, 2Rb and 2Rd which are segregating in the population.

Association analysis: Gambian dataset

We tested SNPs for outdoor-biting association using a logistic regression under three inheritance models (1df): additive, dominant and recessive. Generally, we found that there was good agreement of the data with what was expected under the null hypothesis of no association (Figure S13). This is consistent with our estimates of the genomic inflation factor (λ = 0.96-1.018), which indicate low possibility of false-positive associations resulting from population stratification.

Regarding the GOI, weak associations (p = 0.002-0.05) were identified at *OR65*, *OBP26* and *OBP47,* consistent with the results of the F_{st} analysis but these did not meet the threshold cut-off criteria. However, as mentioned previously, the fact that these genes were represented by a small number of SNPs (n = 1-263) may have decreased the potential to detect strong associations.

There were 73 SNPs located in 51 genomic regions that passed the threshold cut-off of $p < 10^{-5}$ (Table 6.5; Figure 6.14 & Figure S14). All except two SNPs (2L:28257115 and 3R:3831638) were found in non-coding regions (intergenic, downstream and upstream), potentially with regulatory functions on nearby genes (Hangauer *et al.*, 2013). Furthermore, many of these fell within genomic regions displaying elevated levels of differentiation between the two populations – as determined by the F_{st} analysis (Table 6.4; Figure 6.12).

Using Vectorbase, we identified the nearest genes (within 200 kb) for each of the 51 genomic regions and inferred their function based on their gene ontology annotations. We found that these genes had diverse biological functions including transcription, translation, metabolism and proteolysis (Figure 6.13; Appendix 8: Supplementary results). The largest group of genes had no known function (14

genes), while genes involved in transcription formed the second largest group with seven transcription factors and two transcription regulator genes. The strongest signal of association (dominant model, p = 2.92 x 10^{-7}) was in a non-coding region, approximately 27 kb downstream of *DPR* (SNP ID: 3L:17529965), a gene involved in the gustatory response to salt (Nakamura et al., 2002). Another signal was on *PNPLA15* (SNP ID: 3L:35174543, additive model, p = 4.17 x 10^{-7}), a gene involved in lipid breakdown that is essential for development and providing energy during periods of starvation (Dupont *et al.*, 2014). However without prior knowledge of genes implicated in host-seeking behaviour it is difficult to assess the implication of these associations.

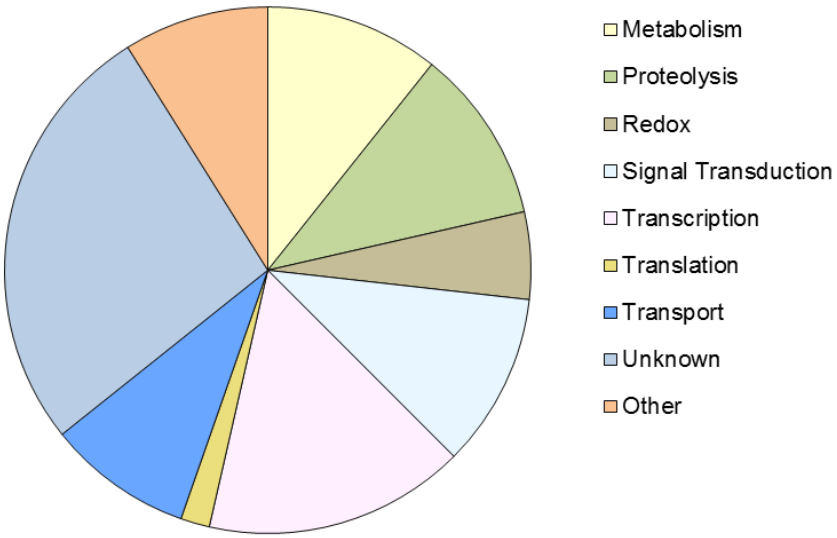

Figure 6.13. Manual annotation of genes found in genomic regions of significant association. Annotations were based on the gene ontology terms obtained from the Vectorbase database. Genes belonging to the "other" category consist of genes encoding storage proteins, DNA-binding proteins and molecules involved in axonal guidance.

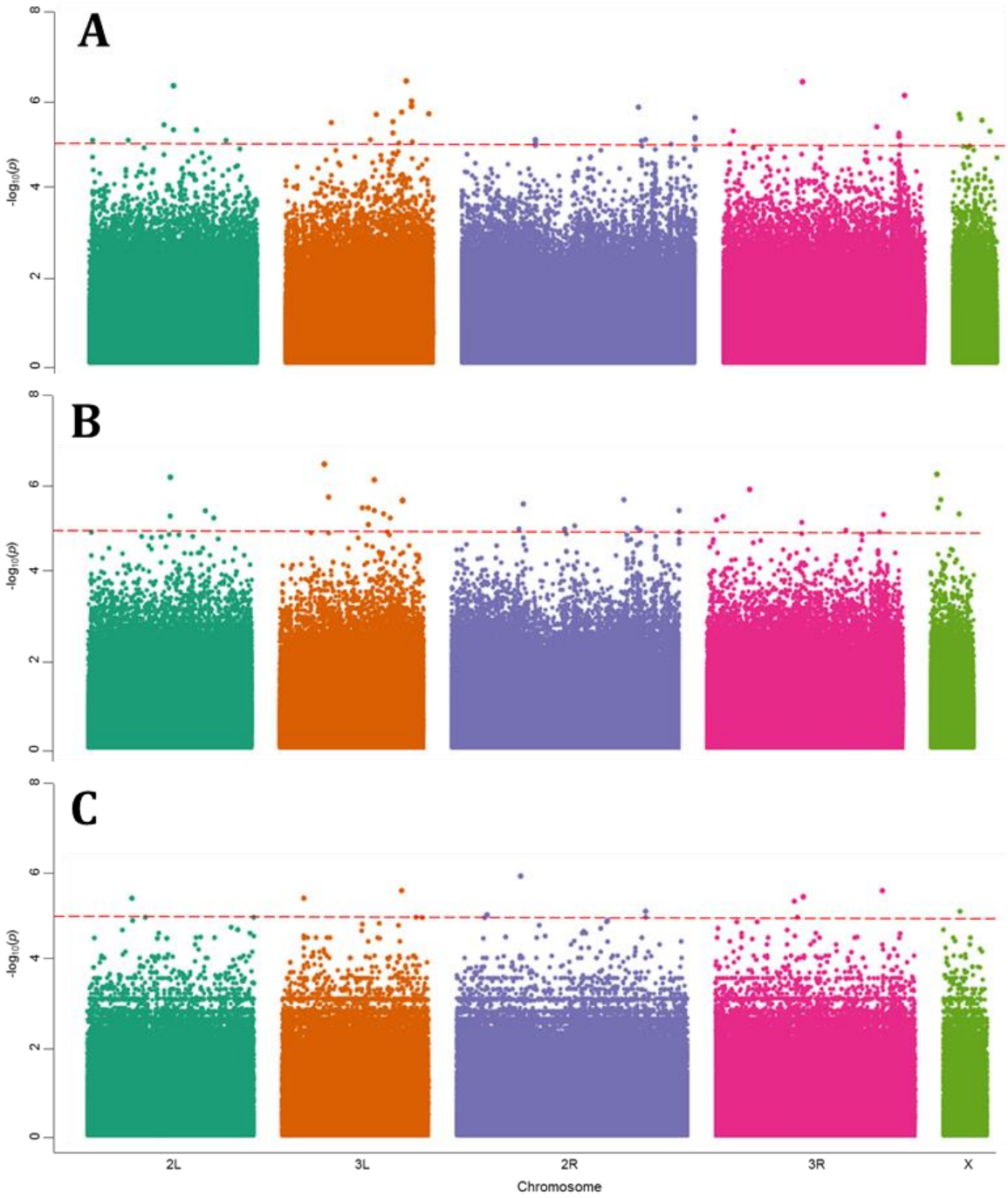

Figure 6.14. Genome-wide signals of association with outdoor-biting behaviour. Logistic regression under an additive (A), dominant (B) and recessive (C) model. Each solid circle represents a single SNP. Red dashed line represents the cut-off threshold ($p \leq 10^{-5}$).

Table 6.5. Regions of the genome showing association with outdoor-biting behaviour.

Chr	Region (Mb)	# SNPs	SNP	Allele	Indoor AF	Outdoor AF	*p*	OR	95% CI	Model	Nearby Genes
2L	3.67-3.68	1	2L:3670888	A	0.16	0.02	9.08×10^{-6}	0.12	0.03-0.39	Add	AGAP004798 (*COPB2)*
2L	15.32-15.33	1	2L:15326158	T	0.39	0.65	9.07×10^{-6}	2.63	1.67-4.13	Add	AGAP005412
2L	16.77-16.78	1	2L:16769668	A	0.26	0.4	3.26×10^{-6}	4	1.88-8.50	Rec	AGAP005536
2L	20.02-20.03	1	2L:20020494	T	0.24	0.4	8.99×10^{-6}	3.84	1.80-8.17	Rec	AGAP005737
2L	25.46-25.47	1	2L:25464003	A	0.06	0.23	4.07×10^{-6}	4.89	2.33-10.25	Add	AGAP006031 (*Nup54)*
2L	28.25-28.26	2	2L:28257115	G	0.19	0.03	5.35×10^{-7}	0.11	0.04-0.32	Add	AGAP006208, AGAP006207
2L	34.63-34.64	1	2L:34633056	T	0.06	0.23	5.31×10^{-6}	4.84	2.33-10.05	Add	AGAP006598
2L	38.11-38.12	1	2L:38112955	G	0.17	0.03	3.45×10^{-6}	0.32	0.17-0.59	Dom	AGAP006773
2L	40.49-40.50	1	2L:40496599	G	0.44	0.4	5.06×10^{-6}	0.44	0.31-0.64	Dom	AGAP006991
2L	42.67-42.68	1	2L:42671746	A	0.36	0.6	9.17×10^{-6}	2.74	1.71-4.40	Add	AGAP007100
2R	6.82-6.83	1	2R:6826788	A	0.16	0.32	9.13×10^{-6}	4.98	1.79-13.95	Rec	AGAP001616
2R	7.47-7.748	2	2R:7471983	G	0.34	0.53	7.89×10^{-6}	2.73	1.68-4.44	Rec	AGAP001657 (*Hexamerin*)
2R	17.04-17.05	1	2R:17043652	A	0.4	0.57	9.06×10^{-6}	2.41	1.56-3.73	Dom	AGAP002178 (*hth*)
2R	18.03-18.04	3	2R:18031010	T	0.19	0.41	2.42×10^{-6}	2.15	1.55-3.00	Dom	AGAP002238
2R	27.53-27.54	2	2R:27532509	G	0.18	0.04	9.08×10^{-6}	0.36	0.21-0.62	Dom	AGAP002793 (*Slit1*)
2R	29.59-29.60	1	2R:29597194	T	0.31	0.5	7.71×10^{-6}	2.18	1.51-3.15	Dom	AGAP002917
2R	42.84-43.50	2	2R:42843109	G	0.17	0.02	1.64×10^{-6}	0.1	0.03-0.34	Add	AGAP003746, AGAP013391
2R	44.78-44.79	1	2R:44788020	A	0.2	0.04	8.74×10^{-6}	0.17	0.06-0.44	Add	AGAP003871
2R	46.66-46.67	1	2R:46589775	C	0.17	0.34	8.53×10^{-6}	2.05	1.48-2.85	Dom	AGAP003940
2R	46.73-47.28	3	2R:46726249	A	0.28	0.43	6.56×10^{-6}	3.38	1.79-6.39	Rec	AGAP003945 (*PTBP2*), AGAP003971
2R	61.42-61.44	3	2R:61418659	C	0.18	0.03	3.42×10^{-6}	0.35	0.21-0.60	Dom	AGAP004675 (*GPRMAC2*)
3L	11.68-11.69	1	3L:11681894	C	0.29	0.41	3.26×10^{-6}	4	1.88-8.50	Rec	AGAP010833 (*CLIPB14*)
3L	17.52-18.59	2	3L:17529965	C	0.12	0.32	2.92×10^{-7}	2.35	1.67-3.30	Dom	AGAP011128 (*DPR*), AGAP011187 (*TOLL10*)
3L	27.03-27.04	1	3L:27038709	A	0.23	0.06	2.97×10^{-6}	0.4	0.26-0.62	Dom	AGAP011544
3L	28.35-28.36	1	3L:28354730	A	0.2	0.04	2.38×10^{-6}	0.14	0.05-0.37	Add	AGAP011580 (*EIF3K*)
3L	28.41-28.41	2	3L:28410034	C	0.37	0.55	7.17×10^{-6}	2.52	1.56-4.01	Dom	AGAP011584
3L	29.89-29.90	2	3L:29895649	A	0.18	0.37	6.74×10^{-7}	2.25	1.61-3.13	Dom	AGAP028207
3L	32.00-32.01	1	3L:32006693	G	0.46	0.27	4.06×10^{-6}	0.47	0.34-0.66	Dom	AGAP011712
3L	32.22-32.23	2	3L:32221764	A	0.39	0.66	3.48×10^{-6}	2.86	1.78-4.61	Add	AGAP011719
3L	33.52-33.53	1	3L:33525031	A	0.08	0.27	5.06×10^{-6}	2.26	1.57-3.25	Dom	AGAP011812
3L	34.15-35.17	2	3L:35174543	T	0.35	0.63	4.17×10^{-7}	3.33	1.98-5.58	Add	AGAP011925 (*PNPLA5*), AGAP011852
3L	36.41-36-42	4	3L:36413185	T	0.04	0.21	1.19×10^{-6}	6.41	2.80-14.68	Add	AGAP012019
3L	36.72-36.73	1	3L:36725585	G	0.17	0.37	9.87×10^{-6}	3.35	1.90-5.92	Add	AGAP012031
3L	40.09-40.10	1	3L:40090717	G	0.48	0.29	9.14×10^{-6}	0.29	0.14-0.60	Rec	AGAP012312
3L	40.47-4.48	1	3L:40476664	A	0.5	0.25	2.30×10^{-6}	0.31	0.18-0.52	Add	AGAP001238 (*RASGRF2*)
3R	2.28-2.29	1	3R:2283582	G	0.06	0.22	5.61×10^{-6}	5.34	2.51-11.38	Add	AGAP007859
3R	3.83-3.84	1	3R:3831638	T	0.09	0.24	4.67×10^{-6}	2.24	1.57-3.21	Dom	AGAP007989
3R	9.90.-9.91	1	3R:9908789	T	0.35	0.51	1.12×10^{-6}	2.51	1.66-3.81	Dom	AGAP028019 (*CYP4H18*)
3R	19.01-19.02	1	3R:19019495	T	0.43	0.57	4.33×10^{-7}	3.25	1.99-5.33	Add	AGAP028025
3R	23.76-23.77	1	3R:23769190	T	0.22	0.2	6.39×10^{-6}	0.48	0.35-0.67	Dom	AGAP009022
3R	34.58-34.59	1	3R:34584956	G	0.12	0.39	9.66×10^{-6}	2.08	1.48-2.93	Dom	AGAP009482 (*ZNF500*)
3R	37.56-39.14	1	3R:37563683	C	0.09	0.31	4.53×10^{-6}	0.26	0.15-0.46	Add	AGAP009646 (*CDX*), AGAP009700 (*SWD2*)

3R	41.56-41.57	1	3R:41563691	A	0.37	0.62	2.19×10^{-6}	2.35	1.63-3.41	Rec	AGAP009725
3R	43.37-43.38	4	3R:43376739	G	0.32	0.12	6.31×10^{-7}	0.26	0.13-0.49	Add	AGAP009792
3R	44.74-44.75	1	3R:44747427	T	0.58	0.31	8.87×10^{-7}	0.31	0.9-0.52	Add	AGAP009886 (*MED15*)
X	2.95-3.96	1	X:2954520	C	0.13	0.34	4.98×10^{-7}	2.29	1.61-3.21	Dom	AGAP012978
X	3.49-3.50	1	X:3492955	A	0.19	0.04	2.96×10^{-6}	0.37	0.22-0.60	Dom	AGAP000214 (*GUCY1B*)
X	4.54-4.55	1	X:4548626	C	0.18	0.04	1.92×10^{-6}	0.34	0.20-0.58	Dom	AGAP000244
X	7.15-7.16	1	X:7151127	G	0.24	0.47	6.56×10^{-6}	3.38	1.79-6.39	Rec	AGAP000389
X	12.10-12.11	1	X:12109761	A	0.21	0.04	4.10×10^{-6}	0.39	0.24-0.62	Dom	AGAP000678 (*WDR68*)
X	15.87-15.88	1	X:15878258	A	0.26	0.96	5.70×10^{-6}	2.87	1.77-4.66	Add	AGAP000862

Shown are regions with ≥SNPs with $p < 10^{-5}$. SNPs within 250 kb of each other were grouped together into a single region, and the SNP with the strongest signal and the model which gave the signal are reported. Each odds ratio (OR) and allele frequency (AF) is defined for the allele specified. We report the most significant model, either additive (add), dominant (dom) and recessive (rec). The nearby gene (Vectorbase ID) is defined as the closest gene within 200kb of the SNP. CI= Confidence interval.

Meta-analysis

A total of 59 SNPs were analysed for outdoor-biting association in the Ugandan dataset, but only five showed a significant effect (p <0.05) (Table S6 & Table S7). This may be explained by the low allele frequencies of most of the SNPs. For example, the SNP X:3492955 has a MAF of 0.007 in the indoor samples which is considerably lower than what was observed in the Gambian samples (0.19). Low allele frequencies lead to low power to detect associations, and the result was consistent with an association of modest effect but not statistically significant (OR = 1.75, 95% CI = 0.05-0.36, p = 0.29).

Of the five SNPs that showed a significant effect in the Ugandan dataset, only three remained significant in the meta-analysis. For the remaining two SNPs, we observed a lack of significance because the alleles had opposing association relationships across the two datasets, termed “flip-flop” (Lin *et al.*, 2007). For example, allele C at loci 3L:35174543 was positively associated with outdoor-biting behavior in the Gambian dataset but was negatively associated in the Ugandan dataset. Therefore the combined analysis did not yield a statistically significant effect (OR = 1.34, 95% CI = 0.95-1.89, p = 0.09) despite a significant, but reversed effect in both the Gambian ($p = 4.17 \times 10^{-7}$) and Ugandan dataset (p = 0.02). We consider this observation genuine (i.e. not a false positive) as the flanking SNP, 3L: 35174648 also exhibited allele flipping with a borderline significant effect (OR = 1.37, 95% CI = 0.98-1.91, p = 0.06). Given that the two datasets represent two different populations and species, it is likely that they have different linkage disequilibrium architectures and haplotype frequencies, which explain the flip-flop associations observed here (Lin *et al.*, 2007, Zaykin and Shibata, 2008). Regardless of the direction of association, it is interesting that

these SNPs are associated with outdoor-biting behav ior in both *An. coluzzii* and *An. gambiae,* indicative of possible parallel selection.

The meta-analysis showed a significant association with outdoor-biting behaviour in 23 SNPs (p = 0.05-5.96 x 10^{-6}). SNPs, 3R:44747427 (additive model, OR = 0.33, p = 5.96 x 10^{-6}) and 2R:18031010 (dominant model, OR = 1.66, p = 7.62 x 10^{-6}) showed the strongest evidence of association. 3R:44747427 is found within the intronic region of the *MED15* gene. *MED15* encodes a component of the Mediator complex which is responsible for regulating the expression of many genes by promoting the interaction of RNA polymerase II with several transcription factors (Kwon *et al.*, 1999). In contrast, 2R:18031010 is located in an intergenic region approximately 3.5 kb upstream of a transcription factor gene, AGAP002238. Although little is known about the function of intergenic regions, studies have shown that they contain non-coding RNA genes with regulatory functions (Hangauer *et al.*, 2013), it is therefore conceivable that the upstream region of AGAP002238 regulates its activity.

The analysis identified additional genes involved in transcription regulation. For example, we observed significant associations at SNPs found within the homeobox transcription factor, *CDX* (additive model, OR = 0.49, p = 2.75 x 10^{-4}). Similarly, we observed a signal at a non-coding region approximately 8kb downstream of *SWD2* (dominant model, OR = 2.05, p = 1.67 x 10^{-5}). This gene encodes *SWD2,* a component of the COMPASS complex, which plays a role in transcription activation through chromatin remodeling (Miller *et al.*, 2001).

8 out of 23 SNPs exhibiting significant associations were found within or near genes with catalytic functions. This includes a guanylate cyclase subunit

(*GUCY1B*), a diacylglycerol kinase (AGAP000519), a carboxylesterase (*COEunkn*), an acetyl-CoA carboxylase (AGAP011616), serine-type endopeptidases (AGAP004644), a sulfotransferase (AGAP009809), and a short-chain dehydrogenease/reductase (AGAP011852). Although little is known regarding these genes, it is possible that they play a role in the downstream molecular signaling underlying biting behaviour. For instance, *GUCY1B* synthesises the cGMP signaling molecule than in turn regulates the activity of several molecules. In particular, cGMP regulates the activity of the protein kinase, *for,* which confers different foraging traits in *Drosophila* (Renger *et al.*, 1999).

The most notable associations were within three genes of known biological significance – gustatory receptors, *Gr11* and *Gr12* as well as, the GOI, *OR1. OR1* plays a critical role in establishing host preferences -exhibiting greater binding affinity to human-associated odours (Fox *et al.*, 2001, Hallem *et al.*, 2004). The role of gustatory receptors (*GRs*) in host finding is less clear as their main function is taste perception, but it has been proposed that *GRs* may also function as odour receptors (Scott *et al.*, 2001). Taken together, these findings suggest that variants in genes involved in both upstream (*ORs* and *GRs*) and downstream signalling (transcription-related and catalytic genes) act in concert to confer outdoor-biting behaviour.

Table 6.6. Significant associations from the combined analysis of the genome-wide association and replication experiment.

		GWAS experiment: Gambian dataset				Follow-up experiment: Ugandan dataset				Combined				
SNP	**Allele**	**AF Indoor**	**AF Outdoor**	**OR**	***p* (2-sided)**	**AF Indoor**	**AF Outdoor**	**OR**	***p* (1-sided)**	**OR**	**95% CI**	***p***	**Model**	**Nearby Gene**
2L:25464003	G	0.94	0.77	0.2	4.07×10^{-6}	0.84	0.84	0.99	0.49	0.55	0.37-0.84	5.87×10^{-3}	Add	AGAP006031 (*Nup54*)
2L:3636997*	C	0.66	0.84	1.85	1.66×10^{-4}	0.92	0.94	1.07	0.37	1.49	1.15-1.93	2.43×10^{-3}	Dom	AGAP004797
2R:18031010	T	0.19	0.41	2.15	2.42×10^{-6}	0.17	0.29	1.34	0.03	1.66	0.133-2.07	7.62×10^{-6}	Dom	AGAP002238
2R:18031141	T	0.16	0.38	3.06	9.69×10^{-6}	0.14	0.15	1.06	0.4	1.61	1.15-2.26	5.87×10^{-3}	Add	AGAP002238
2R:283739*	G	0.84	0.95	3.19	2.04×10^{-3}	0.90	0.92	1.2	0.27	1.72	1.08-2.73	2.33×10^{-2}	Add	AGAP011509 (*COEunkn*)
2R:47283021*	T	0.12	0.31	1.9	1.11×10^{-4}	0.10	0.07	0.83	0.23	1.47	1.11-1.94	6.19×10^{-3}	Dom	AGAP003971
2R:47284139*	A	0.17	0.31	2.19	3.75×10^{-3}	0.09	0.09	0.99	0.48	1.5	1.01-2.22	4.25×10^{-2}	Add	AGAP003971
2R:58955948*	A	0.7	0.84	2.27	3.29×10^{-3}	0.99	0.99	1.15	0.45	2.2	1.33-3.57	2.33×10^{-3}	Add	AGAP004644
3L:29895651	C	0.2	0.41	2.12	3.41×10^{-6}	0.34	0.26	0.85	0.13	1.24	1.00-1.52	0.048	Dom	AGAP001616
3L:33525031	A	0.08	0.27	2.26	5.06×10^{-6}	0.13	0.12	0.95	0.37	1.39	1.09-1.78	7.19×10^{-3}	Dom	AGAP011812
3L:34158609	G	0.63	0.85	3.82	2.11×10^{-6}	0.79	0.80	1.07	0.38	1.74	1.24-2.46	1.55×10^{-3}	Add	AGAP011852
3R:37461899¥	T	0.43	0.37	1.25	0.318	0.17	0.09	2	0.01	1.48	1.03-2.12	3.07×10^{-2}	Add	AGAP009640 (*Or1*)
3R:37462227¥	A	0.35	0.27	0.78	0.110	0.19	0.11	0.68	0.01	0.73	0.58-0.92	8.27×10^{-3}	Dom	AGAP009640 *(Or1)*
3R:37557083	T	0.24	0.1	0.26	4.53×10^{-6}	0.05	0.04	0.84	0.26	0.49	0.33-0.72	2.75×10^{-4}	Add	AGAP009646 (*CDX2*)
3R:37563683	C	0.12	0.31	3.86	4.53×10^{-6}	0.12	0.14	1.19	0.26	1.96	0.13-2.93	9.53×10^{-4}	Add	AGAP009646(*CDX2*)
3R:39145574	C	0.14	0.31	2.07	1.66×10^{-5}	0.00	0.02	1.83	0.16	2.05	1.48-2.84	1.67×10^{-5}	Dom	AGAP009700 (*SWD2*)
3R:43665034*	T	0.41	0.22	0.38	1.55×10^{-4}	0.68	0.69	1.06	0.4	0.7	0.51-0.98	3.86×10^{-2}	Add	AGAP009801
3R:43668381*	T	0.53	0.74	0.39	1.09×10^{-4}	0.12	0.14	1.25	0.21	0.66	0.45-0.96	2.81×10^{-6}	Add	AGAP009802 (*Gr12*)
3R:43669141*	T	0.46	0.37	0.61	5.18×10^{-3}	0.82	0.75	0.71	0.36	0.67	0.54-0.83	2.87×10^{-4}	Dom	AGAP009803 (*Gr11*)
3R:43669258*	C	0.45	0.39	0.61	6.52×10^{-3}	0.47	0.46	0.94	0.34	0.79	0.63-0.99	4.13×10^{-2}	Dom	AGAP009803 (*Gr11*)
3R:43728754	C	0.79	0.93	1.93	4.18×10^{-4}	0.96	0.95	0.9	0.34	1.45	1.07-1.97	1.80×10^{-2}	Dom	AGAP009809
3R:44744808	A	0.67	0.52	2.27	3.5×10^{-6}	0.20	0.20	1.02	0.48	1.56	1.11-2.18	1.08×10^{-2}	Add	AGAP009886 (*MED15*)
3R:44747427	T	0.42	0.31	0.31	8.87×10^{-7}	0.99	0.98	0.59	0.26	0.33	0.21-0.54	5.96×10^{-6}	Add	AGAP009886 (*MED15*)
3R:45900850*	C	0.63	0.57	2.06	7.26×10^{-4}	0.34	0.33	0.96	0.41	1.36	1.01-1.82	3.97×10^{-2}	Add	AGAP009947 (*Beat*)
X:3492955	A	0.18	0.04	0.13	2.96×10^{-6}	0.01	0.01	1.75	0.29	0.23	0.09-0.55	1.01×10^{-3}	Add	AGAP000214 (*GUCY1B*)
X:9227517	T	0.94	0.8	0.28	1.47×10^{-4}	0.98	0.97	0.66	0.22	0.35	0.20-0.61	2.02×10^{-3}	Add	AGAP000519

SNPs showing significant association in the combined dataset of samples from The Gambia and Uganda. Results were combined using a fixed-effect model by inverse-variance weighting of log odd ratios. The odds ratios (OR) and allele frequency (AF) are defined for the allele specified. The nearby gene (Vectorbase ID) is defined as the closest gene within 200kb of the SNP. CI= Confidence interval. *SNPs that were chosen because they have F_{st} estimates that corresponded to the upper 0.1-5% of the distribution. ¥Denotes a SNP found within our genes of interest (GOI).

Conclusion

Substantial reductions in malaria transmission have been achieved in the last decade owing to the mass distribution of ITNs that target the predominantly indoor-biting *An. gambiae s.s.*. However, there are increasingly frequent observations that vectors are diverting their host-seeking activity outdoors, highlighting an important limitation in vector control. In this study, we re-sequenced the whole genome of 171 indoor- and outdoor-collected *An. gambiae s.s.* from The Gambia to generate 3,470,851 high-quality SNPs. Such a dense dataset allowed the identification of several genes potentially under positive selection and variants associated with the outdoor-biting phenotype.

Using PCA, we found that that the inversion states of 2La, 2Rb and 2Rd exhibited frequencies that were similar to those reported previously from other sites in the Wali Kunda area (Bryan *et al.*, 1982). It is also typical for populations found in dry environments because carriers of the 2Rb and 2La inverted arrangements confer an adaptive advantage (Coluzzi *et al.*, 1979). These inversions however, do not appear to play a role in determining indoor- and outdoor-biting as the frequency all three inversions conformed to HW expectations.

It is perhaps not surprising that the PCA and ADMIXTURE analysis did not detect a genetic difference between the two populations. While both analyses are typically sufficient in identifying population structure, it is less robust for identifying a genetic difference between freely interbreeding populations. Although indoor- and outdoor-biting populations may exhibit behavioural divergence, we assume that both populations still share the same mating and breeding sites. Therefore, we expected to see genetic differences at a few loci under positive selection only.

If such loci are indeed very few, it is unlikely that PCA and ADMIXTURE analysis would have identified them especially if the genetic variation within the population is high.

Overall we found no associations and little evidence of genetic differentiation at GOI, except for two genes. *OBP47* and *OR65* showed weak evidence of differentiation (F_{st}= 0.003 and 0.001, respectively) between the indoor and outdoor population in the Gambian dataset. We also found weak evidence of association at these two genes, although they did not meet our threshold cut off of $p < 10^{-5}$. However we were unable to investigate both genes further during the follow-up study as they were excluded during the filtering stage. Surprisingly, the meta-analysis identified a significant association at *OR1.* Both datasets displayed an effect size in the same direction, although it did not reach the significance level in the Gambian dataset. The function of *OR1* in host-seeking is well document due to its strong response to components of human sweat (Fox *et al.*, 2001, Hallem *et al.*, 2004). Taken together, there is suggestive evidence that *OR1* is a candidate gene that needs to be further investigated.

The meta-analysis also identified significant associations at *Gr11* and *Gr12.* Both genes displayed modest effects in the Gambian dataset (p = 6.52 x 10^{-3} - 1.09 x 10^{-4}) and were identified by the F_{st} analysis because they were found within windows corresponding to the upper 0.1% of the distribution. Like *OR1*, *Gr11* and *Gr12* are strong candidate genes because they have been implicated in the detection of CO_2, a human-associated stimulus, and have a predicted function as *ORs* (Bohbot *et al.*, 2014, Sparks *et al.*, 2013).

Consistent with most GWAS studies, an overwhelming majority of the significant associations identified in the Gambian dataset were located in regions that do not encode a protein, suggesting that their association with outdoor-biting behaviour may occur either due to changes in regulatory function or by chance (Boyle *et al.*, 2012). To highlight the potential functional significance of the observed associations with outdoor-biting behaviour, we identified the nearest genes closest to these signals. While most genes had no known, the remaining genes have roles in transcription regulation, catalysis and signal transduction. Our findings therefore suggest that the mechanisms that underlie indoor-outdoor biting tendency is complex, likely involving the regulation of several genes that function in the downstream signalling that determines biting behaviour.

In summary, both the F_{st} and GWAS analysis have identified candidate gene regions for outdoor-biting behaviour that are implicated in diverse downstream signalling functions as well as, olfactory and neuronal signalling (e.g. *Beat*). Although we lack the power to detect higher level of significance ($p > 10^{-8}$) because of the small sample size of the Gambian dataset, we have demonstrated the validity of our findings both by carrying out a follow-up analysis in the Ugandan dataset and by performing a meta-analysis. Further investigations of these associations, as well as following-up on all the other regions identified by the F_{st} and GWAS, will be needed to confirm our findings in future studies.

7. Discussion

Control measures implemented to curb malaria transmission have been limited primarily as a result of the ability of *Anopheles* mosquitoes to adapt to such measures. To achieve an early and well-informed response to control measures failing, it is essential that vectors are monitored intensively. This thesis aimed to investigate whether there is a genetic difference between indoor- and outdoor-biting populations of *An. gambiae.* Using entomological data collected from the field and published literature, we examined the evidence for all *Anopheles* vectors. Next, we focused on the major vectors of malaria in sub-Saharan Africa, *An. coluzzii* and *An. gambiae* (formerly referred to as “M” and “S” molecular forms, respectively) and investigated whether there is genetic evidence for an increased tendency to bite outdoors (“exophagy”).

In 2000, a target was set as part of the United Nations’ Millennium Development Goals to halve the burden of malaria by 2015. Mass distribution of long-lasting insecticidal nets (LLINs) was at the forefront of this effort, leading to an increased coverage of LLINs from 3% to 50% in sub-Saharan Africa where malaria is most prevalent (The Global Fund, 2013). Many hypothesised that such mass-scale use of insecticides would create a strong selection pressure for dominant *Anopheles* vectors to develop resistance mechanisms through physiological, metabolic and behavioural changes for survival and reproduction. The development of physiological and metabolic resistance to insecticides have been well-documented in several *Anopheles* species (see: IR Mapper [www.irmapper.com]) yet, confirmed examples of behavioural resistance have been rare. Predicted

outcomes include changes in their host preference, time of biting and their propensity to bite and rest indoors or outdoors (Muirhead-Thomson, 1960).

