Genomic epidemiology of malaria vectors in the *Anopheles gambiae* species complex

A thesis submitted for the degree of Doctor of Philosophy

Alistair Miles

Linacre College

University of Oxford

15th September 2021

Abstract

In this thesis I ask, how can the study of genome variation within malaria vector populations contribute to the control of malaria in sub-Saharan Africa.

In the **first chapter** I provide an introduction to the current situation in malaria control in sub-Saharan Africa, and the role played by large-scale mosquito control programmes using insecticide-based interventions. I also introduce high-throughput whole-genome sequencing and its potential applications to the study and surveillance of malaria vectors.

In the **second chapter** I introduce the *Anopheles gambiae* species complex, and provide historical context by describing how the species complex was discovered, which marked the introduction of genetic methods into the study and surveillance of African malaria vectors. I conclude that there are important parallels between past and present efforts towards malaria elimination, but also new opportunities afforded by genomic epidemiology.

In the **third chapter** I describe the production of a genome variation data resource derived from whole-genome sequencing of *Anopheles gambiae* and *Anopheles coluzzii* mosquitoes from 8 African countries, carried out as part of the first phase of the *Anopheles gambiae* 1000 Genomes (Ag1000G) Project. This chapter establishes and validates methods for robust discovery of nucleotide variation from Illumina deep whole-genome sequencing of individual mosquitoes, and confirms that *Anopheles* mosquitoes are among the most genetically diverse organisms in the natural world. Subsequent chapters all perform analyses using this data resource.

In the **fourth chapter** I identify genetically distinct populations among the mosquitoes sampled in Ag1000G phase 1, and quantify genetic diversity within and differentiation between these populations. I show that there is strong population structure and marked differences in diversity between populations, suggesting important heterogeneities in popu-

lation size and rates of gene flow. These results are an essential foundation on which to build analyses of recent evolution in subsequent chapters.

In the **fifth chapter** I search for signals of recent positive selection among the populations sampled in Ag1000G phase 1, to identify which genes are most important in generating an adaptive response to the use of insecticides in malaria vector control. I show that there are strong signals of recent selection both at known insecticide resistance genes and at previously unknown genes with a plausible link to insecticide resistance.

In the **sixth chapter** I perform a detailed analysis of the voltage-gated sodium channel gene, where genetic changes cause target-site resistance to pyrethroid insecticides, the main ingredient in insecticide-treated bednets. I identify previously unknown mutations within this gene, and use haplotype data to show that resistance mutations have spread over large geographical distances and between mosquito species.

In the **final chapter**, I discuss the potential translation of genome sequencing into operational malaria vector surveillance and insecticide resistance management systems.

Acknowledgments

The work described in this thesis was carried out in the context of a broader collaboration involving the members of the *Anopheles gambiae* 1000 Genomes Consortium and the MalariaGEN Resource Centre team. I would like to thank everyone involved in this collaboration for patiently educating me in the field of malaria vector biology, and for many enlightening conversations and discussions of analytical approaches and results. Chapters 3, 4, 5 and 6 include content published in The Anopheles gambiae 1000 Genomes Consortium (2017). Chapter 6 also includes content published in Clarkson et al. (2018). All the content included in this thesis was written/produced entirely by myself, except where specifically indicated within the corresponding section.

I would like to thank my supervisor Dominic Kwiatkowski for the unique opportunity to work on the *Anopheles gambiae* 1000 Genomes Project, for the continuous support and faith, and for sharing a wealth of experience and perspective. I would like to thank Victoria Cornelius, whose unfailing enthusiasm for this thesis helped to sustain me on many occasions. I would like to thank my wife Laura and my three daughters Eva, Lola and Ruby, two of whom arrived while this thesis was under construction, for travelling with me on this journey, and for their love and support.

Source code files for this thesis are available from GitHub[1].

References

The Anopheles gambiae 1000 Genomes Consortium (2017). 'Genetic diversity of the African malaria vector Anopheles gambiae'. In: *Nature* 552.7683, pp. 96–100. DOI:

[1] `https://github.com/alimanfoo/dphil`

10.1038/nature24995

CS Clarkson et al. (2018). 'The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors Anopheles gambiae and Anopheles coluzzii'. In: *bioRxiv*. DOI: 10.1101/323980

Contents

1 General introduction

In this chapter I provide an overview of the present situation regarding malaria control in sub-Saharan Africa, the critical role played by large-scale mosquito control programmes, and the current challenges of insecticide resistance. I then introduce whole-genome sequencing and describe its potential role in malaria mosquito surveillance and accelerating efforts towards the development of new mosquito control tools. These factors provided the background and motivation for the establishment of the Anopheles gambiae 1000 Genomes (Ag1000G) Project, an international collaboration to study genetic variation among malaria-transmitting mosquitoes collected from natural populations across sub-Saharan Africa.

