Genomic surveillance of the African malaria mosquito, *Anopheles gambiae*

Thesis submitted in accordance with the requirements of the Liverpool School of Tropical Medicine for the degree of Doctor in Philosophy by

Sanjay Curtis Nagi[1], **BSc (Hons), MSc, MRes**

Department of Vector Biology

Liverpool School of Tropical Medicine

January 2023

[1] sanjay.c.nagi@gmail.com

Abstract

In this thesis, I discuss population genomics of the major malaria vector, *Anopheles gambiae*, with a focus on the evolution of insecticide resistance. In particular, I present a series of computational tools, applicable to both genomic and transcriptomic technologies, which I hope will enhance the research community's ability to perform effective genomic surveillance.

In the **first chapter**, I provide an introduction to malaria as a disease and its vector, the African malaria mosquito. I describe how we attempt to control the mosquito, the problem of insecticide resistance and how genomic surveillance may help to mitigate this problem.

In the **second chapter**, I present *RNA-Seq-Pop*, a reproducible computational snakemake pipeline applicable to Illumina RNA-Sequencing data of any organism. The workflow performs typical transcriptomic analyses, but also calls SNPs and can extract population genomic signals, relating to genetic diversity, selection, ancestry and karyotypes, extending the utility of RNA-Seq and bridging the gap between transcriptomic and genomic studies.

In the **third chapter,** I present a python package, AgamPrimer and an accompanying Google Colaboratory notebook, which allows users to design primers and hybridisation probes for *An. gambiae* whilst considering genetic (SNP) variation in primer/probe binding sites.

In the **fourth chapter**, I describe an in-depth population genomic analysis of a locus under intense selection in *An. gambiae*, which displays evidence of parallel evolution in *An. arabiensis*, *An. coluzzii,* and *Culex pipiens.* I also present a snakemake workflow, locusPocus, which implements many of the analyses contained therein, including phylogenetic analysis, multi-allele phasing, indel calling and more.

In the **fifth chapter,** I describe a population genomic analysis of 485 whole-genome sequenced *An. gambiae* mosquitoes collected from a small region of Obuasi, Central Ghana. I identify ultra-fine-scale population structure and summarise the genomic basis of

insecticide resistance in Obuasi. I present a snakemake workflow, Probe, which automates a subset of the analyses performed.

In the **final chapter**, I discuss the impacts of this work, and what work could be done to take this research forward. I highlight some areas in mechanisms of resistance research that may be interesting to explore in the future.

In the **Appendix,** I briefly describe a new software tool, AnoExpress, which summarises gene expression across published RNA-Sequencing experiments into insecticide resistance in *Anopheles* mosquitoes.

Acknowledgements

Firstly, I would like to thank every individual that was involved in the generation of data analysed in this thesis. Sample collectors, field scientists, lab scientists and everyone in between. All of those involved in every aspect of the Anopheles 1000 genomes project. I have been incredibly fortunate in my academic career so far to have fantastic datasets to work with. Similarly, I thank those who have contributed their valuable time and efforts into developing the analytical tools that this thesis is built upon. Particularly scikit-allel, which in my eyes, will always be the greatest python package of all time.

I would particularly like to thank my primary supervisor, Professor Martin J Donnelly, whom I feel incredibly lucky to have had as a supervisor. I also thank Dave Weetman, an unofficial supervisor of this thesis. The belief I feel from both of you has made this PhD experience an enjoyable one. I thank Eric Lucas, whos expertise, humour, and good nature have been invaluable over the past four years and I am extremely grateful to Alistair Miles, whose mentorship, confidence and encouragement has kept me motivated.

I am extremely fortunate to have the friendship of Daniel McDermott. The image of completing this doctorate without Dan is a very lonely one and a much less enjoyable one. I am grateful for Hilary Ranson, Pat Pignatelli, and my good friend Rhiannon A. E. Logan, without them, I may never have joined LSTM after my MSc. Rhi, your presence has been sorely missed over the last couple of years.

I am very grateful for all of the Leeds people, and to all of my family, I thank you for your constant, unconditional love and support over the past 29 years. I have been so blessed to have such a wonderful family and friends in this life. I dedicate this work to my grandfather, whom I lost in the final few months of completing this thesis - Philip Chapman.

Declaration

I hereby declare that the contents of this thesis are original and have not been submitted in whole or part for consideration to any other degree of qualification, in this, or any other university. This thesis is my own work, and contains nothing which is the outcome of work performed in collaboration with others, except where explicitly stated. This thesis does not exceed the maximum permitted word length; it contains fewer than 100,000 words including appendices and footnotes.

Sanjay Curtis Nagi

January 2023

Contents

Chapter 1 - Introduction

Chapter 2 - RNA-Seq-Pop

Chapter 3 - AgamPrimer

Chapter 4 - Parallel evolution and adaptive introgression in mosquito vectors – a story of insecticide resistance

Chapter 5 - A population genomic analysis of *Anopheles gambiae* from Obuasi, Ghana

Chapter 6 - Discussion

Appendix A - AnoExpress

1

Introduction

1.1 Malaria

Malaria is a vector-borne disease caused by protozoan parasites of the genus *Plasmodium*. As well as humans, *Plasmodium* parasites infect a diverse array of taxa, including other mammals, birds and even reptiles. Six *Plasmodium* species are known to cause human malarial morbidity (Sato, 2021), however, the majority of the global disease burden is caused by *Plasmodium falciparum*, primarily in sub-Saharan Africa (World Health Organization, 2022).

It is estimated that since 2000, malaria control programmes have managed to halve the deaths from malaria (Bhatt *et al.*, 2015) and averted 663 million clinical malaria cases between the years 2000 and 2015. However there is evidence that the numbers of malaria cases and deaths have plateaued in recent years and may have even increased in some high-burden countries (World Health Organization, 2022).

1.2 Vectors of malaria

The life cycle of the human *Plasmodium* parasite is intimately coupled with that of its definitive host, the *Anopheline* mosquito - in humans, only *Anopheles* mosquitoes have been implicated as vectors of the parasite (Cohuet *et al.*, 2010). Over 430 species of *Anopheles* have been classified, however, only around 50 of these species are known vectors of human malaria (Pages *et al.*, 2007), and only a handful are thought to contribute substantially to the global disease burden. The most important vectors of human malaria tend to be the species which have adapted to be highly anthropophagic - they preferentially bite humans over other organisms (Macdonald, 1952; Besansky *et al.*, 2004;

Cohuet *et al.*, 2010). These species tend to also be anthropophilic - they live and reproduce in close proximity to human habitats. In Africa, these mosquito species are *Anopheles gambiae s.l* (specifically *An. gambiae, An. coluzzii and An. arabiensis)*, and *An. funestus* (World Health Organization, 2022).

1.2.1 The *Anopheles gambiae* species complex

The species *Anopheles gambiae* was first described by Giles in 1902, named after the location of its discovery, the Gambia Valley, in The Gambia, West Africa (Giles, 1902). It was in the 1960s and 70s that this species was recognised to be a complex of multiple closely-related sibling species, through laboratory crosses exploiting incompatibilities in hybridisation (Davidson, 1964), and concurrently, with cytogenetic techniques (Coluzzi *et al.*, 1967) which exploited differences in fixed chromosomal inversions between species. After extensive research and correspondence between the entomologists Hugh Paterson, George Davidson, Mario Coluzzi and other colleagues (Miles, 2021), these species were formally defined (Mattingly, 1977). Despite its etymology, members of the *An. gambiae* species complex are found widely across sub-Saharan Africa (Figure 1)

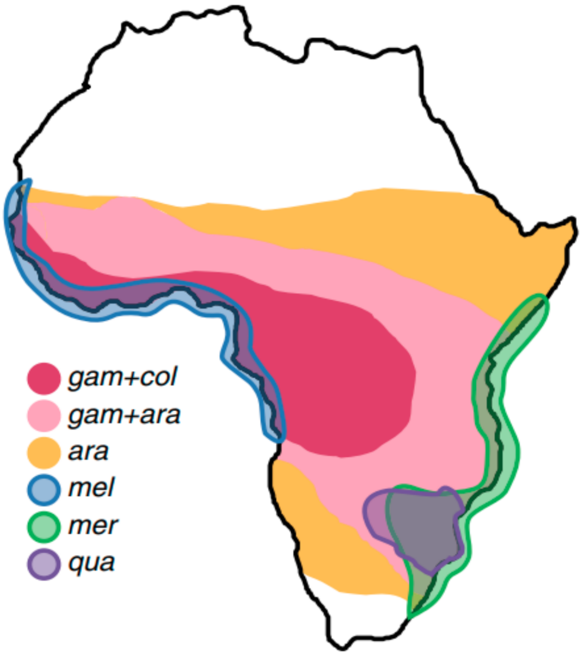


Figure 1. The distributions of primary members of the *Anopheles gambiae* species complex in sub-Saharan Africa. Adapted from (Fontaine *et al.*, 2015).

It was later recognised that within *An. gambiae* there existed five "chromosomal forms" which were morphologically identical, but mated assortatively in areas of sympatry and exhibited markedly different frequencies of chromosomal and molecular markers (Favia *et al.*, 1997; della Torre *et al.*, 2001). These were referred to as Mopti, Savannah, Bamako, Forest and Bissau (Coluzzi *et al.*, 1985). Mopti and Savannah were later designated as M and S forms of *An. gambiae*, and after it was demonstrated that they exhibited marked differentiation across the genome (Lawniczak *et al.*, 2010), these forms were later elevated to species status, with the S-form retaining the name of *An. gambiae* and the M-form becoming *Anopheles coluzzii* (Coetzee et al., 2013). These species are known to occasionally hybridise in the wild, permitting the exchange of haplotypes across species boundaries (Fontaine *et al.*, 2015). This is particularly relevant to public health, as this process of gene flow can be adaptive, with beneficial alleles that contribute to insecticide resistance crossing from one species to another (Grau-Bové, Tomlinson, *et al.*, 2020; Grau-Bové, Lucas, *et al.*, 2020).

After further discoveries of cryptic species in the complex, the *Anopheles gambiae* complex is now considered to have eight recognised sibling species (Barrón *et al.*, 2019; Tennessen *et al.*, 2021), although the true number is likely to be higher. Within the complex, *An. gambiae*, *An. coluzzii*, and *An. arabiensis* are the major malaria vectors, with other species having much more limited geographical ranges, ecological niches and anthropophagic behaviour.

1.3 Controlling the insect vector

In 1998, Roll Back Malaria was established, a concerted, global partnership to substantially reduce the malaria disease burden (Nabarro *et al.*, 1998). The initial aim was to halve the number of deaths from malaria in ten years, which was followed by the Millennium Development Goals, setting further ambitious targets to achieve by the year 2015. These campaigns heavily emphasised the mass delivery of insecticidal nets to control the insect vector. Long-lasting insecticidal bed nets (LLINs) are now the cornerstone of malaria control - since the year 2004, over 2 billion insecticide-treated nets have been delivered worldwide (Roll Back Malaria, 2020). They provide both a physical barrier to blood-feeding and a community-wide killing effect, thereby reducing the size of the vector population (Unwin *et al.*, 2022; Sherrard-Smith *et al.*, 2022).

These initiatives were considered broadly successful, with modelling suggesting a halving in the number of deaths from malaria between 2000 and 2015 (Bhatt *et al.*, 2015). This study also demonstrated that the most impactful strategies to reduce malarial disease had targeted the insect vector rather than the parasite itself, with LLINs the primary driver of the reduction in malaria (Bhatt *et al.*, 2015). Until 2017, pyrethroid-only LLINs were the only type of net approved for use by the WHO, however, pyrethroid-synergist nets, namely Piperonyl butoxide (PBO), as well as nets with dual active ingredients (pyrethroids and chlorphenapyr) have now come to market. Figure 2 shows the increase in LLIN distribution from the years 2004 onwards and the current transition towards novel net types. PBO and Dual-AI nets now make up more than half of the LLIN market globally.

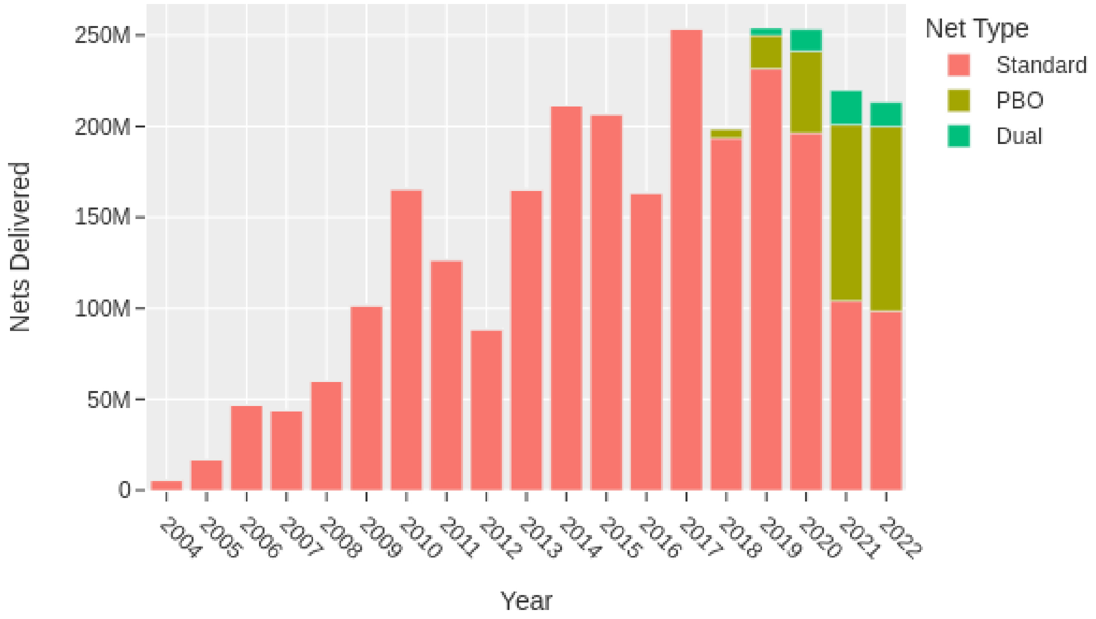


Figure 2. The number of distributed LLINs globally by net type, 2004-2022. Data from (The Alliance for Malaria Prevention, 2022). The year 2022 only includes data from Q1 to Q3. Standard nets refer to pyrethroid-only nets, PBO refer to combination Pyrethroid-PBO nets, and Dual nets refers to combination pyrethroid-chlorphenapyr nets.

The second widely-used vector control technique is indoor residual spraying (IRS). This involves the application of an insecticide to the inside walls of houses in order to kill the female mosquito when it is resting. Typically more expensive and logistically challenging than LLIN distribution, IRS use tends to be concentrated in areas of high transmission (Tangena *et al.*, 2020). Unlike in LLINs, the WHO has now approved five different insecticide classes for use in IRS, with more in the pipeline (WHO, 2022).

Despite their success, both LLINs and IRS are limited in the sense that they typically target endophilic mosquitoes, with LLINs only targeting night-time biters. Considering both the persistence of malaria transmission even in areas of high LLIN and IRS coverage (Killeen, 2014) and escalating insecticide resistance (Hancock *et al.*, 2020), it is widely recognised that new vector control tools are urgently needed.

1.4 The problem of insecticide resistance

Highly anthropophilic vectors, such as *An. gambiae* and *An. coluzzii*, have been exposed to extreme selection pressures due to ubiquitous, intensive insecticide use (Ranson *et al.*, 2011). This, alongside remarkable levels of standing genetic variation, large effective population sizes, and a short reproductive cycle has led to rapid adaptation (Miles *et al.*, 2017). This emergence and spread of resistance threatens to curtail the impact of malaria control programmes (Hemingway *et al.*, 2016).

In to order to manage and mitigate insecticide resistance, alternative tools and chemicals are required. Until recent years, only four classes of insecticides were approved for use in public health, between them exhibiting just two modes of action. This limited arsenal of vector control tools has slowed the adoption of insecticide resistance management (IRM) strategies in public health (WHO, 2012; Chanda *et al.*, 2017), and now, resistance is widespread geographically (Ranson *et al.*, 2016). Resistance to pyrethroid insecticides in particular has been well documented in *Anopheles* mosquitoes, however, emerging resistance to other classes, such as organophosphates and carbamates, is also a major concern (Killeen *et al.*, 2018).

An early demonstration of the potential impact of insecticide resistance was in KwaZulu-Natal, South Africa, in 1996. Due to pressure from environmental activist groups, DDT – which had been used as an indoor residual spray for decades - was phased out. Deltamethrin replaced DDT, and by the year 2000, the regions were suffering from a major malaria epidemic (Craig *et al.*, 2004). It was later realised that the cessation of DDT use had allowed pyrethroid-resistant *Anopheles funestus* to re-invade the area, and alongside emergent drug resistance, circulate malaria transmission. The epidemic was halted by the re-introduction of DDT and effective anti-malarials, which reduced malaria cases from 42,000 in 2000, to less than 2100 in 2002 (Maharaj *et al.*, 2005).

Recently, evidence for the impact of insecticide resistance has come from cluster randomised-controlled trials of pyrethroid LLINs that incorporate the synergist piperonyl-butoxide. This compound inhibits cytochrome P450s, a family of enzymes commonly responsible for insecticide resistance. Trials in both Tanzania and Uganda have shown substantial protective effects of PBO nets vs non-PBO, suggesting that resistance may be hampering control efforts (Protopopoff *et al.*, 2018; Staedke *et al.*, 2020), however, the PBO nets tested in these trials also happen to also include a higher concentration of pyrethroids themselves, making any attribution to the PBO alone challenging.

Between 2019 and 2021, a resurgence of malaria was observed in five districts in Uganda, despite LLIN distributions and regular IRS campaigns (Epstein *et al.*, 2022). This resurgence seemed to match the time which the active ingredient in IRS was changed as part of a resistance management strategy, suggesting that pre-existing insecticide resistance to a novel compound may be resulting in IRS failure. Further work needs to be done to establish causality, however. Prior to this resurgence, IRS had been working effectively, reducing malaria by an estimated 86% after four years (Namuganga *et al.*, 2021). A similar story was found during a trial in Sudan, in which mosquitoes were resistant to pyrethroids but susceptible to carbamates. Switching from deltamethrin to bendiocarb IRS was associated with an increase in protection, suggesting that pyrethroid-resistance may have an effect on pyrethroid-based IRS (Kafy *et al.*, 2017).

Despite this, a recent multi-country study found no association between insecticide resistance and malaria prevalence, demonstrating that the barrier effect of bed nets can still be effective in the presence of insecticide-resistant mosquito populations (Kleinschmidt *et al.*, 2018). However, it may be that the current levels of resistance are not impacting malaria control at this time, or that the phenotyping procedures themselves are flawed and do not capture epidemiologically relevant metrics. Its possible that WHO-based metrics of resistance, are not quantitative enough compared to dose-response assays. Although there are now strong indications that insecticide resistance is compromising vector control, the full epidemiological implications are still unclear (Katureebe *et al.*, 2016; Epstein *et al.*, 2022).

1.5 Mechanisms of insecticide resistance

Mechanisms of insecticide resistance can be classified into four categories; Target-site, metabolic, cuticular, and behavioural resistance (Ranson *et al.*, 2016). In *An. gambiae* and *An. coluzzii*, resistance intensities vary substantially in geographic space, with typically the most extreme levels of resistance found in West Africa (Hancock *et al.*, 2020). Some resistance mechanisms have spread throughout sub-Saharan Africa, while others remain localised to specific regions.

1.5.1 Target-site resistance

Target-site resistance involves one or more mutations in the target site of an insecticide, which reduces the binding affinity of the insecticide, allowing the organism to avoid lethal or sub-lethal effects (Hemingway *et al.*, 2004). In some cases, overexpression of the insecticide target can also confer resistance, allowing the organism to tolerate more of the insecticide at any one time, without losing function (Sun *et al.*, 2012).

One of the earliest observations of insecticide resistance in *Anopheles gambiae* was made in 1954 during a malaria elimination programme in Northern Nigeria (Elliott *et al.*, 1956). Approximately 18 months after the introduction of dieldrin, an organochloride, resistant

Anopheles gambiae mosquitoes were detected. It was 42 years later that the responsible locus was mapped to within the 2la inversion (Zheng *et al.*, 1996), and another decade before the specific alleles were revealed (Du *et al.*, 2005). Two mutations at the GABA receptor subunit were discovered, A296G and A296S, which arose via two hard selective sweeps and introgressed between species (from *An. gambiae* and *An. arabiensis* to *An. coluzzii),* and also across karyotypes of the 2La inversion (Grau-Bové, Tomlinson, *et al.*, 2020). Despite the fact that dieldrin was banned in 1974, the *Rdl* mutations remain at moderate frequencies in many *An. gambiae* populations (Grau-Bové, Tomlinson, *et al.*, 2020). Given the widely held assumption that derived target-site mutations will be deleterious in the absence of insecticidal selection pressure (Ffrench-Constant *et al.*, 2017), it is surprising that this is the case. In many insects including *Aedes aegypti*, *Rdl* is a target for the anthelmintic ivermectin (Meng *et al.*, 2019; Wang *et al.*, 2022). *Rdl* may also, therefore, be an avermectin target in *An. gambiae* and the *Rdl* mutations may also confer resistance to ivermectin or other avermectins widely used in agriculture. In support of this hypothesis, a study in *Helicoverpa armigera* looked at the role of the GABA receptors in exposure to multiple classes of insecticides, including cyclodienes and avermectins (Wang *et al.*, 2020). In *H. armigera*, two *Rdl* subunits exist, one which contains a 'wild-type' alanine residue at the homologous 296 position (*HaRdl-1)*, and another subunit that contains a 'resistant' glycine residue (*HaRdl-2)*. Knockout of the 'susceptible' *HaRdl-1* slightly increased resistance to both abamectin and emamectin benzoate, however, *HaRdl-2* knockout did not significantly affect susceptibility. This phenotypic pattern fits with the fact the *Rdl* mutations are thought to be dominant or semi-dominant (Davidson *et al.*, 1962).

In response to selection pressures from DDT and pyrethroids, mutations have arisen at the voltage-gated sodium channel (VGSC) across the Insecta, with mutations at the 1014 codon (*Musca domestica* codon numbering), labelled *kdr* (knockdown-resistant), found in many species (Zlotkin, 1999; Khambay *et al.*, 2005; Syafruddin *et al.*, 2010). In *An. gambiae*, two mutations at this codon have spread throughout sub-Saharan Africa to high frequency, L995F (initially in West Africa, formerly *kdr*-west) and L995S (initially in East Africa, formerly *kdr*-east), with both mutations now segregating in central African populations (Clarkson *et al.*, 2021). The two mutations have been detected so far on 5 distinct haplotypic backgrounds each. An asparagine to tyrosine mutation, N1570Y, has arisen on the L995F background and is found in both *An. coluzzii* and *An. gambiae* in West Africa (Jones *et al.*, 2012; Edi *et al.*, 2017). In combination with L995F, this mutation provides a synergistic effect, increasing the insensitivity of the sodium channel to pyrethroids (Wang *et al.*, 2015). Despite the highly conserved nature of the Vgsc, a number of secondary mutations have now proliferated on the background of these haplotypes, such as P1874L and P1874S, which are likely to enhance resistance or ameliorate fitness costs (Clarkson *et al.*, 2021).

In West Africa, a new haplotype not carrying an L995F/S mutation, but carrying V402L and I1527T mutations, is seemingly spreading and replacing the L995F haplotypes in *An. coluzzii (Williams et al., 2022)*. The two mutations are in complete linkage, they are only found together (Clarkson *et al.*, 2021). Using CRISPR/Cas9, Williams and colleagues expressed the V402L mutant, and showed that it confers resistance to pyrethroids, but at a lower level than L995F (Williams *et al.*, 2022). There were no fitness costs associated with the mutant, however, in contrast with that of L995F (Grigoraki *et al.*, 2021). Unfortunately, the authors were unable to generate an I1527T mutant in conjunction with the V402L mutation, and therefore it is not yet clear what phenotype the double mutants display. A study in guinea also found I1527T to be associated with permethrin resistance (Collins *et al.*, 2019). The V402L mutation is homologous with *Ae. aegypti* V410L, which is known to confer resistance to pyrethroid insecticides, and 1527 is homologous with the *Ae. aegypti* 1532 codon, two codons distant from F1534C, which is also associated with V410L and with resistance in *Ae. aegypti.*

In a recent RNA-Sequencing study I applied a computational pipeline, *RNA-Seq-Pop* (presented in chapter 2 of this thesis), and observed the V402L and I1527T mutations at moderate frequencies as far east as Chad and Niger; the eastern edge of the range of *An. coluzzii (Wiebe et al., 2017; Ibrahim et al., 2022)*. Given that this haplotype was first discovered in Burkina Faso, this data suggests that the novel haplotype is now at appreciable frequencies across the whole range of the species. It has not yet been found in

An. gambiae, but given the history of insecticide resistance mutations introgressing between species of the gambiae complex, it remains likely that this will occur in the near future if it has not already.

The target of organophosphate and carbamate insecticides, *Ace-1*, is the site of both non-synonymous mutations and gene duplications that confer resistance. The G280S mutation (previously G119S - *Torpedo californica* codon numbering), occurs across many taxa and is found in *Anopheles gambiae s.l (Weill et al., 2003; Grau-Bové, Lucas, et al., 2020).* In isolation, this mutation has a high fitness cost, with the mutant enzyme exhibiting low catalytic activity of the neurotransmitter, acetylcholine (Cheung *et al.*, 2017). Heterologous gene duplications have now occurred, pairing resistant Serine with wild-type Glycine alleles, allowing individuals to retain both the wild-type catalytic activity and the resistant phenotype (Labbé *et al.*, 2007; Assogba *et al.*, 2015). This duplication originally spanned a large genomic region and 11 genes, but multiple subsequent deletions of the other duplicated genes are now spreading throughout West Africa, reducing the fitness costs thought to be associated with gene dosage imbalance (Assogba *et al.*, 2018). These alleles arose on a unique haplotype that has introgressed from An. gambiae into An. coluzzii, and is now common in West Africa, at different levels of copy number (Grau-Bové, Lucas, *et al.*, 2020).

1.5.2 Metabolic resistance

Metabolic resistance involves the upregulation of detoxification enzymes that degrade or eliminate the insecticide, typically cytochrome P450s, carboxylesterases or glutathione-S-transferases (Djouaka *et al.*, 2008; Kwiatkowska *et al.*, 2013). As well as overexpression, non-synonymous mutations may also enhance the catalytic activity of the enzyme against a particular compound. This was initially demonstrated in Drosophila, where Amichot and colleagues showed that a single point mutation in the CYP6A2 gene could confer DDT resistance (Amichot *et al.*, 2004). It has since been shown in *Anopheles funestus* - where common polymorphisms in the CYP6P9a and CYP6P9b alleles enhance the metabolism of pyrethroids (Riveron *et al.*, 2014; Ibrahim *et al.*, 2015). In *Anopheles gambiae*, non-synonymous mutations conferring resistance have been rare.

The first detoxification enzymes to be linked to insecticide resistance in *An. gambiae* were glutathione-S-transferases (GSTs). Early work identified GSTs that mapped to a DDT-resistance locus on chromosome 3R (Ranson *et al.*, 2001), containing a cluster of GSTs. Evidence in *Ae. aegypti* suggested *Gste2*, an epsilon class glutathione-S-transferase as an important gene in DDT resistance (Lumjuan *et al.*, 2005). This was later confirmed to also be true in *An. gambiae*, where two mutations, *Gste2-114T* and *Gste2-119V* have been associated with resistance (Mitchell *et al.*, 2014; Lucas, Rockett, *et al.*, 2019). *Gste2-114T* was found to be associated with a gene duplication (Lucas, Rockett, *et al.*, 2019), and in a later genome-wide scan for copy number variants, at least 11 independent amplifications were found encompassing this gene (Lucas, Miles, *et al.*, 2019), with some also encompassing other genes in the *Gste* cluster.

Esterases have also been associated with insecticide resistance in many insects. Esterases are involved in a diverse number of biological processes, including the synthesis of essential hormones and pheromones, which are integral to insect development and reproduction (Oakeshott *et al.*, 2005). Chemically, esterases catalyse the hydrolysis of ester bonds, a common bond found in organic compounds and in many insecticides (Aldridge, 1993). The most well documented example of metabolic resistance in mosquitoes comes from esterases in *Culex pipiens (Raymond et al., 1998)*, where multiple independent gene amplifications around two genes *Est-2* and *Est-3* have spread around the world, potentially in response to organophosphate insecticides (Raymond *et al.*, 1996; Guillemaud *et al.*, 1997). In chapter 4, we show that the orthologous genes are also under selection in *An. gambiae*, and explore their role in insecticide resistance.

In many organisms, cytochrome P450s are the front line of defence against xenobiotics. One of the largest gene families across all organisms, as well as protecting against xenobiotics have a vast array of functions (Feyereisen, 1999). There are 111 annotated P450s in the *An. gambiae* genome (Ranson *et al.*, 2002), compared to around 105 in *Ae. aegypti* and 160 in *Cx. pipiens* (Liu, 2015). Since the development of the first *An. gambiae* microarray chip (David *et al.*, 2005), P450s have been repeatedly implicated as playing a role in insecticide resistance. In field populations, the P450s CYP6P3, CYP6AA1, CYP6M2, CYP9K1, CYP6Z1, and CYP6Z2 have all been reported as overexpressed and validated to metabolise insecticides in vitro (Müller *et al.*, 2008; Mitchell *et al.*, 2012; Chandor-Proust *et al.*, 2013; Vontas *et al.*, 2018; Njoroge *et al.*, 2022). They are primarily implicated to metabolise pyrethroids, however, activity against other classes, such as organochlorines, carbamates, juvenile hormone analogues, and novel classes of insecticides has also been demonstrated (Mitchell *et al.*, 2012; Edi *et al.*, 2014; Yunta *et al.*, 2016; 2019).

ABC-transporters have also been linked to resistance in various insect species (Denecke *et al.*, 2017). They are transmembrane ATP-dependent efflux pumps, which mediate the transport of compounds in and out of the cell. They can transport a wide range of endogenous and exogenous compounds, and also have a diverse range of functions (Denecke *et al.*, 2017). Expression of a cluster of ABC transporters has been shown to be enriched in the legs of *An. gambiae* mosquitoes, and may be involved in lipid biosynthesis at the cuticle (Pignatelli *et al.*, 2018; Kefi *et al.*, 2021).

Beyond the role of the major detoxification families, there are many other families of genes that consistently appear in transcriptomic studies of insecticide resistance. Ingham et al., found that specific chemosensory proteins were induced in the legs after exposure to pyrethroids and that these genes, particularly *Sap2*, bound strongly to pyrethroids to reduce mortality (Ingham *et al.*, 2020). A meta-analysis of microarray studies found Hexamerins, UGTs and alpha-crystallins repeatedly implicated (Ingham *et al.*, 2018). A meta-analysis of more recent RNA-Sequencing studies is briefly described in Appendix A and should enhance the discovery of resistance-associated genes.

1.5.3 Copy Number Variants

Copy number variants (CNVs) are a type of mutation that involves an amplification or deletion of a genomic region. These can be duplications, where two copies of the gene or genomic region are carried on a single haplotype, or the genomic region can be amplified many times to produce multiple copies. When amplifications occur, these may occur in tandem, or in reverse orientation, and CNVs may be ‘clean’, or they may also have extra insertions of genetic sequence at the CNV breakpoints (van Binsbergen, 2011).

In the context of insecticide resistance, CNVs can play two major roles in sustaining a resistant phenotype. Most commonly, the amplification of detoxification genes can lead to overexpression of the amplified genes, meaning that more of the insecticide can be detoxified (Njoroge *et al.*, 2022). Alternatively, or in combination with this process, CNVs can pair wild-type alleles with mutant alleles, allowing them to co-exist on the same haplotype in a state of permanent heterozygosity (Weetman *et al.*, 2018; Lucas, Rockett, *et al.*, 2019). Notably, this has occurred at the *Ace-1* locus in *Anopheles gambiae* (Grau-Bové, Lucas, *et al.*, 2020), and the *Rdl* gene in *Drosophila melanogaster (Remnant et al., 2013)*, in which functional wild-type alleles are paired with ‘resistant’ mutant alleles.

In *An. gambiae* and *An. coluzzii*, recent sequencing efforts have highlighted large numbers of CNVs at metabolic resistance genes (Lucas, Miles, *et al.*, 2019). At the Cyp6 locus, where *Cyp6p3* had previously been the major candidate gene, 16 independent amplifications were found, of which the majority covered the entirety of the *Cyp6aa1* gene, which had previously been overlooked in *An. gambiae*.

1.5.4 Cuticular resistance

Often thought to work in synergism with metabolic mechanisms, cuticular resistance is a thickening or change in the composition of the insect cuticle, slowing the rate of insecticide penetration (Bass *et al.*, 2016; Yahouédo *et al.*, 2017; Simma *et al.*, 2019). The cuticle is the outermost part of the insect, and as well as providing structural support, protects against desiccation, and is important in sensory perception of the environment (Balabanidou *et al.*, 2018).

Two P450s, CYP4G16 and CYP4G17, have been implicated in the cuticular hydrocarbon synthesis pathway (Balabanidou *et al.*, 2016), and further work has uncovered these processes in greater detail, focusing on oenocytes, specialised cells at the cuticle surface (Grigoraki *et al.*, 2020). Grigoraki and colleagues performed RNA-Sequencing specifically on oenocytes and revealed a set of genes working in concert to produce cuticular hydrocarbons, including fatty acid synthases, reductases and elongases. Another transcriptomic study focused on legs, and found that insecticide detoxification is likely to be occurring in the legs alongside cuticular modifications (Kefi *et al.*, 2021). In many insect species, cuticular resistance has been observed to lead to high resistance ratios (Ahmad *et al.*, 2006).

1.5.5 Behavioural resistance

Behavioural adaptations allow mosquitoes to evade or minimise contact with insecticides, independent of physiological or biochemical changes (Gatton *et al.*, 2013). Female *Anopheles gambiae* mosquitoes are thought to host-seek and blood-feed almost exclusively indoors at night when humans are sleeping. A change from indoor to outdoor biting, and shifts to crepuscular biting patterns where humans are not protected by bed nets represents a major obstacle to malaria control (Reddy *et al.*, 2011; Killeen *et al.*, 2014).

In a recent study from urban Bangui, Central African Republic, Ayala and colleagues performed the Human landing catch (HLCs) over a period of 48 continuous hours, rather than the typical convention of 12 hours between 6pm to 6am. They detected a high proportion of *Anopheles* bites occurring inside, during the day (Sangbakembi-Ngounou *et al.*, 2022). Other studies have also found *An. funestus* biting much later than conventionally thought (Moiroux *et al.*, 2012; Sougoufara *et al.*, 2014), however, it is not clear whether this behaviour has changed due to exposure to insecticidal interventions.

Strong evidence for behavioural resistance is lacking, particularly due to the need for robust temporal data, ideally before and after an intervention. New technologies, such as infrared video-tracking of mosquitoes may shine a light on these processes (Parker et al., 2015). Although this thesis primarily focuses on aspects of physiological resistance rather than behavioural resistance, the evolution of novel and insecticide-avoiding behaviours is a major concern and should be a focus of future studies of malaria vectors.

1.6 Managing insecticide resistance

The evolution of insecticide resistance can be managed and mitigated (Roush *et al.*, 2012; Sparks *et al.*, 2015), with strategic usage of control interventions. These techniques were pioneered in agricultural systems in which crop pests threaten food security (Tabashnik, 1989), however, in vector control, the limited number of available chemical classes makes insecticide resistance management (IRM) even more essential. In 2012 the WHO established the GPIRM, the global plan for insecticide resistance management (WHO, 2012), aiming to prolong the usefulness of the current and future vector control arsenal. This document summarised the problem of insecticide resistance and contained information for malaria-endemic countries on how to perform insecticide resistance management (IRM). For malaria vectors, the main strategies are outlined below:

1. **Reduce the use of insecticides.** If we are able to reduce mosquito populations without the use of chemicals, for example, with larval source management, the selection pressure to evolve resistance will be removed. This is a primary objective of integrated vector management (IVM) (Chanda *et al.*, 2017).
2. **Ensure a killing dose of insecticide.** If an insect is exposed to a high, killing dose of insecticide, it can not go on to pass on genetic material to the next generation. However, if mosquitoes are exposed to moderate or weak doses of insecticides, some will survive, and those that survive are likely to contain mutations which increase their ability to survive insecticides. This allows resistance to spread in a population.
3. **Use insecticides in rotations, mixtures, or mosaics.** Rotations and mosaics aim to reduce the evolution of resistance by providing regions in time or space in which the mosquito is unexposed to the chemical. As the insecticide resistance mutation should theoretically have a fitness cost in the absence of insecticide, its spread will be slowed. The idea behind mixtures of insecticides is that the additive effect of the AIs should kill the insect, or that at least one of them will.

The looming threat of pyrethroid resistance has led malaria stakeholders to search for new active ingredients and combinations of chemicals that mitigate the risk of resistance. As described in chapter 2.3, Dual AI Pyrethroid LLINs have recently come to market, which integrate the pyrrole chlorfenapyr or the juvenile hormone analogue, pyriproxyfen. In 2022, PBO and Dual AI nets now represent more than half of the LLINs delivered globally. Novel active ingredients for vector control are also in development (IVCC, 2021). The introductions of these new products to the vector control market should allow for pre-emptive insecticide resistance management (IRM) strategies to be employed.

Effective insecticide resistance management programmes will need to detect resistance rapidly in vector populations. Robust systems for regular phenotyping will play an important role, however, genomic approaches have the potential to provide an early warning system for the detection of novel and known resistance variants.

1.7 Genomic surveillance

1.7.1 Genomic surveillance of infectious diseases

The Covid-19 pandemic irrefutably demonstrated the value of genomic surveillance (Mercer *et al.*, 2021). Genomic sequencing of Coronavirus strains allowed scientists to study the spread, evolution and pathogenicity of the virus with unprecedented resolution (Robishaw *et al.*, 2021). Nationwide genomic surveillance networks were rapidly established (COVID-19 Genomics UK, 2020; Msomi *et al.*, 2020; Michaelsen *et al.*, 2022) to accelerate research on the virus - to track transmission, and to identify viral mutations and risk factors for the disease in a concerted manner (Robishaw *et al.*, 2021). These efforts have been underpinned by open data-sharing principles and standardised lineage definitions (Rambaut *et al.*, 2020), allowing scientific organisations to define stable transmission lineages. At the time of writing, 14,247,918 SARS-CoV-2 genome sequences have been submitted to the GISAID, the Global Initiative on Sharing Influenza Data.

Similar approaches have been applied to malaria research. In 2005, the Malaria Genomic Epidemiology Network (MalariaGEN) was established, a data-sharing network of partners in over 40 countries, who build and share large-scale human, malaria and mosquito resources (MalariaGEN, 2008). Through successive phases of the parasite project, over 21,000 samples of *Plasmodium falciparum* have been whole-genome sequenced and released to the research community (Manske *et al.*, 2012; MalariaGEN *et al.*, 2021; MalariaGEN, 2023). By comparing genetic variation data, researchers can estimate the relatedness of parasites in different hosts, reconstructing its spread to investigate transmission dynamics (Amato *et al.*, 2018), or explore population structure (Hamilton *et al.*, 2019; Amambua-Ngwa *et al.*, 2019). The genomic data has been used to study resistance to front-line antimalarials (Miotto *et al.*, 2013; 2015; MalariaGEN *et al.*, 2015), and how the parasite may avoid the human immune system (Claessens *et al.*, 2014). These studies are now beginning to feed back into malaria control programmes, highlighting how if performed in a timely manner, genomic surveillance can provide policymakers with actionable evidence to make decisions.

1.7.2 Genomic surveillance of malaria mosquitoes

In an analogous manner to surveillance of the *Plasmodium* parasite, it is possible to discover and track insecticide and gene-drive resistance mutations in the *Anopheles* mosquito. However, in vectors like *An. gambiae*, genomic surveillance is challenging. Partly this is due to the sheer size of the genome, which makes sequencing a single specimen more costly - *Plasmodium* genomes are around 18 to 30 Mb, whereas the *Anopheles gambiae* PEST reference genome is approximately 278 Mb in size (Holt *et al.*, 2002; Bushell *et al.*, 2017). Also, the number of genes is higher, around 13,000 in *Anopheles* compared to less than 5,000 in *Plasmodium* species. In addition, like all mosquito species, *An. gambiae* are diploid (Matthews *et al.*, 1994), and sexually recombine. Lastly, the sheer scale of genetic diversity in the major malaria vector (Leffler *et al.*, 2012; Miles *et al.*, 2017) can mean analytical techniques developed for the analysis of other organisms with much lower levels of diversity, typically humans, may not be directly applicable.

After the initial studies on population structure in *An. gambiae* using polymorphic chromosomal inversions (Coluzzi *et al.*, 1979; 1985), research began to focus on DNA sequence data. Early work on the genetics of the *An. gambiae* species complex was suggestive of gene flow between *An. gambiae* and *An. arabiensis*, and that these species were sister taxa (Besansky *et al.*, 1994; 1997). A number of studies found low rates of differentiation across most of the range of *An. gambiae (Lehmann et al., 1996; 1997; Donnelly et al., 1999; 2004)*, with an exception for populations on either side of the rift valley (Lehmann *et al.*, 2003), with similar patterns of population structure found in the other primary Afrotropical vector, *An. funestus* (Michel *et al.*, 2005).

The genome sequence of *Anopheles gambiae* was sequenced and published in 2002 (Holt *et al.*, 2002), and updated in 2007 (Sharakhova *et al.*, 2007). After this period, studies into genomic DNA moved towards higher-resolution SNP arrays, and eventually whole-genome sequencing (Neafsey *et al.*, 2010; Lawniczak *et al.*, 2010). These studies mostly confirmed that of the earlier genetic studies, finding low genetic differentiation in collinear regions of the genome, and strong differentiation in regions of chromosomal inversions (Reidenbach *et al.*, 2012; Cheng *et al.*, 2012). Fontaine *et al.*, performed phylogenomic analyses across this *An. gambiae* species complex, finding signals of extensive introgression (Fontaine *et al.*, 2015), as was found in the *An. funestus* species complex (Small *et al.*, 2020).

Around this time, MalariaGEN launched *The Anopheles 1000 genomes project*, a multi-country partnership across 20 research institutions to provide a high-resolution view of genetic diversity of the major malaria mosquito, *Anopheles gambiae s.l.* The first phase of the project sequenced 765 *An. gambiae* and *An. coluzzii* mosquitoes, producing a high-quality database of single-nucleotide variation (Miles *et al.*, 2017). The project revealed complex population structure, localised demographic histories and large signals of selection at known and novel loci. The second phase of the project included 1142 whole-genomes and expanded the data resource to 13 countries, focusing on isolation-by-distance analyses, and integrating copy number variant calls (Lucas, Miles, *et al.*, 2019; Ag1000G, 2020). The third and final phase is unpublished at time of writing but includes representation from *An. arabiensis* and brings the total number of sequenced mosquitoes to 2784 from 19 countries. I utilise this Phase 3 data resource in chapters 3 and 4 of this thesis. In conjunction with phase 3, an extensive python package, malariagen_data, has been developed, which allows users to explore the variation data and easily perform a wide range of population genetic analyses in the cloud (Miles *et al.*, 2022).

Genomic surveillance has enhanced the discovery of variants involved in resistance to insecticides (Clarkson *et al.*, 2018). Since the release of the Ag1000G, studies have illuminated the architecture of target-site resistance at the target of pyrethroids and DDT, the *Vgsc*, the organophosphate and carbamate target *Ace-1*, and the GABA receptor, *Rdl* (Grau-Bové, Lucas, *et al.*, 2020; Grau-Bové, Tomlinson, *et al.*, 2020; Clarkson *et al.*, 2021). Whole-genome sequencing has also enabled the discovery of genes which had previously been overlooked based on transcriptomic studies alone (Njoroge *et al.*, 2022).

As well as whole-genome sequencing, it is likely that targeted sequencing will play a major role in genomic surveillance of malaria vectors in the future, due to its much-reduced cost and throughput capabilities. Recently, amplicon sequencing panels have been developed which can accurately identify species across the entire *Anopheles* genus (Boddé *et al.*, 2022; Makunin *et al.*, 2022). In a similar way that has been done for drug resistance markers in malaria parasites (Girgis *et al.*, 2022), targeted sequencing could be used to track resistance mutations in *Anopheles* mosquitoes.

Similarly, transcriptomic studies may even provide opportunities for genomic surveillance beyond gene expression. The publication of the *An. gambiae* genome allowed microarrays to assay for gene expression to be developed (David *et al.*, 2005), eventually widely employed primarily to understand physiological changes in response to insecticide exposure. Since the early 2010s, microarray use has been gradually replaced with short-read RNA-Sequencing, which provides a more unbiased, agnostic view into the transcriptome (Zhao *et al.*, 2014). These RNA-Sequencing studies into insecticide resistance are performed regularly, but often with no corresponding whole-genome sequence data. In chapter 2 of this thesis, I present a reproducible computational pipeline, *RNA-Seq-Pop*, which exploits the SNP data within RNA-Seq to extend the reach of genomic surveillance to RNA-Sequencing.

1.7.3 Reproducibility and open source software principles in computational research

Reproducibility is vital to generate trust in scientific results, however, the reproducibility crisis in the life sciences extends beyond that of laboratory research. Multiple studies on computational reproducibility have found that often, reproducing findings from papers is fraught with difficulty (Grüning *et al.*, 2018; Reinecke *et al.*, 2022). Methods sections may be incomplete, source code is unavailable, or data is not stored in a centralised, open repository (Grüning *et al.*, 2018). Even if the source code is complete, it may be written or organised in such a way as to make the task of replication either a) time-consumingly unfeasible or b) impossible. Researchers are therefore looking for reproducible and standardised ways of analysing data.

Concurrent to the reproducibility crisis, genomics capacity and skills are lacking, and unable to meet the vast demand of modern science (H3Africa Consortium *et al.*, 2014). It follows that the development of software tools which standardise and enable reproducible research should be of value. This is particularly true in tropical medicine, as the bioinformatic capacity of institutes is not distributed equitably throughout the world. Fortunately, the development of workflow management systems over the past decade, such as Snakemake (Köster *et al.*, 2012; Mölder *et al.*, 2021) and Nextflow (Di Tommaso *et al.*, 2017), has made it easier for computational researchers to develop reproducible tools and pipelines.

In the following chapters, I present genomic analyses of the major malaria vector *Anopheles gambiae*. These analyses are contained within, or accompanied by software tools which aim to make analyses for other researchers both more effective and reproducible. I hope that these projects will aid scientists in the field of malaria and beyond to perform their own research. To aid transparency and in the ethos of open-source software development, the code for all analyses are stored within Github or Google Colaboratory.

1.8 References

Ag1000G (2020) 'Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii', *Genome Res*, pp. 1–14.

Ahmad, M., Denholm, I. and Bromilow, R.H. (2006) 'Delayed cuticular penetration and enhanced metabolism of deltamethrin in pyrethroid-resistant strains of Helicoverpa armigera from China and Pakistan', *Pest management science*, 62(9), pp. 805–810.

Aldridge, W.N. (1993) 'The esterases: perspectives and problems', *Chemico-biological interactions*, 87(1-3), pp. 5–13.

Amambua-Ngwa, A., Amenga-Etego, L., Kamau, E., Amato, R., Ghansah, A., Golassa, L., Randrianarivelojosia, M., Ishengoma, D., Apinjoh, T., Maïga-Ascofaré, O., Andagalu, B., Yavo, W., Bouyou-Akotet, M., Kolapo, O., Mane, K., *et al.* (2019) 'Major subpopulations of *Plasmodium falciparum* in sub-Saharan Africa', *Science*, 365(6455), pp. 813–816.

Amato, R., Pearson, R.D., Almagro-Garcia, J., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Drury, E., Stalker, J., Miotto, O., Fairhurst, R.M. and Kwiatkowski, D.P. (2018) 'Origins of the current outbreak of multidrug-resistant malaria in southeast Asia: a retrospective genetic study', *The Lancet infectious diseases*, 18(3), pp. 337–345.

Amichot, M., Tarés, S., Brun-Barale, A., Arthaud, L., Bride, J.M. and Bergé, J.B. (2004) 'Point mutations associated with insecticide resistance in the Drosophila cytochrome P450 Cyp6a2 enable DDT metabolism', *European journal of biochemistry / FEBS*, 271(7), pp. 1250–1257.

Assogba, B.S., Alout, H., Koffi, A., Penetier, C., Djogbénou, L.S., Makoundou, P., Weill, M. and Labbé, P. (2018) 'Adaptive deletion in resistance gene duplications in the malaria vector Anopheles gambiae', *Evolutionary applications*, 11(8), pp. 1245–1256.

Assogba, B.S., Djogbénou, L.S., Milesi, P., Berthomieu, A., Perez, J., Ayala, D., Chandre, F., Makoutodé, M., Labbé, P. and Weill, M. (2015) 'An ace-1 gene duplication resorbs the fitness cost associated with resistance in Anopheles gambiae, the main malaria mosquito', *Scientific reports*, 5(August), pp. 19–21.

Balabanidou, V., Grigoraki, L. and Vontas, J. (2018) 'Insect cuticle: a critical determinant of insecticide resistance', *Current Opinion in Insect Science*, 27, pp. 68–74.

Balabanidou, V., Kampouraki, A., MacLean, M., Blomquist, G.J., Tittiger, C., Juárez, M.P., Mijailovsky, S.J., Chalepakis, G., Anthousi, A., Lynd, A., Antoine, S., Hemingway, J., Ranson, H., Lycett, G.J. and Vontas, J. (2016) 'Cytochrome P450 associated with insecticide resistance catalyzes cuticular hydrocarbon production in Anopheles gambiae', *Proceedings of the National Academy of Sciences*, 113(33), p. 201608295.

Barrón, M.G., Paupy, C., Rahola, N., Akone-Ella, O., Ngangue, M.F., Wilson-Bahun, T.A., Pombi, M., Kengne, P., Costantini, C., Simard, F., González, J. and Ayala, D. (2019) 'A new species in the major malaria vector complex sheds light on reticulated species evolution', *Scientific reports*, 9(1), p. 14753.

Bass, C. and Jones, C.M. (2016) 'Mosquitoes boost body armor to resist insecticide attack', *Proceedings of the National Academy of Sciences*, 113(33), pp. 9145–9147.

Besansky, N.J., Hill, C.A. and Costantini, C. (2004) ‘No accounting for taste: host preference in malaria vectors’, *Trends in parasitology*, 20(6), pp. 249–251.

Besansky, N.J., Lehmann, T., Fahey, G.T., Fontenille, D., Braack, L.E., Hawley, W.A. and 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(4), pp. 1817–1828.

Besansky, N.J., Powell, J.R., Caccone, A., Hamm, D.M., Scott, J.A. and Collins, F.H. (1994) ‘Molecular phylogeny of the Anopheles gambiae complex suggests genetic introgression between principal malaria vectors’, *Proceedings of the National Academy of Sciences of the United States of America*, 91(15), pp. 6885–6888.

Bhatt, S., Weiss, D.J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C.L., Henry, A., Eckhoff, P.A., Wenger, E.A., Briët, O., Penny, M.A., Smith, T.A., Bennett, A., *et al.* (2015) ‘The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015’, *Nature*, 526(7572), pp. 207–211.

van Binsbergen, E. (2011) ‘Origins and breakpoint analyses of copy number variations: up close and personal’, *Cytogenetic and genome research*, 135(3-4), pp. 271–276.

Boddé, M., Makunin, A., Ayala, D., Bouafou, L., Diabaté, A., Ekpo, U.F., Kientega, M., Le Goff, G., Makanga, B.K., Ngangue, M.F., Omitola, O.O., Rahola, N., Tripet, F., Durbin, R. and Lawniczak, M.K.N. (2022) ‘High-resolution species assignment of Anopheles mosquitoes using k-mer distances on targeted sequences’, *eLife*, 11. doi:10.7554/eLife.78775.

Bushell, E., Gomes, A.R., Sanderson, T., Anar, B., Girling, G., Herd, C., Metcalf, T., Modrzynska, K., Schwach, F., Martin, R.E., Mather, M.W., McFadden, G.I., Parts, L., Rutledge, G.G., Vaidya, A.B., *et al.* (2017) ‘Functional Profiling of a Plasmodium Genome Reveals an Abundance of Essential Genes’, *Cell*, 170(2), pp. 260–272.e8.

Chanda, E., Ameneshewa, B., Bagayoko, M., Govere, J.M. and Macdonald, M.B. (2017) ‘Harnessing Integrated Vector Management for Enhanced Disease Prevention’, *Trends in Parasitology*, pp. 30–41. doi:10.1016/j.pt.2016.09.006.

Chandor-Proust, A., Bibby, J., Régent-Kloeckner, M., Roux, J., Guittard-Crilat, E., Poupardin, R., Riaz, M.A., Paine, M., Dauphin-Villemant, C., Reynaud, S. and David, J.-P. (2013) ‘The central role of mosquito cytochrome P450 CYP6Zs in insecticide detoxification revealed by functional expression and structural modelling’, *The Biochemical journal*, 455(1), pp. 75–85.

Cheng, C., White, B.J., Kamdem, C., Mockaitis, K., Costantini, C., Hahn, M.W. and Besansky, N.J. (2012) ‘Ecological genomics of anopheles gambiae along a latitudinal cline: A population-resequencing approach’, *Genetics*, 190(4), pp. 1417–1432.

Cheung, J., Mahmood, A., Kalathur, R., Liu, L. and Carlier, P.R. (2017) ‘Structure of the G119S Mutant Acetylcholinesterase of the Malaria Vector Anopheles gambiae Reveals Basis of Insecticide Resistance’, *Structure* , pp. 1–7.

Claessens, A., Hamilton, W.L., Kekre, M., Otto, T.D., Faizullabhoy, A., Rayner, J.C. and Kwiatkowski, D. (2014) ‘Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of Var genes during mitosis’, *PLoS genetics*, 10(12), p. e1004812.

Clarkson, C.S., Miles, A., Harding, N.J., O’Reilly, A.O., Weetman, D., Kwiatkowski, D. and Donnelly, M.J. (2021) ‘The genetic architecture of target-site resistance to pyrethroid insecticides in the

African malaria vectors Anopheles gambiae and Anopheles coluzzii', *Molecular ecology*, 30(21), pp. 5303–5317.

Clarkson, C.S., Temple, H.J. and Miles, A. (2018) 'The genomics of insecticide resistance: insights from recent studies in African malaria vectors', *Current opinion in insect science*, 27, pp. 111–115.

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

Cohuet, A., Harris, C., Robert, V. and Fontenille, D. (2010) 'Evolutionary forces on Anopheles: what makes a malaria vector?', *Trends in parasitology*, 26(3), pp. 130–136.

Collins, E., Vaselli, N.M., Sylla, M., Beavogui, A.H., Orsborne, J., Lawrence, G., Wiegand, R.E., Irish, S.R., Walker, T. and Messenger, L.A. (2019) 'The relationship between insecticide resistance, mosquito age and malaria prevalence in Anopheles gambiae s.l. from Guinea', *Scientific reports*, 9(1), p. 8846.

Coluzzi, M., Petrarca, V. and di Deco, M.A. (1985) 'Chromosomal inversion intergradation and incipient speciation in Anopheles gambiae', *Bollettino di zoologia*, 52(1-2), pp. 45–63.

Coluzzi, M. and Sabatini, A. (1967) 'Cytogenetic observations on species A and B of the Anopheles gambiae complex', *Parassitologia* [Preprint]. Available at: https://www.semanticscholar.org/paper/e874b633d46c527042f40cf8c6ba28883839ea36 (Accessed: 9 January 2023).

Coluzzi, M., Sabatini, A., Petrarca, V. and 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(5), pp. 483–497.

COVID-19 Genomics UK (2020) 'An integrated national scale SARS-CoV-2 genomic surveillance network', *The Lancet. Microbe*, 1(3), pp. e99–e100.

Craig, M.H., Kleinschmidt, I., Nawn, J.B., Le Sueur, D. and Sharp, B.L. (2004) 'Exploring 30 years of malaria case data in KwaZulu-Natal, South Africa: part I. The impact of climatic factors', *Tropical medicine & international health: TM & IH*, 9(12), pp. 1247–1257.

David, J.-P., Strode, C., Vontas, J., Nikou, D., Vaughan, A., Pignatelli, P.M., Louis, C., Hemingway, J. and Ranson, H. (2005) 'The Anopheles gambiae detoxification chip: a highly specific microarray to study metabolic-based insecticide resistance in malaria vectors', *Proceedings of the National Academy of Sciences of the United States of America*, 102(11), pp. 4080–4084.

Davidson, G. (1964) 'ANOPHELES GAMBIAE, A COMPLEX OF SPECIES', *Bulletin of the World Health Organization*, 31(5), pp. 625–634.

Davidson, G. and Hamon, J. (1962) 'A Case of Dominant Dieldrin Resistance in Anopheles gambiae Giles', *Nature*, 196(4858), pp. 1012–1012.

Denecke, S., Fusetto, R. and Batterham, P. (2017) 'Describing the role of Drosophila melanogaster ABC transporters in insecticide biology using CRISPR-Cas9 knockouts', *Insect biochemistry and molecular biology* [Preprint]. doi:10.1016/j.ibmb.2017.09.017.

Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E. and Notredame, C. (2017)

‘Nextflow enables reproducible computational workflows’, *Nature biotechnology*, 35(4), pp. 316–319.

Djouaka, R.F., Bakare, A.A., Coulibaly, O.N., Akogbeto, M.C., Ranson, H., Hemingway, J. and Strode, C. (2008) ‘Expression of the cytochrome P450s, CYP6P3 and CYP6M2 are significantly elevated in multiple pyrethroid resistant populations of Anopheles gambiae s.s. from Southern Benin and Nigeria’, *BMC genomics*, 9(1), p. 538.

Donnelly, M.J., Cuamba, N., Charlwood, J.D., Collins, F.H. and Townson, H. (1999) ‘Population structure in the malaria vector, Anopheles arabiensis patton, in East Africa’, *Heredity*, 83 (Pt 4), pp. 408–417.

Donnelly, M.J., Pinto, J., Girod, R., Besansky, N.J. and Lehmann, T. (2004) ‘Revisiting the role of introgression vs shared ancestral polymorphisms as key processes shaping genetic diversity in the recently separated sibling species of the Anopheles gambiae complex’, *Heredity*, 92(2), pp. 61–68.

Du, W., Awolola, T.S., Howell, P., Koekemoer, L.L., Brooke, B.D., Benedict, M.Q., Coetzee, M. and Zheng, L. (2005) ‘Independent mutations in the Rdl locus confer dieldrin resistance to Anopheles gambiae and An. arabiensis’, *Insect molecular biology*, 14(2), pp. 179–183.

Edi, A.V.C., N’Dri, B.P., Chouaibou, M., Kouadio, F.B., Pignatelli, P., Raso, G., Weetman, D. and Bonfoh, B. (2017) ‘First detection of N1575Y mutation in pyrethroid resistant Anopheles gambiae in Southern Côte d’Ivoire’, *Wellcome Open Research*, 2(0), pp. 1–10.

Edi, C.V., Djogbénou, L., Jenkins, A.M., Regna, K., Muskavitch, M.A.T., Poupardin, R., Jones, C.M., Essandoh, J., Kétoh, G.K., Paine, M.J.I., Koudou, B.G., Donnelly, M.J., Ranson, H. and Weetman, D. (2014) ‘CYP6 P450 Enzymes and ACE-1 Duplication Produce Extreme and Multiple Insecticide Resistance in the Malaria Mosquito Anopheles gambiae’, *PLoS genetics*, 10(3). doi:10.1371/journal.pgen.1004236.

Elliott, R. and Ramakrishna, V. (1956) ‘Insecticide resistance in Anopheles gambiae Giles’, *Nature*, 177(4507), pp. 532–533.

Epstein, A., Maiteki-Sebuguzi, C., Namuganga, J.F., Nankabirwa, J.I., Gonahasa, S., Opigo, J., Staedke, S.G., Rutazaana, D., Arinaitwe, E., Kamya, M.R., Bhatt, S., Rodríguez-Barraquer, I., Greenhouse, B., Donnelly, M.J. and Dorsey, G. (2022) ‘Resurgence of malaria in Uganda despite sustained indoor residual spraying and repeated long lasting insecticidal net distributions’, *PLOS Global Public Health*, 2(9), p. e0000676.

Favia, G., della Torre, A., Bagayoko, M., Lanfrancotti, A., Sagnon, N., Touré, Y.T. and Coluzzi, M. (1997) ‘Molecular identification of sympatric chromosomal forms of Anopheles gambiae and further evidence of their reproductive isolation’, *Insect molecular biology*, 6(4), pp. 377–383.

Feyereisen, R. (1999) ‘Insect P450 Enzymes’, *Annual review of entomology*, 44, pp. 507–533.

Ffrench-Constant, R.H. and Bass, C. (2017) ‘Does resistance really carry a fitness cost?’, *Current Opinion in Insect Science*, 21, pp. 39–46.

Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V., Jiang, X., Hall, A.B., Catteruccia, F., Kakani, E., Mitchell, S.N., Wu, Y.-C., Smith, H.A., Love, R.R., Lawniczak, M.K.N., *et al.* (2015) ‘Extensive introgression in a malaria vector species complex revealed by phylogenomics’, *Science*, 347(6217), p. 1258522.

Gatton, M.L., Chitnis, N., Churcher, T., Donnelly, M.J., Ghani, A.C., Godfray, H.C.J., Gould, F., Hastings, I., Marshall, J., Ranson, H., Rowland, M., Shaman, J. and Lindsay, S.W. (2013) 'The importance of mosquito behavioural adaptations to malaria control in Africa', *Evolution; international journal of organic evolution* [Preprint]. doi:10.1111/evo.12063.

Giles (1902) *A handbook of the gnats or mosquitoes; giving the anatomy and life history of the Culicidæ together with descriptions of all species noticed up to the present date*. London, J. Bale, sons & Danielsson, ltd, 1902, p. 586.

Girgis, S.T., Adika, E., Nenyewodey, F.E., Senoo Jnr, D.K., Ngoi, J.M., Bandoh, K., Lorenz, O., van de Steeg, G., Nsoh, S., Judge, K., Pearson, R.D., Almagro-Garcia, J., Saiid, S., Atampah, S., Amoako, E.K., *et al.* (2022) 'Nanopore sequencing for real-time genomic surveillance of Plasmodium falciparum', *bioRxiv*. doi:10.1101/2022.12.20.521122.

Grau-Bové, X., Lucas, E., Pipini, D., Rippon, E., van't Hof, A., Constant, E., Dadzie, S., Egyir-Yawson, A., Essandoh, J., Chabi, J., Djogbénou, L., Harding, N., Miles, A., Kwiatkowski, D., Donnelly, M., *et al.* (2020) *Resistance to pirimiphos-methyl in West African Anopheles is spreading via duplication and introgression of the Ace1 locus*, pp. 1–34.

Grau-Bové, X., Tomlinson, S., O'Reilly, A.O., Harding, N.J., Miles, A., Kwiatkowski, D., Donnelly, M.J., Weetman, D. and Anopheles gambiae 1000 Genomes Consortium (2020) 'Evolution of the Insecticide Target Rdl in African Anopheles Is Driven by Interspecific and Interkaryotypic Introgression', *Molecular biology and evolution*, 37(10), pp. 2900–2917.

Grigoraki, L., Cowlishaw, R., Nolan, T., Donnelly, M., Lycett, G. and Ranson, H. (2021) 'CRISPR/Cas9 modified An. gambiae carrying kdr mutation L1014F functionally validate its contribution in insecticide resistance and combined effect with metabolic enzymes', *PLoS genetics*, 17(7), p. e1009556.

Grigoraki, L., Grau-Bové, X., Carrington Yates, H., Lycett, G.J. and Ranson, H. (2020) 'Isolation and transcriptomic analysis of Anopheles gambiae oenocytes enables the delineation of hydrocarbon biosynthesis', *eLife*, 9, pp. 1–24.

Grüning, B., Chilton, J., Köster, J., Dale, R., Soranzo, N., van den Beek, M., Goecks, J., Backofen, R., Nekrutenko, A. and Taylor, J. (2018) 'Practical Computational Reproducibility in the Life Sciences', *Cell Systems*, 6(6), pp. 631–635.

Guillemaud and Makate (1997) 'Esterase gene amplification in Culex pipiens', *Insect molecular biology* [Preprint]. Available at: http://www.webmail.evolutionhumaine.fr/pdf_articles/guillemaud_1997_insect_molecular_biology.pdf.

H3Africa Consortium, Rotimi, C., Abayomi, A., Abimiku, A. 'le, Adabayeri, V.M., Adebamowo, C., Adebiyi, E., Ademola, A.D., Adeyemo, A., Adu, D., Affolabi, D., Agongo, G., Ajayi, S., Akarolo-Anthony, S., Akinyemi, R., *et al.* (2014) 'Research capacity. Enabling the genomic revolution in Africa', *Science*, 344(6190), pp. 1346–1348.

Hamilton, W.L., Amato, R., van der Pluijm, R.W., Jacob, C.G., Quang, H.H., Thuy-Nhien, N.T., Hien, T.T., Hongvanthong, B., Chindavongsa, K., Mayxay, M., Huy, R., Leang, R., Huch, C., Dysoley, L., Amaratunga, C., *et al.* (2019) 'Evolution and expansion of multidrug-resistant malaria in southeast Asia: a genomic epidemiology study', *The Lancet infectious diseases*, 19(9), pp. 943–951.

Hancock, P.A., Hendriks, C.J.M., Tangena, J.-A., Gibson, H., Hemingway, J., Coleman, M., Gething, P.W., Cameron, E., Bhatt, S. and Moyes, C.L. (2020) ‘Mapping trends in insecticide resistance phenotypes in African malaria vectors’, *PLoS biology*, 18(6), p. e3000633.

Hemingway, J., Hawkes, N.J., McCarroll, L. and Ranson, H. (2004) ‘The molecular basis of insecticide resistance in mosquitoes’, in *Insect Biochemistry and Molecular Biology*. doi:10.1016/j.ibmb.2004.03.018.

Hemingway, J., Ranson, H., Magill, A., Kolaczinski, J., Fornadel, C., Gimnig, J., Coetzee, M., Simard, F., Roch, D.K., Hinzoumbe, C.K., Pickett, J., Schellenberg, D., Gething, P., Hoppé, M. and Hamon, N. (2016) ‘Averting a malaria disaster: Will insecticide resistance derail malaria control?’, *The Lancet* [Preprint]. doi:10.1016/S0140-6736(15)00417-1.

Holt, R.A., Mani Subramanian, G., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M.C., Wides, R., Salzberg, S.L., Loftus, B., Yandell, M., Majoros, W.H., Rusch, D.B., *et al.* (2002) ‘The genome sequence of the malaria mosquito Anopheles gambiae’, *Science*, 298(5591), pp. 129–149.

Ibrahim, S.S., Muhammad, A., Hearn, J., Weedall, G.D., Nagi, S.C., Mukhtar, M.M., Fadel, A.N., Mugenzi, L.J., Patterson, E.I., Irving, H. and Wondji, C.S. (2022) ‘Molecular drivers of insecticide resistance in the Sahelo-Sudanian populations of a major malaria vector’, *bioRxiv* [Preprint]. doi:10.1101/2022.03.21.485146.