In Chapter 2, we performed a systematic review of studies investigating such changes and found limited evidence of behavioural resistance because the data available was, at times, difficult to interpret. A number of studies investigating resting behaviour and host preference did not sample mosquitoes outdoors, and were therefore liable to sampling bias that almost certainly underrepresented the overall behaviour of the vector population. Other methodological issues include the presence of confounding factors, inadequate outcome reporting, short follow-up surveillance of vector behaviour after the introduction of interventions and the surveillance of vectors in the presence of an intervention. In the presence of an intervention, it is not possible to determine if any proportional changes in endophilly and endophagy represent true behavioural resistance rather than an artefact of the killing/repellent effect of the insecticide. As almost all studies in the review failed to control for this factor, we cannot conclude that behavioural resistance has occurred in these studies.

Our findings are consistent with the current literature concerning behavioural resistance. Gatton *et al.* (2013) conducted a similar review and concluded that the observational data available are often ambiguous and liable to publication bias. They suggested that apparent behavioural changes could be explained by shifts in species composition within the complex or simply reflect the consequence of unsuccessful feeding on the prior evening. That is, unfed mosquitoes initiate their host search soon after dusk and feed opportunistically on any hosts encountered outdoors on the way to a more suitable host indoors. Drawing upon recent experimental evidence demonstrating associative learning in mosquitoes (Chilaka

et al., 2012), Killeen and Chitnis (2014) suggested that mosquitoes learn the optimal conditions for survival and blood-feeding and adjust their host-seeking activities accordingly. In such a case they propose associative learning as an additional or alternative explanation for observed behavioural changes. Although there appears to be no clear consensus concerning the origins of behavioural changes, left unchecked, behavioural resistance could have a very dramatic impact on the efficiency of current vector control interventions. According to Ranson *et al.* (2011), changes towards outdoor-biting behaviour could potentially exceed the significance of physiological resistance since we have few tools that target outdoor vectors. However, the impact of such a change may not be as great if the majority of malaria transmission still occurs indoors as demonstrated recently by Bradley *et al.* (2015).

Chapters 2 and 3 demonstrate the difficulties associated with using entomological data to investigate behavioural resistance. Yet, laboratory-based behavioural studies are limited in their ability to represent natural conditions. To bypass these problems, we aimed to investigate behavioural resistance using genetic information by capitalising on recent developments in both DNA sequencing and statistical methods. Specifically we aimed to investigate whether there is evidence of a genetic change in indoor-outdoor biting preference. We hypothesised that outdoor-collected vectors would be genetically differentiated from indoor-collected vectors, resulting from the strong selection towards exophagic behaviour in response to vector control.

As a first step, we adopted a candidate gene approach and genotyped 13 odour receptors (*ORs*) and odour-binding proteins (*OBPs*) in *An. coluzzii* and *An.*

gambiae samples from The Gambia and Uganda respectively. Given the major role of olfaction in determining biting behaviour, we posited that genes of interest (GOI) involved in the divergence of indoor- and outdoor-biting mosquitoes should include olfactory-related genes such as *ORs* and *OBPs* given their primary role in host-seeking behaviour. GOI were therefore chosen on the basis of published experimental evidence demonstrating a strong response to human-associated odours. Using a dataset of approximately 30 single nucleotide polymorphisms (SNPs) representing these GOI, the analysis found substantial interbreeding between *An. coluzzii/ An. gambiae* and *An. arabiensis*, except at a few loci where divergence was apparent. In contrast, we found few genetic differences between the indoor- and outdoor-biting populations of *An. coluzzii* and *An. gambiae*.

There are several possible explanations for why a genetic difference between indoor- and outdoor-collected samples was not detected. Simply, there may indeed be no genetic difference between the two populations. This is especially likely because both populations share the same mating and breeding sites, precluding the potential for rapid divergence aided by reproductive barriers. Whole genome divergence can be observed following a strong selective sweep across reproductively isolated populations. In the absence of reproduction isolation, it is still possible that the early-stages of divergence may be detected at a few loci under selection. This has been demonstrated in the wild by the divergence of *An. gambiae* and *An. coluzzii* despite incomplete reproductive isolation driven primarily by differential adaptation to larval habitats (Diabaté *et al.*, 2008, Lehmann and Diabaté, 2008), although pre-mating reproductive barriers (Dabire *et al.*, 2013, Gimonneau *et al.*, 2010, Diabaté *et al.*, 2009) also contribute to their divergence. Similarly, the identification of an outdoor-resting subgroup of *An. gambiae s.s.,*

called Goundry in Burkina Faso provides a more recent example that genetic differentiation and the formation of subpopulations occurs frequently in *An. gambiae* despite the presence of ongoing gene flow (Riehle *et al.*, 2011). The larval and adult ecology of Goundry remains poorly understood, but it is thought that their exploitation of outdoor-resting sites and specialisation of novel larval habitats facilitate their ecological separation (Crawford *et al.*, 2012). Such phenomena are proof of concept and further corroborate our hypothesis regarding indoor-outdoor divergence. Given that we were only able to test a limited number of markers, it is possible that the evidence of early-stage divergence could have been overlooked during the analysis in Chapter 4.

Chapter 4 highlighted the importance of using a large dataset to find evidence of selection and divergence. Regardless of improvements to the technique, it is limited fundamentally by the bias in candidate gene selection. In contrast, genome-wide analysis increases the potential to identify genes or variants under positive selection regardless of whether their function was known before. In Chapter 6 we performed a genome wide analysis of approximately 3 million SNPs, genotyped in indoor- and outdoor-biting samples of *An. coluzzii* collected in The Gambia. We identified few genetic differences between indoor- and outdoor-samples following both a principal component analysis (PCA) and cluster analysis, indicating that the samples belong to a single population. Furthermore, we found that inversions 2Rb, 2Rd and 2La were in agreement with Hardy-Weinberg equilibrium, suggesting that these inversions do not play a role in determining indoor-outdoor biting behaviour. Taken together, these results support our expectation that there is little genetic difference between the two populations overall.

Despite our inability to detect an overall genetic difference between the two populations, we identified several genes potentially under positive selection. A number of biologically relevant genes were identified on the basis of estimates corresponding to the upper 0.1% of the F_{st} distribution. Notably, gustatory receptors (*Gr11* and *Gr12*) and a putative chemosensory receptor were identified, as well as several genes with a predicted role in learning, memory and circadian rhythm. We concluded that these require further investigation, by functional study or using a larger dataset consisting of more samples, to confirm their role in biting behaviour.

The results of the F_{st} analysis were further substantiated by the findings of the genome-wide association study (GWAS). Using a threshold threshold cut-off of $p < 10^{-5}$, we identified 73 SNPs; the majority of these were located within genomic regions of elevated differentiation (corresponding to the upper 0.1%-10% of the F_{st} distribution). When we investigated the genes closest to these SNPs (within 200 kb) we found that the genes identified had predicted roles in transcription, translation, metabolism and proteolysis, although the majority had no confirmed function. The finding of associations in regions that do not encode a protein is not uncommon in a GWAS. Indeed, approximately 88% of identified SNPs in human GWAS studies are found in intergenic or intronic regions (Hindorff *et al.*, 2009).

There are several explanations for a significant association in these regions. The simplest explanation is that the association is a false positive and the loci identified do not have a causal relationship with the trait under study. Another commonly cited explanation is that the true causal SNP is not genotyped but that the loci identified is in LD with the causal SNP (Hindorff *et al.*, 2009). Alternatively, these regions may contain regulators of gene expression and affect it through

changes in transcription factor binding affinity at enhancer or promoter elements; disruption of chromatin interactions; alternative splicing and post-translational modifications. With the availability high-throughput sequencing technologies there is a movement towards identifying these regulatory regions by, for example, epigenomic profiling and RNA-seq (Knight, 2014). Notably, the ENCODE (ENCODE, 2012), Roadmap Epigenetics (Bernstein *et al.*, 2010) and the FANTOM5 (Andersson *et al.*, 2014) projects have made great progress towards creating an epigenomic map of the human genome. Similarly, efforts have been made for the *An. gambiae* genome (e.g. Vannini *et al.* (2014), Gómez-Díaz *et al.* (2014)) and some have already been integrated into the VectorBase database although it is still in its early stages.

Overall, the meta-analysis showed that a total of 23 SNPs were significantly associated with outdoor-biting behaviour. A number of genes with functions related to transcription and catalysis were identified upon investigation of the genes closest to these SNPs*. We were able to replicate the signal at the second highest GWAS hit, 2R:18031010 in the Ugandan dataset (p = 0.03). This SNP is located in an intergenic region approximately 3.5 kb downstream of the transcription factor gene, AGAP002238. This result lends further strengthens this SNP as a candidate that contributes to the outdoor-biting phenotype, possibly by exerting a regulatory function on AGAP002238.The second strongest combined signal of association ($p = 5.96 \times 10^{-6}$) was found within the intronic region of the *MED15* gene which encodes a component of the Mediator complex, responsible for regulating the expression of many genes (Kwon *et al.*, 1999), although this signal did not reach the significance level in the Ugandan dataset.

* Again, the SNPs were mapped to the nearest genes (within 200 kb).

Perhaps more notable, significant associations were identified in gustatory receptors, *Gr11* and *Gr12* ($p = 10^{-4}$-10^{-6}) as well as, the odour receptor, *OR1* ($p = 10^{-2}$-10^{-3}). In Chapter 4, we investigated whether *OR1* was involved in the divergence of indoor- and outdoor-biting populations due to its high affinity to bind human-associated odour compounds. *Gr11* and *Gr12* have no clear role in olfaction, functioning mainly in taste perception, although they are ancestrally related to *ORs* and may thus play a role in odour perception (Scott *et al.*, 2001, Robertson *et al.*, 2003). The identification of these genes is supported by comparative genomic analyses of different anopheline species, conferring varying vectorial capacities, which suggest that *ORs* and *Grs* modulate their differences in host preference through functional divergence (Neafsey *et al.*, 2013). Antennal transcriptional analysis of *An. gambiae* and its non-vector sibling, *An. quadriannulatus* also demonstrate that differences in *OR* expression contribute towards their different host-seeking behaviour (Rinker *et al.*, 2013). Taken together, there is evidence that *ORs and Grs* are strong candidate genes for the potential divergence of indoor- and outdoor-biting populations. Our findings also suggest that the mechanisms that underlie indoor-outdoor biting behaviour are determined by a complex interaction of genes involved in both upstream (*ORs* and *GRs*) and downstream signaling (transcription-related and catalytic genes). Although further work is required before any firm conclusions can be made.

Due to limited resources, the follow-up analysis was limited to 59 SNP markers. We also prioritised the follow-up based on GOI, to address our aim of investigating whether olfactory-related genes contribute towards the potential genetic divergence of outdoor mosquitoes from indoor mosquitoes. This meant that it was not possible to follow-up the majority of the significant associations observed in

the GWAS ($p \leq 10^{-5}$), including the top SNP, 3L:35174543. It is desirable to investigate these SNPs in an experiment involving a larger sample size. SNP, 3L:35174543 is located in a non-coding region, close to the *DPR* gene. The encoded protein of *DPR* is a putative determinant of gustatory responses to salt (Nakamura *et al.*, 2002). Besides its role in electrolyte homeostasis, salt is also thought to stimulate or inhibit feeding in many insects including *Drosophila,* depending on its concentration (Hiroi *et al.*, 2004). Therefore, further investigation of this locus may provide some insight into the mechanisms linking taste perception and biting behaviour.

None of the association signals from the GWAS reached high levels of significance ($p > 10^{-8}$). We attributed this to the small sample size rather than the lack of significance being true. The importance of validating association signals using larger sample sizes to increase statistical power has been demonstrated frequently by several meta-analyses (e.g. Barrett *et al.* (2008), Zeggini *et al.* (2008)). Therefore, I would recommend future follow-up studies to use larger sample sizes to allow the detection of stronger association signals. I would also recommend future research to investigate rare variants (minor allele frequency <10%) which are thought to additionally explain disease-risk and trait heritability that has not been accounted for by common variants (Lee *et al.*, 2014).

Using simulated data derived from case-control studies of the HapMap project, Spencer *et al.* (2009) demonstrated that approximately a million SNPs and 10,000 samples (5,000 cases and 5,000 controls) are required to have 80% power to detect an association with a rare allele that has a small effect on disease risk. While the cost of such an experiment may preclude the analysis of rare variants at

present, it may not be long before the cost of whole-genome re-sequencing declines to allow such studies. In such instance, such an investigation may be limited by the time and cost required for genomic DNA extraction. Although we have attempted to address this issue in Chapter 5 by demonstrating the feasibility of using intact legs as direct PCR templates for genome amplification and subsequently, genotyping, further advances in sequencing technologies will be required before it can be applied to whole genome sequencing.

Here we have attempted to uncover the importance of the candidate regions we identified by examining their gene function and corresponding biological processes or gene networks. This approach is often the next step taken in a GWAS but it is important to note that without replication and validation, these regions remain as candidates. That is that they warrant further investigation before they are considered as being true positives. This point has been demonstrated by Pavlidis *et al.* (2012)*;* using a simulated sample of neutral genomes they showed that it is possible to obtain a biologically meaningful narrative out of a set of false-positive results through literature mining or gene ontology enrichment analysis. They suggest that genome-wide analyses are particularly vulnerable to this because the lack of an *a priori* hypothesis means that there are potentially unlimited ways of interpreting the results. We have attempted to control for this by prioritising candidate olfactory genes during the replication experiment and then validating our results by conducting a meta-analysis.

The role of the transcriptome (i.e. the collection of transcribed RNA sequences) in determining biting behaviour also warrants investigation. Non-coding DNA sequences are believed to play a role in gene regulation by way of RNA splicing or

by inhibiting gene expression through translation repression. Indeed, micro RNAs (miRNA), which act as post-transcriptional regulators of gene expression, are transcribed from such regions. To date, approximately 130 miRNAs have been described in *An. gambiae* and of these, several miRNAs were differentially expressed in bloodfed and non-bloodfed mosquitoes (Biryukova *et al.*, 2014). The functional significance of miRNAs as well as other RNA regulators, in determining gene expression highlights the complex interaction between the genome and transcriptome in ultimately determining behaviour.

Our results have provided a framework for identifying the evolution of exophagy, highlighting genomic regions of interest for further investigations. Targeted and whole-genome resequencing of large numbers of samples will likely play a key role in this process. Similarly, further advances in new technologies and analytical approaches will be required to determine the function of associated variants located in non-coding regions. Here, we provide the first evidence for the possible early-stage divergence of indoor- and outdoor-biting *Anopheles* populations, although future work is required to confirm these findings.

Appendices

Appendix 1: Informed consent form

INFORMED CONSENT FORM FOR ENTOMOLOGICAL STAFF PARTICIPATING IN HUMAN LANDING MOSQUITO COLLECTIONS

(Note: the form will be translated into Mandinka, Wollof and Fula which is the local language in the study area, and provided to the volunteer staff)

Project title: Genetic profile of indoor and outdoor biting *An. gambiae s.s.* in Walikunda, The Gambia

Name of the Institute and address: Medical Research Council Laboratories Unit, The Gambia

Names of the responsible Investigators: Project Coordinator: Prof. Umberto D'Alessandro, Theme Leader Disease Control and Elimination; Co- Investigators: Prof Charles Godfray, Department of Zoology, Oxford; Prof Steve Lindsay, School of Biological and Biomedical Sciences, Durham University; Musa Jawara Entomologist MRC Labs The Gambia.

Date:

_____

Information Sheet

1. Introduction

Malaria is a major disease in The Gambia. For control of malaria, the government either spray houses with insecticides or distribute insecticide treated nets. There is growing concern that these interventions may lead a change in the mosquitoes that transmit malaria and some of them will start biting outside the houses. If this occurs it will make malaria control more difficult since we have few effective tools for the control of mosquitoes biting outside the houses.

2. Purpose of the research intervention

We would like to compare mosquitoes that bite inside the houses with those biting outside, to show if there is any difference between them. The research outcome will be important to understand the change that may have occurred among the mosquitoes transmitting malaria after the large scale implementation of insecticide spraying and insecticide treated nets.

3. Participant selection

We would like to include you in the study as a trained entomological volunteer for this research.

4. Type of research

This is a study on the mosquitoes transmitting malaria and the comparison of those biting inside and outside the houses. For this purpose we will have to capture them both inside and outside the houses at the time they are about to bite. You will be asked to stay either inside a chosen house or outside and collect for a

certain period all mosquitoes landing on your legs. Collected mosquitoes will be identified, processed and send to England for further analysis.

5. Voluntary participation

Your participation in self-bait mosquito collection will be voluntary. You will still receive all the services you usually do whether you choose to participate or not.

6. Participant protection

The risk of getting malaria is expected to be low. Nevertheless, you will receive malaria medicine every day since the start of the study until 2 weeks after you finish the collection. This should decrease the risk of malaria even further.

8. Description of the process, procedures and protocol

We will use the standard self-bait method recommended by WHO, you will be required to work during the night of mosquito collection, which will be done up to 20 nights during this rainy season. You will sit on a chair with your lower legs bare and will collect mosquitoes as soon as they land on your exposed legs (but before probing) with the help of an aspirator and torchlight. The mosquitoes will be put in glass tubes. You will work from 7 PM to 10PM on the night of collection with short breaks to use facilities and take drinks and food. During the night of work you will not be permitted to smoke. A supervisor will remain close to the house during the night of work as a team member.

9. Adverse effects

There are no adverse effects of collection of mosquitoes landing on human volunteers and caught during the act of probing but before being bitten.

Nevertheless, should you observe any adverse effects, you may contact MRC/NMCP on the address/telephone given below.

10. Risks and discomforts

By participating in this research you may experience some discomfort as explained above. There is a risk of contracting malaria infection by accidental bite of infected mosquito. We will try to minimise the chances of any sort of discomfort from occurring, but if an untoward event does occur, we will provide you with highest available medical care and the project will pay for medical costs fully.

11. Benefits and incentives

If you participate in this research activity, you will receive a daily wage at the rate decided by the MRC personnel office. There may not be any other benefit for you but your participation is likely to help us find the answer to the research question, that is, whether there is a change in the mosquitoes transmitting malaria. This information will be useful as it will guide the control efforts in The Gambia.

12. Right to refuse or withdraw

Your participation in the programme is entirely voluntary. You are not under any obligation to participate and you also have the right to refuse this invitation. If at any point in time during the study you take the decision not to participate any further, you are free to do so immediately without any further discussion and it will have no consequences.

13. Who to Contact

If I have any questions about the study, you should contact Mr Musa Jawara Entomologist, MRC Labs Fajara Ph.: 0220 4495 443/6 X 4009 Mobile: 9938105 / 7719129 / 6550959 / 3938105

Part II. Certificate of Consent

(This is an integral part of the information sheet and not a stand-alone document)

I have read the information sheet / or it has been read to me in my own language. I have had the opportunity to ask questions about it and any questions that I have asked have been answered to my satisfaction. I consent voluntarily to participate in the mosquito collection activity in this study and understand that I have the right to withdraw from the study at any time without in any way affecting my medical care. I have been provided with a copy of this consent form. I agree to participate in the study voluntarily and not under duress or pressure or for remuneration.

Print Name of ParticipantDate and Signature of participant

_____ (dd/mm/yy)

I have accurately read the consent form to the potential participant, and the individual has had the opportunity to ask questions. I confirm that the individual has given consent freely.

Print Name of ResearcherDate and Signature of Researcher

_____ (dd/mm/yy)

Appendix 2: Human landing catch (HLC) protocol

Man-landing catches remain the gold standard protocol for directly sampling host-seeking mosquitoes. Typically this involves volunteers acting both as baits and catchers who sit down at a selected site and aspirate mosquitoes landing on their exposed limbs. The procedure is as follows (Doolan, 2002):

- Pairs of collectors will be seated with their arms and legs exposed to catch landing/biting mosquitoes from their exposed limbs using a hand-held mouth aspirator. In addition to background lighting, flashlights will also be provided in order to allow better visualisation of landing/biting mosquitoes.
- Each aspirated mosquito is placed in individual tubes and placed in a bag with the following information: location, date, site of collection (i.e. indoor, outdoor) and hour of collection.
- Collections will be made for the whole night (19:00-7:00 am) or part of the night (19:00-24:00).
- Pairs of collectors will be rotated every two hours to allow each member to rest. Collectors will also swap positions (e.g. indoor collectors will change to performing collections outdoors) in order to control for any bias introduced by individual-specific attractiveness.
- Collected mosquitoes will be morphologically identified and killed by freezing. Identified *An. gambiae s.l.* will be stored in Eppendorf tubes containing RNAlater or dried in silica gel until processing. Specimens should be labelled with their collection site, day and hour of collection.

Appendix 3: Human landing collection form

Serial No: Village : Date of collection:

	Outdoors						
Time	*A. gambiae s.l.*	*A. funestus*	*A. rufipis*	*A. pharoensis*	Other anophelines	*Mansonia species*	*Culicines*
19.00-20.00							
20.00-21.00							
21.00-22.00							
Total							
	Indoors						
19.00-20.00							
20.00-21.00							
21.00-22.00							
22.00-23.00							
23.00-24.00							
24.00-1.00							
1.00-2.00							
2.00-3.00							
3.00-4.00							
4.00-5.00							
5.00-6.00							
Total							

Appendix 4: *An. gambiae s.l.* species complex PCR procedure by Scott *et al.* (1993)

This PCR procedure identifies the *An. gambiae s.l* to their sibling species identity using species-specific primers. The expected product sizes are as follows: *An. gambiae s.s.* (GA; 390 bp), *An. arabiensis* (AR; 315 bp) and *An. merus* (ME; 466 bp) (Table S1).

Table S1. PCR reaction mixture for a 10 µl reaction to determine the *An. gambiae* complex to sibling species level.

Volume	Reagent
5 µl	REDTaq ReadyMix master mix
0.4 µl	UN (25 pmol/µl) [GTGTGCCCCTTCCTCGATGT]
0.4 µl	GA (25 pmol/µl) [CTGGTTTGGTCGGCACGTTT]
0.4 µl	AR (25 pmol/µl) [AAGTGTCCTTCTCCATCCTA]
0.4 µl	ME (25 pmol/µl) [TGACCAACCCACTCCCTTGA]
As appropriate	H_20 - fill up to total reaction volume
10 µl	**Total (add 1-3 legs or 1 µl of template DNA)**

PCR conditions

95^{o}C/5 mins x 1 cycle

(95^{o}C/30 secs, 50^{o}C/30 secs, 72^{o}C/ 30 secs) x 35 cycles

72^{o}C/ 5 mins x 1 cycle

Load 10 µl PCR product on 2% agarose gel.

Appendix 5: *An. gambiae sl.* molecular form PCR procedure by Wilkins *et al.* (2006)

This PCR procedure identifies the *An. gambiae s.s.* mosquitoes as *An. coluzzii* (formerly M molecular form) and *An. gambiae* (formerly S molecular form), based on primers that bind within species-specific SNPs of the internal transcribed spacer regions. The expected product sizes are as follows: *An. coluzzii* (427 bp) and *An. gambiae* (335 bp) (Table S2).

Table S2. PCR reaction mixture for a 10 µl reaction to identify *An. coluzzii* and *An. gambiae*.

Volume	Reagent
5 µl	REDTaq ReadyMix master mix
0.4 µl	M5 (25 pmol/µl) [CTTGGTCTGGAGACCGTTCCaTA]
0.4 µl	M3 (25 pmol/µl) [GACACGTCAACTAAGTCAACACATtAC]
0.4 µl	S5 (25 pmol/µl) [GCCCCTTCCTCGATGGaGC]
0.4 µl	S3 (25 pmol/µl) [CAACCGGCCCAAACGGcTT]
2.4 µl	H_20
10 µl	**Total (add 1 µl of template DNA)**

PCR conditions

95^{o}C/5 mins x 1 cycle

(95^{o}C/30 secs, 54^{o}C/30 secs, 72^{o}C/30 secs) x 30 cycles

72^{o}C/5 mins x 1 cycle

Load 10 µl PCR product on 2% agarose gel.

Appendix 6: Chapter 3 – Supplementary information

Supplementary figure

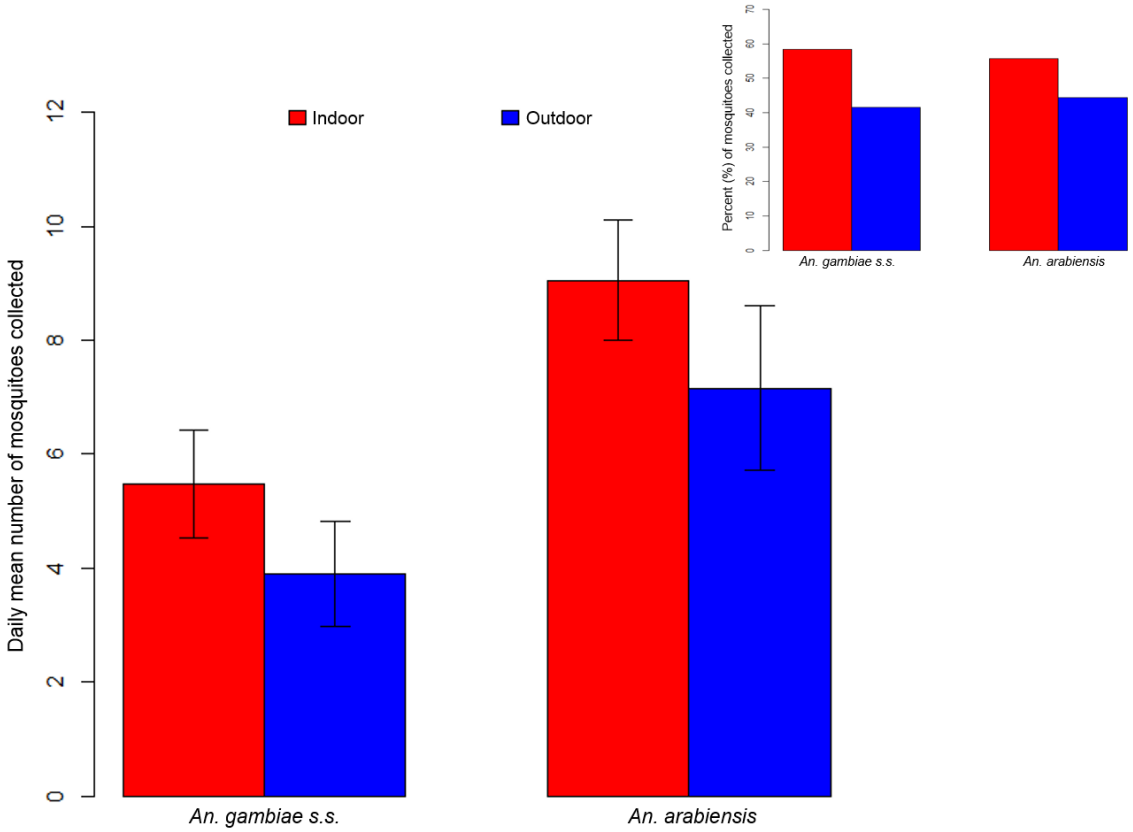

Figure S1. Daily mean number (main) and overall percentage (inset) of *An. gambiae s.s.* and *An. arabiensis* mosquitoes captured indoors and outdoors. Error bars represent the standard error of proportions.

Appendix 7: Chapter 4 – Supplementary information

Supplementary figures

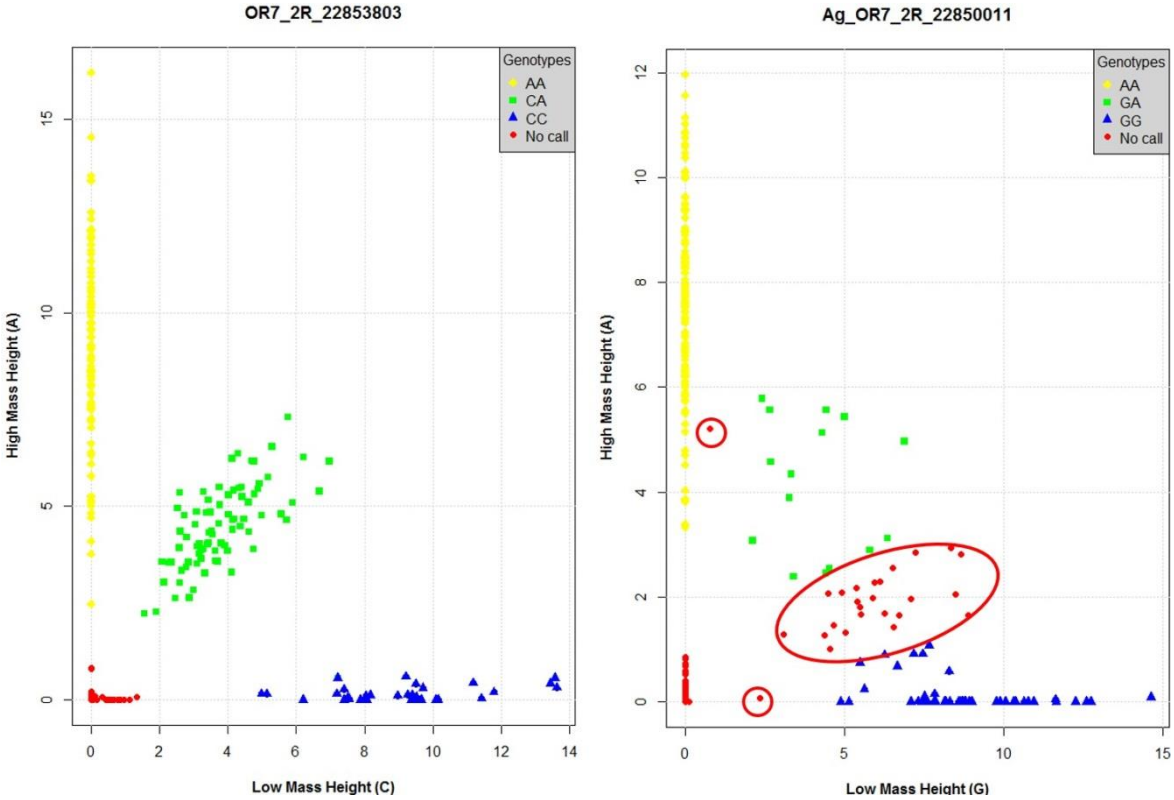

Figure S2. Examples of genotype class clustering using the Sequenom platform. The X and Y axes represent the mass height measurement for the two alleles, C, A (low mass allele versus high mass allele) at SNP, OR7_2R_22853803 and for the two alleles, G,A (low mass allele versus high mass allele) at SNP, OR7_2R_22850011. Each point represents a single sample. Points are coloured according to their genotype calls, for example, at OR7_2R_22853803, yellow points corresponds to samples with an AA genotype. Clear clustering between genotype calls were observed at the OR7_2R_22853803 SNP (left) whereas ambiguous genotype calls (circled red) were observed at OR7_2R_22850011. Ambiguous calls that could be assigned as homozygous (GG) OR heterozygous (GA) were re-assigned as null alleles. In both plots, only genotypes from the Gambian dataset are depicted.

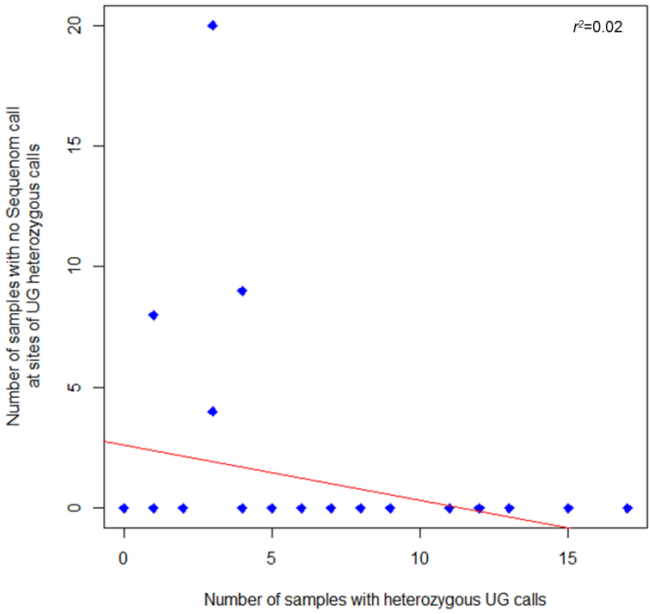

Figure S3. Correlation between heterozygous UG calls and missing data in Sequenom. Each point represents a SNP genotyped by Sequenom. There was no relationship between the likelihood of a heterozygous genotype (based on GATK's Unified Genotyper) and the likelihood of Sequenom failing to call a genotype.

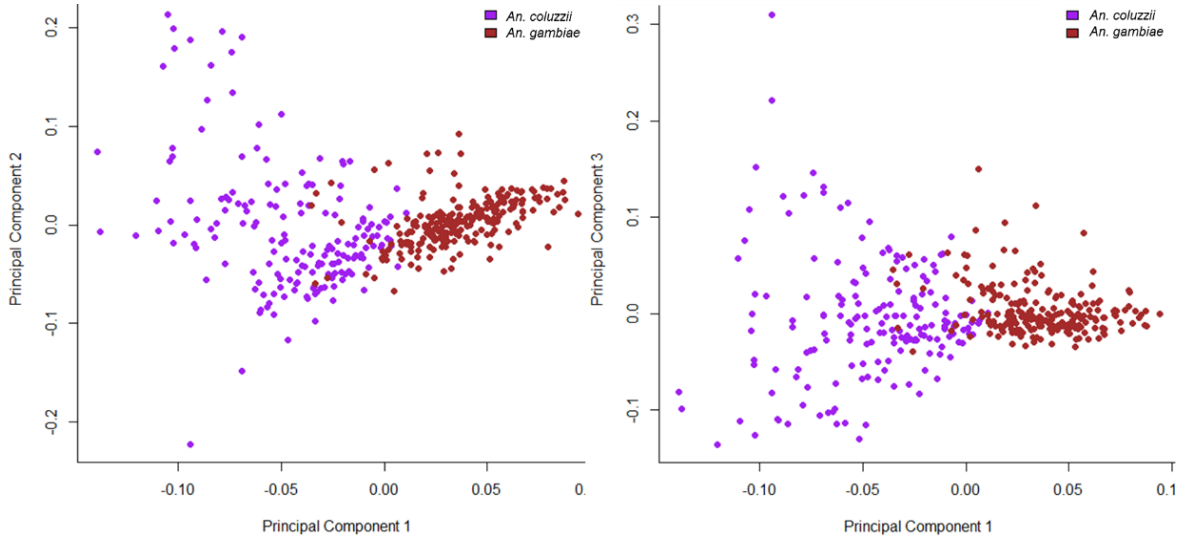

Figure S4. Incipient species relationship of *An. coluzzii* and *An. gambiae* from Uganda and Gambia using 34 autosomal SNPs. Results show the first three principal components of PCA. Each sample is represented by a single point. *An. coluzzii* appears to have greater genetic diversity than *An. gambiae,* showing a greater spread of individuals than *An. gambiae*.