Malaria vector control in sub-Saharan Africa

Malaria is an infectious disease caused by eukaryotic parasites of the genus *Plasmodium* and transmitted by blood-feeding mosquitoes of the genus *Anopheles*. The vast majority of malaria occurs in sub-Saharan Africa, with more than 200 million cases and 400,000 deaths annually (WHO, 2019). The most effective methods of controlling malaria prevent disease transmission by targeting the mosquito vector (malaria vector control). Two methods of malaria vector control are widely used in public health programmes in Africa: mass distribution of long-lasting insecticidal bednets (LLINs) (Carnevale and Gay, 2019; Okumu, 2020) and indoor residual spraying of houses with insecticides (IRS) (WHO, 2006; Pluess et al., 2010; Choi et al., 2019). At the turn of the millennium, a concerted international campaign and partnership was launched with the goal of dramatically increasing the scale and scope of malaria vector control programmes in sub-Saharan Africa (Nabarro and Tayler, 1998). Since then, more than 2 billion LLINs have been distributed in total, exceeding 100

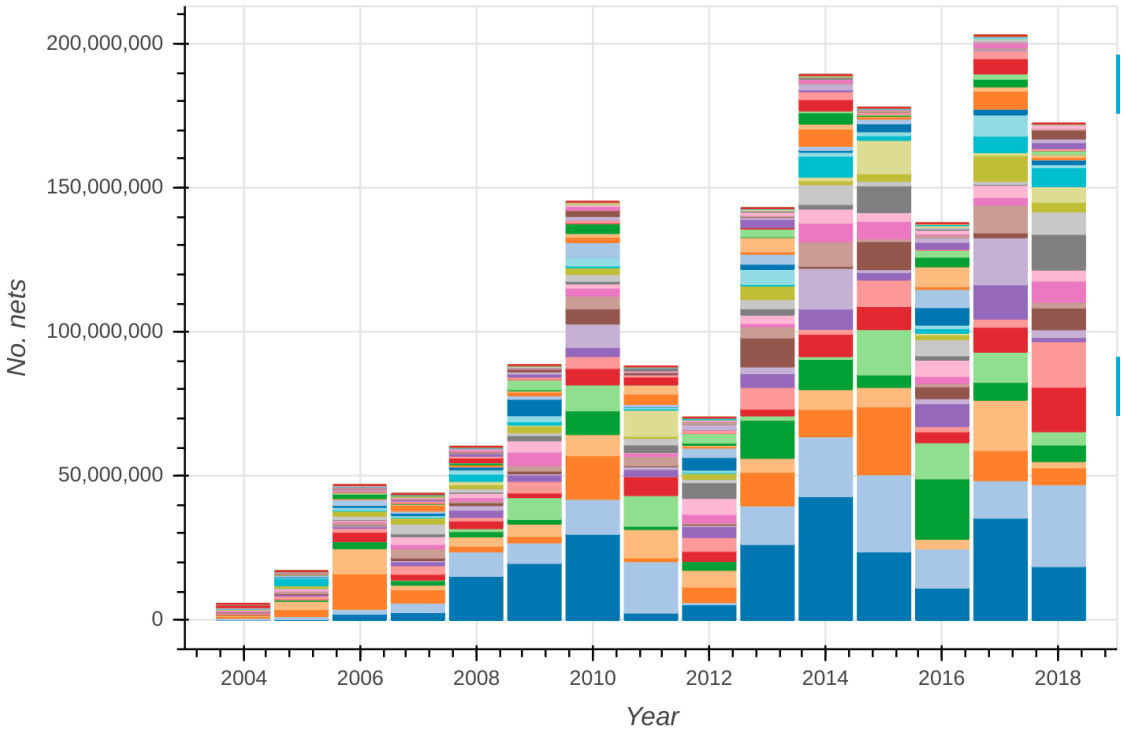

Figure 1.1. Long-lasting insecticidal bednets (LLINs) distributed in sub-Saharan Africa. Bar colours denote countries, ordered vertically by the total number of nets distributed since 2000, largest at the bottom. The eight countries with the largest numbers of nets are: Nigeria (dark blue); Democratic Republic of the Congo (light blue); Ethiopia (dark orange); Kenya (light orange); Uganda (dark green); Tanzania (light green); Ghana (dark red); Madagascar (light red). Data from AMP (2020).

million LLINs annually (AMP, 2020) (Fig. 1.1) with more than 40% of the population at risk sleeping under an LLIN from 2014 onwards (Bhatt et al., 2015; WHO, 2019). IRS programmes are more targeted and reserved for focal areas of higher malaria transmission, but are nevertheless substantial, protecting more than 10% of the population at risk in 2010, although declining to lower levels since (Bhatt et al., 2015; WHO, 2019; Tangena et al., 2020). The impact of these vector control programmes has been substantial. The prevalence of *Plasmodium falciparum* malaria infections in Africa halved between 2000 and 2015, averting approximately 663 million clinical cases, of which 68% can be attributed to LLIN and IRS interventions (Bhatt et al., 2015).