Ibrahim, S.S., Riveron, J.M., Bibby, J., Irving, H., Yunta, C., Paine, M.J.I. and Wondji, C.S. (2015) ‘Allelic Variation of Cytochrome P450s Drives Resistance to Bednet Insecticides in a Major Malaria Vector’, *PLoS genetics*, 11(10), pp. 1–25.

Ingham, V.A., Anthousi, A., Douris, V., Harding, N.J., Lycett, G., Morris, M., Vontas, J. and Ranson, H. (2020) ‘A sensory appendage protein protects malaria vectors from pyrethroids’, *Nature*, 577(7790), pp. 376–380.

Ingham, V., Wagstaff, S. and Ranson, H. (2018) ‘Transcriptomic meta-signatures identified in Anopheles gambiae populations reveal previously undetected insecticide resistance mechanisms’, *Nature communications* [Preprint]. doi:10.1038/s41467-018-07615-x.

IVCC (2021) ‘Annual Report 2020-21’. Available at: https://www.ivcc.com/wp-content/uploads/2021/12/3547_IVCC_Annual-Report_A2_LR.pdf.

Jones, C.M., Liyanapathirana, M., Agossa, F.R., Weetman, D., Ranson, H., Donnelly, M.J. and Wilding, C.S. (2012) ‘Footprints of positive selection associated with a mutation (N1575Y) in the voltage-gated sodium channel of Anopheles gambiae’, *Proceedings of the National Academy of Sciences*, 109(17), pp. 6614–6619.

Kafy, H.T., Ismail, B.A., Mnzava, A.P., Lines, J., Abdin, M.S.E., Eltaher, J.S., Banaga, A.O., West, P., Bradley, J., Cook, J., Thomas, B., Subramaniam, K., Hemingway, J., Knox, T.B., Malik, E.M., *et al.* (2017) ‘Impact of insecticide resistance in *Anopheles arabiensis* on malaria incidence and prevalence in Sudan and the costs of mitigation’, *Proceedings of the National Academy of Sciences*, p. 201713814.

Katureebe, A., Zinszer, K., Arinaitwe, E., Rek, J., Kakande, E., Charland, K., Kigozi, R., Kilama, M., Nankabirwa, J., Yeka, A., Mawejje, H., Mpimbaza, A., Katamba, H., Donnelly, M.J., Rosenthal, P.J., *et al.* (2016) ‘Measures of Malaria Burden after Long-Lasting Insecticidal Net Distribution and Indoor Residual Spraying at Three Sites in Uganda: A Prospective Observational Study’, *PLoS medicine* [Preprint]. doi:10.1371/journal.pmed.1002167.

Kefi, M., Charamis, J., Balabanidou, V., Ioannidis, P., Ranson, H., Ingham, V.A. and Vontas, J. (2021) ‘Transcriptomic analysis of resistance and short-term induction response to pyrethroids, in Anopheles coluzzii legs’, *BMC genomics*, 22(1), p. 891.

Khambay, B.P.S. and Jewess, P.J. (2005) ‘Pyrethroids’, *Comprehensive Molecular Insect Science*, pp. 1–29.

Killeen, G.F. (2014) ‘Characterizing, controlling and eliminating residual malaria transmission’, *Malaria journal*, 13, p. 330.

Killeen, G.F. a. b. and Chitnis, N. c. d. e. (2014) ‘Potential causes and consequences of behavioural resilience and resistance in malaria vector populations: A mathematical modelling analysis’, *Malaria journal*, 13(1), pp. 1–16.

Killeen, G.F. and Ranson, H. (2018) ‘Insecticide-resistant malaria vectors must be tackled’, *The Lancet*, 391(10130), pp. 1551–1552.

Kleinschmidt, I., Bradley, J., Knox, T.B., Mnzava, A.P., Kafy, H.T., Mbogo, C., Ismail, B.A., Bigoga, J.D., Adechoubou, A., Raghavendra, K., Cook, J., Malik, E.M., Nkuni, Z.J., Macdonald, M., Bayoh, N., *et al.* (2018) ‘Implications of insecticide resistance for malaria vector control with long-lasting insecticidal nets: a WHO-coordinated, prospective, international, observational cohort study’, *The Lancet infectious diseases*, 18(6), pp. 640–649.

Köster, J. and Rahmann, S. (2012) ‘Snakemake-a scalable bioinformatics workflow engine’, *Bioinformatics* , 28(19), pp. 2520–2522.

Kwiatkowska, R.M., Platt, N., Poupardin, R., Irving, H., Dabire, R.K., Mitchell, S., Jones, C.M., Diabaté, A., Ranson, H. and Wondji, C.S. (2013) ‘Dissecting the mechanisms responsible for the multiple insecticide resistance phenotype in Anopheles gambiae s.s., M form, from Vallée du Kou, Burkina Faso’, *Gene*, 519(1), pp. 98–106.

Labbé, P., Berthomieu, A., Berticat, C., Alout, H., Raymond, M., Lenormand, T. and Weill, M. (2007) ‘Independent duplications of the acetylcholinesterase gene conferring insecticide resistance in the mosquito Culex pipiens’, *Molecular biology and evolution*, 24(4), pp. 1056–1067.

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., Farmer, C., Chinwalla, A., Yang, S.-P., Minx, P., Nelson, J., *et al.* (2010) ‘Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences’, *Science*, 330(6003), pp. 512–514.

Leffler, E.M., Bullaughey, K., Matute, D.R., Meyer, W.K., Ségurel, L., Venkat, A., Andolfatto, P. and Przeworski, M. (2012) ‘Revisiting an old riddle: what determines genetic diversity levels within species?’, *PLoS biology*, 10(9), p. e1001388.

Lehmann, T., Besansky, N.J., Hawley, W.A., Fahey, T.G., Kamau, L. and Collins, F.H. (1997) ‘Microgeographic structure of Anopheles gambiae in western Kenya based on mtDNA and microsatellite loci’, *Molecular ecology*, 6(3), pp. 243–253.

Lehmann, T., Hawley, W.A., Kamau, L., Fontenille, D., Simard, F. and Collins, F.H. (1996) ‘Genetic differentiation of Anopheles gambiae populations from East and west Africa: comparison of microsatellite and allozyme loci’, *Heredity*, 77 (Pt 2), pp. 192–200.

Lehmann, T., Licht, M., Elissa, N., Maega, B.T.A., Chimumbwa, J.M., Watsenga, F.T., Wondji, C.S., Simard, F. and Hawley, W.A. (2003) ‘Population Structure of Anopheles gambiae in Africa’, *The Journal of heredity*, 94(2), pp. 133–147.

Liu, N. (2015) ‘Insecticide resistance in mosquitoes: impact, mechanisms, and research directions’, *Annual review of entomology*, 60, pp. 537–559.

Lucas, E.R., Miles, A., Harding, N.J., Clarkson, C.S., Lawniczak, M.K.N., Kwiatkowski, D.P., Weetman, D., Donnelly, M.J. and Anopheles gambiae 1000 Genomes Consortium (2019) ‘Whole-genome sequencing reveals high complexity of copy number variation at insecticide resistance loci in malaria mosquitoes’, *Genome research*, 29(8), pp. 1250–1261.

Lucas, E.R., Rockett, K.A., Lynd, A., Essandoh, J., Grisales, N., Kemei, B., Njoroge, H., Hubbart, C., Rippon, E.J., Morgan, J., Van’t Hof, A., Ochomo, E.O., Kwiatkowski, D.P., Weetman, D. and Donnelly, M.J. (2019) ‘A high throughput multi-locus insecticide resistance marker panel for tracking resistance emergence and spread in Anopheles gambiae’, *bioRxiv* [Preprint], (March 2020). doi:10.1101/592279.

Lumjuan, N., McCarroll, L., Prapanthadara, L.-A., Hemingway, J. and Ranson, H. (2005) ‘Elevated activity of an Epsilon class glutathione transferase confers DDT resistance in the dengue vector, Aedes aegypti’, *Insect biochemistry and molecular biology*, 35(8), pp. 861–871.

Macdonald, G. (1952) ‘The analysis of the sporozoite rate’, *Tropical diseases bulletin*, 49(6), pp. 569–586.

Maharaj, R., Mthembu, D.J. and Sharp, B.L. (2005) ‘Impact of DDT re-introduction on malaria transmission in KwaZulu-Natal’, *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde*, 95(11), pp. 871–874.

Makunin, A., Korlević, P., Park, N., Goodwin, S., Waterhouse, R.M., von Wyschetzki, K., Jacob, C.G., Davies, R., Kwiatkowski, D., St Laurent, B., Ayala, D. and Lawniczak, M.K.N. (2022) ‘A targeted amplicon sequencing panel to simultaneously identify mosquito species and Plasmodium presence across the entire Anopheles genus’, *Molecular ecology resources*, 22(1), pp. 28–44.

MalariaGEN (2008) ‘A global network for investigating the genomic epidemiology of malaria’, *Nature*, 456(7223), pp. 732–737.

MalariaGEN (2023) ‘Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples’, *Wellcome Open Research*, 8(22). doi:10.12688/wellcomeopenres.18681.1.

MalariaGEN, Ahouidi, A., Ali, M., Almagro-Garcia, J., Amambua-Ngwa, A., Amaratunga, C., Amato, R., Amenga-Etego, L., Andagalu, B., Anderson, T.J.C., Andrianaranjaka, V., Apinjoh, T., Ariani, C., Ashley, E.A., Auburn, S., *et al.* (2021) ‘An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples’, *Wellcome open research*, 6, p. 42.

MalariaGEN, Band, G., Rockett, K.A., Spencer, C.C.A. and Kwiatkowski, D.P. (2015) ‘A novel locus of resistance to severe malaria in a region of ancient balancing selection’, *Nature*, 526(7572), pp. 253–257.

Manske, M., Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen, G., O’Brien, J., Djimde, A., Doumbo, O., Zongo, I., Ouedraogo, J.-B., Michon, P., Mueller, I., Siba, P., Nzila, A., *et al.* (2012) ‘Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing’, *Nature*, 487(7407), pp. 375–379.

Matthews, T.C. and Munstermann, L.E. (1994) ‘Chromosomal Repatterning and Linkage Group Conservation in Mosquito Karyotypic Evolution’, *Evolution; international journal of organic evolution*, 48(1), pp. 146–154.

Mattingly, P.F. (1977) ‘Names for the Anopheles gambiae Complex’, *Mosquito Systematics*, 9(3).

Meng, X., Yang, X., Zhang, N., Jiang, H., Ge, H., Chen, M., Qian, K. and Wang, J. (2019) ‘Knockdown of the GABA receptor RDL genes decreases abamectin susceptibility in the rice stem borer, Chilo suppressalis’, *Pesticide biochemistry and physiology*, 153, pp. 171–175.

Mercer, T.R. and Salit, M. (2021) ‘Testing at scale during the COVID-19 pandemic’, *Nature reviews. Genetics*, 22(7), pp. 415–426.

Michaelsen, T.Y., Bennedbæk, M., Christiansen, L.E., Jørgensen, M.S.F., Møller, C.H., Sørensen, E.A., Knutsson, S., Brandt, J., Jensen, T.B.N., Chiche-Lapierre, C., Collados, E.F., Sørensen, T., Petersen, C., Le-Quy, V., Sereika, M., *et al.* (2022) ‘Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark’, *Genome medicine*, 14(1), p. 47.

Michel, A.P., Ingrasci, M.J., Schemerhorn, B.J., Kern, M., Le Goff, G., Coetzee, M., Elissa, N., Fontenille, D., Vulule, J., Lehmann, T., Sagnon, N. ’f, Costantini, C. and Besansky, N.J. (2005) ‘Rangewide population genetic structure of the African malaria vector Anopheles funestus’, *Molecular ecology*, 14(14), pp. 4235–4248.

Miles, A. (2021) *Genomic epidemiology of malaria vectors in the Anopheles gambiae species complex*.

Miles, A., Clarkson, C., Hart, L., Murie, K., Kranjc, N., Bennett, K. and Nagi, S. (2022) *malariagen_data* [Python]. Available at: https://github.com/malariagen/malariagen-data-python.

Miles, A., Harding, N.J., Bottà, G., Clarkson, C.S., Antão, T., Kozak, K., Schrider, D.R., Kern, A.D., Redmond, S., Sharakhov, I., Pearson, R.D., Bergey, C., Fontaine, M.C., Donnelly, M.J., Lawniczak, M.K.N., *et al.* (2017) ‘Genetic diversity of the African malaria vector Anopheles gambiae’, *Nature*, 552, pp. 96–100.

Miotto, O., Almagro-Garcia, J., Manske, M., Macinnis, B., Campino, S., Rockett, K.A., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Anderson, J.M., Duong, S., Nguon, C., Chuor, C.M., Saunders, D., *et al.* (2013) ‘Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia’, *Nature genetics*, 45(6), pp. 648–655.

Miotto, O., Amato, R., Ashley, E.A., MacInnis, B., Almagro-Garcia, J., Amaratunga, C., Lim, P., Mead, D., Oyola, S.O., Dhorda, M., Imwong, M., Woodrow, C., Manske, M., Stalker, J., Drury, E., *et al.* (2015) ‘Genetic architecture of artemisinin-resistant Plasmodium falciparum’, *Nature genetics*, 47(3), pp. 226–234.

Mitchell, S.N., Rigden, D.J., Dowd, A.J., Lu, F., Wilding, C.S., Weetman, D., Dadzie, S., Jenkins, A.M., Regna, K., Boko, P., Djogbenou, L., Muskavitch, M.A.T., Ranson, H., Paine, M.J.I., Mayans, O., *et al.* (2014) ‘Metabolic and target-site mechanisms combine to confer strong DDT resistance in Anopheles gambiae’, *PloS one*, 9(3). doi:10.1371/journal.pone.0092662.

Mitchell, S.N., Stevenson, B.J., Muller, P., Wilding, C.S., Egyir-Yawson, A., Field, S.G., Hemingway, J., Paine, M.J.I., Ranson, H. and Donnelly, M.J. (2012) ‘Identification and validation of a gene causing cross-resistance between insecticide classes in Anopheles gambiae from Ghana’, *Proceedings of the National Academy of Sciences*, 109(16), pp. 6147–6152.

Moiroux, N., Gomez, M.B., Pennetier, C., Elanga, E., Djènontin, A., Chandre, F., Djègbé, I., Guis, H. and Corbel, V. (2012) ‘Changes in anopheles funestus biting behavior following universal coverage of long-lasting insecticidal nets in benin’, *The Journal of infectious diseases* [Preprint]. doi:10.1093/infdis/jis565.

Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S. and Köster, J. (2021) ‘Sustainable data analysis with Snakemake [version 1 ; peer review : awaiting peer review]’, pp. 1–17.

Msomi, N., Mlisana, K., de Oliveira, T., Msomi, N., Mlisana, K., Willianson, C., Bhiman, J.N., Goedhals, D., Engelbrecht, S., Van Zyl, G., Preiser, W., Hardie, D., Hsiao, M., Mulder, N., Martin, D., *et al.* (2020) ‘A genomics network established to respond rapidly to public health threats in South Africa’, *The Lancet Microbe*, 1(6), pp. e229–e230.

Müller, P., Warr, E., Stevenson, B.J., Pignatelli, P.M., Morgan, J.C., Steven, A., Yawson, A.E., Mitchell, S.N., Ranson, H., Hemingway, J., Paine, M.J.I. and Donnelly, M.J. (2008) ‘Field-caught permethrin-resistant Anopheles gambiae overexpress CYP6P3, a P450 that metabolises pyrethroids’, *PLoS genetics* [Preprint]. doi:10.1371/journal.pgen.1000286.

Nabarro, D.N. and Tayler, E.M. (1998) ‘The “Roll Back Malaria” Campaign’, *Science*, 280(5372), pp. 2067–2068.

Namuganga, J.F., Epstein, A., Nankabirwa, J.I., Mpimbaza, A., Kiggundu, M., Sserwanga, A., Kapisi, J., Arinaitwe, E., Gonahasa, S., Opigo, J., Ebong, C., Staedke, S.G., Shililu, J., Okia, M., Rutazaana, D., *et al.* (2021) ‘The impact of stopping and starting indoor residual spraying on malaria burden in Uganda’, *Nature communications*, 12(1), p. 2635.

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., Collins, F.H., Lander, E.S., Wirth, D.F., Kafatos, F.C., Besansky, N.J., *et al.* (2010) ‘SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes’, *Science*, 330(6003), pp. 514–517.

Njoroge, H., Van’t Hof, A., Oruni, A., Pipini, D., Nagi, S.C., Lynd, A., Lucas, E.R., Tomlinson, S., Grau-Bove, X., McDermott, D., Wat’senga, F.T., Manzambi, E.Z., Agossa, F.R., Mokuba, A., Irish, S., *et al.* (2022) ‘Identification of a rapidly-spreading triple mutant for high-level metabolic insecticide resistance in Anopheles gambiae provides a real-time molecular diagnostic for antimalarial intervention deployment’, *Molecular ecology*, 31(16), pp. 4307–4318.

Oakeshott, J.G., Claudianos, C., Campbell, P.M., Newcomb, R.D. and Russell, R.J. (2005) ‘Biochemical Genetics and Genomics of Insect Esterases’, *Comprehensive Molecular Insect Science*, 5-6, pp. 309–381.

Pages, F., Orlandi-Pradines, E. and Corbel, V. (2007) ‘Vecteurs du paludisme: biologie, diversité, contrôle et protection individuelle’, *Medecine et maladies infectieuses*, 37(3), pp. 153–161.

Pignatelli, P., Ingham, V.A., Balabanidou, V., Vontas, J., Lycett, G. and Ranson, H. (2018) ‘The Anopheles gambiae ATP-binding cassette transporter family: phylogenetic analysis and tissue localization provide clues on function and role in insecticide resistance’, *Insect molecular biology*, 27(1), pp. 110–122.

Protopopoff, N., Mosha, J.F., Lukole, E., Charlwood, J.D., Wright, A., Mwalimu, C.D., Manjurano, A., Mosha, F.W., Kisinza, W., Kleinschmidt, I. and Rowland, M. (2018) ‘Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions , separately and together , against malaria transmitted by pyrethroid-resistant mosquitoes : a cluster , randomised controlled , trial', *The Lancet*, 391(10130), pp. 1577–1588.

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L. and Pybus, O.G. (2020) 'A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology', *Nature Microbiology*, 5(11), pp. 1403–1407.

Ranson, H., Claudianos, C., Ortelli, F., Abgrall, C., Hemingway, J., Sharakhova, M.V., Unger, M.F., Collins, F.H. and Feyereisen, R. (2002) 'Evolution of supergene families associated with insecticide resistance', *Science*, 298(5591), pp. 179–181.

Ranson, H. and Lissenden, N. (2016) 'Insecticide Resistance in African Anopheles Mosquitoes: A Worsening Situation that Needs Urgent Action to Maintain Malaria Control', *Trends in Parasitology*, pp. 187–196. doi:10.1016/j.pt.2015.11.010.

Ranson, H., N'Guessan, R., Lines, J., Moiroux, N., Nkuni, Z. and Corbel, V. (2011) 'Pyrethroid resistance in African anopheline mosquitoes: What are the implications for malaria control?', *Trends in parasitology*, 27(2), pp. 91–98.

Ranson, H., Rossiter, L., Ortelli, F., Jensen, B., Wang, X., Roth, C.W., Collins, F.H. and Hemingway, J. (2001) 'Identification of a novel class of insect glutathione S-transferases involved in resistance to DDT in the malaria vector Anopheles gambiae', *The Biochemical journal*, 359(Pt 2), pp. 295–304.

Raymond, M., Chevillon, C., Guillemaud, T., Lenormand, T. and Pasteur, N. (1998) 'An overview of the evolution of overproduced esterases in the mosquito Culex pipiens', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 353(1376), pp. 1707–1711.

Raymond, M., Qiao, C.L. and Callaghan, A. (1996) 'Esterase polymorphism in insecticide susceptible populations of the mosquito Culex pipiens', *Genetical research*, 67(1), pp. 19–26.

Reddy, M.R., Overgaard, H.J., Abaga, S., Reddy, V.P., Caccone, A., Kiszewski, A.E. and 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, p. 184.

Reidenbach, K.R., Neafsey, D.E., Costantini, C., Sagnon, N. 'fale, Simard, F., Ragland, G.J., Egan, S.P., Feder, J.L., Muskavitch, M.A.T. and 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(12), pp. 1202–1212.

Reinecke, R., Trautmann, T., Wagener, T. and Schüler, K. (2022) 'The critical need to foster computational reproducibility', *Environmental research letters: ERL [Web site]*, 17(4), p. 041005.

Remnant, E.J., Good, R.T., Schmidt, J.M., Lumb, C., Robin, C., Daborn, P.J. and Batterham, P. (2013) 'Gene duplication in the major insecticide target site, *Rdl*, in *Drosophila melanogaster*', *Proceedings of the National Academy of Sciences of the United States of America*, 110(36), pp. 14705–14710.

Riveron, J.M., Ibrahim, S.S., Chanda, E., Mzilahowa, T., Cuamba, N., Irving, H., Barnes, K.G., Ndula, M. and Wondji, C.S. (2014) 'The highly polymorphic CYP6M7 cytochrome P450 gene partners with the directionally selected CYP6P9a and CYP6P9b genes to expand the pyrethroid resistance front in the malaria vector Anopheles funestus in Africa', *BMC genomics*, 15(1), p. 817.

Robishaw, J.D., Alter, S.M., Solano, J.J., Shih, R.D., DeMets, D.L., Maki, D.G. and Hennekens, C.H.

(2021) ‘Genomic surveillance to combat COVID-19: challenges and opportunities’, *The Lancet. Microbe*, 2(9), pp. e481–e484.

Roll Back Malaria (2020) *2 billion mosquito nets delivered worldwide since 2004*. Available at: https://endmalaria.org/news/2-billion-mosquito-nets-delivered-worldwide-2004 (Accessed: 24 January 2023).

Roush, R. and Tabashnik, B.E. (2012) *Pesticide Resistance in Arthropods*. Springer Science & Business Media.

Sangbakembi-Ngounou, C., Costantini, C., Longo-Pendy, N.M., Ngoagouni, C., Akone-Ella, O., Rahola, N., Cornelie, S., Kengne, P., Nakouné, E.R., Komas, N.P. and Ayala, D. (2022) ‘Diurnal biting of malaria mosquitoes in the Central African Republic indicates residual transmission may be “out of control”’, *Proceedings of the National Academy of Sciences*, 119(21), p. e2104282119.

Sato, S. (2021) ‘Plasmodium-a brief introduction to the parasites causing human malaria and their basic biology’, *Journal of physiological anthropology*, 40(1), p. 1.

Sharakhova, M.V., Hammond, M.P., Lobo, N.F., Krzywinski, J., Unger, M.F., Hillenmeyer, M.E., Bruggner, R.V., Birney, E. and Collins, F.H. (2007) ‘Update of the Anopheles gambiae PEST genome assembly’, *Genome biology*, 8(1), p. R5.

Sherrard-Smith, E., Ngufor, C., Sanou, A., Guelbeogo, M.W., N’Guessan, R., Elobolobo, E., Saute, F., Varela, K., Chaccour, C.J., Zulliger, R., Wagman, J., Robertson, M.L., Rowland, M., Donnelly, M.J., Gonahasa, S., *et al.* (2022) ‘Inferring the epidemiological benefit of indoor vector control interventions against malaria from mosquito data’, *Nature communications*, 13(1), p. 3862.

Simma, E.A., Dermauw, W., Balabanidou, V., Snoeck, S., Bryon, A., Clark, R.M., Yewhalaw, D., Vontas, J., Duchateau, L. and Van Leeuwen, T. (2019) ‘Genome-wide gene expression profiling reveals that cuticle alterations and P450 detoxification are associated with deltamethrin and DDT resistance in Anopheles arabiensis populations from Ethiopia’, *Pest management science*, 75(7), pp. 1808–1818.

Small, S.T., Labbé, F., Lobo, N.F., Koekemoer, L.L., Sikaala, C.H., Neafsey, D.E., Hahn, M.W., Fontaine, M.C. and Besansky, N.J. (2020) ‘Radiation with reticulation marks the origin of a major malaria vector’, *Proceedings of the National Academy of Sciences of the United States of America*, 117(50), pp. 31583–31590.

Sougoufara, S., Diédhiou, S.M., Doucouré, S., Diagne, N., Sembène, P.M., Harry, M., Trape, J.-F., Sokhna, C. and Ndiath, M.O. (2014) ‘Biting by Anopheles funestus in broad daylight after use of long-lasting insecticidal nets: a new challenge to malaria elimination’, *Malaria journal*, 13, p. 125.

Sparks, T.C. and Nauen, R. (2015) ‘IRAC: Mode of action classification and insecticide resistance management’, *Pesticide biochemistry and physiology*, 121, pp. 122–128.

Staedke, S.G., Gonahasa, S., Dorsey, G., Kamya, M.R., Maiteki-Sebuguzi, C., Lynd, A., Katureebe, A., Kyohere, M., Mutungi, P., Kigozi, S.P., Opigo, J., Hemingway, J. and Donnelly, M.J. (2020) ‘Effect of long-lasting insecticidal nets with and without piperonyl butoxide on malaria indicators in Uganda (LLINEUP): a pragmatic, cluster-randomised trial embedded in a national LLIN distribution campaign’, *The Lancet*, 395(10232), pp. 1292–1303.

Sun, L., Cui, L., Rui, C., Yan, X., Yang, D. and Yuan, H. (2012) ‘Modulation of the expression of ryanodine receptor mRNA from Plutella xylostella as a result of diamide insecticide application’, *Gene*, 511(2), pp. 265–273.

Syafruddin, D., Hidayati, A.P.N., Asih, P.B.S., Hawley, W.A., Sukowati, S. and Lobo, N.F. (2010) ‘Detection of 1014F kdr mutation in four major Anopheline malaria vectors in Indonesia’, *Malaria journal*, 9, p. 315.

Tabashnik, B.E. (1989) ‘Managing Resistance with Multiple Pesticide Tactics: Theory, Evidence, and Recommendations’, *Journal of economic entomology*, 82(5), pp. 1263–1269.

Tangena, J.-A.A., Hendriks, C.M.J., Devine, M., Tammaro, M., Trett, A.E., Williams, I., DePina, A.J., Sisay, A., Herizo, R., Kafy, H.T., Chizema, E., Were, A., Rozier, J., Coleman, M. and Moyes, C.L. (2020) ‘Indoor residual spraying for malaria control in sub-Saharan Africa 1997 to 2017: an adjusted retrospective analysis’, *Malaria journal*, 19(1), p. 150.

Tennessen, J.A., Ingham, V.A., Toé, K.H., Guelbéogo, W.M., Sagnon, N. ’falé, Kuzma, R., Ranson, H. and Neafsey, D.E. (2021) ‘A population genomic unveiling of a new cryptic mosquito taxon within the malaria-transmitting Anopheles gambiae complex’, *Molecular ecology*, 30(3), pp. 775–790.

The Alliance for Malaria Prevention (2022) ‘Net Mapping Project, Current ITN Global Delivery Quarterly Report, Q3 2022’, *The Alliance for Malaria Prevention* [Preprint]. Available at: https://allianceformalariaprevention.com/itn-dashboards/net-mapping-project/ (Accessed: 11 January 2023).

della Torre, A., Fanello, C., Akogbeto, M., Dossou-yovo, J., Favia, G., Petrarca, V. and Coluzzi, M. (2001) ‘Molecular evidence of incipient speciation within Anopheles gambiae s.s. in West Africa’, *Insect molecular biology*, 10(1), pp. 9–18.

Unwin, H.J.T., Smith, E.S., Churcher, T.S. and Ghani, A.C. (2022) ‘Quantifying the direct and indirect protection provided by insecticide treated bed nets against malaria’, *medRxiv*. doi:10.1101/2022.01.21.22269650.

Vontas, J., Grigoraki, L., Morgan, J., Tsakireli, D., Fuseini, G., Segura, L., Niemczura de Carvalho, J., Nguema, R., Weetman, D., Slotman, M.A. and Hemingway, J. (2018) ‘Rapid selection of a pyrethroid metabolic enzyme CYP9K1 by operational malaria control activities’, *Proceedings of the National Academy of Sciences*, (21), p. 201719663.

Wang, J., Zhao, X., Yan, R., Wu, S., Wu, Y. and Yang, Y. (2020) ‘Reverse genetics reveals contrary effects of two Rdl-homologous GABA receptors of Helicoverpa armigera on the toxicity of cyclodiene insecticides’, *Pesticide biochemistry and physiology*, 170, p. 104699.

Wang, L., Nomura, Y., Du, Y., Liu, N., Zhorov, B.S. and Dong, K. (2015) ‘A mutation in the intracellular loop III/IV of mosquito sodium channel synergizes the effect of mutations in helix IIS6 on pyrethroid resistance’, *Molecular pharmacology*, 87(3), pp. 421–429.

Wang, Q., Wang, H., Zhang, Y., Chen, J., Upadhyay, A., Bhowmick, B., Hang, J., Wu, S., Liao, C. and Han, Q. (2022) ‘Functional analysis reveals ionotropic GABA receptor subunit RDL is a target site of ivermectin and fluralaner in the yellow fever mosquito, Aedes aegypti’, *Pest management science*, 78(10), pp. 4173–4182.

Weetman, D., Djogbenou, L.S. and Lucas, E. (2018) ‘Copy number variation (CNV) and insecticide resistance in mosquitoes: evolving knowledge or an evolving problem?’, *Current Opinion in Insect Science*, 27, pp. 82–88.

Weill, M., Lutfalla, G., Mogensen, K., Chandre, F., Berthomieu, A., Berticat, C., Pasteur, N., Philips, A., Fort, P. and Raymond, M. (2003) 'Comparative genomics: Insecticide resistance in mosquito vectors', *Nature*, 423(6936), pp. 136–137.

WHO (2012) 'Global plan for insecticide resistance management in malaria vectors', *World Health Organization press*, p. 13.

WHO (2022) *Prequalified Vector Control Products*, *WHO - Prequalification of Medical Products (IVDs, Medicines, Vaccines and Immunization Devices, Vector Control)*. Available at: https://extranet.who.int/pqweb/vector-control-products/prequalified-product-list (Accessed: 24 January 2023).

Wiebe, A., Longbottom, J., Gleave, K., Shearer, F.M., Sinka, M.E., Massey, N.C., Cameron, E., Bhatt, S., Gething, P.W., Hemingway, J., Smith, D.L., Coleman, M. and Moyes, C.L. (2017) 'Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance', *Malaria journal*, 16(1), p. 85.

Williams, J., Cowlishaw, R., Sanou, A., Ranson, H. and Grigoraki, L. (2022) '*In vivo* functional validation of the V402L voltage gated sodium channel mutation in the malaria vector *An. gambiae*', *Pest Management Science*, pp. 1155–1163. doi:10.1002/ps.6731.

World Health Organization (2022) *World malaria report 2022*. World Health Organization.

Yahouédo, G.A., Chandre, F., Rossignol, M., Ginibre, C., Balabanidou, V., Mendez, N.G.A., Pigeon, O., Vontas, J. and Cornelie, S. (2017) 'Contributions of cuticle permeability and enzyme detoxification to pyrethroid resistance in the major malaria vector Anopheles gambiae', *Scientific reports*, 7(1), p. 11091.

Yunta, C., Grisales, N., Nasz, S., Hemmings, K., Pignatelli, P., Voice, M., Ranson, H. and Paine, M.J.I. (2016) 'Pyriproxyfen is metabolized by P450s associated with pyrethroid resistance in An. gambiae', *Insect biochemistry and molecular biology*, 78, pp. 50–57.

Yunta, C., Hemmings, K., Stevenson, B., Koekemoer, L.L., Matambo, T., Pignatelli, P., Voice, M., Nász, S. and Paine, M.J.I. (2019) 'Cross-resistance profiles of malaria mosquito P450s associated with pyrethroid resistance against WHO insecticides', *Pesticide biochemistry and physiology* [Preprint]. doi:10.1016/j.pestbp.2019.06.007.

Zhao, S., Fung-Leung, W.P., Bittner, A., Ngo, K. and Liu, X. (2014) 'Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells', *PloS one*, 9(1). doi:10.1371/journal.pone.0078644.

Zheng, L., Benedict, M.Q., Cornel, A.J., Collins, F.H. and Kafatos, F.C. (1996) 'An integrated genetic map of the African human malaria vector mosquito, Anopheles gambiae', *Genetics*, 143(2), pp. 941–952.

Zlotkin, E. (1999) 'The insect voltage-gated sodium channel as target of insecticides', *Annual review of entomology*, 44, pp. 429–455.

2

RNA-Seq-Pop

This chapter has now been published in Molecular Ecology Resources.

DOI:10.1111/1755-0998.13759

RNA-Seq-Pop: Exploiting the sequence in RNA-Seq - a Snakemake workflow reveals patterns of insecticide resistance in the malaria vector *Anopheles gambiae*

Sanjay C Nagi[1*], Ambrose Oruni[2], David Weetman[1], Martin J Donnelly[1]

[1]Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, UK.

[2]Uganda Virus Research Institute, Entebbe 256, Uganda

[*]Corresponding author: Email: Sanjay.Nagi@lstmed.ac.uk

Keywords: Transcriptomics, Insecticide resistance, Anopheles, Snakemake, population genetics

2.1 Abstract

We provide a reproducible and scalable Snakemake workflow, called RNA-Seq-Pop, which provides end-to-end analysis of RNA-Seq data sets. The workflow allows the user to perform quality control, differential expression analyses, and call genomic variants. Additional options include the calculation of allele frequencies of variants of interest, summaries of genetic variation and population structure, and genome wide selection scans, together with clear visualisations. RNA-Seq-Pop is applicable to any organism, and we demonstrate the utility of the workflow by investigating pyrethroid resistance in selected strains of the major malaria mosquito, Anopheles gambiae. The workflow provides additional modules specifically for *An. gambiae*, including estimating recent ancestry and determining the karyotype of common chromosomal inversions.

The Busia lab-colony used for selections was collected in Busia, Uganda, in November 2018. We performed a comparative analysis of three groups: a parental G24 Busia strain; its deltamethrin-selected G28 offspring; and the susceptible reference strain Kisumu. Measures of genetic diversity reveal patterns consistent with that of laboratory colonisation and selection, with the parental Busia strain exhibiting the highest nucleotide diversity, followed by the selected Busia offspring, and finally, Kisumu. Differential expression and variant analyses reveal that the selected Busia colony exhibits a number of distinct mechanisms of pyrethroid resistance, including the *Vgsc*-995S target-site mutation, upregulation of SAP genes, P450s, and a cluster of carboxylesterases. During deltamethrin selections, the 2La chromosomal inversion rose in frequency (from 33% to 86%), supporting a previous link with pyrethroid resistance. RNA-Seq-Pop is hosted here github.com/sanjaynagi/rna-seq-pop. We anticipate that the workflow will provide a useful tool to facilitate reproducible, transcriptomic studies in *An. gambiae* and other taxa.

2.2 Introduction

Transcriptomics is central to our understanding of how genetic variation influences phenotype (Stark et al., 2019). In recent years, RNA-Sequencing has replaced microarray technologies for whole-transcriptome profiling, providing a relatively unbiased view of transcript expression (Zhao et al., 2014) with associated higher sensitivity and greater dynamic range (Lowe et al., 2017). The utility of RNA-seq is exemplified by the vast amounts of data accruing (Van den Berge et al., 2019), and in the many discoveries it has revealed – such as the extent of alternative splicing, and the biology of non-coding RNAs (Stark et al., 2019; Wang et al., 2010; Wang & Burge, 2008).

In recent years, various computational workflows have been developed to analyse RNA-Seq data in a reproducible manner (Lataretu & Hölzer, 2020; Zhang & Jonassen, 2019), however, these workflows are designed with the primary aim of differential expression analysis (DEA) and leave a large amount of untapped sequence-based information. A study previously detected population genomic signals in RNA-sequencing data, however, this specific application remains rare (Thorstensen et al., 2020). In our own area of research, vector genomics, a scan of the literature revealed thirty-three RNA-Sequencing studies (supplementary table 1), of which only five interrogated the sequence data (Bonizzoni et al., 2015; David et al., 2014; Faucon et al., 2017; Kang et al., 2021; Messenger et al., 2021). A barrier to exploiting the full range of information contained within RNA-Seq data sets has been the absence of comprehensive, user-friendly pipelines which permit easily reproducible analysis (Grüning et al., 2018) and enable comparisons across studies.

In this study, using the workflow management system Snakemake (Mölder et al., 2021), we present a reproducible computational workflow, RNA-Seq-Pop, for the analysis of short-read RNA-Sequencing datasets of any organism. The workflow is applicable to single or paired-end RNA-Sequencing data, such as those produced on Illumina or MGI (DNB-Seq) instruments. However, we also present modules specifically of interest in the analysis of the major malaria mosquito, *Anopheles gambiae s.l.*, and demonstrate their use in a study of pyrethroid-resistance in a strain of *An. gambiae* from Busia, Uganda.

Pyrethroids are the most widely used class of insecticide in malaria control, and over the past two decades, resistance in malaria vectors has spread throughout sub-Saharan Africa, posing a threat to vector control efforts (Ranson, 2017). In this period, the incrimination of genes involved in insecticide-resistant phenotypes of Anopheles gambiae has been primarily based on transcriptomic studies. For many years, these were performed using microarrays; synthesis of which has highlighted the repeatable overexpression of a handful of genes involved in detoxification, confirming well-established cytochrome P450s as candidates, whilst also implicating more diverse genes such as ABC transporters and sensory appendage proteins (Ingham et al., 2018). Yet to date, relatively few diagnostic markers have been identified, and important genes have been missed by standard transcriptomic analyses (Njoroge et al., 2021). These shortcomings illustrate the need for a more comprehensive approach to marker discovery. While whole-genome sequencing is providing valuable information on known and novel resistance variants (Clarkson et al., 2021; The Anopheles gambiae 1000 Genomes Consortium, 2020) exploiting the sequence data within RNA-Seq can help bridge the step from transcriptomics to genomics.

In Uganda, pyrethroid resistance has escalated in recent years (Lynd et al., 2019; Tchouakui et al., 2021). As well as the *Vgsc*-995S mutation, which has repeatedly been associated with pyrethroid-resistance, recent genomic studies from this region have shown that a triple-mutant haplotype, linking a transposable element, a gene duplication (Cyp6aa1) and a non-synonymous mutation Cyp6p4-I236M, is an important marker of pyrethroid resistance (Njoroge et al., 2021). A SNP-array based GWAS also demonstrated the Cyp4J5-L43F mutation to be a useful marker for insecticide resistance, whilst also implicating the 2La inversion karyotype as a potential marker (Weetman et al., 2018). We use RNA-Seq-Pop to uncover patterns of insecticide resistance in Ugandan *An. gambiae*, monitoring these resistance-associated mutations, whilst performing differential expression analyses, summarising genetic variation and ancestry, and karyotyping chromosomal inversions.

2.3 Materials & Methods

2.3.1 RNA-Seq-Pop implementation

We designed the RNA-Seq-Pop workflow according to Snakemake best practices (Köster, 2022). RNA-Seq-Pop is constructed with a single configuration file in human-readable yaml format (the config file), alongside a simple tab-separated text file containing sample metadata (the sample sheet). The overall RNA-Seq-Pop workflow is shown in figure 1.

Dependencies are internally managed by the package manager Conda; to install all required software, specify the --use-conda directive at the command line, and Conda will automatically create isolated software environments in which to run. As of v1.0.4, RNA-Seq-Pop modules are written in Python (81.2% of the codebase) and R (18.8%), and internally, the workflow utilises a library (RNASeqPopTools) which providethe infrastructure to the Python codebase, to ensure readability. We provide documentation to guide users on how to set up and run RNA-Seq-Pop, located here https://sanjaynagi.github.io/rna-seq-pop.

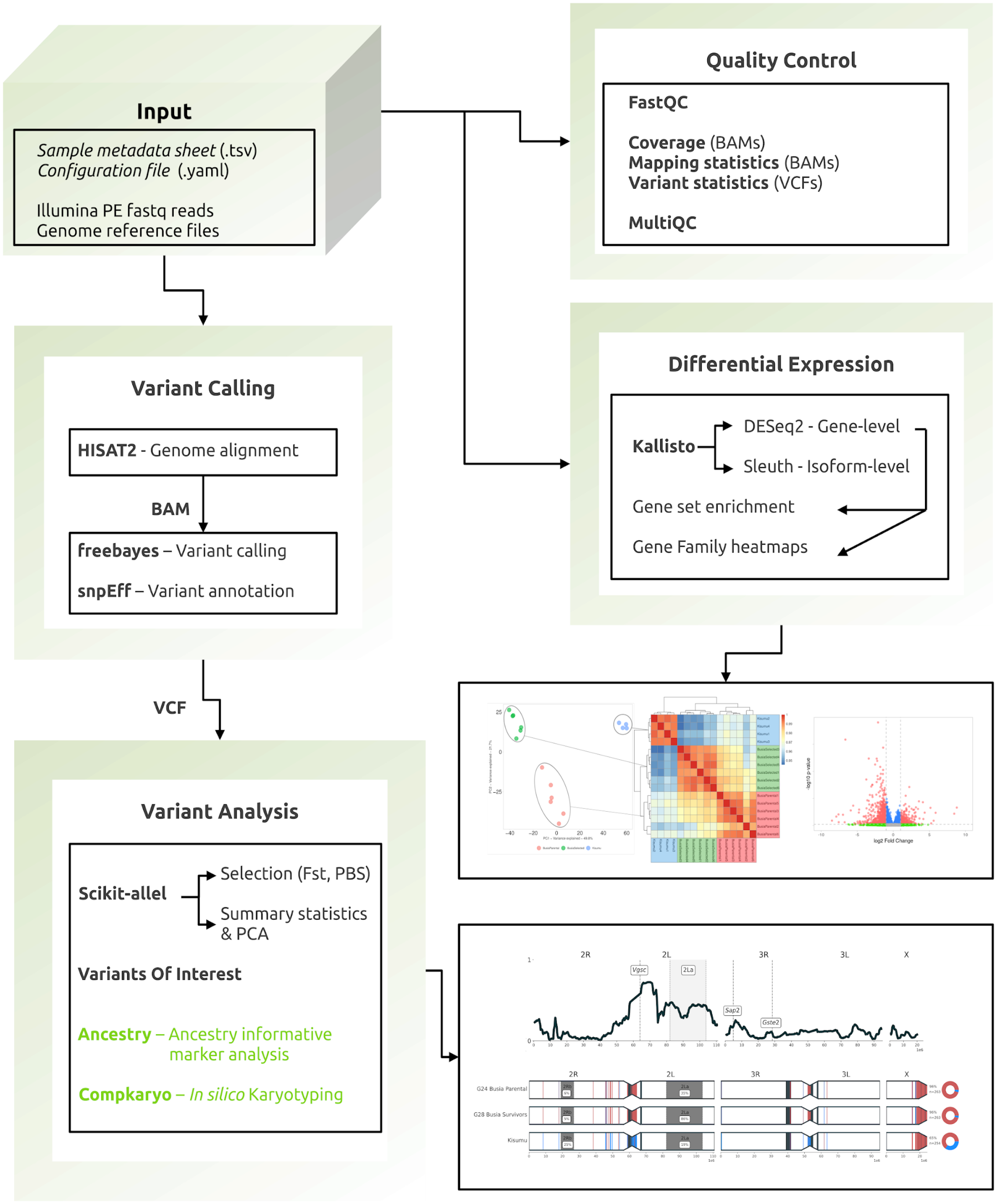


Figure 1: The *RNA-Seq-Pop* workflow and example outputs. The workflow has been designed for ease of use, requiring only a configuration file to set up workflow choices and a sample sheet to provide sample metadata. Modules highlighted in green are specific to *An. gambiae s.l.*

2.3.2 Quality Control

The workflow begins by checking concordance between the user-provided sample metadata, configuration file, and reference and fastq files. Quality control metrics of fastq files are calculated with fastqc (Andrews, 2010), and logs and statistics from eight tools in the workflow are integrated into a report with MultiQC (Ewels et al., 2016). Raw fastq reads may be optionally trimmed with cutadapt (Martin, 2011), with the option of specifiying custom adaptor sequences.

2.3.3 Differential expression

Trimmed reads are aligned to the reference transcriptome with Kallisto (Bray et al., 2015) and differential expression performed at the gene-level with DESeq2 (Love et al., 2014) and at the isoform-level with sleuth (Pimentel et al., 2017). The gene-level counts are normalised to account for sequencing depth, and principal components analysis (PCA) and Pearson's correlation performed among all samples, and on subsets of the user-selected treatment groups used in differential expression analysis. Plots of these analyses are useful for exploratory data visualisation, providing an additional quality control step to ensure expected relationships between samples. RNA-Seq-Pop combines differential expression results from multiple pairwise comparisons into an Excel spreadsheet for the user, as well as generating individual .csv files, volcano plots, and identifying the number of differentially expressed genes at various FDR-adjusted p-value thresholds. The user may create venn diagrams for multiple comparisons and generate heatmaps if a list of genes is provided. We use the hypergeometric test for GO term enrichment analysis, on genes that are significantly over-expressed based on FDR-adjusted p-values and, and optionally, the top 5 percentile of FST values.

2.3.4 Variant calling

Reads are aligned to the reference genome with HISAT2 (Kim et al., 2019) and read duplicates marked with samblaster (Faust & Hall, 2014) producing binary alignment files (BAM), which are sorted by genomic coordinate and indexed with SAMtools v1.19

(Danecek et al., 2021). SNPs are then called with the Bayesian haplotype-based caller freebayes v1.3.2 (Garrison & Marth, 2012). SNPs are called jointly on all samples, with different treatment groups called as separate populations, at the ploidy level provided by the user in the configuration file. The workflow internally parallelises freebayes by splitting the genome into small regions, greatly reducing overall computation time. The separated genomic regions are then concatenated with bcftools v1.19 (Danecek et al., 2021) and the final VCF piped through vcfuniq (Garrison et al., 2021), to filter out any duplicate calls that may occur at the genomic intervals between chunks. Called variants are then annotated using snpEff v5.0 (Cingolani et al., 2012).

2.3.5 Variant analysis & selection

RNA-Seq-Pop can then perform analyses on the variants called by freebayes. We apply filters to the data, including restricting to SNPs (excluding indel calls) and applying missingness and quality filters. We recommend using a quality score of 30 and a missingness proportion of 1, meaning a variant call (reference or alternate allele) must be present in each sample, i.e there are no missing allele calls. For each pairwise comparison specified in the config file, the workflow can perform a windowed Hudson's FST scan (Bhatia et al., 2013; Hudson et al., 1992) along each chromosomal arm, outputting windowed FST estimates and genome-wide plots. Population branch statistic (PBS) scans may also be performed, conditional on the presence of three suitable populations for the phenotype(s) of interest (Yi et al., 2010). It is also possible to run Hudson's FST and PBS scans, taking the average for each protein-coding gene, rather than in windows. All population genetic statistics are calculated in scikit-allel v1.2.1. (Miles & Harding, 2017). We also provide a script (geneScan.py) to interrogate the VCF files, reporting missense variants from any gene of the user's choice. A tab-separated file of variants of interest can be provided, from which the workflow will produce allele frequency heatmaps for each biological replicate and averaged across treatment groups. We define the expressed allele balance as the allele frequency at a genomic location in the aligned read data – for this analysis, RNA-Seq-Pop does not use variants called by freebayes, but instead calculates the proportion of each allele directly in bam files to ease intepretation. An example variant of interest file for *An. gambiae* is provided in the RNA-Seq-Pop GitHub repository.

All analyses described thus far can be conducted across all eukaryotes of any ploidy, requiring only a reference genome (.fa), transcriptome (.fa), and genome annotation files (.gff3).

2.3.6 Anopheles gambiae s.l specific analyses

For Anopheles gambiae s.l datasets we have exploited the Anopheles gambiae 1000 genomes resource (Miles et al., 2017; The Anopheles gambiae 1000 Genomes Consortium, 2020), to incorporate H12 and iHS (Garud et al., 2015) genomic selective sweep analysis. The workflow outputs the differentially expressed gene's genomic location, the specific sweep signals present in the Ag1000g resource at that genomic location, and whether the region is a known insecticide resistance-associated locus. We caution that this kind of analysis is exploratory: many genes will be contained within selective sweeps, and may not have a causal link to phenotypic variation.

2.3.7 Population structure, ancestry and karyotyping

To investigate population structure, we apply SNP quality and missingness filters to the SNP data, which can be configured by the user. Multiple measures of population genetic diversity are estimated for each sample, such as nucleotide diversity (π), Watterson's θ (Watterson, 1975), and inbreeding coefficients. We then prune SNPs in high linkage by excluding variants above an R2 threshold of 0.01 in sliding windows of 500 SNPs with a step size of 250 SNPs, and perform a PCA on the remaining SNPs. If the analysed species is *An. gambiae*, An. coluzzii, or An. arabiensis, the pipeline can implement an analysis of putative ancestry informative markers (AIMs). The AIMs were derived from two different datasets. The *An. gambiae/An. coluzzii* AIMs derive from the 16 genomes project (Neafsey et al., 2015) and in West Africa may distinguish between individuals with *An. gambiae* or *An. coluzzii* ancestry. The *An. gambcolu/An. arabiensis* AIMs are derived from phase 3 of the Anopheles gambiae 1000 genomes project, and distinguish between individuals with either *An. gambiae* or *An. coluzzii* ancestry from *An. arabiensis*. The relative proportion of ancestry is reported and visualised for the whole genome by chromosome. We modified the program compkaryo (Love et al., 2019) to enable the identification of common inversions on chromosome 2 in pooled samples.

2.3.8 Busia RNA-Seq

2.3.8.1 Mosquito lines

As a case-study to the workflow, we sequenced a pyrethroid-resistant colony of Anopheles gambiae s.s from Busia, Uganda, alongside the standard multi-insecticide-susceptible reference strain, Kisumu. After 24 generations in colony, we stored RNA from unexposed Busia individuals (G24 Busia parental). Unexposed, age-matched Kisumu females were used as controls. We then selected the remaining Busia G24 colony using 0.05% deltamethrin papers in WHO tube assays for 4 generations (full details of the selection regime can be found in the supplementary text 2). We exposed females from the selected generation (G28) for one hour to 0.05% deltamethrin WHO papers using standard protocols, left for 24 hours post-exposure, and survivors were stored at -80°C prior to RNA extraction (G28 Busia selected survivors). Selections were perfomed post-mating.

2.3.8.2 Library prep

We extracted RNA from pools of five, 4-day old female mosquitoes using a Picopure RNA isolation kit (Arcturus, Applied Biosystems, USA). We performed six replicates for each Busia-derived treatment group, and four for Kisumu. Library quality and quantity were determined on a Tapestation 2200 (Agilent, UK) using high sensitivity RNA screentape. Paired-end 150bp RNA-Sequencing libraries were prepared and sequenced by Novogene (https://en.novogene.com/), on an Illumina NovaSeq 6000 system.

2.4 Results

2.4.1 Busia resistance phenotyping

The parental G24 *An. gambiae* Busia strain had lost much of its pyrethroid resistance during the time in culture and exhibited susceptibility to deltamethrin (100% mortality, 96.3-100 95% CIs) and low-prevalence resistance to permethrin (92.6% mortality, 85.6-96.4 95% CIs). Four generations of deltamethrin selections, demonstrated this loss to be readily reversible and resulted in a G28 selected Busia strain that showed increased resistance to both deltamethrin (69.7% mortality, 63.2-75.6 95% CIs) and permethrin (21.7% mortality, 14.9-30.5 95% CIs) when exposed for one hour in WHO tube assays. Further colony information can be found in supplementary table 2. We compared the G24 Busia Parental strain with the G28 Busia selected strain, and both Busia strains to the pyrethroid-susceptible reference strain, Kisumu.

2.4.2 RNA-Sequencing

As an illustrative example of the modules and output of the RNA-Seq-Pop workflow, we will describe the analysis of the Busia RNA-Seq dataset.

2.4.3 Quality control

We used RNA-Seq-Pop to import FASTQ data files into FastQC (Andrews, 2010) to determine levels of adaptor content, quality scores, sequence duplication levels and GC content in the raw read data. After genome alignment, we applied rseqQC and SAMtools to collect mapping statistics from the resulting BAM files. We then integrated MultiQC into the workflow, which collates statistics and results from eight tools to generate a convenient, interactive (.html) quality control report. Figure 2 shows reports generated by multiQC on the Busia *An. gambiae* dataset.

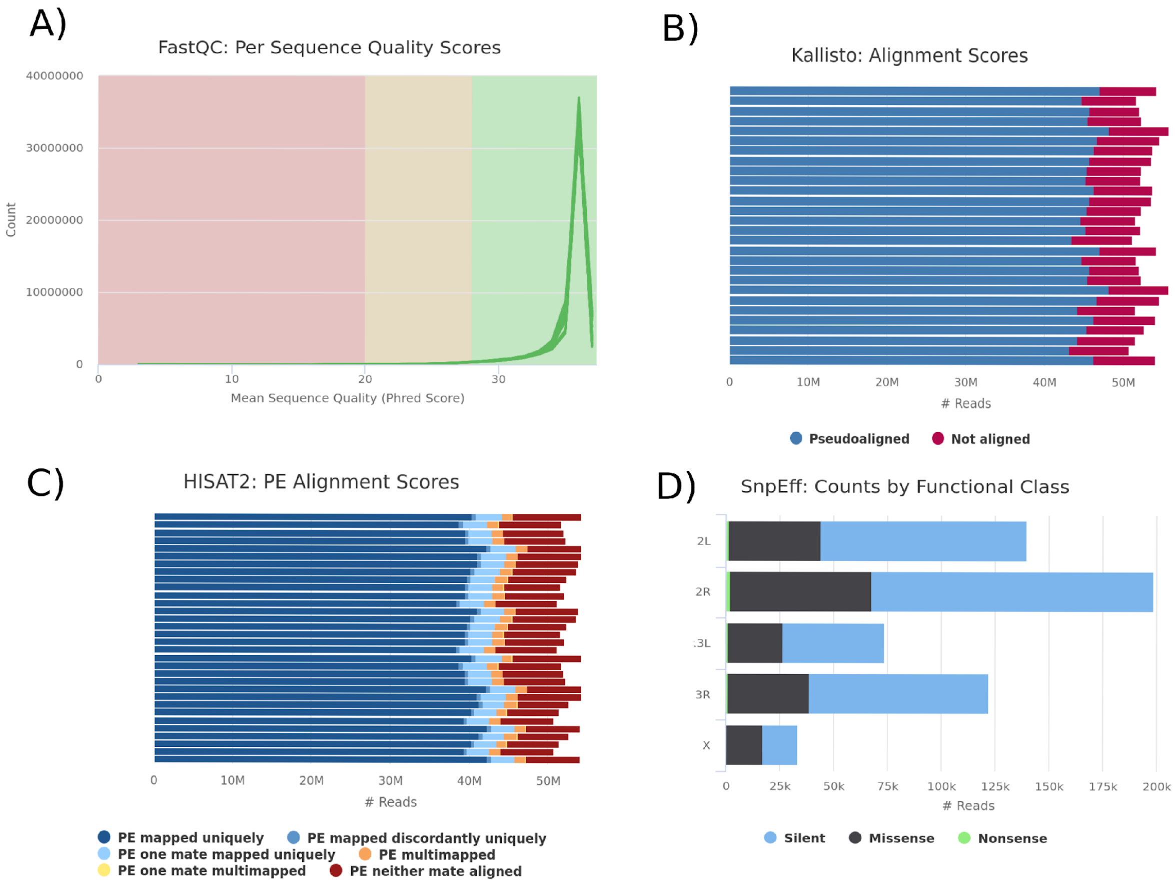


Figure 2: MultiQC captures quality control statistics from across the *RNA-Seq-Pop* workflow. a) per-base sequence content as calculated by FASTQC b) Total reads and number of successfully aligned reads to the reference transcriptome by Kallisto. c) The number of reads that were successfully mapped to the reference genome with HISAT2 d) The proportion of missense, synonymous and nonsense SNPs reported by snpEff.

We removed adapter sequences and aligned paired-end reads to the Anopheles gambiae PEST reference transcriptome (AgamP4.12) (Figure 2b). 844.25 million reads were processed in total, with 727.84 million successfully aligned, giving an overall 85.58% alignment rate (+/- 0.206% standard error) across sixteen total replicates. The breakdown of reads counted per sample can be found in supplementary Figure 3.

As a further quality control step, and to uncover the overarching relationships of gene expression between samples, RNA-Seq-Pop performs a principal components analysis (Figure 3a), and a sample-to-sample correlation heatmap (Figure 3b) on the DESeq2 normalised read count data. In both analyses, biological replicates of each treatment group clustered together, supporting robust replication in these samples.

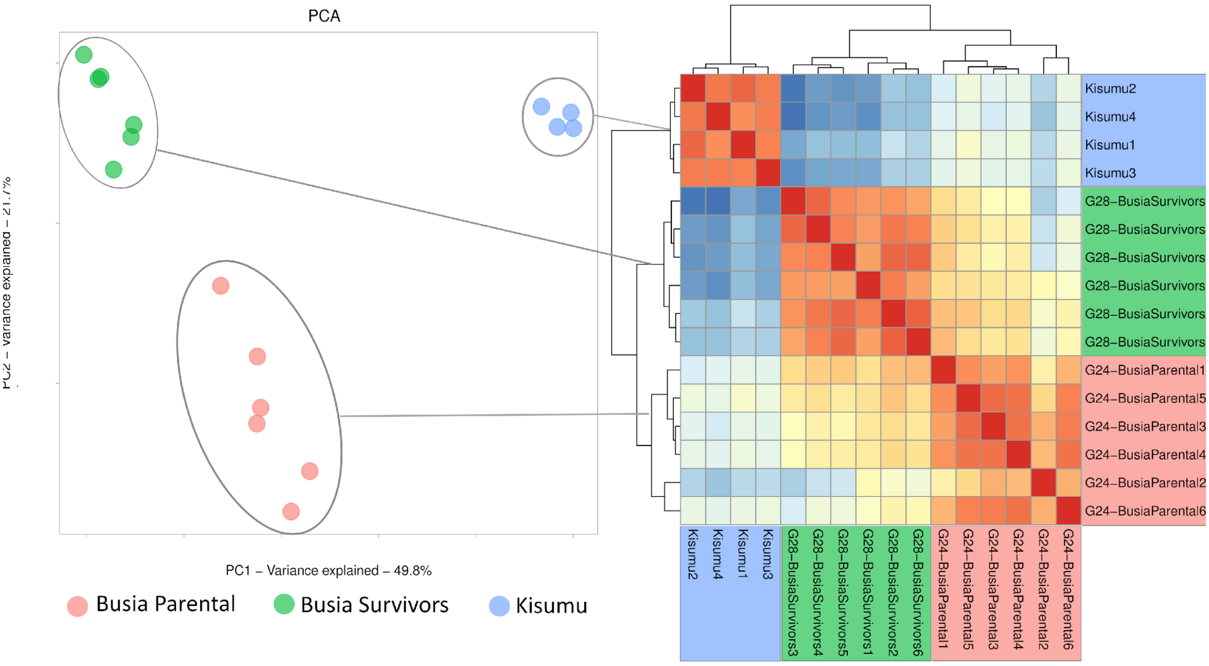


Figure 3: Exploratory sample clustering. a) Principal Components Analysis of the normalised read count data, showing clear separation between conditions b) A sample-to-sample Pearson's correlation heatmap of normalised read counts assigned to each biological replicate, dendrograms show heirarchical clustering applied directly to Pearson's correlations

2.4.4 Differential expression

We compared the G24 parental Busia strain to the G28 Busia survivors, and also to the lab-susceptible Kisumu, which provides a cross-reference with earlier studies, as well as an extra level of filtering to identify candidate genes. Using an adjusted p-value threshold of 0.05, our DESeq2 differential expression analysis (Wald test) identified 5416 differentially expressed genes between Kisumu and the parental Busia line and 5657 between the parental Busia and the G28 selected Busia survivors. The full table of differentially expressed genes in all comparisons can be found in the supplementary file S1, and volcano plots in supplementary figures 4a, b, c.

After four generations of selections plus insecticide exposure, a number of genes belonging to candidate detoxification families were significantly differentially expressed between the G24 Busia Parental and G28 Busia Selected strains, for example, 51 cytochrome P450s, 23 carboxylesterases and 20 ABC transporters. All three sensory appendage protein (Sap) genes in the *An. gambiae* genome were significantly overexpressed in the G28 Busia selected survivors compared to the parental Busia line. Sap2 showed 10.7 fold overexpression (6.5-17.5 95% CIs), while Sap1 exhibited 1.8-fold (1.36-2.44 95% CIs) and Sap3 2-fold (1.58–2.51 95% CIs) overexpression.

We also provide a module which summarises gene expression in specific gene families if provided with a table mapping genes to pfam domains. We provide an example mapping file. Figure 4 shows a summary of expression data in the Glutathione-S-Transferase (GST) gene family, known to be associated with insecticide resistance. The normalised read counts for each gene (blue squares) are clustered and ordered with hierarchical clustering. We then plot the clustering dendrogram alongside a summary of differential expression in each comparison, and the normalised read counts for each biological replicate. Further plots for other gene families are provided in the supplementary file S3. The user may specify their own pfam domains of interest. The default settings apply the analysis to cytochrome P450s, GSTs, carboxylesterases, ABC-transporters, odorant binding proteins, olfactory receptors, Ionotropic recepters, gustatory receptors, and cuticle-related genes. In addition to this module, the user may also supply a list of transcripts, and RNA-Seq-Pop will produce a heatmap on the normalised read count data (supplementary figure 5).

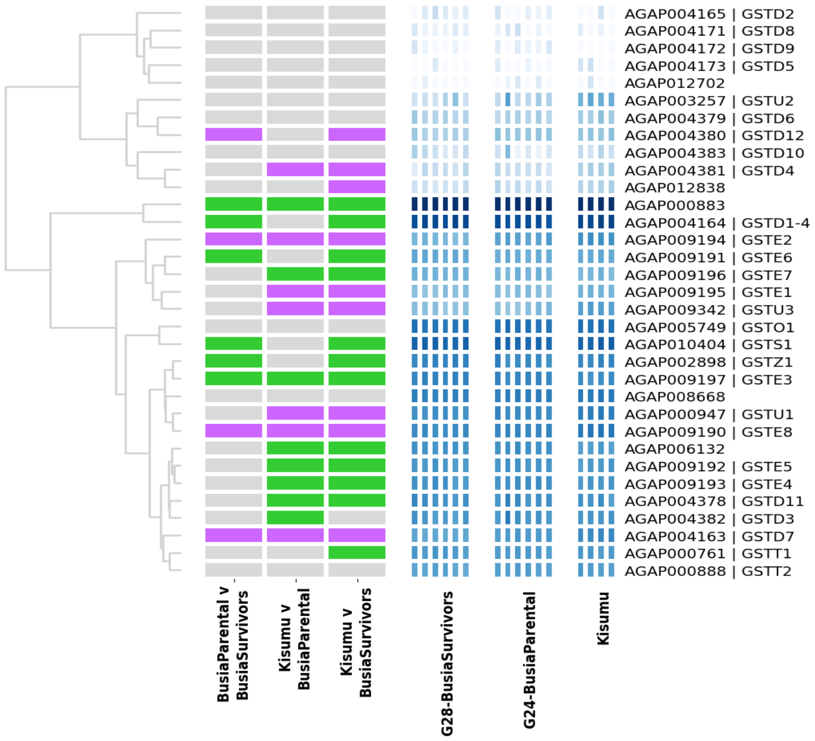


Figure 4: Summarising gene expression in the GST gene family. Using PFAM domains, we extract genes from specific gene families, and summarise fold-change (left) and normalised read count data (right). Genes are ordered by hierarchical clustering of the normalised read count data, clustering results are displayed by the dendrogram (far left). In the fold-change plot (left), green=overexpression in second group, purple=underexpression.

To investigate the similarity of differential expression comparisons, RNA-Seq-Pop provides a venn diagram module which displays the number of shared up or downregulated genes between multiple comparisons (supplementary Figure 6). Using a separate option within RNA-Seq-Pop that looks for genes expressed in the same direction across multiple comparisons, we identified a cluster of carboxylesterases which were overexpressed in G24 Busia Parental vs Kisumu and in G28 Busia Selected Survivors vs G24 Busia Parental. In the latter comparison, Coebe2c showed a fold change of 1.69 (1.3-2.1 95% CIs), Coebe3c 3.05 (1.6-5.9 95% CIs) and Coebe4c 1.61 (1.2-2.2 95% CIs). We examined whether any selective sweeps were observed around these loci in the Ag1000g phase 1 data set and identified one in a population of *An. gambiae* from Gabon, though not in Uganda.

RNA-Seq-Pop also performs differential expression at the isoform-level with sleuth. As an example, we examined isoform level variation at the voltage-gated sodium channel (*Vgsc*), the target of pyrethroid insecticides. As the *Vgsc* contains 13 annotated transcripts and 39 exons, there is a large potential for alternative splicing, which could be an important but as of yet overlooked mechanism of target-site resistance. Between the Busia G24 Parental strain and the Busia G28 survivors, we find no significant difference in expression of any *Vgsc* transcript. However, when comparing the susceptible Kisumu strain to the G24 Busia Parental strain, five *Vgsc* transcripts are differentially expressed – AGAP004707-RA (adjusted pval=0.0059), AGAP004707-RD (adjusted pval=0.0096), AGAP004707-RI (adjusted pval=0.0095), AGAP004707-RL (adjusted pval=1.4e-12) and AGAP004707-RM (adjusted pval=4.36e-07). Given the minimal phenotypic difference between these two strains, it is not clear whether these differences are related to pyrethroid resistance or if this variation is natural between strains.

2.4.5 Variant calling

We enabled RNA-Seq-Pop to call genomic variants with freebayes and output data in VCF format. Across all chromosomes, and after filtering, RNA-Seq-Pop called 734,269 variants. Figure 5 shows a visual representation of genome composition in the Anopheles gambiae PEST reference genome, and the proportion of SNPs covered by each genomic feature in our genotype calls. The *An. gambiae* genome consists of 54% intergenic and 46% genic sequence (of which 14% are exonic, and 32% intronic). Given the nature of RNA-Seq, we expected to primarily find SNPs in coding regions of the genome, which are expressed. Indeed, of these 734,269 variants, we find 73% residing within exons, 11% in introns, and 16% in intergenic regions. The finding of 16% of SNPs in intergenic regions is likely to be explained by expression of non-coding RNAs, and the misannotation of transcripts – particularly 5' and 3' UTRs. The workflow automatically annotates the called variants with snpEff - across all exons, 16.4% of variants were annotated as non-synonymous, and 58.1% as synonymous.

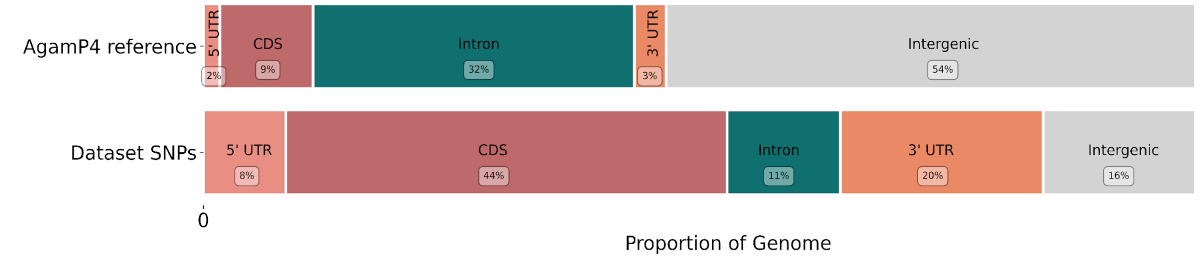


Figure 5: SNPs from RNA-Seq are enriched in transcribed regions. Illustration of the proportion of SNPs found within each genomic feature in the AgamP4 reference genome (Upper panel) and in the combined Busia and Kisumu *Anopheles gambiae* RNA-Seq dataset (Lower panel).