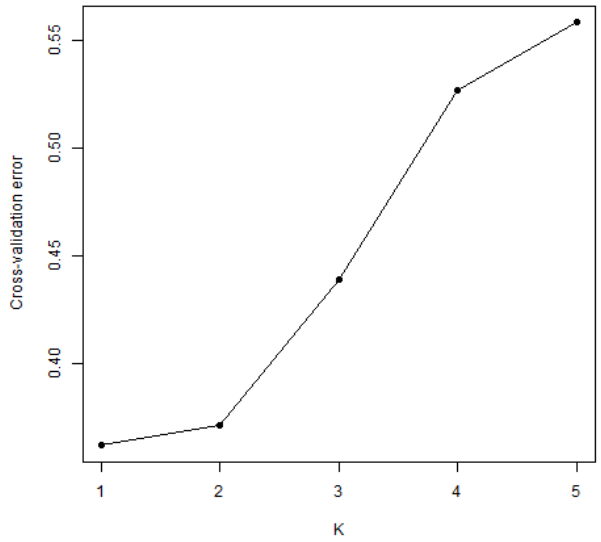

Figure S5. Results of ADMIXTURE analysis of indoor- and outdoor-collected *An. coluzzii* and *An. gambiae.* The graph shows the cross-validation (CV) error as a function of the number of simulated populations (*K*). Results suggests that a single population was the optimal number of populations, as indicated by *K* =1 having the lowest CV error

.

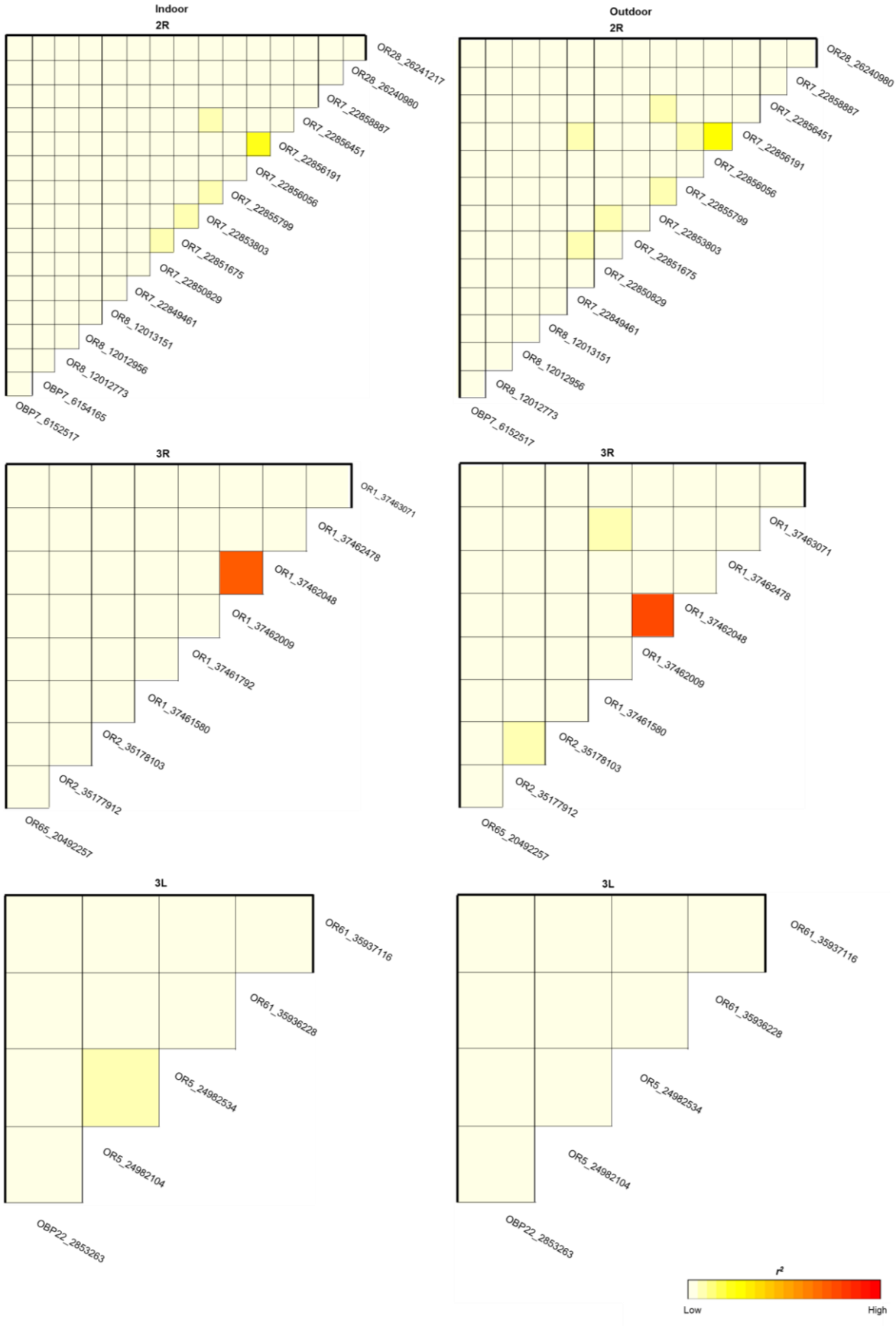

Figure S6. Linkage disequilibrium, r^2 heatmap for the indoor and outdoor population of *An. coluzzii* in Wali Kunda, Gambia. Pairwise LD between SNPs was calculated separately for each chromosome. Each square represents pairwise r^2 values between SNP markers; r^2 values range from 0 (white) to 1 (red).

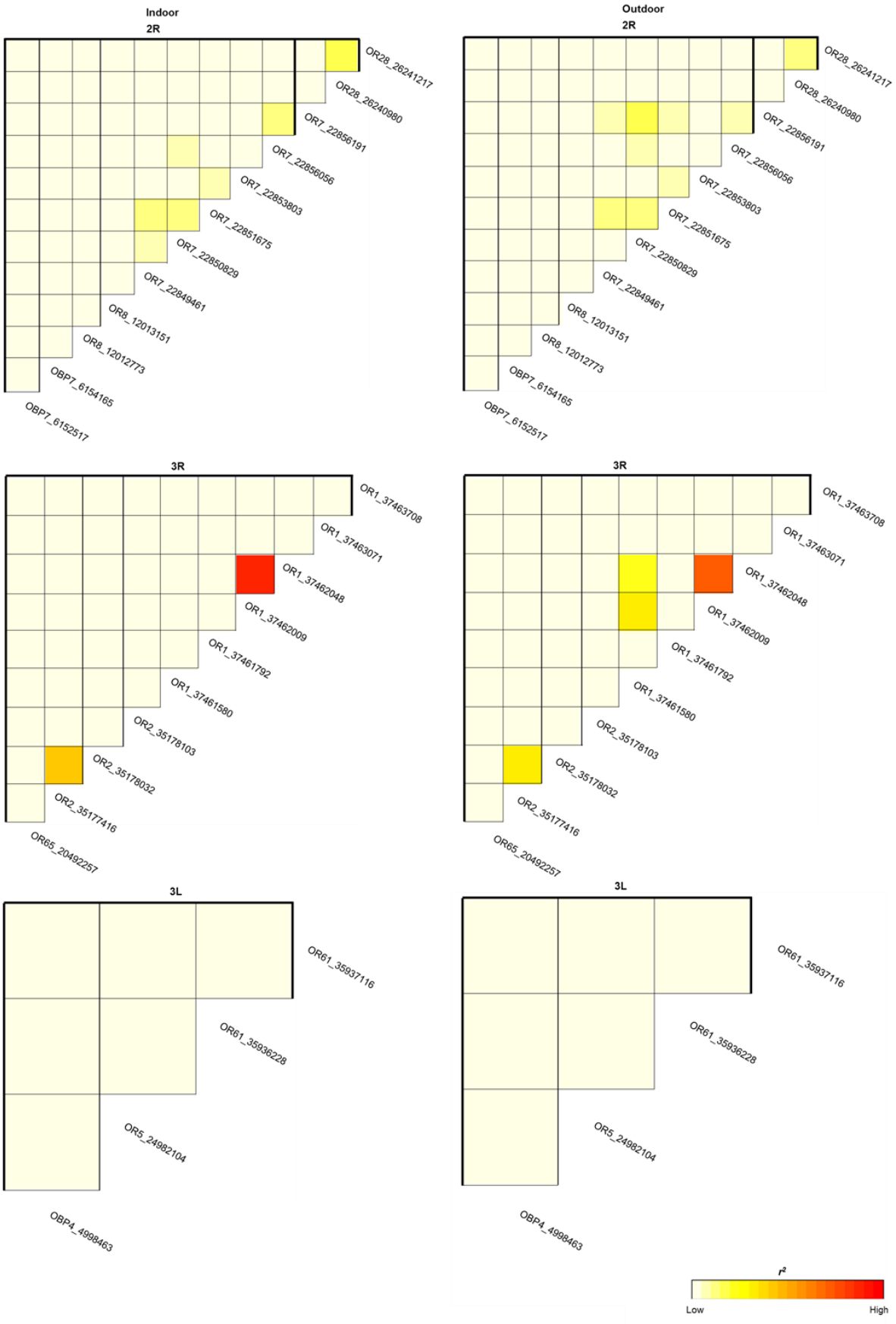

Figure S7. Linkage disequilibrium, r^2 heatmap for the indoor and outdoor population of *An. gambiae* in Uganda. Pairwise LD between SNPs was calculated separately for each chromosome. Each square represents pairwise r^2 values between SNP markers; r^2 values range from 0 (white) to 1 (red).

Supplementary table

Table S3. List of 78 SNPs that were assayed.

Multiplex ID	Chr	Coor	Gene	IUPAC	1st-PCR Primer	2nd-PCR Primer	Ext. primer direction	Ext. Primer	Ext. Primer Mass
W1	3R	35177638	*OR2*	R	ACGTTGGATGAGCAGTGCGCACAGCATCAT	ACGTTGGATGCGCGTGCACGTACAGATATG	R	TCACTCGTCACCCAT	4447.9
	2R	26240980	*OR28*	R	ACGTTGGATGGTTGAAGTTTGAAATCGAGC	ACGTTGGATGGTTCGACATGACTCTTTCCG	R	GGCACTCAGATCCAG	4562
	3R	35178032	*OR2*	R	ACGTTGGATGGTTGCTTTAGGGCCGCTATC	ACGTTGGATGTGCATGTACATCCCGTTCAC	R	CGTTCACCAGCTTCTA	4792.1
	3L	24982104	*OR5*	S	ACGTTGGATGCGTACTTCGGTGGACTAAAG	ACGTTGGATGTGGATTCTCATCGCCACAAG	R	TTGCCGAACAGTACTT	4856.2
	2R	22858638	*OR7*	R	ACGTTGGATGCGAATGTTCGGCATCAGGTC	ACGTTGGATGAACGACCACTACAAACGAGC	R	GCAGATGCAAGTCCAG	4915.2
	3R	35178695	*OR2*	Y	ACGTTGGATGGTGTGGCTGTTCTGGTCGTA	ACGTTGGATGAGTACAGCTTCAGGAACTGG	F	gCCGGGATGCAGCCGA	4932.2
	2R	12012840	*OR8*	R	ACGTTGGATGAGTGTTTGTGTGTGCGTTGG	ACGTTGGATGCGAAAATCATGTTAGCAAACC	R	AACAGCACAAACGAAAA	5198.4
	2R	22855174	*OR7*	Y	ACGTTGGATGCTCTCCATTTAATCTATC	ACGTTGGATGGTGGTACGGTGGGCCATTC	F	GCATAGTTTCAGGCGTA	5225.4
	2R	22853803	*OR7*	M	ACGTTGGATGAAGCAGGAAATGATGGTGCG	ACGTTGGATGCTGGAGGCTTTACACCACAC	F	TTTACACCACACCTTAGA	5402.5
	2R	22858492	*OR7*	S	ACGTTGGATGTCAGCTCGTTCACGTCGTC	ACGTTGGATGATCCGCAAGGTGTACTCCTG	R	GCTCGCCATGGTGCTGAT	5506.6
	2R	26241662	*OR28*	R	ACGTTGGATGGCAGTACATCTTACTCCTAC	ACGTTGGATGGCGCCTACAGGGTATGCAAT	R	aTGATTGATGGTTTGCAG	5584.6
	2R	22853351	*OR7*	K	ACGTTGGATGCACCCGTATCGATTTTCCAC	ACGTTGGATGACACAAGGACACAAAGGACG	F	tggTTTTGGGGGAGCTTT	5607.6
	3R	37462009	*OR1*	R	ACGTTGGATGGCTTATTTCAAGCTAAATACC	ACGTTGGATGGGAAGTGTCTTGAAGGTTAC	R	CGCTAAATCTTCACACATT	5706.7
	2R	22850829	*OR7*	R	ACGTTGGATGAGTTTTGTAACACGGATTGC	ACGTTGGATGATTGTACTACCAACGCAGTC	F	cCCAACGCAGTCATTAAAC	5725.8
	2R	22849461	*OR7*	K	ACGTTGGATGTTTGTCAGCAGTGCCAGAAG	ACGTTGGATGTTTTGCCAGCCTGTTCTGTG	F	aCGAAACGGTGAAAAACTT	5853.9
	3R	37463416	*OR1*	Y	ACGTTGGATGGAAGATACGGATCAGGCAAC	ACGTTGGATGGTATAGGGCTTGCGTTAGAG	F	GCGTACAGATGTAGAAACA	5869.9
	2R	22853925	*OR7*	S	ACGTTGGATGTTGATCTGAGCGGCATCTAC	ACGTTGGATGTTCGGATTTCCGTTCCTGCC	F	CCGTTCCTGCCATTCTCGTC	5970.9
	2R	22857729	*OR7*	Y	ACGTTGGATGGTACCTGGTAGATGAAAGAG	ACGTTGGATGGACGTACACGGTGGATATAC	R	ccccATGCAATGAGCGGACC	6072
	2R	26241217	*OR28*	R	ACGTTGGATGCCAGCTCGATTTTGGTTTGC	ACGTTGGATGTCATTCCATCCGAACACCTG	R	GCTACCATAGTTAGCTCATA	6076
	2R	22855656	*OR7*	Y	ACGTTGGATGGTCGTACGGATGATGGTTTT	ACGTTGGATGGGACTCAAAGAGAGCATAAC	R	cTTTCAATGACCTTAACACTC	6300.1
	3R	20492257	*OR65*	R	ACGTTGGATGGAGTCATTGTACTGCACAGG	ACGTTGGATGAGCTGCTGAATTTTGGACAC	R	TTTGTAGAAAAACTCCATCTT	6379.2
	3L	24982534	*OR5*	Y	ACGTTGGATGCAAGTGTTACGGTTAGAGGG	ACGTTGGATGTGGAAACCTACGGATTCTGC	F	aTACGGATTCTGCTACTTTGG	6427.2
	2R	22849144	*OR7*	K	ACGTTGGATGTTCGCTCGTTCGGTTGTTTG	ACGTTGGATGCTGTAAAAGAAAACTTGGG	F	GAAAACTTGGGATTATTTGTT	6490.2
	2R	22850011	*OR7*	R	ACGTTGGATGGCAAAACGAGAGAAAGAGAG	ACGTTGGATGTGGGATTGGTTTTTGCAGCG	R	cACATTCGCACAAAACCGTCCT	6608.3
	3R	35177912	*OR2*	R	ACGTTGGATGTGTGATACTTTAGACACTCC	ACGTTGGATGGATAGCGGCCCTAAAGCAAC	R	cCAGCGCCGGCACACTGTTCGC	6657.3
	2R	22858887	*OR7*	W	ACGTTGGATGCGCAAATCAGGTCACAAAGC	ACGTTGGATGAAATGCTTGGAGCCAAACGG	F	gggagACGGTAAAATCTCTCGT	6799.4
	3R	35177836	*OR2*	R	ACGTTGGATGTAGACGCTAGTAGACTCGAC	ACGTTGGATGAATGGCTTCCGGTTTGTGTG	F	ccccaCTCCCCTCCCCGAGGGAA	6875.5
	2R	22851675	*OR7*	R	ACGTTGGATGTGCGTTGGGTGTTTGAATAG	ACGTTGGATGGAAAAAAAACGCCTCGCTCG	R	AACGCCTCGCTCGCACAAAAACA	6955.6
	3R	35178957	*OR2*	R	ACGTTGGATGTGTGGAAGAAGGGTGTGTTG	ACGTTGGATGGAAGCATATAATTTGCTCA	R	CATATAATTTGCTCAAAAAGTCA	7014.6
	2R	22856056	*OR7*	R	ACGTTGGATGGATGTCATGTTCTGCTCCTG	ACGTTGGATGGCGGATGTAAAGACAACCAC	F	gagtcAAGACAACCACGTTTGCT	7032.6
	3R	37461580	*OR1*	R	ACGTTGGATGAAAAGCGCCCTTGCCATAAC	ACGTTGGATGGATAGTTTATAGCACCGTCG	R	tCGCACCTGATTTTATCTTTATCT	7234.7
	3R	37462441	*OR1*	Y	ACGTTGGATGACAGCTTCGGCGATGAAGTG	ACGTTGGATGAGCTTGCAGCAGTGTCATAC	F	TGTCATACAAATGATAATTACAGA	7367.8
	3R	37461792	*OR1*	R	ACGTTGGATGGCATGGAATCAGAGTAATGG	ACGTTGGATGGGCTAATTGAAAACTAGCGG	F	aTAATTGAAAACTAGCGGGCTTAC	7384.8
	3L	35936228	*OR61*	M	ACGTTGGATGGTTGTGGTTTCGGTTTCTAC	ACGTTGGATGGTGAATGGCCGATATCGTTG	R	agaaaGAAGTAGTCTGAACTGTCG	7449.9
	2R	12012956	*OR8*	M	ACGTTGGATGGCTAACATGATTTTCGGTTTG	ACGTTGGATGACCGTTACTTACATCGCATC	R	gTACATCGCATCTAGTAGTCATCGT	7632
	3L	35937116	*OR61*	R	ACGTTGGATGAGGTAGATCATACCGTTCGC	ACGTTGGATGGCCCATTTTCCTTCGTACAG	F	TCCTTCGTACAGTACTATGGTACCTT	7887.1
	3R	37463132	*OR1*	Y	ACGTTGGATGGCAAGCTTAACTGCACACTG	ACGTTGGATGGGAGAGTCGTTTGTTAGCAC	F	GTTAGCACTCTTTTTAGTCATATTTC	7892.1
	3R	37463708	*OR1*	K	ACGTTGGATGACCGGCTAAATACAGAGCAG	ACGTTGGATGCTGGAAACAGTACTTGGTT	F	CTGGAAACAGTACTTGGTTTCAATTT	7975.2
	3R	35178780	*OR2*	R	ACGTTGGATGTACGACCAGAACAGCCACAC	ACGTTGGATGGAACATAAACCGCAACCCAC	R	ACCGCAACCCACAGCCGAAAATGCTGAT	8520.6
	X	**153411172**	***G6PD***	**S**	**ACGTTGGATGAGGTGGAAGACCATCCACTG**	**ACGTTGGATGCATTCTCTCCCATTGTGGAC**	**F**	**GCGCCCAGAGCTGAAGGGACAGTTG**	**7757**
W2	2R	22855496	*OR7*	M	ACGTTGGATGCACTACATCCGTTCGTCCAG	ACGTTGGATGACCGGTATTTCAGGATGAGG	R	ACAATCCCCACCCAG	4450.9
	3L	24983049	*OR5*	R	ACGTTGGATGACCCTTTGACTTTTTAAG	ACGTTGGATGTATGATTAGTCGAGGAAACG	R	GCTCGACAACGGTCT	4553
	3L	40214764	*OBP26*	S	ACGTTGGATGTACCACTTTTTGCTTGCCCC	ACGTTGGATGAGCGAAGTTGATGCACTTCC	F	AGTGCATAGTGCACG	4617
	2R	22857774	*OR7*	Y	ACGTTGGATGTGCTCGATAAGGCAACCAAC	ACGTTGGATGCATTGCATTCCACGGATACC	F	GGACTTGATGGGCAG	4673
	2R	22849536	*OR7*	R	ACGTTGGATGTGTTTTCACTCTCCCATACC	ACGTTGGATGACGAAGGTTTTTGCCTCCTC	F	cTGCCTCCTCGGACCC	4754.1
	2R	22852931	*OR7*	Y	ACGTTGGATGCACACACGTGAAGGGAAAGC	ACGTTGGATGTCATCGCGTCAGGATGATTC	F	cCAATTTGCCGCGCTG	4833.1
	2R	26239795	*OR28*	R	ACGTTGGATGAAGACAAACGGGTAACGGTG	ACGTTGGATGACTACCCTGTACTGGCTTTC	R	CCGGCCGGCATCTTACC	5107.3
	2R	12013151	*OR8*	S	ACGTTGGATGTGTTCGTGATGTTACGCTCC	ACGTTGGATGAACAGGAAATAGCTCTCCCC	F	cctCTCTCCCCCTACCTT	5257.4
	3R	37462478	*OR1*	Y	ACGTTGGATGAGCTTGCAGCAGTGTCATAC	ACGTTGGATGTTTCACAGCTTCGGCGATG	R	CGGCGATGAAGTGCAGG	5300.5
	3R	37462048	*OR1*	S	ACGTTGGATGGGACTCTTAAAGATGTTCAC	ACGTTGGATGGTGTGAAGATTTAGCGTAACC	F	AACCTTCAAGACACTTCC	5387.5
	2R	22855261	*OR7*	S	ACGTTGGATGTTGTAGGGACGGATAAGCTC	ACGTTGGATGAGCAATTAGAGCAACACGGC	F	ccGCAACACGGCACTAAG	5462.6
	3L	24981519	*OR5*	W	ACGTTGGATGAACGACGCTATCGCTACAAG	ACGTTGGATGTCTCTAAATCTGGATAGCCG	R	AGAAGCTTAAAAATGCCA	5524.6
	2R	22855799	*OR7*	Y	ACGTTGGATGATCAGCTCCGATTTGGAACC	ACGTTGGATGCGGGCAGGGTATTATGCGAT	R	aagcCCGGCCCAACTCTTC	5693.7

2R	35177416	*OR2*	R	ACGTTGGATGCAGCGCCATTACCTGCTC	ACGTTGGATGATAGAGCAATCAGCTGGCAC	R	CGCAGATGTTTGCCTTCTA	5769.8
2R	22850930	*OR7*	Y	ACGTTGGATGCAGTTTTTTTTTGTAAGCGG	ACGTTGGATGTGCAATCCGTGTTACAAAAC	F	TCTATTTGAGCGTTAACAG	5817.8
3R	35178103	*OR2*	S	ACGTTGGATGTGAACGGGATGTACATGCAG	ACGTTGGATGACCGTACGGCGTCACGATAC	R	cccccACCCCGACCTACCAG	5912.8
3L	24982615	*OR5*	M	ACGTTGGATGAACGGCGATACCAGAGGCTA	ACGTTGGATGTTCCCCTCTAACCGTAACA	F	CTCTAACCGTAACACTTGTA	6036
2R	22856666	*OR7*	M	ACGTTGGATGCACCGCCATAACAATTGGG	ACGTTGGATGCGTGCGAAATAATTTATTCC	R	CATCCCCATTAAAACAAAGT	6038
2R	12012773	*OR8*	Y	ACGTTGGATGATCCCACGGCTGGATATTG	ACGTTGGATGCGCACACACAAACACTAGAG	R	aACAAACACTAGAGAACACG	6121
2L	45011466	*OBP47*	W	ACGTTGGATGAACGGCGTGACGCAACATTC	ACGTTGGATGCAATCCCTGCCTTAAAGGAC	F	TTTCCACTACTATTTCCACTC	6242.1
2R	6152517	*OBP7*	R	ACGTTGGATGATGGAACCGTACCGACAGC	ACGTTGGATGTCCGGGATGATGTAAAGCAG	F	caTCGCCTCGTCCACCACGTC	6263.1
2R	22851925	*OR7*	R	ACGTTGGATGTGTAAAATGAGCCACGCGAC	ACGTTGGATGACGGCTTACGGTTTCGGGC	F	AACAAGAATAAAACCACCTCC	6345.2
2R	22858426	*OR7*	R	ACGTTGGATGAGAAGTTCTCCGAGTTGACC	ACGTTGGATGACGACGTGAACGAGCTGAC	R	agttAACGAGCTGACCGCCAA	6424.2
2R	22854881	*OR7*	R	ACGTTGGATGGACCTGGAGCGGTTATTTAA	ACGTTGGATGTCTCATGTTTTGCCTGACGG	R	gGAAAAAAGAGCGATTTCTTC	6478.2
3L	2853263	*OBP22*	W	ACGTTGGATGGTATGGAGTCGATTTGCAAG	ACGTTGGATGCTTAGCTTCTAGACAGGCAC	F	cCATCATCACTGTTCCCCAGAT	6590.3
2R	22854102	*OR7*	R	ACGTTGGATGCATTTTGAGGTTTACACAT	ACGTTGGATGGATCCTTTTCTGTGGTAGCG	F	cttacTGTAACAATCAGGCGTC	6694.4
2R	26241481	*OR28*	M	ACGTTGGATGCAACAAATTACCGAGGAGGC	ACGTTGGATGGTAGGCTTCTGGGCTTTAAC	R	aaTGGGCTTTAACCATGAATGT	6773.4
2R	22856191	*OR7*	Y	ACGTTGGATGACGGGATAATCCGTTGGGAC	ACGTTGGATGCCATCGAAAACAGCAGGAAG	F	TGAAGCATAAAATGTGTTGTAG	6837.5
2R	12014631	*OR8*	K	ACGTTGGATGTTGGATGAGAGGGTGTGTGG	ACGTTGGATGCAACAAAAAGCGGCACGTTC	F	tcaCGCCAATTATTGATTGATTT	7003.6
3L	4998463	*OBP4*	Y	ACGTTGGATGTCGAAAAGCTGATTTCACCC	ACGTTGGATGAGGAACTGTGAGTATGCTCG	F	aaATGGTATGGTTTCCATTCTTC	7019.6
3R	37463071	*OR1*	Y	ACGTTGGATGCCCACATGATGATGGTGAAG	ACGTTGGATGACTCTCCTCCAAATGTAGCC	R	gatcaAAATGTAGCCCCGTTTTAC	7311.8
2R	22855688	*OR7*	R	ACGTTGGATGCAGACGGACTCAAAGAGAGC	ACGTTGGATGCCACTGTCTCGCTTAGCTTG	F	aagtTCATTGAAATGAACTACCAG	7368.8
2R	12013957	*OR8*	K	ACGTTGGATGACGGTGTGTACCAGTACAAC	ACGTTGGATGGCAGCAGGTTAAACATCCAG	F	ggaggTCGAGAACGATTTCATCAC	7401.8
2R	6154165	*OBP7*	R	ACGTTGGATGGTATTCGCACAGATTGGAAG	ACGTTGGATGTGCTTTCGGCCGTAAGTTTG	R	aaGGCCGTAAGTTTGATAAGCACG	7425.8
2R	22856451	*OR7*	Y	ACGTTGGATGGCTGGTTTAGCTTTAGCTC	ACGTTGGATGCGCTTCGGGATGTCATAAAG	R	ccttGGTGTTTTTAACATTACTTCC	7563.9
3R	20492001	*OR65*	W	ACGTTGGATGTTTGCACGGTACCACCAGAG	ACGTTGGATGCGTCGTTGCGTATGATTTTG	F	ttCGTATGATTTTGTTCATTTTTCG	7625
2R	20493567	*OR65*	M	ACGTTGGATGTACAACGAACTACTGCAATG	ACGTTGGATGAAATGGGGAGAATGCAACGG	R	TTCTACTTATCTATCTATCTACCCAGT	8110.3
13	**1445**	***PfTRAP***	**Y**	**ACGTTGGATGTGGCTTTTCATGTTCTTCCC**	**ACGTTGGATGGAAACACGTCCACATGGTAG**	**R**	**TTCATGTTCTTCCCTTTCAGGAT**	**6947**

Columns include the chromosome (chr), coordinates (coor) and gene of the assayed SNP. Primer sequences, the direction, mass of the extension primer and the assay pass rate are also included. Each multiplex contained 39 SNP assays. The IUPAC base code represents the two possible alleles. Ext = extension.

Table S4. Observed heterozygosity (He) and inbreeding coefficient, F_{is} estimates in indoor and outdoor-collected *An. coluzzii* and *An. gambiae*.

		An. coluzzii						*An. gambiae*					
		H_o			F_{is}			H_o			F_{is}		
Chr	SNP	Indoor	Outdoor	All	Indoor	Outdoor	All	Indoor	Outdoor	All	Indoor	Outdoor	All
2R	OBP7_2R_6152517	0.027	0.056	0.038	-0.014	-0.029	-0.02	0.164	0.138	0.154	-0.01	0.068	0.018
2R	OBP7_2R_6154165	0.009	-	0.006	-0.005	-	-0.003	0.07	0.106	0.083	-0.036	-0.056	-0.043
2R	OR8_2R_12012773	0.5	0.441	0.478	-0.112	0.014	-0.064	0.378	0.482	0.417	0.212**	-0.0001	0.133
2R	OR8_2R_12012956	0.036	0.028	0.033	-0.019	0.652**	0.383**	-	-	-	-	-	-
2R	OR8_2R_12013151	0.009	0.042	0.022	0.795	-0.022	0.489**	0.043	0.047	0.045	-0.022	-0.024	-0.023
2R	OR7_2R_22849461	0.209	0.197	0.204	-0.038	0.185	0.061	0.07	0.069	0.07	-0.036	-0.036	-0.036
2R	OR7_2R_22850829	0.38	0.4	0.388	0.152	0.1667	0.161	0.199	0.19	0.196	0.148	0.004	0.103
2R	OR7_2R_22851675	0.259	0.319	0.282	0.201	0.226	0.22**	0.285	0.192	0.251	0.178	0.469***	0.286***
2R	OR7_2R_22853803	0.378	0.338	0.363	0.124	0.26	0.18*	0.468	0.563	0.504	0.056	-0.126	-0.012
2R	OR7_2R_22855799	0.468	0.38	0.434	-0.164	-0.066	-0.13	-	-	-	-	-	-
2R	OR7_2R_22856056	0.231	0.186	0.213	-0.058	0.13	0.015	0.112	0.126	0.117	0.051	-0.067	0.006
2R	OR7_2R_22856191	0.318	0.296	0.309	-0.012	0.258	0.117	0.056	0.092	0.07	0.171	-0.048	0.075
2R	OR7_2R_22856451	0.055	0.099	0.072	-0.028	-0.05	-0.037	-	-	-	-	-	-
2R	OR7_2R_22858887	0.009	0.044	0.023	0.904*	0.643***	0.788*	-	-	-	-	-	-
2R	OR28_2R_26240980	0.144	0.141	0.143	0.033	-0.076	-0.006	0.458	0.483	0.467	-0.048	0.028	0.003
2R	OR28_2R_26241217	0.009	-	0.006	-0.005	-	-0.003	0.507	0.547	0.523	-0.063	-0.097	-0.07
3R	OR65_3R_20492257	0.009	-	0.006	0.795***	1*	0.854***	0.014	0.034	0.02	0.66***	0.382	0.534***
3R	OR2_3R_35177416	-	-	-	-	-	-	0.112	0.138	0.122	-0.06	-0.074	-0.065
3R	OR2_3R_35177912	0.126	0.155	0.137	-0.067	-0.084	-0.074	-	-	-	-	-	-
3R	OR2_3R_35178032	-	-	-	-	-	-	0.077	0.046	0.066	-0.04	-0.024	-0.034
3R	OR2_3R_35178103	0.27	0.261	0.267	0.119	-0.053	0.065	0.246	0.259	0.25	-0.09	-0.067	-0.08
3R	OR1_3R_37461580	0.119	0.157	0.134	0.413***	0.067	0.293**	0.014	0.011	0.013	-0.007	-0.006	-0.007
3R	OR1_3R_37461792	0.036	-	0.022	-0.018	-	-0.011	0.035	0.034	0.035	-0.018	-0.018	-0.018
3R	OR1_3R_37462009	0.291	0.371	0.322	0.058	-0.103	-0.007	0.056	0.034	0.048	-0.029	-0.018	-0.025
3R	OR1_3R_37462048	0.279	0.394	0.324	0.136	-0.083	0.046	0.063	0.046	0.057	-0.032	-0.024	-0.029
3R	OR1_3R_37462478	0.054	0.056	0.055	-0.028	-0.029	-0.028	-	-	-	-	-	-
3R	OR1_3R_37463071	0.108	0.085	0.099	0.085	-0.044	0.048	0.301	0.322	0.309	-0.059	-0.024	-0.044
3R	OR1_3R_37463708	-	0.032	0.012	-	-0.016	-0.006	0.035	0.046	0.039	-0.018	-0.024	-0.02
3L	OBP22_3L_2853263	0.109	0.07	0.094	-0.058	-0.037	-0.049	-	-	-	-	-	-
3L	OBP4_3L_4998463	-	-	-	-	-	-	0.145	0.149	0.148	0.08	-0.081	0.025
3L	OR5_3L_24982104	0.509	0.529	0.517	-0.14	-0.095	-0.116	0.414	0.395	0.407	0.027	0.063	0.04
3L	OR5_3L_24982534	0.102	0.085	0.095	0.586***	0.525***	0.568*	-	-	-	-	-	-
3L	OR61_3L_35936228	0.156	0.169	0.161	-0.085	0.049	-0.024	0.014	0.057	0.03	-0.007	0.256	0.207
3L	OR61_3L_35937116	0.236	0.268	0.249	-0.004	-0.062	-0.027	0.119	0.08	0.104	-0.063	-0.042	-0.055

All =indoor and outdoor samples combined into a single population. Significant departures from HWE under a chi-square test are represented: $^{*}p$ <0.05, ** p <0.01, *** p <0.001.