The threat of insecticide resistance

All LLINs approved for public health use by the World Health Organization (WHO) use a pyrethroid insecticide (Elliott, 1989) as the primary active ingredient (WHO, 2020). The majority of countries with IRS programmes have also primarily used pyrethroid insecticides (WHO, 2019; Tangena et al., 2020). The heavy reliance on a single insecticide class has inevitably led to a substantial increase in the prevalence of resistance to pyrethroid insecticides among African malaria vectors (Hemingway et al., 2016). Pyrethroid resistance is conventionally monitored via standardised bioassays, and resistance is detected when the proportion of mosquitoes killed by a diagnostic dose falls below a standard threshold (WHO, 2018). Many countries routinely perform these insecticide resistance bioassays as part of entomological monitoring programmes. When these bioassay data have been aggregated, there is a clear trend of increasing pyrethroid resistance since 2000 within the major African malaria vector species (Hancock et al., 2020). This trend is particularly acute in West Africa where, for example, it is estimated that more than 80% of the region has resistance to the pyrethroid deltamethrin since 2010.

Resistance to pyrethroid insecticides in malaria vectors clearly has the potential to reduce the efficacy of LLIN and IRS programmes. In practice, however, it has not been straightforward to demonstrate or measure this reduction in efficacy. A WHO-coordinated observational study in four African countries and India found no evidence that pyrethroid resistance had any effect on malaria infection prevalence or disease incidence in areas using pyrethroid LLINs as the primary intervention (Kleinschmidt et al., 2018). On the other hand, several studies have found that switching IRS programmes from pyrethroids to a different insecticide class has led to a significant reduction in disease (Hargreaves et al., 2000; Kafy et al., 2017). Using non-pyrethroid IRS in addition to pyrethroid LLINs, or using newer LLINs that combine a pyrethroid with the synergist piperonyl butoxide (PBO) that counteracts some forms of pyrethroid resistance, have also been shown to reduce disease relative to pyrethroid-only LLINs alone (Protopopoff et al., 2018). Mathematical modelling of aggregated data from bioassays and experimental hut trials has also provided evidence that non-pyrethroid IRS substantially reduces malaria relative

to pyrethroid IRS (Sherrard-Smith et al., 2018) and that PBO LLINs are more effective than pyrethroid-only LLINs in areas with pyrethroid resistance (Churcher et al., 2016). There is a growing consensus that pyrethroid resistance now poses a significant threat to malaria vector control programmes, and is at least in part responsible for the fact that malaria prevalence has not substantially reduced in sub-Saharan Africa since 2014 and has increased in some countries (Hemingway et al., 2016; WHO, 2019).

Insecticide resistance management

Recognising this trend of rising pyrethroid resistance and anticipating the threat to malaria vector control, the WHO developed a global plan for insecticide resistance management (GPIRM) (WHO, 2012). The ultimate goal of insecticide resistance management (IRM) is to maintain the effectiveness of insecticides by preventing or delaying the evolution and spread of resistance mutations in mosquito populations. IRM principles and strategies have been developed and established in agricultural pest management, and these have largely formed the basis for malaria vector IRM recommendations (Georghiou, 2005; Sternberg and Thomas, 2018). In general, there are three main IRM strategies:

1. **Management by moderation**. Reducing insecticide use by using them in conjunction with a range of non-insecticide-based control measures. For malaria vectors, this is also known as integrated vector management (IVM), and includes measures such as larval source management.

2. **Management by saturation**. Ensure that insecticides are delivered in a way that overwhelms the insect's natural defenses. For malaria vectors, this includes using synergists such as PBO that inhibit resistance mechanisms within LLINs, and IRS formulations that use microencapsulation to increase the dose received on contact.

3. **Management by multiple attack**. Employ multiple insecticides with different modes of action. Specific strategies available for malaria vectors include LLINs impregnated with two different insecticides, IRS formulations that are mixtures of two different insecticides, and IRS programmes that regularly rotate treatments

between different insecticides. This can also include deployment of both LLINs and IRS in the same location using different insecticides.

These strategies form pillar I of the GPIRM.