There was a positive correlation between read counts per gene, and the number of called SNPs per gene when controlling for gene size (GLM - coef=0.135, pval=2.2e-36, supplementary table 5). A PCA based upon read count data, was not qualitatively different from the PCA on expression data (Figure 3 and supplementary Figure 8).

2.4.6 Genetic diversity

Table 1 shows genome-wide nucleotide diversity (π) and Watterson's θ, averaged across 20kb non-overlapping windows. To standardise sample size we down-sampled both Busia strains from six to four replicates. Both measures of genetic diversity were significantly lower in the Kisumu strain compared to the two Busia strains, as would be expected after a long history of laboratory colonisation. The G28 selected Busia survivors also show a reduction in genetic diversity compared to the unexposed the G24 parental Busia colony.

Table 1: Genetic Diversity. Average measures of genetic diversity, calculated in 20kb overlapping windows, across chromosomal arms. a) π, Nucleotide diversity b) Θ, Watterson's theta

	π (95% CIs)	**Θ** (95% CIs)
Busia Parental	1.04×10^{-3} (1.02×10^{-3}-1.07×10^{-3})	7.4×10^{-4} (7.23×10^{-4}-7.57×10^{-4})
Busia Selected	7.07×10^{-4} (6.87×10^{-4}-7.27×10^{-4})	5.51×10^{-4} (5.37×10^{-4}-5.65×10^{-4})
Kisumu	6.18×10^{-4} (6.0×10^{-4}-6.35×10^{-4})	4.06×10^{-4} (3.95×10^{-4}-4.18×10^{-4})

2.4.7 Known insecticide resistance variants of interest

If provided with a list of user-defined variants of interest, RNA-Seq-Pop will generate reports and plots of allele balance (the allele frequency found in the read alignments). For our variants of interest, we curated a selection of SNPs which have been associated with insecticide resistance in previous studies. Figure 6 shows allele frequencies of variants of interest across all samples. We show that over the four generations of selections and after insecticide exposure, the frequency of the *Vgsc*-995S kdr allele increased from 25% (95% CIs: 21.5-29.8%) in G24 to 100% in the G28 Busia survivors. In agreement with recent work from the Ag1000g project, we found no known secondary kdr mutations alongside the *Vgsc*-995S allele (Clarkson et al., 2021).

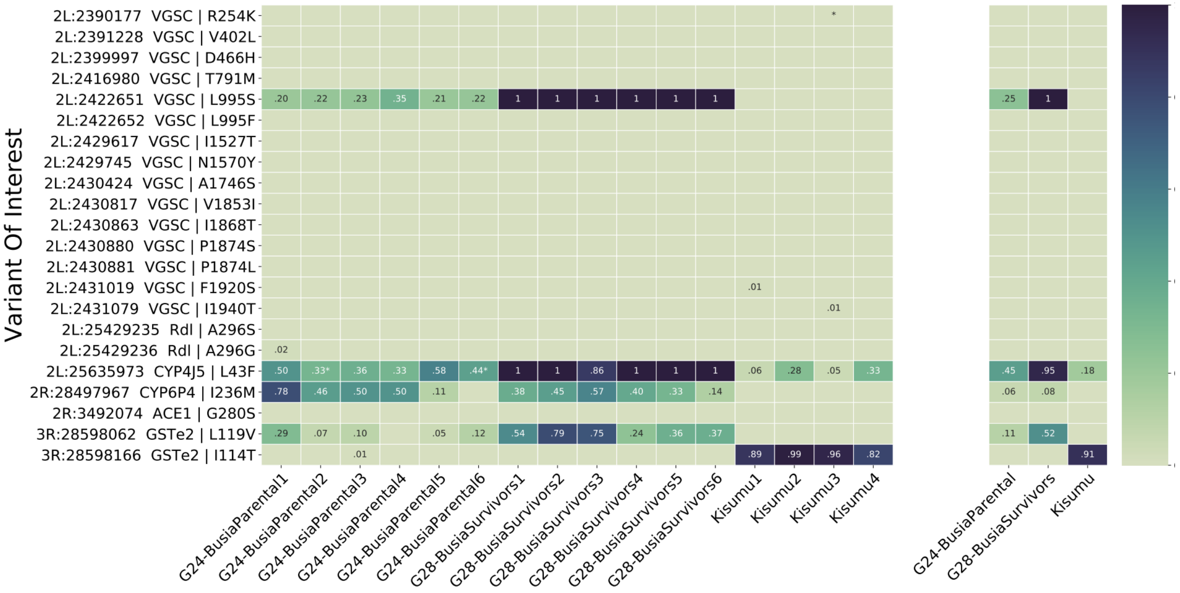


Figure 6: Variants of Interest. A heatmap showing allele frequencies of variants of interest found in read data in a) each sample and b) overall average allele frequency across strains. Blank cells indicate that the mutant allele was not detected despite reads across that genomic position. White cells would indicate zero reads.

In addition, the Cyp4j5-L43F mutation, previously associated with insecticide resistance in Uganda, showed a large increase in frequency after the selection regime and exposure, increasing from an average frequency of 45% (95% CIs: 32-54%) in the G24 Busia to 95% (95% CIs: 93-100%) in the G28 Busia survivors. The Gste2-I114T mutation, associated with DDT resistance, was absent in both Busia strains, however surprisingly, it was present at high frequency (92%) in the pyrethroid susceptible Kisumu reference strain. Another mutation, Gste2-L119V, increased in frequency from 11% (95% CIs: 9-13%) to 52% (95% CIs: 47-58%). The Cyp6p4-I236M mutation, linked to a swept haplotype in Uganda, was present in Busia samples, but there was no significant difference in frequency between the parental (39%, 95% CIs: 29-53%) and selected survivors (38%, 95% CIs: 26-52%). In agreement with these differences in frequency of known insecticide-resistance variants, we find Fst values in both the Vgsc and Cyp4J5 genes in the top 5% percentile between the G24 parental Busia strain and the G28 selected Busia survivors, but not in Cyp6P4 (89th percentile).

The Ace-1-G280S mutation was absent from all samples. A single allele of the rdl-A296G mutation was detected in the Parental Busia strain. Complete allele balance data for all variants of interest can be found in the supplementary file S2. We looked within the primary candidate gene from differential expression analysis, Sap2, for allele frequency changes, but no non-synonymous variants were present in the data.

2.4.8 Selection

The workflow permits calculation of Fst and the population branch statistic (PBS) both in windows as genome-wide selection scans (GWSS) and within each gene. In the context of insecticide resistance, finding regions of high genetic differentiation between susceptible and resistant mosquito populations can allow us to identify loci or variants that contribute to the phenotype. We found high overall levels of Fst between the G24 parental Busia and the G28 selected Busia survivors, however, Fst on chromosomal arm 2L was especially elevated as compared to the other arms (Figure 7), with large Fst signals around the *Vgsc* and 2La inversion. In other datasets from F1 *An. gambiae* (examples in supplementary figure 13), the genome-wide selection scans are able to capture signals at sites of known selective sweeps.

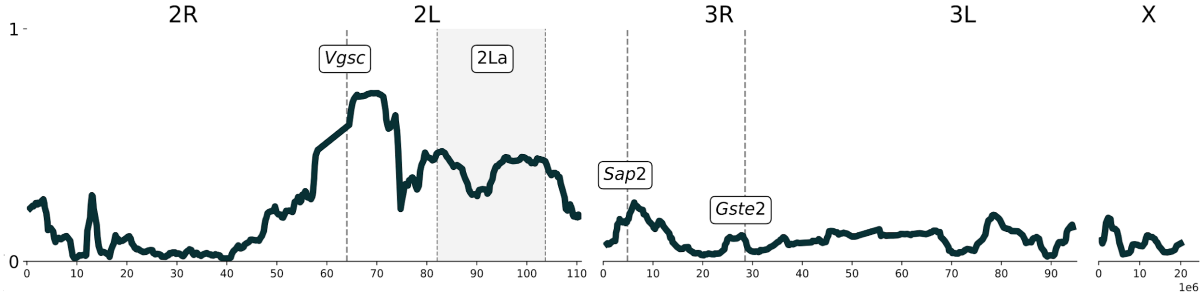


Figure 7: Selection. Hudson F_{st} calculated in windows across the genome, comparing the G24 Busia Parental strain to the G28 Busia survivors. High genetic differentiation can be observed on the 2L chromosomal arm.

2.4.9 Chromosomal Inversions

We estimated the karyotype of the samples with compkaryo for the 2La and 2Rb chromosomal inversions, by extracting karyotype-tagging SNPs. Karyotyping tagging SNPs are alleles in located within the inversion breakpoints, which show fixed (or almost fixed) differences between the inverted and standard karyotypes. We focus on these two inversions because both contain a large number of tagging SNPs, providing confidence in the overall calls. Figure 7 shows a diagram of the *An. gambiae* genome, with the location and average karyotype frequency per group. The 2La inversion was present at a frequency of 86% in the G28 Busia survivors, compared to 33% in the G24 Busia Parental strain (Mann-Whitney U, Adjusted P-value = 0.014), where 0% means no 2La alleles across all tag SNP loci, and 100% means all 2La alleles across all tag SNP loci. The frequency of the 2Rb inversion was also significantly different between Kisumu and both Busia colonies (Mann-Whitney U, Adjusted P-values < 0.05). Supplementary figure 12 shows the per-replicate karyotype frequency.

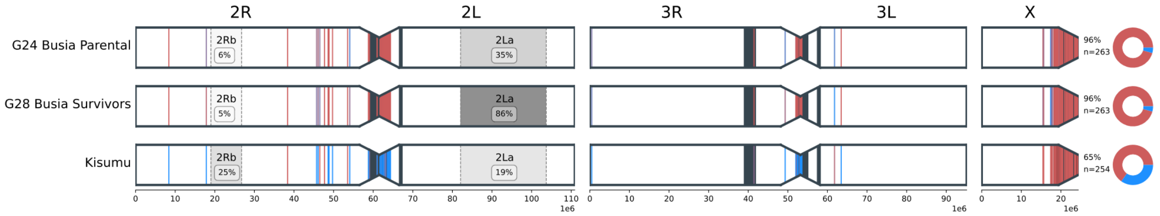


Figure 8: Ancestry and karyotyping. Left) A diagram of the mosquito chromosomal arms, including heterochromatin regions (black). Ancestry informative markers that are indicative of either *An. gambiae* (red) or *An. coluzzii* (blue) are displayed as vertical lines. The major inversions 2La and 2Rb are displayed, along with their respective average frequency amongst treatment groups, as called by the program compkaryo. Far right) A donut chart of the proportion of ancestry informative markers that are indicative of either *An. gambiae* (red) or *An. coluzzii* (blue) ancestry for each sample. The overall proportion of gambiae alleles (%) and the number of called AIMs (n=) per group is labelled.

2.4.10 Ancestry

Ancestry informative markers (AIMs) are SNPs which show fixed (or almost fixed) differences between species. RNA-Seq-Pop can utilise sets of Ancestry Informative markers to investigate the proportion of ancestry for each chromosome assigned to either *An.*

gambiae, *An. coluzzii* or *An. arabiensis*. Figure 8 shows the position of called AIM alleles that map to either *An. gambiae* or An. coluzzii across the genome. This shows that the Busia samples were primarily of *An. gambiae* s.s ancestry across all chromosomes, in concordance with the X chromosome-based SINE species ID assay (Santolomazza et al., 2008). However, the pattern was markedly different for the susceptible reference strain, Kisumu, which showed a large degree of putative *An. coluzzii* ancestry on the autosomes (supplementary table 11).

2.5 Discussion

2.5.1 RNA-Seq-Pop Implementation

RNA-Seq-Pop encompasses a complete workflow for RNA-Sequencing analysis, from quality control and read trimming, to transcript quantification and differential expression analysis (DEA). However, as well as conducting traditional differential expression analyses at both the gene and isoform level, RNA-Seq-Pop exploits useful, but often ignored sequence data.

RNA-Seq-Pop is designed for ease of use, requiring only a sample metadata sheet and a yaml format configuration file. A single command in the terminal will automatically install all dependencies and run the workflow, which is scaled by Snakemake to run on a personal computer, cluster or cloud environment. The workflow is applicable to single or paired-end RNA-Sequencing data from any organism, allowing for variation in ploidy; including haploid, diploid, or pooled samples. We have written RNA-Seq-Pop in accordance with Snakemake best practices (Köster, 2022), and hope that it is an intuitive program, readily configured by the user to allow reproducible transcriptomic analyses. We present documentation with guidance on how to set up and run the workflow. To increase accessibility RNA-Seq-Pop is written in python and R, the two most popular programming languages in the life sciences.

Decreasing sequencing costs have facilitated the proliferation of genomic surveillance in infectious disease research (Neafsey et al., 2021). The specific modules within RNA-Seq-Pop, which are readily adapted to other organisms, allow us to investigate novel variants that may be involved in our phenotype of interest (insecticide resistance), while simultaneously producing data on known resistance variants which can provide actionable information for malaria control programme personnel. For *An. gambiae s.l*, we provide a versioned variants of interest file in the GitHub repository, which we will update with additional resistance or disease transmission-related variants. As well as highlighting known variants, RNA-Seq-Pop can also perform genome-wide selections scans, using Fst (Bhatia et al., 2013) and the Population Branch statistic, PBS (Yi et al., 2010), highlighting known and novel regions of the genome that may be involved in the phenotype of interest.

For the major malaria vector, *An. gambiae s.l*, RNA-Seq-Pop can determine the karyotype frequency of chromosomal inversions, utilising the program compKaryo (Love et al., 2019). *An. gambiae s.l* has been shown to harbour a number of segregating chromosomal inversions, which have been associated with environmental heterogeneity, susceptibility to Plasmodium infection, and with insecticide resistance (Coluzzii et al., 1979, Riehle et al., 2017, Weetman et al., 2018). Typically, we can only detect these inversions through molecular PCR-based assays (of which many do not exist for the range of inversions karyotyped by compkaryo) or laborious and technically challenging cytologic experiments (Coluzzi et al., 2002, White et al., 2007), although recent approaches using tagging SNP panels appear promising (Love et al., 2020).

We can also illuminate the putative ancestry of our samples. This is of particular interest as the two recently-diverged sibling species *An. gambiae* and *An. coluzzii*, may often hybridise, and have undergone extensive introgression in the recent past (Fontaine et al., 2015; Vicente et al., 2017), allowing resistance alleles to cross from one species to another (Clarkson et al., 2014; Grau-Bové et al., 2020, 2021). Despite this, molecular assays typically target only a single marker on the X chromosome, ignoring the potential for admixture elsewhere in the genome (Caputo et al., 2021; Chabi et al., 2019; Santolamazza et al., 2008). As genome resources advance in other vectors, such as Aedes aegypti and

Culex pipiens, we will expand the ancestry informative marker analysis of *RNA-Seq-Pop* to these species complexes.

2.5.2 Patterns of resistance in the Busia dataset

The differential expression analysis highlighted a multitude of detoxification genes overexpressed in the selected Busia survivors, including cytochrome P450s, carboxylesterases, chemosensory proteins, and ABC transporters, reflecting the broad nature of the response to insecticide exposure. Many P450 genes were ≈2 fold overexpressed and it is not known whether this is due to constitutive differences between the strains, or induction by deltamethrin exposure in the G28 Busia strain. The Sap2 gene in particular was highly overexpressed (10.7 fold after deltamethrin selections and exposure), and thus serves as a strong candidate for pyrethroid resistance outside of the West African An. coluzzii populations in which it was originally identified (Ingham et al., 2020). Sap2 is known to be induced upon insecticide exposure, although the relative importance of selection versus induction by exposure cannot be determined from this experimental design.

The increase in *Vgsc*-995S kdr allele frequency (previously 1014S) following selections and segregation post-exposure is as predicted given its known association with pyrethroid resistance. Interestingly, the G28 selected Busia strain showed a much stronger phenotype against permethrin than deltamethrin (supplementary table 2A), which could partially be a result of this mutation. Earlier studies have shown a stronger protective effect of the *Vgsc*-995S allele against permethrin than deltamethrin (Lynd et al., 2010). In agreement with this difference in *Vgsc*-995S frequency, we find high Fst in the *Vgsc* between the G24 parental and G28 selected Busia survivors. The *Vgsc* gene is not differentially expressed between the parental Busia strain and the selected Busia survivors, meaning this result would have been missed using differential expression analyses alone.

The 2La inversion was at much greater frequency in the G28 survivors, suggesting an association with deltamethrin resistance in Busia. Associations between the 2La inversion and insecticide resistance have been previously reported in Uganda (Weetman et al., 2018).

We also find a large difference in Cyp4J5-L43F mutation frequency and there is very high Fst in this gene (0.59), which lies within the 2La inversion. Interestingly, the gene is also differentially expressed, perhaps suggesting that the 2La haplotypic background results in over-transcription of the gene when compared to 2L+a haplotypes. It is not clear whether Cyp4J5 is causative, or if there are other variants on the 2La haplotype(s) that are driving this shift in 2La. In agreement with this and the difference in kdr, we find high overall Fst between the G24 parental and G28 Busia survivors on the 2L chromosomal arm (supplementary table 10).

Interestingly, RNA-Seq-Pop revealed that the Kisumu reference strain, exhibits a large proportion of putative An. coluzzii ancestry. The Kisumu reference strain was colonised from Kisumu, Kenya in 1975 (Williams et al., 2019) from an area where An. coluzzii has not been recorded. The most parsimonious explanation is that the colony has been contaminated through hybridization in the insectary during its long colonisation. The X chromosome is typically resistant to introgression, and consistent with a theory of a lab contamination event, no An. coluzzii variants are found on the X chromosome. The X chromosome of Kisumu also has a particularly low estimate of Watterson's Θ compared to the autosomes, which may reflect admixture present on the autosomes (supplementary table 7A). In addition, we also find that the Kisumu strain contains the Gste2-114T mutation at high frequency. In agreement with this finding, recent data shows occasional low-level resistance to DDT in this strain (Williams et al., 2019). We also observe some putative An. coluzzii alleles in the two Busia strains. Whilst we cannot rule out other explanations, this set of ancestry informative markers were derived from Mali, and therefore it is likely that some may not be truly informative of ancestry outside of this population.

In this study, we exposed the G28 selected strain in order to maximise the resistance phenotype and strengthen the genotype-phenotype association. This design choice, however, may mean estimates of allele and karyotype frequencies are overestimates and not necessarily reflective of either the unexposed G28 or G29 Busia strains. This is because susceptible G28 mosquitoes have not survived, and G28 survivors have likely already mated, meaning susceptible alleles may be passed on to the next generation, affecting allele frequencies. Equally, insecticide survivors may not go on to produce offspring. In general, we recommend sequencing appropriate controls where possible – for example, in our case, including a G28 Busia unexposed group.

When analysing RNA-Sequencing data we only have read coverage in expressed parts of the genome, primarily in exons, and so we can only call genetic variants in these regions. Although not ideal, given that we expect the majority of functional variants to exist in expressed regions of the genome (Choi et al., 2009), this is not a necessarily a major drawback. Indeed, this is the premise of exome sequencing, in which only protein-coding parts of the genome are targeted to sequence. Additionally, estimated population allele frequencies derived from RNA-Seq data may not accurately reflect DNA-based allele frequencies. Allele-specific expression is one cause of this, where two or more alleles in a diploid or polyploid may be expressed at different levels, causing an imbalance. Despite this, previous studies have shown a strong correlation between expressed and true allele frequencies, particularly at higher sequencing depth (Jehl et al., 2021; Lopez-Maestre et al., 2016; Oikkonen & Lise, 2017; Quinn et al., 2013). In this study, we performed RNASeq at a high sequencing depth, and therefore can have more confidence overall in our genotype calls and subsequent analyses. We recommend generally that for differential expression analyses, low coverage RNA-Sequencing is sufficient (10-25 million reads, or 5-13.5X coverage for *An. gambiae*), with a higher number of biological replicates (≥ 4). For variant analyses, higher coverage is preferred (25-60 million reads, or 13.5-32.4X coverage for *An. gambiae*), with a high number of individuals pooled per replicate (≥ 10).

Although other studies present strategies to call variants from RNA-Sequencing data (Brouard & Bissonnette, 2022; Jehl et al., 2021; Piskol et al., 2013; Quinn et al., 2013), none of these studies present convenient, reproducible bioinformatic pipelines to implement their suggested strategies, instead requiring the user to manually perform each step. In addition to that, we found no studies that present pipelines to call variants and also perform analyses on the resulting SNP data. Although a previous study showed that population genomic signals can be extracted from RNA-Sequencing data, they did not package their analysis into any usable workflow, and the analyses themselves are limited in comparison to RNA-Seq-Pop (Thorstensen et al., 2020). Based on the lack of comparable, easy-to-use workflows, we envisage that RNA-Seq-Pop will prove useful to many researchers in a range of biological fields.

2.5.3 Data accessibility

The workflow is hosted at https://github.com/sanjaynagi/rna-seq-pop. We welcome and encourage any feedback or contributions to RNA-Seq-Pop. The variant of interest file is versioned and is included in the GitHub repository. Raw sequence data is deposited at the ENA under BioProject PRJNA748581. The modified version of compKaryo is found here https://github.com/sanjaynagi/compkaryo.

2.6 References

Akbari, O. S., Antoshechkin, I., Amrhein, H., Williams, B., Diloreto, R., Sandler, J., & Hay, B. A. (2013). The Developmental Transcriptome of the Mosquito *Aedes aegypti,* an Invasive Species and Major Arbovirus Vector. G3 Genes|Genomes|Genetics, 3(9), 1493–1509. https://doi.org/10.1534/g3.113.006742

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. Genome Research, 23(9), 1514–1521. https://doi.org/10.1101/gr.154831.113

Bonizzoni, M., Afrane, Y., Dunn, W. A., Atieli, F. K., Zhou, G., Zhong, D., Li, J., Githeko, A., & Yan, G. (2012). Comparative Transcriptome Analyses of Deltamethrin-Resistant and -Susceptible *Anopheles gambiae* Mosquitoes from Kenya by RNA-Seq. PLoS ONE, 7(9), 1–11. https://doi.org/10.1371/journal.pone.0044607

Bonizzoni, M., Dunn, W. A., Campbell, C. L., Olson, K. E., Dimon, M. T., Marinotti, O., & James, A. A. (2011). RNA-seq analyses of blood-induced changes in gene expression in the mosquito vector species, *Aedes aegypti*. BMC Genomics, 12(1), 1–13. https://doi.org/10.1186/1471-2164-12-82

Bonizzoni, M., Ochomo, E., Dunn, W. A., Britton, M., Afrane, Y., Zhou, G., Hartsel, J., Lee, M.-C., Xu, J., Githeko, A., Fass, J., & Yan, G. (2015). RNA-seq analyses of changes in the *Anopheles gambiae* transcriptome associated with resistance to pyrethroids in Kenya: Identification of candidate-resistance genes and candidate-resistance SNPs. Parasites & Vectors, 8(1), 474. https://doi.org/10.1186/s13071-015-1083-z

Boonkaew, T., Mongkol, W., Prasert, S., Paochan, P., Yoneda, S., Nguitragool, W., Kumpitak, C., Sattabongkot, J., & Kubera, A. (2020). Transcriptome analysis of *Anopheles dirus* and *Plasmodium* *vivax* at ookinete and oocyst stages. Acta Tropica, 207, 105502. https://doi.org/10.1016/j.actatropica.2020.105502

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2015). Near-optimal probabilistic RNA-Seq quantification. Nature Biotechnology, 1–21.

Caputo, B., Pichler, V., Bottà, G., De Marco, C., Hubbart, C., Perugini, E., Pinto, J., Rockett, K. A., Miles, A., & della Torre, A. (2021). Novel genotyping approaches to easily detect genomic admixture between the major Afrotropical malaria vector species, *Anopheles coluzzii* and *An. gambiae*. Molecular Ecology Resources, 21(5), 1504–1516. https://doi.org/10.1111/1755-0998.13359

Cassone, B. J., Kay, R. G. G., Daugherty, M. P., & White, B. J. (2017). Comparative Transcriptomics of Malaria Mosquito Testes: Function, Evolution, and Linkage. G3 (Bethesda, Md.), 7(4), 1127–1136. https://doi.org/10.1534/g3.117.040089

Chabi, J., Van't Hof, A., N'dri, L. K., Datsomor, A., Okyere, D., Njoroge, H., Pipini, D., Hadi, M. P., De Souza, D. K., Suzuki, T., Dadzie, S. K., & Jamet, H. P. (2019). Rapid high throughput SYBR green assay for identifying the malaria vectors *Anopheles arabiensis, Anopheles coluzzii* and *Anopheles gambiae s.s.* Giles. PLoS ONE, 14(4), 1–11. https://doi.org/10.1371/journal.pone.0215669

Chen, B., Zhang, Y.-J., He, Z., Li, W., Si, F., Tang, Y., He, Q., Qiao, L., Yan, Z., Fu, W., & Che, Y. (2014). De novo transcriptome sequencing and sequence analysis of the malaria vector *Anopheles sinensis* (Diptera: Culicidae). Parasites & Vectors, 7(1), 1–12. https://doi.org/10.1186/1756-3305-7-314

Choi, Y.-J., Aliota, M. T., Mayhew, G. F., Erickson, S. M., & Christensen, B. M. (2014). Dual RNA-seq of Parasite and Host Reveals Gene Expression Dynamics during Filarial Worm–Mosquito Interactions. PLOS Neglected Tropical Diseases, 8(5), e2905. https://doi.org/10.1371/journal.pntd.0002905

Cingolani, P., Platts, A., Wang, L. 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, 6(2), 80–92. https://doi.org/10.4161/fly.19695

Clarkson, C. S., Miles, A., Harding, N. J., O'Reilly, A. O., Weetman, D., Kwiatkowski, D., & Donnelly, M. J. (2021). The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors *Anopheles gambiae* and *Anopheles coluzzii*. Molecular Ecology, 30(21), 5303–5317. https://doi.org/10.1111/mec.15845

Clarkson, C. S., Weetman, D., Essandoh, J., Yawson, A. E., Maslen, G., Manske, M., Field, S. G., Webster, M., Antão, T., MacInnis, B., Kwiatkowski, D., & Donnelly, M. J. (2014). Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation. Nature Communications, 5(May). https://doi.org/10.1038/ncomms5248

Coatsworth, H., Caicedo, P. A., Winsor, G., Brinkman, F., Ocampo, C. B., & Lowenberger, C. (2021). Transcriptome comparison of dengue-susceptible and -resistant field derived strains of Colombian *Aedes aegypti* using RNA-sequencing. Memórias Do Instituto Oswaldo Cruz, 116. https://doi.org/10.1590/0074-02760200547

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2). https://doi.org/10.1093/gigascience/giab008

David, J.-P., Faucon, F., Chandor-Proust, A., Poupardin, R., Riaz, M., Bonin, A., Navratil, V., & Reynaud, S. (2014). Comparative analysis of response to selection with three insecticides in the dengue mosquito Aedes aegypti using mRNA sequencing. BMC Genomics, 15(1), 174. https://doi.org/10.1186/1471-2164-15-174

De Marco, L., Sassera, D., Epis, S., Mastrantonio, V., Ferrari, M., Ricci, I., Comandatore, F., Bandi, C., Porretta, D., & Urbanelli, S. (2017). The choreography of the chemical defensome response to insecticide stress: Insights into the *Anopheles stephensi* transcriptome using RNA-Seq. Scientific Reports, 7(1), 41312. https://doi.org/10.1038/srep41312

Djouaka, R. F., Bakare, A. A., Coulibaly, O. N., Akogbeto, M. C., Ranson, H., Hemingway, J., & Strode, C. (2008). Expression of the cytochrome P450s, CYP6P3 and CYP6M2 are significantly elevated in multiple pyrethroid resistant populations of *Anopheles gambiae* s.s. From Southern Benin and Nigeria. BMC Genomics, 9(1), 538. https://doi.org/10.1186/1471-2164-9-538

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Faucon, F., Gaude, T., Dusfour, I., Navratil, V., Corbel, V., Juntarajumnong, W., Girod, R., Poupardin, R., Boyer, F., Reynaud, S., & David, J. P. (2017). In the hunt for genomic markers of metabolic resistance to pyrethroids in the mosquito *Aedes aegypti*: An integrated next-generation sequencing approach. PLoS Neglected Tropical Diseases, 11(4), 1–20. https://doi.org/10.1371/journal.pntd.0005526

Faust, G. G., & Hall, I. M. (2014). SAMBLASTER: Fast duplicate marking and structural variant read extraction. Bioinformatics, 30(17), 2503–2505. https://doi.org/10.1093/bioinformatics/btu314

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y.-C., Smith, H. A., Love, R. R., Lawniczak, M. K. N., Hahn, M. W., & Besansky, N. J. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science, 347(6217), 1258522. https://doi.org/10.1126/science.1258522

Garrison, E., Kronenberg, Z. N., Dawson, E. T., & Pedersen, B. S. (2021). Vcflib and tools for processing the VCF variant call format. BioRxiv, 1–15.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. ArXiv, 1–9.

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. PLoS Genetics, 11(2), 1–32. https://doi.org/10.1371/journal.pgen.1005004

Grau-Bové, X., Lucas, E., Pipini, D., Rippon, E., van 't Hof, A. E., Constant, E., Dadzie, S., Egyir-Yawson, A., Essandoh, J., Chabi, J., Djogbénou, L., Harding, N. J., Miles, A., Kwiatkowski, D., Donnelly, M. J., Weetman, D., & The Anopheles gambiae 1000 Genomes Consortium. (2021). Resistance to pirimiphos-methyl in West African *Anopheles* is spreading via duplication and introgression of the Ace1 locus. PLOS Genetics, 17(1), e1009253. https://doi.org/10.1371/journal.pgen.1009253

Grau-Bové, X., Tomlinson, S., O'Reilly, A. O., Harding, N. J., Miles, A., Kwiatkowski, D., Donnelly, M. J., Weetman, D., & and The Anopheles gambiae 1000 Genomes Consortium. (2020). Evolution of the Insecticide Target Rdl in African *Anopheles* Is Driven by Interspecific and Interkaryotypic Introgression. Molecular Biology and Evolution, 37(10), 2900–2917. https://doi.org/10.1093/molbev/msaa128

Grüning, B., Chilton, J., Köster, J., Dale, R., Soranzo, N., van den Beek, M., Goecks, J., Backofen, R., Nekrutenko, A., & Taylor, J. (2018). Practical Computational Reproducibility in the Life Sciences. Cell Systems, 6(6), 631–635. https://doi.org/10.1016/j.cels.2018.03.014
Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics, 132(2), 583–589. https://doi.org/10.1093/genetics/132.2.583

Ingham, V. A., Anthousi, A., Douris, V., Harding, N. J., Lycett, G., Morris, M., Vontas, J., & Ranson, H. (2020). A sensory appendage protein protects malaria vectors from pyrethroids. Nature, 577(7790), 376–380. https://doi.org/10.1038/s41586-019-1864-1

Ingham, V. A., Tennessen, J. A., Lucas, E. R., Elg, S., Yates, H. C., Carson, J., Guelbeogo, W. M., Sagnon, N., Hughes, G. L., Heinz, E., Neafsey, D. E., & Ranson, H. (2021). Integration of whole genome sequencing and transcriptomics reveals a complex picture of the reestablishment of insecticide resistance in the major malaria vector *Anopheles coluzzii*. PLOS Genetics, 17(12), e1009970. https://doi.org/10.1371/journal.pgen.1009970

Ingham, V., Wagstaff, S., & Ranson, H. (2018). Transcriptomic meta-signatures identified in *Anopheles gambiae* populations reveal previously undetected insecticide resistance mechanisms. Nature Communications. https://doi.org/10.1038/s41467-018-07615-x

Jehl, F., Degalez, F., Bernard, M., Lecerf, F., Lagoutte, L., Désert, C., Coulée, M., Bouchez, O., Leroux, S., Abasht, B., Tixier-Boichard, M., Bed'hom, B., Burlot, T., Gourichon, D., Bardou, P., Acloque, H., Foissac, S., Djebali, S., Giuffra, E., … Lagarrigue, S. (2021). RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. Frontiers in Genetics, 12. https://www.frontiersin.org/article/10.3389/fgene.2021.655707

Jiang, X., Hall, A. B., Biedler, J. K., & Tu, Z. (2017). Single molecule RNA sequencing uncovers trans-splicing and improves annotations in *Anopheles stephensi*. Insect Molecular Biology, 26(3), 298–307. https://doi.org/10.1111/imb.12294

Kang, D. S., Kim, S., Cotten, M. A., & Sim, C. (2021). Transcript Assembly and Quantification by RNA-Seq Reveals Significant Differences in Gene Expression and Genetic Variants in Mosquitoes of the *Culex pipiens* (Diptera: Culicidae) Complex. Journal of Medical Entomology, 58(1), 139–145. https://doi.org/10.1093/jme/tjaa167

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Köster, J. (2022). Snakemake: Best Practices. https://snakemake.readthedocs.io/en/stable/snakefiles/best_practices.html

Lataretu, M., & Hölzer, M. (2020). Rnaflow: An effective and simple rna-seq differential gene expression pipeline using nextflow. Genes, 11(12), 1–17. https://doi.org/10.3390/genes11121487

Li, Y., Piermarini, P. M., Esquivel, C. J., Drumm, H. E., Schilkey, F. D., & Hansen, I. A. (2017). RNA-Seq Comparison of Larval and Adult Malpighian Tubules of the Yellow Fever Mosquito *Aedes aegypti* Reveals Life Stage-Specific Changes in Renal Function. Frontiers in Physiology, 8. https://www.frontiersin.org/article/10.3389/fphys.2017.00283

Lien, N. T. K., Ngoc, N. T. H., Lan, N. N., Hien, N. T., Tung, N. V., Ngan, N. T. T., Hoang, N. H., & Binh, N. T. H. (2019). Transcriptome Sequencing and Analysis of Changes Associated with Insecticide Resistance in the Dengue Mosquito (*Aedes aegypti*) in Vietnam. The American Journal of Tropical Medicine and Hygiene, 100(5), 1240–1248. https://doi.org/10.4269/ajtmh.18-0607

Lopez-Maestre, H., Brinza, L., Marchet, C., Kielbassa, J., Bastien, S., Boutigny, M., Monnin, D., Filali, A. E., Carareto, C. M., Vieira, C., Picard, F., Kremer, N., Vavre, F., Sagot, M. F., & Lacroix, V. (2016). SNP calling from RNA-seq data without a reference genome: Identification, quantification, differential analysis and impact on the protein sequence. Nucleic Acids Research, 44(19), 1–13. https://doi.org/10.1093/nar/gkw655

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 1–21. https://doi.org/10.1186/s13059-014-0550-8

Love, R. R., Redmond, S. N., Pombi, M., Caputo, B., Petrarca, V., Della Torre, A., & Besansky, N. J. (2019). In Silico Karyotyping of Chromosomally Polymorphic Malaria Mosquitoes in the *Anopheles gambiae* Complex. G3 (Bethesda, Md.), 9(10), 3249–3262. https://doi.org/10.1534/g3.119.400445

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. PLoS Computational Biology, 13(5), 1–23. https://doi.org/10.1371/journal.pcbi.1005457

Lv, Y., Wang, W., Hong, S., Lei, Z., Fang, F., Guo, Q., Hu, S., Tian, M., Liu, B., Zhang, D., Sun, Y., Ma, L., Shen, B., Zhou, D., & Zhu, C. (2016). Comparative transcriptome analyses of deltamethrin-susceptible and -resistant *Culex pipiens pallens* by RNA-seq. Molecular Genetics and Genomics, 291(1), 309–321. https://doi.org/10.1007/s00438-015-1109-4

Lynd, A., Gonahasa, S., Staedke, S. G., Oruni, A., Maiteki-Sebuguzi, C., Dorsey, G., Opigo, J., Yeka, A., Katureebe, A., Kyohere, M., Hemingway, J., Kamya, M. R., & Donnelly, M. J. (2019). LLIN Evaluation in Uganda Project (LLINEUP): A cross-sectional survey of species diversity and insecticide resistance in 48 districts of Uganda. Parasites & Vectors, 12(1), 94. https://doi.org/10.1186/s13071-019-3353-7

Lynd, A., Weetman, D., Barbosa, S., Egyir Yawson, A., Mitchell, S., Pinto, J., Hastings, I., & Donnelly, M. J. (2010). Field, genetic, and modeling approaches show strong positive selection acting upon an insecticide resistance mutation in *Anopheles gambiae s.s.* Molecular Biology and Evolution, 27(5), 1117–1125. https://doi.org/10.1093/molbev/msq002

Mackenzie-Impoinvil, L., Weedall, G. D., Lol, J. C., Pinto, J., Vizcaino, L., Dzuris, N., Riveron, J., Padilla, N., Wondji, C., & Lenhart, A. (2019). Contrasting patterns of gene expression indicate differing pyrethroid resistance mechanisms across the range of the New World malaria vector *Anopheles albimanus*. PLoS ONE, 14(1), 1–27. https://doi.org/10.1371/journal.pone.0210586

Martin, M. (2011). Cutadapt removes adaptor sequences from high-throughput sequencing reads. EMBnet Journal, 17.

Martynova, T., Kamanda, P., & Sim, C. (2022). Transcriptome profiling reveals sex-specific gene expressions in pupal and adult stages of the mosquito Culex pipiens. Insect Molecular Biology, 31(1), 24–32. https://doi.org/10.1111/imb.12735

Messenger, L. A., Impoinvil, L. M., Derilus, D., Yewhalaw, D., Irish, S., & Lenhart, A. (2021). A whole transcriptomic approach provides novel insights into the molecular basis of organophosphate and pyrethroid resistance in *Anopheles arabiensis* from Ethiopia. Insect Biochemistry and Molecular Biology, 139(July), 103655. https://doi.org/10.1016/j.ibmb.2021.103655

Miles, A., & Harding, N. J. (2017). Scikit-allel. https://doi.org/10.5281/zenodo.3935797

Miles, A., Harding, N. J., Bottà, G., Clarkson, C. S., Antão, T., Kozak, K., Schrider, D. R., Kern, A. D., Redmond, S., Sharakhov, I., Pearson, R. D., Bergey, C., Fontaine, M. C., Donnelly, M. J., Lawniczak, M. K. N., Ayala, D., Besansky, N. J., Burt, A., Caputo, B., … Kwiatkowski, D. P. (2017). Genetic diversity of the African malaria vector *Anopheles gambiae*. Nature, 552, 96–100. https://doi.org/10.1038/nature24995

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. F1000Research, 1–17.

Mongkol, W., Nguitragool, W., Sattabongkot, J., & Kubera, A. (2018). Blood-induced differential gene expression in *Anopheles dirus* evaluated using RNA sequencing. Medical and Veterinary Entomology, 32(4), 399–406. https://doi.org/10.1111/mve.12310

Müller, P., Warr, E., Stevenson, B. J., Pignatelli, P. M., Morgan, J. C., Steven, A., Yawson, A. E., Mitchell, S. N., Ranson, H., Hemingway, J., Paine, M. J. I., & Donnelly, M. J. (2008). Field-caught permethrin-resistant *Anopheles gambiae* overexpress CYP6P3, a P450 that metabolises pyrethroids. PLoS Genetics. https://doi.org/10.1371/journal.pgen.1000286

Nag, D. K., Dieme, C., Lapierre, P., Lasek-Nesselquist, E., & Kramer, L. D. (2021). RNA-Seq analysis of blood meal induced gene-expression changes in *Aedes aegypti* ovaries. BMC Genomics, 22(1), 396. https://doi.org/10.1186/s12864-021-07551-z

Neafsey, D. E., Taylor, A. R., & MacInnis, B. L. (2021). Advances and opportunities in malaria population genomics. Nature Reviews Genetics, 22(August), 502–517. https://doi.org/10.1038/s41576-021-00349-5

Neafsey, D., Waterhouse, R., Abai, M., Aganezov, S., Alekseyev, M., Allen, J., Amon, J., Arca, B., Arensburger, P., Artemov, G., Assour, L., Basseri, H., Berlin, A., Birren, B., Blandin, S., Brockman, A., Burkot, T., Burt, A., Chan, C., … Besansky, N. (2015). Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. Science, 347(6217), 1258522–1258522. https://doi.org/10.1126/science.1258522

Neira-Oviedo, M., Tsyganov-Bodounov, A., Lycett, G. J., Kokoza, V., Raikhel, A. S., & Krzywinski, J. (2011). The RNA-Seq approach to studying the expression of mosquito mitochondrial genes. Insect Molecular Biology, 20(2), 141–152. https://doi.org/10.1111/j.1365-2583.2010.01053.x

Njoroge, H., Oruni, A., Pipini, D., Nagi, S. C., Lynd, A., Eric, R., Tomlinson, S., Grau-bove, X., Mcdermott, D., Emile, Z., Agossa, F. R., Mokuba, A., Irish, S., Kabula, B., Mbogo, C., Paine, M. J. I., Weetman, D., Donnelly, M. J., Place, P., … Place, P. (2021). Identification of a rapidly-spreading triple mutant for high-level metabolic insecticide resistance in *Anopheles gambiae* provides a real-time molecular diagnostic for anti-malarial intervention deployment. BioRxiv, 1–23.

Oikkonen, L., & Lise, S. (2017). Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. Wellcome Open Research, 2, 6. https://doi.org/10.12688/wellcomeopenres.10501.2

Pimentel, H., Bray, N. L., Puente, S., Melsted, P., & Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nature Methods, 14(7), 687–690. https://doi.org/10.1038/nmeth.4324

Poelchau, M. F., Reynolds, J. A., Elsik, C. G., Denlinger, D. L., & Armbruster, P. A. (2013). RNA-Seq reveals early distinctions and late convergence of gene expression between diapause and quiescence in the Asian tiger mosquito, *Aedes albopictus*. Journal of Experimental Biology, 216(21), 4082–4090. https://doi.org/10.1242/jeb.089508

Quinn, E. M., Cormican, P., Kenny, E. M., Hill, M., Anney, R., Gill, M., Corvin, A. P., & Morris, D. W. (2013). Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. PLoS ONE, 8(3), e58815. https://doi.org/10.1371/journal.pone.0058815

Santolamazza, F., Mancini, E., Simard, F., Qi, Y., Tu, Z., & della Torre, A. (2008). Insertion polymorphisms of SINE200 retrotransposons within speciation islands of *Anopheles gambiae* molecular forms. Malaria Journal, 7(1), 163. https://doi.org/10.1186/1475-2875-7-163

Simma, E. A., Dermauw, W., Balabanidou, V., Snoeck, S., Bryon, A., Clark, R. M., Yewhalaw, D., Vontas, J., Duchateau, L., & Van Leeuwen, T. (2019). Genome-wide gene expression profiling reveals that cuticle alterations and P450 detoxification are associated with deltamethrin and DDT resistance in *Anopheles arabiensis* populations from Ethiopia. Pest Management Science, 75(7), 1808–1818. https://doi.org/10.1002/ps.5374

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. Nature Reviews Genetics, 20(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Sun, H., Mertz, R. W., Smith, L. B., & Scott, J. G. (2021). Transcriptomic and proteomic analysis of pyrethroid resistance in the CKR strain of *Aedes aegypti*. PLOS Neglected Tropical Diseases, 15(11), e0009871. https://doi.org/10.1371/journal.pntd.0009871

Taxiarchi, C., Kranjc, N., Kriezis, A., Kyrou, K., Bernardini, F., Russell, S., Nolan, T., Crisanti, A., & Galizi, R. (2019). High-resolution transcriptional profiling of *Anopheles gambiae* spermatogenesis reveals mechanisms of sex chromosome regulation. Scientific Reports, 9(1), 14841. https://doi.org/10.1038/s41598-019-51181-1

Tchouakui, M., Mugenzi, L. M. J., D. Menze, B., Khaukha, J. N. T., Tchapga, W., Tchoupo, M., Wondji, M. J., & Wondji, C. S. (2021). Pyrethroid Resistance Aggravation in Ugandan Malaria Vectors Is Reducing Bednet Efficacy. Pathogens, 10(4), 415. https://doi.org/10.3390/pathogens10040415

The Anopheles gambiae 1000 Genomes Consortium. (2020). Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii. Genome Research, 1–14. https://doi.org/10.1101/gr.262790.120.Freely

Thorstensen, M. J., Jeffrey, J. D., Treberg, J. R., Watkinson, D. A., Enders, E. C., & Jeffries, K. M. (2020). Genomic signals found using RNA sequencing show signatures of selection and subtle population differentiation in walleye (*Sander vitreus*) in a large freshwater ecosystem. Ecology and Evolution, 10(14), 7173. https://doi.org/10.1002/ece3.6418

Van den Berge, K., Hembach, K. M., Soneson, C., Tiberi, S., Clement, L., Love, M. I., Patro, R., & Robinson, M. D. (2019). RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. Annual Review of Biomedical Data Science, 2(1), 139–173. https://doi.org/10.1146/annurev-biodatasci-072018-021255

Vedururu, R. kiran, Neave, M. J., Tachedjian, M., Klein, M. J., Gorry, P. R., Duchemin, J.-B., & Paradkar, P. N. (2019). RNASeq Analysis of Aedes albopictus Mosquito Midguts after Chikungunya Virus Infection. Viruses, 11(6), 513. https://doi.org/10.3390/v11060513

Vicente, J. L., Clarkson, C. S., Caputo, B., Gomes, B., Pombi, M., Sousa, C. A., Antao, T., Dinis, J., Bottà, G., Mancini, E., Petrarca, V., Mead, D., Drury, E., Stalker, J., Miles, A., Kwiatkowski, D. P., Donnelly, M. J., Rodrigues, A., Della Torre, A., … Pinto, J. (2017). Massive introgression drives species radiation at the range limit of Anopheles gambiae. Nature Publishing Group. https://doi.org/10.1038/srep46451

Wang, Z., & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. RNA, 14(5), 802–813. https://doi.org/10.1261/rna.876308

Wang, Z., Gerstein, M., & Snyder, M. (2010). RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57–63. https://doi.org/10.1038/nrg2484.RNA-Seq

Watterson, G. A. (1975). On the Number of Segregating Sites in Genetical Models without Recombination. Theoretical Population Biology, 276(7), 256–276.

Weetman, D., Wilding, C. S., Neafsey, D. E., Müller, P., Ochomo, E., Isaacs, A. T., Steen, K., Rippon, E. J., Morgan, J. C., Mawejje, H. D., Rigden, D. J., Okedi, L. M., & Donnelly, M. J. (2018). Candidate-gene based GWAS identifies reproducible DNA markers for metabolic pyrethroid resistance from standing genetic variation in East African Anopheles gambiae. Scientific Reports, 8(1), 2920. https://doi.org/10.1038/s41598-018-21265-5

Williams, J., Flood, L., Praulins, G., Ingham, V. A., Morgan, J., Lees, R. S., & Ranson, H. (2019). Characterisation of Anopheles strains used for laboratory screening of new vector control products. Parasites and Vectors, 12(1), 1–14. https://doi.org/10.1186/s13071-019-3774-3

Williams, J., Ingham, V. A., Morris, M., Toé, K. H., Hien, A. S., Morgan, J. C., Dabiré, R. K., Guelbéogo, W. M., Sagnon, N., & Ranson, H. (2022). Sympatric Populations of the Anopheles gambiae Complex in Southwest Burkina Faso Evolve Multiple Diverse Resistance Mechanisms in Response to Intense Selection Pressure with Pyrethroids. Insects, 13(3), 247. https://doi.org/10.3390/insects13030247

Wondji, C. S., Hearn, J., Irving, H., Wondji, M. J., & Weedall, G. (2022). RNAseq-based gene expression profiling of the Anopheles funestus pyrethroid-resistant strain FUMOZ highlights the predominant role of the duplicated CYP6P9a/b cytochrome P450s. G3 Genes|Genomes|Genetics, 12(1), jkab352. https://doi.org/10.1093/g3journal/jkab352

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, N., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., … Wang,

J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science, 329(5987), 75–78. https://doi.org/10.1126/science.1190371

Zhang, X., & Jonassen, I. (2019). RASflow: An RNA-Seq Analysis Workflow with Snakemake. BioRxiv, 1–9. https://doi.org/10.1101/839191

Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS ONE, 9(1). https://doi.org/10.1371/journal.pone.0078644

Zhou, Y., Fu, W.-B., Si, F.-L., Yan, Z.-T., Zhang, Y.-J., He, Q.-Y., & Chen, B. (2019). UDP-glycosyltransferase genes and their association and mutations associated with pyrethroid resistance in Anopheles sinensis (Diptera: Culicidae). Malaria Journal, 18(1), 62. https://doi.org/10.1186/s12936-019-2705-2

2.7 Appendix

2.7.1 Literature review – published studies of RNA-Sequencing in disease vectors

Databases – Web of Science
Date - 03/05/2022
Search terms –

1) (Anopheles OR Aedes OR Culex OR vector OR tsetse OR sandfly) AND (rna-seq OR rna-sequencing OR rna seq OR expression OR transcriptomics)
2) Mosquito AND (rna-seq OR rna-sequencing OR rna seq OR expression OR transcriptomics)

Title	Citation	Taxon	Phenotype / purpose	Year	Sequence data utilised
The RNA-Seq approach to studying the expression of mosquito mitochondrial genes	(Neira-Oviedo et al., 2011)	*Aedes aegypti, Anopheles gambiae, and Anopheles quadrimaculatus*	Mitochondria	2011	
RNA-seq analyses of blood-induced changes in gene expression in the mosquito vector species, *Aedes aegypti*	(Bonizzoni et al., 2011)	*Aedes aegypti*	Bloodmeal	2011	
Comparative Transcriptome Analyses of Deltamethrin-Resistant and -Susceptible *Anopheles gambiae* Mosquitoes from Kenya by RNA-Seq	(Bonizzoni et al., 2012)	*Anopheles gambiae*	Insecticide resistance (pyrethroid)	2012	
RNA-Seq reveals early distinctions and late convergence of gene expression between diapause and quiescence in the Asian tiger mosquito, *Aedes albopictus*	(Poelchau et al., 2013)	*Aedes albopictus*	Diapause	2013	
The Developmental Transcriptome of the Mosquito *Aedes aegypt*i, an Invasive Species and Major Arbovirus Vector	(Akbari et al., 2013)	*Aedes aegypti*	Life-stage	2013	
De novo transcriptome sequencing and sequence analysis of the malaria vector *Anopheles sinensis*	(Chen et al., 2014)	*Anopheles sinensis*	Annotation	2014	
Comparative analysis of response to selection with three insecticides in the dengue mosquito *Aedes aegypti* using mRNA sequencing	(David et al., 2014)	*Aedes aegypti*	Insecticide resistance (permethrin, imidacloprid, propoxur)	2014	Yes
Dual RNA-seq of Parasite and Host Reveals Gene Expression Dynamics during Filarial Worm–Mosquito Interactions	(Choi et al., 2014)	*Aedes aegypti*	Host-parasite	2014	
RNA-seq analyses of changes in the *Anopheles gambiae* transcriptome associated with resistance to pyrethroids in Kenya: identification of candidate-resistance genes and candidate-resistance SNPs	(Bonizzoni et al., 2015)	*Anopheles gambiae*	Insecticide resistance (pyrethroid)	2015	Yes
Comparative transcriptome analyses of deltamethrin-susceptible and -resistant *Culex pipiens pallens* by RNA-seq	(Lv et al., 2016)	*Culex pipiens*	Insecticide resistance (pyrethroid)	2016	
Single molecule RNA sequencing uncovers trans-splicing and improves annotations in *Anopheles stephensi*	(Jiang et al., 2017)	*Anopheles stephensi*	Annotation	2017	
Comparative Transcriptomics of Malaria Mosquito Testes: Function, Evolution, and Linkage	(Cassone et al., 2017)	*A. gambiae and A. merus*	Spermatogenesis	2017	
In the hunt for genomic markers of metabolic resistance to pyrethroids in the mosquito *Aedes aegypti*: An integrated next-generation sequencing approach	(Faucon et al., 2017)	*Aedes aegypti*	Insecticide resistance (pyrethroid)	2017	Yes
RNA-Seq Comparison of Larval and Adult Malpighian Tubules of the Yellow Fever Mosquito *Aedes aegypti* Reveals Life Stage-Specific Changes in Renal Function	(Li et al., 2017)	*Aedes aegypti*	Life-stage	2017	

The choreography of the chemical defensome response to insecticide stress: insights into the *Anopheles stephensi* transcriptome using RNA-Seq	(De Marco et al., 2017)	*Anopheles stephensi*	Insecticide resistance	2017	
Blood-induced differential gene expression in *Anopheles dirus* evaluated using RNA sequencing	(Mongkol et al., 2018)	*Anopheles dirus*	Bloodmeal	2018	
High-resolution transcriptional profiling of *Anopheles gambiae* spermatogenesis reveals mechanisms of sex chromosome regulation	(Taxiarchi et al., 2019)	*Anopheles gambiae*	Spermatogenesis	2019	
Transcriptome Sequencing and Analysis of Changes Associated with Insecticide Resistance in the Dengue Mosquito (*Aedes aegypti*) in Vietnam	(Lien et al., 2019)	*Aedes aegypti*	Insecticide resistance	2019	
Genome-wide gene expression profiling reveals that cuticle alterations and P450 detoxification are associated with deltamethrin and DDT resistance in *Anopheles arabiensis* populations from Ethiopia	(Simma et al., 2019)	*Anopheles arabiensis*	Insecticide resistance (pyrethroid)	2019	
UDP-glycosyltransferase genes and their association and mutations associated with pyrethroid resistance in Anopheles sinensis (Diptera: Culicidae)	(Zhou et al., 2019)	*Anopheles sinensis*	Insecticide resistance (pyrethroid)	2019	
RNASeq Analysis of *Aedes albopictus* Mosquito Midguts after Chikungunya Virus Infection	(Vedururu et al., 2019)	*Aedes albopictus*	Host-parasite	2019	
Contrasting patterns of gene expression indicate differing pyrethroid resistance mechanisms across the range of the New World malaria vector *Anopheles albimanus*	(Mackenzie-Impoinvil et al., 2019)	*Anopheles albimanus*	Insecticide resistance (pyrethroid)	2019	
Transcriptome analysis of *Anopheles dirus* and *Plasmodium vivax* at ookinete and oocyst stages	(Boonkaew et al., 2020)	*Anopheles dirus*	Plasmodium infection	2020	
Transcript Assembly and Quantification by RNA-Seq Reveals Significant Differences in Gene Expression and Genetic Variants in Mosquitoes of the *Culex pipiens* (Diptera: Culicidae) Complex	(Kang et al., 2021)	*Culex pipiens*	Insecticide resistance (pyrethroid)	2021	Yes
Integration of whole genome sequencing and transcriptomics reveals a complex picture of the reestablishment of insecticide resistance in the major malaria vector *Anopheles coluzzii*	(Ingham et al., 2021)	*Anopheles gambiae*	Insecticide resistance (pyrethroid)	2021	
Transcriptome comparison of dengue-susceptible and -resistant field derived strains of Colombian *Aedes aegypti* using RNA-sequencing	(Coatsworth et al., 2021)	*Aedes aegypti*	Vector competence	2021	
Transcriptomic and proteomic analysis of pyrethroid resistance in the CKR strain of *Aedes aegypti*	(Sun et al., 2021)	*Aedes aegypti*	Insecticide resistance (pyrethroid)	2021	
RNA-Seq analysis of blood meal induced gene-expression changes in *Aedes aegypt*i ovaries	(Nag et al., 2021)	*Aedes aegypti*	Bloodmeal	2021	
Sympatric Populations of the *Anopheles gambiae* Complex in Southwest Burkina Faso Evolve Multiple Diverse Resistance Mechanisms in Response to Intense Selection Pressure with Pyrethroids	(Williams et al., 2022)	*Anopheles gambiae*	Insecticide resistance (pyrethroid)	2022	
RNAseq-based gene expression profiling of the *Anopheles funestus* pyrethroid-resistant strain FUMOZ highlights the predominant role of the duplicated CYP6P9a/b cytochrome P450s	(Wondji et al., 2022)	*Anopheles funestus*	Insecticide resistance (pyrethroid)	2022	
Transcriptome profiling reveals sex-specific gene expressions in pupal and adult stages of the mosquito *Culex pipiens*	(Martynova et al., 2022)	*Culex pipiens*	Life-stage	2022	
A whole transcriptomic approach provides novel insights into the molecular basis of organophosphate and pyrethroid resistance in *Anopheles arabiensis* from Ethiopia	(Messenger et al., 2021)	*Anopheles arabiensis*	Insecticide resistance	2022	Yes

2.7.2 Colony selection regime

Blood fed *Anopheles gambiae s.s* were collected from Busia, Uganda, in November 2018, using a prokopack aspirator. Approximately 200 mosquitoes were collected from 12 different homes in the village of South-Bugwere. Eggs were transported to the Liverpool School of Tropical Medicine and mosquitoes were reared in the insectaries at approximately 75% RH and 27°C, with a 12:12 hour light:dark photoperiod.

Between generations 10 and 24, due to the COVID-19 pandemic, Busia mosquitoes were not selected against deltamethrin to maintain their insecticide resistance status. As a result, the colony had lost resistance by G24, displaying 100% mortality to a 1 hour 0.05% deltamethrin (1X) WHO paper exposure, and 92.6% mortality to permethrin 0.75% (1X). Mosquitoes were then selected on 1X deltamethrin WHO papers for four consecutive generations (G24-G27), initially of 15 minute exposures, and subsequently for one hour.

A) Profiling

Generation	Insecticide	Dead	Total	Mortality (%)
G24	Deltamethrin	99	99	100.0
G24	Permethrin	88	95	92.6
G25	Deltamethrin	47	54	87.0
G25	Permethrin	15	47	31.9
G26	Permethrin	9	30	30.0
G27	Deltamethrin	164	217	75.6
G28	Deltamethrin	145	208	69.7
G28	Permethrin	23	106	21.7

B) Selections

1X deltamethrin (0.05%)

Generation	Exposure.(mins)	Dead	Total	Mortality (%)
G24	15	1111	1205	92.2
G25	15	348	456	76.3
G26	60	265	327	81.0
G27	60	164	217	75.6

2.7.3 Sequencing depth

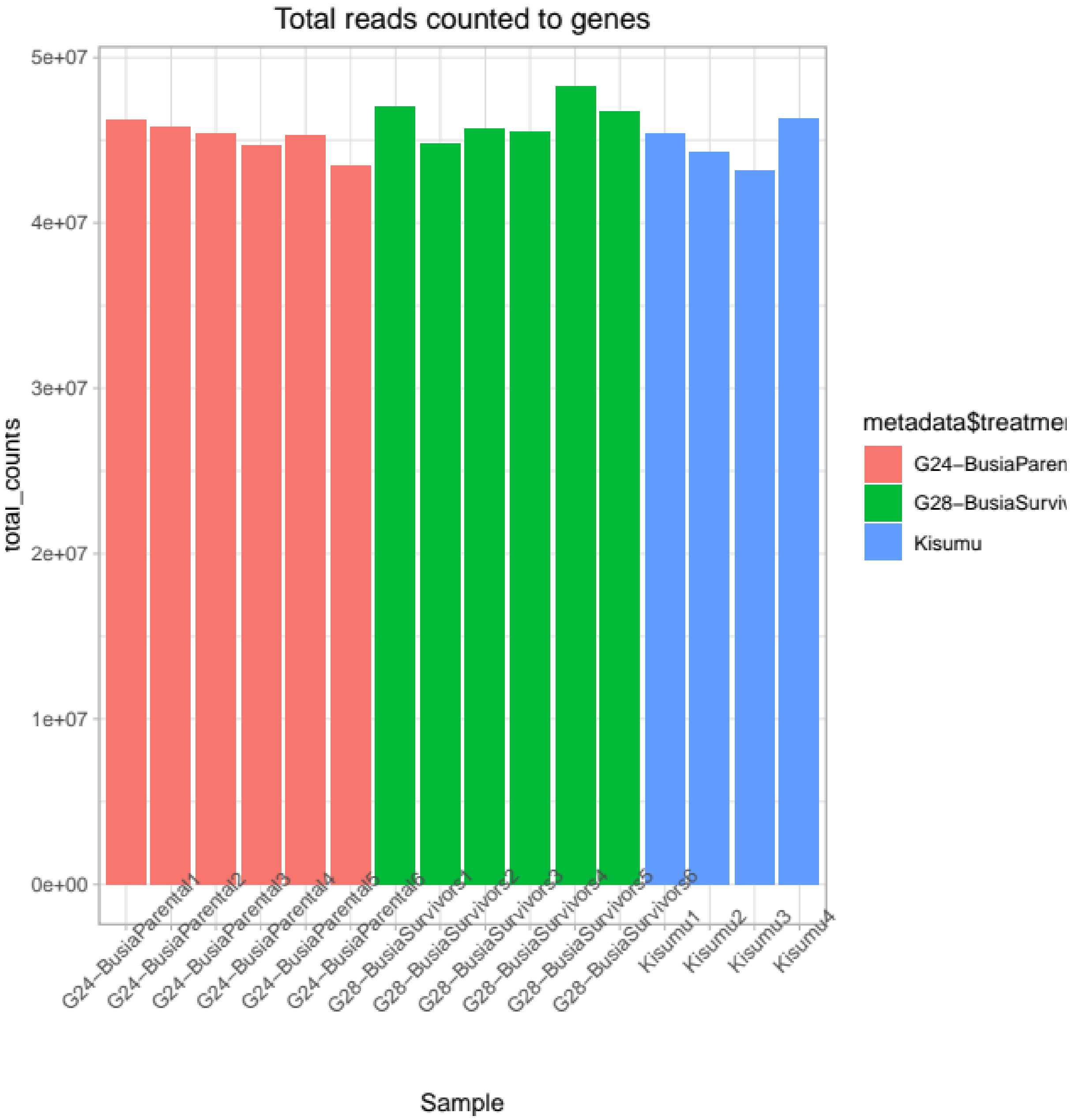
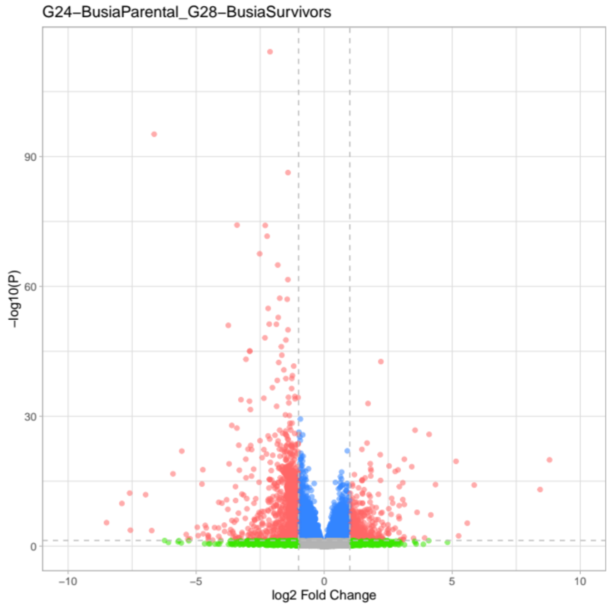


Fig A3. Total reads counted to genes in each sample. Read quantification is performed by kallisto.

2.7.4 Volcano plots



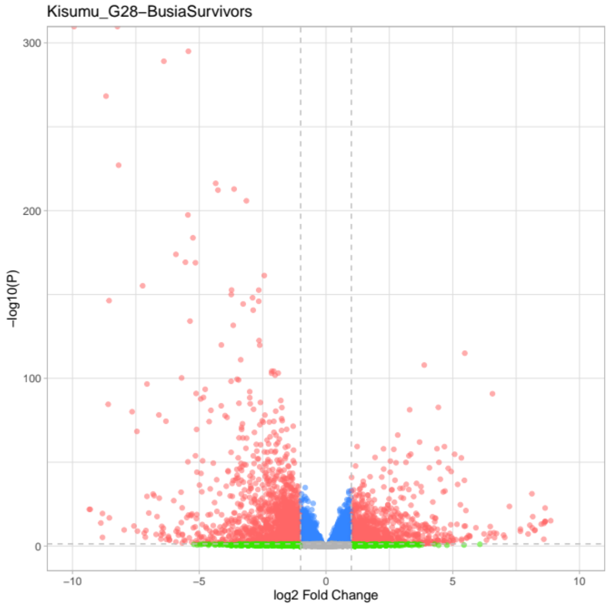
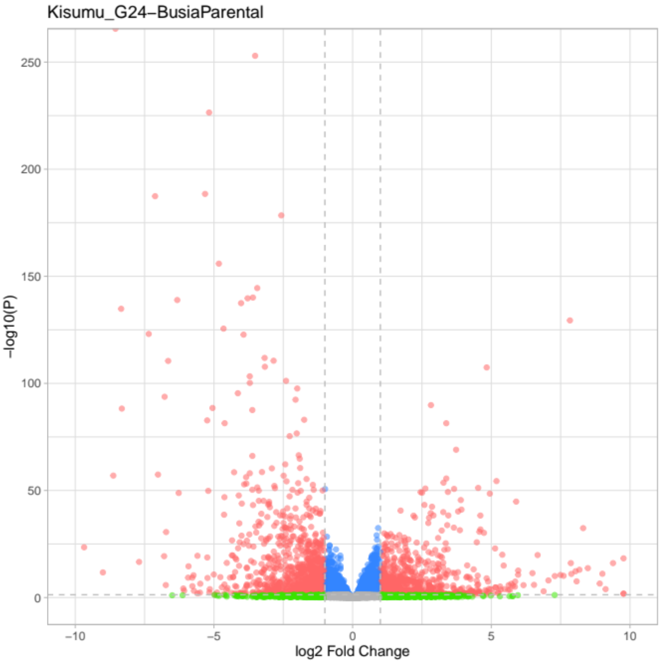


Figure A4: A volcano plot showing gene expression differences between Busia Parental and Busia Selected survivors. -Log10 P-values are plotted against Log2 Fold Change. Red=genes with adjusted p-value < 0.05 and an absolute fold change > 2, green= adjusted pvalue > 0.05 and a fold change > 2, blue= adjusted pvalue > 0.05 and a fold change < 2. An outlier (AGAP012637) has been removed for visualisation purposes.