Table S5. Single-locus F_{st} and gene flow, *Nm* distribution for each species/population comparison.

	An. gambiae vs *An. arabiensis*		*An. coluzii:* Indoor vs Outdoor		*An. gambiae:* Indoor vs Outdoor	
SNP	F_{st}	*Nm*	F_{st}	*Nm*	F_{st}	*Nm*
OBP7_2R_6152517	0.047	5.069	0	-	0	-
OBP7_2R_6154165	0.019	12.908	0	-	-	-
OR8_2R_12012773	0.28	0.643	0	-	0	-
OR8_2R_12012956	-	-	0.003	83.083	0	-
OR8_2R_12013151	0.012	20.583	0	-	0	-
OR7_2R_22849461	0.015	16.417	0	-	0	-
OR7_2R_22850829	0.782	0.07	0	-	0	-
OR7_2R_22851675	0.041	5.848	0.01	24.75	0.002	124.750
OR7_2R_22852931	0.015	16.417	-	-		-
OR7_2R_22853803	0.342	0.481	0	-	0	-
OR7_2R_22855688	0.023	10.62	-	-	-	-
OR7_2R_22855799			0.001	249.75	0	-
OR7_2R_22856056	0.033	7.326	0	-	0	-
OR7_2R_22856191	0.43	0.331	0.01	24.75	0	-
OR7_2R_22856451			0	-	0	-
OR7_2R_22856666	0.029	8.371	-	-	-	-
OR7_2R_22858887			0	-	0	-
OR28_2R_26240980	0.257	0.723	0	-	0.033	7.326
OR28_2R_26241217	0.277	0.653	0.005	49.75	0.007	35.464
OR65_3R_20492257	0.005	49.75	-	-	0	-
OR2_3R_35177416	0.032	7.563	0	-	0	-
OR2_3R_35177912			0	-	0.006	41.417
OR2_3R_35178032	0.014	17.607	0	-	0	-
OR2_3R_35178103	0.079	2.915	0	-	0	-
OR2_3R_35178695	0.048	4.958	-	-	-	-
OR2_3R_35178780	0.014	17.607	-	-	-	-
OR2_3R_35178957	0.013	18.981	-	-	-	-
OR1_3R_37461580			-	-	0.007	35.464
OR1_3R_37461792	0.004	62.25	0.009	27.528	0	-
OR1_3R_37462009	0.008	31	0	-	0	-
OR1_3R_37462048	0.011	22.477	0	-	0	-
OR1_3R_37462441	0.925	0.020	-	-		
OR1_3R_37462478			0	-	0	-
OR1_3R_37463071	0.112	1.982	0	-	0	-
OR1_3R_37463132	0.054	4.38	-	-	-	-
OR1_3R_37463708	0.005	49.75	0	-	0	-
OBP22_3L_2853263	-	-	0	-	0	-
OBP4_3L_4998463	0.045	5.306	-	-	0	-
OR5_3L_24982104	0.152	1.395	0.005	49.75	0	-
OR5_3L_24982534	-	-	0.005	49.75	0	-
OR61_3L_35936228	0.004	62.25	0	-	0.023	10.620
OR61_3L_35937116	0.052	4.558	0	-	0	-

Negative F_{st} values were rounded up to zero and positive values were rounded off to two decimal places.

Appendix 8: Chapter 6 – Supplementary information

Supplementary notes

- Quality by Depth (QD) < 5.0

The quality score (i.e. confidence in the variant) of the SNP normalized for the amount of coverage available. Score below 5 are indicative of false positive calls and artefacts.

- Mapping Quality (MQ) < 40.0

MQ is a measure of how likely a read has been mapped properly to the reference genome. Typically an MQ > 30 is chosen for most experiments.

- Fisher Strand (FS) > 60.0

If there is a disagreement between the genotypes inferred from the forward and reverse strand, it is said as having a strand bias. SNPs with high FS values are more likely to have strand bias. More bias is indicative of false positive calls.

- Read Position Rank Sum Test (ReadPosRankSum) < -8.0

Estimates bias of an allele with respect to their position within the read. Alleles near the ends of reads are likely artefacts of sequencing. Values close to zero indicate that there is no bias while negative values suggest that an alternate allele is found at the ends of the reads more often than the reference allele, indicating a sequencing error.

- Homopolymer Run (HRun) > 4

Homopolymers are sequences of repetitive bases. This filter excludes variants that have homopolymers of more than four bases either side, to limit false positive due to incorrect alignment to the reference genome.

- Coverage (DP) < 4000 & > 7000

DP represents the number of times a given SNP has been sequenced by independent reads across all samples. SNPs with extremely low DP suggest that it is within a poorly sequenced region and thus, likely a sequencing error. In contrast, high DP is characteristic of highly repetitive sites which have a greater chance of being incorrectly mapped to the reference genome. We therefore removed SNPs with extremely low and high DP values; this cut-off was determined based on the distribution of DP across the entire genome.

- Repeat Masked

Repeat Masked sites correspond to regions of the genome that are repetitive. Variants within these sites were excluded.

Supplementary results

This appendix summarises the top GWAS hits, as defined by $p < 10^{-5}$ in the Gambian samples. Not all SNPs were followed-up during the meta-analysis of the Gambian and Ugandan datasets. Hits were investigated via the Vectorbase, Entrez Gene and UniProt online databases as well as, the SnpEff program. To highlight the potential functional significance of the observed associations with outdoor-biting behaviour, the results are presented with regard to the nearest gene (200 kb). SNPs located within a region that correspond to the upper 0.1-10% of the F_{st} distribution are also mentioned.

1. AGAP004798 (*COPB2*)

The SNP, 2L:3670888, is located within a non-coding region of chromosome arm, 2L. It is approximately 25 kb downstream of the gene, *COPB2,* which encodes a subunit of the coatomer protein complex (COP). The COP functions primarily in intracellular protein transport by facilitating Golgi budding and vesicular trafficking. This gene is conserved across a variety of organisms including humans, chimpanzees, *Drosophila*, mosquitoes and *Caenorhabditis elegans*.

2. AGAP005412

2L:15326158 is within a non-coding region of chromosome arm, 2L and is approximately 8.5 kb upstream of the gene, AGAP005412 which has no known function yet.

3. AGAP005536

The nearest gene to 2L:16769668 is AGAP005536, which is located 30 kb upstream. AGAP005536 has a putative role in signal transduction by catalysing the hydrolysis of phosphodiester bonds.

4. AGAP005737

2L:20020494 is located in the intronic region of AGAP005737 which putatively encodes a protein with oxidoreductase activity. This gene is conserved in *Drosophila,* mosquitoes, *Magnaporthe oryzae, Arabidopsis thaliana* and rice.

5. AGAP006031 (*Nup54*)

2L:25464003 is located in the intronic region of *Nup54,* which encodes a nucleoporin protein. This SNP is also found within a genomic region exhibiting an F_{st} estimate that corresponds to the top 10% of the F_{st} distribution. *Nup54* is a component of the nuclear pore complex, a complex required for nuclear transport.

This gene is conserved across a variety of organisms including humans, chimpanzees, fruit flies, mosquitoes and frogs.

6. AGAP006208, AGAP006207

SNPs, 2L:28254734 and 2L:28257115 (r^2 = 0.03) are found 91 bp upstream of AGAP006207 and within the exonic region of AGAP006208, respectively. Both SNPs are also within a region displaying an F_{st} estimate that corresponds to the top 1% of the F_{st} distribution. Given the proximal location of 2L:2825473, we posit that it is within the promoter region of AGAP006207. 2L:28257115 is a synonymous variant and thus, is likely to have a low impact on AGAP006208. Both genes encode carboxypeptidase B enzymes which play a role in proteolysis.

7. AGAP006598

SNP, 2L:34633056 is 488bp downstream of AGAP006598, which has no known function yet. This gene has higher, non-significant expression in non-blood fed mosquitoes (Marinotti *et al.*, 2006).

8. AGAP006773

2L:38112955 is located in the three prime untranslated (3'UTR) region of AGAP006773; this gene has no known function.

9. AGAP006991

SNP, 2L:40496599 is 222 bp downstream of AGAP006991. This gene has a predicted role chitin binding.

10. AGAP007100

2L:42671746 is 4.6 kb upstream of AGAP007100; this gene has no known function. It is also located within a region corresponding to the top 2.5% of the F_{st} distribution.

11. AGAP001616

2R:6826788 is located within a region corresponding to the top 10% of the F_{st} distribution, and is found upstream (1.9kb) of AGAP001616. The function of AGAP001616 is unknown.

12. AGAP001657 (*Hexamerin*)

SNPs, 2R:7471983 and 2R:7471984 are completely linked (r^2 = 1) and are 65 kb downstream of the hexamerin gene, AGAP001657. Hexamerins are storage proteins that serve as nitrogen and amino acid sources, which are required by pupae and adults during metamorphosis and reproduction (Telfer and Kunkel, 1991).

13. AGAP002178 (*hth*)

2R:17043652 is intronic of the transcription factor, *hth* ("homeobox protein homothorax"). In *Drosophila, hth* has been implicated in determining the patterning of embryonic cuticle, organisation of the peripheral nervous system and antennal identity as well as, regulating the expression of antennal-specific genes such as *sal* and *dac.*

14. AGAP002238

SNPs, 2R:18031010, 2R:18031141 and 2R:18031566 (r^2 = 0.07-0.73) are found within a region corresponding to the top 2.5% of the F_{st} distribution. 2R:18031010

is the second, highest GWAS hit (p= 7.62 x 10^{-6}). All three SNPs are located in an intergenic region and is approximately 3.5 kb downstream of the transcription factor gene, AGAP002238. AGAP002238 displayed higher, although non-significant expression in non-bloodfed mosquitoes.

15. AGAP002793 (*Slit1*)

SNPs, 2R:27532509 and 2R:27532512 (r^2 = 0.84) are intronic of the gene, *Slit1* ("Slit homolog 1 protein"). They are also located within a region corresponding to the top 2.5% of the F_{st} distribution. *Slit1* is protein coding and putatively plays a role in axonal guidance. It is conserved in human, chimpanzees, fruit flies, mosquitoes, *C. elegans* as well as, other organisms.

16. AGAP002916

2R:29597194 is an intergenic SNP with the nearest gene, AGAP002916 being approximately 5kb away. AGAP002916 has no known function but it is located within a region corresponding to the upper 5% of the F_{st} distribution.

17. AGAP003746, AGAP013391

This region contains two unlinked SNPs, 2R:42843109 and 2R:43500726 (r^2 = 0.078). 2R:42843109 is 8.2 kb upstream of AGAP003746 while 2R:43500726 is intronic of AGAP013391. As of yet, both genes have no annotated functions. Both genes are also located within a region corresponding to the upper 10% and 5% of the F_{st} distribution, respectively.

18. AGAP003871

2R:44788020 is found within the intronic region of the putative transcription factor gene, AGAP003871. This gene is conserved in humans, chimpanzee, mouse, rat,

mosquitoes and a variety of other organisms. This SNP is also located within a region corresponding to the upper 5% of the F_{st} distribution.

19. AGAP003940

As well as showing a significant association, 2R:46589775 is also located within a region corresponding to the upper 1% of the F_{st} distribution. 2R:46589775 is intronic of the putative transcription factor gene, AGAP003940. AGAP003940 has conserved structures that are similar to the *Drosophila* transcription factor, *Forkhead* which regulates tissue-specific expression.

20. AGAP003945 (*PTBP2*)

This region contains three SNPs, 2R:46726249, 2R:46726250 and 2R:47281060. SNPs, 2R:46726249, 2R:46726250 are almost, completely linked (r^2 = 0.99) and are intronic of the gene, *PTBP2.* By contrast, 2R:47281060 is located in a non-coding region and is ~3.5kb downstream of the closest gene, AGAP003971. *PTBP2* ("polypyrimidine tract-binding protein 2") and AGAP003971 putatively encode a transcription factor and a serine-type endopeptidase, respectively. Additionally, these SNPs are within a region corresponding to the upper 5% and 1% of the F_{st} distribution, respectively.

21. AGAP004675 (*GPRMAC2*)

SNPs, 2R:61418622, 2R:61418659, 2R:61441980 are almost, completely linked (r^2 = 0.97-0.99) and are located downstream (676 bp – 24 kb) of a putative G-protein coupled receptor, muscarine acetelycholine receptor 2 (*GPRMAC2*).

22. AGAP010833 (*CLIPB14*)

SNP, 3L:11681894 is located within a region corresponding to the upper 5% of the F_{st} distribution, with the nearest gene (29 kb upstream away) encoding a Clip-domain serine protease (*CLIPB14*). *CLIPB14* has a predicted role in proteolysis.

23. AGAP011128 (*DPR*), AGAP011187 (*TOLL10*)

This region contains two SNPs: 3L:17529965 and 3L:18593146. Both SNPs are found in the intergenic region with *DPR* ("defective proboscis extension response") and *TOLL10* as the closest genes, respectively (~26 kb and 33.6 kb downstream, respectively). In *Drosophila,* it has been demonstrated that *DPR* plays a role in the gustatory response to salt (Nakamura et al., 2002). In contrast, *TOLL10* as well as, other *TOLL* genes putatively encodes receptor proteins that play a role in innate immune response signalling and pattern formation during embryogenesis (Luna *et al.*, 2002). Additionally, 3L:17529965 is found within a region with an F_{st} estimate corresponding to the upper 1% of the F_{st} distribution

24. AGAP011544

3L:27038709 is intergenic and 15 kb upstream of gene, AGAP011544. Little is known of the protein that this gene encodes except that it is predicted to bind nucleic acids and metal ions.

25. AGAP011580 (*EIF3K*)

3L:28354730 is intergenic and 17 kb upstream of *EIF3K,* a gene that putatively encodes a component of the Eukaryotic translation initiation factor 3 (EIF-3) complex. The EIF-3 complex plays a role in translation initiation. *EIF3K* is highly

conserved in various organisms including humans, chimpanzees, dogs, cows fruit flies and mosquitoes.

26. AGAP011584

SNPs, 3L:28410032 and 3L:28410034 are completely linked (r^2 = 0.99) and are found in the intronic region of the gene, AGAP011584. AGAP011584 encodes a protein with a predicted function as a serine-type peptidase, catalysing various proteolytic processes through GTPase-mediated signal transduction. This gene is also highly conserved in various organisms including humans, chimpanzees, dogs, cows fruit flies and mosquitoes.

27. AGAP028207

SNPs, 3L:29895649 and 3L:29895651 are both located in an intergenic region with AGAP028207 as the closest gene (~13 kb upstream). Both SNPs are also highly linked (r^2 = 0.85).

28. AGAP011712

3L:32006693 is intergenic, 42.5 kb downstream of AGAP011712 gene which putatively encodes a transcription factor. 3L:32006693 is also found within a region corresponding to the upper 10% of the F_{st} distribution.

29. AGAP011719

SNPs, 3L:32221745 and 3L:32221764 are highly linked (r^2 = 0.95) and are 4.9 kb upstream of AGAP011719; this gene putatively encodes a serine-type peptidase. Both SNPs are also found within a region corresponding to the upper 5% of the F_{st} distribution.

30. AGAP011812

3L:33525031 is 3.9 kb upstream of AGAP011812; this gene encodes a transferase with a predicted role in the elongation of long chain fatty acids.

31. AGAP011925 (*PNPLA5*), AGAP011852

This region contains two unlinked SNPs (r^2 = 0.001), 3L:34158609 and 3L:35174543. 3L:34158609 is intergenic and is 58.6 kb downstream of AGAP011852 while, 3L:35174543 is located in the 3'UTR region of *PNPLA5.* AGAP011852 encodes a protein with a predicted role in redox reactions as a short-chain dehydrogenase/reductase. In contrast, *PNPLA5* ("Patatin-like phospholipase domain containing 5") encodes a protein with a predicted role in lipid breakdown. Both SNPs are also located within regions corresponding to the upper 10% and 5% of the F_{st} distribution, respectively.

32. AGAP012019

This region contains four, highly linked SNPs (3L:36413144, 3L:36413145, 3L:36413160, 3L:36413185; r^2 = 0.97-1). All SNPs are intergenic, with the closest gene, AGAP012019 being 8.3-11.2 kb away (upstream). Although the function of AGAP012019 is unknown, it contains a transmembrane-channel-like (TMC) domain. TMC proteins are predicted modifiers of ion channels and transporters. Mutations in these genes have also been implicated in human conditions such as deafness and epidermodysplasia verruciformis disorder. All SNPs are also located within a region corresponding to the upper 2.5% of the F_{st} distribution.

33. AGAP012031

3L:36725585 is 1.7 kb downstream of AGAP012031; this gene has no known function.

34. AGAP012312

3L:40090717 is intergenic and is 5.6 kb upstream of AGAP012312; this gene has no known function.

35. AGAP001238 (*RASGRF2*)

3L:40476664 is located in a non-coding region with the nearest gene, *RASGRF2* being 71 kb away (downstream). *RASGRF2* ("Ras-specific guanine nucleotide-releasing factor 2"), putatively encodes a protein that plays a role in the regulation of Rho protein signal transduction. Recently, the Ras family of G-proteins have been implicated to play a role in memory formation as well as, in the molecular processes underlying neuronal and behavioural plasticity (Ye and Carew, 2010).

36. AGAP007859

3R:2283582 is intergenic and is 8.6 kb downstream of AGAP007859. This gene contains domains that correspond to the sugar transporter, sugar transporter ERD6-like 6 protein. 3R:2283582 also located within a region corresponding to the upper 10% of the F_{st} distribution.

37. AGAP007989

3R:3831638 is a synonymous variant, found within the AGAP007989 gene. Little is known of the protein that this gene encodes except that it has a predicted nucleic acid binding activity. It also contains domains that correspond to nucleoporin and signalosome proteins.

38. AGAP028019 (*CYP4H18*)

3R:9908789 is 2.9 kb upstream of *CYP4H18*, a gene that encodes a cytochrome P450 protein. *CYP4H18* has a predicted role in oxidation-reduction processes and has shown upregulated expression post blood-feeding. Over-expression of cytochrome P450 proteins have also been implicated in the development of insecticide resistance in many insect species, including *An. gambiae* (Mitchell *et al.*, 2012).

39. AGAP028025

3R:19019495 is intergenic and is ~23.9kb upstream of AGAP028025; this gene has no known function. 3R:19019495 is also found within a region corresponding to the upper 10% of the F_{st} distribution.

40. AGAP009022

3R:23769190 is located within the intronic region of AGAP009022, a gene that putatively encodes a chitinase and plays a role in the breakdown of chitin.

41. AGAP009482 (*ZNF500*)

3R:34584956 is intergentic and is 5.6 kb upstream of the transcription factor, *ZNF500*.

42. AGAP009646 (*CDX*), AGAP009700 (*SWD2*)

3R:37563683 is located within the intronic region of the putative homeobox gene, *CDX. CDX* is a transcription factor, responsible for regulating the transcription of various genes and has been shown to be down-regulated after blood-feeding. In contrast, 3R:39145574 is intergenic with the closest gene being 8kb downstream of *SWD2*. *SWD2* putatively encodes a component of the COMPASS complex

which plays a role in telomere length maintenance and transcription elongation regulation.

43. AGAP009725

3R:41563691 is intergenic and is 12 kb upstream of AGAP009725, a gene that is thought to encode a cadherin protein. Cadherin proteins function as cell adhesion molecules. 3R:41563691 is also found within a region corresponding to the upper 1% of the F_{st} distribution.

44. AGAP009792

This region contains of four, highly linked SNPs: 3R:43376625, 3R:43376646, 3R:43376692, 3R:43376739 (r^2 = 0.90-1) which are 23 kb downstream of AGAP009792, a putative importin subunit alpha-2 protein. Importins play a role in protein transport. AGAP009792 has also been shown to be upregulated after blood-feeding. All SNPs are found within a region that corresponds to the upper 1% of the F_{st} distribution.

45. AGAP009886 (*MED15*)

3R:44747427 is located within the exonic region of *MED15,* a gene that encodes a component of the Mediator complex. The Mediator complex is a coactivator involved in regulating the transcription of RNA polymerase II-dependent genes. 3R:44747427 is also found within a region that correspond to the upper 1% of the F_{st} distribution.

46. AGAP012978

X:2954520 is 2.4 kb upstream of AGAP012978, a gene with no known function.

47. AGAP000214 (*GUCY1B*)

X:3492955 is found within the intronic region of *GUCY1B*, a gene that is thought to encode a guanylate cyclase and plays a role in cGMP biosynthesis processes. This gene is conserved in various organisms including humans, chimpanzees, fruit flies and mosquitoes.

48. AGAP000244

X:4548626 is 2.4 kb upstream of AGAP000244, a gene with no known function. It contains a domain with a phosphopeptide binding motif that is thought to bind phosphorylated protein molecules.

49. AGAP000389

X:7151127 is located within the intron region of AGAP000389, a gene with no known function. However, it is predicted to encode a protein with a protein binding function. X:7151127 is also located within a region that corresponds to the upper 5% of the F_{st} distribution. AGAP000389 is conserved in various organisms including humans, chimpanzees and mosquitoes.

50. AGAP000678 (*WDR68*)

X:12109761 is located within the intron region of *WDR68* ("WD repeat-containing protein 68"), a gene that is thought to play a role in binding other proteins. Proteins with WD40 domains are involved in a variety of processes including signal transduction, pre-mRNA processing and cytoskeletal assemble. *WDR68* is conserved in various organisms such as humans, chimpanzees, fruit flies, mosquitoes and *C. elegans*.

51. AGAP000862

X:15878258 is 3.7 kb downstream of AGAP000862, a gene that putatively encodes an alpha 1,3-glucosidase enzyme for the breakdown of carbohydrate molecules. AGAP000862 is conserved in organisms such as humans, chimpanzees, fruit flies, mosquitoes and frogs. It is also located within a region that corresponds to the upper 2.5% of the F_{st} distribution.

Supplementary figures

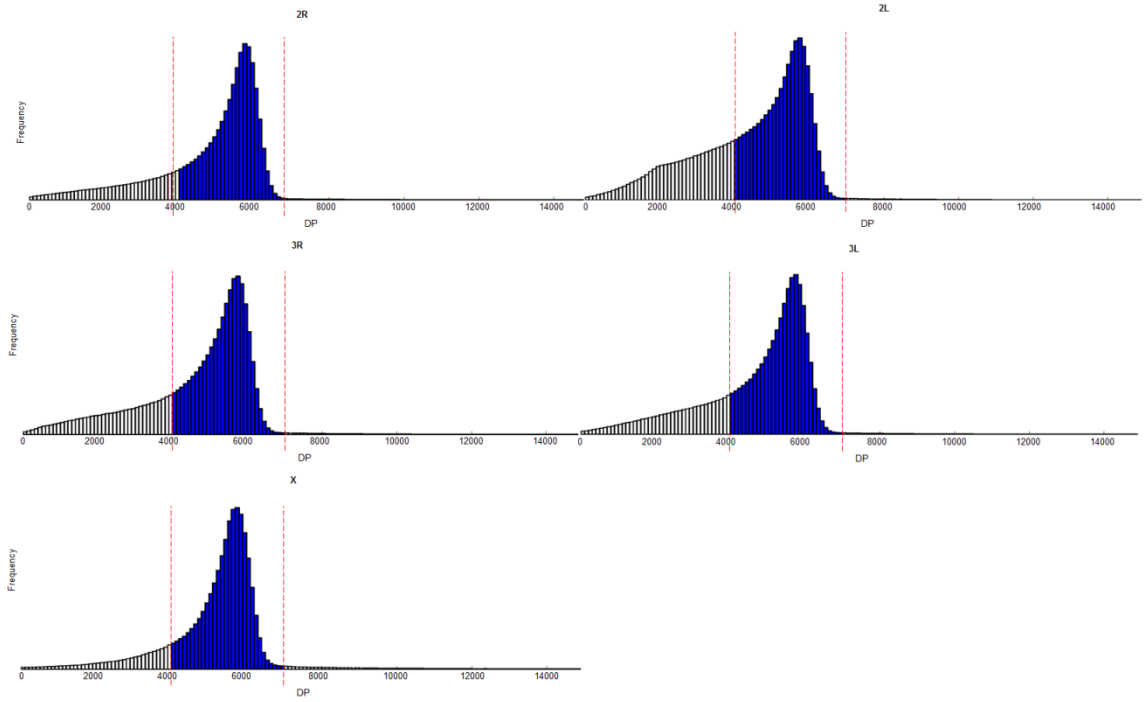

Figure S8. Frequency distribution of genotype coverage (DP) before quality filtering. For each genotype, coverage is calculated as the sum of per-sample coverage. Red dash lines represent the exclusion criteria: DP < 4000 & > 7000. Bars in blue represent the frequency of SNPs that passed this filter.

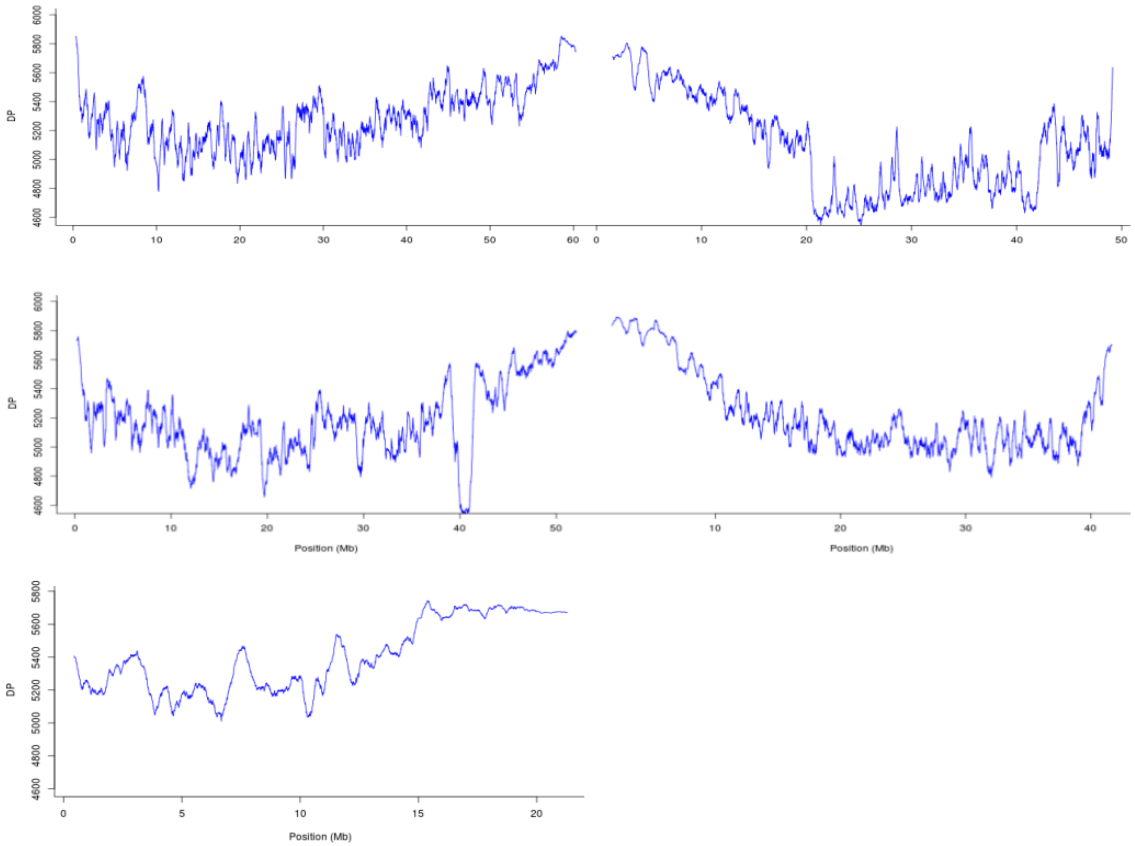

Figure S9. Median genotype coverage (DP) of SNPs across all samples of the Gambian dataset after quality filtering. Median DP was calculated using an overlapping window of 5 kb.

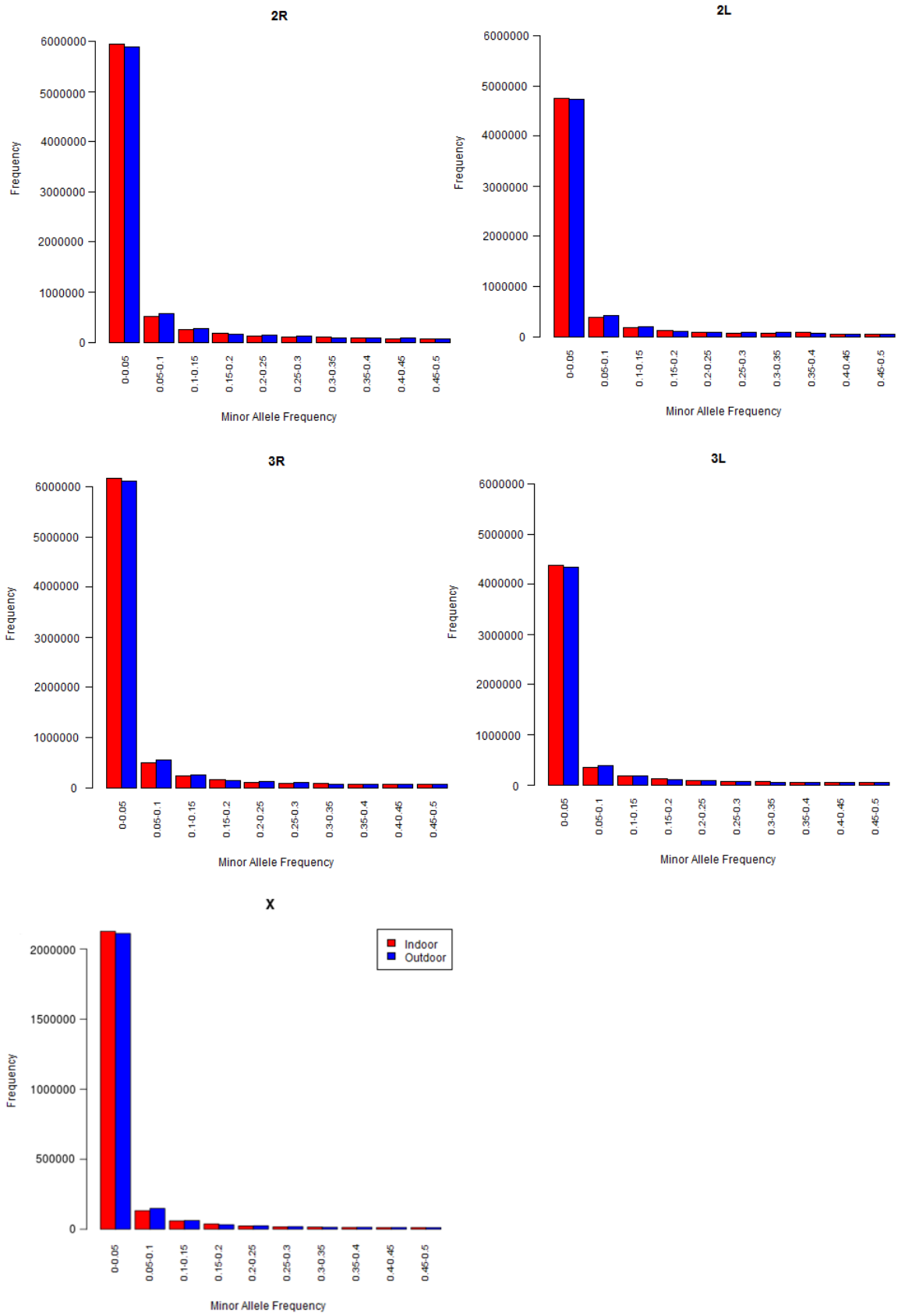

Figure S10. Minor allele frequency (MAF) spectrum of indoor- and outdoor-biting *An. coluzzii* from Wali Kunda, Gambia. The plot represents SNPs that have passed the filtering staged but have not been filtered based on the maf criteria (exclusion of SNPs with maf < 10%). The majority of the SNPs across the entire genome are low frequency (maf = 0.1-0.05%).