When the GPIRM was published in 2012, there were practical limitations on the range of possible IRM strategies for malaria vector control, because the only LLIN products available all used a single pyrethroid active ingredient, and a limited range of IRS products were available spanning four insecticide classes and two modes of action. More recently, a "next-generation" of dual active ingredient LLIN products have been approved for public health use, including PBO LLINs (Gleave et al., 2018) and LLINs that combine a pyrethroid with a second insecticide with a different mode of action (Bayili et al., 2017; Tiono et al., 2018). A "next-generation" of IRS products have also become available, including a microencapsulation formulation of the organophosphate insecticide pyrimiphos methyl (Oxborough et al., 2014), and products using the neonicotinoid insecticide clothianidin not previously used in public health (Oxborough et al., 2019). These new products have opened up new possibilities for malaria vector IRM. Malaria control programmes began rotating IRS programmes to use next-generation IRS products from 2016 (Tangena et al., 2020) and several initiatives are working to accelerate the deployment of next-generation LLINs[1,2]. These changes represent the most significant upheaval in malaria vector control strategy since the turn of the millennium, and present both new opportunities and new challenges. In particular, the use of new insecticides will inevitably exert new selective pressures on malaria vector populations, and the emergence and spread of new forms of resistance is a major concern. Also, next-generation LLIN and IRS products are more expensive and can be logistically more demanding (ten Brink et al., 2018). Thus, choices need to be made about where, when and how to deploy these new products to get maximum impact from limited resources without reducing the number of people protected from malaria infection (WHO, 2017).

[1] http://unitaid.org/call-for-proposal/catalyzing-market-introduction-next-generation-long-lasting-insecticidal-nets-llins/

[2] https://www.ivcc.com/market-access/new-nets-project/

Surveillance as a core intervention

Data are thus needed to guide decisions regarding optimal vector control strategies in different settings, to evaluate the success and impact of interventions and resistance management strategies, and to provide early warning of new evolutionary adaptations in response to the deployment of next-generation LLIN and IRS products. Recognising this need, the WHO Global Technical Strategy for Malaria 2016–2030 advocates that malaria surveillance should be transformed into a core intervention as one of three pillars of the strategy (WHO, 2015). Here "surveillance" covers a broad range of data gathering activities, encompassing both epidemiological and entomological variables, in addition to intervention coverage, available resources, trends in health service utilisation, and more. Vector surveillance, also known as entomological monitoring, is a critical component of this, gathering data on malaria vector populations. Vector surveillance programmes are established to a varying degree in most malaria-endemic countries and typically collect data on which malaria vector species are present, their abundance and seasonality, their behaviour in terms of time and location of biting and host preference, as well as the susceptibility of each vector species to different insecticides (Russell et al., 2020). The GPIRM emphasizes the need for insecticide resistance monitoring as pillar II of its action plan (WHO, 2012).

Current malaria vector surveillance programmes have two key limitations. The first is that surveillance programmes are not sufficiently resourced, and thus there is a lack of trained personnel and facilities (Russell et al., 2020). The second is that surveillance programmes generally do not have the capability to gather molecular data on malaria vector populations, and where they do the data are limited in resolution. These data gaps mean that a number of key operational questions cannot be effectively answered, including but not limited to the following:

1. **Detecting cryptic vector species**. Malaria vector species occur within cryptic species complexes, where morphological identification is not sufficient to resolve species identity (Davidson, 1964; Coetzee et al., 2013). Failure to differentiate known species or recognise the presence of previously unknown species means that other

data variables become muddled, because different species within the same complex may have different behaviours and/or insecticide resistance adaptations.

2. **Differentiating molecular mechanisms of insecticide resistance**. The presence of insecticide resistance can be detected by standard WHO or CDC bioassays, but these do not provide information regarding the underlying molecular mechanisms of resistance that are present in a given malaria vector population (WHO, 2018). Given that a number of new vector control products such as PBO LLINs are designed to target a specific mechanism of resistance, the absence of these data means there is no way to determine whether such a product should be deployed.

3. **Providing early warning of novel insecticide resistance adaptations**. There is limited ability to obtain early warning of novel insecticide resistance adaptations, such as those emerging in response to deployment of a new IRS or LLIN products, because resistance typically has to reach a relatively high frequency within a mosquito population before it becomes evident via conventional bioassays (Roush and Miller, 1986; Sternberg and Thomas, 2018).

4. **Monitoring of insecticide resistance allele frequencies**. IRM strategies such as IRS rotation depend on managing the frequency of insecticide resistance alleles within mosquito populations, switching products at the right time and using them for the right duration to ensure resistance does not reach fixation (South and Hastings, 2018). However, without any data on resistance allele frequencies, there is no way to know if IRM strategies are working as intended.

5. **Tracking the spread of insecticide resistance**. Mosquitoes move and have the capability to spread insecticide resistance over the course of multiple generations of movement and reproduction (Service, 1997; Huestis et al., 2019). Without genetic data, it is not possible to resolve whether resistance has originated locally or spread from elsewhere. Thus, it is hard to determine which interventions or conditions are driving the emergence of resistance, and whether IRM strategies can be designed locally or need to be coordinated nationally or even internationally to be effective.