2.7.5 Example heatmaps

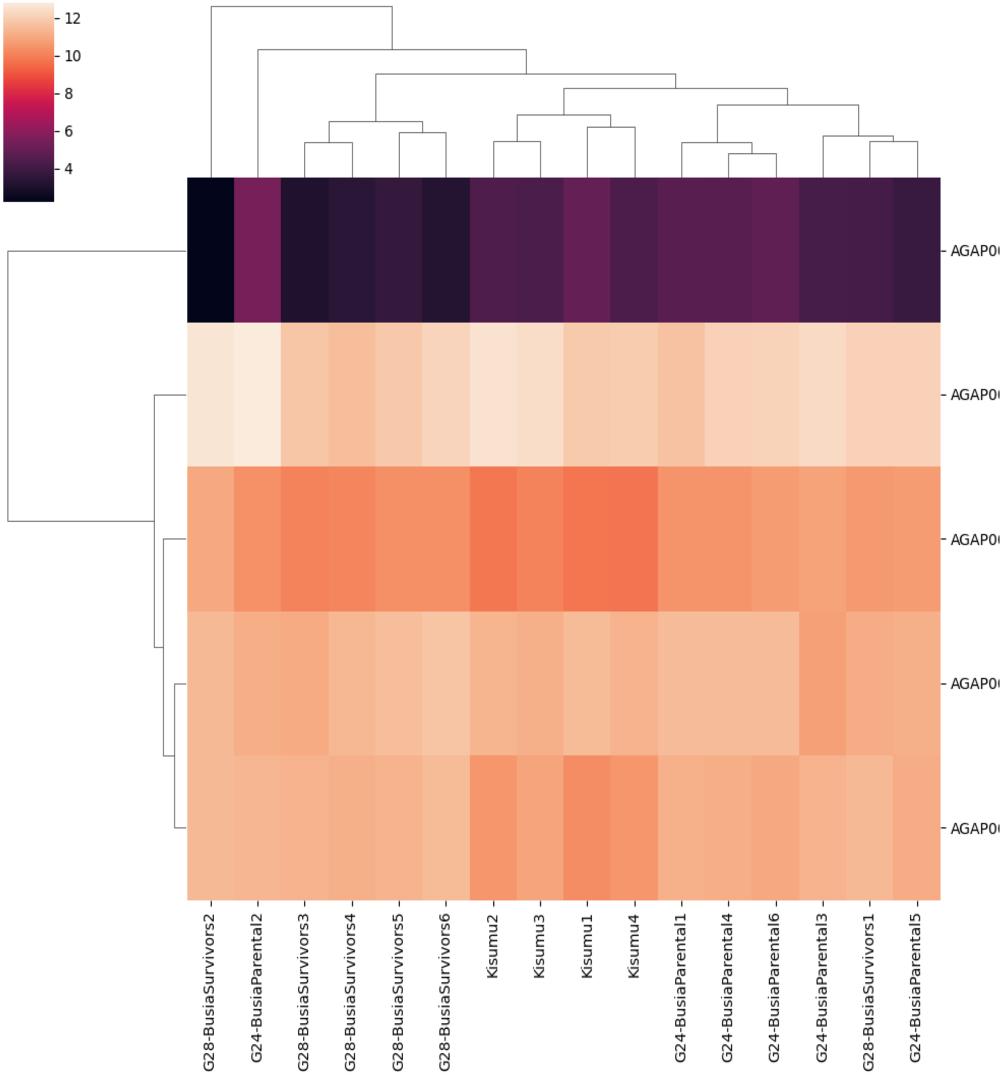


Figure A5. An example heatmap showing read counts in each sample.

2.7.6 Venn diagrams

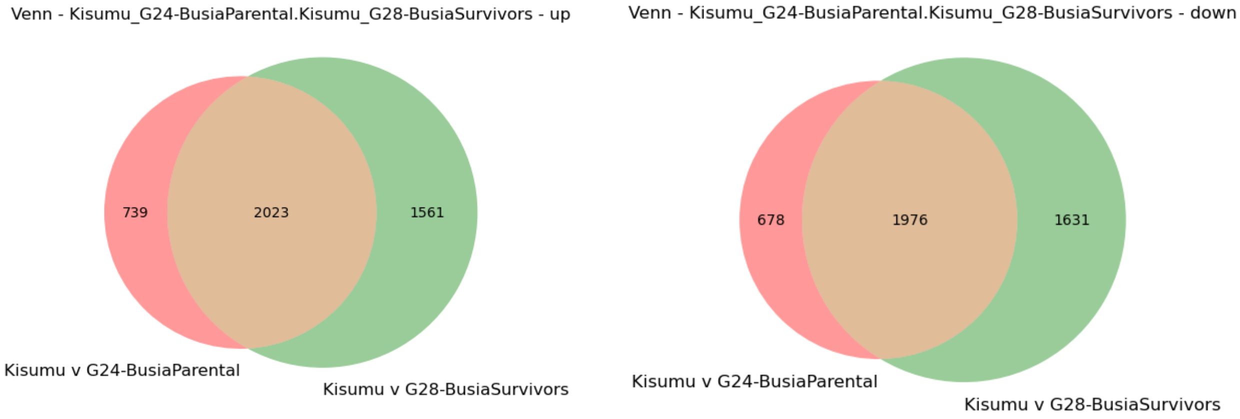
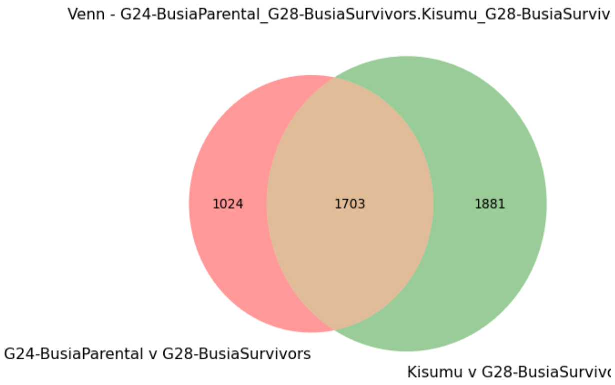


Figure A6. Venn diagrams of differentially expressed genes (left: upregulated, right: downregulated).



2.7.7 OLS of SNPs per gene

Regression of log2 number of SNPs per gene, total read counts per gene, and gene size in (bp).

```
 OLS Regression Results
==============================================================================
Dep. Variable:              nSNPs     R-squared:                  0.431
Model:                        OLS               Adj. R-squared:            0.430
Method:              Least Squares    F-statistic:                449.4
Date:              Fri, 21 Jan 2022             Prob (F-statistic):        5.48e-146
Time:                      13:11:34             Log-Likelihood:            -1734.4
No. Observations:              1189   AIC:                        3475.
Df Residuals:                 1186              BIC:                       3490.
Df Model:                        2
Covariance Type:            nonrobust
==============================================================================
              coef    std err        t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -2.4515     0.257     -9.529     0.000     -2.956     -1.947
Readcounts    0.1462     0.011     13.698     0.000      0.125      0.167
GeneSize      0.4661     0.017     26.811     0.000      0.432      0.500
==============================================================================
Omnibus:                  87.964                Durbin-Watson:             1.592
Prob(Omnibus):               0.000    Jarque-Bera (JB):           106.486
Skew:                     -0.697                Prob(JB):                  7.53e-24
Kurtosis:                  3.457                Cond. No.                  160.
==============================================================================
```

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

2.7.8 PCA on genotypes

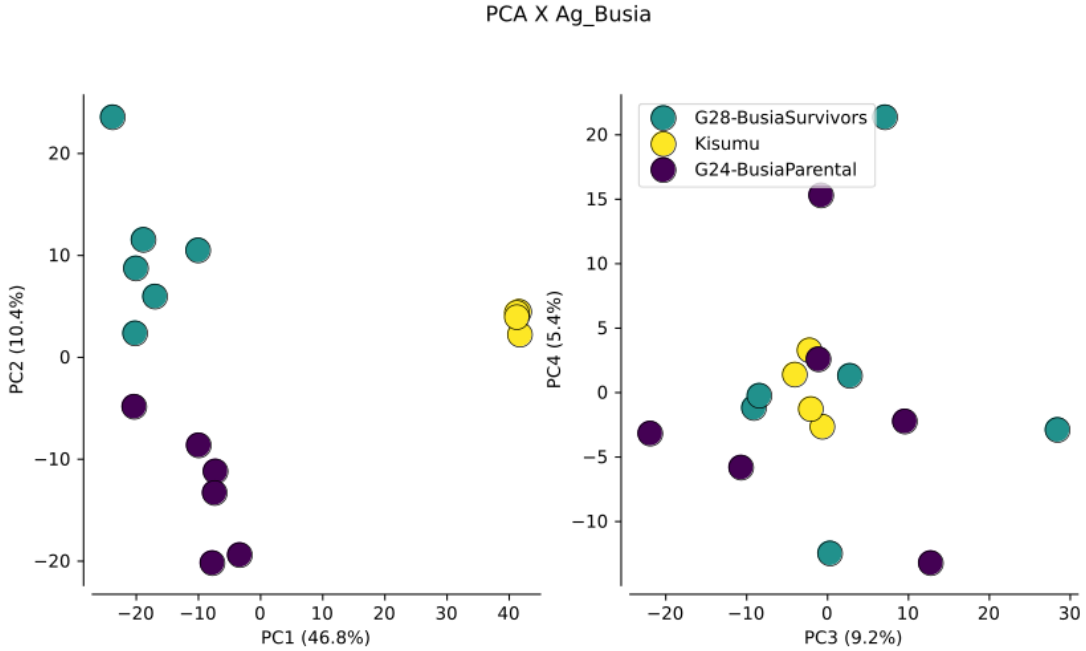
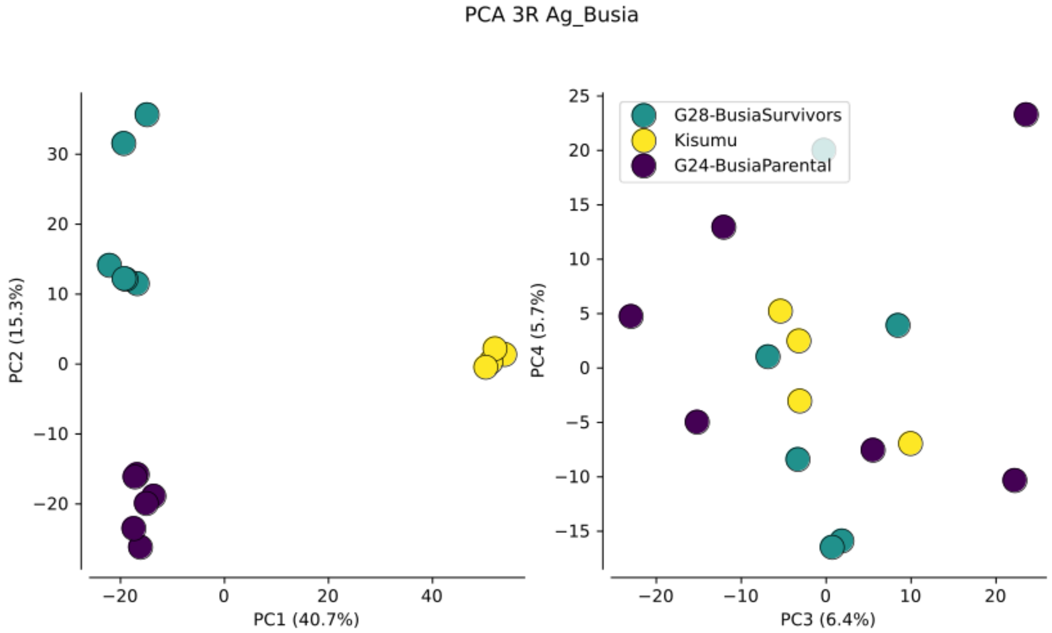
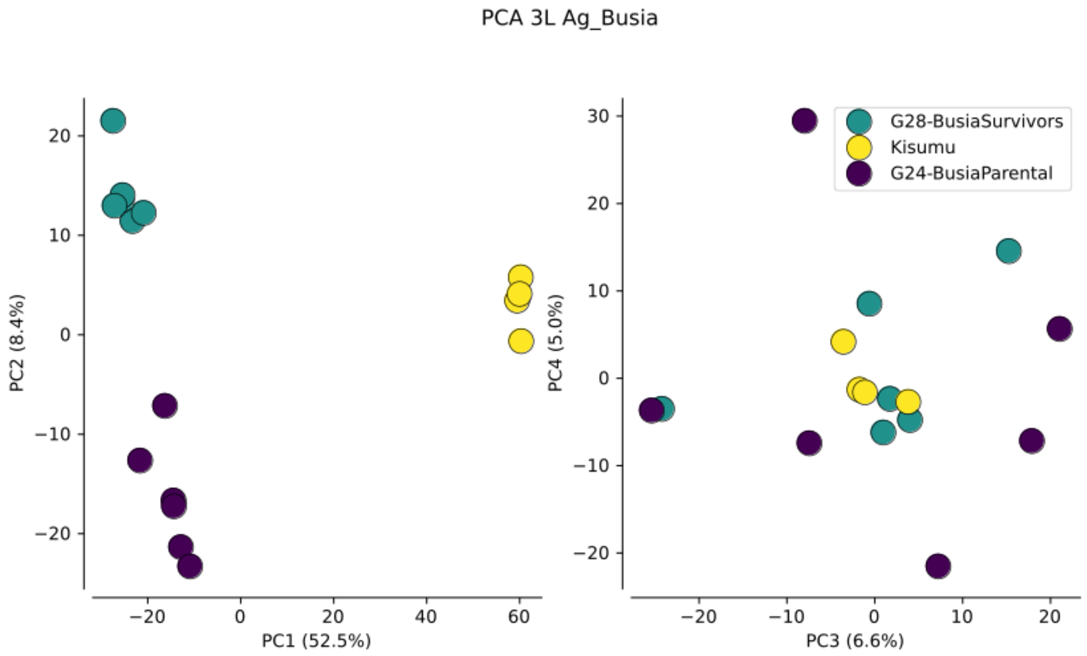
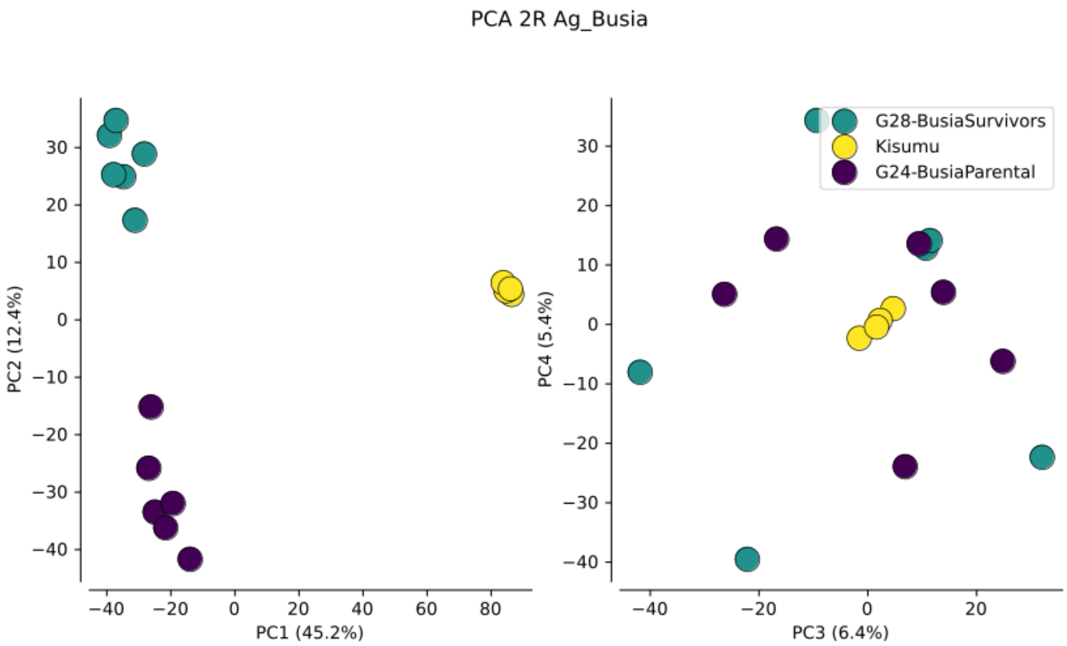
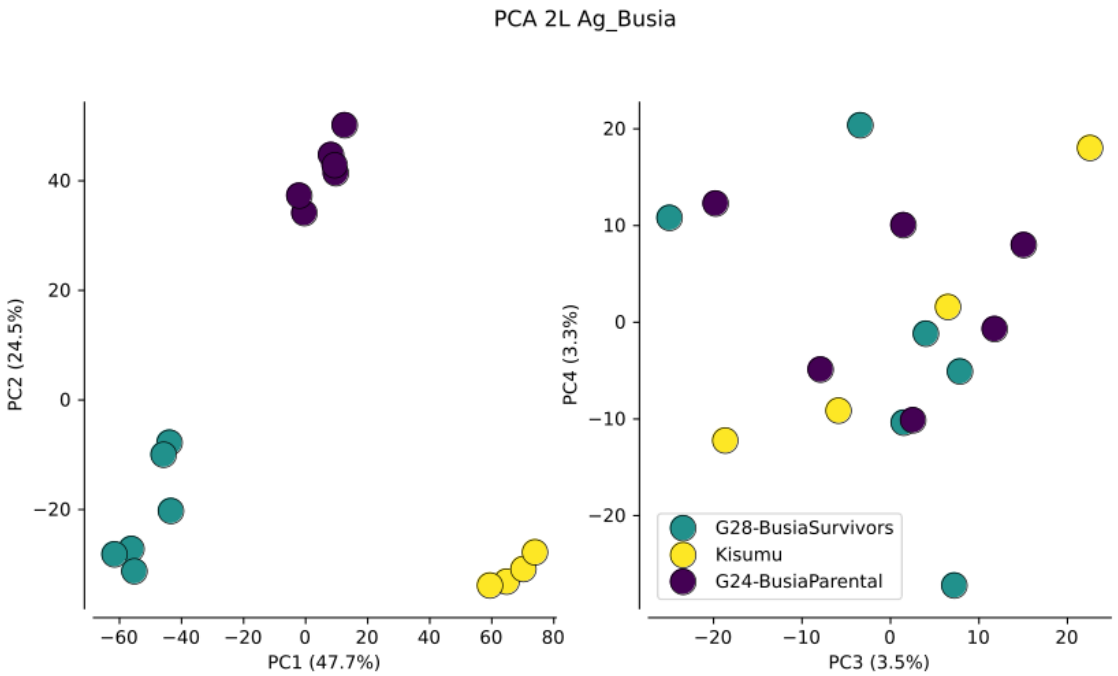


Figure A8. Principal components analysis (PCA) on called genotypes on the X chromosome.




PCA 3L Ag_Busia
PC2 (8.4%)
PC1 (52.5%)
20
10
0
−10
−20
−20
0
20
40
60
PC4 (5.0%)
PC3 (6.6%)
30
20
10
0
−10
−20
−20
−10
0
10
20
G28-BusiaSurvivors
Kisumu
G24-BusiaParental


PCA 2R Ag_Busia
PC2 (12.4%)
PC1 (45.2%)
30
20
10
0
−10
−20
−30
−40
−40
−20
0
20
40
60
80
PC4 (5.4%)
PC3 (6.4%)
30
20
10
0
−10
−20
−30
−40
−40
−20
0
20
G28-BusiaSurvivors
Kisumu
G24-BusiaParental



PCA 2L Ag_Busia

2.7.9 Genetic diversity

A) Wattersons Theta

Contig	G28 Busia Survivors	G24 Busia Parental	Kisumu
2L	0.00057	0.00068	0.00056
2R	0.00066	0.00083	0.00044
3L	0.00034	0.00059	0.00036
3R	0.00053	0.00067	0.00043
X	0.00068	0.00075	0.00025

B) Nucleotide Diversity

Contig	G28 Busia Survivors	G24 Busia Parental	Kisumu
2L	0.00053	0.00102	0.00084
2R	0.00097	0.00116	0.00065
3L	0.00038	0.00069	0.00048
3R	0.00075	0.00092	0.00064
X	0.00097	0.00123	0.00034

2.7.10 Hudson Fst Per chromosome (G24 Busia Parental v G28 Busia survivors)

Contig	Fst
2L	0.431
2R	0.109
3R	0.093
3L	0.11
X	0.106

2.7.11 Proportion of ancestry per chromosome

Strain	Contig	AIM fraction gambiae	AIM fraction coluzzii	n_aims
G28 Busia survivors	2L	0.981	0.000	53
G28 Busia survivors	2R	0.796	0.093	18
G28 Busia survivors	3L	0.871	0.062	15
G28 Busia survivors	3R	0.836	0.164	16
G28 Busia survivors	X	0.946	0.035	161
G24 Busia Parental	2L	0.981	0.000	53
G24 Busia Parental	2R	0.778	0.111	18
G24 Busia Parental	3L	0.887	0.047	15
G24 Busia Parental	3R	0.836	0.164	16
G24 Busia Parental	X	0.945	0.036	161
Kisumu	2L	0.105	0.875	51
Kisumu	2R	0.185	0.703	18
Kisumu	3L	0.057	0.877	15
Kisumu	3R	0.170	0.830	16
Kisumu	X	0.966	0.014	154

2.7.12 Karyotype frequencies calculated by compKaryo.

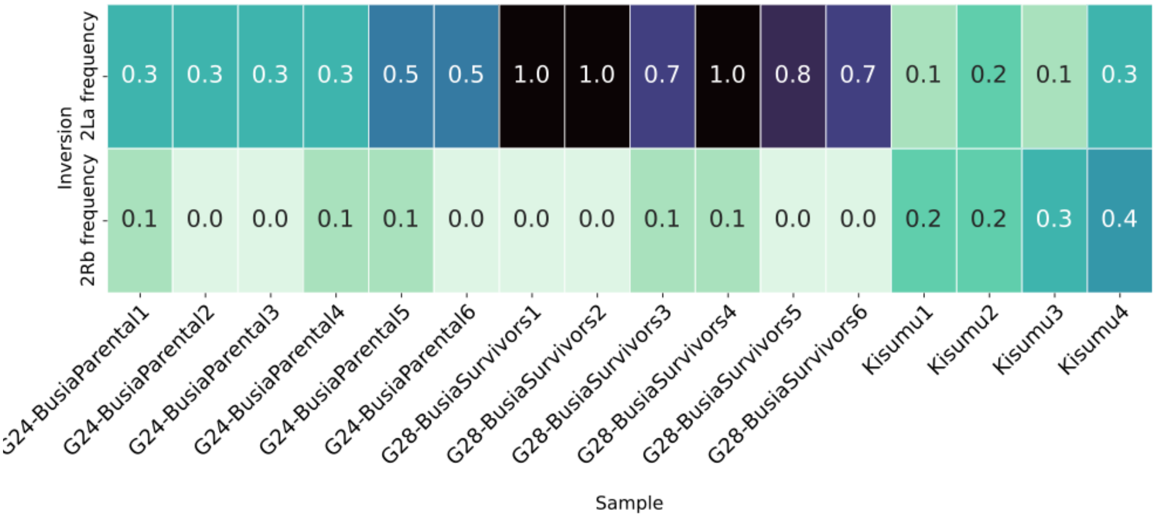


Figure A12: Karyotype frequencies. The frequency of the 2La and 2Rb karyotypes in a) each biological replicate and b) averaged across experimental conditions

3

AgamPrimer

This work is now on bioRxiv and has been prepared for submission to Wellcome Open Research in the near future as a software tool. It follows the format Introduction-Methods and Implementation-Discussion. Preprint doi: https://doi.org/10.1101/2022.12.31.521737

AgamPrimer: Primer Design in *Anopheles gambiae* informed by range-wide genomic variation

Sanjay Curtis Nagi[1*], Alistair Miles[2], Martin J Donnelly[1]

[1]Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, UK.

[2]Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SD, UK

[*]Corresponding author: Email: Sanjay.Nagi@lstmed.ac.uk

Keywords: PCR, Anopheles, Primers, Probes, Molecular biology

3.1 Abstract

The major malaria mosquito, *Anopheles gambiae s.l*, is one of the most studied organisms in medical research and also one of the most genetically diverse. When designing polymerase chain reaction (PCR) or hybridisation-based molecular assays, reliable primer and probe design is crucial. However, single nucleotide polymorphisms (SNPs) in primer binding sites can prevent primer binding, leading to null alleles, or bind suboptimally, leading to preferential amplification of specific alleles. Given the extreme genetic diversity of *An. gambiae*, researchers need to consider this genetic variation when designing primers and probes to avoid amplification problems. In this note, we present a python package, AgamPrimer, which exploits the AG1000G dataset and allows users to rapidly design primers in *An. gambiae*, whilst summarising genetic variation in the primer binding sites and visualising the position of primer pairs. AgamPrimer allows the design of both genomic DNA and cDNA primers and hybridisation probes. By coupling this python package with Google Colaboratory, AgamPrimer is an open and accessible platform for primer and probe design, hosted in the cloud for free. AgamPrimer is available here https://github.com/sanjaynagi/AgamPrimer and we hope it will be a useful resource for the community to design probe and primer sets that can be reliably deployed across the *An. gambiae* species range.

3.2 Introduction

The polymerase chain reaction (PCR) is ubiquitous in molecular biology, providing template sequence for a wide array of techniques, such as detecting the presence or absence of particular DNA sequences, quantifying the abundance of transcripts, or in Sanger and next-generation sequencing. Primers - short, single-strand DNA sequences which bind to the template and facilitate amplification - are crucial to effective PCR reactions and must be designed to be robust, reliable and consistent across experimental conditions.

Single nucleotide polymorphisms (SNPs) in primer binding sites can affect both the stability of the primer-template duplex, as well as the efficiency with which DNA polymerases can extend the primer (Letowski *et al.*, 2004; Wu *et al.*, 2009). In some cases, this can completely prevent primer binding and amplification of the template DNA, often referred to as null alleles or allelic dropout (Carlson *et al.*, 2006). On most genotyping platforms, these alleles are problematic and difficult to detect, as null allele heterozygotes will be indistinguishable from true homozygous individuals. Allelic dropout is known to cause problems in human genetic testing (Zajícková *et al.*, 2003; Silva *et al.*, 2017). Null allele homozygotes could be suggested if a sample repeatedly fails to amplify, however, when performing PCR on pooled samples we would not observe this failure, and therefore can never know whether all samples amplified successfully. Ensuring genetic markers do not violate Hardy-Weinberg equilibrium (HWE) is one way to partially safeguard against this problem (Chapuis and Estoup, 2007), however, this is not always performed in practice, and excluding such markers may lead to loss of information when HWE deviation has another cause.

Another problematic scenario occurs if primers do bind but with unequal efficiency against different genetic variants. In this case, any molecular assay that is quantitative, such as qPCR for gene expression, could be severely affected and lead to biases in the estimation of sequence abundance between genetic variants or strains (Lefever *et al.*, 2013). A previous study found that single mismatches can introduce a range of impacts on Cq values, ranging from relatively minor (<1.5) to major (>7.0) (Stadhouders *et al.*, 2010). The impact of a variant on primer binding depends on multiple factors but mismatches within the last 5 nucleotides at the 3’ end can disrupt the nearby polymerase active site, and so these mismatches tend to have a much greater impact (Stadhouders *et al.*, 2010; Martins *et al.*, 2011). Primers should therefore be designed to avoid these sites or if unavoidable, to contain degenerate bases at the sites of SNPs, in order to maximise the robustness of molecular experiments (Quinlan and Marth, 2007).

The *Anopheles gambiae* 1000 genomes project has revealed staggering amounts of genetic variation in the major malaria mosquito, *Anopheles gambiae s.l (Miles et al., 2017)* with a segregating SNP in less than every 2 bases of the accessible genome (Ag1000G, 2020). Despite this, the vast majority of primers designed to target the *Anopheles gambiae* *s.l* genome do not consider SNP variation. In the past, this was not straightforward, as it would require both handling large genomic datasets and matching designed primers to genomic positions. Thanks to recent advances in cloud computing and the malariagen_data API, we can now design primers in the cloud whilst checking for genetic variation in the *Anopheles gambiae* 1000 genomes project. In this note we present AgamPrimer, a python package which is coupled with a Google Colaboratory notebook, allowing users to easily design primers and probes in the cloud whilst considering genetic variation in *Anopheles gambiae s.l.*

3.3 Methods and Implementation

AgamPrimer is a two-phase process, first designing sets of primers and probes and secondly investigating SNP variation in the targeted sites. AgamPrimer uses Primer3 as the core primer design engine, in the form of Primer3-py. Primer3 is open-source and has become the *de-facto* standard for primer design for molecular biology. Primer3-py is a set of recently developed python bindings for the Primer3 program (Untergasser *et al.*, 2012), which can be run readily in a Google Colaboratory environment. To load genetic variation data from the *Anopheles* 1000 genomes project, we integrate the malariagen_data API, which allows rapid download and analysis of genomic data from the cloud. Integration of the PyData stack in malariagen_data allows users to perform rapid genomic analysis on large datasets where compute resources are modest, such as in Google Colaboratory notebooks. Google Colaboratory is a proprietary version of Jupyter Notebook and is provided for free alongside CPU and GPU access to anyone with a Google account.

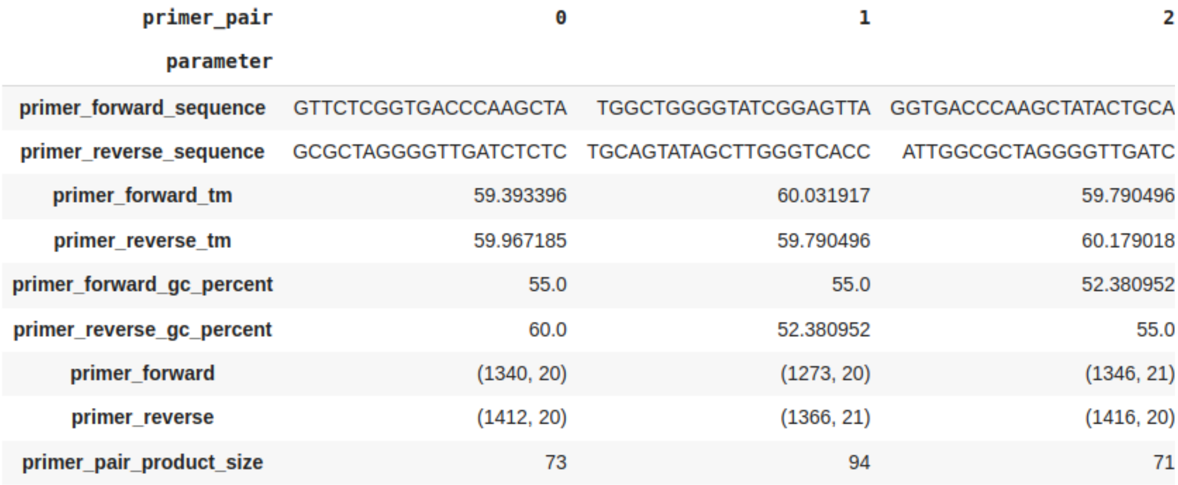
AgamPrimer can be run in two ways, either running the full Colaboratory notebook in a step-wise fashion, or in a single command which produces all outputs, which may be preferred in more high-throughput primer design settings. Users may select primer design parameters by providing a python dictionary, or the primer3 default parameters can be used.

3.3.1 Primer Design with primer3

AgamPrimer allows the design of genomic DNA primers, hybridisation probes or cDNA primers (for gene expression purposes). In the case of cDNA primers, one of the forward or reverse primers will be designed to span an exon-exon junction where available, to prevent the amplification of genomic DNA in the sample.

Table 1 shows the output from the initial phase of primer design. AgamPrimer reports the primer sequences, along with information on melting point, GC content, amplicon size and position in the target sequence, though the full Primer3 output is accessible to the user. The user may specify the number of desired primer pairs to design. After the Primer3 run, AgamPrimer will print out run statistics which may be useful for troubleshooting.

Table 1. Primer3 results: A pandas dataframe generated by AgamPrimer. Useful information from about each primer set are stored, such as the sequence, melting temperature and GC content.



primer_pair	0	1	2
parameter			
primer_forward_sequence	GTTCTCGGTGACCCAAGCTA	TGGCTGGGGTATCGGAGTTA	GGTGACCCAAGCTATACTGCA
primer_reverse_sequence	GCGCTAGGGGTTGATCTCTC	TGCAGTATAGCTTGGGTCACC	ATTGGCGCTAGGGGTTGATC
primer_forward_tm	59.393396	60.031917	59.790496
primer_reverse_tm	59.967185	59.790496	60.179018
primer_forward_gc_percent	55.0	55.0	52.380952
primer_reverse_gc_percent	60.0	52.380952	55.0
primer_forward	(1340, 20)	(1273, 20)	(1346, 21)
primer_reverse	(1412, 20)	(1366, 21)	(1416, 20)
primer_pair_product_size	73	94	71

3.3.2 Interrogating the ag3 resource

The malariagen_data python package pulls in Ag1000g data from the cloud, facilitating rapid analysis of over 10,000 *An. gambiae s.l* whole genomes from throughout sub-Saharan Africa. In step 1 of the primer design process, we record the genomic positions of the designed primers, and in step 2 use these coordinates to extract SNP allele frequency information for given Ag1000g samples of choice. In the Colaboratory notebook, we generate a summary table of the Ag1000g inventory, counting samples by taxon, sample set and country, to guide users in selecting an appropriate cohort. Through the use of sample set identifiers, and sample queries (following standard pandas syntax), users may select any group of samples in the dataset to interrogate. Alternatively, the default settings will use every available mosquito genome. A sample query can be performed on any column of the sample metadata, such as species (taxon), country, year or location, amongst other metadata.

We then generate an interactive plot (Figure 1) which shows SNP variation in designed primer binding sites, in the user-selected Ag1000g cohort. The user can hover over points, which returns the exact frequencies of each nucleotide at that genomic position, which may be useful in the case where the user would prefer to design degenerate primers, as opposed to avoiding that primer set entirely. The plot also highlights the 3' and 5' prime ends, as well as the genomic span, GC content and melting temperature, allowing the user to easily and rapidly identify suitable oligonucleotides.

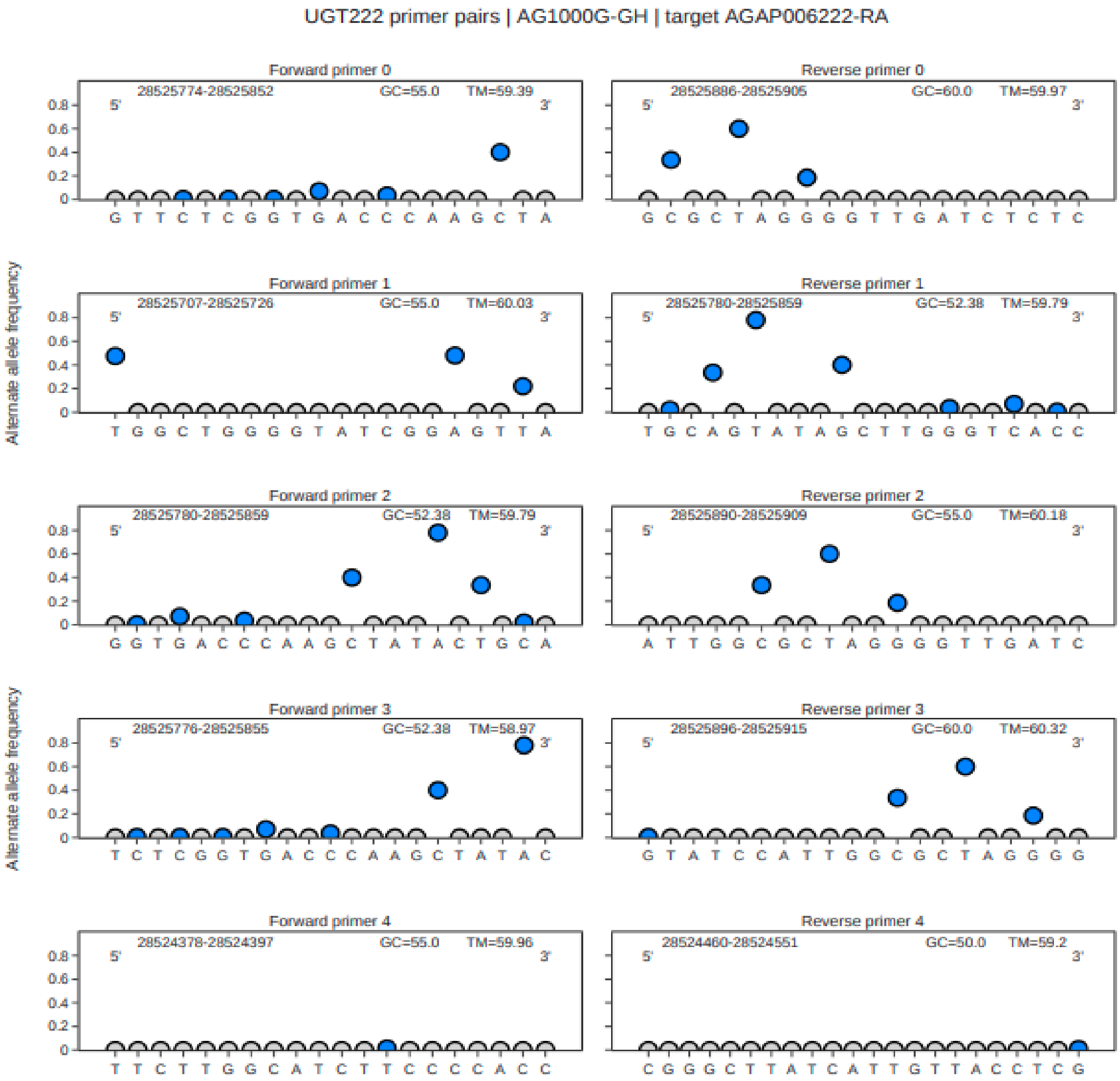


Figure 1. Illustrative plots showing allele frequencies in primer binding sites targeting the AGAP006222-RA transcript in specimens of *An. gambiae ss.* and *An. coluzzii* from Ghana. An interactive plot with Plotly displays the primer or probe sequences from 5' to 3', with circles indicating the summed alternate allele frequency at that genomic position. Blue circles indicate segregating SNPs, and grey circles indicate sites which are invariant in the ag3 cohort of choice. The genomic span of each oligo is displayed alongside the GC content and Tm.

3.3.3 Genomic location of primers

AgamPrimer then plots the position of the primer in the genome, in relation to any nearby exons. In Figure 2, we can see that all but one primer pair were designed at the Exon 4 and

5 boundary. Primer pair 4, which targets the Exon 1 and 2 junction, contains much less SNP variation than the other primers.

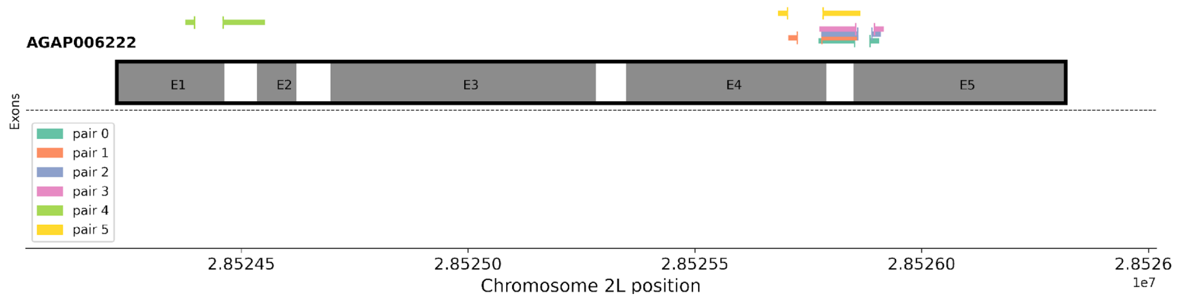


Figure 2. The genomic locations of designed primer sets in relation to nearby exons. Primers spanning exons are shown as expanded to clearly illustrate span of the whole junction for visualisation purposes, and only contain sequence at each extremity.

To ensure the specificity of the designed primer and probes for only one genomic location, we align oligonucleotides to the AgamP3 genome with BLAT, using the gget python package API.

3.3.4 Checking existing oligonucleotides for variation

During the development of AgamPrimer, we presented the package and notebook to stakeholders and prospective users within a series of bioinformatics training workshops for partners at PAMCA. A stakeholder remarked that they have primer sets which fail in some populations of *An. gambiae*, and wondered whether in AgamPrimer it would be possible to check existing primers for variation in the Ag1000g.

We present functionality which takes as input an oligonucleotide DNA sequence, such as a primer or probe, and will return the frequencies of any SNPs across that sequence in the ag3 cohort of choice. We use the gget package to rapidly align the oligo sequence to the AgamP3 genome using BLAT, to retrieve genomic coordinates. We then use Plotly to produce base frequency plots as in the main AgamPrimer workflow.

3.4 Discussion

AgamPrimer integrates Primer3-py and the malariagen_data API to rapidly and conveniently design variation-informed primers and probes for molecular biology. Through the use of forms in Colaboratory, users are able to define their own parameters, which means that the AgamPrimer notebook does not require programming skills. This is an extremely important point, as we hope the tool will be useful for all researchers including molecular biologists who may not have programming experience. We provide a walkthrough video linked at the beginning of the notebooks to guide users.

Genomic surveillance of malaria mosquitoes is becoming increasingly important, with a number of high throughput amplicon sequencing panels having been developed to identify species across the entire *Anopheles* genus (Boddé *et al.*, 2022; Makunin *et al.*, 2022), within the *An. gambiae* complex (Caputo *et al.*, 2021), and to karyotype samples (Love *et al.*, 2020). In the near future, it is likely that other amplicon sequencing panels will be designed to target phenotypes of interest, such as insecticide resistance, gene drive resistance, or vector competence. Just as in more standard genotyping assays, in amplicon sequencing, robust primer design is crucial, and therefore having a computational framework to design primers will prove invaluable.

AgamPrimer is open-source and free to use, and we encourage suggestions and contributions to the software. The package is available here https://github.com/sanjaynagi/AgamPrimer and on the python package index (PyPi).

3.5 References

Ag1000G (2020) 'Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii', *Genome Res*, pp. 1–14.

Boddé, M., Makunin, A., Ayala, D., Bouafa, L., Diabaté, A., Ekpo, U.F., Kientega, M., Le Goff, G., Makanga, B.K., Ngangue, M.F., Omitola, O.O., Rahola, N., Tripet, F., Durbin, R. and Lawniczak, M.K.N. (2022) 'High resolution species assignment of Anopheles mosquitoes using k-mer distances on targeted sequences', *bioRxiv*. doi:10.1101/2022.03.18.484650.

Caputo, B., Pichler, V., Bottà, G., De Marco, C., Hubbart, C., Perugini, E., Pinto, J., Rockett, K.A., Miles,

A. and Della Torre, A. (2021) ‘Novel genotyping approaches to easily detect genomic admixture between the major Afrotropical malaria vector species, Anopheles coluzzii and An. gambiae’, *Molecular ecology resources*, 21(5), pp. 1504–1516.

Carlson, C.S., Smith, J.D., Stanaway, I.B., Rieder, M.J. and Nickerson, D.A. (2006) ‘Direct detection of null alleles in SNP genotyping data’, *Human molecular genetics*, 15(12), pp. 1931–1937.

Chapuis, M.-P. and Estoup, A. (2007) ‘Microsatellite null alleles and estimation of population differentiation’, *Molecular biology and evolution*, 24(3), pp. 621–631.

Lefever, S., Pattyn, F., Hellemans, J. and Vandesompele, J. (2013) ‘Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative PCR assays’, *Clinical chemistry*, 59(10), pp. 1470–1480.

Letowski, J., Brousseau, R. and Masson, L. (2004) ‘Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays’, *Journal of microbiological methods*, 57(2), pp. 269–278.

Love, R.R., Pombi, M., Guelbeogo, M.W., Campbell, N.R., Stephens, M.T., Dabire, R.K., Costantini, C., Della Torre, A. and Besansky, N.J. (2020) ‘Inversion Genotyping in the Anopheles gambiae Complex Using High-Throughput Array and Sequencing Platforms’, *G3* , 10(9), pp. 3299–3307.

Makunin, A., Korlević, P., Park, N., Goodwin, S., Waterhouse, R.M., von Wyschetzki, K., Jacob, C.G., Davies, R., Kwiatkowski, D., St Laurent, B., Ayala, D. and Lawniczak, M.K.N. (2022) ‘A targeted amplicon sequencing panel to simultaneously identify mosquito species and Plasmodium presence across the entire Anopheles genus’, *Molecular ecology resources*, 22(1), pp. 28–44.

Martins, E.M., Vilarinho, L., Esteves, S., Lopes-Marques, M., Amorim, A. and Azevedo, L. (2011) ‘Consequences of primer binding-sites polymorphisms on genotyping practice’, *Open journal of genetics*, 01(02), pp. 15–17.

Miles, A., Harding, N.J., Bottà, G., Clarkson, C.S., Antão, T., Kozak, K., Schrider, D.R., Kern, A.D., Redmond, S., Sharakhov, I., Pearson, R.D., Bergey, C., Fontaine, M.C., Donnelly, M.J., Lawniczak, M.K.N., *et al.* (2017) ‘Genetic diversity of the African malaria vector Anopheles gambiae’, *Nature*, 552, pp. 96–100.

Quinlan, A.R. and Marth, G.T. (2007) ‘Primer-site SNPs mask mutations’, *Nature methods*, 4(3), p. 192.

Silva, F.C., Torrezan, G.T., Brianese, R.C., Stabellini, R. and Carraro, D.M. (2017) ‘Pitfalls in genetic testing: a case of a SNP in primer-annealing region leading to allele dropout in BRCA1’, *Molecular genetics & genomic medicine*, 5(4), pp. 443–447.

Stadhouders, R., Pas, S.D., Anber, J., Voermans, J., Mes, T.H.M. and Schutten, M. (2010) ‘The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5’ nuclease assay’, *The Journal of molecular diagnostics: JMD*, 12(1), pp. 109–117.

Untergasser, Cutcutache and Koressaar (2012) ‘Primer3—new capabilities and interfaces’, *Nucleic acids and molecular biology* [Preprint]. Available at: https://academic.oup.com/nar/article-abstract/40/15/e115/1223759.

Wu, J.-H., Hong, P.-Y. and Liu, W.-T. (2009) ‘Quantitative effects of position and type of single mismatch on single base primer extension’, *Journal of microbiological methods*, 77(3), pp. 267–275.

Zajícková, K., Krepelová, A. and Zofková, I. (2003) ‘A single nucleotide polymorphism under the reverse primer binding site may lead to BsmI mis-genotyping in the vitamin D receptor gene’, *Journal of bone and mineral research: the official journal of the American Society for Bone and Mineral Research*, 18(10), pp. 1754–1757.

4

Parallel evolution and adaptive introgression in mosquito vectors – a story of insecticide resistance

4.1 Abstract

The primary control methods for the African malaria mosquito, *Anopheles gambiae*, are based on insecticidal interventions. Emerging resistance to these compounds is therefore of major concern to malaria control programmes. The recently introduced organophosphate, pimiriphos-methyl, is widely used in indoor residual spray programmes and is essential to combat widespread resistance present in malaria vectors. Here, we use a population genomic approach to examine novel mechanisms of resistance to pirimiphos-methyl in *Anopheles gambiae s.l* mosquitoes. In multiple cohorts, we find large and repeated signals of selection at a locus containing a cluster of detoxification enzymes, some of whose orthologs are known to confer resistance to organophosphates in *Culex pipiens*. Close examination reveals a complex and diverse pattern of haplotypes under selection in An. gambiae, *An. coluzzii* and *An. arabiensis*. As in *Cx. pipiens*, copy number variation seems to play a role in the evolution of insecticide resistance at this locus. We use haplotype and phylogenetic approaches to examine whether these signals arise from parallel evolution or adaptive introgression. Finally, using whole-genome sequenced phenotyped samples, we find that multiple haplotypes under selection are associated with resistance to pirimiphos-methyl. Overall, we demonstrate a striking example of contemporary parallel evolution which has important implications for malaria control programmes.

4.2 Introduction

The spread of organophosphate resistance in the common house mosquito, *Culex pipiens*, is a textbook example of contemporary evolution in response to anthropogenic pressures (Raymond et al., 1998). In this species, mutations around two alpha-esterases enhanced the ability of the mosquito to detoxify and eliminate organophosphate (OP) insecticides used in larviciding campaigns (Guillemaud et al., 1997). This locus was termed *Ester*, with independent gene duplications and transposable element insertions at the *Est2* and *Est3* resulting in at least 16 distinct haplotypes across the mosquitoes' worldwide range (Raymond et al., 1996).

Conversely, in the major malaria vector, *Anopheles gambiae*, resistance to organophosphates is associated primarily with the *Ace1* locus, the target of OP and carbamate insecticides. At this locus, a complex combination of heterozygous gene duplications and the *Ace1*-G119S non-synonymous mutation, confer varying levels of resistance and fitness costs (Edi *et al.*, 2014; Assogba *et al.*, 2018; Grau-Bové *et al.*, 2020). Organophosphates are recommended by the WHO for use in indoor residual spraying (IRS) campaigns as the formulation Actellic CS, and whilst resistance has been slow to develop, it is emerging (Grau-Bové *et al.*, 2020).

In this study, we investigate whether there is evidence for additional OP resistance mechanisms in *An. gambiae*. Using data from the *An. gambiae* 1000 genomes project (Miles *et al.*, 2017; Clarkson *et al.*, 2020), we found evidence of large, repeated signals of selection at the locus orthologous to the *Culex pipiens Ester* locus. We integrate expression data from studies across sub-Saharan Africa and perform an extensive haplotypic analysis of this region, highlighting the *Coeae1f* and *Coeae2f* locus as being important for insecticide resistance in malaria vectors. We find distinct, novel CNVs have arisen at this locus in both *An. gambiae* and *An. arabiensis* and that putative adaptive haplotypes have introgressed between *An. gambiae* and *An. coluzzii*. These data demonstrate the importance of parallel evolution and introgression in the evolution of adaptively important traits in insect disease vectors.

4.3 Results

4.3.1 A novel insecticide resistance locus

In the first phase of the Anopheles 1000 genomes project (Miles *et al.*, 2017), a large signal of selection was observed at a region, (≅28.5 Mb) on the 2L chromosomal arm. This signal was found solely in populations of *An. gambiae* from West Africa, specifically in samples from Burkina Faso, Ghana, Mali and Guinea. The signals of selection were very broad, with haplotype homozygosity extending beyond one megabase, suggesting that selection at the locus may have occurred relatively recently prior to sample collections.

A closer examination of the region reveals a cluster of 7 putative detoxification genes residing within 20 Kb of the suspected selection signal peak, including two alpha-esterases, C*oeae1f*, and C*oeae2f*. These alpha-esterases sit in reverse orientation, 495 bases apart, and despite their recent shared ancestry, contain a varying number of exons (*Coeae1f*- n=7*; Coeae2f*- n=4). At the amino acid level, sequence similarity is 50.93% between the two genes. We performed reciprocal orthology searches, revealing that these carboxylesterases are one-to-one orthologs with the *Est3 (Coeae1f)* and *Est2 (Coeae2f)* carboxylesterases of *Culex pipiens.* This provided tentative evidence that the *Coeae1f/2f* may be driving the signals of selection. *Coeae1f* shares 69.02% amino acid similarity with *Est3*, and *Coeae2f* shares 63.84% amino acid similarity with *Est2*.

Two of the genes in the cluster (AGAP006222, AGAP006223) are UDP-glucosyltransferases (UGTs), whilst the remaining three are annotated as aldehyde oxidases. UGTs are known to be involved in the phase 2 detoxification of pyrethroids (Ismail *et al.*, 2013), and aldehyde oxidases have been associated with resistance to neonicotinoids (Hemingway *et al.*, 2000; Shi *et al.*, 2009). Despite the evidence from *Culex pipiens* for the alpha-esterases' involvement in insecticide resistance, the close proximity of these other detoxification genes made it difficult to confidently assign any gene as causal of the selective sweep. In comparison, none of the downstream genes have an obvious link to insecticide resistance based on their genome annotations.

To further elucidate the drivers of the selective sweeps we examined signals of selection from whole genome sequencing data in a larger cohort of *An. gambiae s.l* from throughout sub-Saharan Africa, collected between 2012 and 2018 (see methods for further details of sample collections and sequencing). Given the existence of multiple segregating selective sweeps in these populations we applied the H123 statistic calculated in 1200bp stepping windows across the 2L chromosomal arm (Figure 1)(Garud *et al.*, 2015). We find selection signals at this locus across most of these populations (Figure 1), as well as at two insecticide target sites (Vgsc and Gaba). The peak of the majority of the selection signals was over the two alpha-esterases.

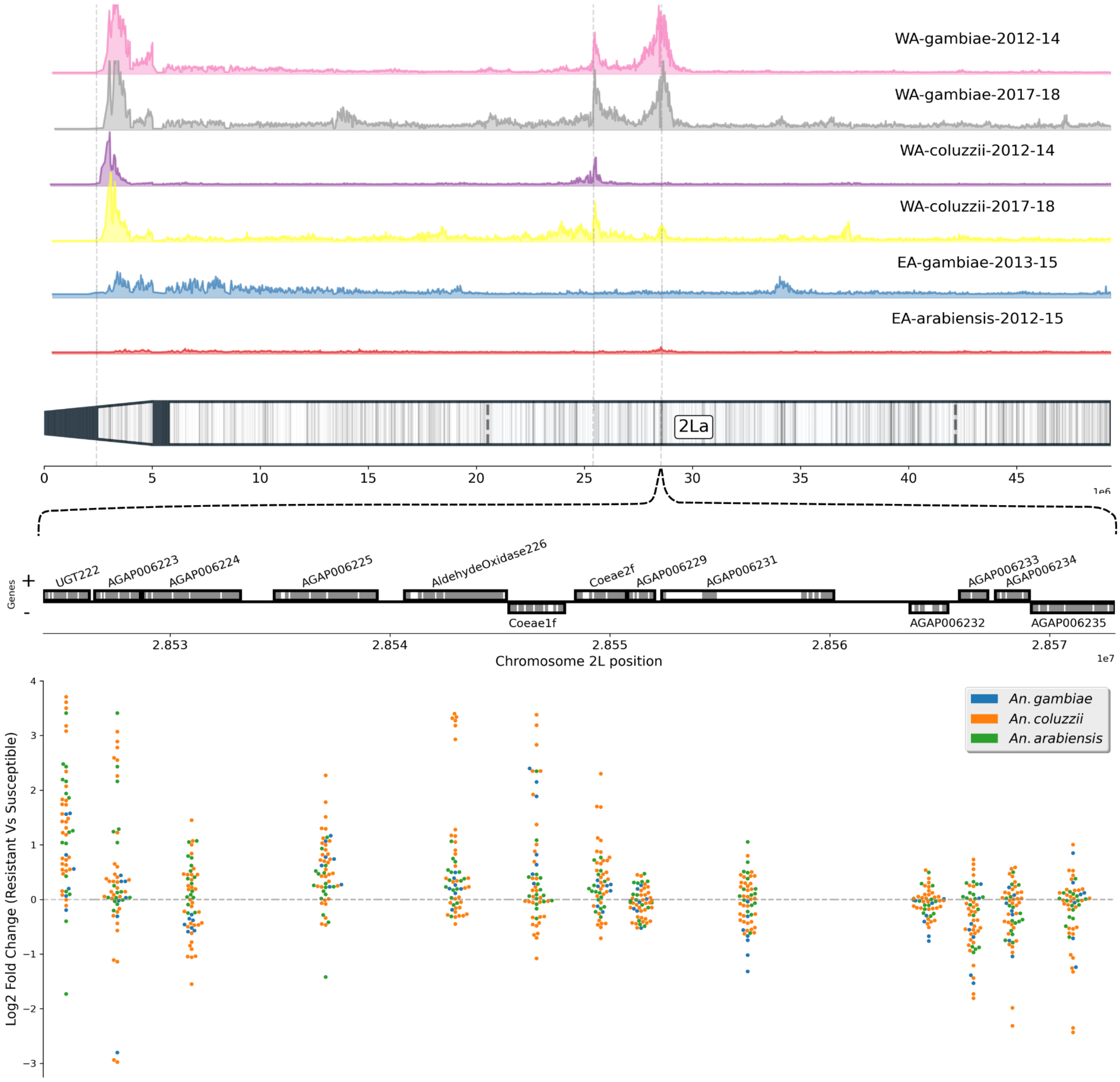


Figure 1. Evidence for selection around the *Coeae1f/2f* locus. Upper: Genome-wide H123 selection scans in cohorts from the Ag1000g show evidence of selection on the 2L chromosomal arm. WA=West Africa, EA=East Africa. A diagram of the 2L chromosomal arm is shown, with banding reflecting gene density, and black regions showing heterochromatin. The 2la inversion region is demarcated by dotted lines. **Lower:** For each gene, the Log2 Fold change is displayed from RNA-Sequencing and microarray studies performed into insecticide resistance from across sub-Saharan Africa. Each dot is a separate differential expression comparison.

Insecticide resistance is often associated with Increased expression of insecticide-detoxifying genes. To determine whether there was any evidence for resistance-associated differential expression across genes flanking 25 Kb on either side of *Coeae1f*, we integrated gene expression data from 55 microarray and RNA-Sequencing experiments (Ingham *et al.*, 2018). Figure 1 shows swarm plots of log2 fold changes coloured by species, from transcriptomic experiments in which an insecticide-resistant strain of *An. gambiae s.l* was compared to a susceptible strain (see methods).

We hypothesised that detoxification genes contributing to insecticide resistance may show evidence of over-expression, which could implicate them in a role in resistance at this locus. Firstly, there is a clear difference in the patterns of expression for the detoxification gene cluster (*UGT222* to *Coeae2f*), and the genes downstream (up to AGAP006235). The detox cluster clearly shows more occurrences of high fold changes, which may suggest an involvement in resistance in these genes.

UGT222 is a glucosyl transferase that is consistently upregulated in populations of *An. coluzzii* across Africa, and is speculated to be involved to some degree in phase 2 detoxification of insecticides (Ingham *et al.*, 2018). In the other UGT, AGAP006223, we observe both some highly positive and negative fold changes. The gene has very low expression in RNA-Sequencing data, which may contribute to high variability in differential expression. Additionally, the presence of a gene deletion (see CNV section), which is polymorphic across many species and geographical regions could be contributing to the extreme values seen here, if a resistant or susceptible strain used in comparison carried the deletion. Two out of three aldehyde oxidases show little differential expression between resistant and susceptible colonies. In *An. coluzzii*, AGAP006226 is heavily upregulated in five transcriptomics experiments and so has a high mean expression, however, its upregulation is not consistent across experiments. In *An. gambiae, Coeae1f* shows positive mean fold changes in *An. gambiae (2.72), An. coluzzii (1.71) and An. arabiensis (1.40)*. The expression data is less convincing for *Coeae2f*, however, this gene is highly expressed at a base level, and so could still contribute to the insecticide-resistant phenotype without large fold changes.

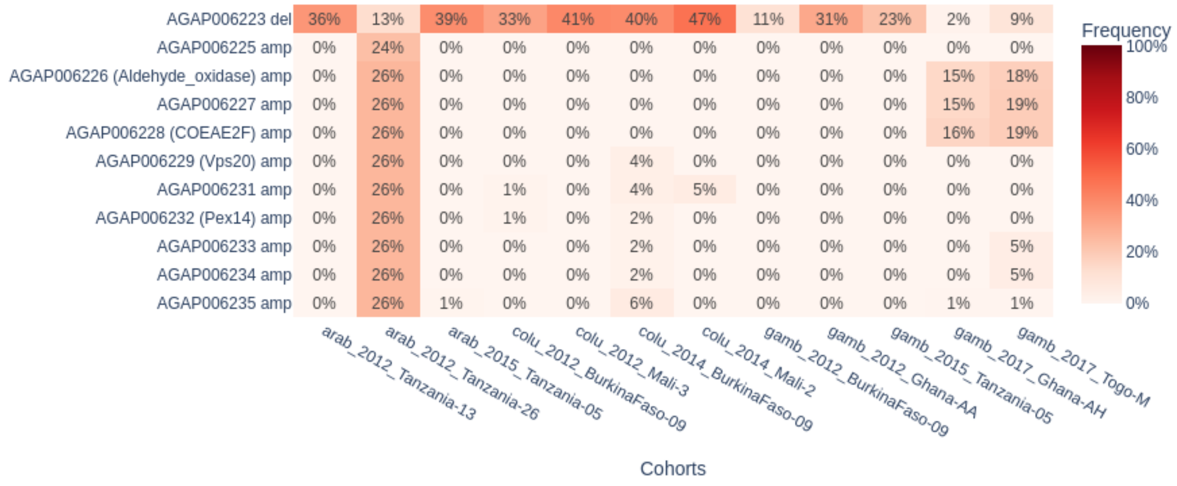
4.3.2 Copy number variation

A common mechanism of metabolic resistance is copy number variation (Lucas *et al.*, 2018). Gene amplifications in *An. gambiae* have been associated with increased gene expression and insecticide resistance (Njoroge *et al.*, 2022). In *Cx. pipiens*, haplotypes at the *ester* locus have spread around the world, and the majority of these haplotypes are associated with amplifications which cover one or both of the *Coeae1f* and *Coeae2f* orthologs (Raymond *et al*., 1991).

Given the gene expression data and the presence of copy number variation at this locus in *Cx. pipiens*, we speculated whether copy number variants might also exist in *An. gambiae s.l.* In order to identify CNVs, we calculated the modal copy number at each gene in the region and then computed the frequency of amplifications or deletions in the dataset. Table 1 shows a summary of the frequencies of CNVs at the locus. For this analysis, we split the dataset into cohorts corresponding to a specific species, year of collection, and district in each country.

In the arab_2012_Tanzania-26 cohort, a CNV which we term Dup2 can be observed, spanning 10 genes, as well as another CNV in West African populations of *An. gambiae*, which spans the two alpha-esterases *Coeae1f* and *Coeae2f*. A deletion in AGAP006223 is polymorphic across species and across sub-Saharan Africa, which agrees with the results from the gene expression meta-analysis, in which the most extreme negative values were observed.

Table 1. Copy number variant frequencies at the *Coeae1f* locus. Frequencies refer to the presence of an amplification within an individual mosquito. Genes with zero copy number variants are not shown in the table. *Coeae1f*=AGAP006227, but is not labelled as its gene name is not annotated in the *Agam* P4.12 PEST reference genome.



	arab_2012_Tanzania-13	arab_2012_Tanzania-26	arab_2015_Tanzania-05	colu_2012_BurkinaFaso-09	colu_2012_Mali-3	colu_2014_BurkinaFaso-09	colu_2014_Mali-2	gamb_2012_BurkinaFaso-09	gamb_2012_Ghana-AA	gamb_2015_Tanzania-05	gamb_2017_Ghana-AH	gamb_2017_Togo-M
AGAP006223 del	36%	13%	39%	33%	41%	40%	47%	11%	31%	23%	2%	9%
AGAP006225 amp	0%	24%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
AGAP006226 (Aldehyde_oxidase) amp	0%	26%	0%	0%	0%	0%	0%	0%	0%	0%	15%	18%
AGAP006227 amp	0%	26%	0%	0%	0%	0%	0%	0%	0%	0%	15%	19%
AGAP006228 (COEAE2F) amp	0%	26%	0%	0%	0%	0%	0%	0%	0%	0%	16%	19%
AGAP006229 (Vps20) amp	0%	26%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%
AGAP006231 amp	0%	26%	0%	1%	0%	4%	5%	0%	0%	0%	0%	0%
AGAP006232 (Pex14) amp	0%	26%	0%	1%	0%	2%	0%	0%	0%	0%	0%	0%
AGAP006233 amp	0%	26%	0%	0%	0%	2%	0%	0%	0%	0%	0%	5%
AGAP006234 amp	0%	26%	0%	0%	0%	2%	0%	0%	0%	0%	0%	5%
AGAP006235 amp	0%	26%	1%	0%	0%	6%	0%	0%	0%	0%	1%	1%

Figure 2 shows traces of coverage in 300bp windows and the CNV HMM prediction for two example individuals from Ghana and Tanzania, respectively. Both CNVs cover the two carboxylesterases *Coeae1f* and *Coeae2f*, and the CNV in Ghana and Togo only covers these two genes - also amplifying a truncated version of a putative aldehyde oxidase, AGAP0006226. Interestingly, gene amplifications at this locus in *Cx. pipiens* have also been reported to contain a partial copy of an aldehyde oxidase (Buss *et al.*, 2004). No other CNVs were noted covering detoxification genes at the locus.

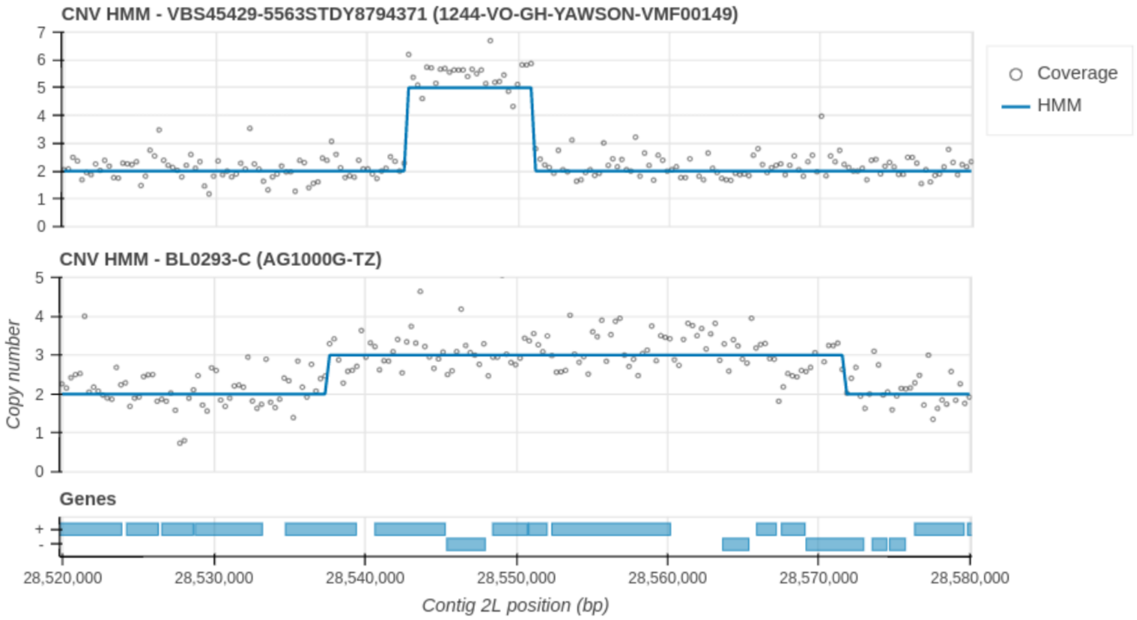


Figure 2. Example coverage traces for Dup1 (upper, *An. gambiae female* collected in Obuasi, Ghana) and Dup2 (lower, *An. arabiensis female* from Moshi, Tanzania). Dots show coverage as calculated in 300 Bp stepping windows, and the line represents the HMM prediction of copy number state described in (Lucas *et al.*, 2019).

4.3.3 Haplotype clustering of the *Coeae1f* locus

To identify haplotype clusters that have been driven to appreciable frequency, we performed hierarchical clustering on haplotypes from the *An. gambiae s.l* dataset. Figure 3 shows the results of this clustering, aligned with metadata from the individual which contributed the haplotype, such as sample country, and whether a haplotype's parent individual harbours at least one CNV at this locus. Based on the evidence from orthology and copy number variation, we also plot non-synonymous mutations in the *Coeae1f* and *Coeae2f* genes. We extracted 4862 haplotypes from 2431 individual mosquitoes, and calculated the number of SNP differences in this ~5kb region, using a cut-off of 1 SNP to separate clusters. We then used a minimum threshold of 70 haplotypes in order to designate a cluster as swept, any smaller clusters or non-clustered haplotypes were designated as wild-type. We manually selected these values to ensure haplotype clusters which exhibited little within-cluster variation and to avoid spurious small clusters which may be linked to demographic events. To confirm that the observed haplotype clusters are due to selection as opposed to population demography (which would impact the entire chromosome) we examined regions 2 Mb up and downstream and did not observe large clusters of haplotypes as seen in Figure 3, indicating that the haplotype clusters at this locus are indeed due to selection.