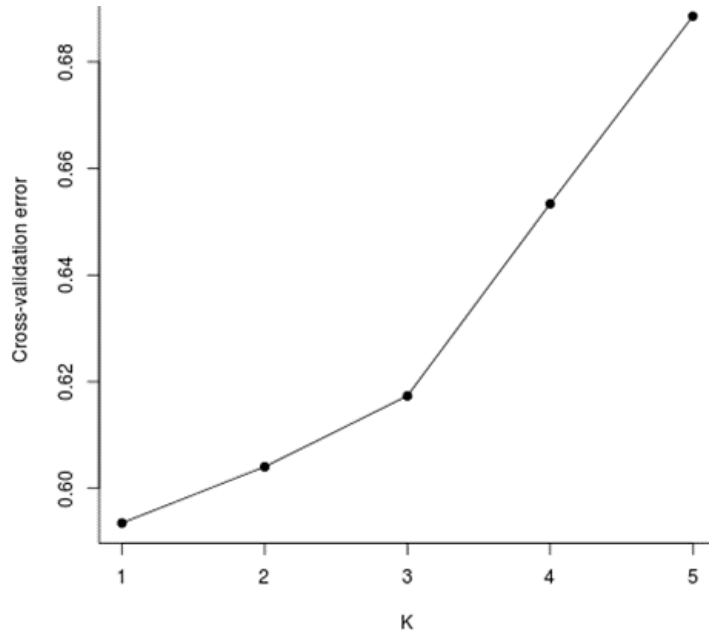

Figure S11. ADMIXTURE cross-validation error as a function of the number of simulated populations (*K*).

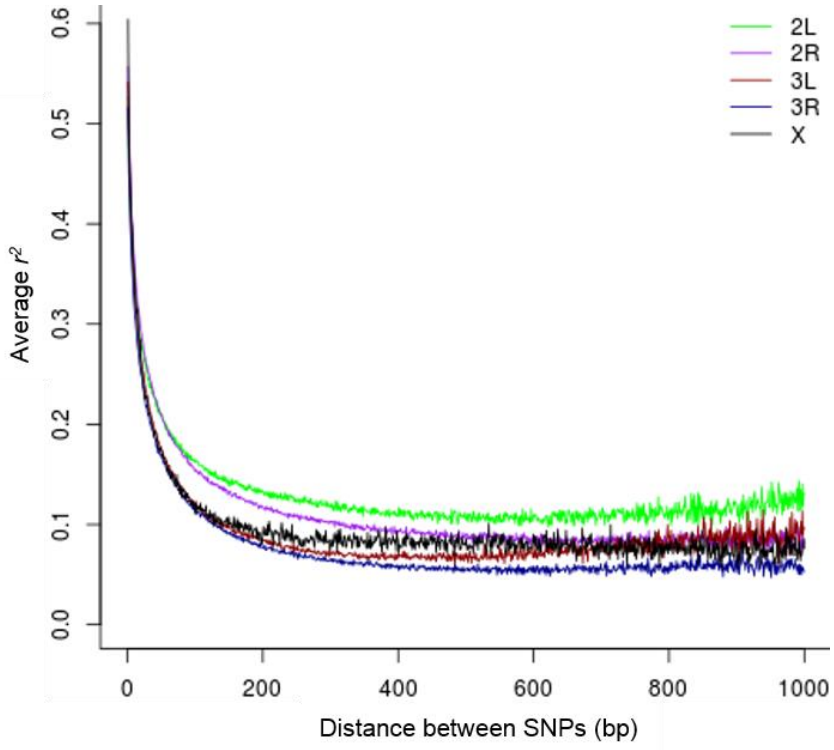

Figure S12. Average linkage decay plot (r^2) between SNPs across the entire genome of the Gambian dataset. Linkage disequilibrium (LD) decays quickly, dropping below r^2 0.2 within 100 bp in all chromosomes.

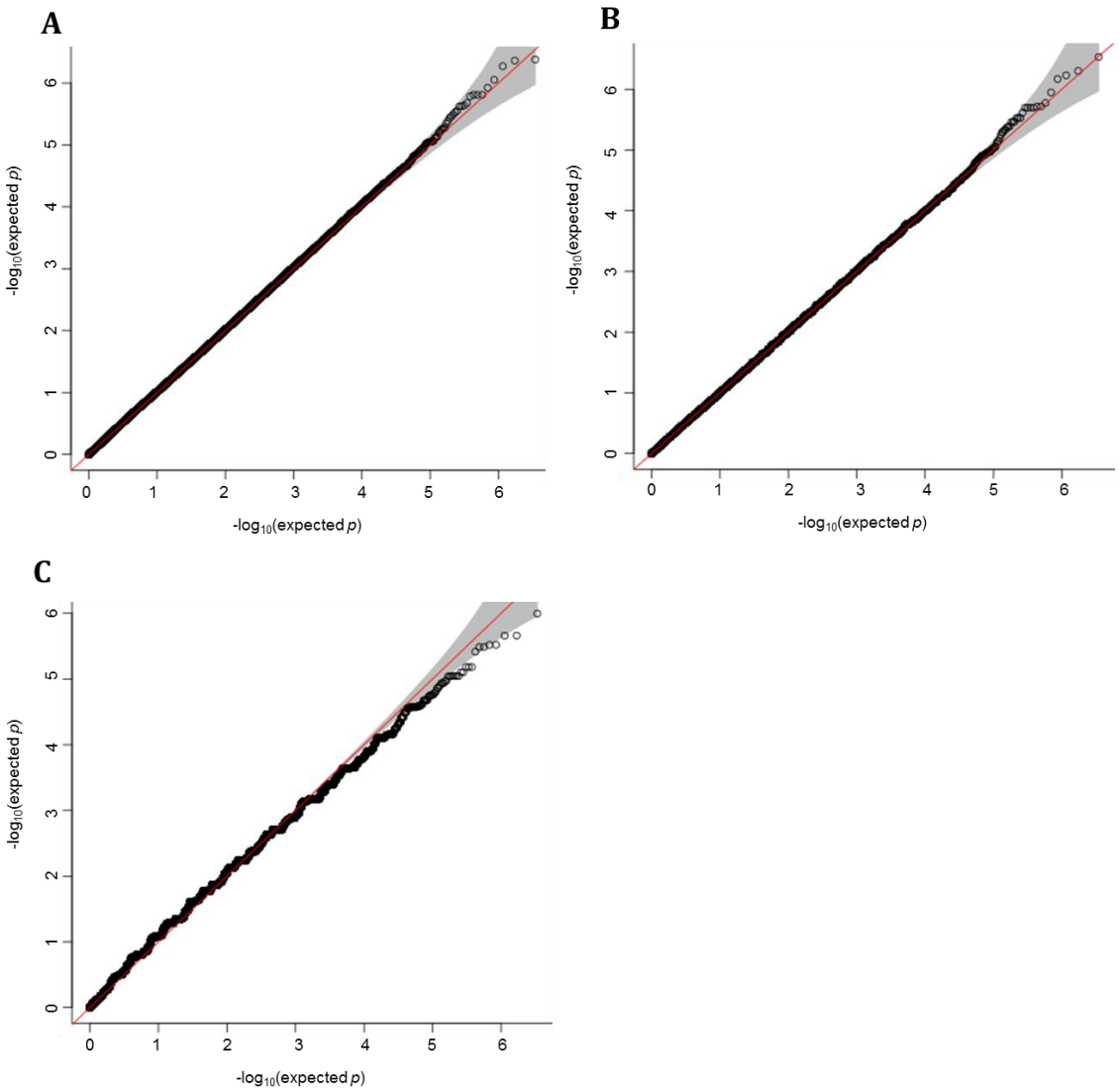

Figure S13. Quantile-quantile plots representing the distribution of logistic regression *p-values* under the additive (A), dominant (B) and recessive models (C). The black circles represent the observed *p*-value against the expected *p-value* under the null hypothesis of no association. The grey, shaded band represents the 95% confidence interval values while the red, diagonal line represents the expected distribution if the observed *p*-value equals the expected values. Genomic inflation factors were (λ) 0.96, 1 and 1.02 for the additive, dominant and recessive models, respectively. Generally, the data conform to the null hypothesis, with the exception of a few SNPs that are deviating from null expectations, thereby representing strong associations.

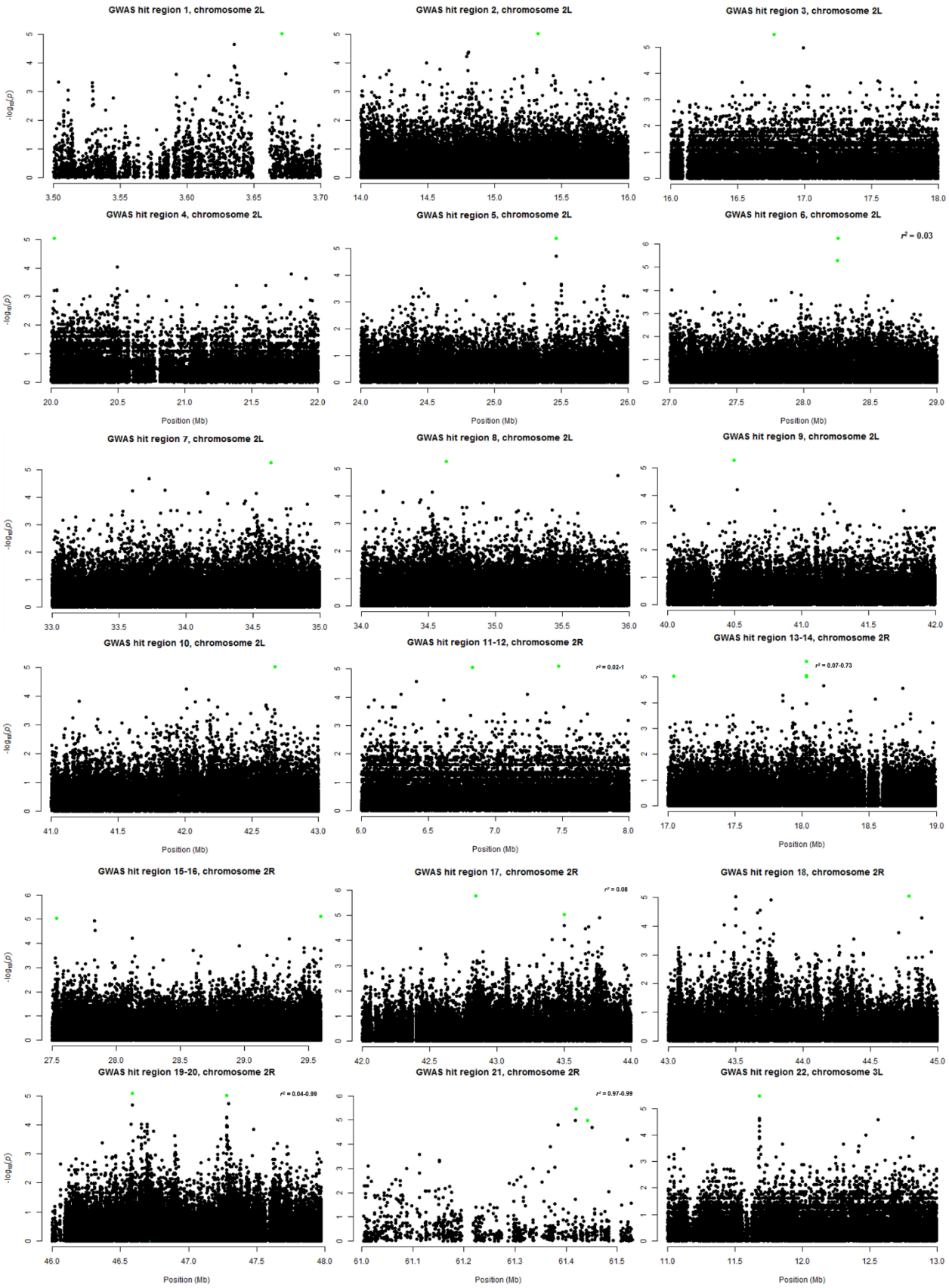

Figure S14 continued in the next page

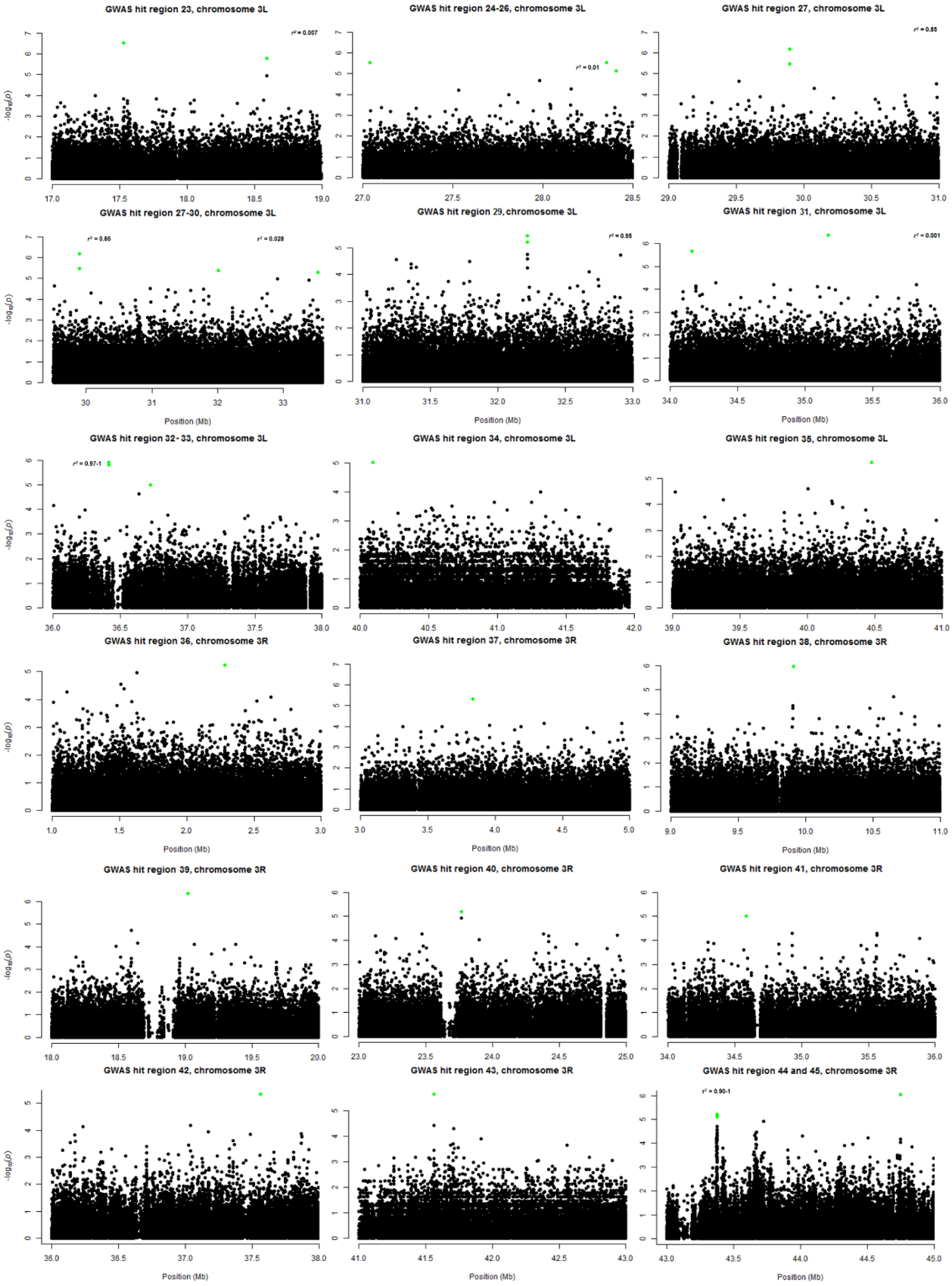

Figure S14 continued in the next page

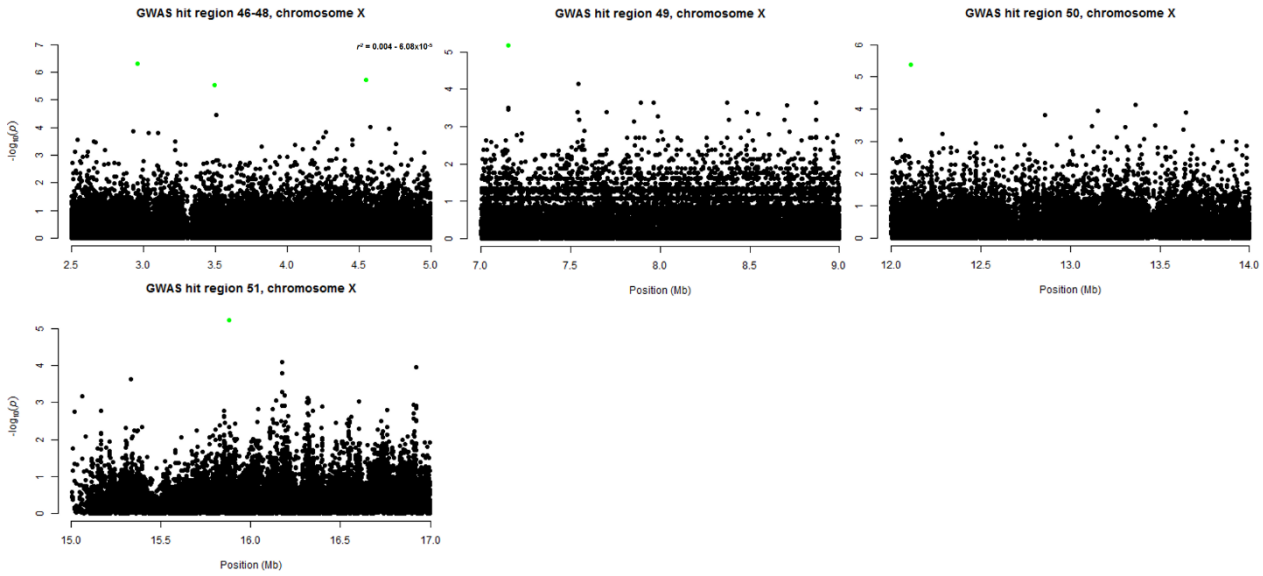

Figure S14. Association plots of the 51 GWAS hit regions. Observed *p* values are the –log10 transformed *p* values for association with outdoor-biting behaviour in the Gambian dataset. Each black, solid circle represents a single SNP while green, solid circles represent SNPs with significant associations ($p \leq 10^{-5}$). The level of LD between significant SNPs, as measured by r^2, is also indicated in the figure.

Supplementary tables

Table S6. Sequenom assay designs for the 177 SNPs that were assayed in the Ugandan dataset.

Multiplex ID	Chr	Coor	Gene	IUPAC base code	1st-PCR Primer	2nd-PCR Primer	Ext. primer direction	Ext. Primer	Ext. Primer Mass
W1	3R	35177638	*OR2*	R	ACGTTGGATGAGCAGTGCGCACAGCATCAT	ACGTTGGATGCGCGTGCACGTACAGATATG	R	TCACTCGTCACCCAT	4447.9
	3L	24982104	*OR5*	S	ACGTTGGATGCGTACTTCGGTGGACTAAAG	ACGTTGGATGTGGATTCTCATCGCCACAAG	R	TTGCCGAACAGTACTT	4856.2
	2R	22858638	*OR7*	R	ACGTTGGATGCGAATGTTCGGCATCAGGTC	ACGTTGGATGAACGACCACTACAAACGAGC	R	GCAGATGCAAGTCCAG	4915.2
	3R	35178695	*OR2*	Y	ACGTTGGATGGTGTGGCTGTTCTGGTCGTA	ACGTTGGATGAGTACAGCTTCAGGAACTGG	F	gCCGGGATGCAGCCGA	4932.2
	2R	12012840	*OR8*	R	ACGTTGGATGAGTGTTTGTGTGTGCGTTGG	ACGTTGGATGCGAAAATCATGTTAGCAAACC	R	AACAGCACAAACGAAAA	5198.4
	2R	22855174	*OR7*	Y	ACGTTGGATGCTCTCCATTTAATCTATC	ACGTTGGATGGTGGTACGGTGGGCCATTC	F	GCATAGTTTCAGGCGTA	5225.4
	2R	22853803	*OR7*	M	ACGTTGGATGAAGCAGGAAATGATGGTGCG	ACGTTGGATGCTGGAGGCTTTACACCACAC	F	TTTACACCACACCTTAGA	5402.5
	2R	22858492	*OR7*	S	ACGTTGGATGTCAGCTCGTTCACGTCGTC	ACGTTGGATGATCCGCAAGGTGTACTCCTG	R	GCTCGCCATGGTGCTGAT	5506.6
	2R	26241662	*OR28*	R	ACGTTGGATGGCAGTACATCTTACTCCTAC	ACGTTGGATGGCGCCTACAGGGTATGCAAT	R	aTGATTGATGGTTTGCAG	5584.6
	2R	22853351	*OR7*	K	ACGTTGGATGCACCCGTATCGATTTTCCAC	ACGTTGGATGACACAAGGACACAAAGGACG	F	tggTTTTGGGGGAGCTTT	5607.6
	3R	37462009	*OR1*	R	ACGTTGGATGGCTTATTTCAAGCTAAATACC	ACGTTGGATGGGAAGTGTCTTGAAGGTTAC	R	CGCTAAATCTTCACACATT	5706.7
	2R	22850829	*OR7*	R	ACGTTGGATGAGTTTTGTAACACGGATTGC	ACGTTGGATGATTGTACTACCAACGCAGTC	F	cCCAACGCAGTCATTAAAC	5725.8
	2R	22849461	*OR7*	K	ACGTTGGATGTTTGTCAGCAGTGCCAGAAG	ACGTTGGATGTTTTGCCAGCCTGTTCTGTG	F	aCGAAACGGTGAAAAACTT	5853.9
	3R	37463416	*OR1*	Y	ACGTTGGATGGAAGATACGGATCAGGCAAC	ACGTTGGATGGTATAGGGCTTGCGTTAGAG	F	GCGTACAGATGTAGAAACA	5869.9
	2R	22853925	*OR7*	S	ACGTTGGATGTTGATCTGAGCGGCATCTAC	ACGTTGGATGTTCGGATTTCCGTTCCTGCC	F	CCGTTCCTGCCATTCTCGTC	5970.9
	2R	22857729	*OR7*	Y	ACGTTGGATGGTACCTGGTAGATGAAAGAG	ACGTTGGATGGACGTACACGGTGGATATAC	R	ccccATGCAATGAGCGGACC	6072
	2R	22855656	*OR7*	Y	ACGTTGGATGGTCGTACGGATGATGGTTTT	ACGTTGGATGGGACTCAAAGAGAGCATAAC	R	cTTTCAATGACCTTAACACTC	6300.1
	3L	24982534	*OR5*	Y	ACGTTGGATGCAAGTGTTACGGTTAGAGGG	ACGTTGGATGTGGAAACCTACGGATTCTGC	F	aTACGGATTCTGCTACTTTGG	6427.2
	2R	22849144	*OR7*	K	ACGTTGGATGTTCGCTCGTTCGGTTGTTTG	ACGTTGGATGCTGTAAAAGAAAACTTGGG	F	GAAAACTTGGGATTATTTGTT	6490.2
	2R	22850011	*OR7*	R	ACGTTGGATGGCAAAACGAGAGAAAGAGAG	ACGTTGGATGTGGGATTGGTTTTTGCAGCG	R	cACATTCGCACAAAACCGTCCT	6608.3
	3R	35177912	*OR2*	R	ACGTTGGATGTGTGATACTTTAGACACTCC	ACGTTGGATGGATAGCGGCCCTAAAGCAAC	R	cCAGCGCCGGCACACTGTTCGC	6657.3
	2R	22858887	*OR7*	W	ACGTTGGATGCGCAAATCAGGTCACAAAGC	ACGTTGGATGAAATGCTTGGAGCCAAACGG	F	gggagACGGTAAAATCTCTCGT	6799.4
	3R	35177836	*OR2*	R	ACGTTGGATGTAGACGCTAGTAGACTCGAC	ACGTTGGATGAATGGCTTCCGGTTTGTGTG	F	ccccaCTCCCCTCCCCGAGGGAA	6875.5
	3R	35178957	*OR2*	R	ACGTTGGATGTGTGGAAGAAGGGTGTGTTG	ACGTTGGATGGAAGCATATAATTTGCTCA	R	CATATAATTTGCTCAAAAAGTCA	7014.6
	2R	22856056	*OR7*	R	ACGTTGGATGGATGTCATGTTCTGCTCCTG	ACGTTGGATGGCGGATGTAAAGACAACCAC	F	gagtcAAGACAACCACGTTTGCT	7032.6
	3R	37461580	*OR1*	R	ACGTTGGATGAAAAGCGCCCTTGCCATAAC	ACGTTGGATGGATAGTTTATAGCACCGTCG	R	tCGCACCTGATTTTATCTTTATCT	7234.7
	3R	37462441	*OR1*	Y	ACGTTGGATGACAGCTTCGGCGATGAAGTG	ACGTTGGATGAGCTTGCAGCAGTGTCATAC	F	TGTCATACAAATGATAATTACAGA	7367.8
	2R	12012956	*OR8*	M	ACGTTGGATGGCTAACATGATTTTCGGTTTG	ACGTTGGATGACCGTTACTTACATCGCATC	R	gTACATCGCATCTAGTAGTCATCGT	7632
	3L	35937116	*OR61*	R	ACGTTGGATGAGGTAGATCATACCGTTCGC	ACGTTGGATGGCCCATTTTCCTTCGTACAG	F	TCCTTCGTACAGTACTATGGTACCTT	7887.1
	3R	37463132	*OR1*	Y	ACGTTGGATGGCAAGCTTAACTGCACACTG	ACGTTGGATGGGAGAGTCGTTTGTTAGCAC	F	GTTAGCACTCTTTTTAGTCATATTTC	7892.1
	3R	35178780	*OR2*	R	ACGTTGGATGTACGACCAGAACAGCCACAC	ACGTTGGATGGAACATAAACCGCAACCCAC	R	ACCGCAACCCACAGCCGAAAATGCTGAT	8520.6
W2	2R	22855496	*OR7*	M	ACGTTGGATGCACTACATCCGTTCGTCCAG	ACGTTGGATGACCGGTATTTCAGGATGAGG	R	ACAATCCCCACCCAG	4450.9
	3L	24983049	*OR5*	R	ACGTTGGATGACCCTTTGACTTTTTAAG	ACGTTGGATGTATGATTAGTCGAGGAAACG	R	GCTCGACAACGGTCT	4553
	3L	40214764	*OBP26*	S	ACGTTGGATGTACCACTTTTTGCTTGCCCC	ACGTTGGATGAGCGAAGTTGATGCACTTCC	F	AGTGCATAGTGCACG	4617
	2R	22849536	*OR7*	R	ACGTTGGATGTGTTTTCACTCTCCCATACC	ACGTTGGATGACGAAGGTTTTTGCCTCCTC	F	cTGCCTCCTCGGACCC	4754.1
	2R	26239795	*OR28*	R	ACGTTGGATGAAGACAAACGGGTAACGGTG	ACGTTGGATGACTACCCTGTACTGGCTTTC	R	CCGGCCGGCATCTTACC	5107.3
	3R	37462478	*OR1*	Y	ACGTTGGATGAGCTTGCAGCAGTGTCATAC	ACGTTGGATGTTTCACAGCTTCGGCGATG	R	CGGCGATGAAGTGCAGG	5300.5
	3R	37462048	*OR1*	S	ACGTTGGATGGGACTCTTAAAGATGTTCAC	ACGTTGGATGGTGTGAAGATTTAGCGTAACC	F	AACCTTCAAGACACTTCC	5387.5
	2R	22855261	*OR7*	S	ACGTTGGATGTTGTAGGGACGGATAAGCTC	ACGTTGGATGAGCAATTAGAGCAACACGGC	F	ccGCAACACGGCACTAAG	5462.6
	3L	24981519	*OR5*	W	ACGTTGGATGAACGACGCTATCGCTACAAG	ACGTTGGATGTCTCTAAATCTGGATAGCCG	R	AGAAGCTTAAAAATGCCA	5524.6
	2R	22855799	*OR7*	Y	ACGTTGGATGATCAGCTCCGATTTGGAACC	ACGTTGGATGCGGGCAGGGTATTATGCGAT	R	aagcCCGGCCCAACTCTTC	5693.7
	2R	22850930	*OR7*	Y	ACGTTGGATGCAGTTTTTTTTTGTAAGCGG	ACGTTGGATGTGCAATCCGTGTTACAAAAC	F	TCTATTTGAGCGTTAACAG	5817.8
	3R	35178103	*OR2*	S	ACGTTGGATGTGAACGGGATGTACATGCAG	ACGTTGGATGACCGTACGGCGTCACGATAC	R	cccccACCCCGACCTACCAG	5912.8
	3L	24982615	*OR5*	M	ACGTTGGATGAACGGCGATACCAGAGGCTA	ACGTTGGATGTTCCCCTCTAACCGTAACA	F	CTCTAACCGTAACACTTGTA	6036
	2R	22856666	OR7	M	ACGTTGGATGCACCGCCATAACAATTGGG	ACGTTGGATGCGTGCGAAATAATTTATTCC	R	CATCCCCATTAAAACAAAGT	6038