While these limitations remain, it is difficult for malaria control programmes to adapt vector control strategies to different settings or respond to changing circumstances, and it is equally difficult for advisory agencies such as WHO to formulate clear criteria for policy change (WHO, 2017). An awareness of these gaps, coupled with the rapid advancement of high throughput technologies for genome sequencing and other molecular diagnostics, and the development and maturation of genomic epidemiology as a field of research, has driven an interest in the development of molecular methods and particularly genome sequencing for malaria vector surveillance.

New vector control tools are needed

Even if next-generation LLIN and IRS products are widely deployed, IRM strategies are implemented, and these are supported by improved vector surveillance capabilities, there is a consensus that this will not be sufficient to reach malaria elimination. New vector control tools will be required, and the need for expanding research to accelerate the development of new tools is recognised as pillar III of the GPIRM (WHO, 2012) and as supporting element 1 of the WHO Global Technical Strategy for Malaria 2016–2030 (WHO, 2015). This includes further repurposing of agricultural insecticides, as well as the development of new insecticides with novel modes of action designed specifically for public health use (Hemingway et al., 2006; Lees et al., 2019). This also includes the development of novel control tools that do not rely on insecticides, such as genetic control tools (Davidson, 1974; Burt, 2003). In particular, CRISPR/Cas9 gene editing technology had enabled the development of highly effective gene drives, which are selfish genetic elements with the capability to spread through mosquito populations via super-Mendelian inheritance and cause population suppression or modification (Burt, 2003; Kyrou et al., 2018). The development of these novel vector control tools has benefited greatly from and continues to rely upon high throughput molecular tools such as whole-genome sequencing, and the availability of high quality open access molecular data resources such as reference genome sequences for malaria vectors (Holt et al., 2002; Sharakhova et al., 2007; Lawniczak et al., 2010; Neafsey et al., 2014). Increasingly, the value of and need for data on natural

genetic variation in targeted mosquito populations is also becoming evident. For example, CRISPR/Cas9 gene drives need to target regions of low genetic diversity to minimise the evolution of resistance (Kyrou et al., 2018).

Next-generation sequencing as a tool for genomic epidemiology and infectious disease surveillance

High throughput "next-generation" genome sequencing (NGS) technologies have advanced rapidly in the last two decades, opening up new applications for the study and control of infectious diseases (Goodwin et al., 2016). The fundamental innovation of NGS is to perform sequencing of up to billions of small fragments of DNA in parallel. For example, in 2005 the Illumina Genome Analyzer was the first commercially available NGS instrument, capable of generating 1 gigabase (Gb) of data per run. The Illumina HiSeq instruments available from 2011 extended this capability by two orders of magnitude, generating up to 600 Gb per run (Illumina, 2017). These leaps in data generating capability, coupled with the corresponding reduction in per-unit sequencing costs, allow for the sequencing of many individuals of a given species, and thus the study of genetic variation within and between natural populations. The 1000 Genomes Project pioneered the use of NGS for the study of genetic variation, sequencing the genomes of 2,504 individuals from 26 human populations, and making the data openly available as a resource for the research community (The 1000 Genomes Project Consortium, 2015). Applied to infectious diseases, NGS allows pathogen genomes from many individual infections to be sequenced and compared, providing opportunities to discover genetic variation underlying important traits such as virulence or drug resistance (Armstrong et al., 2019). Comparing genome sequences also reveals patterns of relatedness between pathogens in different hosts, which can be used to investigate outbreaks of bacterial and viral diseases, or the transmission dynamics of endemic parasite diseases such as malaria (Robinson et al., 2013; Daniels et al., 2015; Wohl et al., 2016; Neafsey and Volkman, 2017; Wesolowski et al., 2018; Armstrong et al., 2019).

To promote the application of genomics to malaria research, the Malaria Genomic

Epidemiology Network (MalariaGEN)[3] was established in 2005, and provides a framework for equitable collaboration and data sharing between multiple research centres and public health laboratories with access to different sampling and sequencing capabilities. The first large-scale study of genome variation in malaria parasites carried out by MalariaGEN sequenced 227 samples from three continents and discovered 86 thousand polymorphisms within gene coding regions (Manske et al., 2012). Subsequent studies expanded this dataset and applied it to identify multiple populations of drug resistant parasites in Cambodia and discover genetic variants associated with antimalarial drug resistance (Miotto et al., 2013; Miotto et al., 2015). The most recent data release includes high quality genotype calls on 3 million single nucleotide polymorphisms (SNPs) and short insertion/deletions (indels) in 7,000 worldwide samples of *P. falciparum* infection (MalariaGEN et al., 2019). Given these successes with use of NGS to study infectious disease pathogens, it is natural to ask whether a similar approach could be used to study disease vectors such as *Anopheles* mosquitoes that transmit malaria.