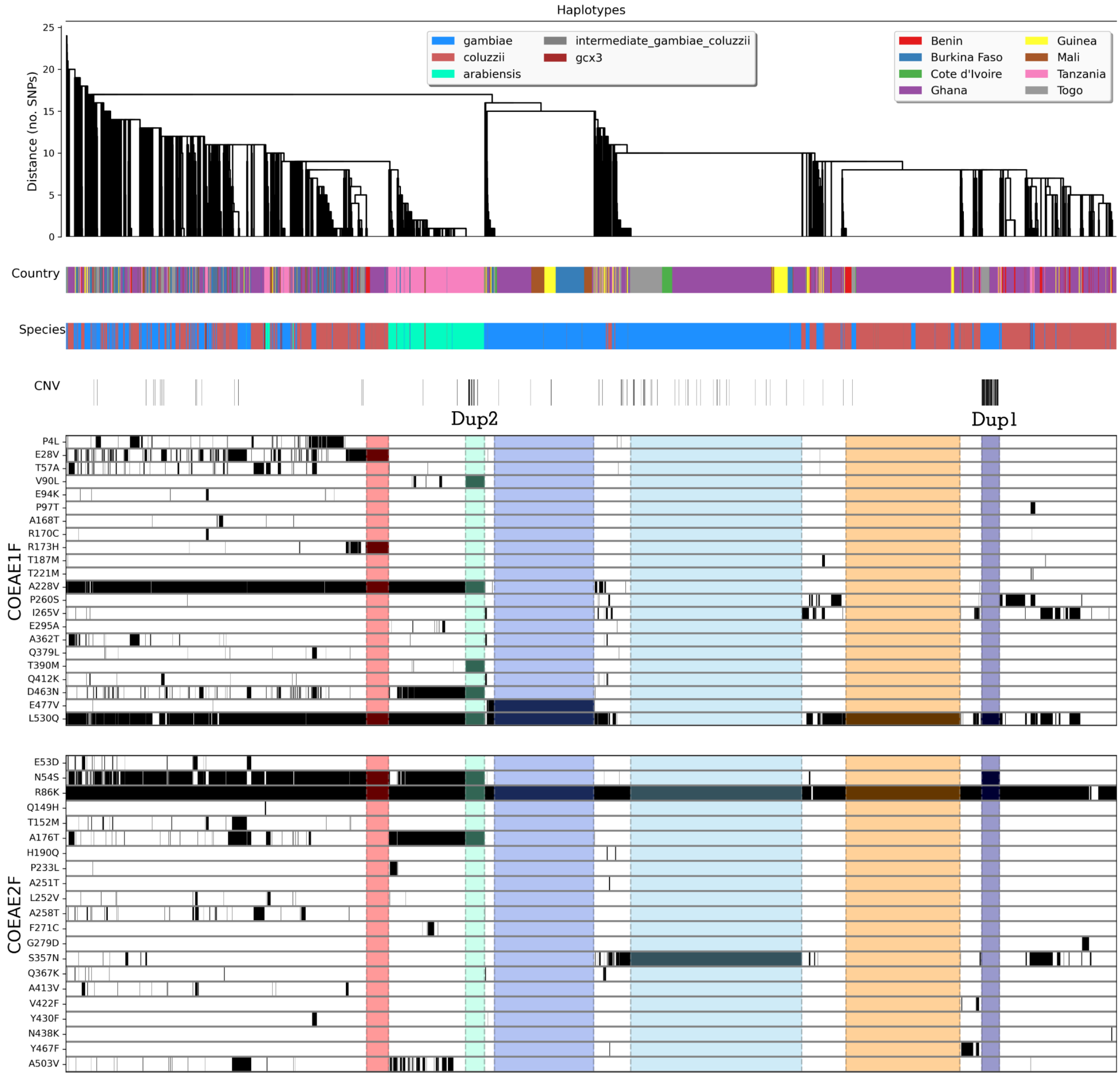
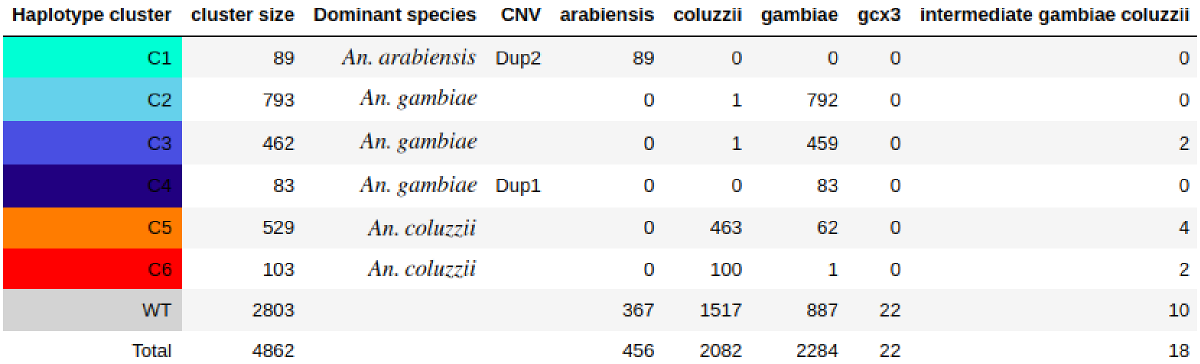


Figure 3. Haplotype clustering over the Coeae1f and Coeae2f region and amino acid variation. Each column is a haplotype ordered by the dendrogram with hierarchical clustering, with a distance metric of hamming distance (converted to distance in SNPs) and single linkage. The coloured rectangles represent haplotype clusters that are <=1 allele different over this region and are indicative of a selective sweep. CNVs; A CNV is strongly associated with a sweep found in *An. gambiae* (C4, navy), and another one less tightly associated with a sweep in Tanzanian *An.*

arabiensis (C1, teal).

Cutting the dendrogram at 1 SNP over the 5kb window, we detect six distinct selective sweeps in the dataset. Table 2 describes the haplotype clusters in the dataset. The colour for each haplotype cluster was chosen according to the predominant species - greens for *An. arabiensis*, blues for *An. gambiae* and reds for *An. coluzzii*, and are used throughout the study.

Table 2. A description of each haplotype cluster detected, ordered by predominant species and cluster size.



Haplotype cluster	cluster size	Dominant species	CNV	arabiensis	coluzzii	gambiae	gcx3	intermediate gambiae coluzzii
C1	89	*An. arabiensis*	Dup2	89	0	0	0	0
C2	793	*An. gambiae*		0	1	792	0	0
C3	462	*An. gambiae*		0	1	459	0	2
C4	83	*An. gambiae*	Dup1	0	0	83	0	0
C5	529	*An. coluzzii*		0	463	62	0	4
C6	103	*An. coluzzii*		0	100	1	0	2
WT	2803			367	1517	887	22	10
Total	4862			456	2082	2284	22	18

Figure 4 shows two maps of the sample collection sites in our dataset, and the frequency of each haplotype cluster. The C1 sweep is found only in *An. arabiensis* in east Africa, whereas the other five clusters are found primarily in West Africa. In the first phase of Ag1000G collections ranging from 2012-2014, the C3 sweep was the predominant haplotype cluster. Other haplotype clusters are only at appreciable frequency in later years of sampling, however, the lack of real longitudinal data from single sites makes it difficult to infer that haplotypes have increased in frequency.

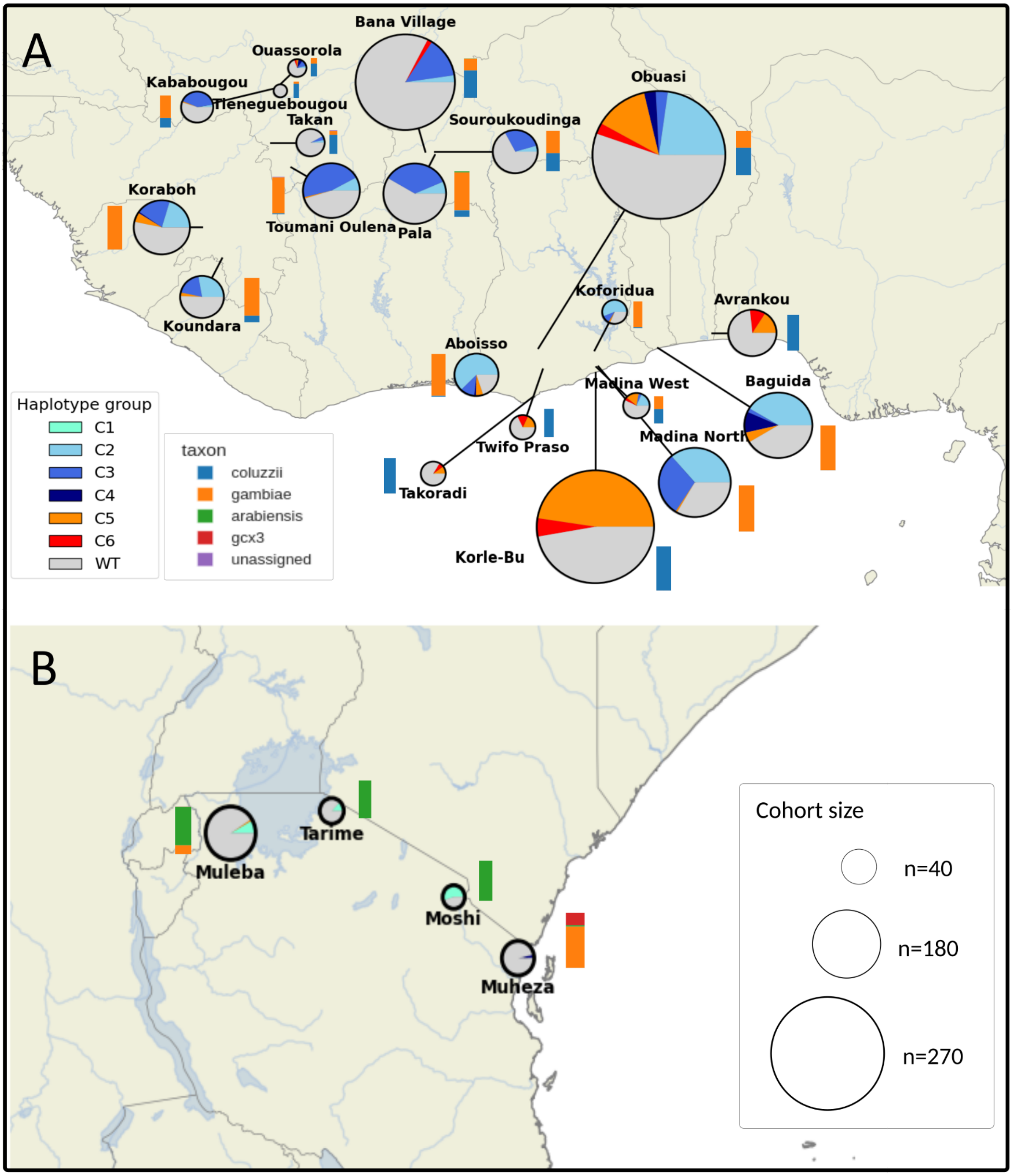


Figure 4. Maps of sample locations and cluster frequencies. A: West Africa **B:** Tanzania. The size of the pie is proportional to the sample size (bottom right legend), and the proportion of the pie corresponds to the haplotype frequency. The species composition of each cohort is shown with stacked bar plots adjacent to pie charts.

The haplotype clustering provides evidence of adaptive gene flow. According to haplotype clustering, the C5 cluster shows strong evidence of adaptive gene flow between species. Although 463 haplotypes out of 529 in this cluster are from individuals which are assigned as *An. coluzzii*, 62 are assigned as *An. gambiae*, with 4 putative hybrids. Given it consists primarily of *An. coluzzii* individuals, the directionality of introgression is likely to be *An. coluzzii* -> *An. gambiae*, though further analyses are required to confirm this finding. Interestingly, the C2 and C3 clusters also contain one *An. coluzzii* haplotype and the C6 cluster contains one *An. gambiae* haplotype. The apparent frequency of haplotype sharing between species observed here is striking and may be commonplace in regions of the genome under intense selection like the *Coeae1f/2f* locus. Interestingly, we also observe more hybrid individuals in the clusters which have evidence of introgression than without, although given the low numbers of hybrids our resolution may be low.

The C6, C1 and C3 sweeps are only found on the background of the 2La inversion polymorphism. Whilst this inversion is fixed in *An. arabiensis*, it is polymorphic in *An. gambiae* and *An. coluzzii* which may restrict the spread of the sweep, as recombination is suppressed between 2La heterozygotes, although limited recombination can occur as previously evidenced (Grau-bové *et al.*, 2019). Moreover, the *Coeae1f/2f* locus is towards the centre of the inversion (8.02Mb from the beginning of the breakpoint, and 13.62Mb from the end), which makes inter-karyotypic introgression more likely than if close to the endpoints (Cheng *et al.*, 2012). The hierarchical clustering resolves to two clades corresponding to 2L+a and 2La haplotypes. This would be expected given that the inversion predates the split between *An. gambiae* and *An. coluzzii* (Ayala *et al.*, 2011), and therefore population structure in the region of the inversion is driven by inversion karyotype, rather than species (Miles *et al.*, 2017).

To confirm the above patterns of introgression, we perform a phylogenetic analysis of haplotypes, described in Appendix A. By accounting for molecular evolution, phylogenetic analysis should be more robust than hierarchical clustering of haplotypes. We confirm the signals of inter-species introgression, and also find some evidence for the exchange of haplotypes between the 2L+a and 2La karyotypes in the C3 cluster.

With haplotype clusters resolved, we examined the association of CNVs with these clusters. From the clustering analyses, it looks as though the *An. arabiensis* C1 sweep is only weakly associated with the presence of a CNV, in comparison with the *An. gambiae* C4 sweep, in which most haplotypes seem to be associated with a CNV. To examine this in more detail, Figure 5 shows the copy number of the *Coeae1f* and *Coeae2f* genes, plotted against the number of haplotypes an individual has in the C1 and C4 clusters, respectively. In agreement with the results from haplotype clustering, the C4 cluster and Dup1 show a strong correlation, indicating the sweep and duplication are closely linked. In contrast, the correlation of the C1 cluster and Dup2 is weak, and so many haplotypes which fall into the C1 cluster are not associated with a CNV.

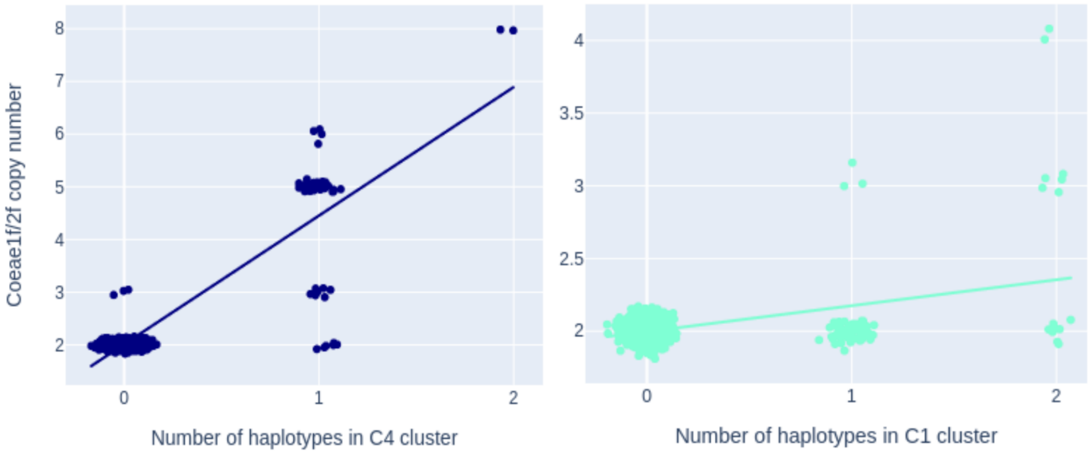


Figure 5. The C1 and C4 haplotypes are associated with CNV alleles. Plots show for each individual the number of haplotypes in each haplotype cluster and the Coeae1f/2f copy number. **Left:** Dup1 and cluster C4. C4 is closely associated with a CNV, termed Dup1, which exists at a variable copy number. **Right:** Dup2 and cluster C1. C1 is weakly associated with a CNV, Dup2, which only exists as one extra copy. Points are jitted to avoid overlapping data.

We re-examined the coverage traces for the example *An. gambiae* female from Obuasi, Ghana, which is positive for Dup1 (Figure 2). The modal copy number in this individual is five, and the individual is heterozygous for the C4 sweep, suggesting that Dup1 can exist as a four-copy amplification. In agreement with this observation, the highest copy number we observe in this dataset is eight, in two individuals homozygous for the C4 sweep. The lower plot of Figure 2 shows an *An. arabiensis* female from Moshi, Tanzania, with the larger Dup2

CNV outlined. In figure 5, we only observe Dup2-positive individuals with either one or two extra copies, suggesting that it only exists as a duplication.

As the Dup2 copy number variant segregates in the C1 haplotype cluster, we also examined patterns of extended haplotype homozygosity within the cluster, grouping haplotypes by whether or not the parent individual of each haplotype was positive for at least one Dup2 copy (Appendix B1). If we were to phase Dup2, this would provide greater power, however, Dup2 is present at variable copy numbers which makes phasing challenging. We found that C1 haplotypes which were not associated with Dup2, had faster rates of haplotype homozygosity decay, reflecting the fact that Dup2 must have arisen on the C1 haplotypic background, and therefore less time has passed to break down the haplotypes.

4.3.4 Amino acid variation

The six identified selective sweeps could be driven by amino-acid alterations that affect detoxification efficiency, or CNVs and other cis-regulatory mutations that affect the expression of proximal genes. As a first step to identifying potential causal mutations in each selective sweep, we calculated allele frequencies at biallelic non-synonymous sites in the dataset overall and in each haplotype cluster. Figure 2 visualises non-synonymous variation in *Coeae1f/2f* across all haplotypes in the dataset, with the six haplotype clusters highlighted.

In *Coeae2f,* R86K is almost fixed across all samples, and *Coeae1f*-A228V and *Coeae1f*-L530Q are almost fixed in 2La haplotypes, therefore we will not discuss them below. The C1 cluster - which consists solely of *An. arabiensis* haplotypes - contains few distinguishing amino acid mutations, but does carry the E28V and R173H mutations in *Coeae1f*. In *Coeae2f*, C1 carries A176T. The largest sweep in terms of cohort size, C2, has no biallelic amino acid mutations in *Coeae1f* but does carry the *Coeae2f*-S357N allele, which is rarely found outside of this cluster. The C3 sweep contains a mutation *Coeae1f*-E477V, which is again rare on haplotypes outside of this cluster, and the mutation serves as a tagging SNP, found by the haplotype tagging SNP algorithms (section 4.3.9).

The C4 and C5 clusters carry no distinguishing biallelic mutations in *Coeae1f* or in *Coeae2f*, whereas C6 carry *Coeae1f*-E28V and *Coeae1f*-R173H in *Coeae1f*.

From these mutations, the E477V mutation in the C3 cluster and the S357N mutation in the large C2 cluster are candidates for a possible causal role in selection. This is due to the fact that both variants are otherwise rare, and visualisation of an *in silico* protein model of *Coeae1f* (PEST) indicates that they are in regions of the enzyme which are in close proximity to the active site.

We also produced haplotype clustering plots for each detoxification gene in the *Coeae1f/2f* region (Appendix C). Though we do not describe polymorphisms in these genes in detail, it remains possible that mutations in or near these genes could instead be driving the selective sweeps, rather than *Coeae1f* or *Coeae2f*. This is probably more likely in selective sweeps involving *An. coluzzii*, as the UGT222 gene in particular often shows substantial over-expression in this species (Figure 1).

4.3.5 The haplotype clusters are under positive selection

To investigate the degree of positive selection acting on the detected haplotype clusters, we analysed extended haplotype homozygosity (EHH) (Sabeti *et al.*, 2002). Haplotypes which contain a beneficial allele and spread through a population will be shared across many individuals. Due to the speed at which the haplotypes spread to an appreciable frequency, there will not be time for recombination to break the haplotypes down. This means that around the focal region, within a haplotype cluster, haplotype homozygosity will be elevated, and this elevation will decay the further we move away from the focal region. We defined a focal point in the intergenic region between *Coeae1f* and *Coeae2f* and extended the region 50kb in either direction (100kb window). We selected a large window size as the H12 signals at this locus are broad. We used the six haplotype clusters we identified in the haplotype clustering analysis.

Throughout the study we define wild-type as any haplotype that did not fall into one of the six haplotype clusters, however, for EHH analysis we further defined three wild-type cohorts within this group. Using a wild-type cohort which contains a mixture of species and geographical regions would bias the EHH analysis, as we would expect haplotype homozygosity decay to be extreme in a highly diverse group of haplotypes. We, therefore, selected three control populations consisting of wild-type haplotypes that belonged to either Tanzanian *An. arabiensis* from 2015, Burkinabe *An. coluzzii* from 2014, or Guinean *An. gambiae* from 2012.

Compared to the three wild-type populations, all six haplotype clusters displayed extremely elevated levels of haplotype homozygosity, which combined with their appreciable frequencies in the dataset, suggests positive selection acting on these haplotypes (Figure 6). The difference between wild-type groups and the six haplotype clusters is stark. However, we have not yet formally tested for significance by calculating the length of shared haplotypes and deriving confidence intervals.

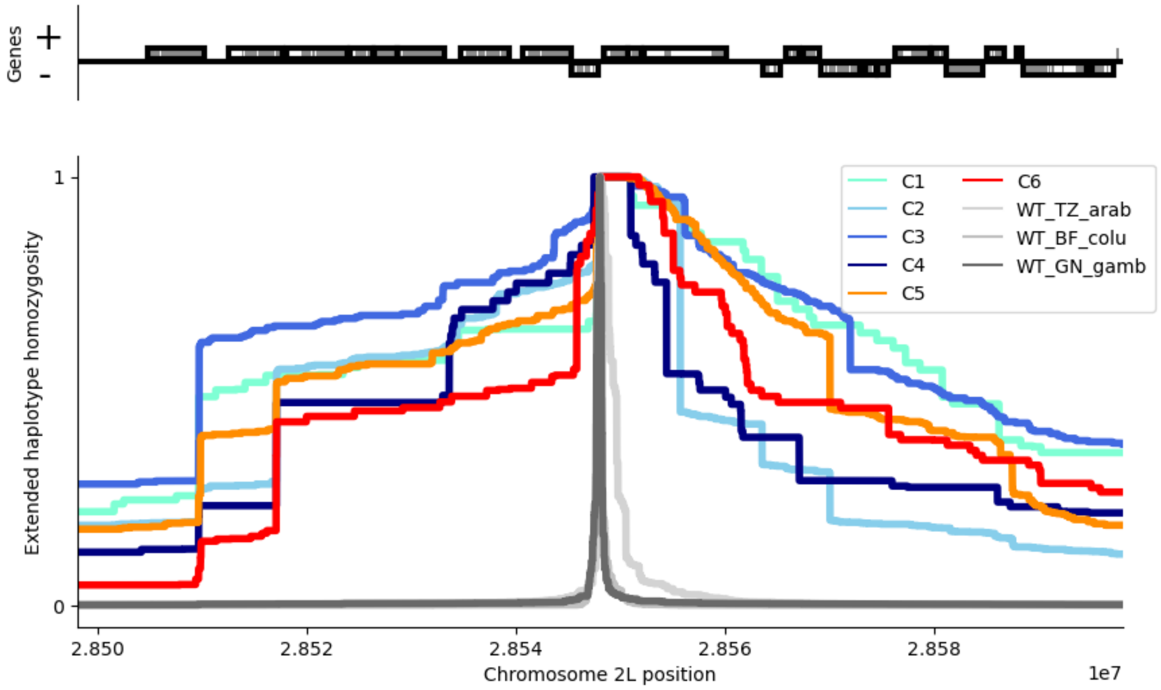


Figure 6. Positive selection of haplotype clusters. Extended haplotype homozygosity, moving 50kb on either side of the focal intergenic region between *Coeae1f* and *Coeae2f*. Three wild-type cohorts were selected from Tanzania (*An. arabiensis)*, Burkina Faso (*An. coluzzii)*, and Guinea (*An. gambiae)*.

4.3.6 The *Coeae1f/Coeae2f* locus is associated with resistance to Pirimiphos-methyl

Given evidence that the six haplotype clusters are under positive selection, we hypothesised that they may confer resistance to insecticides used in vector control. Pirimiphos-methyl is the active ingredient in Actellic-300CS, an organophosphate used widely in indoor residual spraying (IRS) programmes. Resistance to the compound has been slow to arise, and the *Ace1* locus remains the only validated resistance marker (Grau-Bové *et al.*, 2020). The pyrethroid Deltamethrin is the most widely used insecticide in long-lasting insecticide-treated nets (LLINs), and due to the ubiquitous use of pyrethroids, resistance has spread through sub-Saharan Africa.

To determine which insecticides the sweep may confer resistance to, we took the subset of samples from our cohorts which were phenotyped by WHO tube assays to either Pirimiphos-methyl or Deltamethrin, prior to sample collection. This amounted to 973 individuals, all either *An. gambiae* or *An. coluzzii* collected from 2017-18 in West Africa. Figure 7 shows the odds ratios, confidence intervals and sample sizes from the binomial GLM. We found significant associations between the C2 (OR=1.30, 1.02-1.64), C3 (OR=2, 1.23-3.26), and C5 (OR=1.53, 1.08-2.14) haplotypes with Pirimiphos-methyl survivorship, and between the C3 haplotype and Deltamethrin survivorship (OR=1.58, 1.04-2.4).

The C1 cluster of *An. arabiensis*, was not represented in this subset of phenotyped data, and C4 and C6 had low sample sizes. Despite not reaching significance, C4 (in which some haplotypes carry Dup1) did have positive odds ratios for both insecticides (Delta OR=1.46, 0.7-2.9, PM OR=1.67, 0.88-3.16).

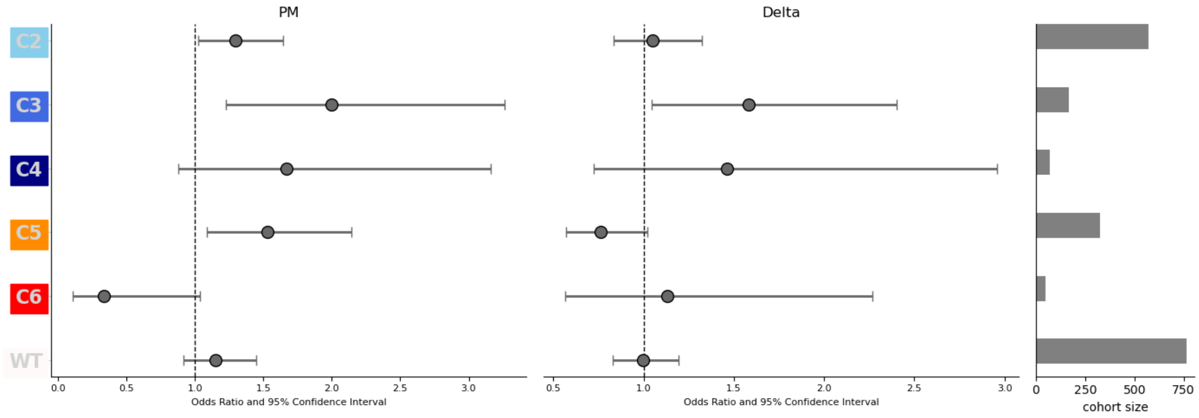


Figure 7. Haplotype association tests using a binomial GLM. Odds rations and confidence intervals are plotted for each haplotype cluster for Deltamethrin and Pirimiphos-methyl, along with the sample size for each cluster. The C2, C3 and C5 haplotypes are all associated with resistance to pirimiphos-methyl. C3 is also associated with resistance to Deltamethrin.

4.3.7 *Coeae1f/2f* are regulated by the transcription factor Maf-S

Although there are a large number of non-synonymous mutations found in the haplotype clusters, it is unclear whether these are driving the sweeps, or whether there are as yet unidentified factors, such as insertions or deletions at the locus that may affect gene expression or metabolism. COEAE1F and COEAE2F are controlled by the MAF-S transcription factor, demonstrated by downregulations of each gene during a MAF-S knockout (Ingham *et al.*, 2017).

We extracted motifs from the JASPAR database and used BioPython to search for CnC binding sites in the region under selection. We detected CnC binding sites in the *Coeae1f*, *Coeae2f* and AGAP006225 genes. Further work to examine SNP variation at these sites is ongoing.

Interestingly, an independent selective sweep is occurring on chromosome 2R, at 40.95Mb, in the same populations, but also in a couple of West African *An. coluzzii* populations. The most likely target of this sweep is KEAP1 (Harding et al., unpublished) a gene which forms a complex with MAF-S, and is therefore instrumental in the pathway.

4.3.8 Haplotype tagging SNPs

In order to track the selected haplotype clusters in time and space, we identified haplotype tagging SNPs, using machine learning approaches. We applied a decision tree (CART), random forest and lasso regression (Appendix D), implemented in scikit-learn. Results were highly similar for all models and so we report only the CART results. In order to reduce the number of haplotype tags, we ran an initial iteration of each model, retaining only informative markers. With this reduced marker set, we then performed recursive feature elimination with cross-validation (RFECV), repeatedly training the models, whilst removing one marker from the model in each iteration until only one haplotype tag remained, measuring recall and precision at each iteration (Appendix D1). Accuracy for every model is greater than 0.986 using a maximum of two haplotype tagging SNPs. We then use the python package AgamPrimer to design sets of primers to target each haplotype tag. AgamPrimer allows users to design primers and probes in *An. gambiae*, whilst considering genetic variation in the Ag1000g in primer binding sites (Nagi *et al.*, 2023). We have also developed a Locked Nucleic acid qPCR probe assay to detect the E477V mutation, which is a tagging SNP for the C3 cluster (Appendix E).

4.3.9 The locusPocus workflow

We designed a snakemake workflow to reproduce and implement a subset of the above analyses, allowing users to perform targeted analysis of specific regions of the *An. gambiae s.l* genome, including haplotype clustering, multi-allele phasing and indel calling. The workflow has been designed according to Snakemake best practices and uses the package manager Conda to automatically install all required software. It is located here: https://github.com/sanjaynagi/locusPocus.

4.4 Discussion

In the common house mosquito, *Culex pipiens*, the *ester* locus, orthologous to the *Coeae1f/2f* locus, is a textbook example of contemporary anthropogenic selection. Through exposure to toxic organophosphate insecticides, at least 16 haplotypes spread throughout the range of the mosquito, competing with each other and providing varying combinations of insecticide resistance and fitness cost. Many of these haplotypes are associated with gene amplifications around the two alpha-esterases.

In this study, we used haplotype data from 2431 individual mosquitoes to demonstrate parallel evolution at the orthologous locus in malaria mosquitoes. Intense selection is acting upon the locus in *An. gambiae, An. coluzzii, and An. arabiensis*, with multiple haplotypes rising to high frequencies, some of which contain CNVs. We demonstrate that some of these haplotypes are also associated with resistance to organophosphates. As with Dup1 and Dup2 in *An. gambiae s.l*, there is one gene amplification in *Culex pipiens* which covers both alpha esterases and a partial copy of the neighbouring aldehyde oxidase gene, and another which covers the alpha-esterases and the aldehyde oxidase.

As well as evidence of parallel evolution, the study also provides important information for malaria control programmes. We reveal a novel locus in *An. gambiae* which contributes to resistance to the organophosphate pirimiphos-methyl. Pirimiphos methyl, formulated as Actellic CS 500, is widely used in indoor residual spraying (IRS) campaigns throughout sub-Saharan Africa. Prior to this study, only one locus in *An. gambiae s.l* had been associated with resistance to this compound - *Ace1*, which is also the target site of organophosphate and carbamate insecticides. Unlike LLINs which still provide a physical barrier and so protect even against insecticide-resistant vectors, IRS is arguably more prone to resistance-mediated failure. An example of this is in Kwazulu-Natal, where pyrethroid-resistant *An. funestus* caused a malaria outbreak after a change in IRS product from DDT to pyrethroids (Maharaj *et al.*, 2005) or more recently, from recent IRS campaigns in Uganda (Epstein *et al.*, 2022). It is therefore even more important to detect emerging mechanisms of resistance early, so that insecticide resistance management (IRM) practices can be employed, such as rotating out the insecticide. With new IRS formulations such as clothianidin, broflanilide and chlorfenapyr on the market, this is more feasible now than ever.

Despite the analyses conducted in this study, it is still not clear which causal genes and mutations selection is acting upon. Given the evidence from *Culex pipiens*, and the fact that three of the haplotype clusters here are also protective against an organophosphate, we may expect that the alpha-esterases, *Coeae1f* and *Coeae2f*, are driving the selective sweeps. The presence of CNVs that span the two genes also adds weight to this hypothesis, as well as favourable expression evidence, particularly for *Coeae1f*. However, the other genes at the locus may also be important. In particular, AGAP006222, a UGT, is one of the most frequently overexpressed when comparing resistant *An. gambiae s.l* strains to susceptible strains, particularly in *An. coluzzii* (Ingham *et al.*, 2018). The AGAP006223 and AGAP006224 genes are generally very lowly expressed in RNA-Sequencing data, with AGAP006223 often deleted entirely, so it is unlikely that these genes are important for resistance. AGAP006226 does show some favourable expression evidence in *An. coluzzii*. The Dup2 duplication includes AGAP006226, whereas Dup1 causes a partial copy of this gene, which we assume to be non-functional.

Another limitation of the study is that we do not identify the causal mutation in any of the haplotype clusters. Finding and validating putative causal mutations is generally arduous, and this becomes prohibitive when as in our case, we are analysing multiple distinct selective sweeps. Although we highlight amino acid mutations and copy number variation, a limitation of the study is that we do not identify small indels or larger insertions which could well be driving the sweeps. In previous work, insertions containing transposable elements have been found at the *ester* locus in *Culex pipiens*, the CYP6aa locus in *An. gambiae*, and the *Cyp6p9a/b* locus in *An. funestus (Buss et al., 2004; Weedall et al., 2020; Njoroge et al., 2022)*. For the purpose of surveillance, however, the information on causal genes and mutations could be considered superfluous - the priority is that we know which insecticides each haplotype cluster is protective against, and that we can identify the haplotype clusters.

It will be important to track these haplotypes in future studies to determine and validate their function. The CART, random forest and lasso algorithms used to find haplotype tagging SNPs, were highly effective at distinguishing between haplotype clusters using only one or two haplotype tagging SNPs (Appendix D1). The combination of haplotype tag SNP selection and AgamPrimer may prove useful when designing amplicon sequencing panels to track the spread of resistance haplotypes.

Most of the methods presented here are packaged into a snakemake workflow, which allows users to reproduce our analysis, and apply the methods to other regions of the *An. gambiae s.l* genome. In addition, for the python-based analyses, rather than python scripts, we use Jupyter notebooks which are parameterised by Papermill. Papermill is a tool which allows users to run a whole Jupyter notebook but pass parameters through from the command line (for example, we may want to pass the genomic region through as a parameter). This is not possible with standard Jupyter notebooks. The benefit of this is that users can then extract specific notebooks from the snakemake workflow, and run these interactively on their own data, rather than having to run the snakemake workflow itself.

4.5 Methods

4.5.1 Data collection

We used a subset of whole genome sequence variation data from phase 3 of the *Anopheles* 1000 genomes project and a recent GWAS in West Africa (Clarkson *et al.*, 2020; Lucas *et al.*, 2023). The data contained 2431 individual *Anopheles* mosquitoes, of which there were 1142 *An. gambiae*, 1041 *An. coluzzii*, 228 *An. arabiensis*, 11 of a cryptic taxon termed gcx3, and 9 hybrid *An. gambiae / An. coluzzii* individuals. Specimens were all collected between 2013 and 2018. Sample provenance and methods for genome sequencing and analysis are described elsewhere (Clarkson *et al.*, 2020; Lucas *et al.*, 2023).

4.5.2 Genome-wide selection scans

We first performed genome-wide selection scans with the H123 statistic (Garud *et al.*, 2015). This statistic captures the haplotype frequency spectrum of the three highest frequency haplotype clusters and is particularly powerful to detect soft selective sweeps, or where there are multiple distinct haplotype clusters at the same locus.

We first extracted phased biallelic haplotypes from the 2431 individual mosquitoes and split the populations into cohorts of West African (WA) *An. coluzzii*, WA *An. gambiae*, East African (EA) *An. gambiae and EA An. arabiensis*. We further split the WA cohorts into early (2012-2014) and later collections (2017-2018). We then took a random sample of 200 haplotypes from each cohort and used scikit-allel v1.3.5 to compute H123 in 1200 SNP sliding windows with a step size of 600 SNPs.

4.5.3 Gene expression data

We integrated gene expression data from multiple sources to help identify genes putatively involved in insecticide resistance. First, we used data from a meta-analysis of microarray studies, which compare resistant to susceptible populations of *An. gambiae s.l (Ingham et al., 2018)*. We then integrated a further 24 differential expression comparisons of RNA-Sequencing data (Thesis Appendix A), in which raw read counts were normalised and differential expression analysis performed using DESeq2 (Love *et al.*, 2014).

4.5.4 Haplotype clustering

We extracted phased biallelic haplotypes from the start of the *Coeae1f* gene (2L:28,545,396) to the end of *Coeae2f* (2L:28,550,748), and clustered the haplotypes with single-linkage hierarchical clustering, as implemented in Scipy. We use hamming distance as the distance metric between haplotypes, which we multiply by the total number of SNPs in the genomic window to convert to the number of SNP differences. We then determined which non-synonymous variants were present on each haplotype and plotted this data alongside the CNV status, population and taxon of the individual to whom the haplotype belongs.

4.5.5 Phylogenetic analysis of haplotypes

We extracted genomic alignments of SNPs in the phased biallelic haplotype data, from four regions around the *Coeae1f/2f* locus, with the following coordinates: 1) The focal region (2L:28,545,767 ± 10,000 kb, 6819 variants) 2) the downstream region (2L:30,545,767 ± 10,000 kb, 5131 variants) 3) the upstream region (2L:26,545,767 ± 10,000 kb, 4359 variants). To ease the interpretation of the resulting phylogenetic trees, we restricted the analysis to haplotypes from individuals that were homozygous for either the 2La inversion or the standard (2L+a) arrangement, as calculated from karyotype tagging SNPs (Love *et* *al*., 2019). In addition to the aforementioned populations, we included sequence data from *An. melas, An. merus,* and *An. quaddrianulatus* as outgroups.

We then used IQ-TREE v2.2.0 to compute maximum-likelihood phylogenies for each genomic alignment (Nguyen *et al*., 2015). The best-fitting nucleotide substitution model for each alignment was chosen according to the ModelFinder algorithm (Kalyaanamoorthy *et al*., 2017). We calculated branch statistical supports using the UF bootstrap procedure (Minh *et al*., 2013; Hoang *et al*., 2018) and refined the tree for up to 10,000 iterations until convergence was achieved. Trees were visualised in R using the Ape 5.6 library (Paradis *et al*., 2018), and each phylogeny was midpoint-rooted with phytools 1.0.3 (Revell, 2011).

4.5.6 Positive selection in haplotype clusters

For each cluster of selected haplotypes, and the remaining haplotypes which we designate as wild-type, we analysed signals of positive selection by calculating extended haplotype homozygosity (EHH) moving away from the focal locus (2L:28548072) for 50 kb on either flank (using the *ehh_decay* function in scikit-allel). We used three wild-type cohorts as control groups - these were haplotypes that were not part of the C1-C6 haplotype clusters and consisted of *An. gambiae* haplotypes from Guinea collected in 2012, *An. coluzzii* haplotypes from Burkina Faso in 2014, and *An. arabiensis* from Tanzania, collected in 2015.

4.5.7 Haplotype association tests

Using phased haplotypes from 973 phenotyped individuals from West Africa (Lucas et al. 2023) we performed haplotype association tests. These individuals were phenotypes against either Pirimiphos-methyl or Deltamethrin, using WHO tube bioassays (World Health Organization, 2016). These insecticides are widely used in malaria vector control. We first assigned haplotype clusters to individual haplotypes, and performed haplotype association tests with a single binomial GLM, with each cluster as a predictor variable and phenotype (dead or alive) as the response variable.

4.5.8 Haplotype tagging SNPs

We used the scikit-learn v1.0.2 implementations of a decision tree classifier, a random forest classifier and a penalised logistic regression (L1 penalty; lasso) classifier, to predict haplotype cluster membership from integrated phased biallelic and multiallelic SNP markers in the *Coeae1f/2f* region. We ran the first iteration of each model, in order to ascertain informative markers. We then subset the input data to this informative marker set and re-ran the models, performing recursive feature elimination with cross-validation (RFECV), wherein we removed one SNP at a time from the model whilst measuring recall and precision, until only 1 SNP remains. This allows us to determine an optimal number of feasible haplotype tagging SNPs, to use in future research.

4.5.9 Code availability

LocusPocus is available here https://github.com/sanjaynagi/locusPocus. All other code for this chapter are stored here https://github.com/sanjaynagi/PhD_thesis_code/coeae1f.

4.6 References

Assogba, B.S., Alout, H., Koffi, A., Penetier, C., Makoundou, P., Weill, M. and Labbé, P. (2018) 'Adaptive deletion in resistance gene duplications in the malaria vector *Anopheles gambiae*', (February), pp. 1245–1256.

Ayala, D., Fontaine, M.C., Cohuet, A., Fontenille, D., Vitalis, R. and Simard, F. (2011) 'Chromosomal inversions, natural selection and adaptation in the malaria vector *Anopheles funestus*', *Molecular biology and evolution*, 28(1), pp. 745–758.

Buss, D.S. and Callaghan, A. (2004) 'Molecular comparisons of the Culex pipiens (L.) complex esterase gene amplicons', *Insect biochemistry and molecular biology*, 34(5), pp. 433–441.

Cheng, C., White, B.J., Kamdem, C., Mockaitis, K., Costantini, C., Hahn, M.W. and Besansky, N.J. (2012) 'Ecological genomics of anopheles gambiae along a latitudinal cline: A population-resequencing approach', *Genetics*, 190(4), pp. 1417–1432.

Clarkson, C.S., Miles, A., Harding, N.J., Lucas, E.R., Battey, C.J., Amaya-Romero, J.E., Kern, A.D., Fontaine, M.C., Donnelly, M.J., Lawniczak, M.K.N. and Others (2020) 'Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii', *Genome research*, 30(10), pp. 1533–1546.

Edi, C.V., Djogbénou, L., Jenkins, A.M., Regna, K., Muskavitch, M.A.T., Poupardin, R., Jones, C.M., Essandoh, J., Kétoh, G.K., Paine, M.J.I., Koudou, B.G., Donnelly, M.J., Ranson, H. and Weetman, D. (2014) 'CYP6 P450 Enzymes and ACE-1 Duplication Produce Extreme and Multiple Insecticide Resistance in the Malaria Mosquito Anopheles gambiae', *PLoS genetics*, 10(3). doi:10.1371/journal.pgen.1004236.

Epstein, A., Maiteki-Sebuguzi, C., Namuganga, J.F., Nankabirwa, J.I., Gonahasa, S., Opigo, J., Staedke, S.G., Rutazaana, D., Arinaitwe, E., Kamya, M.R., Bhatt, S., Rodríguez-Barraquer, I., Greenhouse, B., Donnelly, M.J. and Dorsey, G. (2022) 'Resurgence of malaria in Uganda despite sustained indoor residual spraying and repeated long lasting insecticidal net distributions', *PLOS Global Public Health*, 2(9), p. e0000676.

Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V., Jiang, X., Hall, A.B., Catteruccia, F., Kakani, E., Mitchell, S.N., Wu, Y.-C., Smith, H.A., Love, R.R., Lawniczak, M.K.N., *et al.* (2015) 'Extensive introgression in a malaria vector species complex revealed by phylogenomics', *Science*, 347(6217), p. 1258522.

Garud, N.R., Messer, P.W., Buzbas, E.O. and Petrov, D.A. (2015) 'Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps', *PLoS genetics*, 11(2), pp. 1–32.

Grau-Bové, X., Lucas, E., Pipini, D., Rippon, E., van't Hof, A., Constant, E., Dadzie, S., Egyir-Yawson, A., Essandoh, J., Chabi, J., Djogbénou, L., Harding, N., Miles, A., Kwiatkowski, D., Donnelly, M., *et al.* (2020) *Resistance to pirimiphos-methyl in West African Anopheles is spreading via duplication and introgression of the Ace1 locus*, pp. 1–34.

Grau-bové, X., Tomlinson, S., Reilly, A.O.O., Harding, N.J., Miles, A., Kwiatkowski, D., Donnelly, M.J., Weetman, D. and Anopheles, T. (2019) 'Resistance to dieldrin evolution in African malaria vectors is driven by interspecific and interkaryotypic introgression', pp. 1–32.

Guillemaud and Makate (1997) 'Esterase gene amplification in Culex pipiens', *Insect molecular biology* [Preprint]. Available at: http://www.webmail.evolutionhumaine.fr/pdf_articles/guillemaud_1997_insect_molecular_biology.pdf.

Health Organization, W. (2016) *Test procedures for insecticide resistance monitoring in malaria vector mosquitoes*. apps.who.int. Available at: https://apps.who.int/iris/bitstream/handle/10665/250677/9789241511575-eng.pdf (Accessed: 19 September 2022).

Hemingway, J., Coleman, M., Paton, M., McCarroll, L., Vaughan, A. and DeSilva, D. (2000) 'Aldehyde oxidase is coamplified with the World's most common Culex mosquito insecticide resistance-associated esterases', *Insect molecular biology*, 9(1), pp. 93–99.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. and Vinh, L.S. (2018) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular biology and evolution*, 35(2), pp. 518–522.

Ingham, V.A., Pignatelli, P., Moore, J.D., Wagstaff, S. and Ranson, H. (2017) 'The transcription factor Maf-S regulates metabolic resistance to insecticides in the malaria vector Anopheles gambiae', *BMC genomics*, 18(1), p. 669.

Ingham, V., Wagstaff, S. and Ranson, H. (2018) 'Transcriptomic meta-signatures identified in Anopheles gambiae populations reveal previously undetected insecticide resistance mechanisms', *Nature communications* [Preprint]. doi:10.1038/s41467-018-07615-x.

Ismail, H.M., O'Neill, P.M., Hong, D.W., Finn, R.D., Henderson, C.J., Wright, A.T., Cravatt, B.F., Hemingway, J. and Paine, M.J.I. (2013) 'Pyrethroid activity-based probes for profiling cytochrome P450 activities associated with insecticide interactions', *Proceedings of the National Academy of Sciences*, 110(49), pp. 19766–19771.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A. and Jermiin, L.S. (2017) 'ModelFinder: Fast model selection for accurate phylogenetic estimates', *Nature methods*, 14(6), pp. 587–589.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome biology*, 15(12), pp. 1–21.

Love, R.R., Redmond, S.N., Pombi, M., Caputo, B., Petrarca, V., della Torre, A. and Besansky, N.J. (2019) 'In Silico Karyotyping of Chromosomally Polymorphic Malaria Mosquitoes in the Anopheles gambiae Complex', *G3: Genes, Genomes, Genetics*, 9(10), pp. 3249–3262.

Lucas, E.R., Miles, A., Harding, N.J., Clarkson, C.S., Lawniczak, M.K.N., Kwiatkowski, D.P., Weetman, D., Donnelly, M.J. and Anopheles gambiae 1000 Genomes Consortium (2019) 'Whole-genome sequencing reveals high complexity of copy number variation at insecticide resistance loci in malaria mosquitoes', *Genome research*, 29(8), pp. 1250–1261.

Lucas, E.R., Miles, A., Harding, N.J., Clarkson, C.S., Mara, K., Lawniczak, N., Kwiatkowski, D.P., Weetman, D., Donnelly, M.J. and Place, P. (2018) 'Whole genome sequencing reveals high complexity of copy number variation at insecticide resistance loci in malaria mosquitoes', pp. 1–28.

Lucas, E.R., Nagi, S.C., Egyir-Yawson, A., Essandoh, J., Dadzie, S., Chabi, J., Djogbenou, L.S., Medjigbodo, A.A., Edi, C.V., Ketoh, G.K., Koudou, B.G., Van't Hof, A.E., Rippon, E.J., Pipini, D., Harding, N.J., *et al.* (2023) 'Genome-wide association studies reveal novel loci associated with pyrethroid and organophosphate resistance in Anopheles gambiae s.l', *bioRxiv*. doi:10.1101/2023.01.13.523889.

Maharaj, R., Mthembu, D.J. and Sharp, B.L. (2005) 'Impact of DDT re-introduction on malaria transmission in KwaZulu-Natal', *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde*, 95(11), pp. 871–874.

Miles, A., Harding, N.J., Bottà, G., Clarkson, C.S., Antão, T., Kozak, K., Schrider, D.R., Kern, A.D., Redmond, S., Sharakhov, I., Pearson, R.D., Bergey, C., Fontaine, M.C., Donnelly, M.J., Lawniczak, M.K.N., *et al.* (2017) 'Genetic diversity of the African malaria vector Anopheles gambiae', *Nature*, 552, pp. 96–100.

Minh, B.Q., Nguyen, M.A.T. and von Haeseler, A. (2013) 'Ultrafast approximation for phylogenetic bootstrap', *Molecular biology and evolution*, 30(5), pp. 1188–1195.

Nagi, S.C., Miles, A. and Donnelly, M.J. (2023) 'AgamPrimer: Primer Design in Anopheles gambiae informed by range-wide genomic variation', *bioRxiv*. doi:10.1101/2022.12.31.521737.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q. (2015) 'IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular biology and evolution*, 32(1), pp. 268–274.

Njoroge, H., Van't Hof, A., Oruni, A. and Pipini, D. (2022) 'Identification of a rapidly-spreading triple mutant for high-level metabolic insecticide resistance in Anopheles gambiae provides a real-time molecular diagnostic for ...', *Molecular* [Preprint]. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16591.

Paradis, E. and Schliep, K. (2018) 'ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R', *Bioinformatics* , 35(3), pp. 526–528.

Raymond, M., Callaghan, A., Fort, P. and Pasteur, N. (1991) 'Worldwide migration of amplified insecticide resistance genes in mosquitoes', *Nature*, 350(6314), pp. 151–153.

Raymond, M., Chevillon, C., Guillemaud, T., Lenormand, T. and Pasteur, N. (1998) 'An overview of the evolution of overproduced esterases in the mosquito Culex pipiens', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 353(1376), pp. 1707–1711.

Raymond, M., Qiao, C.L. and Callaghan, A. (1996) 'Esterase polymorphism in insecticide susceptible populations of the mosquito Culex pipiens', *Genetical research*, 67(1), pp. 19–26.

Revell, L.J. (2011) 'phytools: an R package for phylogenetic comparative biology', *Methods in Ecology and Evolution* [Preprint]. doi:10.1111/j.2041-210X.2011.00169.x.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooperk, R., Kwiatkowski, D., *et al.* (2002) 'Detecting recent positive selection in the human genome from haplotype structure', *Nature*, 419(October). doi:10.1038/nature01027.

Shi, X., Dick, R.A., Ford, K.A. and Casida, J.E. (2009) 'Enzymes and inhibitors in neonicotinoid insecticide metabolism', *Journal of agricultural and food chemistry*, 57(11), pp. 4861–4866.

Weedall, G.D., Riveron, J.M., Hearn, J., Irving, H., Kamdem, C., Fouet, C., White, B.J. and Wondji, C.S. (2020) 'An Africa-wide genomic evolution of insecticide resistance in the malaria vector Anopheles funestus involves selective sweeps, copy number variations, gene conversion and transposons', *PLoS genetics*, 16(6), p. e1008822.

4.7 Appendix

4.7.1 Appendix A - Phylogenetic analysis

To examine the relationships between haplotypes whilst accounting for molecular evolution we performed a phylogenetic analysis of haplotypes at three loci around *Coaeae1f and Coeae2f*. We selected a 10kb window, focused on the intergenic region between the two genes, and also two loci 2Mb upstream and downstream from this focal locus. To ease the interpretation of the resulting phylogeny, we excluded all haplotypes in which the parent individual was heterozygous for the 2La inversion. The 2La inversion cannot be reliably phased (due to switch errors on the haplotypes themselves), and so labelling heterozygotes on the phylogeny can obscure signals.

As with the haplotype clustering, two main clades are revealed by the phylogeny, corresponding to haplotypes with either the 2La or 2L+a inversion allele. This is expected as population structure in this genomic region is driven by karyotype and not species (Miles *et al.*, 2017), due to the inversion predating the radiation of the *An. gambiae* complex (Fontaine *et al.*, 2015). Figure A1 shows a phylogeny of all haplotypes in the data, coloured by species, haplotype cluster and karyotype. As with haplotype clustering, An. *arabiensis* haplotypes reside in the 2La clade. The six detected selective sweeps are scattered throughout the phylogeny, with no obvious pattern or clustering.

One interesting note, observed both in haplotype clustering and in the phylogenies, is that *An. gambiae* and *An. coluzzii* haplotypes seem to be more intermixed in the 2La clade, as opposed to the 2L+a clade, in which branches are more distinct for each species. It is not clear why this is the case, though it could be related to varying levels of hybridisation in coastal West Africa where 2L+a dominates, compared to the Sahel, in which 2La is predominant. This is particularly notable in the upstream and downstream region (Appendix A2). We can see that in the downstream and upstream regions, as we move away from the focal locus, haplotype clusters no longer cluster together, as recombination has broken the haplotypes down.

We also observe evidence of recombination between karyotypes. In the right-hand plot of Figure A1, multiple 2l+a haplotypes (as determined by tagging SNPs mostly at either end of the inversion) are found in the opposing clade, and vice versa. There is one *An. gambiae* haplotype in cluster C3 that is assigned a 2La karyotype, but contains the C3 sweep which is only found in 2l+a haplotypes.

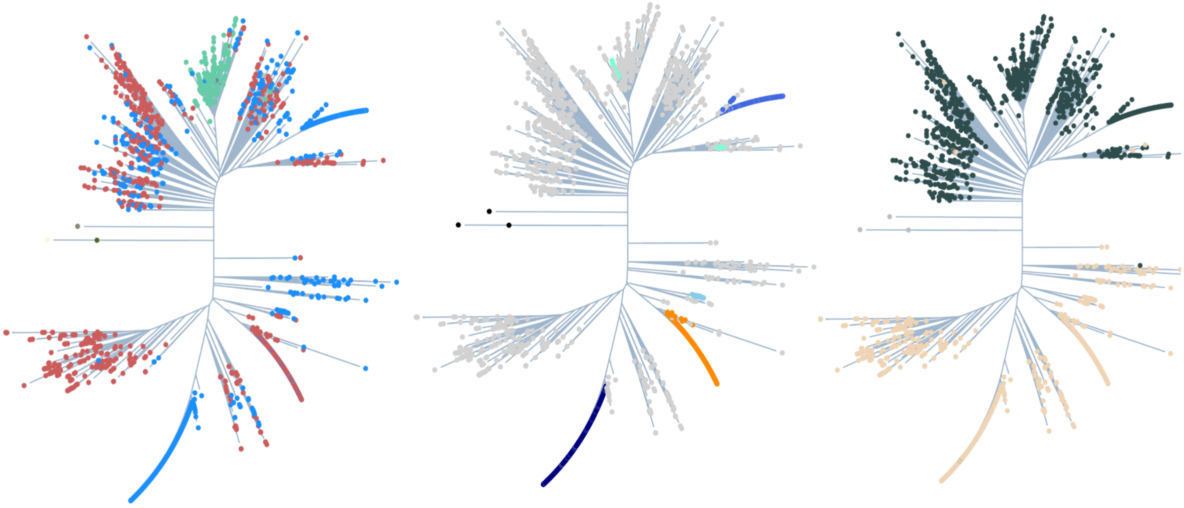
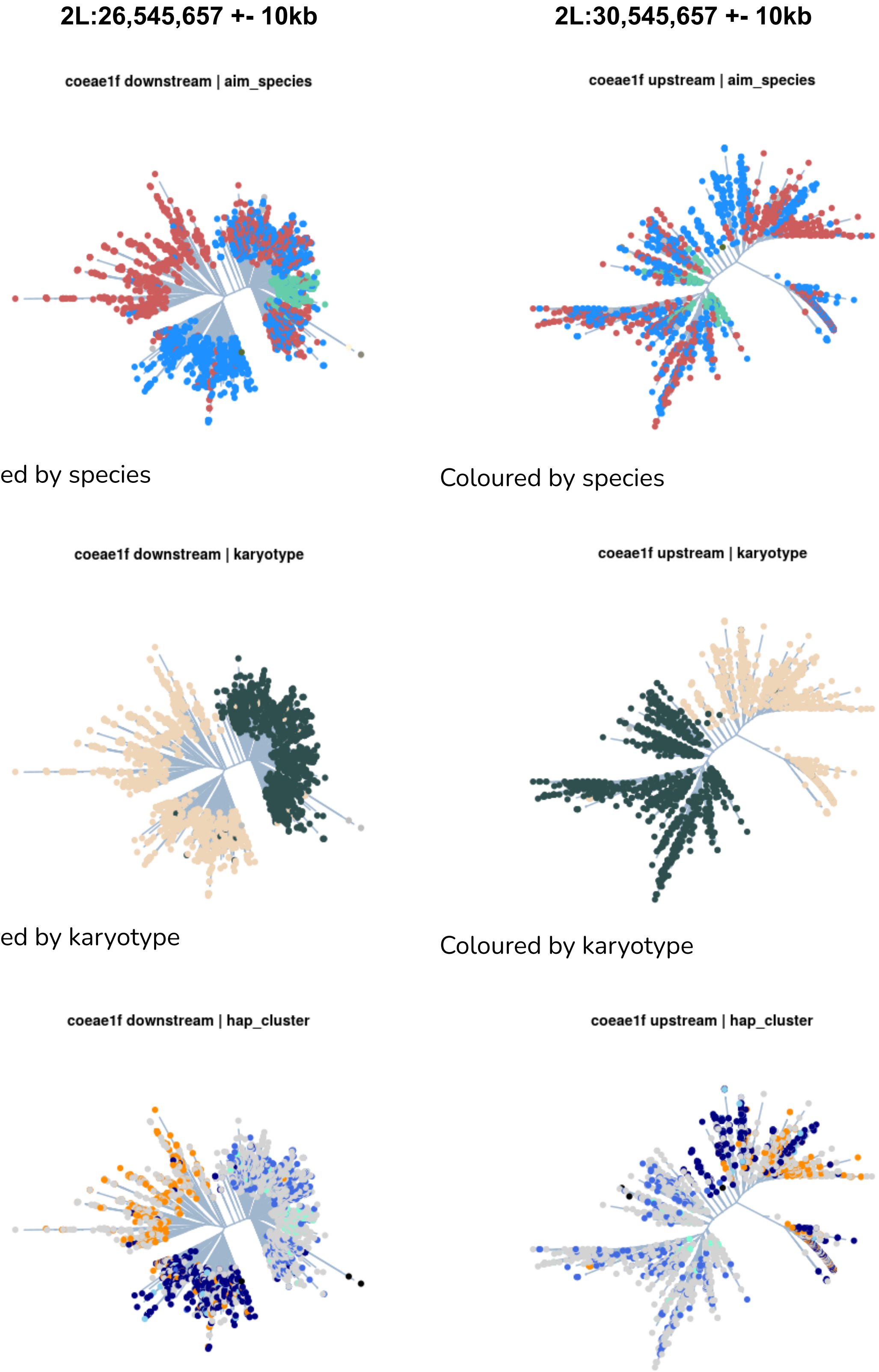


Figure A1. Phylogenetic tree of the *Coeae1f/Coeae2f* region. Left: coloured by species, blue=*gambiae*, red=*coluzzii*, teal=*arabiensis*. **Middle**: coloured by haplotype cluster. **Right**: coloured by karyotype, gold=2l+a, dark=2La. Nodes are coloured by species - *gambiae* = red, *coluzzii* = blue, *arabiensis* = teal. Multiple sweeps can be observed (lines of nodes next to each other) - we have set a minimum distance between neighbouring nodes so that they do not fully overlap. Dark-coloured points are the outgroups; *An. melas*, *An. merus, and An. quaddrianulatus*.



Coloured by species

Coloured by species

Coloured by karyotype

Coloured by karyotype

Coloured by haplotype cluster

Coloured by haplotype cluster

Figure A2-A7. Haplotype phylogenies of the downstream (2L:26,545,657 +- 10kb) and upstream (2L:30,545,657 +- 10kb) regions to the *Coeae1f* locus.

4.7.2 Appendix B - Cluster C1 and Dup2 EHH

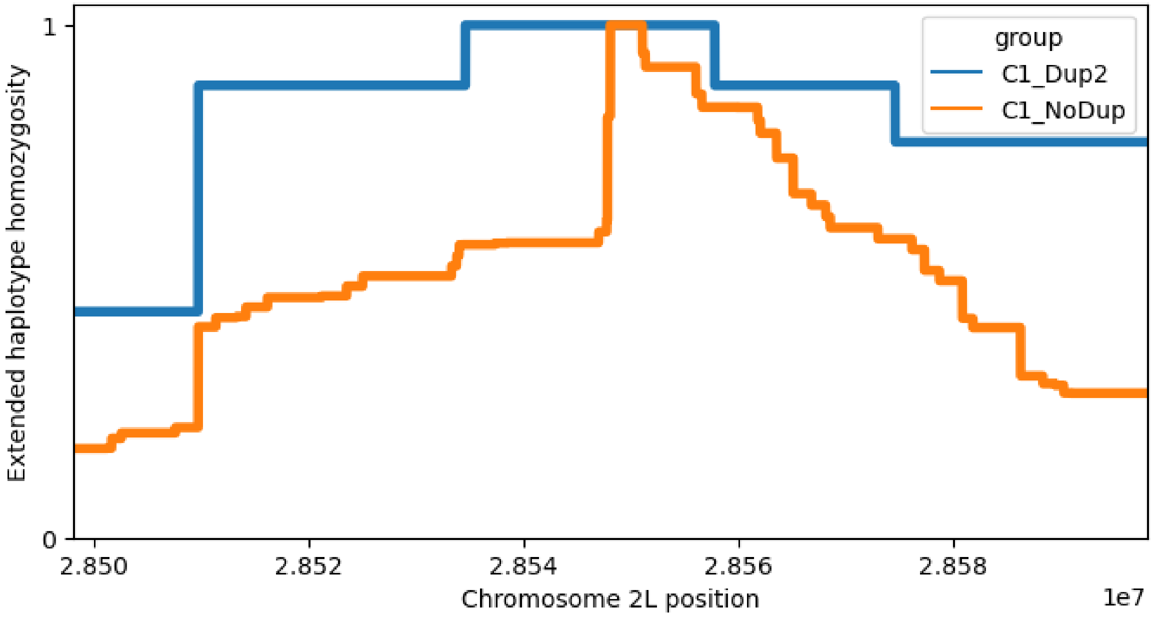


Figure B1. Positive selection of haplotypes in the C1 cluster, split by CNV status. Extended haplotype homozygosity, moving 50kb on either side of the focal intergenic region between *Coeae1f* and *Coeae2f*.

4.7.3 Appendix C - Haplotype clustering and amino acid variation

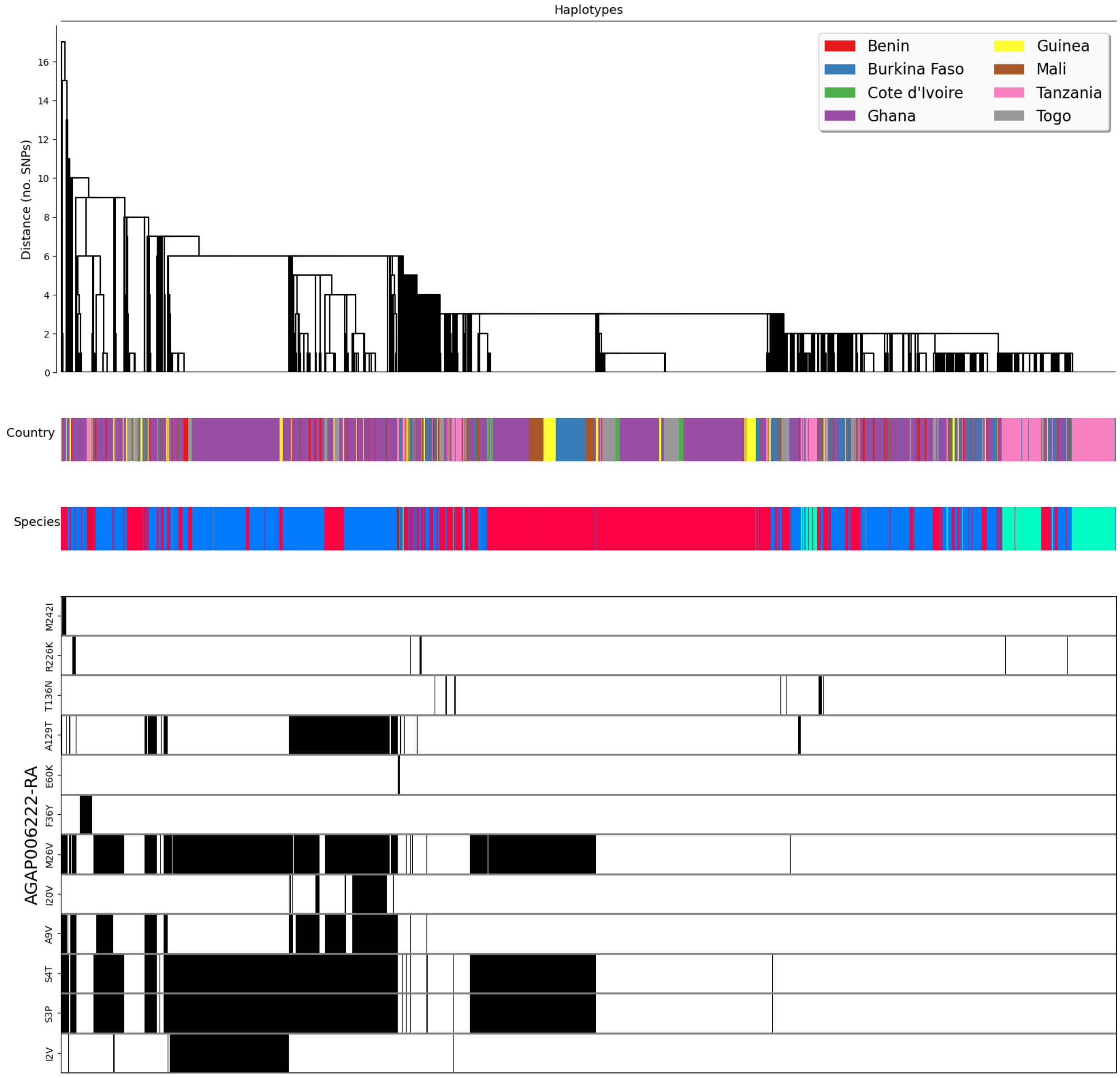


Figure C1. Haplotype cluster plot for AGAP006222-RA.