	2R	12012773	*OR8*	Y	ACGTTGGATGATCCCACGGCTGGATATTG	ACGTTGGATGCGCACACACAAACACTAGAG	R	aACAAACACTAGAGAACACG	6121
	2L	45011466	*OBP47*	W	ACGTTGGATGAACGGCGTGACGCAACATTC	ACGTTGGATGCAATCCCTGCCTTAAAGGAC	F	TTTCCACTACTATTTCCACTC	6242.1
	2R	22851925	*OR7*	R	ACGTTGGATGTGTAAAATGAGCCACGCGAC	ACGTTGGATGACGGCTTACGGTTTCGGGC	F	AACAAGAATAAAACCACCTCC	6345.2
	2R	22858426	*OR7*	R	ACGTTGGATGAGAAGTTCTCCGAGTTGACC	ACGTTGGATGACGACGTGAACGAGCTGAC	R	agttAACGAGCTGACCGCCAA	6424.2
	2R	22854881	*OR7*	R	ACGTTGGATGGACCTGGAGCGGTTATTTAA	ACGTTGGATGTCTCATGTTTTGCCTGACGG	R	gGAAAAAAGAGCGATTTCTTC	6478.2
	3L	2853263	*OBP22*	W	ACGTTGGATGGTATGGAGTCGATTTGCAAG	ACGTTGGATGCTTAGCTTCTAGACAGGCAC	F	cCATCATCACTGTTCCCCAGAT	6590.3
	2R	22854102	*OR7*	R	ACGTTGGATGCATTTTGAGGTTTACACAT	ACGTTGGATGGATCCTTTTCTGTGGTAGCG	F	cttacTGTAACAATCAGGCGTC	6694.4
	2R	26241481	*OR28*	M	ACGTTGGATGCAACAAATTACCGAGGAGGC	ACGTTGGATGGTAGGCTTCTGGGCTTTAAC	R	aaTGGGCTTTAACCATGAATGT	6773.4
	2R	22856191	*OR7*	Y	ACGTTGGATGACGGGATAATCCGTTGGGAC	ACGTTGGATGCCATCGAAAACAGCAGGAAG	F	TGAAGCATAAAATGTGTTGTAG	6837.5
	2R	12014631	*OR8*	K	ACGTTGGATGTTGGATGAGAGGGTGTGTGG	ACGTTGGATGCAACAAAAAGCGGCACGTTT	F	tcaCGCCAATTATTGATTGATTT	7003.6
	2R	12013957	*OR8*	K	ACGTTGGATGACGGTGTGTACCAGTACAAC	ACGTTGGATGGCAGCAGGTTAAACATCCAG	F	ggaggTCGAGAACGATTTCATCAC	7401.8
	2R	22856451	*OR7*	Y	ACGTTGGATGGCTGGTTTAGCTTTAGCTC	ACGTTGGATGCGCTTCGGGATGTCATAAAG	R	ccttGGTGTTTTTAACATTACTTCC	7563.9
	3R	20492001	*OR65*	W	ACGTTGGATGTTTGCACGGTACCACCAGAG	ACGTTGGATGCGTCGTTGCGTATGATTTTG	F	ttCGTATGATTTTGTTCATTTTTCG	7625
	3R	20493567	*OR65*	M	ACGTTGGATGTACAACGAACTACTGCAATG	ACGTTGGATGAAATGGGGAGAATGCAACGG	R	TTCTACTTATCTATCTATCTACCCAGT	8110.3
W3	3R	37461899	*OR1*	K	ACGTTGGATGAACACCATTACTCTGATTCC	ACGTTGGATGTTATGAAGCTATCGTACTCC	R	cCTATCGTACTCCTATCTGGC	6308.1
	3R	37461993	*OR1*	R	ACGTTGGATGACATCAAGGTGGGAAGTGTC	ACGTTGGATGGCTTATTTCAAGCTAAATACC	F	ggATTTCAAGCTAAATACCACAGCAT	7947.2
	3R	37462107	*OR1*	W	ACGTTGGATGCTGCAAATGTGAGATAGCGG	ACGTTGGATGCTTCCCACCTTGATGTGAAC	F	AGAACGTATTTAATGTACTGAC	6757.4
	3R	37462227	*OR1*	R	ACGTTGGATGGGACACGTAGAGGAATTAAA	ACGTTGGATGGGAAAACCCAACAAACTCTG	F	CTCTGTAAATTTATCCGTCTGTAAC	7581.9
	3R	37462530	*OR1*	R	ACGTTGGATGGTGTGTGCTGTTTCACAAGG	ACGTTGGATGCGAAGCTGTGAAAAGTTGGG	F	AGTTGGGTCAAGCTTTTTAAATG	7108.6
	3R	37462690	*OR1*	W	ACGTTGGATGTTTCCATTCCGCATCCGTTG	ACGTTGGATGGGGAGTTTGTTTCCGTTATG	R	TTGGACATGACGTCCC	4857.2
	3R	37463567	*OR1*	K	ACGTTGGATGTCAGTGCCACTGTTTAACCG	ACGTTGGATGACTGCAACCAGAACGAATCC	F	gacCCGTGGGTTCAGTTT	5521.6
	3R	43668933	*Gr11*	M	ACGTTGGATGCGTCTTTCGTACCTAAACAC	ACGTTGGATGGGCACTCATCTTAATGAACC	F	tacTTAGGTTCGGGTCCG	5521.6
	3R	43669141	*Gr12*	W	ACGTTGGATGGTCTGCTAGATCACTTGAAC	ACGTTGGATGCTGCGGAATGTTCGATATAG	R	gTTCATATGGTAAGCTGCTA	6147
	3R	43671604	*Gr10*	M	ACGTTGGATGAATCGTCAACCTCGCCAATC	ACGTTGGATGACGCAAGTGAAGAGTTGGAG	R	TTTGGACTACTTCTTGCATC	6048.9
	3L	35174543	*PNPLA5*	Y	ACGTTGGATGAGACGAGAGCCAGCAGTAG	ACGTTGGATGAATGGCGCAGGCAAACGAG	R	GGCAAACGAGAATCGG	4964.3
	3L	35184869	*PNPLA5*	R	ACGTTGGATGGCCCAAAATCGCATTTCTTC	ACGTTGGATGTAAGCGGGTTTTGTGCGTTG	R	ACCCACAACCCGTGA	4490.9
	3L	35174648	*PNPLA5*	Y	ACGTTGGATGAAAAGTCTGGCTCTGTACCG	ACGTTGGATGACCCTGCTGCTGAGCAAAC	R	cccTCCTACTCGATCGAGATGTA	6959.5
	2L	22305714	-	K	ACGTTGGATGAGAACCATCAAAAAAGGCTG	ACGTTGGATGGATCCATCAGTGGCTCTAGG	R	gAGTTGCGCTAGAAGAG	5299.5
	2L	34623415	AGAP006598	R	ACGTTGGATGTGTGTGGCAGTTGGTTTGTG	ACGTTGGATGGCTCACCAGGAGTGTAATTC	F	AGACATGCCTGCCCC	4497.9
	2L	40496599	AGAP006991	R	ACGTTGGATGTTACCAAATGGCCCGCTTTC	ACGTTGGATGAGACACGTACGTACTAACGG	F	ctTATCGTTCGCGTTGTATCG	6394.2
	3R	37563683	*CDX*	Y	ACGTTGGATGGCGAAATTGGCTGTGGAAAG	ACGTTGGATGACTTGCGCTACACACCTTTG	F	tgGTTGTTTTTGCTGTTGTAC	6455.2
	3R	37557083	*CDX*	Y	ACGTTGGATGCACTTTTTTTAAACCCGCGA	ACGTTGGATGAGCCATCAAACATGCAAAGG	R	GGAATCGATTGTTGAGAG	5618.7
	3R	43724416	AGAP009809	R	ACGTTGGATGAGGTTCTGGAGTAATTTTGG	ACGTTGGATGGTTCAAGGCTTCTAATTTTC	F	TACAATCAAACACGCTACT	5724.8
	3R	43728733	AGAP009809	Y	ACGTTGGATGATGGGATAAGTAGCCCAACG	ACGTTGGATGCGTTGGTGCGATGTGTACAG	F	ACAGTGCAAAAACATTCG	5500.6
	3R	43727408	AGAP009809	H	ACGTTGGATGTACGATGGAGTCACGGTTAG	ACGTTGGATGGTGCAAGTGTGCGTAACATC	R	TGCGTAACATCTTCTCT	5111.3
	3R	43723827	AGAP009809	R	ACGTTGGATGTTGGGTTAGAATTGCCTTGG	ACGTTGGATGGTGTGATTGCACCATCGTTC	F	gggGTGCGAAGGTACTCAG	5933.9
	3R	44747427	*MED15*	W	ACGTTGGATGAAAACACCGGCTAGCGTATC	ACGTTGGATGACGTGAGATGAGTAAGTGTC	R	cccTGATTGAACACACACACAC	6617.3
	3R	44747202	*MED15*	R	ACGTTGGATGAGGCAACAGCAGCCAGCAG	ACGTTGGATGGGTTCTGCAACGCATTGATG	F	gttgtGATCGGGCATACCGCT	6453.2
	3R	44745000	*MED15*	R	ACGTTGGATGTTAATGTCCGGCCCGAACAG	ACGTTGGATGGTCTCGAGCAGTCTGCAGG	R	ttCCACGGGCAACCACAC	5413.5
	3R	44744808	*MED15*	W	ACGTTGGATGAGAATGTGCGGTATCTCCTG	ACGTTGGATGACCACGACAGCGGCCGCTA	R	CGGCCGCTACGACTGG	4883.2
	X	3492955	AGAP000214	W	ACGTTGGATGGGTCGCTAAATGTTAGCGAT	ACGTTGGATGTTAGAGCATTGGACAGGCAG	F	CATTGGACAGGCAGCAACCCAAAA	7348.8
	X	7151127	AGAP000389	S	ACGTTGGATGCATCGTAACTGTGTTCAATC	ACGTTGGATGTCCAGCATTGTAGGATTCCC	R	agTTAAACGAGTCTTTCAATCATC	7310.8
	X	7139849	AGAP000389	R	ACGTTGGATGTCTGTGTCCACTAGCAACTC	ACGTTGGATGAACCAATGAAAGTCCCATCG	R	atAAAGTCCCATCGAGAATT	6109
	X	7135098	AGAP000389	M	ACGTTGGATGGGTAGCTTGATGATCAATGC	ACGTTGGATGAGCTTACGTACGTTGGGTTC	R	cccTCATGACGCATTTTCATT	6307.1
	2L	3637030	AGAP004797	R	ACGTTGGATGCTGTCTCCTTTTCTGTTGGC	ACGTTGGATGGGAGCAAACGTGAATAAAAC	R	cccACTACTCGTTCCTCCAT	5923.9
	2L	49227722	AGAP007717	W	ACGTTGGATGGTTTATGTACACCACAAACA	ACGTTGGATGAACTATACACGCATGGTTTC	R	CACGCATGGTTTCAAATTAATTATAAAT	8569.6
	2R	55229842	*Beat lib*	R	ACGTTGGATGTTTCAGACGTACTTCACCCC	ACGTTGGATGCCACGCTTTGTGAATATCCC	F	cccccTTAATAGACGCCCCACTA	6873.5
	2R	55261800	*Beat lib*	M	ACGTTGGATGCATTGTCGTTCTGTTCGAAG	ACGTTGGATGGAACAAAGACTCTTCCGCTG	F	AATTTCTAGAATAAATATTGCTCAATG	8280.4
	3L	26103443	-	Y	ACGTTGGATGATGTTTCCTTGTGGTCACCC	ACGTTGGATGTGTTTAAGGGTGGGAAACGC	R	agggtAAACGCGTTGGACAAA	6528.3
	3L	32221745	-	M	ACGTTGGATGACGGAATCTGATTGCAGAGG	ACGTTGGATGGTTCGATTGCGATCGAATGG	F	cctCGGGTATGATTTTATCACAAAG	7656
	3L	33525031	-	R	ACGTTGGATGCGTACTGCGCAGTGTTTTG	ACGTTGGATGTAAGGGTAAGCTGCAACCAC	F	ggaccGGTTGATCATTAATCGGTG	7423.8
	2R	18031010	-	W	ACGTTGGATGCACTGATTGGCCCAAGAGTT	ACGTTGGATGCCGCTGCGTAAACCCGTCA	F	cccctCGTAAACCCGTCAACTGTT	7208.7
	2L	37280520	-	R	ACGTTGGATGTTGGGTGCGTGTCTTTTCTG	ACGTTGGATGACTAGCTGCACTCAATCCTG	F	ggggaACTCAAGCGTTAAAGGTTT	7456.9
W4	3R	37462118	*OR1*	W	ACGTTGGATGGAGTCCTGAAATAGAGAACG	ACGTTGGATGAGCGTACCAGCCAAGCAATG	R	GCAAATGTGAGATAGCGG	5612.7
	3R	43667573	*Gr12*	Y	ACGTTGGATGAAGCGTTGCTGGTTTGTCTG	ACGTTGGATGGAATACGCAGCGAGTTGAAG	R	TGAAGCGAGGAAGTGG	5035.3

	3R	43667673	*Gr12*	R	ACGTTGGATGACCGCCACATTAACGTCTTG	ACGTTGGATGTGGCGACGGACCATATTAAC	R	AACAATTCATTCGCTCTC	5393.5
	3R	43667990	*Gr12*	R	ACGTTGGATGTCGTTCACCAAATTCACGGC	ACGTTGGATGTTTGGCGCATCGGTATACAC	R	GCCCGACAGTGGTGC	4594
	3R	43668381	*Gr12*	Y	ACGTTGGATGTCCTGCAAATGTTGCACACG	ACGTTGGATGTCGCACACGCTTGGTAAATG	F	GCATATACACGATGGTTAC	5811.8
	3R	43668947	*Gr11*	R	ACGTTGGATGGGCACTCATCTTAATGAACC	ACGTTGGATGCGTCTTTCGTACCTAAACAC	R	CCTAAACACAAACCTTAACA	6007
	3R	43669258	*Gr11*	Y	ACGTTGGATGAAACCGGTGTATCTGCGTTG	ACGTTGGATGAACAAACGGCTCAACTCGTG	R	tcCGATCGAAATCGGAACTC	6086
	3R	43671450	*Gr10*	R	ACGTTGGATGTACACCGTTTGCCACACTTG	ACGTTGGATGTAGAAACGGGCAGGATGATG	R	GATGCCGGTTGCTGA	4624
	2L	28257115	AGAP006208	R	ACGTTGGATGTGCTCAGGATATGCAGGTTG	ACGTTGGATGTCGAATCGGAGCTGGTAGAG	F	GAGGAGAACGTACAGTC	5268.4
	2L	28257381	AGAP006208	R	ACGTTGGATGACGAGAACCGTTCGATACTG	ACGTTGGATGATGCACACCCGCATCGACCA	R	gggtGCGCGTCTGATTGATG	6220
	2L	22334428	-	Y	ACGTTGGATGCAGCACAAACTAGCGATGAG	ACGTTGGATGCTAAATCAACAACACGCCCC	F	GACAAGCTATTGTAGGCTG	5867.8
	3R	43728754	AGAP009809	Y	ACGTTGGATGATGGGATAAGTAGCCCAACG	ACGTTGGATGCGTTGGTGCGATGTGTACAG	F	ACAAAACGAACAGTGCC	5181.4
	3R	44747407	*MED15*	Y	ACGTTGGATGACGTGAGATGAGTAAGTGTC	ACGTTGGATGAAAACACCGGCTAGCGTATC	F	CGTATCCTTCCGGTC	4494.9
	3R	44747655	*MED15*	R	ACGTTGGATGAAGATGTACGTACGATGCGG	ACGTTGGATGAAAGGAGAGGATGGAGAGAG	F	AAAACGTTCACGCGAAA	5196.4
	3R	44745300	*MED15*	R	ACGTTGGATGGATACTTCTCCCGGTACAGC	ACGTTGGATGTCGTTAAACACGCCCGGCCA	R	GTGGCGCCGTGCCCAG	4899.2
	X	7150927	AGAP000389	K	ACGTTGGATGCACACGAAACCACTGTACTG	ACGTTGGATGCACCATCAGATTAGGGTGGC	F	cctcGTAGCTCCCATTAGTAGG	6686.3
	X	12110046	*WDR68*	W	ACGTTGGATGGAGTTCAGTATCTTTATCCG	ACGTTGGATGGCTCAGGCAAATGAGTTTCG	F	ccTTCTTTGCTAGAAAATGCAAGAAA	7962.2
	2L	28255280	AGAP006207	W	ACGTTGGATGAAATACAGCACGATCGAGGC	ACGTTGGATGAGAAGTGCTCGAAATCGACC	R	cccGTACTCCTGATCGTACTT	6308.1
	2L	28256236	AGAP006207	R	ACGTTGGATGATAATTTCCGTGGCAGGCAG	ACGTTGGATGATGATTATGTGGCCGGTGTG	F	ggaCGGTGTGATCGGCGCGCGGTA	7490.8
	2L	3864566	AGAP004819	M	ACGTTGGATGGCCTCCATCATAAATCACCG	ACGTTGGATGCGTCTTGCAAGGGAAAATGG	R	gtcgCGAGCAAGAGATGCATATTTT	7721
	2L	3619546	AGAP004795	R	ACGTTGGATGATTTCCCTAGTGCTGGAACG	ACGTTGGATGCATGACAATTATTTGCATC	R	TGCATCTATAATATTTGATCAATAATG	8271.4
	2L	3620033	AGAP004795	R	ACGTTGGATGAAGGAACAGGATCACGCAGG	ACGTTGGATGCAACCGCCAACTCTTGTTAC	R	TGTGAATAAACTTAGGTTTATGG	7132.7
	2L	3625276	AGAP004795	R	ACGTTGGATGAGGATGATTTTGGAGATGTC	ACGTTGGATGAAACACTCTCTCCCATAGCG	F	tcGGCTTTTCAATTCGCAA	5753.8
	2L	3637015	AGAP004797	Y	ACGTTGGATGATAAAACTACTCGTTCCTCC	ACGTTGGATGGCCTGTCTCCTTTTCTGTTG	F	cTTTCTCACTGCTGCC	4759.1
	2L	49228419	AGAP007717	K	ACGTTGGATGGATACACAAAGTTCAAGGAC	ACGTTGGATGTAACACTAAATCAAATTAGG	F	TCAAATTAGGATATCTTTAGGAATA	7703.1
	2R	278043	AGAP001109	W	ACGTTGGATGACCATTACGTACGCATGTCC	ACGTTGGATGCGACGCCAATTGACCATAAC	R	gggtGGGTACGGTTAGTATCTG	6877.5
	2R	283739	AGAP001109	R	ACGTTGGATGGACTTTCACTGCAATTTTAC	ACGTTGGATGCGATTGAAGGCACGTGTAAG	F	cctcAACAATTAAGAGGTCAAAGAC	7652
	3R	43664830	AGAP009801	Y	ACGTTGGATGTCTAACCAAACGGACGACAG	ACGTTGGATGTACAGCTCGATCCCTTTTGC	F	aaaCTTTTGCTATGCATAGGC	6420.2
	3R	43665034	AGAP009801	W	ACGTTGGATGAGACTTTGTGGCAAAACGGG	ACGTTGGATGTTTGCTGTTGCTGTACCCTG	R	aGGGTGATGTACCCAATTGAGG	6839.5
	3R	45892033	*Beat*	M	ACGTTGGATGTTCATAGAGACTTTGAGCTG	ACGTTGGATGTGCGTGCATTTTTGTTTTGG	R	gagcTGTCACAAAGGCGAAC	6160
	3R	45893405	*Beat*	R	ACGTTGGATGTGTTATGCTTCTTTGAGCCC	ACGTTGGATGTCGATACCACAGAGGCAAAG	R	AGCAGAAAGGAAACATTTTATAC	7088.7
	3R	45900850	*Beat*	Y	ACGTTGGATGCTTGAACAACATTTGCCCCG	ACGTTGGATGGTATTCACCGTGTGCGGTAG	R	cgTTTGTGTTCTTGTTCTATACTGC	7610.9
	3L	10792362	*HPX14*	M	ACGTTGGATGAAACAACTATCCCGCCGAAG	ACGTTGGATGCAGACAAGTTGGTGTAGACG	F	atTTGGTGTAGACGATTTTGG	6522.2
	3L	10792796	*HPX14*	K	ACGTTGGATGCAGCTGTAATCGAAAAACGG	ACGTTGGATGGTTACTCGTGACATCGACAG	R	caCAGCGCAACGAAAGTATTACA	7034.6
	3L	5842042	AGAP010540	Y	ACGTTGGATGCCTACAAAAATGTGGTGATAG	ACGTTGGATGGGTTATCCTCCGCAAACTTG	F	ataCGCAAACTTGTGAATTTC	6404.2
	3L	5861278	AGAP010540	M	ACGTTGGATGAACGCAGATTGTGTAGCGTG	ACGTTGGATGGCAAACACTTTCTCGTACCG	R	cACCGTGTCTACTCTGAG	5450.6
	2R	40476664	-	R	ACGTTGGATGAGCGGATGCTGGAACGGC	ACGTTGGATGCAAAAAGTCCTGATTATCGC	R	cctctTGATTATCGCTTCCCAG	6612.3
	2R	61418659	-	Y	ACGTTGGATGGTTTATAACCGAAGTTGAAG	ACGTTGGATGCGTGTAAACGGCGTTTGTAG	F	TTAAATAAAAACAACATGATATCATGG	8307.5
	2R	47284139	-	M	ACGTTGGATGACAAAGTCCGCACACGATTC	ACGTTGGATGGAAGATCTCCCATAGAACGC	F	actaCCCCCTCCCCCTGCT	5580.6
	3R	9908789	-	K	ACGTTGGATGCGTAAGCTGTAACGCTTGAG	ACGTTGGATGTATCACAACTGGAAGCGCTC	F	CTGGAAGCGCTCATTAATATTTAAAGGAA	8948.9
W5	2L	25464003	AGAP006031	R	ACGTTGGATGTTGCCGATTGTTTGGCCAAG	ACGTTGGATGTCAACACAGTGGCGCAATAG	R	TGCTCCCCTCCCAGG	4464.9
	X	9227517	AGAP000519	Y	ACGTTGGATGCCGTTTGTTTCTCTGAGGG	ACGTTGGATGAGGCTGGAGCAAACAGTCAC	F	CACCCGGAAAGGGCG	4612
	3R	43664774	AGAP009801	S	ACGTTGGATGTTTAGCTCCTGAAGCTGCTC	ACGTTGGATGTCGAGCTGTATGCATACACC	R	tCGCGCTTCCCCACTG	4769.1
	3L	29895651	-	Y	ACGTTGGATGAATGTACGATGTCCCGAGTG	ACGTTGGATGTCCGTCCATTTTCCAGCAAC	R	aTCCAGCAACCCAAAG	4828.2
	3R	45881796	*Beat*	R	ACGTTGGATGCCGGACCGTAAAAGAGTTTG	ACGTTGGATGTAATACATAGCCTCGGCGTG	R	TCGGCGTGCTAATCGT	4888.2
	2L	3868552	AGAP004819	M	ACGTTGGATGCTATACCTTTGTCCTGTTTGG	ACGTTGGATGTCCATTTGTAGTGTCGATTG	R	GAAGGGTCATTTCCGT	4912.2
	3L	5263600	AGAP010510	R	ACGTTGGATGAGACTCACCTTACTCACAAG	ACGTTGGATGCGGTAAAAGTTAGATAGACC	F	cTCCATTTCAGCAACCA	5074.3
	2L	28255328	AGAP006207	W	ACGTTGGATGGTCCAGAAGTGCTCGAAATC	ACGTTGGATGAAGTACGATCAGGAGTACCG	F	ctTGCATCCGGTCGGGC	5178.4
	3L	36413144	-	Y	ACGTTGGATGGTAATCCCTGAGTGGGTGG	ACGTTGGATGGTAACGTCTCTAGCAACACC	F	GCAACACCTTTTTTCTGC	5400.5
	2L	3636997	AGAP004797	Y	ACGTTGGATGATAAAACTACTCGTTCCTCC	ACGTTGGATGCTTGCCTGTCTCCTTTTCTG	F	tTCCTTTTCTGTTGGCTT	5428.5
	2R	18031141	-	K	ACGTTGGATGCGCCAGGTTATTTATACTGC	ACGTTGGATGTTGCGTACCTGCATGCAAAA	F	gTGCATGCAAAATGTTCA	5522.6
	2L	3639220	*COPB2*	M	ACGTTGGATGTCGTCCAGATTCAGCGCTTC	ACGTTGGATGTGTCAGCAGCGCAAGTAAAG	R	AAAGCGGAAAAGTTCACT	5540.6
	2R	42843109	-	S	ACGTTGGATGCGGATGTTTACAATGTACGG	ACGTTGGATGCGCGTCCTAAAGTAGAACTG	R	GAGCCGTACGTTTGCACCA	5788.8
	3R	15531185		R	ACGTTGGATGACACAACTCTCTATGCCACC	ACGTTGGATGGCCTTTTTTGGAGACGCATC	R	gtGTGCGGCGCGGTCATGT	5891.8
	X	12109761	*WDR68*	R	ACGTTGGATGGCAGTTGTTAGAACGAGAAG	ACGTTGGATGGTACGGTATTTGTAGAGCTG	R	TAACTTCTATGTCAAACACC	6020
	2R	47283021	-	W	ACGTTGGATGCCAGGTGGAATAGAAAAGCG	ACGTTGGATGGCTCGAATAACTGCAAAGAG	F	gtGTGTTTCATTCCCTTTCG	6055.9
	3L	5263001	AGAP010510	W	ACGTTGGATGGACAGAAATGCTTGGGAAGG	ACGTTGGATGACCTCAGACAAAAAAAGGGC	R	GACAAAAAAAGGGCAACAGC	6186.1

2L	28254734	-	Y	ACGTTGGATGCTAACCAGTGGCGGTACAAT	ACGTTGGATGCATAGCCGTCCCGGTTTATT	F	cGATGAGATTAACTGCCTTCA	6405.2
3R	2283582	-	R	ACGTTGGATGCGTGGTCATGTGTTTTTTGC	ACGTTGGATGATGCTTGTGTTGTGTGTGAC	F	gGTGTTGTGTGTGACAACACC	6477.2
3R	39145523	-	M	ACGTTGGATGCTTGTAATAGCCGCGTTTCG	ACGTTGGATGTACGTTGTCATGGGAAACGG	R	aatAAACGGTGATTTTGACAG	6493.3
3R	43376739	-	S	ACGTTGGATGACCATAAGCTTTGCTTACTG	ACGTTGGATGATTTGTGGCCCATTTTTGTG	R	ccctCCATTTTTGTGGCCATTA	6627.3
3L	5851676	AGAP010510	Y	ACGTTGGATGATACAAATTTCAGTTTATAT	ACGTTGGATGGTCTTTCGATAAGCGGTCTG	R	gCTGTTTCAAAATCAATCACAT	6677.4
2L	49227823	AGAP007717	Y	ACGTTGGATGCGAACACAAGACATACTGAAG	ACGTTGGATGGTCGCCTGAACCATGATTAC	F	TGATTACATTCAACACTGTTAT	6683.4
3R	43657013	AGAP009801	M	ACGTTGGATGCCGCGACAGTGTATCTTATG	ACGTTGGATGCTGAACTCGTGAAATATTC	R	AACTCGTGAAATATTCACTAAT	6701.4
2L	28254678	-	W	ACGTTGGATGAATAAACCGGGACGGCTATG	ACGTTGGATGAGCCGCACATCATTACACAC	F	GGTTGGAAATTTTGGCTTTTAT	6801.4
2R	55234215	*Beat lib*	R	ACGTTGGATGAGACGGTACTGAATAGCAAC	ACGTTGGATGATGAGCAACTTTAACGAGCC	F	cGAATACTTTTAATTCACCTCTTC	7212.7
2L	2422651	*kdr*	Y	ACGTTGGATGCTTGGCCACTGTAGTGATAG	ACGTTGGATGAGCACCTGCAAAACAATGTC	R	ctTTAATTTGCATTACTTACGACT	7267.7
2R	55232273	*Beat lib*	R	ACGTTGGATGTCTCAACGTACTCGGTTGTG	ACGTTGGATGCGGACAACAACAAAAAGCGT	F	TGCAATTTGTTAGTATTTATCTCC	7298.8
2R	58955948	AGAP004644	M	ACGTTGGATGGACGATTACTAACTGACTAAC	ACGTTGGATGGAATTAAGGATCAGCATTTG	R	AGGATCAGCATTTGTTAAACTTAC	7350.8
2L	28254333	-	R	ACGTTGGATGGTACACGCCTATCGTGAATG	ACGTTGGATGAAGGTTTCCCCTTCTCAGAC	F	aTTCTCAGACAGTATGCATATCTTT	7606
2R	47283557	-	R	ACGTTGGATGCGACTCCGAAATACTCTAAC	ACGTTGGATGCGACTCGACTCCGGTTATTA	R	gCAGTTAAATCTAAACCATTAGCTC	7609
X	9236853	AGAP000519	R	ACGTTGGATGAAAGTGGCATTTCCGGGAAG	ACGTTGGATGACGTTCTATCACCTTTCCCG	R	tGAAAACCGGGACCAAAAGTTATTA	7707.1
2R	47283445	-	W	ACGTTGGATGAAAGGCCATCCTTCGAAGTG	ACGTTGGATGTTGGTCATTATAGGACCGGC	R	GCATTTCATTTTATTTAGTTTTTTCG	7913.2
X	11282305	-	Y	ACGTTGGATGCGGATCGGTTTGATATATT	ACGTTGGATGCAATTTCAGATTGCGCCAAG	R	ggGCGCATAAATTTTCAAGATTATAT	8008.2
3R	39145574	-	S	ACGTTGGATGGACAACGTAAAATCGAGCAG	ACGTTGGATGATTCTTGCCATCCAGACTGC	R	ATGGCAAAAATTGCACGAATAAGATTC	8324.5
3R	45891826	*Beat*	K	ACGTTGGATGGTAAGGTTCGTCTCACAATC	ACGTTGGATGTTGAATTTGAAGAAATTTAC	R	TGAAGAAATTTACATAACAATATCATAA	8595.7
X	9241714	AGAP000519	Y	ACGTTGGATGAGTTGTTGAGCCTCTAATT	ACGTTGGATGTCATGCGTATTAAACACAG	F	ACAATAGTCTCTACAATATTATCTATTTT	8799.8
2L	42671746	-	R	ACGTTGGATGGCAACTGACGTTATCACGAG	ACGTTGGATGTAGATTTGTTCTTGGGCGTC	F	cGATTTGTTCTTGGGCGTCCCCGGCTTTAC	9146.9
3L	34158609	-	K	ACGTTGGATGGCATCATTCAACGTTTGCAG	ACGTTGGATGTCAAATCACGGCGTTTCACC	R	GCGGCACGTGACAGTTCTGATCTGATTATC	9213

The assays were split into five multiplexes (W1-W5). Columns include the chromosome (chr), coordinates (coor) and gene of the assayed SNP. For genes with no common gene name, the Vectorbase ID is provided instead. Primer sequences, the direction and mass of the extension primer are also included. The IUPAC code represents the two possible alleles. Ext = extension.

Table S7. Combined result of the genome-wide association and replication experiment including results that were not significant.

		GWAS experiment: Gambian dataset				Follow-up experiment: Ugandan dataset				Combined				
SNP	Allele	AF Indoor	AF Outdoor	OR	*p* (2-sided)	AF Indoor	AF Outdoor	OR	*p* (1-sided)	OR	95% CI	*p*	Model	Nearby Gene
3R:43668381*	T	0.53	0.74	0.39	1.09×10^{-4}	0.12	0.14	1.25	0.21	0.66	0.45-0.96	2.81×10^{-6}	Add	AGAP009802 (*Gr12*)
3R:44747427	T	0.42	0.31	0.31	8.87×10^{-7}	0.99	0.98	0.59	0.26	0.33	0.21-0.54	5.96×10^{-6}	Add	AGAP009886 (*MED15*)
2R:18031010	T	0.19	0.41	2.15	2.42×10^{-6}	0.17	0.29	1.34	0.03	1.66	0.13-2.07	7.62×10^{-6}	Dom	AGAP002238
3R:39145574	C	0.14	0.31	2.07	1.66×10^{-5}	0	0.02	1.83	0.16	2.05	1.48-2.84	1.67×10^{-5}	Dom	AGAP009700 (*SWD2*)
X:9227517	T	0.94	0.8	0.28	1.47×10^{-4}	0.98	0.97	0.66	0.22	0.35	0.20-0.61	2.02×10^{-4}	Add	AGAP000519
3R:37557083	T	0.24	0.1	0.26	4.53×10^{-6}	0.05	0.04	0.84	0.26	0.49	0.33-0.72	2.75×10^{-4}	Add	AGAP009646 (*CDX2*)
3R:43669141*	T	0.46	0.37	0.61	5.18×10^{-3}	0.82	0.75	0.71	0.36	0.67	0.54-0.83	2.87×10^{-4}	Dom	AGAP009803 (*Gr11*)
3R:37563683	C	0.12	0.31	3.86	4.53×10^{-6}	0.12	0.14	1.19	0.26	1.96	0.13-2.93	9.53×10^{-4}	Add	AGAP009646 (*CDX2*)
X:3492955	A	0.18	0.04	0.13	2.96×10^{-6}	0.01	0.01	1.75	0.29	0.23	0.09-0.55	1.01×10^{-3}	Add	AGAP000214 (*GUCY1B*)
3L:34158609	G	0.63	0.85	3.82	2.11×10^{-6}	0.79	0.8	1.07	0.38	1.74	1.24-2.46	1.55×10^{-3}	Add	AGAP011852
3L:35184869	A	0.83	0.94	1.85	2.16×10^{-3}	0.87	0.93	1.42	0.08	1.67	1.20-2.31	2.32×10^{-3}	Dom	AGAP011925 (*PNLP5*)
2R:58955948*	A	0.7	0.84	2.27	3.29×10^{-3}	0.99	0.99	1.15	0.45	2.2	1.33-3.57	2.33×10^{-3}	Add	AGAP004644
2L:3636997*	C	0.66	0.84	1.85	1.66×10^{-4}	0.92	0.94	1.07	0.37	1.49	1.15-1.93	2.43×10^{-3}	Dom	AGAP004797
2L:22334428	C	0.21	0.39	2.02	1.43×10^{-5}	0.16	0.13	0.93	0.35	1.43	1.12-1.82	4.32×10^{-3}	Dom	AGAP005839, MFS transporter
2R:18031141	T	0.16	0.38	3.06	9.69×10^{-6}	0.14	0.15	1.06	0.4	1.61	1.15-2.26	5.87×10^{-3}	Add	AGAP002238
2L:25464003	G	0.94	0.77	0.2	4.07×10^{-6}	0.84	0.84	0.99	0.49	0.55	0.37-0.84	5.87×10^{-3}	Add	AGAP006031, (*Nup54*)
2L:28254333	G	0.17	0.28	1.55	6.36×10^{-3}	0.13	0.18	1.22	0.14	1.4	1.10-1.77	6.12×10^{-3}	Dom	AGAP006207 (*carboxypeptidase B*)
2R:47283021*	T	0.12	0.31	1.9	1.11×10^{-4}	0.1	0.07	0.83	0.23	1.47	1.11-1.94	6.19×10^{-3}	Dom	AGAP003971
3L:33525031	A	0.08	0.27	2.26	5.06×10^{-6}	0.13	0.12	0.95	0.37	1.39	1.09-1.78	7.19×10^{-3}	Dom	AGAP011812
3R:37462227¥	A	0.35	0.27	0.78	0.11	0.19	0.11	0.68	0.01	0.73	0.58-0.92	8.27×10^{-3}	Dom	AGAP009640 *(Or1)*
3R:44744808	A	0.67	0.52	2.27	3.5×10^{-4}	0.20	0.20	1.02	0.48	1.56	1.11-2.18	0.01	Add	AGAP009886
3R:43728754	C	0.79	0.93	1.93	4.18×10^{-4}	0.96	0.95	0.9	0.34	1.45	1.07-1.97	0.02	Dom	AGAP009809
2R:283739*	G	0.84	0.95	3.19	2.04×10^{-3}	0.9	0.92	1.2	0.27	1.72	1.08-2.73	2.33×10^{-2}	Add	AGAP011509 (*COEunkn*)
3R:43668381	T	0.53	0.74	0.39	1.09×10^{-4}	0.12	0.14	1.25	0.21	0.66	0.45-0.96	0.03	Add	AGAP009802 (*Gr12*)
3R:37461899¥	T	0.43	0.37	0.8	0.32	0.17	0.09	0.5	0.01	0.68	0.47-0.96	0.03	Add	AGAP009640 (*OR1*)
3R:43665034*	T	0.41	0.22	0.38	1.55×10^{-4}	0.68	0.69	1.06	0.4	0.7	0.51-0.98	3.86×10^{-2}	Add	AGAP009801
3R:45900850*	C	0.63	0.57	2.06	7.26×10^{-4}	0.34	0.33	0.96	0.41	1.36	1.01-1.82	3.97×10^{-2}	Add	AGAP009947 (*Beat*)
3R:43669258*	C	0.45	0.39	0.61	6.52×10^{-3}	0.47	0.46	0.94	0.34	0.79	0.63-0.99	0.04	Dom	AGAP009803 (*Gr11*)
2R:47284139*	A	0.17	0.31	2.19	3.76×10^{-3}	0.09	0.09	0.99	0.48	1.5	1.01-2.22	0.04	Add	AGAP003971
3L:29895651	C	0.2	0.41	2.12	3.41×10^{-6}	0.34	0.26	0.85	0.13	1.24	1.00-1.52	0.05	Dom	AGAP011616
2L:3864566	C	0.26	0.13	0.44	3.48×10^{-3}	0.16	0.15	0.95	0.42	0.69	0.31-1.07	0.06	Add	AGAP001370
3L:35174648	C	0.64	0.59	2.33	8.96×10^{-5}	0.19	0.13	0.66	0.06	1.37	1.03-1.7	0.07	Add	AGAP011925
3L:26103443	T	0.63	0.58	2.09	4.8×10^{-4}	0.54	0.53	0.97	0.42	1.26	1.01-1.51	0.07	Add	AGAP011509 (*COEunkn*)
3R:45893405	G	0.66	0.54	2.09	6.15×10^{-5}	0.19	0.15	0.88	0.21	1.23	0.99-1.48	0.09	Dom	AGAP009947 (*Beat*)
3L:35174543	T	0.65	0.63	3.33	4.17×10^{-7}	0.26	0.18	0.63	0.02	1.34	1.00-1.69	0.09	Add	AGAP011925
2L:34623415	G	0.61	0.58	2.18	$.47 \times 10^{-4}$	0.33	0.3	0.92	0.32	1.28	0.99-1.57	0.09	Add	AGAP006598
2L:42671746	A	0.64	0.6	2.74	9.17×10^{-6}	0.29	0.22	0.79	0.11	1.28	0.99-1.57	0.1	Add	AGAP007100
2L:28255328	T	0.45	0.4	0.5	3.19×10^{-3}	0.43	0.44	1.05	0.41	0.78	0.48-1.08	0.1	Add	AGAP006207
2L:28257381	A	0.11	0.24	1.74	1.12×10^{-3}	0.16	0.12	0.83	0.15	1.22	0.98-1.46	0.11	Dom	AGAP006208
2L:37280520	A	0.31	0.46	1.75	8.22×10^{-4}	0.15	0.12	0.86	0.18	1.22	0.98-1.45	0.11	Dom	AGAP006723,COEAE2G
3L:5851676	C	0.81	0.93	3.11	1.42×10^{-3}	0.93	0.89	0.63	0.07	0.8	0.96-1.66	0.11	Add	AGAP010540
3R:43667573	T	0.45	0.27	0.64	5.07×10^{-3}	0.43	0.47	1.07	0.33	0.84	0.63-1.06	0.12	Dom	AGAP009947 (*Beat*)
2L:28257115	A	0.81	0.97	3.07	5.86×10^{-7}	0.83	0.81	0.93	0.32	1.23	0.97-1.49	0.12	Dom	AGAP006208 (*carboxypeptidase B*)
3R:44747655	A	0.24	0.15	0.67	0.02	0.24	0.25	1	0.49	0.85	0.63-1.06	0.13	Dom	AGAP009886
3R:37462107	A	0.13	0.13	0.98	1.45×10^{-3}	0.02	0.02	0.75	0.22	1.51	0.93-2.05	0.16	Rec	AGAP009640 (*OR1*)
3R:37462530	G	0.22	0.19	1.08	0.84	0.31	0.32	1.23	0.17	1.29	0.92-1.67	0.16	Rec	AGAP009640 (*OR1*)
3R:37462118	T	0.17	0.18	0.66	0.68	0.02	0.01	1.33	0.24	1.29	0.94-2.09	0.16	Rec	AGAP009640 (*OR1*)

2L:3619546	G	0.34	0.2	0.77	0.39	0.3	0.34	1.2	0.2	1.27	0.93-1.61	0.16	Rec	AGAP004795
X:7139849	G	0.83	0.94	1.81	3.29×10^{-3}	0.96	0.93	0.76	0.2	1.28	0.93-1.64	0.17	Rec	AGAP000389
X:7150927	G	0.24	0.34	1.82	0.09	0.33	0.27	0.83	0.18	1.29	0.93-1.65	0.17	Rec	AGAP000389
3L:5842042	C	0.81	0.95	1.95	1.27×10^{-3}	0.76	0.76	0.95	0.36	1.17	0.93-1.41	0.19	Dom	AGAP010540
2R:278043	T	0.32	0.16	0.38	0.02	0.65	0.69	0.93	0.49	1.25	0.91-1.58	0.2	Rec	AGAP001109
3L:5861278	A	0.82	0.94	3.65	5.64×10^{-4}	0.93	0.89	0.62	0.08	1.3	0.79-1.81	0.31	Add	AGAP010540
X:12110046	T	0.31	0.16	0.63	4.59×10^{-3}	0.28	0.39	1.28	0.04	0.95	0.74-1.16	0.65	Dom	AGAP000678
3R:37462690	T	0.24	0.26	1.1	0.69	0.23	0.24	1.05	0.41	1.07	0.76-1.38	0.66	Add	AGAP009640 (*OR1*)
3R:37461993	G	0.21	0.21	0.99	0.96	0.03	0.02	0.76	0.32	0.95	0.47-1.42	0.82	Add	AGAP009640 (*OR1*)

SNPs are ordered in decreasing order of significance. Results were combined using a fixed-effect model by inverse-variance weighting of log odd ratios. The odds ratios (OR) and allele frequency (AF) are defined for the allele specified. The nearby gene (Vectorbase ID) is defined as the closest gene within 200kb of the SNP. CI= Confidence interval. *SNPs that were chosen because they have F_{st} estimates that corresponded to the upper 0.1-5% of the distribution. ¥Denotes a SNP found within our genes of interest (GOI).