Genomic epidemiology of malaria vectors: opportunities and challenges

The genome of the major Afrotropical malaria vector species, *Anopheles gambiae*, was sequenced in 2002 and updated in 2007 (Holt et al., 2002; Sharakhova et al., 2007). This reference sequence, scaffolded to complete chromosomes, spans a total of 230 Mb across the two autosomes and the X sex chromosome, with an additional 42.6 Mb in unplaced contigs. The availability of a high quality reference sequence makes possible the analysis of natural genome variation via NGS, because short reads from individual mosquitoes can be aligned to the reference genome and variants identified by comparison with the reference sequence. Many questions could be investigated using such data. For example, although the general mechanisms of insecticide resistance in malaria vectors are reasonably well established, the specific genes and genetic variants underlying these resistance mechanisms are for the most part unknown. Analogous to the discovery of drug resistance variants in malaria

[3] https://www.malariagen.net

parasites or antimicrobial resistance variants in bacteria, NGS could be used to discover and build a more complete picture of the genetic basis of insecticide resistance in malaria vectors (Donnelly et al., 2016). Analogous to the analysis of pathogen outbreaks, NGS could be used to analyse the geographical origins and movements of different insecticide resistance mutations between malaria vector populations. The analysis of genetic variation can also be used to make demographic inferences about populations, including patterns of migration and changes in population size. Of particular relevance to malaria vectors would be investigating demographic changes in response to vector control interventions. Many of the gaps in vector surveillance data described above could in principle be filled by NGS of malaria vectors. Open data resources of natural genetic variation could also accelerate many forms of research, including the research and development of new vector control tools.

Although there are parallels between genomic epidemiology of pathogens and vectors, there are also fundamental differences which present unique challenges, including:

1. **Genome size**. Whereas viral genomes are typically less than 50 kb, bacterial genomes less than 5 Mb, and eukaryotic pathogen genomes such as *Plasmodium* less than 30 Mb, *Anopheles* genomes are an order of magnitude larger at ~300 Mb in size (Neafsey et al., 2014).

2. **Genome complexity**. Viral genomes typically have 10–100 genes, bacteria 100–5,000 genes, *Anopheles gambiae* has 13,057 annotated protein-coding genes (VectorBase, 2019; Giraldo-Calderón et al., 2014). The *Anopheles gambiae* genome also has a diversity of repetitive elements including various classes of transposable element, not generally found in viruses, bacteria or eukaryotic parasites (Tu and Coates, 2004; Fernández-Medina et al., 2011).

3. **Sexual reproduction and frequent recombination**. Viral and bacterial pathogens are not sexually reproducing, and although some do undergo a form of recombination under certain circumstances, it is much less frequent than that which occurs in sexually reproducing eukaryotes. The recombination rate of *Anophe-*

les gambiae is estimated at 1 cM/Mb, approximately one recombination event per chromosome per generation (Pombi et al., 2006).

4. **Diploidy**. *Anopheles gambiae* is diploid and has a heterogametic sex determination system, in common with other dipteran insects, mammals and many other eukaryotic phyla. This means that sequencing the genomic DNA from a single individual provides information about both of the genome sequences present within that diploid individual.

5. **Large effective population size and high levels of genetic diversity**. Previous studies of diversity within *Anopheles* species based on sequencing of individual genes have found nucleotide diversity at synonymous coding sites in the range 0.4-2.9 % depending on species, with *Anopheles gambiae* being the highest among these (Leffler et al., 2012). This is indicative of very large effective population size and is at the extreme end of diversity found across the tree of life (Leffler et al., 2012).

These factors require that analytical methods need to be established and validated that are appropriate to malaria vectors, in order to provide a solid foundation for epidemiological inferences to be made.

The *Anopheles gambiae* 1000 Genomes Project

In 2013, the cost of Illumina whole-genome deep sequencing of individual *Anopheles* mosquitoes had reached the level at which sequencing upwards of 1000 individuals was a feasible goal. MalariaGEN, supported by the Malaria Programme at the Wellcome Trust Sanger Institute, established a project to sequence the genomes of mosquitoes in the *Anopheles gambiae* species complex, collected from natural populations across sub-Saharan Africa. To support this venture, a consortium was established, bringing together representatives from 20 different research institutions, including groups with expertise in field sampling, malaria vector biology and population genomics. The project was named "The *Anopheles gambiae* 1000 Genomes Project", abbreviated to "Ag1000G". The primary goal of the Ag1000G Project was to generate a high quality open access data resource

on natural genetic variation within *Anopheles gambiae* populations, and to make these data available to the research and public health communities. The Ag1000G Project also aimed to use these data to investigate the demography and evolution of *Anopheles gambiae* populations, particularly in relation to malaria vector control interventions. The ultimate aim was to establish a foundation comprising both a high quality data resource and a collection of proven analytical methods, that could then be used as a platform to develop and scale up NGS as an operational tool for malaria vector surveillance.