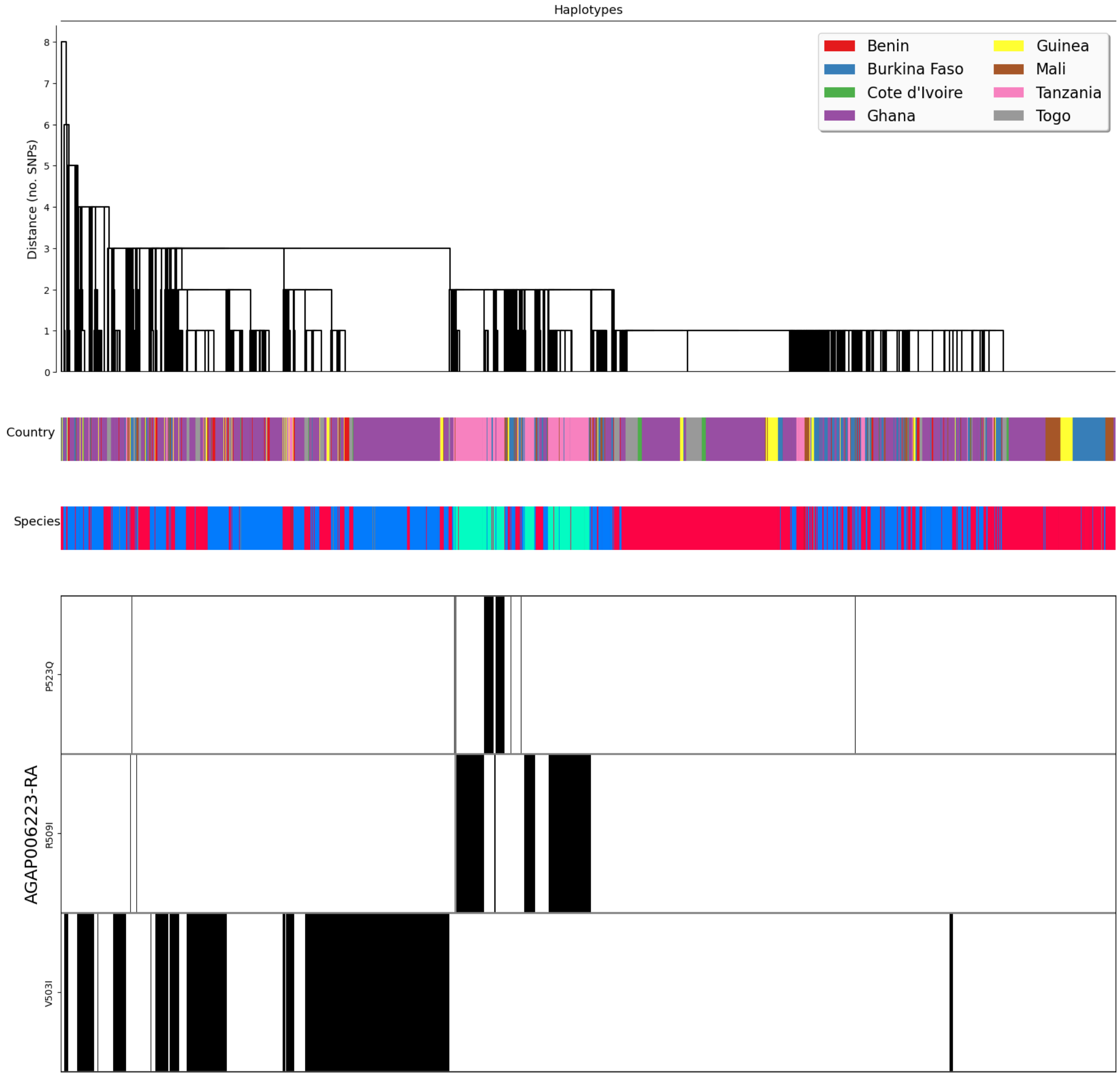


Figure C2. Haplotype cluster plot for AGAP006223-RA

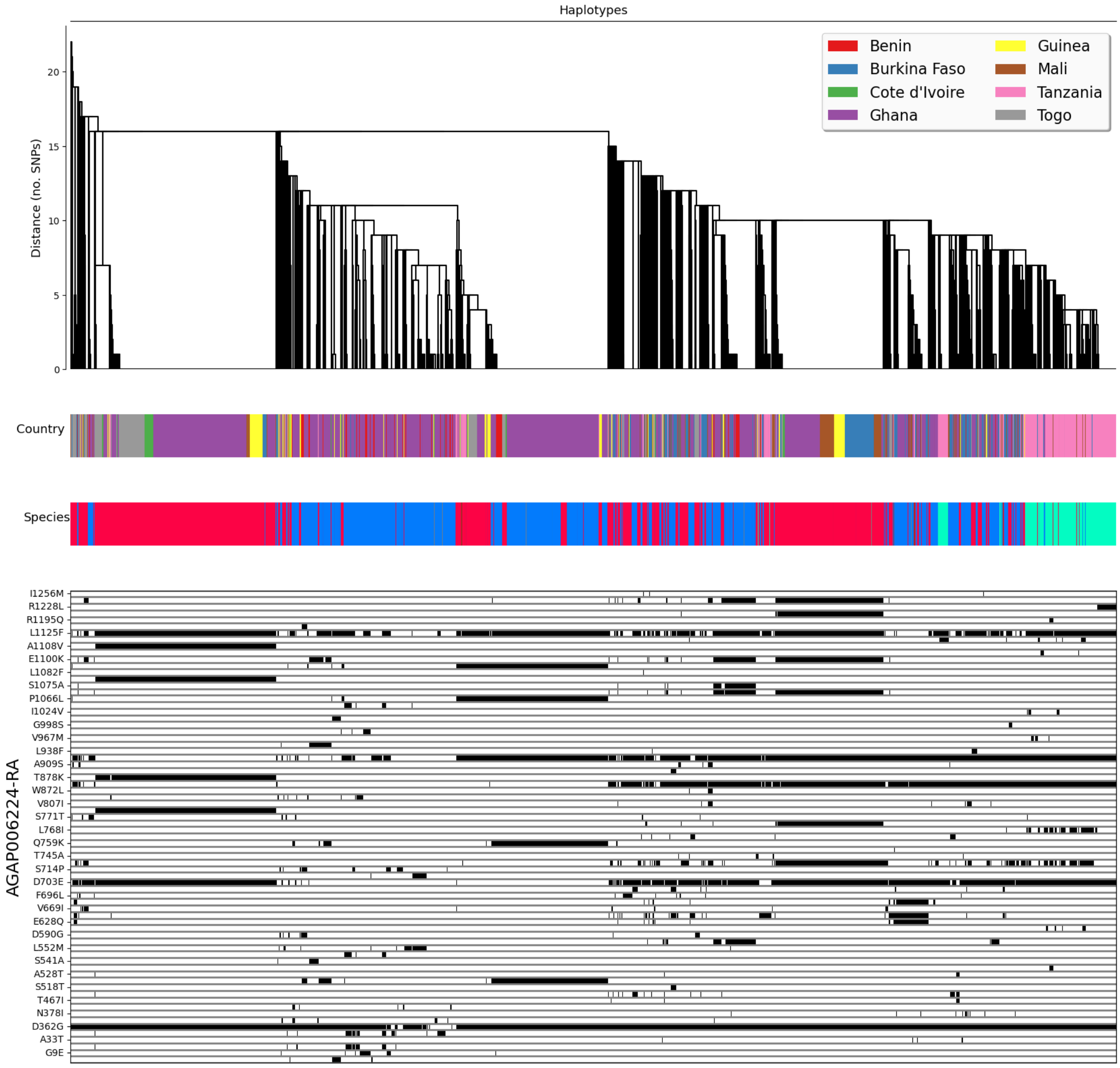


Figure C3. Haplotype cluster plot for AGAP006224-RA.

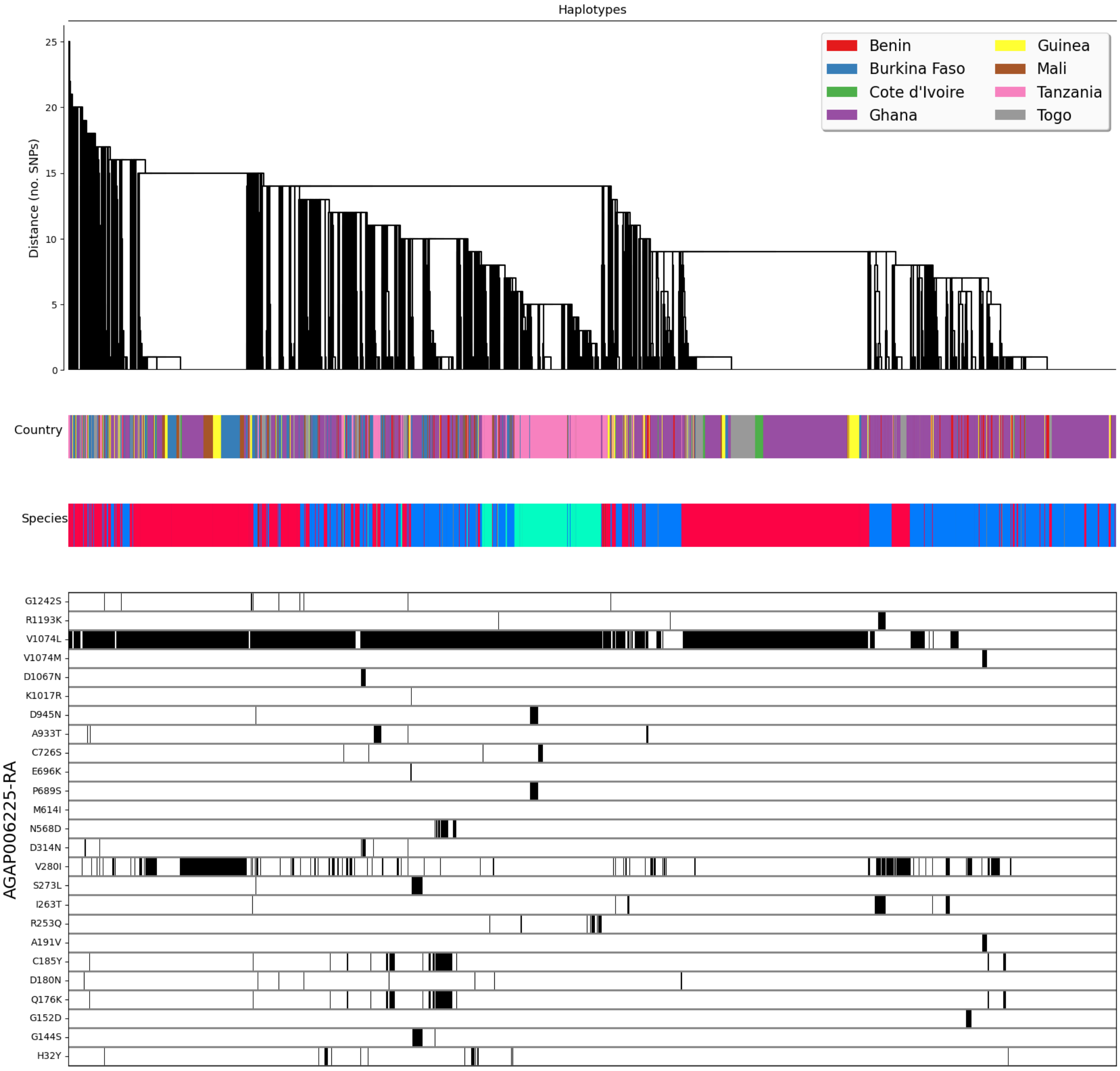


Figure C4. Haplotype cluster plot for AGAP006225-RA.

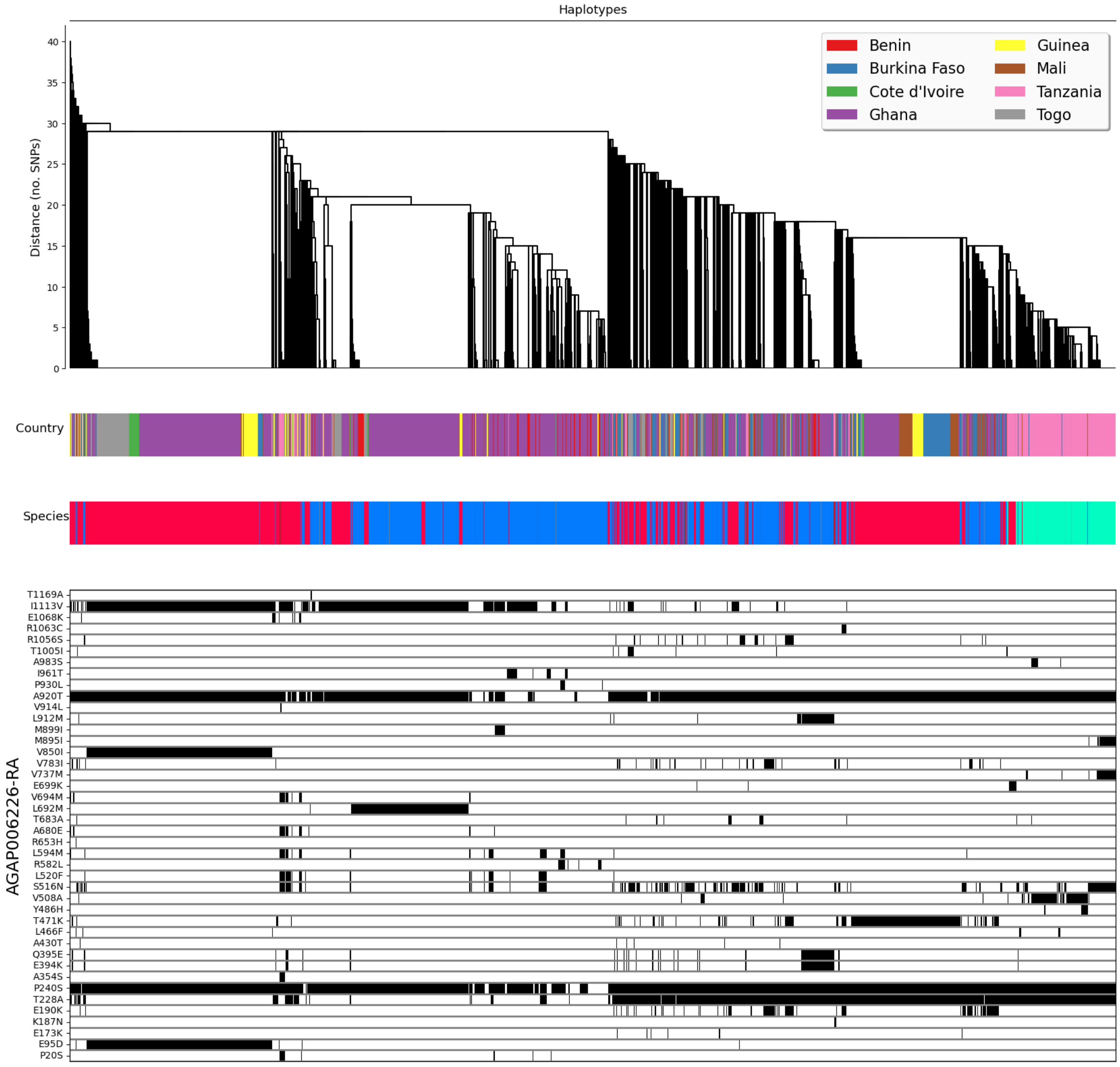


Figure C5. Haplotype cluster plot for AGAP006226-RA.

4.7.4 Appendix D - Haplotype tagging SNPs

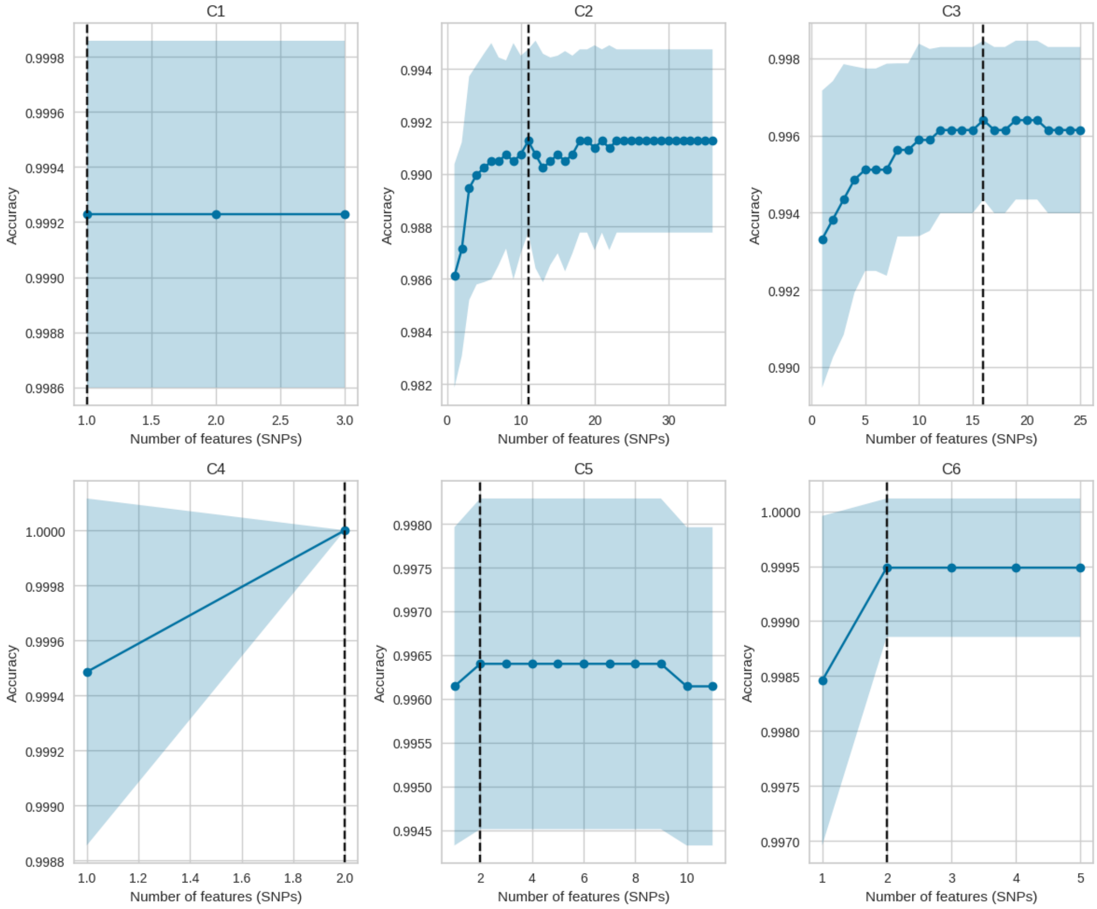
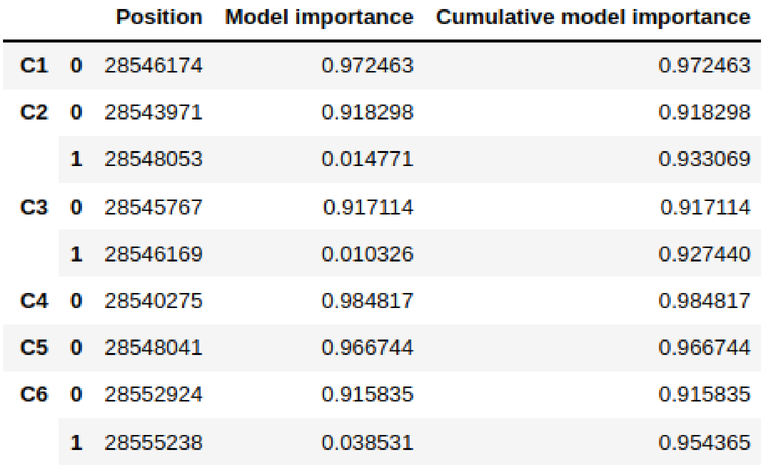


Figure D1. Results from classification and regression tree (CART). The plots show recursive feature elimination with cross-validation, for each haplotype cluster. The accuracy of the model is plotted against the number of SNPs used in the model.

Table D1. The topmost important SNPs for distinguishing clusters, where we request the algorithm to give us a maximum of 2 SNPs.



		Position	Model importance	Cumulative model importance
C1	**0**	28546174	0.972463	0.972463
C2	**0**	28543971	0.918298	0.918298
	1	28548053	0.014771	0.933069
C3	**0**	28545767	0.917114	0.917114
	1	28546169	0.010326	0.927440
C4	**0**	28540275	0.984817	0.984817
C5	**0**	28548041	0.966744	0.966744
C6	**0**	28552924	0.915835	0.915835
	1	28555238	0.038531	0.954365

4.7.5 Appendix E - LNA probes

Table E1. Information on a qPCR LNA probe to target and track the C3 haplotype.

Name	5’ Fluorophore	Sequence	3’ Quencher
E477V 2F		CAACACTGCCGCCAACATTC	
E477V 2R		GACGATGCAAACGCTGGTAA	
E477 – WT	HEX	A+CAA+C+T+CG+A+CC	IBFQ
477V – C3	6-FAM	A+A+CAA+C+A+CG+ACCT	IBFQ

5

A population genomic analysis of *Anopheles gambiae* from Obuasi, Ghana

This chapter is in the format introduction-results/discussion-conclusion-methods as it will be worked on further for submission to Molecular Biology & Evolution.

5.1 Abstract

Although great progress has been made in reducing the burden of malaria, progress may be stalling, likely in part due to insecticide resistance. In Ghana, *An. gambiae s.l* mosquitoes have been documented as highly resistant to pyrethroid insecticides, with emergent resistance to other classes. Monitoring the frequency and spread of insecticide resistance alleles and their spatial heterogeneity is important for malaria control programmes, and the spread of resistance alleles in *An. gambiae s.l* will depend strongly on population structure. We present a descriptive population genomic analysis of micro-spatial population structure and the genomic basis of resistance from whole-genome sequences of 485 *An. gambiae s.l* mosquitoes from Obuasi, central Ghana. We detect isolation-by-distance in *An. coluzzii* at a high spatial resolution, and find that geographic distance, rather than environment, drives patterns of isolation-by-distance. We elucidate the continued evolution of the target of pyrethroid insecticides, the *Voltage-gated sodium channel*, and demonstrate remarkable numbers of haplotypes under selection at other loci important for insecticide resistance. We also present *Probe*, a snakemake workflow which can perform a subset of the population genomic analyses herein on data from VCF files or directly from Ag1000G samples using *malariagen_data*.

5.2 Introduction

Evidence suggests insecticide resistance (IR) may be compromising the efficacy of malaria control interventions (Kafy *et al.*, 2017; Epstein *et al.*, 2022). Repeated and ubiquitous exposure to toxic chemicals has led to rapid adaptation in mosquito vectors, allowing them to survive substantially higher doses (Oumbouke *et al.*, 2020). In Ghana, *An. gambiae s.l* have been documented to be highly resistant to insecticides, including in Obuasi specifically (Pwalia *et al.*, 2019; Hamid-Adiamoh *et al.*, 2020; Mugenzi *et al.*, 2022; Lucas *et al.*, 2023).

Although certain resistance alleles, such as *Kdr*, have spread across much of sub-Saharan Africa (Clarkson *et al.*, 2021), many other mechanisms of resistance seem to remain localised to specific areas (Williams, Ingham, *et al.*, 2022). This may result from patterns of insecticide usage in the direct environment, providing hotspots of selection pressure (Tepa *et al.*, 2022). It could also arise from micro-spatial population structure, with minor barriers to gene flow in which only alleles with large enough selective advantage can cross, or the process may be stochastic in nature. Understanding these processes is important for the prospect of targeted vector control, where interventions can be concentrated in hotspots of malaria transmission (Stresman *et al.*, 2019).

The utility of genomic surveillance has enabled researchers to study the genomic basis of insecticide resistance in recent years (Miles *et al.*, 2017; Lucas *et al.*, 2019; Ag1000G, 2020; Clarkson *et al.*, 2021; Grau-Bové *et al.*, 2021; Lucas *et al.*, 2023). The spread of alleles involved in insecticide resistance in the major malaria vector *An. gambiae* is of major relevance to public health (WHO, 2012). This spread of alleles will depend on multiple factors, including the selective advantage of the allele, its dominance and genetic structure in the vector population (Labbé *et al.*, 2007). The selective advantage will depend on the protective effect conferred against insecticides, as well as any inherent fitness costs, and the degree of insecticide exposure in the mosquito's environment (Wood *et al.*, 1983; Hawkins *et al.*, 2019).

The majority of these genomic studies, however, have taken place over extremely large spatial scales, often continent-wide, leaving researchers unable to explore questions on micro-spatial scales (Ag1000G, 2020). They have also been performed without predefined sampling frameworks - sampling sites have been selected for their convenience and familiarity, rather than any informed design. In 2019, Sedda *et al.* developed a sampling framework for the surveillance of malaria mosquitoes, which incorporates ecological data to optimise the accuracy of abundance estimates (Sedda *et al.*, 2019). Using ecologically-informed sampling frameworks may allow us to understand the contribution of the environment on genotype distributions with greater resolution. Understanding these processes is also central to the use of CRISPR-based gene drives, that either replace or suppress the vector population (Kyrou *et al.*, 2018; Dhole *et al.*, 2020). By calculating kinship between sampled individuals, it is possible to estimate dispersal and migration parameters in the population, as has been done in other mosquito species (Bravington *et al.*, 2016; Filipović *et al.*, 2020; Schmidt *et al.*, 2022).

The ecologically-informed sampling framework developed by Sedda *et al.* (2019) combines a regular lattice with random points as close pairs, in order to maximise spatial coverage, the representativeness of ecological zones and vector spatial autocorrelation. In this framework, 70% of sampling points are in a lattice, with the remaining 30% of points randomly allocated as close-pairs. In order to ensure ecological representativeness, each ecological stratum must contain a number of sampling sites proportional to the stratum size.

In this study, we collect *An. gambiae s.l* mosquitoes using this ecologically-informed sampling framework in a 70km^2 region around Obuasi in central Ghana. We perform a population genomic analysis using whole-genome sequencing of 485 *An. gambiae* mosquitoes, investigating relatedness and population structure at ultra-fine spatial scales. We explore the genomics of insecticide resistance and report allele and haplotype frequencies of known insecticide resistance loci, detecting signals of selection and examples of adaptive introgression.

5.3 Results / Discussion

5.3.1 Sample collections

Mosquitoes were collected from Obuasi, central Ghana. Obuasi is located in the southern Ashanti region, and is known for substantial mining activities; the Obuasi gold mine is one of the largest underground gold mines in the world and has been in use since the seventeenth century (Fougerouse *et al.*, 2017). Mining activities are a mixture of large mines run by multinational corporations or artisanal and small-scale miners (ASM). ASM mines and quarries have been reported to create *Anopheles* breeding sites (Ferring *et al.*, 2019).

To inform sampling design, the spatially explicit sampling framework described earlier was implemented, stratifying regions based on ecological variation to provide a more representative sample of the vector population (Sedda *et al.*, 2019). Mosquitoes were collected indoors using CDC light traps from four houses in each sample site.

5.3.2 Mapping samples

A total of 485 *An. coluzzii* (n=422) and *An. gambiae s.s.* (n-63) samples were whole-genome sequenced to a target coverage of 30X. No *An. arabiensis* or cryptic species were detected in the data. Reads were aligned to the AgamP4 reference genome and analysed as previously described (Ag1000G, 2020). We discovered a total of 43,730,731 segregating SNPs in the dataset across all chromosomes. SNPs were high-confidence, passing all quality filters as previously described (Ag1000G, 2020). The dataset also consists of phased biallelic haplotypes. Figure 1 shows a map of the Obuasi region, with sample collection sites, the relative proportion of *An. gambiae s.s* and *An. coluzzii* shown, and whether ASM mines are present at the site.

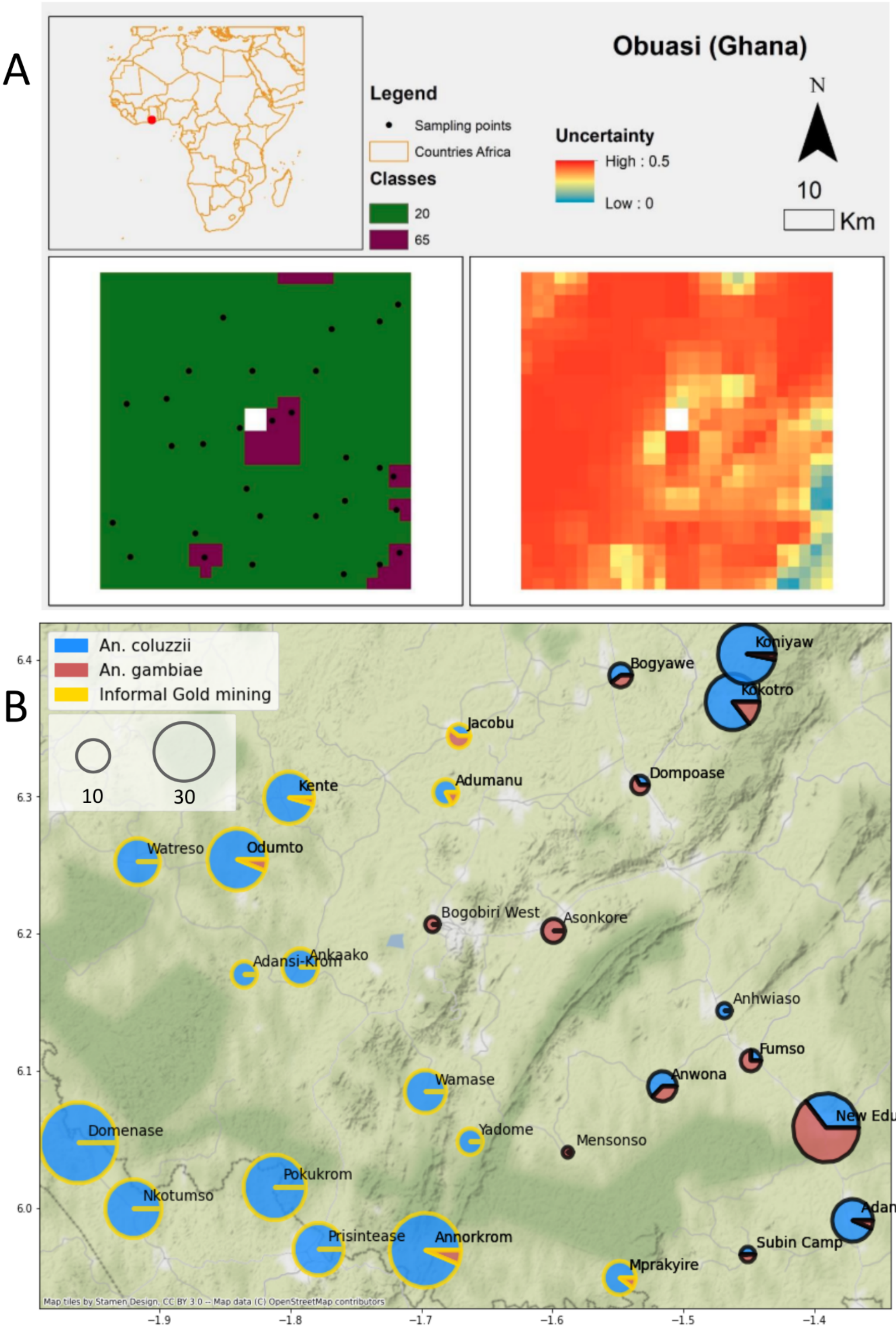


Figure 1. Ecological classification of the Obuasi district and a map of the sampling locations. A) Adapted from (Sedda *et al.*, 2019). The bottom left plot shows sampling locations and the ecological classification based on quadratic discriminant analysis. The dark green (class=20 is described as 'Forest'. The Purple (class=65) is described as a 'mixture of Urban, Tundra, Wetland, Water bodies and Grassland'. Bottom right displays

uncertainty in the ecological classifications of the model. Uncertainty is measured as the sum of the probabilities that a point belongs to any of the other classes. B) Map of whole-genome sequenced samples. Pie charts illustrate the relative proportion of *An. coluzzii* and *An. gambiae s.s.* The size of the circles indicates the sample size. Gold circles indicate whether one or more informal ASM gold mines are located nearby the sampling location.

5.3.3 Population structure

We performed principal components analysis (PCA) on a region of the 3L chromosomal arm between 15 Mb and 44 Mb, selected to avoid chromosomal inversions, regions of low recombination and regions with known selective sweeps. Figure 2 displays the PCAs for both *An. gambiae s.s* and *An. coluzzii* samples combined. As expected there is clear segregation by species, with PC1 capturing variation that distinguishes the two species, and PC2 capturing variation within *An. gambiae s.s.* We can also observe two outliers assigned as *An. gambiae s.s.* We also perform PCA for both species in isolation (Appendix A). Figure A1A shows *An. coluzzii*, in which two pairs of outliers appear distinct from the other individuals. Interestingly, each pair of outliers were sampled from the same village. Removal of these four outliers (Figure A1B) results in four new samples being pulled out by the PCA as outliers, which previously clustered with the rest of the *An. coluzzii* individuals. This indicated to us that there was little population structure in the data.

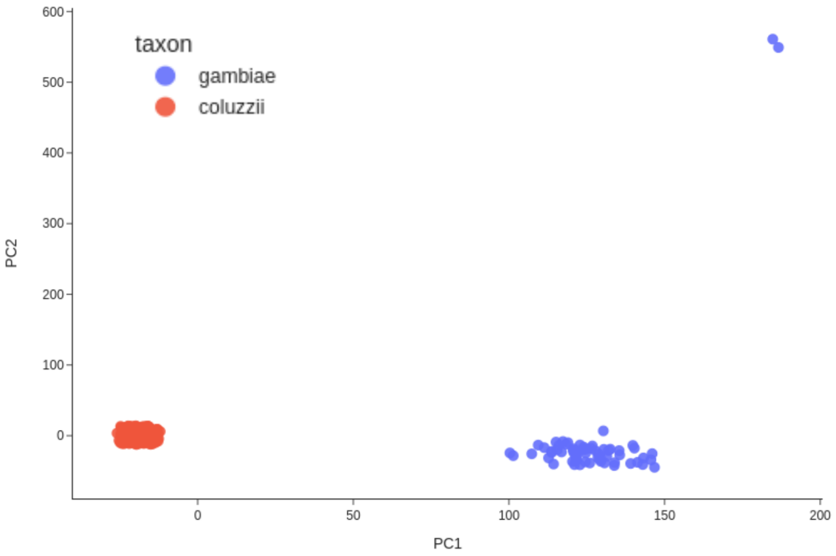


Figure 2. Principal components analysis on genotype data from the 3L chromosomal arm, 15 Mb - 44 Mb. 100,000 SNPs were randomly selected from this region.

In *An. gambiae*, we observed a similar pattern (Appendix A, Figure A1C), with two outlying samples. However, the removal of these two samples resulted in a PCA which was much more clearly without structure (Figure A1D). Overall, the principal components analysis suggests that there is little variation within *An. coluzzii* or *An. gambiae s.s*, in agreement with earlier findings that this region of West Africa is relatively homogeneous (Ag1000G, 2020).

The observation that the outliers of the PCAs came in pairs and that each time they originated from the same village, led us to speculate that perhaps these pairs were related in some way. Sampling mosquitoes at fine-spatial scales allow us to potentially detect kin between sites which is important to avoid confounding in population and statistical genetics, and may be useful to study dispersal (Hoffman, 2013; Schmidt *et al.*, 2022).

To estimate relatedness, we used NgsRelate (Korneliussen *et al.*, 2015) to calculate the KING-robust statistic (Manichaikul *et al.*, 2010) from biallelic SNP markers across the whole-genome for every pair of individual mosquitoes in the dataset. To evaluate how chromosomal inversions affect the sibship inference, chromosomal inversions were also scored with a modified version of compkaryo (Love *et al.*, 2019). Appendix B shows karyotype frequencies for each sample site in Obuasi for the 2La and 2Rb inversions.

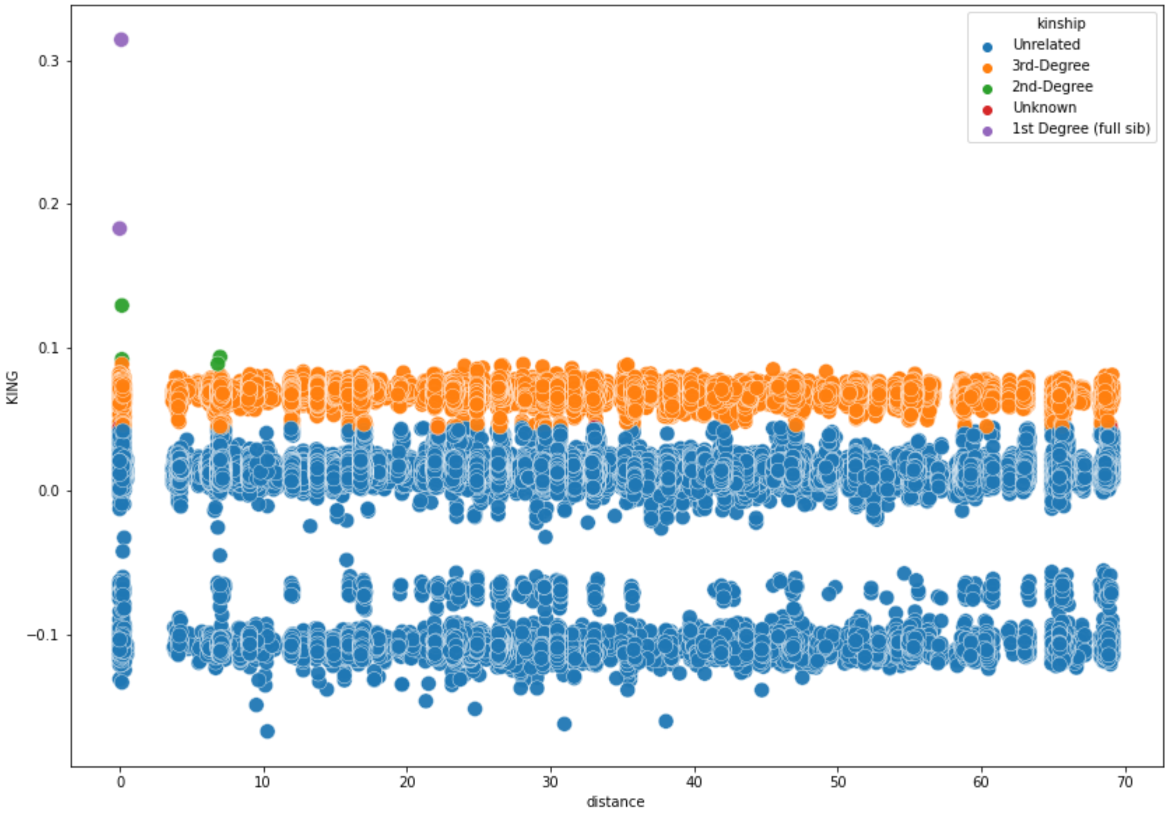


Figure 3. Pairwise relatedness (KING-robust) between samples plotted against geographic distance in kilometres. Points are coloured depending on the thresholds advised by the KING authors. Pairwise relatedness is calculated using NgsRelate on biallelic SNPs across the whole genome. 1st Degree kin equate to full sibling or parent-child relationships, 2nd Degree is aunt, uncle, grandparent, grandchild, niece, nephew, or half-sibling relationship.

Figure 3 shows the results of the kinship analysis. Three clusters of KING-robust values can be observed which interestingly, relate to the pairwise 2La karyotype concordance. The negative values centering around -0.1 correspond to comparisons in which both individuals were homozygous but for alternative 2La inversion karyotypes, the cluster centred around 0 is primarily pairwise comparisons between homozygotes and heterozygotes, and the cluster at ≈0.075 are pairwise comparisons with identical 2La karyotypes (Appendix C). According to the thresholds set by the authors (Manichaikul *et al.*, 2010), almost all intra-karyotypic comparisons are defined as 3rd-degree relatives, regardless of physical distance. This suggests that the KING estimator may have limited resolution at lower degrees of relatedness, at least when inversions are present in the data. It is possible that the unusual patterns of linkage caused by the 2La inversion are confounding the KING analysis. The 2Rb inversion is at much lower frequencies in Obuasi (Appendix A) and is also significantly smaller than 2La, and so does not drive these KING values. In *An. gambiae s.l*, it seems that inversions may have a substantial impact on kinship inference using KING, and inversion regions should be excluded when calculating kinship. At the time of writing, this work is ongoing.

Two 1st degree siblings relationships were, however, detected in the data (Figure 3). The two *An. coluzzii* individuals with the highest KING value (0.314) were designated as full siblings and were sampled from Odumto, but were caught in different houses, approximately 111m away from each other. We explored genome-wide patterns of differentiation between the putative siblings to investigate genome-wide inheritance in detail. Figure 5 shows F_{st} between these two individuals, calculated in 10,000bp stepping windows across the genome. Three levels of F_{st} can be observed, depending on whether siblings inherited zero, one, or both of the same segments of chromosomes from their parents at any given position. F_{st} on the X chromosome is -1, indicating that both individuals received the only paternal X chromosome, and the same maternal X chromosome. Between these two individuals, there is also a negative signal at the 2La inversion (Chromosome 2: ~80-100 Mb), suggesting, and confirmed by *in silico* karyotyping that these individuals have identical inversion karyotypes (both 2La heterozygotes).

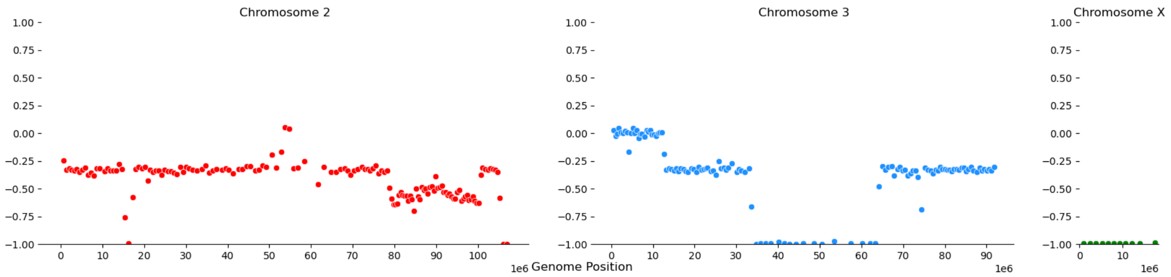


Figure 4. Plots of Hudson's F_{st} across the genome of two individuals designated as full siblings as per the KING statistic (WA-2361, WA-2363, Female *An. coluzzii* collected in Odumto, in separate houses 111m apart). F_{st} was calculated in 10,000 bp stepping windows.

The other pair of full siblings were also *An. coluzzii,* caught in Annorkrom, however, in this case both individuals were captured in the same household. A second-degree relationship was also found between two *An. gambiae s.s* samples, caught in New Edubiase, 156m apart from each other (F_{st} plots; Appendix D3).

We also produced F_{st} plots for the outliers in the PCAs (Figure 2, Figure A1A). Genome-wide F_{st} data between the *An. gambiae* outliers in Appendix A1C are shown in Appendix D1. In the *An. coluzzii* PCA, one set of outliers are actually the 1st-degree siblings with the highest KING value. Genome-wide F_{st} for the other pair (WA-2014 & WA-2009) is shown in Appendix D3. They look related based on the number of recombination breakpoints, but are homozygous but different for the 2La inversion, resulting in a KING value of only 0.03 (classed as ‘Unrelated’).

Using the previously published (Zheng *et al.*, 1996) estimates of the per-base recombination rate in *An. gambiae* of $10^{-8}bp^{-1}$, we calculated the expected number of recombination events per generation for chromosome 2. After 1000 simulations, the mean for this value was 1.14 recombination events (95% CIs: 0.00-3.15) with 30.4% of generations having 0 recombination events. This number of crossover events is slightly less than observed in humans, which is 2-3 events per chromosome (Alberts *et al.*, 2002), although most human chromosomes are much larger than *An. gambiae* chromosomes. *An. gambiae* and other mosquitoes also only have 3 chromosomes compared to humans (n=23 pairs). It is therefore likely that by chance, pairs of *An. gambiae* siblings will have much higher variance in their actual genetic relatedness than human siblings. For example, we expect human siblings to be identical-by-descent across ~50% of their genome, but in *An. gambiae* it could be considerably lower or higher than this value, and siblings could in some cases appear more like genetic cousins, or closer to genetic twins. This is likely to make kinship inference challenging in *An. gambiae*, and may mean that we should use some form of kinship likelihood rather than a single prediction of kinship category, particularly in future statistical methods relating to dispersal.

To investigate population structure further, we calculated genetic diversity metrics from the high-quality genotypes. Specifically, we only used SNP sites at fourfold degenerate codons; SNPs in coding regions of the genome, in which any base change does not modify the resulting codon. These sites were selected as they should be under little to no selection and so are highly reflective of neutral evolution (Perna *et al.*, 1995). As we had many more *An. coluzzii* samples spread over a larger region, we split the *An. coluzzii* samples into two cohorts corresponding to Ashanti district and central Obuasi district.

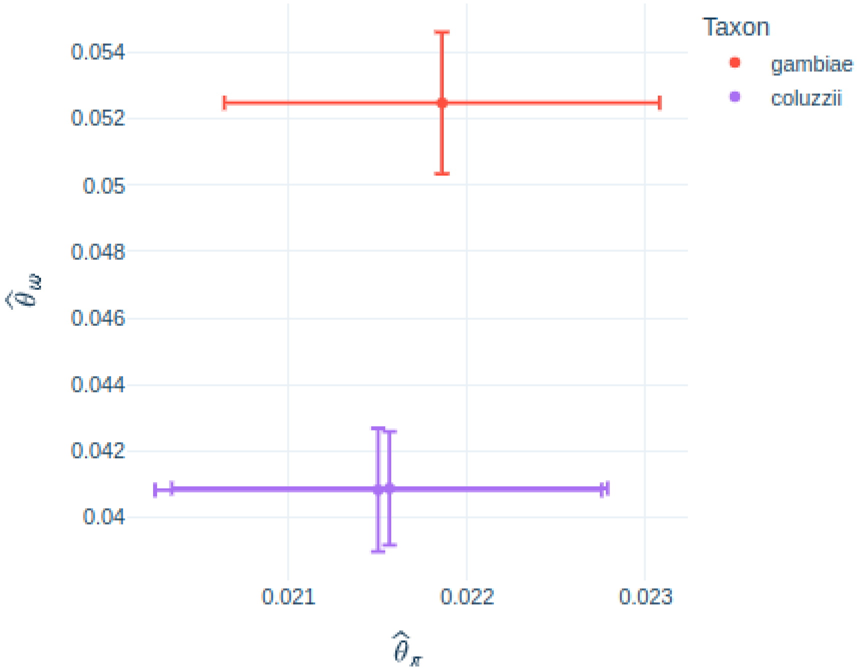


Figure 5. Genetic diversity estimates for *An. gambiae s.s.* and *An. coluzzii from Obuasi, Ghana*. The X-axes show Nucleotide diversity, Y-axes shows estimates of Watterson's Theta. *An. coluzzii* samples have been split into two cohorts, based on administrative boundaries.

Figure 5 shows estimates of Nucleotide diversity and Watterson's theta in both species present. Although patterns of Nucleotide diversity were not significantly different with overlapping confidence intervals between species, values of Watterson's theta were significantly lower in *An. coluzzii* than *An. gambiae s.s.* It is not clear why this is the case.

5.3.4 Isolation by distance

Isolation-by-distance (IBD) is an important parameter in population genetics, describing the tendency for organisms closer in space to be more similar to each other, due to geographically limited dispersal (Wright, 1943). In West Africa, earlier studies have found different rates of IBD when comparing *An. gambiae s.s* or *An. coluzzii*, between the two species (Ag1000G, 2020), with *An. coluzzii* displaying stronger isolation-by-distance, suggesting reduced dispersal. Although we could not compare between species due to limited sample sizes in *An. gambiae s.s,* we performed an analysis of isolation-by-distance in *An. coluzzii*, at a much finer scale than has been previously attempted. We calculated Hudson's F_{st} between individuals from each sampling site (Figure 6). F_{st} was generally low, as was expected from the proximity of sampling locations.

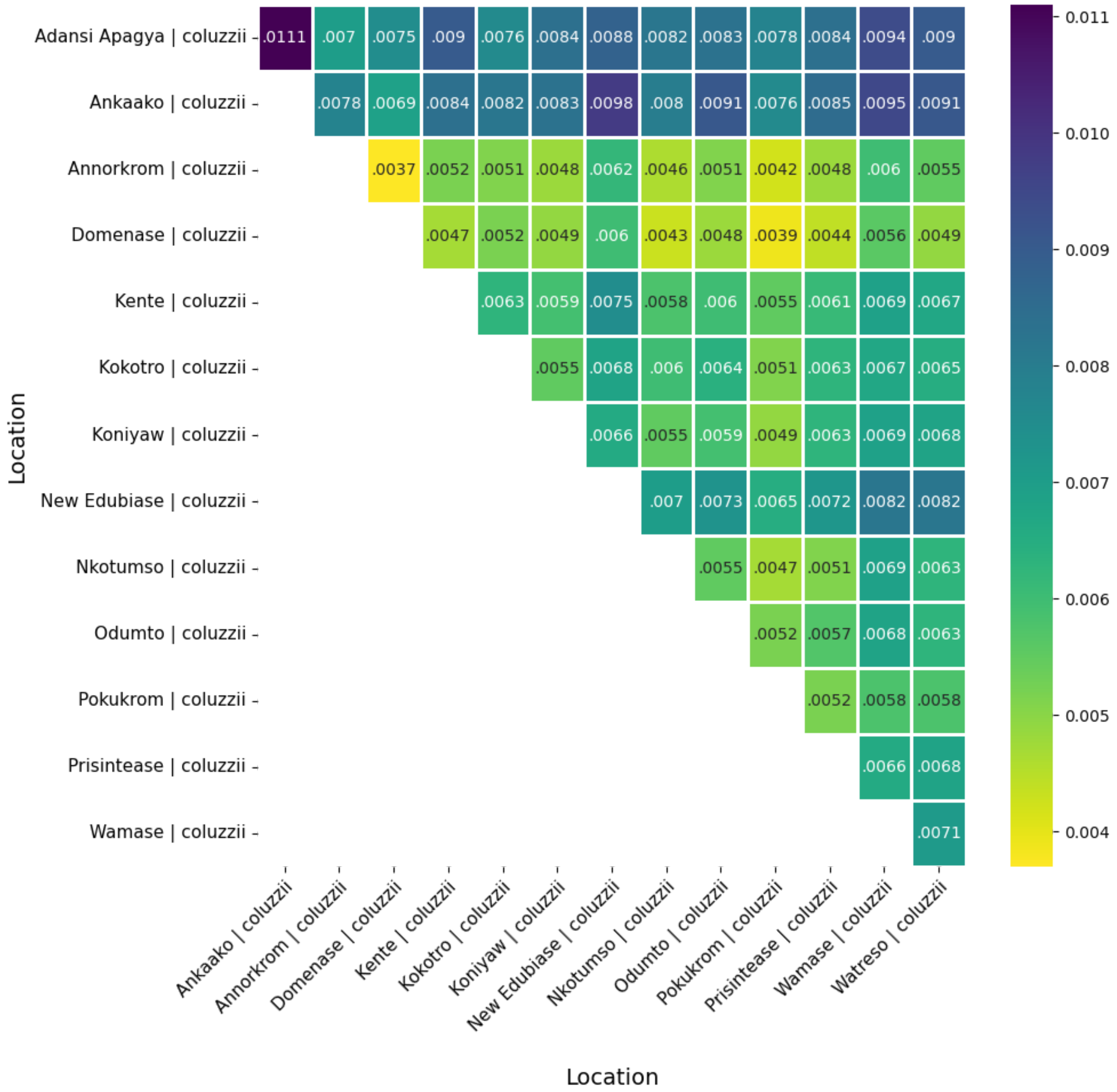


Figure 6. Pairwise Fst between sampling locations of *An. coluzzii.* The one *An. gambiae s.s* sampling site with sufficient numbers was excluded to allow better discrimination in the colour range of *An. coluzzii* study sites. Sites are ordered alphabetically.

Using the pairwise F_{st} estimates between sample sites, we plotted linearised F_{st} against the geographic distance between sites (Figure 7), and fitted a regression line as in the method of Rousset (Rousset, 1997). We initially excluded sampling locations with less than 10 mosquitoes, which resulted in 13 sampling locations in total. We observe a positive and significant slope of the regression line (p=0.0063), suggesting that genetic differentiation does increase with geographic distance in our dataset.

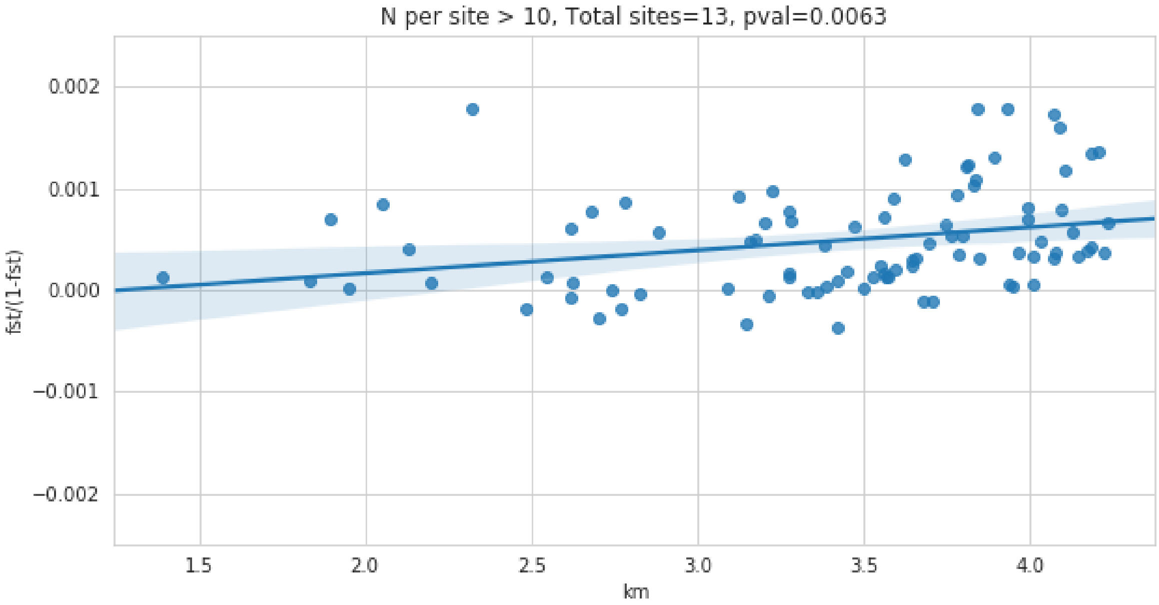


Figure 7. Isolation-by-distance. Genetic distance between each sample site in the form of linearised Hudson's F_{st} (Y-axes) plotted against the log of geographic distance in Kilometres (X-axes). A linear regression line with 95% confidence intervals is displayed. Genetic distance is significantly associated with geographic distance.

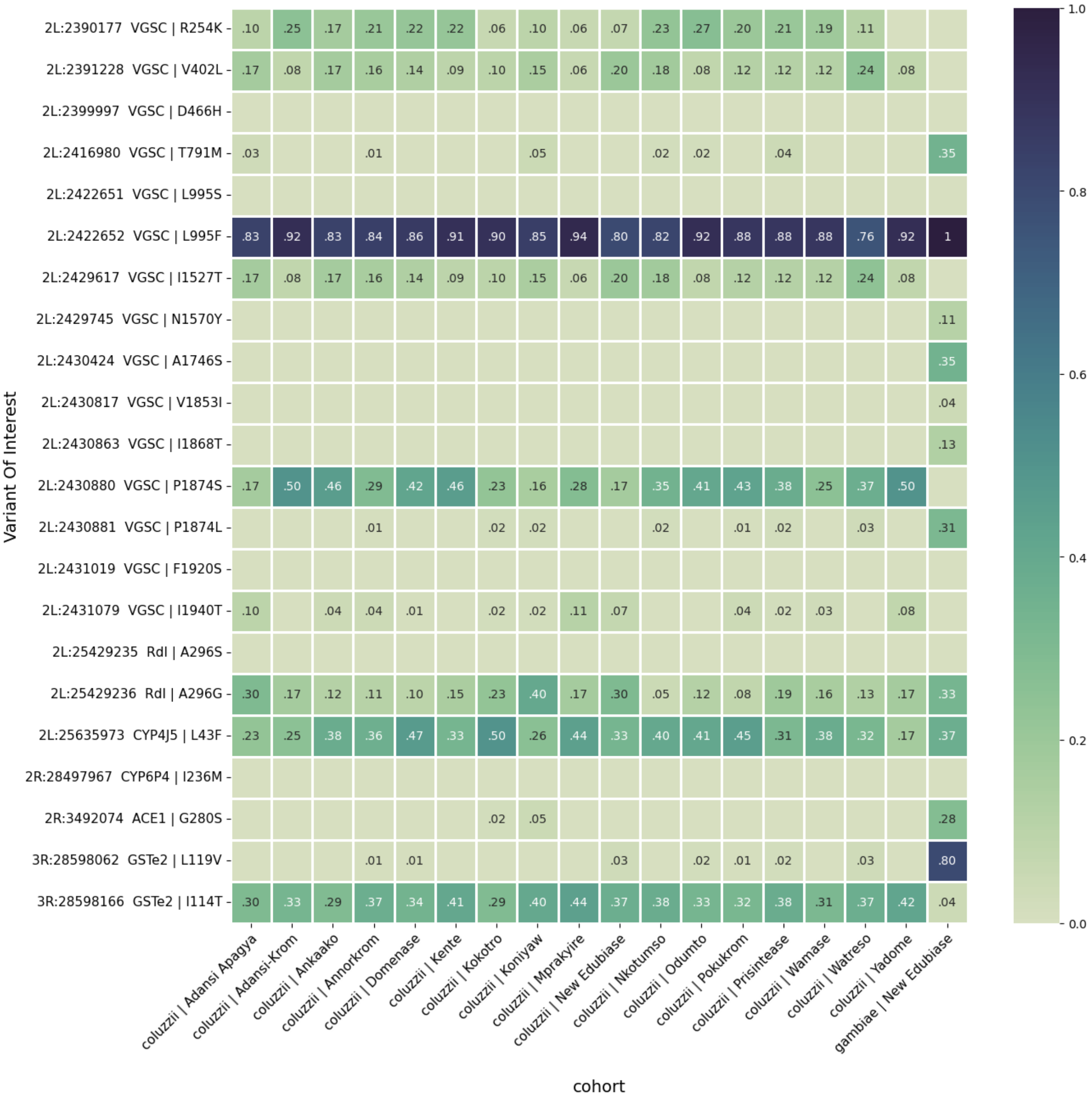
Whilst IBD has been observed between *An. gambiae s.l*, populations before, given the micro-spatial sampling regime which covered only 70km^2, it is novel and surprising to be able to detect isolation by distance within this dataset. To determine the minimum sample size per site when estimating isolation-by-distance, we re-ran the analysis whilst altering the minimum threshold for samples per location to be included. We found that when including sites with less than five mosquitoes, F_{st} estimation was too variable, introducing large amounts of noise into the analysis (Appendix D).

As another measure of population structure, we identified doubleton (f_2) variants, alleles which are only found twice in the dataset and are thought to be indicative of recent mutational events as they have not yet had chance to drift to appreciable frequencies (Mathieson *et al.*, 2014). It was shown by Matheison and McVean that these loci provide a powerful means of detecting fine-scale population structure. As with F_{st}, we found a significant association with the number of shared doubletons between two individuals and with geographic distance (p=4.5e-120). We also find that the individuals with the highest number of shared doubletons also tend to be siblings (Appendix D).

To investigate any effect of environment on isolation-by-distance, we also gathered 13 ecological variables which were used to inform the sampling design (Sedda *et al.*, 2019), (eg. vegetation and land cover, elevation, temperature and precipitation). We removed correlated ecological variables, and calculated an overall "ecological distance" between each sample sites, using euclidean distance. We then performed partial mantel tests on the distance, ecological distance variable and F_{st} matrices, to test for associations between two variables whilst controlling for a third (Manel *et al.*, 2003). When we tested for associations between geographic distance and F_{st} whilst controlling for ecology, the result was significant (p=0.018). However, the reverse was not true - ecological distance did not have a significant effect on genetic distance when controlling for geographic distance (p=0.839), suggesting that geographic distance drives differentiation at this scale, rather than isolation by environment. The distribution of ASM mines to the east of Obuasi precluded any investigation into the impact of ASM gold mining on population structure or insecticide resistance.

5.3.5 The genomics of insecticide resistance

Given the high intensities of insecticide resistance in Ghana (Mugenzi *et al.*, 2022), we then investigated known insecticide resistance alleles and examined signatures of selection across the genome. Figure 8 shows allele frequencies of variants associated with insecticide resistance, in each sampling location that contained more than 10 individuals. Overall, sampling sites show little variation in allele frequencies in *An. coluzzii*, with only the single *An. gambiae s.s* population from New Edubiase clearly distinct.



Variant Of Interest	coluzzii | Adansi Apagya	coluzzii | Adansi-Krom	coluzzii | Ankaako	coluzzii | Annorkrom	coluzzii | Domenase	coluzzii | Kente	coluzzii | Kokotro	coluzzii | Koniyaw	coluzzii | Mprakyire	coluzzii | New Edubiase	coluzzii | Nkotumso	coluzzii | Odumto	coluzzii | Pokukrom	coluzzii | Prisintease	coluzzii | Wamase	coluzzii | Watreso	coluzzii | Yadome	gambiae | New Edubiase
2L:2390177 VGSC | R254K	.10	.25	.17	.21	.22	.22	.06	.10	.06	.07	.23	.27	.20	.21	.19	.11		
2L:2391228 VGSC | V402L	.17	.08	.17	.16	.14	.09	.10	.15	.06	.20	.18	.08	.12	.12	.12	.24	.08	
2L:2399997 VGSC | D466H																		
2L:2416980 VGSC | T791M	.03			.01				.05			.02	.02		.04				.35
2L:2422651 VGSC | L995S																		
2L:2422652 VGSC | L995F	.83	.92	.83	.84	.86	.91	.90	.85	.94	.80	.82	.92	.88	.88	.88	.76	.92	1
2L:2429617 VGSC | I1527T	.17	.08	.17	.16	.14	.09	.10	.15	.06	.20	.18	.08	.12	.12	.12	.24	.08	
2L:2429745 VGSC | N1570Y																		.11
2L:2430424 VGSC | A1746S																		.35
2L:2430817 VGSC | V1853I																		.04
2L:2430863 VGSC | I1868T																		.13
2L:2430880 VGSC | P1874S	.17	.50	.46	.29	.42	.46	.23	.16	.28	.17	.35	.41	.43	.38	.25	.37	.50	
2L:2430881 VGSC | P1874L				.01			.02	.02			.02		.01	.02		.03		.31
2L:2431019 VGSC | F1920S																		
2L:2431079 VGSC | I1940T	.10		.04	.04	.01		.02	.02	.11	.07			.04	.02	.03		.08	
2L:25429235 Rdl | A296S																		
2L:25429236 Rdl | A296G	.30	.17	.12	.11	.10	.15	.23	.40	.17	.30	.05	.12	.08	.19	.16	.13	.17	.33
2L:25635973 CYP4J5 | L43F	.23	.25	.38	.36	.47	.33	.50	.26	.44	.33	.40	.41	.45	.31	.38	.32	.17	.37
2R:28497967 CYP6P4 | I236M																		
2R:3492074 ACE1 | G280S							.02	.05										.28
3R:28598062 GSTe2 | L119V				.01	.01					.03		.02	.01	.02		.03		.80
3R:28598166 GSTe2 | I114T	.30	.33	.29	.37	.34	.41	.29	.40	.44	.37	.38	.33	.32	.38	.31	.37	.42	.04

Figure 8. Allele frequencies of variants associated with insecticide resistance by location. Variants were selected from the literature. Only locations with more than 10 individuals are displayed. Blank cells indicate a frequency of zero.

In *An. coluzzii*, the *Vgsc*-995F mutation remains at high frequencies ranging between 76% and 94%. At each site, the remaining non-L995F alleles correspond to the V402L and I1527T haplotype. Recent data has shown that the 1527T haplotype is increasing in frequency and geographic range (Clarkson *et al.*, 2021; Williams, Cowlishaw, *et al.*, 2022), with a recent transcriptomic study recording the haplotype at frequencies exceeding 50% as far east as Chad and Niger (Ibrahim *et al.*, 2022). Suggesting that the haplotype first found in Guinea, Burkina Faso, and Ghana, has already spread across the range of *An.*

coluzzii. *An. gambiae s.s* remains fixed for L995F haplotypes (Clarkson *et al*., 2021). Between *An. gambiae s.s* and *An. coluzzii*, we can also observe distinct patterns of the secondary mutations occurring with L995F. In *An. coluzzii*, R254K, T791M, P1874S and I1940T are all found at low to moderate frequencies. These alleles are known to have arisen on the background of L995F haplotypes and may increase the resistance level or reduce fitness costs associated with resistance. In *An. gambiae s.s,* however, the secondary mutations are T791M, N1570Y, A1746S, V1853I, I1868T, and P1874L. The *Rdl*-A296G allele, which confers resistance to dieldrin, is also found at every sample site. *Ace-1-G280S* is found at very low rates in *An. coluzzii*, with a higher frequency in the *An. gambiae s.s* cohort from New Edubiase, in agreement with previous work in Ghana (Grau-Bové *et al*., 2021). We also detect the *Gste2-114T* mutation at moderate frequencies in *An. coluzzii and Gste2-119V* at high frequencies in *An. gambiae s.s*.

We also calculated frequencies of amplifications or deletions (copy number variants) over insecticide resistance genes (Figure 9). In general, we observe high frequencies of *Cyp9K1* CNVs in *An. gambiae s.s* but low frequencies of Cyp6aa/P CNVs. The opposite pattern is observed in *An. coluzzii*. Interestingly, in *An. gambiae s.s* we observe a much higher prevalence of *Ace1* amplifications (60%) than the *Ace1-280S* mutation (28%). It is known that CNVs pair wild-type with resistance 280S alleles, and this difference in G280S frequency may highlight how genotyping with a diploid model can become inaccurate as copy number increases.