References

Alexander, D. H., Novembre, J. & Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research,* 19**,** 1655–1664.

Alphey, L., Nimmo, D., O'Connell, S. & Alphey, N. 2008. Insect Population Suppression Using Engineered Insects. *In:* Aksoy, S. (ed.) *Transgenesis and the Management of Vector-Borne Disease.* Springer New York.

Amos, W., Driscoll, E. & Hoffman, J. I. 2010. Candidate genes versus genome-wide associations: which are better for detecting genetic susceptibility to infectious disease? *Proceedings of the Royal Society of London, Series B: Biological Sciences*.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T.*, et al.* 2014. An atlas of active enhancers across human cell types and tissues. *Nature,* 507**,** 455-461.

Antonio-Nkondjio, C., Awono-Ambene, P., Toto, J. C., Meunier, J. Y., Zebaze-Kemleu, S., Nyambam, R., Wondji, C. S., Tchuinkam, T. & Fontenill, D. 2002. High malaria transmission intensity in a village close to Yaounde, the capital city of Cameroon. *J Med Entomol,* 39**,** 350-5.

Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics,* 23**,** 1294-1296.

Baker, D., Nolan, T., Fischer, B., Pinder, A., Crisanti, A. & Russell, S. 2011. A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, *Anopheles gambiae*. *BMC Genomics,* 12**,** 296.

Balenghien, T., Fouque, F., Sabatier, P. & Bicout, D. J. 2006. Horse-, bird-, and human-seeking behavior and seasonal abundance of mosquitoes in a West Nile virus focus of southern France. *Journal of Medical Entomology,* 43**,** 936-946.

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, D., Barmada, M. M.*, et al.* 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet,* 40**,** 955-962.

Bayoh, M. N., Mathias, D. K., Odiere, M. R., Mutuku, F. M., Kamau, L., Gimnig, J. E., Vulule, J. M., Hawley, W. A., Hamel, M. J. & Walker, E. D. 2010. *Anopheles gambiae:* historical population decline associated with regional distribution of insecticide-treated bed nets in western Nyanza Province, Kenya. *Malar Journal,* 9**,** 62.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R.*, et al.* 2008.

Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456**,** 53-59.

Benton, R., Sachse, S., Michnick, S. W. & Vosshall, L. B. 2006. Atypical membrane topology and heteromeric function of *Drosophila* odorant receptors *in vivo*. *PLoS Biol,* 4**,** e20.

Benton, R., Vannice, K. S., Gomez-Diaz, C. & Vosshall, L. B. 2009. Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. *Cell,* 136**,** 149-62.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R.*, et al.* 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol,* 28**,** 1045 - 1048.

Besansky, N. J. 2008. Genome Analysis Of Vectorial Capacity In Major Anopheles Vectors Of Malaria Parasites. *In:* Committee, A. G. C. (ed.).

Besansky, N. J., Lehmann, T., Fahey, G. T., Fontenille, D., Braack, L. E. O., Hawley, W. A. & Collins, F. H. 1997. Patterns of mitochondrial variation within and between African malaria vectors, *Anopheles gambiae* and *An. arabiensis*, suggest extensive gene flow. *Genetics,* 147**,** 1817-1828.

Bette, S., Breer, H. & Krieger, J. 2002. Probing a pheromone binding protein of the silkmoth *Antheraea polyphemus* by endogenous tryptophan fluorescence. *Insect Biochemistry and Molecular Biology,* 32**,** 241-246.

Biessmann, H., Nguyen, Q. K., Le, D. & Walter, M. F. 2005. Microarray-based survey of a subset of putative olfactory genes in the mosquito *Anopheles gambiae*. *Insect Molecular Biology,* 14**,** 575 - 89.

Bird, A. 2007. Perceptions of epigenetics. *Nature,* 447**,** 396-398.

Biryukova, I., Ye, T. & Levashina, E. 2014. Transcriptome-wide analysis of microRNA expression in the malaria mosquito *Anopheles gambiae*. *BMC Genomics,* 15**,** 557.

Black, W. C. & Lanzaro, G. C. 2001. Distribution of genetic variation among chromosomal forms of *Anopheles gambiae s.s*: introgressive hybridization, adaptive inversions, or recent reproductive isolation? *Insect Mol Biol,* 10**,** 3-7.

Bockarie, M. J., Service, M. W., Toure, Y. T., Traore, S., Barnish, G. & Greenwood, B. M. 1993. The ecology and behaviour of the forest form of *Anopheles gambiae s.s. Parassitologia,* 35**,** 5-8.

Bøgh, C., Pedersen, E. M., Mukoko, D. A. & Ouma, J. H. 1998. Permethrin-impregnated bed net effects on resting and feeding behaviour of lymphatic filariasis vector mosquitoes in Kenya. *Medical and Veterinary Entomology,* 12**,** 52 - 59.

Bohbot, J. D., Jones, P. L., Wang, G., Pitts, J. R., Pask, G. M. & Zwiebel, L. J. 2010a. Conservation of indole responsive odorant receptors in mosquitoes reveals an ancient olfactory trait. *Chemical Senses*.

Bohbot, J. D., Lu, T. & Zwiebel, L. J. 2010b. Molecular regulation of olfaction in mosquitoes. *In:* Takken, W. & Knols, B. G. (eds.) *Olfaction in Vector-Host Interactions.* Wageningen Academic Publishers.

Bohbot, J. D., Sparks, J. T. & Dickens, J. C. 2014. The maxillary palp of *Aedes aegypti*, a model of multisensory integration. *Insect Biochem Mol Biol,* 48**,** 29-39.

Bosch, O. J., Geier, M. & Boeckh, J. 2000. Contribution of fatty acids to olfactory host finding of female *Aedes aegypti*. *Chem Senses,* 25**,** 323-30.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S*., et al.* 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research,* 22**,** 1790-1797.

Bradley, J., Lines, J., Fuseini, G., Schwabe, C., Monti, F., Slotman, M., Vargas, D., Garcia, G., Hergott, D. & Kleinschmidt, I. 2015. Outdoor biting by *Anopheles* mosquitoes on Bioko Island does not currently impact on malaria control. *Malar J,* 14**,** 170.

Braginets, O. P., Minakawa, N., Mbogo, C. M. & Yan, G. 2003. Population genetic structure of the African malaria mosquito *Anopheles funestus* in Kenya. *The American Journal of Tropical Medicine and Hygiene,* 69**,** 303-308.

Braimah, N., Drakeley, C. J., Kweka, E., Mosha, F., Helinski, M., Pates, H., Maxwell, C., Massawe, T., Kenward, M. G. & Curtis, C. 2005. Tests of bednet traps (Mbita traps) for monitoring mosquito populations and time of biting in Tanzania and possible impact of prolonged insecticide treated net use. *International Journal of Tropical Insect Science,* 25**,** 208-213.

Bright, A., Tewhey, R., Abeles, S., Chuquiyauri, R., Llanos-Cuentas, A., Ferreira, M., Schork, N., Vinetz, J. & Winzeler, E. 2012. Whole genome sequencing analysis of *Plasmodium vivax* using whole genome capture. *BMC Genomics,* 13**,** 262.

Brown, A. H. D. 1970. The estimation of Wright's fixation index from genotypic frequencies. *Genetica,* 41**,** 399-406.

Bryan, J. H., Di Deco, M. A., Petrarca, V. & Coluzzi, M. 1982. Inversion polymorphism and incipient speciation in *Anopheles gambiae s.s.* in Gambia, West Africa. *Genetica,* 59**,** 167 - 176.

Bryan, J. H., Petrarca, V., Di Deco, M. A. & Coluzzi, M. 1987. Adult behaviour of members of the *Anopheles gambiae* complex in the Gambia with special reference to *An. melas* and its chromosomal variants. *Parassitologia,* 29**,** 221 - 249.

Bugoro, H., Cooper, R. D., Butafa, C., Iro'ofa, C., Mackenzie, D. O., Chen, C. & Russell, T. L. 2011. Bionomics of the malaria vector *Anopheles farauti* in Temotu Province, Solomon Islands: issues for malaria elimination. *Malar Journal,* 10**,** 133.

Caputo, B., Nwakanma, D. C., Jawara, M., Adiamoh, M., Dia, I., Konate, L., Petrarca, V., Conway, D. & della Torre, A. 2008. *Anopheles gambiae* complex along The Gambia river, with particular reference to the molecular forms of *An. gambiae s.s. Malaria Journal,* 7**,** 182.

Cardon, L. R. & Bell, J. I. 2001. Association study designs for complex diseases. *Nat Rev Genet,* 2**,** 91-99.

Carey, A. F., Wang, G., Su, C., Zwiebel, L. R. & Carlson, J. R. 2010. Odorant reception in the malaria mosquito *Anopheles gambiae*. *Nature,* 464**,** 66-71.

Cassone, B. J., Mouline, K., Hahn, M. W., White, B. J., Pombi, M., Simard, F., Costantini, C. & Besansky, N. J. 2008. Differential gene expression in incipient species of *Anopheles gambiae*. *Molecular Ecology,* 17**,** 2491-504.

Cavalli-Sforza, L. L. & Feldman, M. W. 2003. The application of molecular genetic approaches to the study of human evolution. *Nature Genetics,* 33**,** 266-275.

CDC. 2010. *Malaria* [Online]. Centers for Disease Control and Prevention. Available: http://www.cdc.gov/malaria/about/biology/index.html.

Ceesay, S. J., Casals-Pascual, C., Erskine, J., Anya, S. E., Duah, N. O., Fulford, A. J. C., Sesay, S. S. S., Abubakar, I., Dunyo, S., Sey, O.*, et al.* 2008. Changes in malaria indices between 1999 and 2007 in The Gambia: a retrospective analysis. *The Lancet,* 372**,** 1545-1554.

Chandre, F., Darriet, F., Manguin, S., Brengues, C., Carnevale, P. & Guillet, P. 1999. Pyrethroid cross resistance spectrum among populations of *Anopheles gambiae s.s.* from Cote d'Ivoire. *J Am Mosq Control Assoc,* 15**,** 53-59.

Chang, B. L., Cramer, S. D., Wiklund, F., Isaacs, S. D., Stevens, V. L., Sun, J., Smith, S., Pruett, K., Romero, L. M., Wiley, K. E.*, et al.* 2009. Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. *Human Molecular Genetics,* 18**,** 1368-1375.

Charlwood, J. D. & Graves, P. M. 1987. The effect of permethrin-impregnated bednets on a population of *Anopheles farauti* in coastal Papua New Guinea. *Med Vet Entomol,* 1**,** 319 - 327.

Charlwood, J. D., Smith, T., Kihonda, J., Heiz, B., Billingsley, P. F. & Takken, W. 1995. Density-independent feeding success of malaria vectors (Diptera: Culicidae) in Tanzania. *Bull Entomol Res,* 85**,** 29 - 35.

Cheng, Y., Endo, K., Wu, K., Rodan, A. R., Heberlein, U. & Davis, R. L. 2001. *Drosophila fasciclinII* Is required for the formation of odor memories and for normal sensitivity to alcohol. *Cell,* 105**,** 757-768.

Chilaka, N., Perkins, E. & Tripet, F. 2012. Visual and olfactory associative learning in the malaria vector *Anopheles gambiae sensu stricto*. *Malaria Journal,* 11**,** 27.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin),* 6**,** 80-92.

Clarkson, C. S., Weetman, D., Essandoh, J., Yawson, A. E., Maslen, G., Manske, M., Field, S. G., Webster, M., Antão, T., MacInnis, B.*, et al.* 2014. Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation. *Nat Commun,* 5.

Coetzee, M., Craig, M. & le Sueur, D. 2000. Distribution of African malaria mosquitoes belonging to the *Anopheles gambiae* complex. *Parasitol Today,* 16**,** 74 - 77.

Coetzee, M., Hunt, R. H., Wilkerson, R., Della Torre, A., Coulibaly, M. B. & Besansky, N. J. 2013. *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa,* 3619**,** 246 - 274.

Collins, F. H. & Besansky, N. J. 1994. Vector biology and the control of malaria in Africa. *Science,* 264**,** 1874-1875.

Collins, F. H. & Paskewitz, S. M. 1995. Malaria: current and future prospects for control. *Annual Review of Entomology,* 40**,** 195-219.

Coluzzi, M., Petrarca, V. & Dideco, M. A. 1985. Chromosomal inversion intergradation and incipient speciation in *Anopheles gambiae*. *Bolletino di Zoologia,* 52**,** 45 - 63.

Coluzzi, M., Sabatini, A., della Torre, A., Di Deco, M. A. & Petrarca, V. 2002. A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science,* 298**,** 1415 - 1418.

Coluzzi, M., Sabatini, A., Petrarca, V. & Di Deco, M. A. 1977. Behavioural divergences between mosquitoes with different inversion karyotypes in polymorphic populations of the *Anopheles gambiae* complex. *Nature* 266**,** 832-833.

Coluzzi, M., Sabatini, A., Petrarca, V. & Di Deco, M. A. 1979. Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 73**,** 483-497.

Conneely, K. N. 2008. *Methods for Large-scale Genetic Association Studies.* PhD, University of Michigan

Cork, A. & Park, K. C. 1996. Identification of electrophysiologically-active compounds for the malaria mosquito, *Anopheles gambiae,* in human sweat extracts. *Medical and Veterinary Entomology,* 10**,** 269-276.

Costantini, C., Gibson, G., Sagnon, N., Della Torre, A., Brady, J. & Coluzzi, M. 1996. Mosquito responses to carbon dioxide in a west African Sudan savanna village. *Med Vet Entomol,* 10**,** 220 - 227.

Crawford, J. E., Bischoff, E., Garnier, T., Gneme, A., Eiglmeier, K., Holm, I., Riehle, M. M., Guelbeogo, W. M., Sagnon, N., Lazzaro, B. P.*, et al.* 2012. Evidence for population-specific positive selection on immune genes of *Anopheles gambiae*. *G3 (Bethesda),* 2**,** 1505-19.

Crawford, J. E. & Lazzaro, B. P. 2010. The demographic histories of the M and S molecular forms of *Anopheles gambiae s.s*. *Molecular Biology and Evolution,* 27**,** 1739-1744.

Crooks, L., Carlborg, O., Marklund, S. & Johansson, A. M. 2013. Identification of null alleles and deletions from SNP genotypes for an intercross between domestic and wild chickens. *G3,* 3**,** 1253-60.

Cuzin-Ouattara, N., Van den Broek, A. H., Habluetzel, A., Diabate, A., Sanogo-Ilboudo, E., Diallo, D. A., Cousens, S. N. & Esposito, F. 1999. Wide-scale installation of insecticide-treated curtains confers high levels of protection against malaria transmission in a hyperendemic area of Burkina Faso. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 93**,** 473-9.

Dabire, K. R., Sawadodgo, S., Diabate, A., Toe, K. H., Kengne, P., Ouari, A., Costantini, C., Gouagna, C., Simard, F., Baldet, T.*, 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. *Med Vet Entomol,* 27**,** 298-312.

Dahanukar, A., Hallem, E. A. & Carlson, J. R. 2005. Insect chemoreception. *Current Opinion in Neurobiology,* 15**,** 423-430.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T.*, et al.* 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.

Das, S. & Dimopoulos, G. 2008. Molecular analysis of photic inhibition of blood-feeding in *Anopheles gambiae*. *BMC Physiology,* 8**,** 23.

Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. 2001. Rapid amplification of plasmid and phage DNA using phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Research,* 11**,** 1095 - 1099.

Dekker, T. & Takken, W. 1998. Differential responses of mosquito sibling species *Anopheles arabiensis* and *An. quadriannulatus* to carbon dioxide, a man or a calf. *Med Vet Entomol,* 12**,** 136-40.

della Torre, A., Costantini, C., Besansky, N. J., Caccone, A., Petrarca, V., Powell, J. R. & Coluzzi, M. 2002. Speciation within *Anopheles gambiae*-the glass is half full. *Science,* 298**,** 115-7.

della Torre, A., Fanello, C., Akogbeto, M., Dossou-yovo, J., Favia, G., Petrarca, V. & Coluzzi, M. 2001. Molecular evidence of incipient speciation within *Anopheles gambiae s.s.* in West Africa. *Insect Mol Biol,* 10**,** 9 - 18.

della Torre, A., Tu, Z. & Petrarca, V. 2005. On the distribution and genetic differentiation of *Anopheles gambiae s.s.* molecular forms. *Insect Biochem Mol Biol,* 35**,** 755 - 769.

DFID 2010. Malaria: Burden and Interventions. Department for International Development.

Diabaté, A., Baldet, T., Chandre, F., Dabire, R. K., Simard, F., Ouedraogo, J. B., Guillet, P. & Hougard, J. M. 2002. The role of agricultural use of insecticides in resistance to pyrethroids in *Anopheles gambiae s.l.* in Burkina Faso. *Am J Trop Med Hyg,* 67**,** 617 - 622.

Diabaté, A., Dabire, R. K., Heidenberger, K., Crawford, J., Lamp, W. O., Culler, L. E. & Lehmann, T. 2008. Evidence for divergent selection between the molecular forms of *Anopheles gambiae*: role of predation. *BMC Evolutionary Biology,* 8**,** 5.

Diabaté, A., Dao, A., Yaro, A. S., Adamou, A., Gonzalez, R., Manoukis, N. C., Traore, S. F., Gwadz, R. W. & Lehmann, T. 2009. Spatial swarm segregation and reproductive isolation between the molecular forms of *Anopheles gambiae*. *Proc Biol Sci,* 276**,** 4215-22.

Eldridge, B. F. 2005. Mosquitoes, the Culicidae. *In:* Marquardt, W. C. (ed.) *Biology of Disease Vectors* 2nd edition ed. USA: Elsevier Academic Press

ENCODE 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

Falush, D., Stephens, M. & Pritchard, J. K. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics,* 164**,** 1567 - 1587.

Falush, D., Stephens, M. & Pritchard, J. K. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes,* 7**,** 574-578.

Favia, G., della Torre, A., Bagayoko, M., Lanfrancotti, A., Sagnon, N., Toure, Y. T. & Coluzzi, M. 1997. Molecular identification of sympatric chromosomal forms of

Anopheles gambiae and further evidence of their reproductive isolation. *Insect Mol Biol,* 6**,** 377-83.

Fay, J. C. & Wu, C. I. 2000. Hitchhiking under positive Darwinian selection. *Genetics,* 155**,** 1405-1413.

FIMM. 2012. *SNP-Sequenom* [Online]. Available: http://www.fimm.fi/en/technologycentre/genomics/snp_genotyping/sequenom/ [Accessed 22/11 2013].

Fornadel, C. M., Norris, L. C., Glass, G. E. & Norris, D. E. 2010. Analysis of *Anopheles arabiensis* blood feeding behavior in Southern Zambia during the two years after introduction of insecticide-treated bed nets. *The American Journal of Tropical Medicine and Hygiene,* 83**,** 848-853.

Foster, W. A. & Takken, W. 2004. Nectar-related vs. human-related volatiles: behavioural response and choice by female and male *Anopheles gambiae* (Diptera: Culicidae) between emergence and first feeding. *Bulletin of Entomological Research,* 94**,** 145-157.

Fox, A. N., Pitts, R. J., Robertson, H. M., Carlson, J. R. & Zwiebel, L. J. 2001. Candidate odorant receptors from the malaria vector mosquito *Anopheles gambiae* and evidence of down-regulation in response to blood feeding. *Proc Natl Acad Sci U S A,* 98**,** 14693-7.

Gabriel, S., Ziaugra, L. & Tabbaa, D. 2009. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current Protocols in Human Genetics,* 2.

Gamble, C., Ekwaru, P. J., Garner, P. & ter Kuile, F. O. 2007. Insecticide-treated nets for the prevention of malaria in pregnancy: a systematic review of randomised controlled trials. *PLoS Med,* 4**,** e107.

Gatton, M. L., Chitnis, N., Churcher, T., Donnelly, M. J., Ghani, A. C., Godfray, H. C., Gould, F., Hastings, I., Marshall, J., Ranson, H.*, et al.* 2013. The importance of mosquito behavioural adaptations to malaria control in Africa. *Evolution,* 67**,** 1218 - 1230.

Geier, M., Bosch, O. J. & Boeckh, J. 1999. Ammonia as an attractive component of host odour for the yellow fever mosquito, *Aedes aegypti*. *Chem Senses,* 24**,** 647-53.

Gillet, J. D. 1972. *Common Africa Mosquitoes and their Medical Importance,* London, Wilem Heimann Medical Books Ltd.

Gillies, M. T. 1964. Selection for host preference in *Anopheles gambiae*. *Nature,* 203**,** 852-854.

Gillies, M. T. 1980. The role of carbon dioxide in host-finding by mosquitoes (Diptera:Culicidae): a review. *Bull Entomol Res* 70**,** 525-532.

Gillies, M. T. & Coetzee, M. 1987. A supplement to the Anophelinae of Africa south of the Sahara (Afrotropical Region). *American Journal of Tropical Medicine and Hygiene,* 55**,** 1-143.

Gillies, M. T. & De Meillon, B. 1968. The Anophelinae of Africa south of the Sahara (Ethiopian Zoogeographical Region). *Publ S Afr Inst Med Res,* 54**,** 1 - 343.

Gillies, M. T. & Smith, A. 1960. The effect of a residual house-spraying campaign on species balance in the *Anopheles funestus* group: The replacement of *Anopheles gambiae* Giles with *Anopheles rivulorum* Leeson. *Bulletin of Entomological Research* 51**,** 248 - 252.

Gimonneau, G., Bouyer, J., Morand, S., Besansky, N. J., Diabaté, A. & Simard, F. 2010. A behavioral mechanism underlying ecological divergence in the malaria mosquito *Anopheles gambiae*. *Behavioral Ecology,* 21**,** 1087-1092.

Gómez-Díaz, E., Rivero, A., Chandre, F. & Corces, V. G. 2014. Insights into the epigenomic landscape of the human malaria vector *Anopheles gambiae*. *Frontiers in Genetics,* 5**,** 277.

Google Inc. 2013. *Google Earth (Version 7.1.2.2041)* [Online]. Available: http://www.google.co.uk/intl/en_uk/earth/.

Govella, N. J., Okumu, F. O. & Killeen, G. F. 2010. Insecticide-treated nets can reduce malaria transmission by mosquitoes which feed outdoors. *The American Journal of Tropical Medicine and Hygiene,* 82**,** 415 - 419.

Haff, L. A. & Smirnov, I. P. 1997. Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. *Genome Research,* 7**,** 378-88.

Hallem, E. A., Nicole Fox, A., Zwiebel, L. J. & Carlson, J. R. 2004. Olfaction: mosquito receptor for human-sweat odorant. *Nature,* 427**,** 212-3.

Hangauer, M. J., Vaughn, I. W. & McManus, M. T. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genetics,* 9**,** e1003569.

Harrison, G. 1978. *Mosquitoes, Malaria and Man: a history of hostilities since 1880,* London, UK, John Murray.

Hay, S. I., Sinka, M. E., Okara, R. M., Kabaria, C. W., Mbithi, P. M., Tago, C. C., Benz, D., Gething, P. W., Howes, R. E., Patil, A. P.*, et al.* 2010. Developing global maps of the dominant *Anopheles* vectors of human malaria. *PLoS Med,* 7**,** e1000209.

Hays, C. W. 2000. The United States Army and malaria control in World War II. *Parassitologia,* 42**,** 47-52.

Heid, I. M., Lamina, C., Kuchenhoff, H., Fischer, G., Klopp, N., Kolz, M., Grallert, H., Vollmert, C., Wagner, S., Huth, C.*, et al.* 2008. Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in a large epidemiologic sample. *American Journal of Epidemiology,* 168**,** 878-89.

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, A., Maes, G. E., Diopere, E., Carvalho, G. R.*, et al.* 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources,* 11**,** 123-136.

Hill, C. A., Fox, N. A., Pitts, J. R., Kent, L. B., Tan, P. L., Chrystal, M. A., Cravchik, A. l., Collins, F. l. H., Robertson, H. l. M. & Zwiebel, L. I. J. 2002. G Protein-Coupled Receptors in *Anopheles gambiae*. *Science,* 298**,** 176-178.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. & Manolio, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A,* 106**,** 9362-7.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A. & Cox, D. R. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science,* 307**,** 1072-1079.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M. C., Wides, R.*, et al.* 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science,* 298**,** 129-149.

Howard, A., Zhou, G. & Omlin, F. 2007. Malaria mosquito control using edible fish in western Kenya: preliminary findings of a controlled study. *BMC Public Health,* 7**,** 199.

Hudson, R. R., Kreitman, M. & Aguadé, M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics,* 116**,** 153-159.

Hunt, R. H., Coetzee, M. & Fettene, M. 1998. The *Anopheles gambiae* complex: a new species from Ethiopia. *Trans R Soc Trop Med Hyg,* 92**,** 231 - 235.

Iatrou, K. & Biessmann, H. 2008. Sex-biased expression of odorant receptors in antennae and palps of the African malaria vector *Anopheles gambiae*. *Insect Biochemistry and Molecular Biology,* 38**,** 268 - 74.

Ilboudo-Sanogo, E., Cuzin-Ouattara, N., Diallo, D. A., Cousens, S. N., Esposito, F., Habluetzel, A., Sanon, S. & Ouédraogo, A. P. 2001. Insecticide-treated materials, mosquito adaptation and mass effect: entomological observations after five years of vector control in Burkina Faso. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 95**,** 353-360.

Illumina. 2013. An introduction to next-generation sequencing technology Available:http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

IR Mapper. Available: www.irmapper.com [Accessed March 2015].

Jaenisch, R. & Bird, A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet,* 33 Suppl**,** 245-54.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R.*, et al.* 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature,* 451**,** 998-1003.

Jawara, M., Smallegange, R. C., Jeffries, D., Nwakanma, D. C., Awolola, T. S., Knols, B. G. J., Takken, W. & Conway, D. J. 2009. Optimizing odor-baited trap methods for collecting mosquitoes during the malaria season in The Gambia. *PLoS ONE,* 4**,** e8167.

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics,* 24**,** 1403-1405.

Kabbale, F., Akol, A., Kaddu, J. & Onapa, A. 2013. Biting patterns and seasonality of *Anopheles gambiae sensu lato* and *Anopheles funestus* mosquitoes in Kamuli District, Uganda. *Parasites & Vectors,* 6**,** 340.

Kamau, L., Mukabana, W. R., Hawley, W. A., Lehmann, T., Irungu, L. W., Orago, A. A. & Collins, F. H. 1999. Analysis of genetic variability in *Anopheles arabiensis* and *Anopheles gambiae* using microsatellite loci. *Insect Mol Biol,* 8**,** 287-97.

Keller, A., Zhuang, H., Chi, Q., Vosshall, L. B. & Matsunami, H. 2007. Genetic variation in a human odorant receptor alters odour perception. *Nature,* 449**,** 468-472.

Killeen, G. F. & Chitnis, N. 2014. Potential causes and consequences of behavioural resilience and resistance in malaria vector populations: a mathematical modelling analysis. *Malaria Journal,* 13**,** 97.

Killeen, G. F., Kihonda, J., Lyimo, E., Oketch, F., Kotas, M., Mathenge, E., Schellenberg, J., Lengeler, C., Smith, T. & Drakeley, C. 2006. Quantifying behavioural interactions between humans and mosquitoes: Evaluating the protective efficacy of insecticidal nets against malaria transmission in rural Tanzania. *BMC Infectious Diseases,* 6**,** 161.

Killeen, G. F., McKenzie, F. E., Foy, B. D., Bøgh, C. & Beier, J. C. 2001. The availability of potential hosts as a determinant of feeding behaviours and malaria transmission by mosquito populations. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 95**,** 469 - 476.

Killeen, G. F. & Smith, T. A. 2007. Exploring the contributions of bed nets, cattle, insecticides and excitorepellency to malaria control: a deterministic model of mosquito host-seeking behaviour and mortality. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 101**,** 867-880.

Kim, M. S., Repp, A. & Smith, D. P. 1998. LUSH odorant-binding protein mediates chemosensory responses to alcohols in *Drosophila melanogaster*. *Genetics,* 150**,** 711-721.

Kimchi-Sarfaty, C., Oh, J. M., Kim, I. W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V. & Gottesman, M. M. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science,* 315**,** 525-528.

Knight, J. C. 2014. Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Medicine,* 6**,** 92.

Knols, B. G. J., van Loon, J. J. A., Cork, A., Robinson, R. D., Adam, W., Meijerink, J., Jong, R. D. & Takken, W. 1997. Behavioural and electrophysiological responses of the female malaria mosquito *Anopheles gambiae* (Diptera: Culicidae) to Limburger cheese volatiles. *Bulletin of Entomological Research,* 87**,** 151-159.

Koella, J. C., Rieu, L. & Paul, R. E. L. 2002. Stage-specific manipulation of a mosquito's host-seeking behavior by the malaria parasite *Plasmodium gallinaceum*. *Behavioral Ecology,* 13**,** 816-820.

Koella, J. C., Sørensen, F. L. & Anderson, R. A. 1998. The malaria parasite, *Plasmodium falciparum*, increases the frequency of multiple feeding of its mosquito vector, *Anopheles gambiae*. *Proceedings of the Royal Society B: Biological Sciences,* 265**,** 763-768.

Krishnan, B., Dryer, S. E. & Hardin, P. E. 1999. Circadian rhythms in olfactory responses of *Drosophila melanogaster*. *Nature,* 400**,** 375-8.

Kwon, H., Lu, T., Rützler, M. & Zwiebel, L. J. 2006. Olfactory responses in a gustatory organ of the malaria vector mosquito *Anopheles gambiae*. *Proceedings of the National Academy of Sciences, USA,* 103**,** 13526-13531.