For practical purposes, the Ag1000G Project was divided into three phases:

- **Ag1000G phase 1** sequenced 765 individual mosquitoes sampled from natural populations in 8 African countries, with representation of two major malaria vector species within the *Anopheles gambiae complex*, *Anopheles gambiae* and *Anopheles coluzzii*. This phase focused on the genome-wide discovery and analysis of single nucleotide polymorphisms (SNPs) and SNP haplotypes. Primary analyses of those data were published in The Anopheles gambiae 1000 Genomes Consortium (2017).

- **Ag1000G phase 2** expanded the resource to include 1,142 wild-caught mosquitoes from 13 countries, with increased representation of *Anopheles coluzzii*. This phase also studied SNPs and added genome-wide discovery and analysis of copy number variation (CNV). Primary analyses were published recently in The Anopheles gambiae 1000 Genomes Consortium (2020).

- **Ag1000G phase 3** is expanding the resource to include 2,784 wild-caught mosquitoes from 19 countries, with representation of a third vector species, *Anopheles arabiensis*. Analyses of those data are ongoing, with data scheduled for release in 2021.

My role within the Ag1000G Consortium is to coordinate the project as a whole, and to lead data production, curation and analysis. Although this work has been carried out in collaboration with other members of the Ag1000G Consortium, there are specific areas where I made individual contributions, and those are the focus of this thesis. In particular, the majority of this thesis describes contributions I made during first project phase.

Structure of this thesis

In the next chapter I take a brief historical detour, in order to introduce the *Anopheles gambiae* species complex, which is the biological subject of this thesis. Whilst working on the Ag1000G project I was fortunate to come into possession of a collection of previously unpublished letters between entomologists George Davidson and Hugh Paterson, documenting their collaboration during the 1960s which led to the discovery of the *Anopheles gambiae* complex. Their discovery marked the introduction of genetic methods into malaria entomology, and their correspondence provides a unique historical perspective on present malaria elimination efforts, as well as the potential role for genomic epidemiology in future vector control programmes.

In the third chapter I describe the production of the Ag1000G phase 1 genome variation data resource, which comprises high quality data on 52 million single nucleotide polymorphisms (SNPs). The aim of the chapter is to establish robust methodologies for genome-wide discovery and genotyping of SNPs from Illumina high throughput deep sequencing of individual *Anopheles* mosquitoes, and to evaluate the accuracy and sensitivity of the data produced in Ag1000G phase 1 using those methods. The chapter concludes by describing the level of genetic diversity within the Ag1000G phase 1 cohort as a whole, and provides confirmation that *Anopheles gambiae* mosquitoes are indeed among the most genetically diverse organisms in the natural world.

In the fourth chapter I analyse the Ag1000G phase 1 data resource to investigate genetic population structure among wild-caught mosquitoes. I also compare genetic diversity within these populations and quantify genetic differentiation between them. Ag1000G phase 1 sampled mosquito populations from a broad geographical range, crossing the entire span of continental Africa from coast to coast, and there are many contrasts between the ecosystems inhabited by *Anopheles gambiae* mosquitoes across this range, including coastal to inland locations, deep forest to savannah, and urban to rural settings. The aim of the chapter is to uncover heterogeneities among the mosquito populations sampled across these diverse settings, and to begin to form an understanding of how major geographical and ecological features have influenced the shape, size and connectivity of contemporary

populations.

In the fifth chapter I continue the analysis of the Ag1000G phase 1 resource, performing genome-wide scans for signals of recent positive selection. The aim of the chapter is to identify genes experiencing the most extreme selection pressures, and thus which are likely to be involved in the adaptive response to the scale-up of malaria vector control interventions within the last two decades. I discover strong signals of selection at six functionally-validated insecticide resistance genes, providing confirmation that these genes are the central drivers of insecticide resistance in these mosquito populations. I also describe the design and implementation of a web resource to allow exploration and fine-mapping of novel selection signals, and illustrate its use to investigate a novel selection signal at a diacylglycerol kinase gene, a novel candidate with potential links to insecticide resistance.

In the sixth chapter I focus on a single gene encoding the voltage-gated sodium channel protein, which is the physiological binding target of pyrethroid insecticides, and where amino acid substitutions have been shown to cause target-site insensitivity to pyrethroids. I describe the discovery of novel non-synonymous substitutions within the *Vgsc* gene, and present population-genetic evidence that these substitutions are unlikely to be neutral and should be studied further for evidence of an insecticide resistance phenotype. I also use haplotype data from Ag1000G phase 1 to analyse the genetic backgrounds on which known resistance alleles are found, and make inferences about the spread of resistance alleles between different countries and mosquito species.