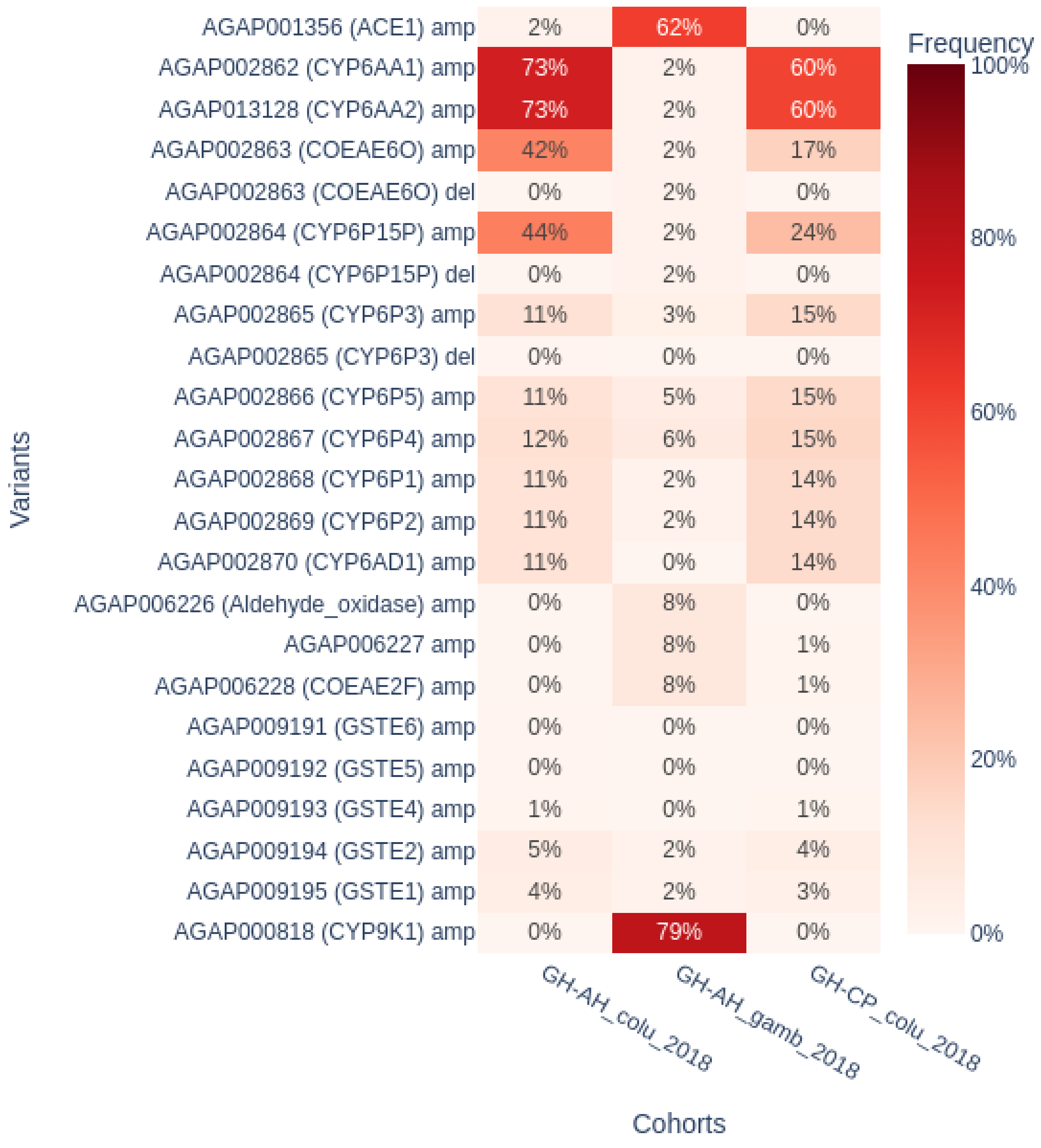


Figure 9. CNV frequencies in the Obuasi cohorts at major insecticide resistance loci. amp=amplification, del=deletion. CNVs were detected using a HMM on read depth (coverage) in 300 bp windows (Lucas, et al., 2019).

We performed haplotype-based genome-wide selection scans in both *An. gambiae s.s* and *An. coluzzii* separately. We used the H12 statistic (Garud, Messer, *et al.*, 2015; Garud and Rosenberg, 2015), and applied it in 1000bp stepping windows across each chromosome. In Figure 10, H12 values almost reaching 1 can be seen at the *Vgsc* (3 Mb) on 2L in both species, with *An. gambiae s.s* also harbouring signals at *Rdl* and *Coeae1f*. In both species, we observe signals at the CYP6aa/P region (28.5 Mb) and at KEAP1 (40.5 Mb) on the 2R chromosomal arm. On the 3R chromosomal arm, the *Gste2* locus has high values of H12 in both species. This is also true for the *Cyp9K1* locus on the X chromosome (15 Mb), however, CNVs are only at high frequency in *An. gambiae s.s* at *Cyp9K1*, suggesting that other mechanisms must be driving selective sweeps at this locus in *An. coluzzii*.

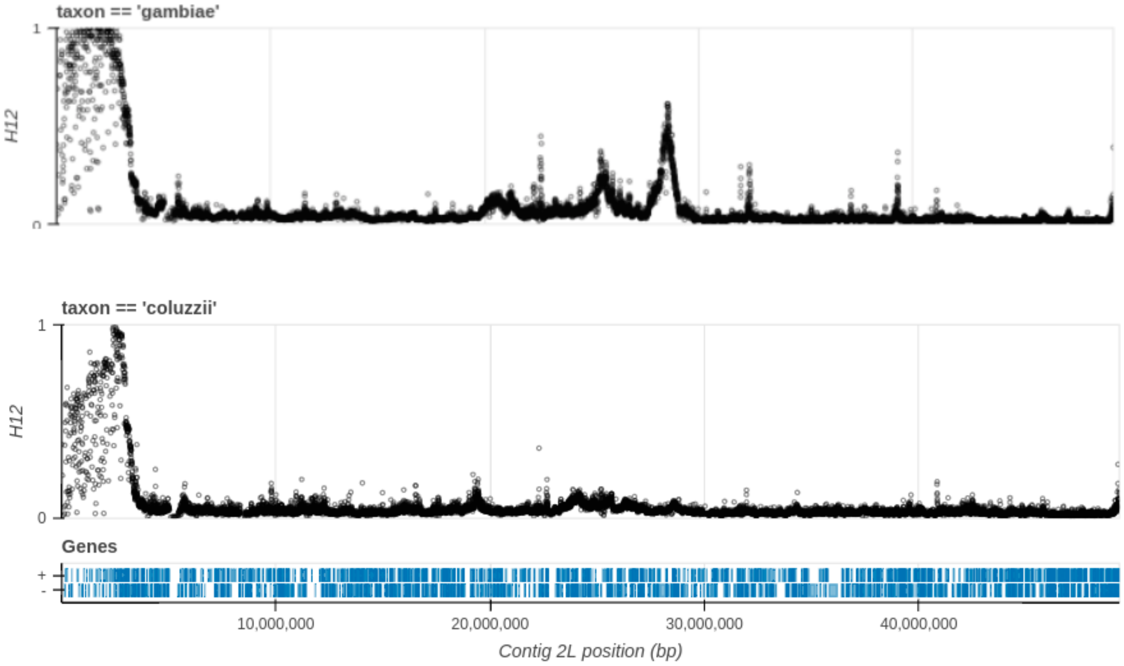


Figure 10. H12 genome-wide selection scan on *An. gambiae* and *An. coluzzii* 2L chromosomal arm in 1000 bp steppings windows. H12 captures the frequency of the two most frequent haplotypes under selection.

A complementary approach using an F_{st} genome-wide selection scan using a window size of 2000 SNPs (Appendix H) replicated the signals at the Gste2 locus at 28.5 MB on 3R, and at the CYP9K1 locus (Chromosome X: 15 Mb).

In the *An. gambiae* species complex, adaptive gene flow is well-documented with insecticide resistance alleles passing between species (Clarkson *et al.*, 2014; Grau-Bové *et al.*, 2020; 2021). To investigate adaptive gene flow between *An. gambiae s.s* and *An. coluzzii,* we performed genome-wide scans with the H1X statistic (Miles, 2021). This statistic calculates the probability that any two haplotypes sampled between two populations are identical, and is useful to identify haplotypes that have introgressed between species and that are also at moderate to high frequencies. We calculate H1X over each chromosomal arm (Appendix I). Figure 11 shows data for the 2L chromosomal arm where most signals were found.

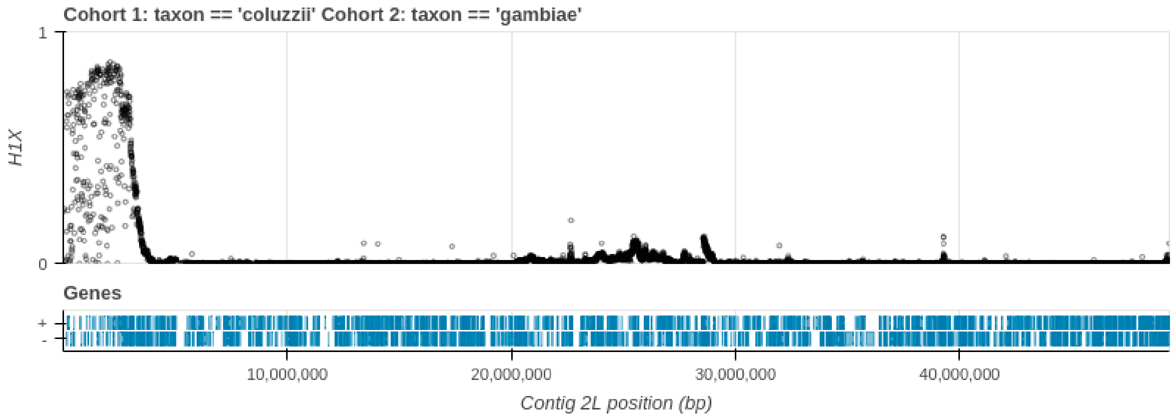


Figure 11. H1X scan across the 2L chromosomal arm, comparing *An. gambiae* and *An. coluzzii*. H1X was calculated in 1000bp stepping windows. The H1X statistic reflects the probability that two haplotypes sampled at random between two populations will be identical.

A large H1X signal can be noted near the centromere of the 2L chromosomal arm. The target of pyrethroid insecticides, the *Vgsc*, is located in this region. It has been shown previously in populations from Ghana that a resistant haplotype introgressed from *An. gambiae s.s* to *An. coluzzii*, and subsequently spread to high frequencies (Clarkson *et al.*, 2014), so this result is expected. We can also note some smaller signals of introgression around 25 Mb (*Rdl* GABA-gated chloride receptor). Again, this is a site of known introgression events (Grau-Bové *et al.*, 2020), with two haplotypes arising via hard selective sweeps and spreading from *An. gambiae s.s* and *An. arabiensis* to *An. coluzzii*. Lastly, a small signal can be noted at 28.5 Mb - this is the site of the *Coeae1f/2f* genes, the focus of chapter 4, in which I also demonstrate haplotype sharing between *An. gambiae s.s* and *An. coluzzii*. A small H1X signal was also noted at approximately 9 Mb into the X chromosome (Appendix H4). No other H1X signals were observed on other chromosomal arms.

Although genome-wide selection scans are useful to detect genomic regions under selection, they generally give little to no information on the nature of selection at that locus. For example, we may not know how many independent selective sweeps are present, what the haplotype frequency distribution looks like, or if any of those sweeps have spread between species or other metadata. To explore these questions, we can focus specifically on the locus, and hierarchically cluster haplotypes to identify groups of highly similar haplotypes. When large groups of haplotypes are highly similar, this is likely indicative of a selective sweep - all haplotypes in a cluster contain a beneficial mutation which has caused the haplotype to spread. To further explore the number and breadth of selective sweeps at known insecticide resistance loci, I implemented an interactive haplotype clustering method in the *malariagen_data* python package, and used it to perform hierarchical clustering on the Obuasi haplotype data at major resistance loci. In Figures 12 to 15, dendrograms are shown which display the distance in number of SNPs (Y-axis) between individual haplotypes (X-axis). Dendrogram leaves are coloured by the species assignment of the individual that bears the haplotype.

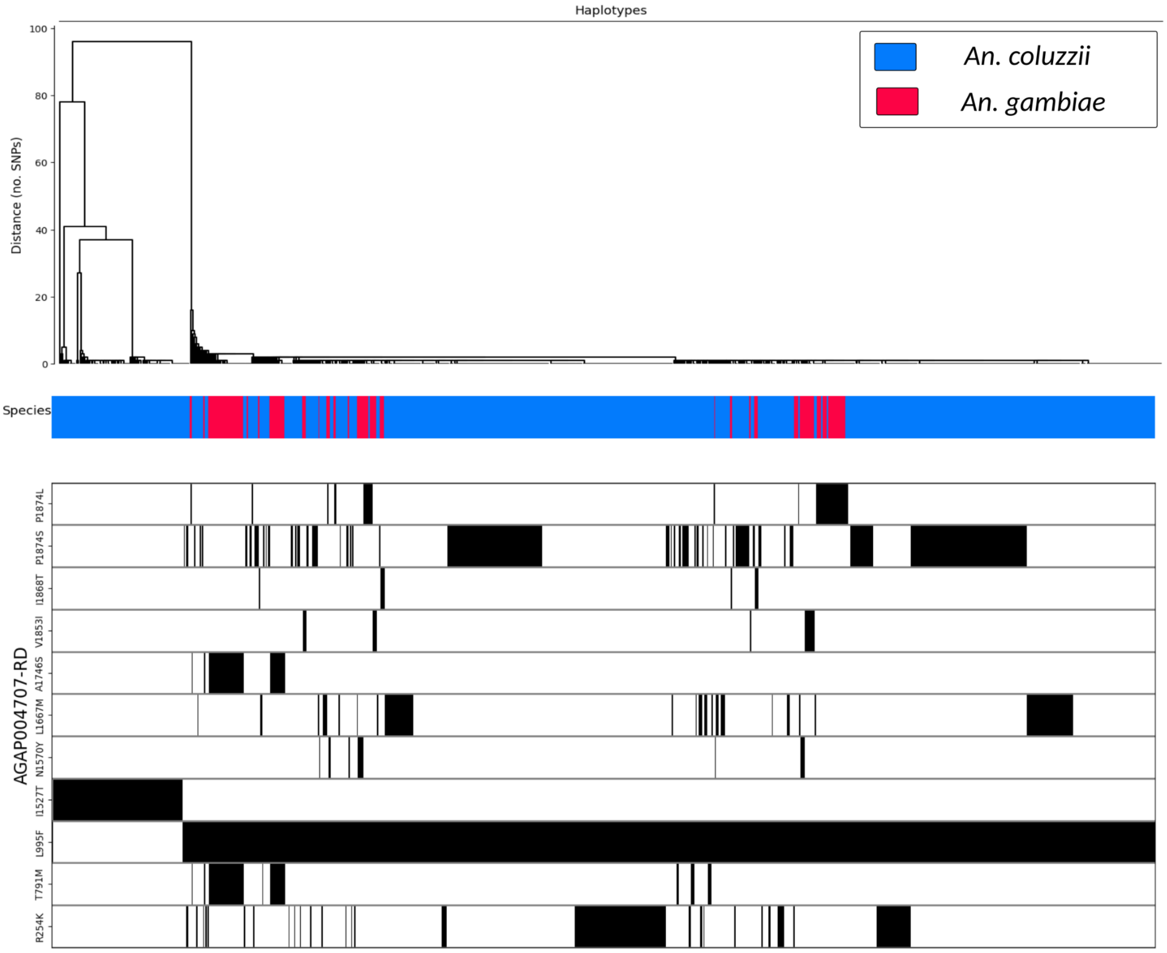


Figure 12. Hierarchical haplotype clustering at the Voltage-gated sodium channel *(Vgsc)*. Biallelic, non-synonymous variants with a frequency greater than 5% are shown on each haplotype. Hamming distance and single linkage were used for hierarchical clustering. The figure was generated with the locusPocus workflow.

Haplotype clustering at the *Vgsc* (Figure 12), results in two overall clusters, with no wild-type haplotypes outside of these two clusters. The smaller cluster contains only individuals assigned as *An. coluzzii*, whereas the larger cluster contains a mixture of *An. coluzzii* and *An. gambiae s.s* haplotypes. Biallelic non-synonymous variants are shown below the dendrogram leaves. The small *An. coluzzii* corresponds to the 402L/1527T haplotype (402L not shown as it is a multiallelic site), and the large mixed species cluster corresponds to 995F-bearing haplotypes. A plethora of secondary mutations can be seen to have arisen on the 995F haplotypic background.

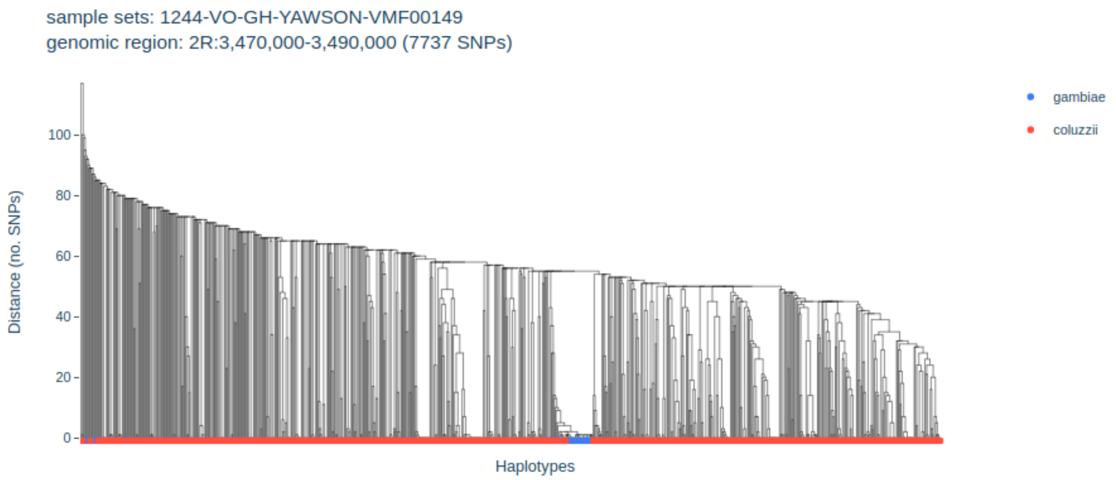


Figure 13. Haplotype clustering at *Ace-1*. The figure was generated with malariagen_data. Haplotype leaves are coloured by taxon. Hierarchical clustering was performed with single linkage and hamming distance.

At *Ace-1*, we observe mostly wild-type haplotypes, as evident from the large distance in SNPs between most pairwise haplotypes (Figure 13). There are however, at least two small clusters of haplotypes which appear to be under selection, and may correspond to the *Ace1-G280S* mutation, which confers resistance to organophosphates (Grau-Bové *et al.*, 2021).

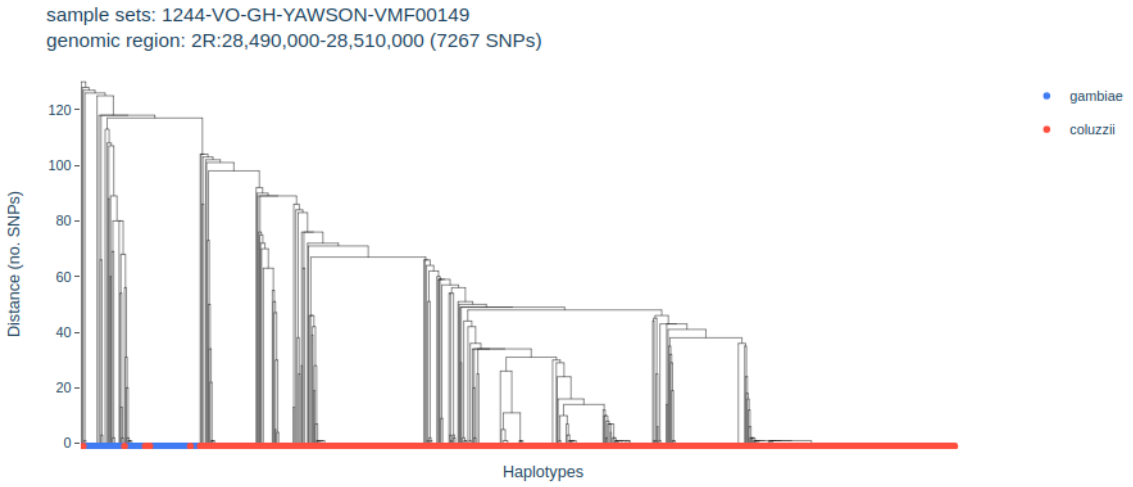


Figure 14. Haplotype clustering at the CYP6aa/P locus. The figure was generated with malariagen_data. Haplotype leaves are coloured by taxon. Hierarchical clustering was performed with single linkage and hamming distance.

At the Cyp6aa/P locus, we can clearly observe multiple independent selective sweeps (Figure 14). One of these sweeps is shared between *An. gambiae s.s* and *An. coluzzii*. The number of independent sweeps at this locus is a clear illustrations of the complexity of insecticide resistance. In the H12 scans for *An. coluzzii*, the signal at the Cyp6aa/P locus is relatively small (Appendix F1). The haplotype clustering dendrogram for this locus then highlights a scenario in which H12 actually has relatively low power. We can observe that most haplotypes are part of a cluster, i.e most haplotypes are likely to be under selection. However, the H12 statistic only emphasises the two most frequent haplotypes, meaning if there is a 3rd or 4th (and so on) haplotype cluster of appreciable frequency, it will not contribute to the H12 value. An ideal selection statistic would capture information across the whole haplotype frequency spectrum.

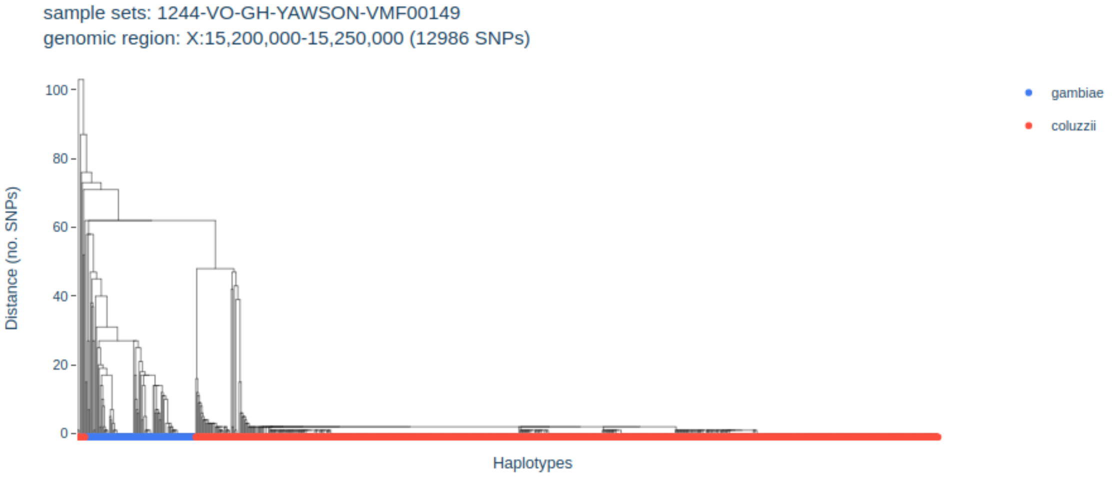


Figure 15. Haplotype clustering at CYP9K1. The figure was generated with malariagen_data. Haplotype leaves are coloured by taxon. Hierarchical clustering was performed with single linkage and hamming distance.

We also explore clustering at the CYP9K1 locus on the X chromosome (Figure 15). In accordance with the extremely high genetic differentiation seen at 15 Mb in F_{st} plots of the X chromosome (Appendix C), we find that both *An. gambiae s.s* and *An. coluzzii* harbour independent selective sweeps at this locus, which are not shared. The lack of haplotype sharing at this locus could be related to the fact that the X may be more resistant to introgression than the autosomes (Fontaine *et al.*, 2015).

5.4 Conclusion

In this study, we investigated population structure and the genomics of insecticide resistance in populations of *An. gambiae s.s* and *An. coluzzii* from Obuasi, Ghana. We were able to detect fine-scale isolation-by-distance in *An. coluzzii*, which was primarily driven by physical distance at this scale rather than the environment. We observe continued evolution of the *Voltage-gated sodium channel*, with secondary mutations spreading on the background of the 995F allele, and the 402L/1527T haplotype spreading in *An. coluzzii*. We find contrasting patterns of copy number variation in the *An. gambiae* and *An. coluzzii* at the CYP6aa and CYP9K1 locus. We found evidence of a high prevalence of haplotypes under selection, with the *Cyp6aa/P* locus in particularly displaying an extremely high number of haplotypes seemingly under selection. We find further evidence of haplotype sharing between IR loci between *An. gambiae* s.s and *An. coluzzii.*

Ideally, it might be more informative to see the context of these analyses against the wider Ghanaian and West African regions, however, given the known lack of population structure, it is unclear what extra information this would add to the already published studies (Miles *et al.*, 2017; Ag1000G, 2020). It will also be necessary to omit the 2La inversion from the relatedness calculations. This work is ongoing. it will be important to assess how well we can determine close kin in *Anopheles* mosquitoes for future research. In chapter 6, we discuss ways in which the KING statistic could be evaluated.

5.4.1 Probe

Many of the analyses conducted here may be conducted using a snakemake workflow, *Probe*, both directly on Ag1000G data via integration of *malariagen_data*, or on a user-provided VCF file, meaning that other species may be analysed. I recently used the workflow to calculate relatedness for a genome-wide association study (Lucas *et al.*, 2023). The workflow is located at https://github.com/sanjaynagi/probe. The analyses include calculating a number of relatedness statistics with NgsRelate, principal components analysis, and Garud's G and H selection statistics. This workflow is no longer being developed or maintained, however, as I believe for population genomic analyses such as in this chapter, a better approach is to use interactive notebooks (particularly as it helps users learn). The workflow is still useful for calculating relatedness, however, as this requires complex writing of Variant Call Format (VCF) from Zarr files, and command line tools which are cumbersome to run in a notebook environment.

5.5 Methods

5.5.1 Sample collection and sequencing

Samples were collected in Obuasi district with CDC light traps, using an ecologically-informed sampling framework (Sedda *et al.*, 2019). Collections took place from October to December 2018. Mosquitoes were stored in ethanol prior to Illumina 150bp paired-end whole-genome sequencing. Sequencing was performed to a target coverage of 30X and bioinformatic analysis was performed as described previously using a GATK-based workflow (Ag1000G, 2020). Copy number variants were called as described elsewhere (Lucas *et al.*, 2019).

5.5.2 Population genetic analysis

All population genomic analyses were performed in python with scikit-allel version 1.2.1 and malariagen_data version 7.0.1 unless explicitly stated otherwise. Only fourfold degenerate SNP sites were used to calculate genetic diversity metrics. PCAs were performed on chromosome 3L, using markers between 15 MB and 44 Mb, to avoid regions of low recombination and known chromosomal inversions. NgsRelate (Korneliussen *et al.*, 2015) was used to calculate kinship on biallelic genotypes from across the whole-genome. Linear regression was used to test for isolation-by-distance through F_{st} and the number of shared doubletons. H12, H1X scans and haplotype clustering were performed on phased haplotype data. H12 and H1X were both calculated in 1000bp stepping windows, whereas windows for haplotype clustering varied depending on the locus. For haplotype clustering, the Scipy implementation of hierarchical clustering was performed, with single linkage and hamming distance converted to the total number of SNP differences as the distance metric between clusters.

5.5.3 Code availability

The probe workflow is located here github.com/sanjaynagi/probe. All other code for this chapter are stored here https://github.com/sanjaynagi/PhD_thesis_code/obuasi.

5.6 References

Ag1000G (2020) ‘Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii’, *Genome Res*, pp. 1–14.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002) *Meiosis*. Garland Science.

Bravington, M.V., Skaug, H.J. and Anderson, E.C. (2016) ‘Close-Kin Mark-Recapture’, *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31(2), pp. 259–274.

Clarkson, C.S., Miles, A., Harding, N.J., O’Reilly, A.O., Weetman, D., Kwiatkowski, D. and Donnelly, M.J. (2021) ‘The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors Anopheles gambiae and Anopheles coluzzii’, *Molecular ecology*, 30(21), pp. 5303–5317.

Clarkson, C.S., Weetman, D., Essandoh, J., Yawson, A.E., Maslen, G., Manske, M., Field, S.G., Webster, M., Antão, T., MacInnis, B., Kwiatkowski, D. and Donnelly, M.J. (2014) ‘Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation’, *Nature communications*, 5(1), pp. 1–10.

Dhole, S., Lloyd, A.L. and Gould, F. (2020) ‘Gene Drive Dynamics in Natural Populations: The Importance of Density Dependence, Space, and Sex’, *Annual review of ecology, evolution, and systematics*, 51(1), pp. 505–531.

Epstein, A., Maiteki-Sebuguzi, C., Namuganga, J.F., Nankabirwa, J.I., Gonahasa, S., Opigo, J., Staedke, S.G., Rutazaana, D., Arinaitwe, E., Kamya, M.R., Bhatt, S., Rodríguez-Barraquer, I., Greenhouse, B., Donnelly, M.J. and Dorsey, G. (2022) ‘Resurgence of malaria in Uganda despite sustained indoor residual spraying and repeated long lasting insecticidal net distributions’, *PLOS Global Public Health*, 2(9), p. e0000676.

Ferring, D. and Hausermann, H. (2019) ‘The Political Ecology of Landscape Change, Malaria, and Cumulative Vulnerability in Central Ghana’s Gold Mining Country’, *Annals of the Association of American Geographers. Association of American Geographers*, 109(4), pp. 1074–1091.

Filipović, I., Hapuarachchi, H.C., Tien, W.-P., Razak, M.A.B.A., Lee, C., Tan, C.H., Devine, G.J. and Rašić, G. (2020) ‘Using spatial genetics to quantify mosquito dispersal for control programs’, *BMC biology*, 18(1), p. 104.

Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V., Jiang, X., Hall, A.B., Catteruccia, F., Kakani, E., Mitchell, S.N., Wu, Y.-C., Smith, H.A., Love, R.R., Lawniczak, M.K.N., *et al.* (2015) ‘Extensive introgression in a malaria vector species complex revealed by phylogenomics’, *Science*, 347(6217), p. 1258522.

Fougerouse, D., Micklethwaite, S., Ulrich, S., Miller, J., Godel, B., Adams, D.T. and Campbell McCuaig, T. (2017) ‘Evidence for Two Stages of Mineralization in West Africa’s Largest Gold Deposit: Obuasi, Ghana', *Economic geology and the bulletin of the Society of Economic Geologists*, 112(1), pp. 3–22.

Garud, N.R., Messer, P.W., Buzbas, E.O. and Petrov, D.A. (2015) 'Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps', *PLoS genetics*, 11(2), pp. 1–32.

Garud, N.R. and Rosenberg, N.A. (2015) 'Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps', *Theoretical population biology*, 102, pp. 94–101.

Grau-Bové, X., Lucas, E., Pipini, D., Rippon, E., van 't Hof, A.E., Constant, E., Dadzie, S., Egyir-Yawson, A., Essandoh, J., Chabi, J., Djogbénou, L., Harding, N.J., Miles, A., Kwiatkowski, D., Donnelly, M.J., *et al.* (2021) 'Resistance to pirimiphos-methyl in West African Anopheles is spreading via duplication and introgression of the Ace1 locus', *PLoS genetics*, 17(1), p. e1009253.

Grau-Bové, X., Tomlinson, S., O'Reilly, A.O., Harding, N.J., Miles, A., Kwiatkowski, D., Donnelly, M.J., Weetman, D. and Anopheles gambiae 1000 Genomes Consortium (2020) 'Evolution of the Insecticide Target Rdl in African Anopheles Is Driven by Interspecific and Interkaryotypic Introgression', *Molecular biology and evolution*, 37(10), pp. 2900–2917.

Hamid-Adiamoh, M., Amambua-Ngwa, A., Nwakanma, D., D'Alessandro, U., Awandare, G.A. and Afrane, Y.A. (2020) 'Insecticide resistance in indoor and outdoor-resting Anopheles gambiae in Northern Ghana', *Malaria journal*, 19(1), p. 314.

Hawkins, N.J., Bass, C., Dixon, A. and Neve, P. (2019) 'The evolutionary origins of pesticide resistance', *Biological reviews of the Cambridge Philosophical Society*, 94(1), pp. 135–155.

Hoffman, G.E. (2013) 'Correcting for population structure and kinship using the linear mixed model: theory and extensions', *PloS one*, 8(10), p. e75707.

Ibrahim, S.S., Muhammad, A., Hearn, J., Weedall, G.D., Nagi, S.C., Mukhtar, M.M., Fadel, A.N., Mugenzi, L.J., Patterson, E.I., Irving, H. and Wondji, C.S. (2022) 'Molecular drivers of insecticide resistance in the Sahelo-Sudanian populations of a major malaria vector', *bioRxiv*. doi:10.1101/2022.03.21.485146.

Kafy, H.T., Ismail, B.A., Mnzava, A.P., Lines, J., Abdin, M.S.E., Eltaher, J.S., Banaga, A.O., West, P., Bradley, J., Cook, J., Thomas, B., Subramaniam, K., Hemingway, J., Knox, T.B., Malik, E.M., *et al.* (2017) 'Impact of insecticide resistance in *Anopheles arabiensis* on malaria incidence and prevalence in Sudan and the costs of mitigation', *Proceedings of the National Academy of Sciences*, p. 201713814.

Korneliussen, T.S. and Moltke, I. (2015) 'NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data', *Bioinformatics* , 31(24), pp. 4009–4011.

Kyrou, K., Hammond, A.M., Galizi, R., Kranjc, N., Burt, A., Beaghton, A.K., Nolan, T. and Crisanti, A. (2018) 'A CRISPR–Cas9 gene drive targeting doublesex causes complete population suppression in caged Anopheles gambiae mosquitoes', *Nature biotechnology*, 36(11), pp. 1062–1066.

Labbé, P., Berticat, C., Berthomieu, A., Unal, S., Bernard, C., Weill, M. and Lenormand, T. (2007) 'Forty years of erratic insecticide resistance evolution in the mosquito Culex pipiens', *PLoS genetics*, 3(11), p. e205.

Love, R.R., Redmond, S.N., Pombi, M., Caputo, B., Petrarca, V., della Torre, A. and Besansky, N.J. (2019) 'In Silico Karyotyping of Chromosomally Polymorphic Malaria Mosquitoes in the Anopheles gambiae Complex', *G3: Genes, Genomes, Genetics*, 9(10), pp. 3249–3262.

Lucas, E.R., Miles, A., Harding, N.J., Clarkson, C.S., Lawniczak, M.K.N., Kwiatkowski, D.P., Weetman, D., Donnelly, M.J. and Anopheles gambiae 1000 Genomes Consortium (2019) ‘Whole-genome sequencing reveals high complexity of copy number variation at insecticide resistance loci in malaria mosquitoes’, *Genome research*, 29(8), pp. 1250–1261.

Lucas, E.R., Nagi, S.C., Egyir-Yawson, A., Essandoh, J., Dadzie, S., Chabi, J., Djogbenou, L.S., Medjigbodo, A.A., Edi, C.V., Ketoh, G.K., Koudou, B.G., Van’t Hof, A.E., Rippon, E.J., Pipini, D., Harding, N.J., *et al.* (2023) ‘Genome-wide association studies reveal novel loci associated with pyrethroid and organophosphate resistance in Anopheles gambiae s.l’, *bioRxiv*. doi:10.1101/2023.01.13.523889.

Manel, S., Schwartz, M.K., Luikart, G. and Taberlet, P. (2003) ‘Landscape genetics: combining landscape ecology and population genetics’, *Trends in ecology & evolution*, 18(4), pp. 189–197.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. and Chen, W.-M. (2010) ‘Robust relationship inference in genome-wide association studies’, *Bioinformatics* , 26(22), pp. 2867–2873.

Mathieson, I. and McVean, G. (2014) ‘Demography and the Age of Rare Variants’, *PLoS genetics*, 10(8). doi:10.1371/journal.pgen.1004528.

Miles, A. (2021) *Genomic epidemiology of malaria vectors in the Anopheles gambiae species complex*.

Miles, A., Harding, N.J., Bottà, G., Clarkson, C.S., Antão, T., Kozak, K., Schrider, D.R., Kern, A.D., Redmond, S., Sharakhov, I., Pearson, R.D., Bergey, C., Fontaine, M.C., Donnelly, M.J., Lawniczak, M.K.N., *et al.* (2017) ‘Genetic diversity of the African malaria vector Anopheles gambiae’, *Nature*, 552, pp. 96–100.

Mugenzi, L.M.J., Akosah-Brempong, G., Tchouakui, M., Menze, B.D., Tekoh, T.A., Tchoupo, M., Nkemngo, F.N., Wondji, M.J., Nwaefuna, E.K., Osae, M. and Wondji, C.S. (2022) ‘Escalating pyrethroid resistance in two major malaria vectors Anopheles funestus and Anopheles gambiae (s.l.) in Atatam, Southern Ghana’, *BMC infectious diseases*, 22(1), p. 799.

Oumbouke, W.A., Pignatelli, P., Barreaux, A.M.G., Tia, I.Z., Koffi, A.A., Ahoua Alou, L.P., Sternberg, E.D., Thomas, M.B., Weetman, D. and N’Guessan, R. (2020) ‘Fine scale spatial investigation of multiple insecticide resistance and underlying target-site and metabolic mechanisms in Anopheles gambiae in central Côte d’Ivoire’, *Scientific reports*, 10(1), p. 15066.

Perna, N.T. and Kocher, T.D. (1995) ‘Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes’, *Journal of molecular evolution*, 41(3), pp. 353–358.

Pwalia, R., Joannides, J., Iddrisu, A., Addae, C., Acquah-Baidoo, D., Obuobi, D., Amlalo, G., Akporh, S., Gbagba, S., Dadzie, S.K., Athinya, D.K., Hadi, M.P., Jamet, H.P. and Chabi, J. (2019) ‘High insecticide resistance intensity of Anopheles gambiae (s.l.) and low efficacy of pyrethroid LLINs in Accra, Ghana’, *Parasites & vectors*, 12(1), p. 299.

Rousset, F. (1997) ‘Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance’, *Genetics*, 145(4), pp. 1219–1228.

Schmidt, T.L., Elfekih, S., Cao, L.-J., Wei, S.-J., Al-Fageeh, M.B., Nassar, M., Al-Malik, A. and Hoffmann, A.A. (2022) ‘Close kin dyads indicate intergenerational dispersal and barriers’, *bioRxiv*. doi:10.1101/2022.01.18.476819.

Sedda, L., Lucas, E.R., Djogbénou, L.S., Edi, A.V.C., Egyir-Yawson, A., Kabula, B.I., Midega, J., Ochomo, E., Weetman, D. and Donnelly, M.J. (2019) ‘Improved spatial ecological sampling using open data and standardization: an example from malaria mosquito surveillance', *Journal of the Royal Society, Interface / the Royal Society*, 16(153), p. 20180941.

Stresman, G., Bousema, T. and Cook, J. (2019) 'Malaria Hotspots: Is There Epidemiological Evidence for Fine-Scale Spatial Targeting of Interventions?', *Trends in parasitology*, 35(10), pp. 822–834.

Tepa, A., Kengne-Ouafo, J.A., Djova, V.S., Tchouakui, M., Mugenzi, L.M.J., Djouaka, R., Pieme, C.A. and Wondji, C.S. (2022) 'Molecular Drivers of Multiple and Elevated Resistance to Insecticides in a Population of the Malaria Vector Anopheles gambiae in Agriculture Hotspot of West Cameroon', *Genes*, 13(7). doi:10.3390/genes13071206.

WHO (2012) 'Global plan for insecticide resistance management in malaria vectors', *World Health Organization press*, p. 13.

Williams, J., Cowlishaw, R., Sanou, A., Ranson, H. and Grigoraki, L. (2022) '*In vivo* functional validation of the V402L voltage gated sodium channel mutation in the malaria vector *An. gambiae*', *Pest Management Science*, pp. 1155–1163. doi:10.1002/ps.6731.

Williams, J., Ingham, V.A., Morris, M., Toé, K.H., Hien, A.S., Morgan, J.C., Dabiré, R.K., Guelbéogo, W.M., Sagnon, N. 'falé and Ranson, H. (2022) 'Sympatric Populations of the Anopheles gambiae Complex in Southwest Burkina Faso Evolve Multiple Diverse Resistance Mechanisms in Response to Intense Selection Pressure with Pyrethroids', *Insects*, 13(3). doi:10.3390/insects13030247.

Wood, R.J. and Cook, L.M. (1983) 'A note on estimating selection pressures on insecticide-resistance genes', *Bulletin of the World Health Organization*, 61(1), pp. 129–134.

Wright, S. (1943) 'ISOLATION BY DISTANCE', *Genetics*, 28(2), pp. 114–138.

Zheng, L., Benedict, M.Q., Cornel, A.J., Collins, F.H. and Kafatos, F.C. (1996) 'An integrated genetic map of the African human malaria vector mosquito, Anopheles gambiae', *Genetics*, 143(2), pp. 941–952.

5.6 Appendix

5.6.1 Appendix A

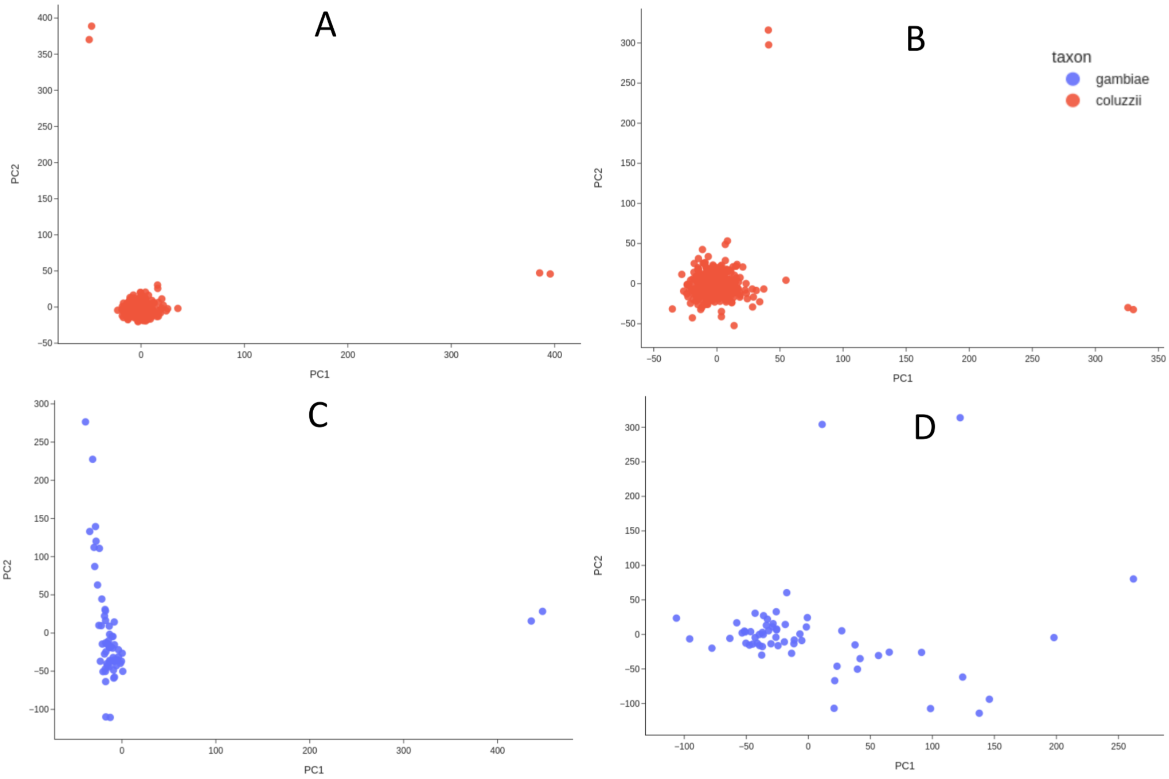


Figure A1. Principal components analysis on chromosomal arm 3L, 15 Mb - 44 Mb. A) *An. coluzzii*, B) *An. coluzzii* after removal of outliers C) *An. gambiae* D) *An. gambiae* after removal of outliers.

5.6.2 Appendix B

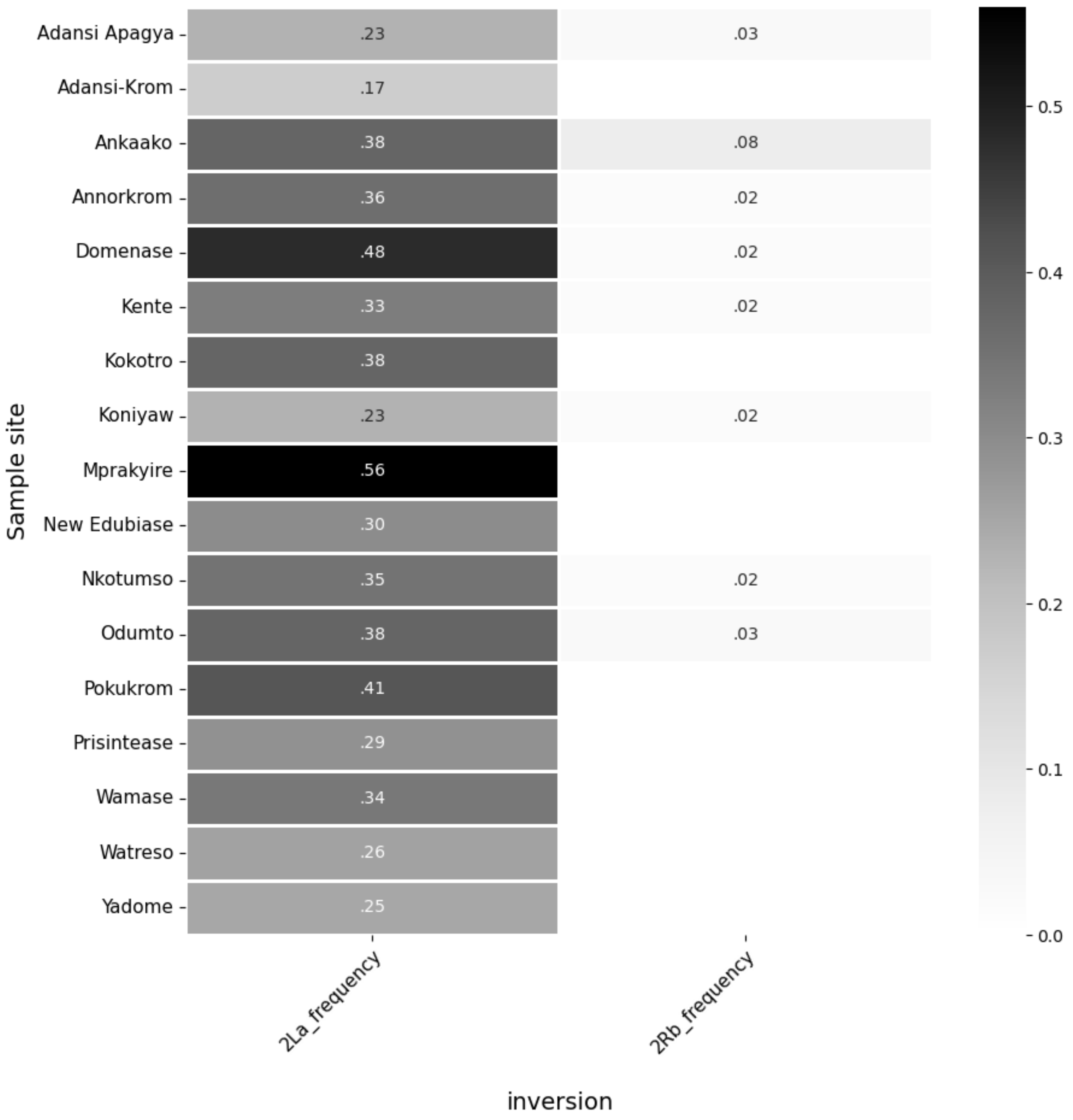


Figure B1) Inversion frequencies of 2La and 2Rb as estimated by compkaryo.

5.6.3 Appendix C

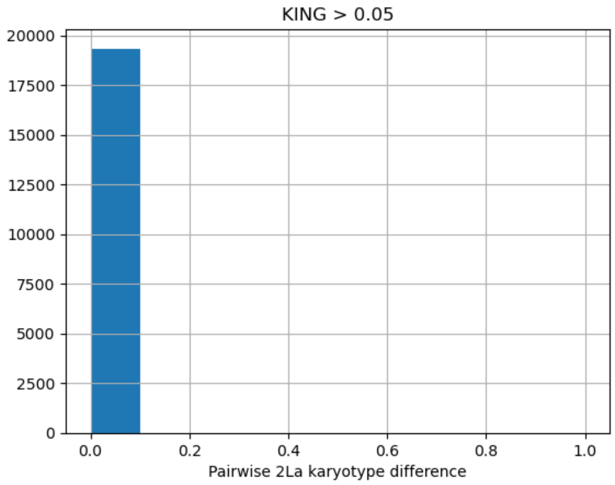


Figure C1. Difference in karyotype between pairwise comparisons with KING value > 0.05

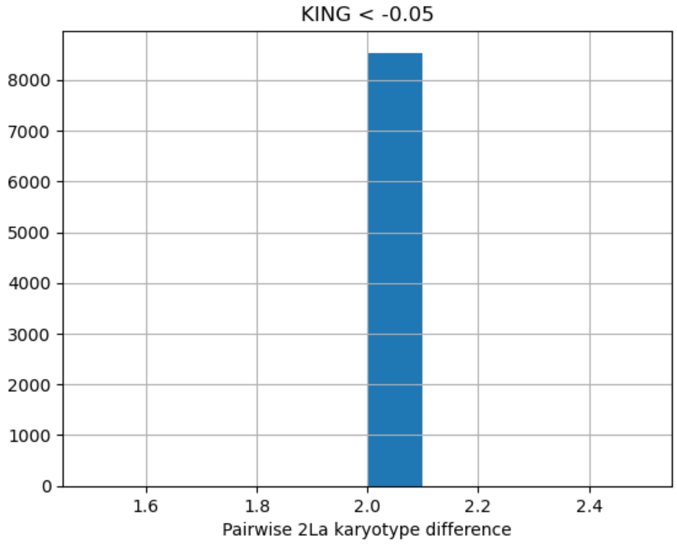


Figure C2. Difference in karyotype between pairwise comparisons with KING value < 0.05

5.6.4 Appendix D

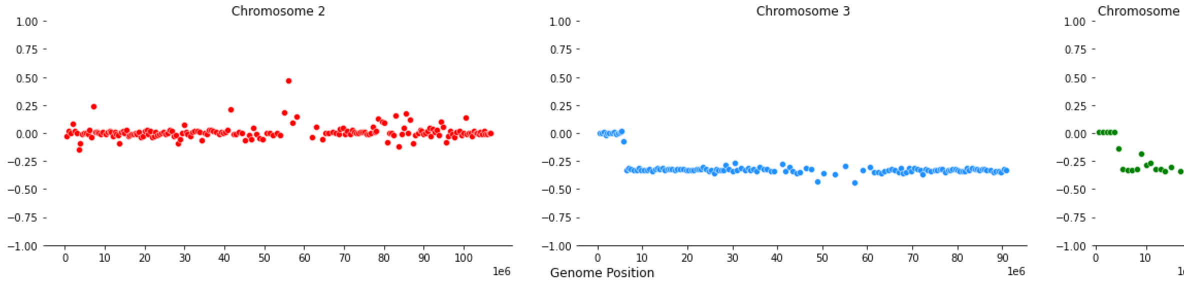


Figure D1. Plots of Hudson's F$_{st}$ across the genome between two *An. gambiae* individuals that were outliers in the *An. gambiae* PCA Figure A1C. WA-2224 v WA 2421. F$_{st}$ was calculated in 10,000 bp stepping windows.

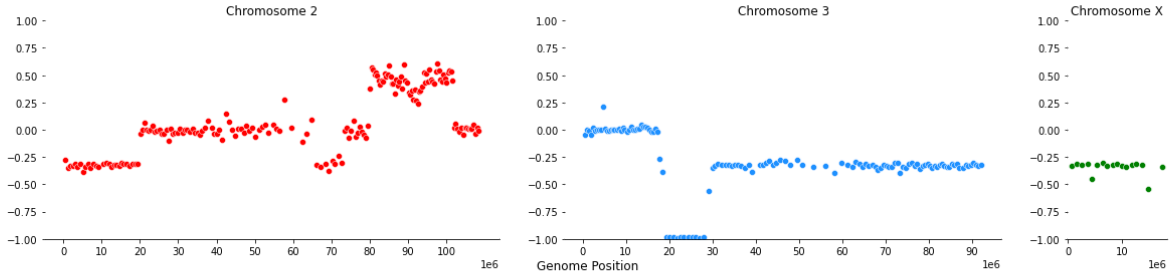


Figure D2. Plots of Hudson's F$_{st}$ across the genome between two samples that are outliers in the *An. coluzzii* PCA Fig A1A and have the fourth most shared doubletons. WA-2014 v WA-2009 Fst. These samples have the 4th most shared doubletons. They look related but are homozygous but different for the 2La inversion, so the KING value is only 0.039. F$_{st}$ was calculated in 10,000 bp stepping windows.

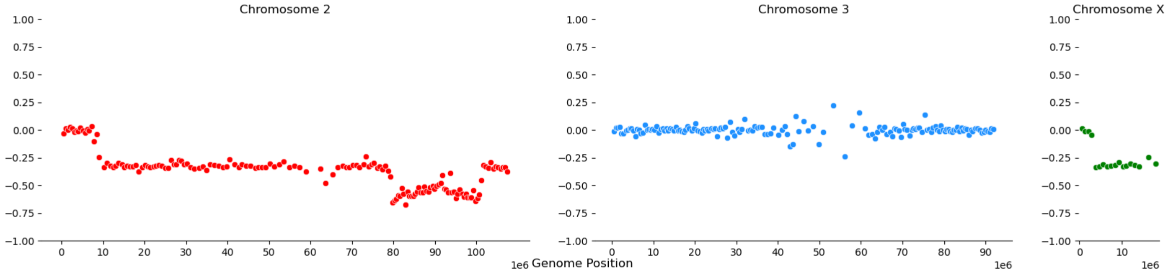


Figure D3. **Plots of Hudson's F$_{st}$ across the genome between two *An. coluzzii* individuals that had the second highest KING value (0.18).** WA-2088 v WA 2285. F$_{st}$ was calculated in 10,000 bp stepping windows.

5.6.5 Appendix E

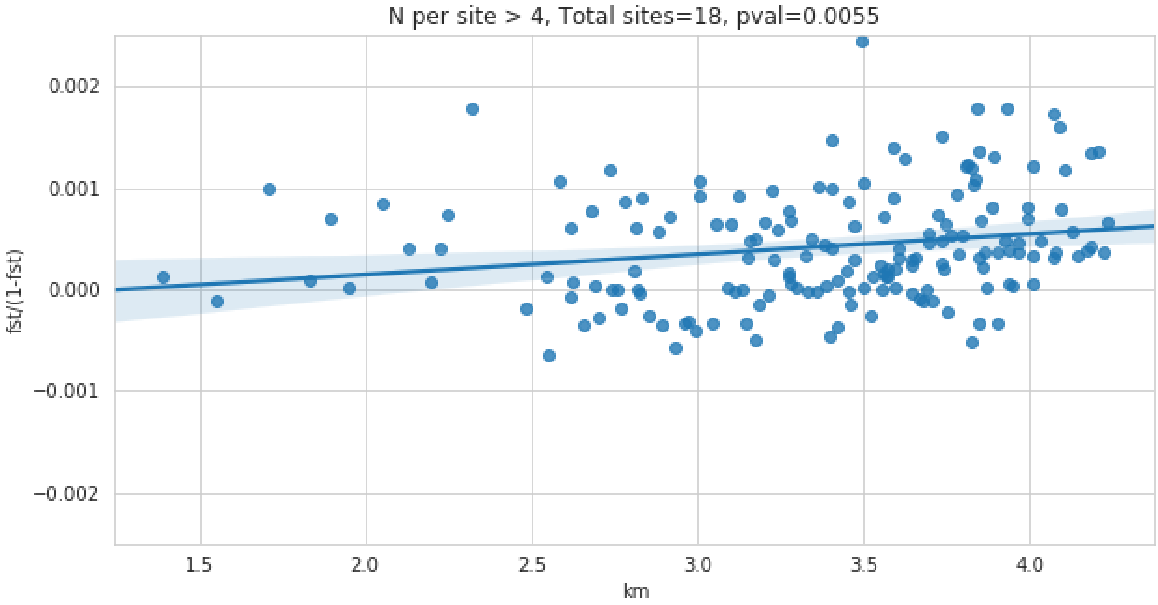


Figure E1. Fst vs geographic distance, with a minimum of five samples per site.

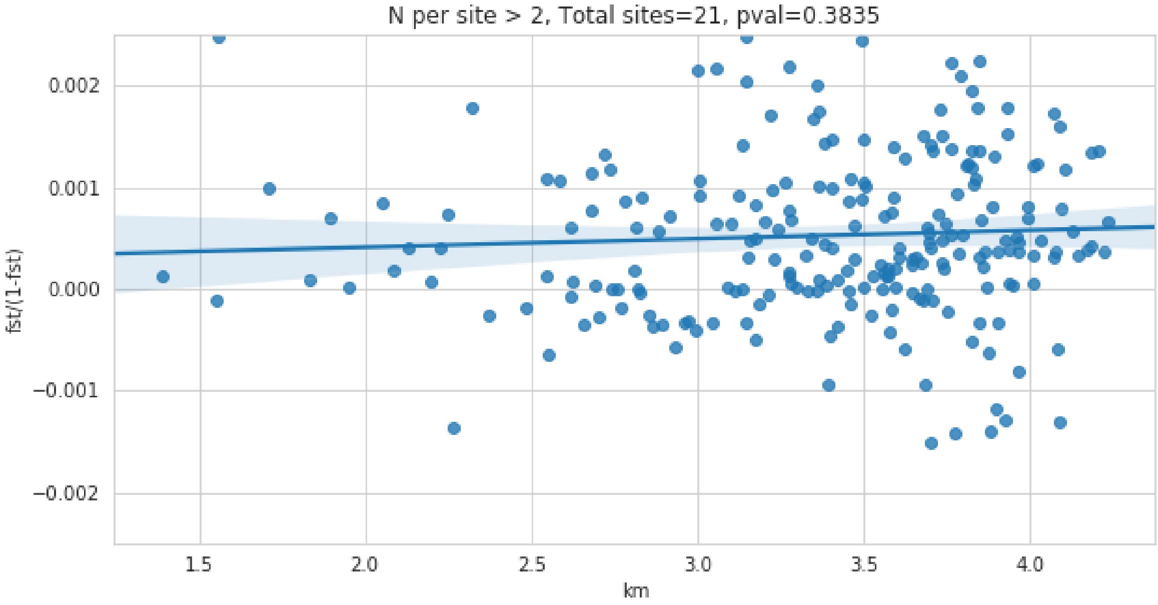


Figure E2. Fst vs geographic distance, with a minimum of three samples per site.

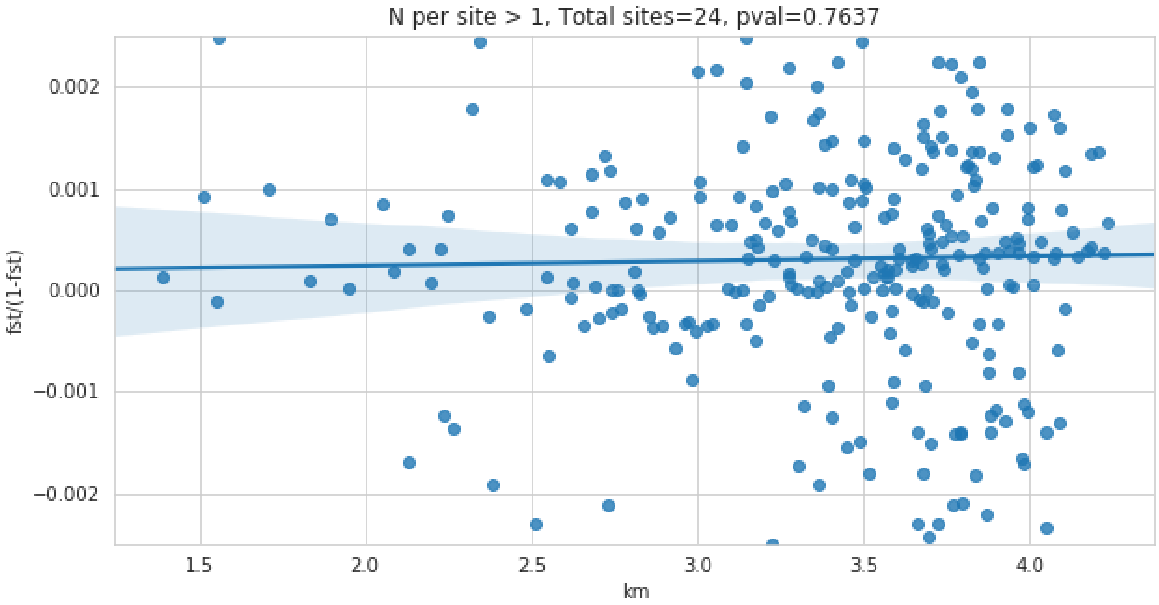


Figure E3. Fst vs geographic distance, with a minimum of two samples per site.

5.6.6 Appendix F

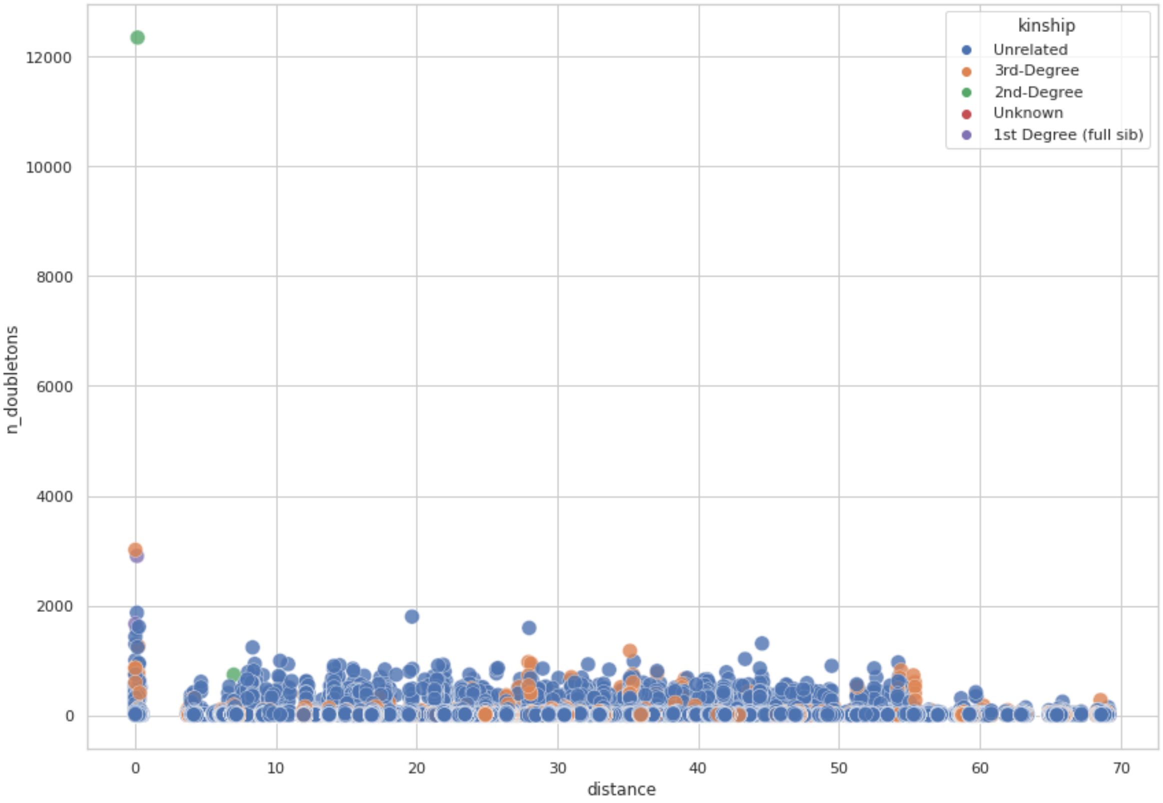


Figure F1. The number of doubletons vs geographic distance between individual samples.