Kwon, J. Y., Park, J. M., Gim, B. S., Han, S. J., Lee, J. & Kim, Y.-J. 1999. *Caenorhabditis elegans* Mediator complexes are required for developmental-specific transcriptional activation. *Proceedings of the National Academy of Sciences of the United States of America,* 96**,** 14990-14995.

Laarman, J. J. 1958. The host-seeking behaviour of Anopheline mosquitoes. *Trop. Geogr. Med.,* 10**,** 293-305.

Lacroix, R., Mukabana, W. R., Gouagna, L. C. & Koella, J. C. 2005. Malaria infection increases attractiveness of humans to mosquitoes. *PLoS Biol,* 3**,** e298.

Lanzaro, G. C., Toure, Y. T., Carnahan, J., Zheng, L., Dolo, G., Traore, S., Petrarca, V., Vernick, K. D. & Taylor, C. E. 1998. Complexities in the genetic

structure of *Anopheles gambiae* populations in west Africa as revealed by microsatellite DNA analysis. *PNAS,* 95**,** 14260 - 14265.

Larsson, M. C., Domingos, A. I., Jones, W. D., Chiappe, M. E., Amrein, H. & Vosshall, L. B. 2004. Or83b encodes a broadly expressed odorant receptor essential for *Drosophila* olfaction. *Neuron,* 43**,** 703-14.

Lawniczak, M. K. N., Emrich, S. J., Holloway, A. K., Regier, A. P., Olson, M., White, B., Redmond, S., Fulton, L., Appelbaum, E., Godfrey, J.*, et al.* 2010. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science,* 330**,** 512-514.

Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. 2014. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics,* 95**,** 5-23.

Lee, Y., Collier, T. C., Sanford, M. R., Marsden, C. D., Fofana, A., Cornel, A. J. & Lanzaro, G. C. 2013. Chromosome inversions, genomic differentiation and speciation in the African malaria mosquito *Anopheles gambiae*. *PLoS ONE,* 8**,** e57887.

Lefèvre, T., Gouagna, L. C., Dabire, K. R., Elguero, E., Fontenille, D., Renaud, F., Costantini, C. & Thomas, F. 2009. Beyond nature and nurture: phenotypic plasticity in blood-feeding behavior of *Anopheles gambiae s.s.* when humans are not readily accessible. *The American Journal of Tropical Medicine and Hygiene,* 81**,** 1023-9.

Lehane, M. J. 2005. The Biology of Blood-Sucking in Insects. Second Edition ed. New York: Cambridge University Press.

Lehmann, T. & Diabaté, A. 2008. The molecular forms of *Anopheles gambiae*: A phenotypic perspective. *Infection, Genetics and Evolution,* 8**,** 737-746.

Lemasson, J. J., Fontenille, D., Lochouarn, L., Dia, I., Simard, F., Ba, K., Diop, A., Diatta, M. & Molez, J. F. 1997. Comparison of behavior and vector efficiency of *Anopheles gambiae* and *An. arabiensis* (Diptera:Culicidae) in Barkedji, a Sahelian area of Senegal. *Journal of Medical Entomology,* 34**,** 396-403.

Lengeler, C. 2004. Insecticide-treated bednets and curtains for preventing malaria (Review). *Cochrane Database Syst Rev,* 2**,** CD000363.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M. W., Cavalli-Sforza, L. L.*, et al.* 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science,* 319**,** 1100-1104.

Lin, P., Vance, J. M., Pericak-Vance, M. A. & Martin, E. R. 2007. No gene is an island: the flip-flop phenomenon. *The American Journal of Human Genetics,* 80**,** 531-538.

Lindsay, S. W., Adiamah, J. H., Miller, J. E. & Armstrong, J. R. M. 1991. Pyrethroid-treated bednet effects on mosquitoes of the *Anopheles gambiae* complex in The Gambia. *Medical and Veterinary Entomology,* 5**,** 477 - 483.

Lindsay, S. W., Alonso, P. L. & Armstrong Schellenberg, J. R. M. 1993a. Impact of permethrin-impregnated bednets on malaria vectors. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 87**,** 45 - 52.

Lindsay, S. W., Alonso, P. L., Armstrong Schellenberg, J. R. M., Hemingway, J., Adiamah, J. H., Shenton, F. C., Jawara, M. & Greenwood, B. M. 1993b. A malaria control trial using insecticide-treated bed nets and targeted chemoprophylaxis in a rural area of The Gambia, West Africa: 7. Impact of permethrin-impregnated bed nets on malaria vectors. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 87, Supplement 2**,** 45-51.

Lindsay, S. W., Shenton, F. C., Snow, R. W. & Greenwood, B. M. 1989. Responses of *Anopheles gambiae* complex mosquitoes to the use of untreated bednets in The Gambia. *Med Vet Entomol,* 3**,** 253-62.

Lines, J. D., Myamba, J. & Curtis, C. F. 1987. Experimental hut trials of permethrin-impregnated mosquito nets and eave curtains against malaria vectors in Tanzania. *Med Vet Entomol,* 1**,** 37-51.

Lipkin, W. I. 2013. The changing face of pathogen discovery and surveillance. *Nat Rev Microbiol,* 11**,** 133-41.

Livak, K. J., Flood, S. J., Marmaro, J., Giusti, W. & Deetz, K. 1995. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl,* 4**,** 357-62.

Lu, T., Qiu, Y. T., Wang, G., Kwon, J. Y., Rutzler, M., Kwon, H., Pitts, R. J., van Loon, J. J. A., Takken, W., Carlson, J. R.*, et al.* 2007. Odor coding in the maxillary palp of the malaria vector mosquito *Anopheles gambiae*. *Current Biology,* 17**,** 1533-1544.

Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics,* 4**,** 981-994.

Lwetoijera, D. W., Harris, C., Kiware, S. S., Dongus, S., Devine, G. J., McCall, P. J. & Majambere, S. 2014. Increasing role of *Anopheles funestus* and *Anopheles arabiensis* in malaria transmission in the Kilombero Valley, Tanzania. *Malar J,* 13**,** 331.

Macpherson, J. M., Sella, G., Davis, J. C. & Petrov, D. A. 2007. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. *Genetics,* 177**,** 2083-2099.

Magesa, S. M., Wilkes, T. J. & Mnzava, A. E. 1991. Trial of pyrethroid impregnated bednets in an area of Tanzania holoendemic for malaria. Part 2. Effects on the malaria vector population. *Acta Tropica,* 49**,** 97 - 108.

Malaria Genomic Epidemiology Network 2014. Reappraisal of known malaria resistance loci in a large multicenter study. *Nature Genetics,* Advance online publication.

MalariaGEN 2013. Using next-generation sequencing to understand the diversity and dynamics of *Anopheles* populations.

Manske, M., Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen, G., O/'Brien, J., Djimde, A., Doumbo, O., Zongo, I.*, et al.* 2012. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature,* 487**,** 375-379.

Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. 2004. The effects of human population structure on large genetic association studies. *Nature Genetics,* 36**,** 512-7.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y., Chen, Z.*, et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-380.

Marinotti, O., Calvo, E., Nguyen, Q. K., Dissanayake, S., Ribeiro, J. M. & James, A. A. 2006. Genome-wide analysis of gene expression in adult *Anopheles gambiae*. *Insect Molecular Biology,* 15**,** 1 - 12.

Marinotti, O., Cerqueira, G. C., de Almeida, L. G. P., Ferro, M. I. T., Loreto, E. L. S., Zaha, A., Teixeira, S. M. R., Wespiser, A. R., Almeida e Silva, A., Schlindwein, A. D.*, et al.* 2013. The genome of *Anopheles darlingi*, the main neotropical malaria vector. *Nucleic Acids Research*.

Martinez-Torres, D., Chandre, F., Williamson, M. S., Darriet, F., Berge, J. B. & Devonshire, A. L. 1998. Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector *Anopheles gambiae s.s. Insect Mol Biol,* 7**,** 179 - 184.

Mathenge, E. M., Gimning, J. E., Kolczak, M., Ombok, M., Irungu, L. W. & Hawley, W. A. 2001. Effect of permethrin- impregnated nets on exiting behaviour, blood feeding success, and time of feeding of malaria mosquitoes (Diptera: Culicidae) in Western Kenya. *Journal of Medical Entomology,* 38**,** 531 - 536.

Matsuo, T., Sugaya, S., Yasukawa, J., Aigaki, T. & Fuyama, Y. 2007. Odorant-binding proteins OBP57d and OBP57e affect taste perception and host-plant preference in *Drosophila sechellia*. *PLoS Biology,* 5**,** e118.

Mbogo, C. N. M., Baya, N. M., Ofulla, A. V. O., Githure, J. I. & Snow, R. W. 1996. The impact of permethrin-impregnated bednets on malaria vectors on the Kenyan coast. *Medical and Veterinary Entomology,* 10**,** 251 - 259.

McBride, C. S. 2007. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proceedings of the National Academy of Sciences,* 104**,** 4996-5001.

McBride, C. S., Baier, F., Omondi, A. B., Spitzer, S. A., Lutomiah, J., Sang, R., Ignell, R. & Vosshall, L. B. 2014. Evolution of mosquito preference for humans linked to an odorant receptor. *Nature,* 515**,** 222-7.

McCormack, J. E., Maley, J. M., Hird, S. M., Derryberry, E. P., Graves, G. R. & Brumfield, R. T. 2012. Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular Phylogenetics and Evolution,* 62**,** 397-406.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* 20**,** 1297-303.

Meijerink, J., Braks, M. A. H. & Brack, A. A. 2000. Identification of olfactory stimulants for *Anopheles gambiae* from human sweat samples. *Journal of Chemical Ecology,* 26**,** 1367-1382.

Moiroux, N., Gomez, M. B., Pennetier, C., Elanga, E., Djenontin, A., Chandre, F., Djegbe, I., Guis, H. & Corbel, V. 2012. Changes in *Anopheles funestus* biting behavior following universal coverage of long-lasting insecticidal nets in Benin. *J Inf Dis,* 206**,** 1622 - 1629.

Muirhead-Thomson, R. C. 1960. The significance of irritability, behaviouristic avoidance and allied phenomena in malaria eradication. *Bulletin of the World Health Organization,* 22**,** 721-734.

Mukabana, W. R., Takken, W., Killeen, G. F. & Knols, B. G. J. 2007. Clinical malaria reduces human attractiveness to mosquitoes. *Proc. Neth. Entomol. Soc.,* 18**,** 125-129.

Mukwaya, L. G. 1977. Genetic control of feeding preferences in the mosquitoes *Aedes (Stegomyia) simpsoni* and *aegypti*. *Physiological Entomology,* 2**,** 133-145.

Murase, S. & Schuman, E. M. 1999. The role of cell adhesion molecules in synaptic plasticity and memory. *Current Opinion in Cell Biology,* 11**,** 549-553.

Murray, C. J. L., Rosenfeld, L. C., Lim, S. S., Andrews, K. G., Foreman, K. J., Haring, D., Fullman, N., Naghavi, M., Lozano, R. & Lopez, A. D. 2012. Global malaria mortality between 1980 and 2010: a systematic analysis. *The Lancet,* 379**,** 413-431.

Mutuku, F. M., King, C. H., Mungai, P., Mbogo, C. N. M., Mwangangi, J., Muchiri, E. M., Walker, E. D. & Kitron, U. 2011. Impact of insecticide-treated bed nets on

malaria transmission indices on the south coast of Kenya. *Malaria Journal,* 10**,** 356.

Nakamura, M., Baldwin, D., Hannaford, S., Palka, J. & Montell, C. 2002. Defective proboscis extension response (DPR), a member of the Ig superfamily required for the gustatory response to salt. *J Neurosci,* 22**,** 3463-72.

Ndiath, M. O., Mazenot, C., Sokhna, C. & Trape, J. 2014. How the malaria vector *Anopheles gambiae* adapts to the use of insecticide-treated nets by African populations. *PLoS ONE,* 9**,** e97700.

Neafsey, D., Christophides, G., Collins, F., Emrich, S., Fontaine, M., Gelbart, W., Hahn, M., Howell, P., Kafatos, F., Lawson, D.*, et al.* 2013. The evolution of the Anopheles 16 genomes project. *G3,* 3**,** 1191 - 1194.

Neafsey, D. E., Lawniczak, M. K. N., Park, D. J., Redmond, S. N., Coulibaly, M. B., Traoré, S. F., Sagnon, N., Costantini, C., Johnson, C., Wiegand, R. C.*, et al.* 2010. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science,* 330**,** 514-517.

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arcà, B., Arensburger, P., Artemov, G.*, et al.* 2015. Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. *Science,* 347.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* 70**,** 3321-3323.

Ng'habi, K. R. N., Knols, B. G., Lee, Y., Ferguson, H. M. & Lanzaro, G. 2011. Population genetic structure of *Anopheles arabiensis* and *Anopheles gambiae* in a malaria endemic region of southern Tanzania. *Malaria Journal,* 10**,** 289.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R.*, et al.* 2008. Genes mirror geography within Europe. *Nature,* 456**,** 98-U5.

Okoye, P. N., Brooke, B. D., Koekemoer, L. L., Hunt, R. H. & Coetzee, M. 2008. Characterisation of DDT, pyrethroid and carbamate resistance in *Anopheles funestus* from Obuasi, Ghana. *Trans R Soc Trop Med Hyg,* 102 591-598.

Oyewole, I. O. & Awolola, T. S. 2006. Impact of urbanisation on bionomics and distribution of malaria vectors in Lagos, southwestern Nigeria. *J Vector Borne Dis,* 43**,** 173-8.

Padonou, G., Gbedjissi, G., Yadouleton, A., Azondekon, R., Razack, O., Oussou, O., Gnanguenon, V., Rock, A., Sezonlin, M. & Akogbeto, M. 2012. Decreased proportions of indoor feeding and endophily in *Anopheles gambiae s.l.* populations following the indoor residual spraying and insecticide-treated net interventions in Benin (West Africa). *Parasites & Vectors,* 5**,** 262.

Paradis, E. 2010. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics,* 26**,** 419-20.

Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W. & Drineas, P. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics,* 3**,** e160.

Pates, H. & Curtis, C. 2005. Mosquito behavior and vector control. *Annual Review of Entomology,* 50**,** 53 - 70.

Pates, H. V., Takken, W., Stuke, K. & Curtis, C. F. 2001. Differential behaviour of *Anopheles gambiae sensu stricto* (Diptera: Culicidae) to human and cow odours in the laboratory. *Bull Entomol Res,* 91**,** 289-96.

Paunio, T., Reima, I. & Syvänen, A. C. 1996. Preimplantation diagnosis by whole-genome amplification, PCR amplification, and solid-phase minisequencing of blastomere DNA. *Clinical Chemistry,* 42**,** 1382-90.

Pavlidis, P., Jensen, J. D., Stephan, W. & Stamatakis, A. 2012. A critical assessment of story-telling: Gene Ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*.

Pelosi, P. 1998. Odorant-binding proteins: structural aspects. *Annals of the New York Academy of Sciences,* 855**,** 281-293.

Pelosi, P., Zhou, J. J., Ban, L. P. & Calvello, M. 2006. Soluble proteins in insect chemical communication. *Cellular and Molecular Life Sciences,* 63**,** 1658-76.

Pevsner, J. & Snyder, S. H. 1990. Odorant-binding protein: odorant transport function in the vertebrate nasal epithelium. *Chemical Senses,* 15**,** 217-222.

Phillips-Howard, P. A., Nahlen, B. L., Kolczak, M. S., Hightower, A. W., Ter Kuile, F. O., Alaii, J. A., Gimnig, J. E., Arudo, J., Vulule, J. M., Odhacha, A.*, et al.* 2003. The efficacy of permethrin-treated bed nets on child mortality and morbidity in western Kenya. *The American Journal of Tropical Medicine and Hygiene,* 68**,** 23-29.

Pipes, G. C., Lin, Q., Riley, S. E. & Goodman, C. S. 2001. The Beat generation: a multigene family encoding IgSF proteins related to the Beat axon guidance molecule in *Drosophila*. *Development,* 128**,** 4545-52.

Pitts, R. J. & Zwiebel, L. J. 2006. Antennal sensilla of two female anopheline sibling species with differing host ranges. *Malar J,* 5**,** 26.

Pool, J. E., Hellmann, I., Jensen, J. D. & Nielsen, R. 2010. Population genetic inference from genomic sequence variation. *Genome Res,* 20**,** 291-300.

Pophof, B. 2004. Pheromone-binding proteins contribute to the activation of olfactory receptor neurons in the silkmoths *Antheraea polyphemus* and *Bombyx mori*. *Chemical Senses,* 29**,** 117-125.

Pritchard, J. K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945 - 959.

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. *Genetics,* 160**,** 1179-1189.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J.*, et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics,* 81**,** 559-575.

Quiňones, M. L., Lines, J. D., Thomson, M. C., Jawara, M., Morris, J. & Greenwood, B. M. 1997. *Anopheles gambiae* gonotrophic cycle duration, biting and exiting behaviour unaffected by permethrin-impregnated bednets in The Gambia. *Medical and Veterinary Entomology,* 11**,** 71-78.

R Development Core Team 2010. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ranson, H., N'guessan, R., Lines, J., Moiroux, N., Nkuni, Z. & Corbel, V. 2011. Pyrethroid resistance in African anopheline mosquitoes: what are the implications for malaria control? *Trends Parasitol,* 27**,** 91 - 98.

Reddy, M. R., Overgaard, H. J., Abaga, S., Reddy, V. P., Caccone, A., Kiszewski, A. E. & Slotman, M. A. 2011. Outdoor host seeking behaviour of *Anopheles gambiae* mosquitoes following initiation of malaria vector control on Bioko Island, Equatorial Guinea. *Malaria Journal,* 10**,** 1475-2875.

Reidenbach, K. R., Neafsey, D. E., Costantini, C., Sagnon, N., Simard, F., Ragland, G. J., Egan, S. P., Feder, J. L., Muskavitch, M. A. T. & Besansky, N. J. 2012. Patterns of genomic differentiation between ecologically differentiated M and S forms of *Anopheles gambiae* in west and central Africa. *Genome Biology and Evolution,* 4**,** 1202-1212.

Renger, J. J., Yao, W. D., Sokolowski, M. B. & Wu, C. F. 1999. Neuronal polymorphism among natural alleles of a cGMP-dependent kinase gene, foraging, in *Drosophila*. *Journal of Neuroscience,* 19**,** Rc28.

Riehle, M. M., Guelbeogo, W. M., Gneme, A., Eiglmeier, K., Holm, I., Bischoff, E., Garnier, T., Snyder, G. M., Li, X., Markianos, K.*, et al.* 2011. A cryptic subgroup of *Anopheles gambiae* is highly susceptible to human malaria parasites. *Science,* 331**,** 596-8.

Rinker, D., Zhou, X., Pitts, R., Consortium, T. A., Rokas, A. & Zwiebel, L. I. J. 2013. Antennal transcriptome profiles of anopheline mosquitoes reveal human host olfactory specialization in *Anopheles gambiae*. *BMC Genomics,* 14**,** 749.

Robertson, H. M., C.G., W. & Carlson, J. R. 2003. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. ,* 100**,** 14537–14542.

Rollmann, S. M., Wang, P., Date, P., West, S. A., Mackay, T. F. C. & Anholt, R. R. H. 2010. Odorant receptor polymorphisms and natural variation in olfactory behavior in *Drosophila melanogaster*. *Genetics,* 186**,** 687-697.

Rongo, C., Whitfield, C. W., Rodal, A., Kim, S. K. & Kaplan, J. M. 1998. LIN-10 is a shared component of the polarized protein localization pathways in neurons and epithelia. *Cell,* 94**,** 751-9.

Rund, S. S., Bonar, N. A., Champion, M. M., Ghazi, J. P., Houk, C. M., Leming, M. T., Syed, Z. & Duffield, G. E. 2013. Daily rhythms in antennal protein and olfactory sensitivity in the malaria mosquito *Anopheles gambiae*. *Sci Rep,* 3.

Rund, S. S. C., Hou, T. Y., Ward, S. M., Collins, F. H. & Duffield, G. E. 2011. Genome-wide profiling of diel and circadian gene expression in the malaria vector *Anopheles gambiae*. *Proceedings of the National Academy of Sciences*.

Russell, T. L., Govella, N. J., Azizi, S., Drakeley, C. J., Kachur, S. P. & Killeen, G. F. 2011. Increased proportions of outdoor feeding among residual malaria vector populations following increased use of insecticide-treated nets in rural Tanzania. *Malaria Journal,* 10**,** 80.

Samarawickrema, W. A., Parkinson, A. D., Kere, N. & Alo, O. G. 1992. Seasonal abundance and biting behaviour of *Anopheles punctulatus* and *An.koliensis* in Malaita Province, Solomon Islands, and a trial of permethrin impregnated bednets against malaria transmission. *Medical and Veterinary Entomology,* 6**,** 371-378.

Sato, K., Pellegrino, M., Nakagawa, T., Nakagawa, T., Vosshall, L. B. & Touhara, K. 2008. Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature,* 452**,** 1002-6.

Scott, J. A., Brogdon, W. G. & Collins, F. H. 1993. Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. *American Journal of Tropical Medicine and Hygiene,* 49**,** 520-529.

Scott, K., Brady Jr, R., Cravchik, A. l., Morozov, P., Rzhetsky, A., Zuker, C. & Axel, R. 2001. A chemosensory gene family encoding candidate gustatory and olfactory receptors in *Drosophila*. *Cell,* 104**,** 661-673.

Selkoe, K. A. & Toonen, R. J. 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters,* 9**,** 615-629.

Sharakhova, M., Hammond, M., Lobo, N., Krzywinski, J., Unger, M., Hillenmeyer, M., Bruggner, R., Birney, E. & Collins, F. H. 2007. Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biology,* 8**,** R5.

Shin, J., Blay, S., McNeney, B. & Graham, J. 2006. LDheatmap: An R function for graphical display of pairwise linkage disequilibria between Single Nucleotide Polymorphisms. *Journal of Statistical Software,* 16.

Sinka, M., Bangs, M., Manguin, S., Coetzee, M., Mbogo, C. N. M., Hemingway, J., Patil, A., Temperley, W., Gething, P., Kabaria, C.*, et al.* 2010. The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis. *Parasites & Vectors,* 3**,** 117.

Slatkin, M. 1985. Gene flow in natural populations. *Annual Review of Ecology and Systematics,* 16**,** 393-430.

Slatkin, M. & Barton, N. H. 1989. A comparison of three indirect methods for estimating average levels of gene flow. *Evolution,* 43**,** 1349-1368.

Smadja, C. M., Canbäck, B., Vitalis, R., Gautier, M., Ferrari, J., Zhou, J.-J. & Butlin, R. K. 2012. Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. *Evolution,* 66**,** 2723-2738.

Smallegange, R., Geier, M. & Takken, W. 2002. Behavioural responses of *Anopheles gambiae* to ammonia, lactic acid and a fatty acid in a y-tube olfactometer. *Proceedings of the section Experimental and Applied Entomology of the Netherlands Entomological Society (NEV)* 13**,** 147-152.

Smith, D. R., Quinlan, A. R., Peckham, H. E., Makowsky, K., Tao, W., Woolf, B., Shen, L., Donahue, W. F., Tusneem, N., Stromberg, M. P.*, et al.* 2008. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research,* 18**,** 1638-1642.

Sparks, J. T., Vinyard, B. T. & Dickens, J. C. 2013. Gustatory receptor expression in the labella and tarsi of *Aedes aegypti*. *Insect Biochemistry and Molecular Biology,* 43**,** 1161-1171.

Spencer, C. C. A., Su, Z., Donnelly, P. & Marchini, J. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet,* 5**,** e1000477.

Stranger, B. E., Stahl, E. A. & Raj, T. 2011. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics,* 187**,** 367-383.

Sun, H., Trabue, S. L., Scoggin, K., Jackson, W. A., Pan, Y., Zhao, Y., Malkina, I. L., Koziel, J. A. & Mitloehner, F. M. 2008. Alcohol, volatile fatty acid, phenol, and methane emissions from dairy cows and fresh manure. *J Environ Qual,* 37**,** 615-22.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123**,** 585-95.

Takken, W. & Knols, B. G. 1999. Odor-mediated behavior of Afrotropical malaria mosquitoes. *Annu Rev Entomol,* 44**,** 131-57.

Takken, W. & Lindsay, S. W. 2003. Factors affecting the vectorial competence of *Anopheles gambiae*: a question of scale. *In:* Takken, D. I. W. & Salt, T. W. (eds.) *Ecological aspects for application of genetically modified mosquitoes.* Dordrecht: Kluwer Academic Publishers.

Takken, W., Stuke, K. & Klowden, M. J. 2002. Biological differences in reproductive strategy between the mosquito sibling species *Anopheles gambiae sensu stricto* and *An. quadriannulatus*. *Entomol. Exp. Appl.,* 103**,** 83-89.

Takken, W., van Loon, J. & Adam, W. 2001. Inhibition of host-seeking response and olfactory responsiveness in Anopheles gambiae following blood feeding. *J Insect Physiol,* 47**,** 303 - 310.

Takken, W. & Verhulst, N. O. 2013. Host preferences of blood-feeding mosquitoes. *Annual Review of Entomology,* 58**,** 433-453.

Taylor, B. 1975. Changes in the feeding behaviour of a malaria vector, *Anopheles farauti* Lav., following use of DDT as a residual spray in houses in the British Solomon Islands Protectorate. *Transactions of the Royal Entomological Society of London,* 127**,** 277-292.

The Global Fund. 2013. Looking back: a decade of unimaginable Pprogress against AIDS, tuberculosis and malaria - looking ahead: a decade of Incredible opportunity.Available: http://www.theglobalfund.org/documents/press/NewsRelease_130923_NewsRelease_en/ [Accessed 04/02/2015].

Tirados, I., Costantini, C., Gibson, G. & Torr, S. J. 2006. Blood-feeding behaviour of the malarial mosquito *Anopheles arabiensis*: implications for vector control. *Medical and Veterinary Entomology,* 20**,** 425-437.

Torr, S. J., Mangwiro, T. N. C. & Hall, D. R. 2006. The effects of host physiology on the attraction of tsetse (Diptera: Glossinidae) and Stomoxys (Diptera: Muscidae) to cattle. *Bulletin of Entomological Research,* 96**,** 71-84.

Tripet, F., Toure, Y. T., Taylor, C. E., Norris, D. E., Dolo, G. & Lanzaro, G. C. 2001. DNA analysis of transferred sperm reveals significant levels of gene flow between molecular forms of *Anopheles gambiae*. *Mol Ecol,* 10**,** 1725 - 1732.

Ulloa, A., I., A.-J. J., Rodriguez, M. H., Fernandez-Salas, I. & Gonzalez-Ceron, L. 2004. Innate host selection in *Anopheles vestipennis* from southern Mexico. *Journal Of The American Mosquito Control Association* 20**,** 337-341.

van den Broek, I. V. F. & den Otter, C. J. 1999. Olfactory sensitivities of mosquitoes with different host preferences (*Anopheles gambiae s.s., An. arabiensis, An. quadriannulatus, An. m. atroparvus*) to synthetic host odours. *Journal of Insect Physiology,* 45**,** 1001-1010.

Vannini, L., Augustine Dunn, W., Reed, T. W. & Willis, J. H. 2014. Changes in transcript abundance for cuticular proteins and other genes three hours after a blood meal in *Anopheles gambiae*. *Insect Biochem Mol Biol,* 44**,** 33-43.

Verhulst, N. O., Beijleveld, H., Knols, B. G. J., Takken, W., Schraa, G., Bouwmeester, H. J. & Smallegange, R. C. 2009. Cultured skin microbiota attracts malaria mosquitoes. *Malaria Journal,* 8**,** 302-302.

Vieira, F. G. & Rozas, J. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene gamilies across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biology and Evolution*.

Vosshall, L. B. & Hansson, B. S. 2011. A Unified nomenclature system for the insect olfactory coreceptor. *Chemical Senses,* 36**,** 497-498.

Wang-Sattler, R., Blandin, S., Ning, Y., Blass, C., Dolo, G., Touré, Y. T., Torre, A., Lanzaro, G. C., Steinmetz, L. M., Kafatos, F. C.*, et al.* 2007. Mosaic genome architecture of the *Anopheles gambiae* species complex. *PLoS ONE,* 2**,** e1249.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J.*, et al.* 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science,* 280**,** 1077-82.

Wang, G., Carey, A. F., Carlson, J. R. & Zwiebel, L. J. 2010a. Molecular basis of odor coding in the malaria vector mosquito *Anopheles gambiae*. *Proceedings of the National Academy of Sciences, USA,* 7**,** 4418–4423.

Wang, P., Lyman, R. F., Mackay, T. F. C. & Anholt, R. R. H. 2010b. Natural variation in odorant recognition among odorant-binding *proteins in Drosophila melanogaster*. *Genetics,* 184**,** 759-767.

Wang, P., Lyman, R. F., Shabalina, S. A., Mackay, T. F. & Anholt, R. R. 2007. Association of polymorphisms in odorant-binding protein genes with variation in olfactory response to benzaldehyde in *Drosophila*. *Genetics,* 177**,** 1655-65.

Weetman, D., Steen, K., Rippon, E. J., Mawejje, H. D., Donnelly, M. J. & Wilding, C. S. 2014. Contemporary gene flow between wild *An. gambiae s.s.* and *An. arabiensis*. *Parasit Vectors,* 7**,** 345.

Weetman, D., Wilding, C. S., Steen, K., Morgan, J. C., Simard, F. & Donnelly, M. J. 2010. Association mapping of insecticide resistance in wild *Anopheles gambiae* populations: major variants identified in a low-linkage disequilbrium genome. *PLoS ONE,* 5**,** e13140.

Weir, B. S. & Cockerham, C. C. 1984. Estimating F-statistics for the analysis of population structure. *Evolution,* 38**,** 1358-1370.

Wetzel, C. H., Behrendt, H. J., Gisselmann, G., Stortkuhl, K. F., Hovemann, B. & Hatt, H. 2001. Functional expression and characterization of a *Drosophila* odorant receptor in a heterologous cell system. *Proceedings of the National Academy of Sciences USA,* 98**,** 9377-80.

White, B. J., Collins, F. H. & Besansky, N. J. 2011. Evolution of *Anopheles gambiae* in Relation to Humans and Malaria. *Annual Review of Ecology, Evolution, and Systematics,* 42**,** 111-132.

Whitlock, M. C. & McCauley, D. E. 1999. Indirect measures of gene flow and migration: F_{ST} not equal to 1/(4*Nm* +1). *Heredity,* 82**,** 117 - 125.

WHO. 2006. Indoor residual spraying, use of indoor residual spraying for scaling up global malaria control and elimination.

WHO 2011. World Malaria Report 2011 Geneva, Switzerland.

WHO. 2013a. Indoor residual spraying: an operational manual for Indoor residual spraying (IRS) for malaria transmission control and elimination. Available: http://apps.who.int/iris/bitstream/10665/80126/1/9789241505123_eng.pdf?ua=1 [Accessed 16/02/15].

WHO 2013b. World Malaria Report 2013. Geneva, Switzerland.

Wilke, A. B., Gomes, A. C., Natal, D. & Marrelli, M. T. 2009. Control of vector populations using genetically modified mosquitoes. *Revista de Saude Publica,* 43 869-874.

Wilkins, E. E., Howell, P. I. & Benedict, M. Q. 2006. IMP PCR primers detect single nucleotide polymorphisms for *Anopheles gambiae* species identification, Mopti and Savanna rDNA types, and resistance to dieldrin in *Anopheles arabiensis. Malaria Journal,* 5**,** 125.

Windbichler, N., Galizzi, R., Burt, A. & Crisanti, A. 2012. Engineering mosquito population for vector control. *Malaria Journal,* 11**,** O44.

Wright, S. 1951. The genetical structure of populations. *Eugenics,* 15**,** 323-354.

Xu, P. X., Zwiebel, L. J. & Smith, D. P. 2003. Identification of a distinct family of genes encoding atypical odorant-binding proteins in the malaria vector mosquito, *Anopheles gambiae*. *Insect Molecular Biology,* 12**,** 549 - 560.

Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'Connell, J. R., Mangino, M.*, et al.* 2011. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet,* 19**,** 807-12.

Youngson, N. A. & Whitelaw, E. 2008. Transgenerational epigenetic effects. *Annual Review of Genomics and Human Genetics,* 9**,** 233-257.

Yukich, J., Tediosi, F. & Lengeler, C. 2007. Comparative cost-effectiveness of ITNs or IRS in sub-Saharan Africa. *Malaria Journal,* 8.

Zaykin, D. V. & Shibata, K. 2008. Genetic flip-flop without an accompanying change in linkage disequilibrium. *Am J Hum Genet,* 82**,** 794-6; author reply 796-7.

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I. W., Abecasis, G. R., Almgren, P., Andersen, G.*, et al.* 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet,* 40**,** 638-645.

Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W. & Arnheim, N. 1992. Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Sciences, USA,* 89**,** 5847-5851.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C. & Weir, B. S. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics,* 28**,** 3326-3328.

Zhou, J. J., He, X. L., Pickett, J. A. & Field, L. M. 2008. Identification of odorant-binding proteins of the yellow fever mosquito *Aedes aegypti:* genome annotation and comparative analyses. *Insect Molecular Biology,* 17**,** 147-63.

Zhou, J. J., Zhang, G. A., Huang, W., Birkett, M. A., Field, L. M., Pickett, J. A. & Pelosi, P. 2004. Revisiting the odorant-binding protein LUSH of *Drosophila melanogaster:* evidence for odour recognition and discrimination. *FEBS Letters,* 558**,** 23-6.

Ziv, E. & Burchard, E. G. 2003. Human population structure and genetic association studies. *Pharmacogenomics,* 4**,** 431-41.

Zwiebel, L. J. & Takken, W. 2004. Olfactory regulation of mosquito-host interactions. *Insect Biochemistry and Molecular Biology,* 34**,** 645 - 652.