5.6.7 Appendix G

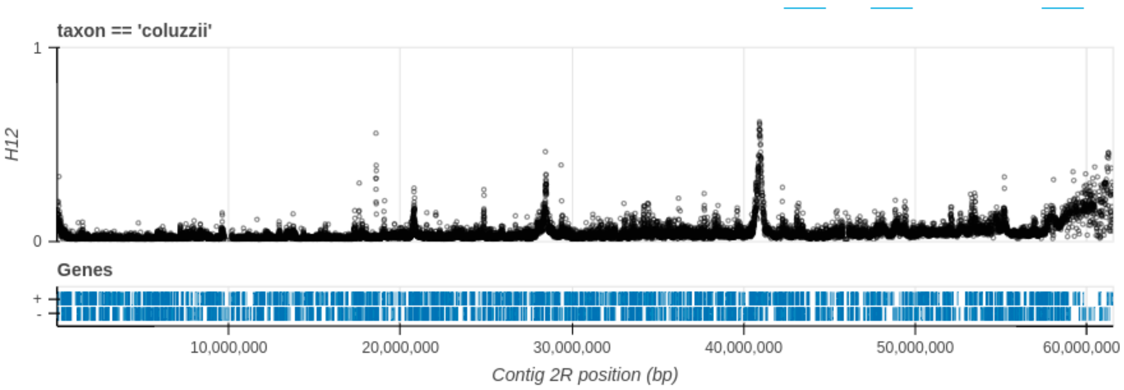


Figure G1. H12 genome-wide selection scan on *An. coluzzii, contig 2R*

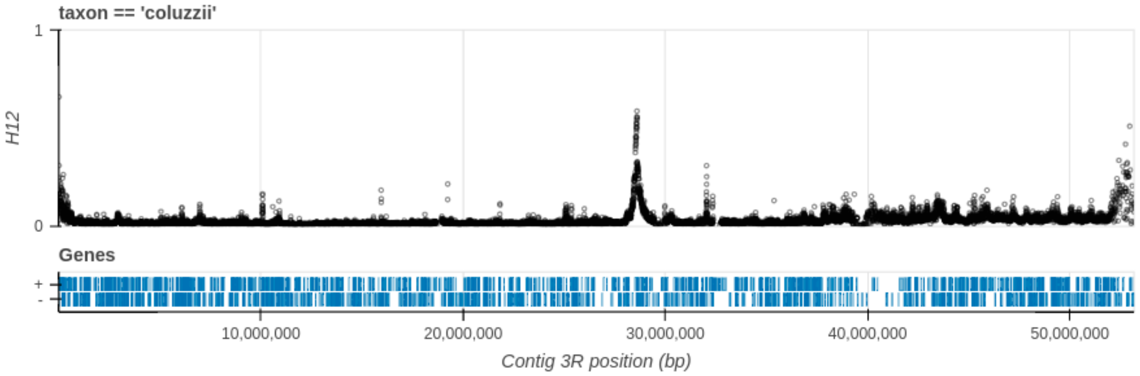


Figure G2. H12 genome-wide selection scan on *An. coluzzii, contig 3R*

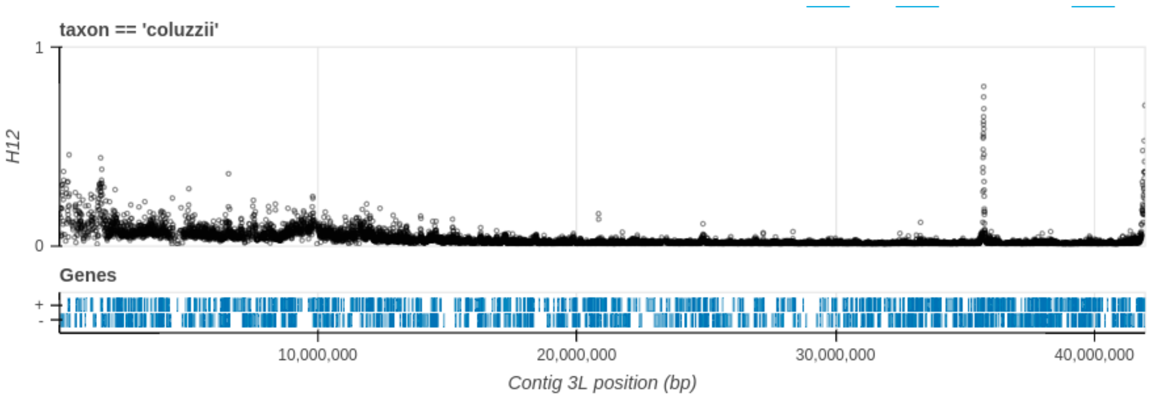


Figure G3. H12 genome-wide selection scan on *An. coluzzii, contig 3L*

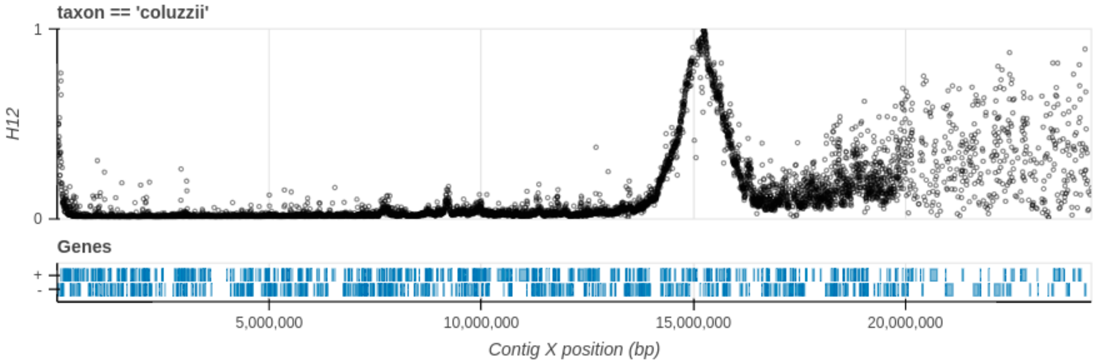


Figure G4. H12 genome-wide selection scan on *An. coluzzii, contig X*

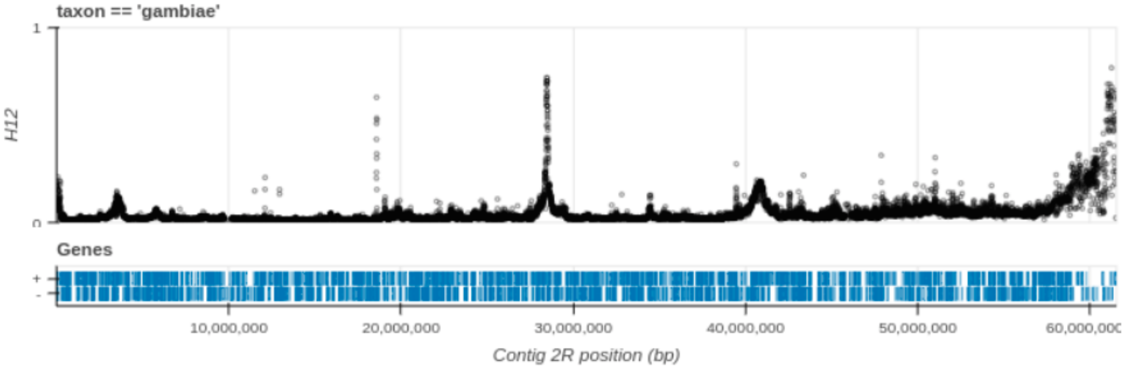


Figure G5. H12 genome-wide selection scan on *An. gambiae*, contig 2R

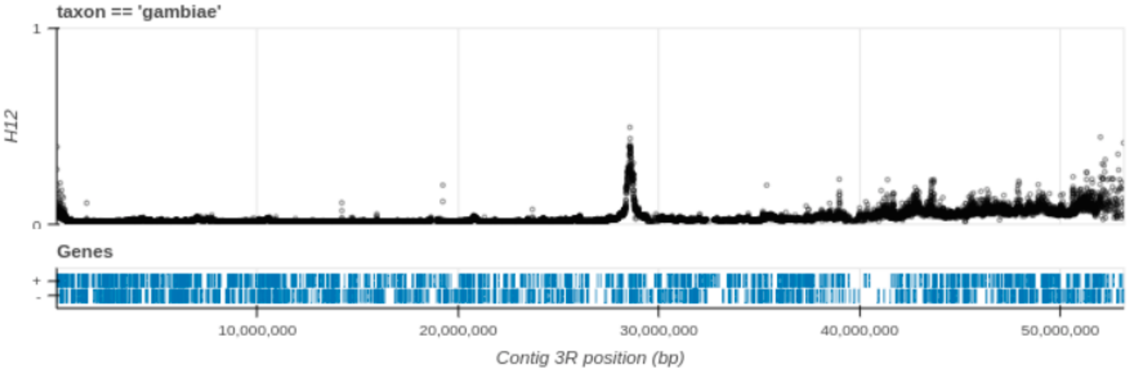


Figure G6. H12 genome-wide selection scan on *An. gambiae*, contig 3R

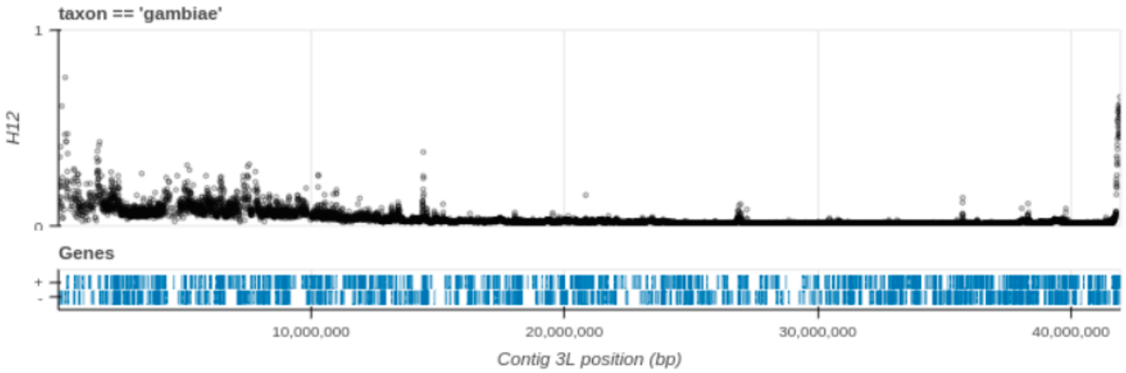


Figure G7. H12 genome-wide selection scan on *An. gambiae*, contig 3L

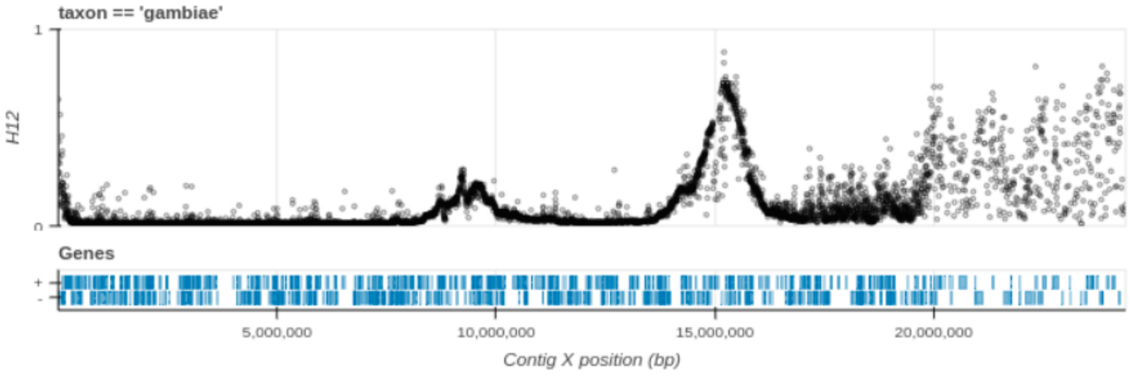


Figure G8. H12 genome-wide selection scan on *An. gambiae*, contig X

5.6.8 Appendix H

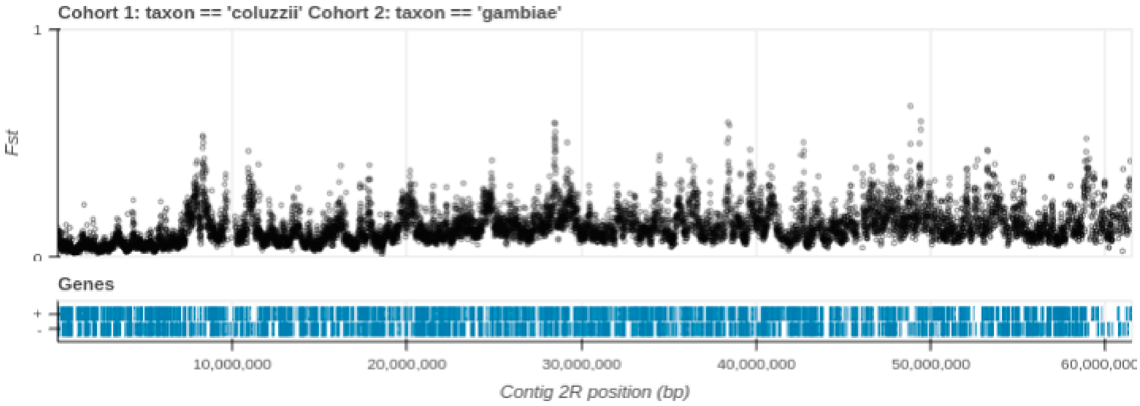


Figure H1. F$_{st}$ genome-wide selection scan on chromosome 2R comparing *An. gambiae to An. coluzzii*

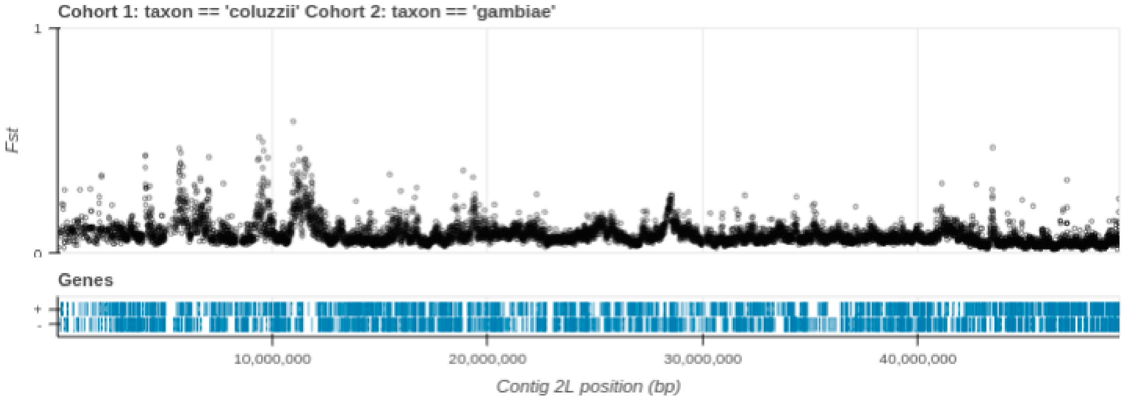


Figure H2. F$_{st}$ genome-wide selection scan on chromosome 2L comparing *An. gambiae to An. coluzzii*

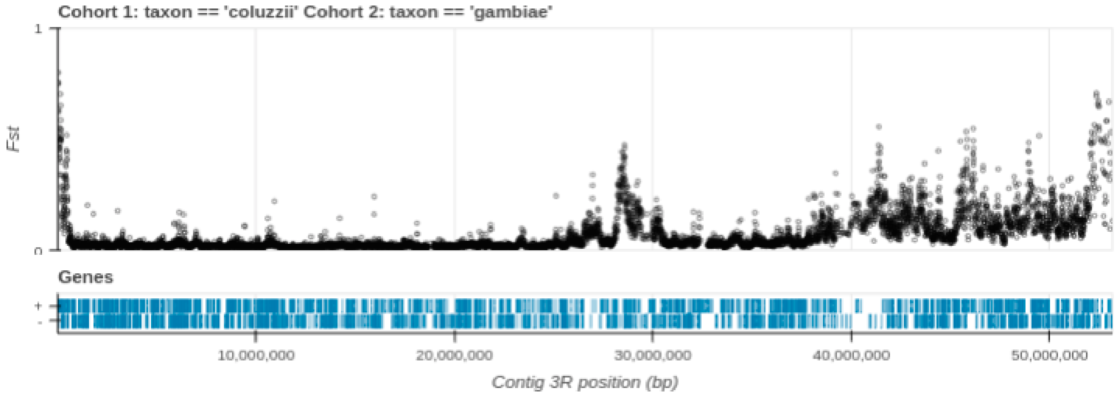


Figure H3. F$_{st}$ genome-wide selection scan on chromosome 3R comparing *An.*

gambiae to An. coluzzii

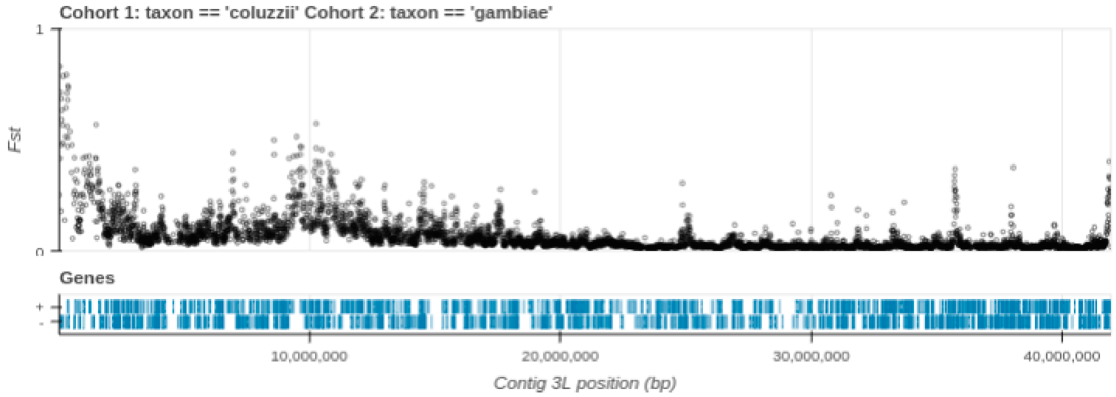


Figure H4. F_{st} genome-wide selection scan on chromosome 3L comparing *An. gambiae to An. coluzzii*

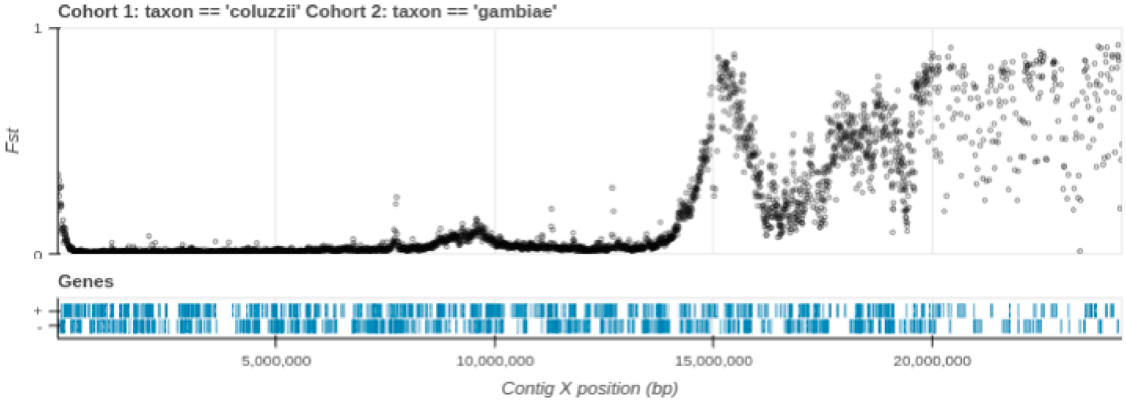


Figure H5. F_{st} genome-wide selection scan on chromosome X comparing *An. gambiae to An. coluzzii*

5.6.9 Appendix I

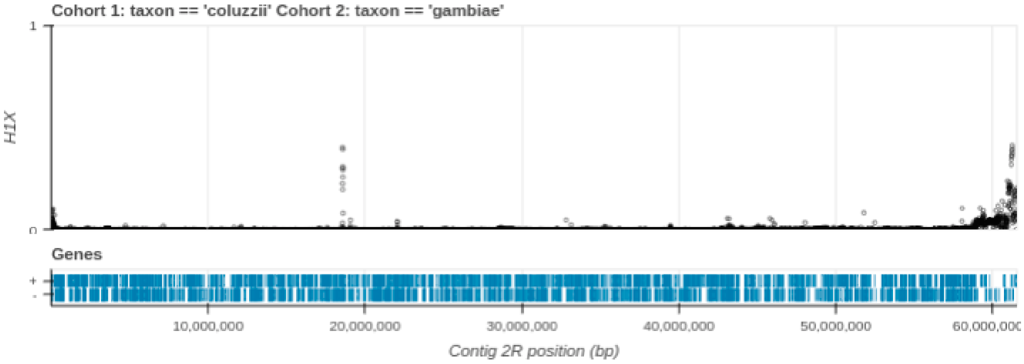


Figure I1. H1X genome-wide scan for shared selective sweeps on contig 2R between An. gambiae and An. coluzzii.

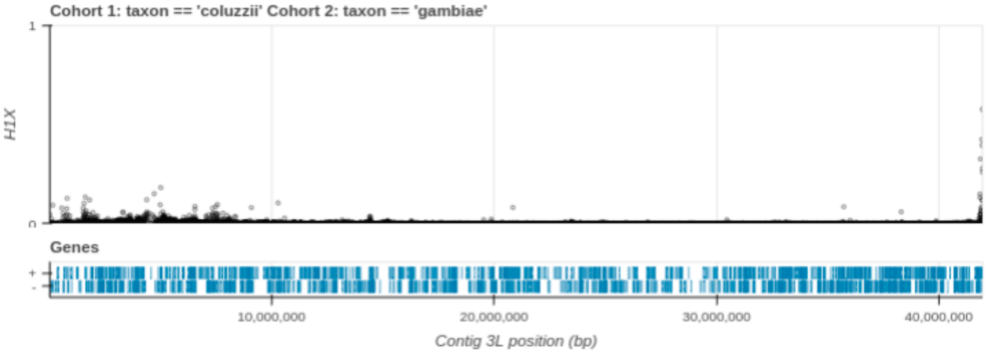


Figure I2. H1X genome-wide scan for shared selective sweeps on contig 3L between An. gambiae and An. coluzzii.

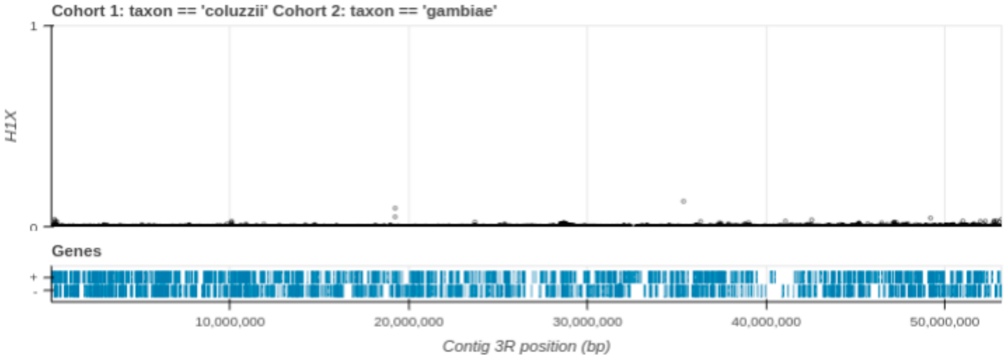


Figure I3. H1X genome-wide scan for shared selective sweeps on contig 3R between An. gambiae and An. coluzzii.

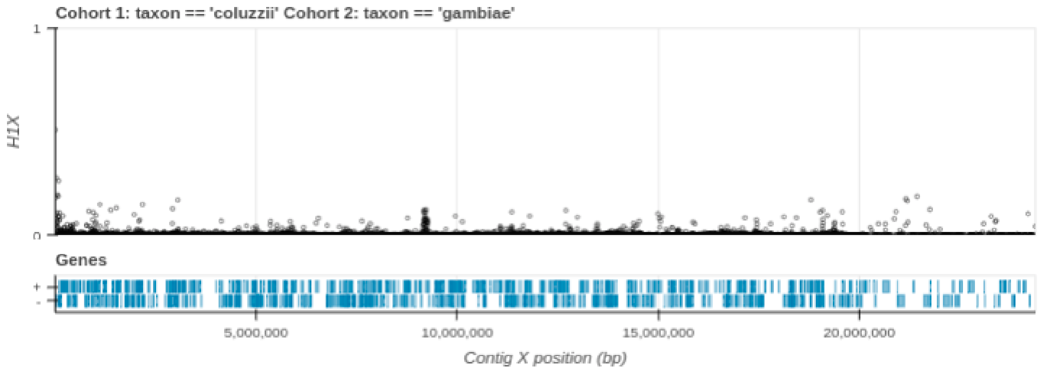


Figure I4. H1X genome-wide scan for shared selective sweeps on contig X between An. gambiae and An. coluzzii.

6

Discussion

The utility of next-generation sequencing (NGS) has led to a proliferation in the scale of sequencing in the biological sciences (Shendure *et al.*, 2017). Scientists have applied NGS to answer a diverse of scientific questions relating to the health of human populations, the ecology and evolution of species, and conservation (Levy *et al.*, 2016). Both individual research groups and large-scale international partnerships, such as the *Anopheles* 1000 genomes project, are generating masses of open, high-quality genomic data (Ag1000G, 2020). Analysing genomic datasets at such a scale, however, brings with it its own challenges, especially in a field where bioinformatics capacity is lacking. In this thesis, I make use of open data derived from next-generation sequencing technologies and develop software tools which may aid researchers to make sense of it all, in a manner which is both reproducible and scalable. I focus primarily on the problem of insecticide resistance - a phenomenon which is both a prime example of contemporary evolution in the face of anthropogenic pressures and also highly relevant to public health.

6.1 Recapitulation and future work

6.1.2 *RNA-Seq-Pop*

In this chapter, we wanted to perform RNA-Sequencing on a pyrethroid-resistant strain of *An. gambiae* collected during a recent large-scale PBO bed-net study in Uganda. These transcriptomic studies are regularly performed in the field of insecticide resistance, and so I thought it would be helpful to develop a reproducible computational pipeline, which we can then apply to other datasets. *RNA-Seq-Pop* can perform substantial quality control and differential expression analyses typical to most RNA-Sequencing studies, however, its major advantage is its analysis of single nucleotide polymorphisms (SNPs). The workflow can call SNPs from the transcriptome data, and perform population genomic analyses, such as calculating measures of diversity and selection statistics. In *An. gambiae*, *RNA-Seq-Pop* users may determine the frequency of common chromosomal inversions, and estimate ancestry across the genome.

In an ideal world, transcriptomic studies for gene expression purposes would be accompanied by whole-genome sequencing. In practice, however, only one published study has done so in *An. gambiae* (Ingham, Tennessen, *et al.*, 2021); generally, transcriptomic studies are designed to have no matching WGS data. By using *RNA-Seq-Pop*, we can actually characterise our transcriptomic samples thoroughly, helping to bridge the gap between transcriptomic and genomic studies.

When we estimate population genetic summary statistics from RNA-Seq data, it will absolutely introduce error as compared to whole-genome sequencing, for reasons mentioned in Chapter 2. When thinking about genetic differentiation (F_{st}) for example, in *RNA-Seq-Pop,* we will be capturing more than just genetic differences between two conditions. Allele-specific expression will bias the allele frequencies compared to whole-genome sequencing, as will the presence of RNA-editing. However, both of these phenomena contribute to the phenotype of the organism, and one could argue, is it not useful to capture these processes in a statistic when we wish to contrast between phenotypes? Could the F_{st} captured by *RNA-Seq-Pop* be more phenotypically relevant than that of whole-genome sequencing? It may not necessarily be the case, but an interesting thought nonetheless.

In our context, *RNA-Seq-Pop* can extend the utility of RNA-Seq itself. For example, the genome of *Aedes aegypti* is over 1 Gb, meaning the cost of whole-genome sequencing is high, compared to *An. gambiae* (278 Mb), which may have contributed to the *Aedes 1200 genomes project* decision to perform low-coverage sequencing. RNA-Seq on the other hand, is broadly equivalent in price between the two species, as the overall number of expressed genes is similar. Utilising the sequence data in RNA-Seq maybe even more important in species with larger genomes. Although I initially developed *RNA-Seq-Pop* to analyse the transcriptomes of disease vectors, the workflow is generalised to be capable of analysing any organism of any ploidy. I envisage that it could be of use to researchers in most biological systems, particularly where samples are taken from the wild.

RNA-Seq-Pop is comprehensive, however, there is still functionality that could be added. I would like to extend the ancestry and karyotyping modules to other major vector species, such as *An. funestus*, *Ae. aegypti* and *Cx. Pipiens*. Large-scale genome projects are ongoing for all of these vectors, which should provide the means to locate ancestry informative markers or karyotype tagging SNPs where appropriate (not all of these species are part of species complexes).

The documentation for *RNA-Seq-Pop* is now adequate, however, further work could be done to develop this into a short training course, which may benefit new users. It may also be possible to develop a graphical user interface (GUI) or a text user interface (TUI), to make the program even more accessible to new users, or to users with little to no experience using the command line. Snakemake itself has its own experimental web browser GUI, however, it is highly limited at the current time of writing and needs substantial development before this is an option.

6.1.3 AgamPrimer

In chapter 3, I present AgamPrimer, a python package coupled with Google Colaboratory, which allows users to design primers and probes with Primer3, whilst checking for genomic variation in primer binding sites (Nagi *et al.*, 2023). SNPs in primer binding sites can cause allelic dropout, where some alleles are not amplified. These problems can go unnoticed, as true heterozygotes can appear as homozygotes, or in quantitative PCR assays, may provide false positives in which samples appear to have different quantities of the nucleic acid of interest, when in fact, one condition is simply amplifying suboptimally. AgamPrimer uses *malariagen_data* and *Primer3-py* to design primer sets and then checks in primer binding sites for SNP variation in the Ag1000G. Users can select specific cohorts or use the whole of Ag1000G phase 3. This allows users to quickly avoid primer sets which may be at risk of failure in *An. gambiae s.l* populations.

In the future, it will be useful to extend AgamPrimer beyond *An. gambiae s.l.* Alongside the ongoing *An. funestus* 1000 genomes project, support for *An. funestus* has been introduced into *malariagen_data*, and so we plan to incorporate this data to further increase the utility of the package, accompanied by a name change to AnoPrimer. At the time of writing, MalariaGEN have just released Pf7, an open dataset of over 20,000 whole-genome sequenced samples (MalariaGEN, 2023). There is also some limited support for Pf7 in *malariagen_data*, and thus it could also be worth adding this functionality into AgamPrimer.

Currently, AgamPrimer is suited to designing primer sets for one locus at a time. For most purposes, this is fine, however, certain use cases exist for which this is limiting. For example, in the design of amplicon sequencing panels, we often wish to target up to or more than 100 amplicons, making the one-by-one design unfeasible, or at least, time-consuming. AgamPrimer requires user judgement on the preferred primer candidates, based on the primer set characteristics and SNP allele frequencies. Ideally, we could build a scoring function, based on the combined position and frequency of any SNPs, so that primer sets can be automatically ranked. This would simplify the automation of complex amplicon panel design.

A limitation of AgamPrimer is that it is solely based on variation in single-nucleotide polymorphisms, as there is no representation of small insertions or deletions (indels) in the Ag1000G. Currently, we also do not integrate CNV calls into AgamPrimer, which may be particularly relevant if a deletion has occurred over a gene of interest.

6.1.4 Parallel evolution at the *Coeae1f/Coeae2f* locus

In chapter 4, I investigate large novel signals of selection on the 2L chromosomal arm at approximately 28.5 Mb. We find two genes *Coeae1f* and *Coeae2f* directly under the selection signal peaks that are orthologs of well-documented genes in *Culex pipiens*, known to confer resistance to organophosphate insecticides. We integrate SNP, copy number variation and gene expression data, and perform a haplotype-based analysis to detect multiple distinct selective sweep events. We identify introgression events between both species and karyotypes. We then use phenotyped data to show that some of these haplotypes under selection are protective against pirimiphos-methyl, with one haplotype also conferring resistance to Deltamethrin. Overall, the analyses are an example of parallel evolution in mosquitoes, and these haplotypes should be monitored as part of the surveillance of organophosphate resistance.

Although haplotype-based approaches using phenotyped data have the potential to circumvent the need for functional validation studies, it would be interesting to dig down and express these the *Coeae1f/2f* enzymes in a heterologous expression system, such as in *E. coli* or *Sf9* cells, and perform binding or metabolism assays. Ideally, the exact haplotype from each selective sweep could be synthesised independently and tested for metabolic activity. In addition or alternatively, the genes could be over-expressed in *Drosophila melanogaster* or *Anopheles gambiae* using the Gal4-UAS system and screened with insecticides to confirm the phenotype associations found here.

For future research, it will be useful to port some of the analyses performed here into *malariagen_data*. For example, modules like finding haplotype-tagging SNPs and haplotype clustering that integrates non-synonymous variation.

6.1.5 Obuasi

In chapter 5, we sample *Anopheles gambiae s.l* from a 70 km^2 region in Obuasi, central Ghana, using an ecologically-informed sampling framework (Sedda *et al.*, 2019). We perform a population genomic analysis on 485 mosquitoes whole-genome sequenced to a 30X target coverage. Previous studies of isolation-by-distance have focused on much larger spatial scales, or used low-resolution genomic markers such as microsatellites (Gélin *et al.*, 2016; Ag1000G, 2020). Using multiple measures of relatedness, including F_{st}, kinship, and the number of pairwise shared doubletons, we detect isolation-by-distance and population structure at an ultra-fine scale. We find that geographic distance, rather than ecological variation, drives genetic differentiation at this scale. It will be useful to extend these ecological analyses to a continent-wide scale. We find continued evolution of variants at the target of pyrethroid insecticides, the *Vgsc*, and a high level of haplotypes under selection at resistance loci, consistent with high rates of phenotypic resistance reported in Ghana (Mugenzi *et al.*, 2022; Lucas *et al.*, 2023).

We do find that the presence of chromosomal inversions obstructs the inference of close-kin in *Anopheles gambiae s.l.* Further work is ongoing to determine this impact, and to re-estimate kinship after the masking of inversion regions. Calculating kinship in mosquito species is an important goal, as it can be used to answer questions such as dispersal, a method known as close-kin mark-recapture (CKMR) (Bravington *et al.*, 2016). This has been done with some success in *Ae. aegypti (Filipović et al., 2020; Sharma et al., 2022)*, however, this species disperses much less far than *An. gambiae* per-generation. More extreme dispersal is challenging, as it means requiring more intensive sampling to recover the same number of relatives. In this study, the 2La inversion seemed to drive kinship signals at low levels of relatedness, suggesting we should mask this region and evaluate the differences. Most likely it will be necessary to mask or remove these regions. Although the number of expected crossover events per chromosome is similar in humans and mosquitoes, due to the few chromosomes in mosquito species, there is a high variance in the length of the genome which is identical by descent in close relatives. This has the effect that, for example, mosquito siblings can vary in their genetic relatedness much more than human siblings would. This phenomenon is likely, in practice, to make exact kinship inference in mosquitoes challenging. It should, however, be possible to perform a simulation study to test the ability of any genomic kinship statistics, such as KING, in detecting siblings and relatives:

1. **Create a synthetic pedigree**. Take the full chromosome haplotypes from two unrelated individuals from an Ag1000G cohort, and designate these as the parents. Repeat this process, using at least 64 distinct total parents in order to be able to produce a large enough pedigree. The chromosome haplotype will contain switch errors, however, in practice, this should not affect the simulation study.

2. **Generate artificial gametes for each parent**, by recombining each parent's chromosomes. To determine recombination breakpoints, draw from the binomial distribution at the *An. gambiae* per-base recombination rate, as many times as the length of the chromosome in bases. The indices of each binomial success, indicate the recombination breakpoints. Concatenate the appropriate segments of each chromosome together. Computationally, this is very simple and only involves splitting and concatenating numerical arrays.
3. **Create offspring** by randomly selecting an artificial gamete from each parent to construct a diploid chromosome.
4. **Repeat this process,** until you have a full pedigree, including parent-child, sibling, cousin, second-cousin, and so on relationships present, and their diploid genomic sequences.
5. **Calculate kinship statistics and evaluate.**

The study also highlights a lack of computational tools for performing population genomics with geographically spaced samples. Previous work in *Anopheles* has generally focused on whole cohorts that come from a specific village, rather than many samples that are distributed somewhat evenly over a region. As genomic surveillance begins to be integrated with intervention trials to detect mechanisms of resistance early, sampling over geographic regions and also through time will also become commonplace. There is a need for software to deal with these kinds of datasets, both pure population genetics methods, and methods to visualise the data (Bradburd *et al*., 2019). For example:

- □ How can we best visualise allele frequencies in geographic space? And in time - how can we visualise the spread of resistance haplotypes, analogous to outbreaks of a pathogen or drug resistance?
- □ How can we visualise relationships between samples in space? For example, sibship?
- □ What are the best approaches to test for differences in allele frequencies in geographic space?

□ How can we calculate and visualise similarities and differences in genome-wide selection signals between locations? Or more specifically, similarities and differences in the frequencies of haplotypes under selection. Can we create a measure of genetic distance specifically relating to IR loci?

As genomic surveillance moves into the mainstream, more attention is now being paid to sampling regimens that can introduce the least bias and produce the most representative data. The proportional lattice plus close pairs design developed by Sedda et al. (2019), provides a method to ensure the representativeness of ecological zones, as well as variable between-pair distances between sites. Lattices are efficient for prediction in spatial models, whilst close pairs are efficient for parameter estimation (Zimmerman, 2006), therefore a combination of both is optimal (Sedda *et al.*, 2019). Similarly, this approach is also advantageous for genomic surveillance. The lattice allows for comprehensive spatial coverage in allele frequency estimates, which should improve the spatial prediction of resistance marker frequencies and their spread, as well as reducing gaps in which novel variants of concern can go undiscovered. The variation in geographic distance provided by close pairs is also important for population genomics, as it allows for greater power and precision in estimations of isolation by distance, informing estimates at which mosquitoes can disperse (Manel *et al.*, 2012). As mosquito populations are likely to show adaptation to ecological zones within which they inhabit, ensuring ecological representativeness across the sampling area ensures the representativeness of mosquito ecotypes (Cheng *et al.*, 2012). Although we use data resulting from this approach in Chapter 5, examination of the sampling framework itself was limited and further work remains in demonstrating the benefits of this sampling framework over traditional convenience and local knowledge-based methods. In addition, the framework does not consider geographic accessibility, which may be useful in reducing the overall cost of surveillance (Longbottom *et al.*, 2020). This may allow the use of more sample sites, which is beneficial in obtaining precise landscape genomic estimates (Aguirre-Liguori *et al.*, 2020).

6.2 Open-source software as a tool for capacity development

Throughout this thesis and during the course of my doctoral degree, I have strived to develop computational tools that other researchers may use to perform their own analyses. Building software that is accessible and easy to use can democratise computational research, empowering researchers to perform complex analyses that may otherwise be unfeasible to learn and complete in the time provided by modern academic settings. Indeed, the development of the population genetic toolkit, scikit-allel, has been instrumental to this thesis and my own research.

A major effort in infectious disease research is to strengthen the capacity of institutes in the global south. Given the prominent role of genomics and bioinformatics in the contemporary biological sciences, building bioinformatics capacity is a specific priority, exemplified in recent projects such as H3abionet, the pan-African bioinformatics network, and the MalariaGEN-PAMCA bioinformatics workshops. As well as being a teaching assistant, I have contributed modules in the MalariaGEN-PAMCA workshops focusing on primer design with AgamPrimer (Workshop 6 Module 4), and haplotype clustering to identify adaptive gene flow (Workshop 7 Module 3). Additionally, I have introduced functionality to *malariagen_data*, a python package to load and analyse data from the Ag1000G and Vector Observatory projects. The PAMCA workshops have been a joy to be a part of, and the efforts by collaborators at MalariaGEN to develop software alongside the course have been herculean and incredibly impressive (though not on my part!). At the current time of writing, there is no other organism on earth with which you can so easily interact and analyse its genomic data than *Anopheles gambiae s.l*, and this is something that every person who’s contributed should be particularly proud of.

6.3 An ecosystem of software for malaria research

In this thesis, I describe multiple software tools for *Anopheles'* research. Of primary interest to the community - *RNA-Seq-Pop*, *AgamPrimer* and *AnoExpress* (Appendix A) - a tool which summarises gene expression across RNA-Sequencing insecticide resistance studies. There is also the immense *malariagen_data*, which underpins *AgamPrimer* and is integrated into the *LocusPocus* and *Probe* snakemake workflows.

One could envisage an ecosystem of python-based software tools for malaria research, which are able to interact with each other. Certainly, the capabilities of *malariagen_data* would fit well with *AnoExpress* - particularly if a user aims to find genes involved in insecticide resistance. Already, users could use *RNA-Seq-Pop* to reproducibly analyse their RNA-Sequencing data to find a gene of interest, explore genomic and copy number variation in that gene with *malariagen_data*, use *AnoExpress* to look at its distribution of expression across sub-Saharan Africa, and design primers and probes with *AgamPrimer* to perform functional validation or track the gene of interest. It is easy to imagine the development of a training course which integrates all of these tools together.

6.4 Future directions in *Anopheles* insecticide resistance research

For brevity's sake, in this section, I will focus on a few areas of research which I myself could feasibly tackle in the future.

Research into insecticide resistance in the major malaria mosquito, *Anopheles gambiae*, has come far since the earliest discoveries of emerging resistance in the 1960s (Davidson *et al.*, 1962). The malaria mosquito has become a model for studying insecticide resistance in insects. There are, however, still a huge amount of unknowns in regard to this phenomenon.

The first phase of the Anopheles 1000 genomes project revealed extensive signals of selection across the genomes of *An. gambiae* and *An. coluzzii*. Work was done on a phase 1 selection atlas, however, it was not completed (Harding *et al.*, 2019). We plan to update this work and build a selection atlas with genome-wide selection scans from the full release of the Ag1000G. With these data, we can identify selective sweeps and clusters of haplotypes, and define or classify these somehow, as was done for SARS-CoV-2 strains, in order to track them in future research. This is likely to require developing novel software or functionality, probably in *malariagen_data*, though more generic software may be needed which can be applied to any organism. We can then begin to track these haplotypes in future research, using phenotyped data to test for associations with insecticides - as was performed in chapter 4. Depending on the number of selective sweeps we find, we could find haplotype-tagging SNPs (as in chapter 4) and feasibly use AgamPrimer as a framework to design a highly multiplexed amplicon sequencing panel with which to identify and track these haplotypes. You could then use the high-throughput nature of amplicon sequencing to test many phenotyped samples and test for associations with insecticide resistance with greatly increased power. This would also reduce the need for laborious functional validation. A nanopore-based amplicon panel was recently developed for malaria drug resistance loci and is worth considering in addition to Illumina sequencing (Girgis *et al.*, 2022).

We do not know which mosquito tissues the detoxification of insecticides primarily occurs in. Most transcriptomic studies are performed on whole-body RNA extracts, providing zero resolution of any tissue specificity in the changes in gene expression. This could be problematic, because if a smaller tissue is important for detoxification, any signal may be drowned out by RNA from the rest of the carcass. A previous study performed tissue-specific microarray experiments on four tissues (Ingham *et al.*, 2014), finding tissue-specific expression of certain candidate genes, however, it would be useful to update this research with RNA-Sequencing and to include more tissues, such as the brain, antennae and legs. It may, however, be prudent to explore the spatial expression of gene expression with the much greater fidelity of modern technologies. In *Plasmodium*, researchers have developed an atlas of gene expression at single-cell resolution in different life stages and in passage through the vector mosquito (Howick *et al.*, 2019; Real *et al.*, 2021). Similarly, researchers in the *Drosophila* community have produced the fly cell atlas (Li *et al.*, 2022), from both sexes in 15 individually dissected tissues. One can imagine designing a similar project in *Anopheles gambiae* using the *Drosophila* experiments as a framework, profiling gene expression and comparing female insecticide-susceptible and resistant strains, rather than sexes. Ideally, these would be from a similar genetic background, though this will probably be difficult to achieve. Another possibility would be to sequence the same strain before and after exposure to insecticides. This would enable a much greater understanding of where in the adult mosquito candidate resistance genes are expressed, and exactly how they are contributing to resistance. It would also provide a much greater resolution with which to build gene regulatory networks (GRNs) (Ingham, Elg, *et al.*, 2021), enhancing the discovery of genes involved in the regulation of resistance.

Traditional genomics makes use of a single linear reference genome assembly, typically from a single individual, or more commonly, a pool of highly inbred individuals. Although convenient, this is limiting, as the lack of variation present in the reference causes reference bias, where difficulties arise in mapping sequence reads to the reference, due to SNPs, indels or larger variants (Paten *et al.*, 2017). In theory, the extreme genetic variation in *An. gambiae s.l* will make reference bias a larger problem than in organisms with lower diversity, such as humans. The developing field of pangenomics has emerged to address and mitigate this bias (Eizenga *et al.*, 2020). Pangenomics involves the construction of reference genomes which incorporate population-level genetic variation, usually represented as variation graphs (Garrison *et al.*, 2018). Although the field of pangenomics will certainly prove to be highly important for DNA sequencing studies, I am specifically interested in the use of pangenomics in transcriptomic research. Recently Sibbeson and colleagues demonstrated that by creating a pan-transcriptome, mapping can be made more efficient, and haplotype-specific transcript expression can be determined (Sibbesen *et al.*, 2023). In theory, it should even be possible to combine the power of pan-transcriptomics with single-cell transcriptomics.

Currently, all published RNA-Sequencing studies into insecticide resistance are based on short-read Illumina sequencing, in which RNA is first reverse transcribed to cDNA and fragmented prior to sequencing. Though this is likely adequate for the quantification of genes, it introduces difficulty when one wishes to analyse alternative splicing, particularly when the parent gene contains many exons and can give rise to many alternative isoforms. An example highly relevant to public health is that of the Voltage-gated sodium channel (*Vgsc)*, the target of pyrethroid insecticides. In *Anopheles gambiae*, the *Vgsc* contains 39 annotated exons with 13 known transcripts. This can make the inference of alternative splicing particularly challenging with short-read data. Recently, Oxford nanopore introduced technologies and protocols to perform long-read sequencing of native RNA transcripts, without cDNA synthesis or PCR (Garalde *et al.*, 2018). This allows researchers to read through and reconstruct full-length isoforms, and reduces bias in the quantification process. Applying direct long-read RNA-Sequencing in the context of insecticide resistance may provide useful insights into mechanisms of insecticide resistance which may otherwise be missed with short-read sequencing.

6.5 References

Ag1000G (2020) ‘Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii’, *Genome Res*, pp. 1–14.

Aguirre-Liguori, J.A., Luna-Sánchez, J.A., Gasca-Pineda, J. and Eguiarte, L.E. (2020) ‘Evaluation of the Minimum Sampling Design for Population Genomic and Microsatellite Studies: An Analysis Based on Wild Maize’, *Frontiers in genetics*, 11, p. 870.

Bradburd, G.S. and Ralph, P.L. (2019) ‘Spatial Population Genetics: It’s About Time’, *Annual review of ecology, evolution, and systematics*, 50(1), pp. 427–449.

Bravington, M.V., Skaug, H.J. and Anderson, E.C. (2016) ‘Close-Kin Mark-Recapture’, *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31(2), pp. 259–274.

Cheng, C., White, B.J., Kamdem, C., Mockaitis, K., Costantini, C., Hahn, M.W. and Besansky, N.J. (2012) ‘Ecological genomics of Anopheles gambiae along a latitudinal cline: a population-resequencing approach’, *Genetics*, 190(4), pp. 1417–1432.

Davidson, G. and Hamon, J. (1962) ‘A Case of Dominant Dieldrin Resistance in Anopheles gambiae Giles’, *Nature*, 196(4858), pp. 1012–1012.

Eizenga, J.M., Novak, A.M., Sibbesen, J.A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J.D., Rounthwaite, R., Ebler, J., Rautiainen, M., Garg, S., Paten, B., Marschall, T., Sirén, J., *et al.* (2020) ‘Pangenome Graphs’, *Annual review of genomics and human genetics*, 21, pp. 139–162.

Filipović, I., Hapuarachchi, H.C., Tien, W.-P., Razak, M.A.B.A., Lee, C., Tan, C.H., Devine, G.J. and Rašić, G. (2020) ‘Using spatial genetics to quantify mosquito dispersal for control programs’, *BMC biology*, 18(1), p. 104.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., *et al.* (2018) ‘Highly parallel direct RNA sequencing on an array of nanopores’, *Nature methods*, 15(3), pp. 201–206.

Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., Paten, B. and Durbin, R. (2018) ‘Variation graph toolkit improves read mapping by representing genetic variation in the reference', *Nature biotechnology*, 36(9), pp. 875–879.

Gélin, P., Magalon, H., Drakeley, C., Maxwell, C., Magesa, S., Takken, W. and Boëte, C. (2016) 'The fine-scale genetic structure of the malaria vectors Anopheles funestus and Anopheles gambiae (Diptera: Culicidae) in the north-eastern part of Tanzania', *International journal of tropical insect science*, 36(4), pp. 161–170.

Girgis, S.T., Adika, E., Nenyewodey, F.E., Senoo Jnr, D.K., Ngoi, J.M., Bandoh, K., Lorenz, O., van de Steeg, G., Nsoh, S., Judge, K., Pearson, R.D., Almagro-Garcia, J., Saiid, S., Atampah, S., Amoako, E.K., *et al.* (2022) 'Nanopore sequencing for real-time genomic surveillance of Plasmodium falciparum', *bioRxiv*. doi:10.1101/2022.12.20.521122.

Harding, N.J., Miles, A., Clarkson, C.S., Lucas, E., Kozak, K., Donnelly, M., Lawniczak, M., Kwiatkowski, D. and Anopheles, T. (2019) 'An atlas of recent positive selection in the African malaria vectors Anopheles gambiae and Anopheles coluzzii', pp. 1–33.

Howick, V.M., Russell, A.J.C., Andrews, T., Heaton, H., Reid, A.J., Natarajan, K., Butungi, H., Metcalf, T., Verzier, L.H., Rayner, J.C., Berriman, M., Herren, J.K., Billker, O., Hemberg, M., Talman, A.M., *et al.* (2019) 'The Malaria Cell Atlas: Single parasite transcriptomes across the complete *Plasmodium* life cycle', *Science*, 365(6455), p. eaaw2619.

Ingham, V.A., Elg, S., Nagi, S.C. and Dondelinger, F. (2021) 'Capturing the transcription factor interactome in response to sub-lethal insecticide exposure', *Current research in insect science*, 1, p. None.

Ingham, V.A., Jones, C.M., Pignatelli, P., Balabanidou, V., Vontas, J., Wagstaff, S.C., Moore, J.D. and Ranson, H. (2014) 'Dissecting the organ specificity of insecticide resistance candidate genes in Anopheles gambiae: known and novel candidate genes', *BMC genomics*, 15(1), p. 1018.

Ingham, V.A., Tennessen, J.A., Lucas, E.R., Elg, S., Yates, H.C., Carson, J., Guelbeogo, W.M., Sagnon, N. 'fale, Hughes, G.L., Heinz, E., Neafsey, D.E. and Ranson, H. (2021) 'Integration of whole genome sequencing and transcriptomics reveals a complex picture of the reestablishment of insecticide resistance in the major malaria vector Anopheles coluzzii', *PLoS genetics*, 17(12), p. e1009970.

Levy, S.E. and Myers, R.M. (2016) 'Advancements in Next-Generation Sequencing', *Annual review of genomics and human genetics*, 17, pp. 95–115.

Li, H., Janssens, J., Waegeneer, M.D., Kolluru, S.S., Davie, K., Gardeux, V., Saelens, W., David, F.P.A., Brbić, M., Spanier, K., Leskovec, J., McLaughlin, C.N., Xie, Q., Jones, R.C., Brueckner, K., *et al.* (2022) 'Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly', *Science*, 375(6584), p. eabk2432.

Longbottom, J., Krause, A., Torr, S.J. and Stanton, M.C. (2020) 'Quantifying geographic accessibility to improve efficiency of entomological monitoring', *PLoS neglected tropical diseases*, 14(3), p. e0008096.

Lucas, E.R., Nagi, S.C., Egyir-Yawson, A., Essandoh, J., Dadzie, S., Chabi, J., Djogbenou, L.S., Medjigbodo, A.A., Edi, C.V., Ketoh, G.K., Koudou, B.G., Van't Hof, A.E., Rippon, E.J., Pipini, D., Harding, N.J., *et al.* (2023) 'Genome-wide association studies reveal novel loci associated with pyrethroid and organophosphate resistance in Anopheles gambiae s.l', *bioRxiv*. doi:10.1101/2023.01.13.523889.

MalariaGEN (2023) 'Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples', *Wellcome Open Research*, 8(22). doi:10.12688/wellcomeopenres.18681.1.

Manel, S., Albert, C.H. and Yoccoz, N.G. (2012) ‘Sampling in landscape genomics’, *Methods in molecular biology* , 888, pp. 3–12.

Mugenzi, L.M.J., Akosah-Brempong, G., Tchouakui, M., Menze, B.D., Tekoh, T.A., Tchoupo, M., Nkemngo, F.N., Wondji, M.J., Nwaefuna, E.K., Osae, M. and Wondji, C.S. (2022) ‘Escalating pyrethroid resistance in two major malaria vectors Anopheles funestus and Anopheles gambiae (s.l.) in Atatam, Southern Ghana’, *BMC infectious diseases*, 22(1), p. 799.

Nagi, S.C., Miles, A. and Donnelly, M.J. (2023) ‘AgamPrimer: Primer Design in Anopheles gambiae informed by range-wide genomic variation’, *bioRxiv*. doi:10.1101/2022.12.31.521737.

Paten, B., Novak, A.M., Eizenga, J.M. and Garrison, E. (2017) ‘Genome graphs and the evolution of genome inference’, *Genome research*, 27(5), pp. 665–676.

Real, E., Howick, V.M., Dahalan, F.A., Witmer, K., Cudini, J., Andradi-Brown, C., Blight, J., Davidson, M.S., Dogga, S.K., Reid, A.J., Baum, J. and Lawniczak, M.K.N. (2021) ‘A single-cell atlas of Plasmodium falciparum transmission through the mosquito’, *Nature communications*, 12(1), pp. 1–13.

Sedda, L., Lucas, E.R., Djogbénou, L.S., Edi, A.V.C., Egyir-Yawson, A., Kabula, B.I., Midega, J., Ochomo, E., Weetman, D. and Donnelly, M.J. (2019) ‘Improved spatial ecological sampling using open data and standardization: an example from malaria mosquito surveillance’, *Journal of the Royal Society, Interface / the Royal Society*, 16(153), p. 20180941.

Sharma, Y., Bennett, J.B., Rašić, G. and Marshall, J.M. (2022) ‘Close-kin mark-recapture methods to estimate demographic parameters of mosquitoes’, *bioRxiv*. doi:10.1101/2022.02.19.481126.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A. and Waterston, R.H. (2017) ‘DNA sequencing at 40: Past, present and future’, *Nature*, 550(7676). doi:10.1038/nature24286.

Sibbesen, J.A., Eizenga, J.M., Novak, A.M., Sirén, J., Chang, X., Garrison, E. and Paten, B. (2023) ‘Haplotype-aware pantranscriptome analyses using spliced pangenome graphs’, *Nature methods*, pp. 1–9.

Zimmerman, D.L. (2006) ‘Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction’, *Environmetrics*, 17(6), pp. 635–652.

*Appendix A - Ano*Express

This work is a collaboration between I, Sanjay Curtis Nagi and Victoria A Ingham. I have performed all data analyses contained therein. In this appendix, I do not describe the results of the study itself but instead the software tool I have developed to go alongside it.

7.1 Abstract

Gene expression plays a large role in producing insecticide-resistant phenotypes in the African malaria vectors *An. gambiae s.l* and *An. funestus*. Traditionally, microarray chips were used to assay the gene expression on a transcriptome-wide scale, however, this has been replaced with RNA-Sequencing in recent years. We collate raw read count data from all published studies which use RNA-Sequencing to investigate insecticide resistance in the two major vectors and perform a meta-analysis of the data, highlighting the role of known and novel candidate genes. We have developed an online Google Colaboratory-based tool, *Ano*Express, which allows users to rapidly load, query, and visualise the meta-analysis expression data, to enhance the discovery of genes involved in insecticide resistance. The tool is located here https://github.com/sanjaynagi/AnoExpress.

7.2 Introduction

Transcriptomic studies have driven the discovery of genes involved in insecticide resistance since the first microarray in *An. gambiae*, the detoxification chip (David *et al.*, 2005). Since then, RNA-Sequencing has become the de facto technology for transcriptomics. In 2018 Ingham et al., presented an R-shiny tool, IR-Tex, which summarised transcript expression from 31 insecticide-resistance microarray studies in sub-Saharan Africa of *Anopheles gambiae s.l* (Ingham *et al.*, 2018). We present a python-based successor to this tool, which instead summarises gene expression from recently published RNA-Sequencing studies. It includes studies published in *An. gambiae s.l* and also *An. funestus*. The tool is named *AnoExpress*, for *Anopheles* gene expression in resistance studies.

7.3 Methods

VectorBase was used to retrieve orthologs to *An. gambiae* PEST and those with one-to-many relationships were extracted, protein sequences retrieved and BLASTed against a custom PEST4.13 protein database using command line BLAST 2.90. A custom R script was then used to define orthologs using the following parameters in priority order: An e value of 0; multiple e values of 0 within 10% of the top percentage identity; an e value of < 1e-80; multiple e values of < 1e-80 within 10% of the top percentage identity. PEST4.13 genes without orthologs were then back BLASTed using the above parameters against AcolN1.2, AfunF3.2 and AraD1.12.

We perform differential expression analysis within studies to eliminate the possibility of batch effects and reduce errors due to large differences between studies in sequencing depth. Differential expression analysis was performed in R with DESeq2 v2.3.1, and hypothesis testing was performed with the wald test. For enrichment analysis, differentially expressed genes (DEGs) were selected by taking the top 5% percentile of genes when ranked by a median fold change. A custom hypergeometric test was implemented, using *An. gambiae* PEST GO terms and PFAM domain annotations. Plots are produced with plotly and are interactive, to aid interpretation for the user.

7.4 Results / Discussion

In order to integrate RNA-Sequencing datasets from four different species where different reference genomes had been used, one-to-many orthologs were found between the *An. gambiae PEST* reference genome, and the *An. coluzzii, An. arabiensis, and An. funestus* reference genomes. As integrating each successive species reduced the number of genes with orthologs, I performed four separate analyses - ‘gamb_colu’ (*An. gambiae and An. coluzzii)*, ‘gamb_colu_arab’ (adding *An. arabiensis)*, ‘gamb_colu_arab_fun’ (adding *An.*

funestus), and ‘fun’ (*An. funestus alone)*. The purpose of providing multiple datasets was that if users cannot find their genes in the full dataset (‘gamb_colu_arab_fun’), they may still be present in the others.

At the time of writing, AnoExpress is broken up into multiple python notebooks, intended to be run in Google Colaboratory. The main notebook, allows users to plot an interactive summary of gene expression in the users' genes of interest. Figure 1 shows an example plot for *Coeae1f*, *Coeae2f*, and the UGT AGAP006222. Figures are produced with Plotly and are interactive, allowing users to hover over a data point to bring up further information about that specific experiment. Users may rank the plots by the median or mean fold-change, or by the AGAP identifier. A function is provided for users to load their own gene lists.

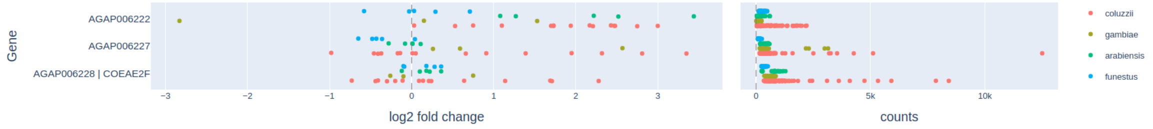
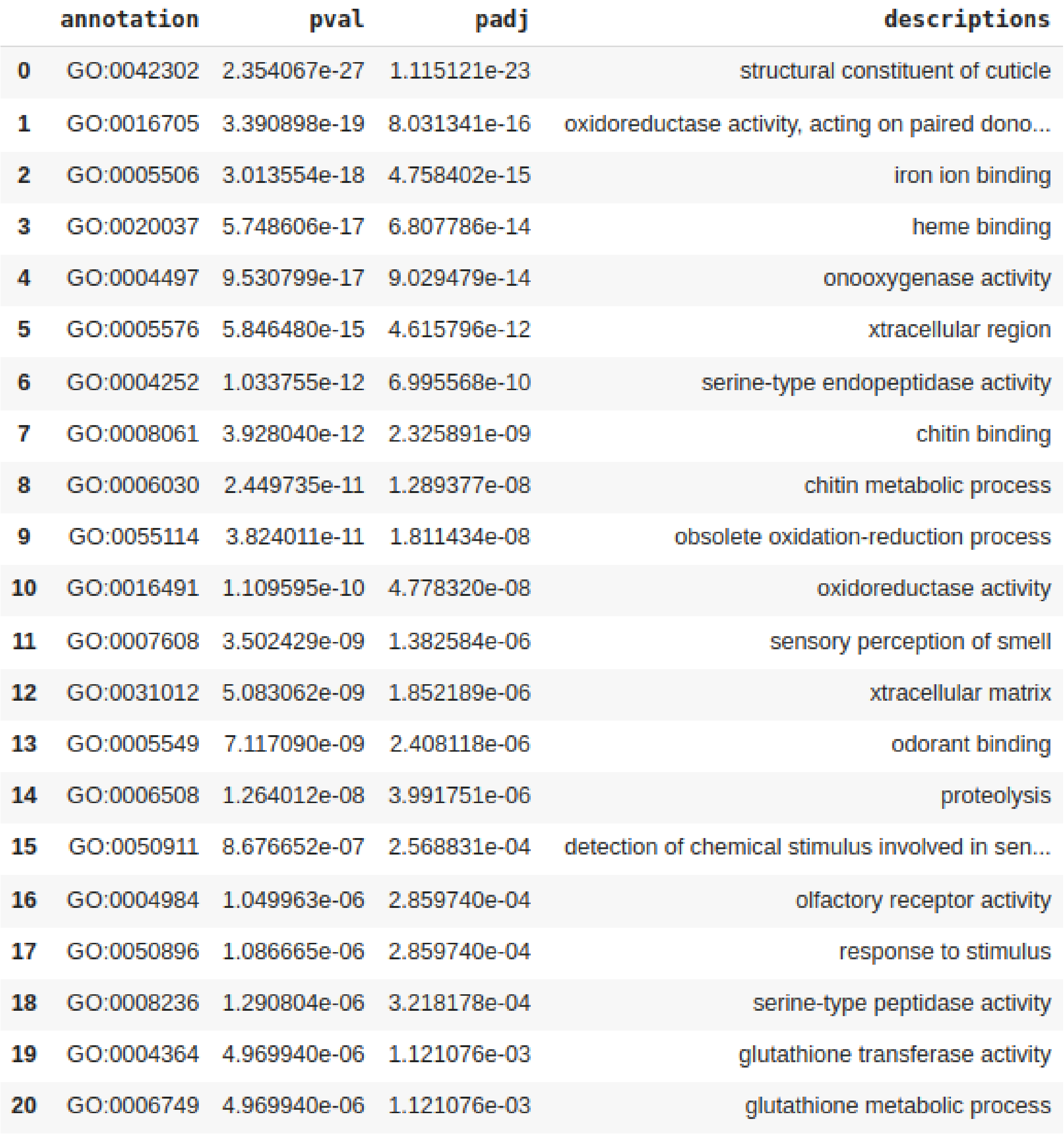


Figure 1. Gene expression plots of 35 meta-analysed RNA-Sequencing comparisons between insecticide susceptible and resistant mosquitoes.

In a separate notebook, users can produce the same gene expression plot but for any group of genes with a given GO term or PFAM domain, for example, plotting the expression of all olfactory receptors, or all cytochrome P450s.

Table 1 shows the results from the enrichment analysis, which is available as a notebook. The results are highly consistent with the types of genes reported in the literature to be involved in insecticide resistance, adding confidence to the meta-analysis results.

Table 1. Enrichment analyses with the hypergeometric test on the full ‘gamb_colu_arab_fun’ dataset.



	annotation	pval	padj	descriptions
0	GO:0042302	2.354067e-27	1.115121e-23	structural constituent of cuticle
1	GO:0016705	3.390898e-19	8.031341e-16	oxidoreductase activity, acting on paired dono...
2	GO:0005506	3.013554e-18	4.758402e-15	iron ion binding
3	GO:0020037	5.748606e-17	6.807786e-14	heme binding
4	GO:0004497	9.530799e-17	9.029479e-14	onooxygenase activity
5	GO:0005576	5.846480e-15	4.615796e-12	xtracellular region
6	GO:0004252	1.033755e-12	6.995568e-10	serine-type endopeptidase activity
7	GO:0008061	3.928040e-12	2.325891e-09	chitin binding
8	GO:0006030	2.449735e-11	1.289377e-08	chitin metabolic process
9	GO:0055114	3.824011e-11	1.811434e-08	obsolete oxidation-reduction process
10	GO:0016491	1.109595e-10	4.778320e-08	oxidoreductase activity
11	GO:0007608	3.502429e-09	1.382584e-06	sensory perception of smell
12	GO:0031012	5.083062e-09	1.852189e-06	xtracellular matrix
13	GO:0005549	7.117090e-09	2.408118e-06	odorant binding
14	GO:0006508	1.264012e-08	3.991751e-06	proteolysis
15	GO:0050911	8.676652e-07	2.568831e-04	detection of chemical stimulus involved in sen...
16	GO:0004984	1.049963e-06	2.859740e-04	olfactory receptor activity
17	GO:0050896	1.086665e-06	2.859740e-04	response to stimulus
18	GO:0008236	1.290804e-06	3.218178e-04	serine-type peptidase activity
19	GO:0004364	4.969940e-06	1.121076e-03	glutathione transferase activity
20	GO:0006749	4.969940e-06	1.121076e-03	glutathione metabolic process

In another notebook, users may rank genes from each analysis based on their median or mean expression, to identify candidate genes involved in insecticide resistance. The results of this analysis will be discussed elsewhere. Lastly, users can plot heatmaps of fold-change data for any gene or gene family of interest. Figure 2 shows a plot for all P450s with greater than 1.8 fold expression.

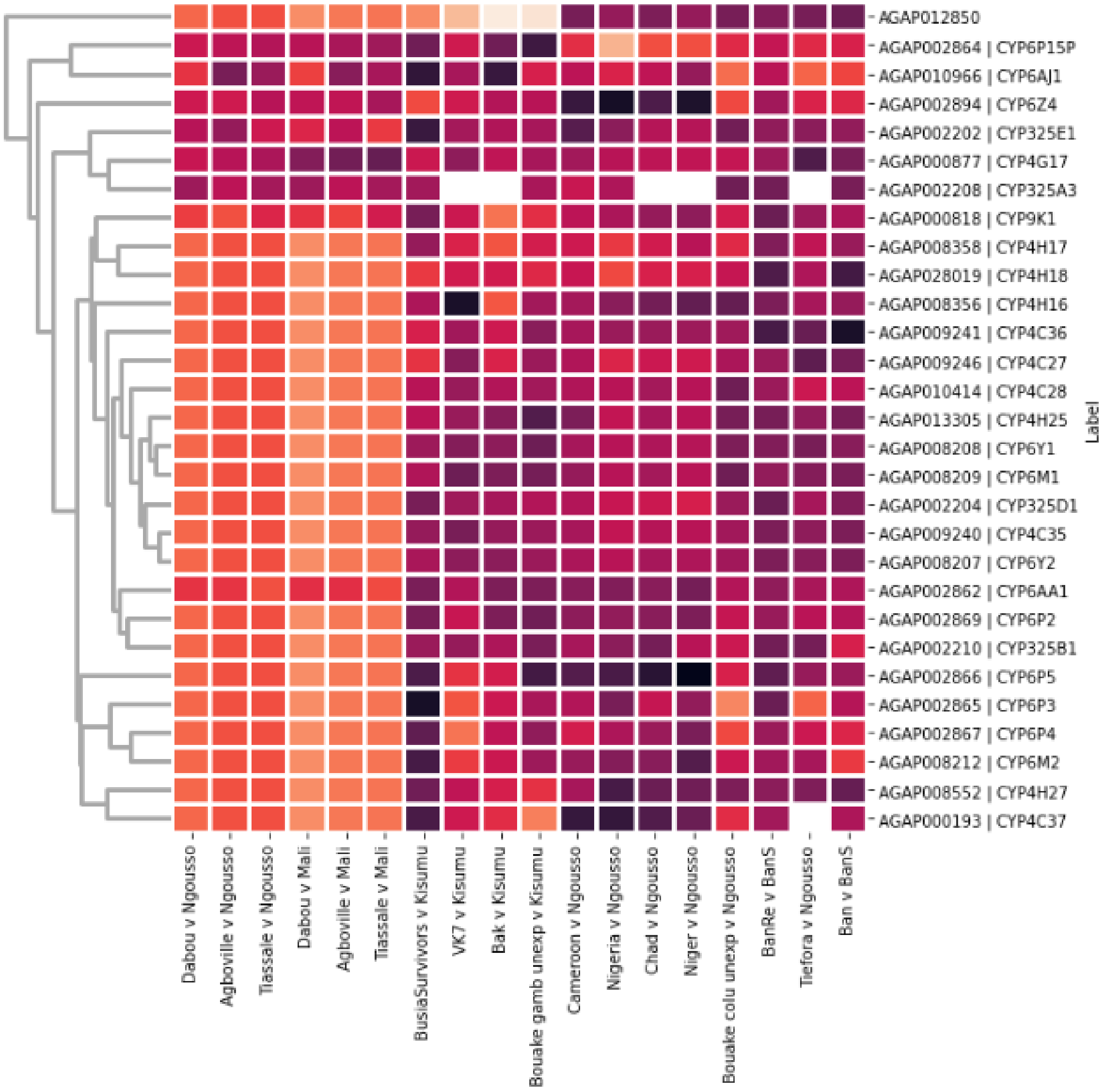


Figure 2. Heatmaps and clustering of cytochrome P450s with greater than 1.8 median fold-change.

In order to allow full integration with other python packages for *Anopheles* genomics research, such as *malariagen_data* and *AgamPrimer*, I will develop the functions in *AnoExpress* into a python package. This will allow us to build functionality further, and build methods in *malariagen_data* or *AnoExpress* which fully integrate genomic and transcriptomic data, to enhance *Anopheles* insecticide resistance research. I hope that *AnoExpress* will be a useful tool for the community.

7.5 References

David, J.-P., Strode, C., Vontas, J., Nikou, D., Vaughan, A., Pignatelli, P.M., Louis, C., Hemingway, J. and Ranson, H. (2005) 'The Anopheles gambiae detoxification chip: a highly specific microarray to study metabolic-based insecticide resistance in malaria vectors', *Proceedings of the National Academy of Sciences of the United States of America*, 102(11), pp. 4080–4084.

Ingham, V., Wagstaff, S. and Ranson, H. (2018) 'Transcriptomic meta-signatures identified in Anopheles gambiae populations reveal previously undetected insecticide resistance mechanisms', *Nature communications* [Preprint]. doi:10.1038/s41467-018-07615-x.