In the final chapter I look forward to future applications of genome sequencing within operational vector surveillance programmes, exploring the potential pathway to translation.

References

AMP (2020). *Net Mapping Project, Current ITN Global Delivery Quarterly Report, Q2 2020*. Tech. rep. Alliance for Malaria Prevention.

Armstrong, GL, DR MacCannell, J Taylor, HA Carleton, EB Neuhaus, RS Bradbury, JE Posey and M Gwinn (2019). 'Pathogen Genomics in Public Health'. In: *N. Engl. J. Med.* 381.26, pp. 2569–2580. DOI: 10.1056/nejmsr1813907.

Bayili, K, S N'do, M Namountougou, R Sanou, A Ouattara, RK Dabiré, AG Ouédraogo, D Malone and A Diabaté (2017). 'Evaluation of efficacy of Interceptor® G2, a long-lasting insecticide net coated with a mixture of chlorfenapyr and alpha-cypermethrin, against pyrethroid resistant Anopheles gambiae s.l. in Burkina Faso'. In: *Malar. J.* 16.1. DOI: 10.1186/s12936-017-1846-4.

Bhatt, S et al. (2015). 'The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015'. In: *Nature* 526.7572, pp. 207–211. DOI: 10.1038/nature15535.

Burt, A (2003). 'Site-specific selfish genes as tools for the control and genetic engineering of natural populations'. In: *Proc. R. Soc. B Biol. Sci.* 270.1518, pp. 921–928. DOI: 10.1098/rspb.2002.2319.

Carnevale, P and F Gay (2019). 'Insecticide-treated mosquito nets'. In: *Malaria Control and Elimination*. Ed. by F Ariey, F Gay and R Ménard. New York: Springer, pp. 221–232. DOI: 10.1007/978-1-4939-9550-9_16.

Choi, L, J Pryce and P Garner (2019). 'Indoor residual spraying for preventing malaria in communities using insecticide-treated nets'. In: *Cochrane Database Syst. Rev.* DOI: 10.1002/14651858.CD012688.pub2.

Churcher, TS, N Lissenden, JT Griffin, E Worrall and H Ranson (2016). 'The impact of pyrethroid resistance on the efficacy and effectiveness of bednets for malaria control in Africa'. In: *Elife* 5. DOI: 10.7554/eLife.16090.

Coetzee, M, RH Hunt, R Wilkerson, AD Torre, MB Coulibaly and NJ Besansky (2013). 'Anopheles coluzzii and anopheles amharicus, new members of the anopheles gambiae complex'. In: *Zootaxa* 3619.3, pp. 246–274. DOI: 10.11646/zootaxa.3619.3.2.

Daniels, RF et al. (2015). 'Modeling malaria genomics reveals transmission decline and rebound in Senegal'. In: *Proc. Natl. Acad. Sci. U. S. A.* 112.22, pp. 7067–7072. DOI: 10.1073/pnas.1505691112.

Davidson, G (1964). 'Anopheles gambiae, a complex of species'. In: *Bull. World Health Organ.* 31.5, pp. 625–634.

Davidson, G (1974). *Genetic Control of Insect Pests*. 1st ed. London and New York: Academic Press.

Donnelly, MJ, AT Isaacs and D Weetman (2016). 'Identification, Validation, and Application of Molecular Diagnostics for Insecticide Resistance in Malaria Vectors'. In: *Trends Parasitol.* 32.3, pp. 197–206. DOI: 10.1016/j.pt.2015.12.001.

Elliott, M (1989). 'The pyrethroids: Early discovery, recent advances and the future'. In: *Pestic. Sci.* 27.4, pp. 337–351. DOI: 10.1002/ps.2780270403.

Fernández-Medina, RD, CJ Struchiner and JM Ribeiro (2011). 'Novel transposable elements from Anopheles gambiae'. In: *BMC Genomics* 12.1. DOI: 10.1186/1471-2164-12-260.

Georghiou, G (2005). 'Principles of insecticide resistance management'. In: *Phytoprotection* 75.4, pp. 51–59. DOI: 10.7202/706071ar.

Giraldo-Calderón, GI et al. (2014). 'VectorBase: An updated Bioinformatics Resource for invertebrate vectors and other organisms related with human diseases'. In: *Nucleic Acids Res.* 43.D1, pp. D707–D713. DOI: 10.1093/nar/gku1117.

Gleave, K, N Lissenden, M Richardson, L Choi and H Ranson (2018). 'Piperonyl butoxide (PBO) combined with pyrethroids in insecticide-treated nets to prevent malaria in Africa'. In: *Cochrane Database Syst. Rev.* DOI: 10.1002/14651858.CD012776.pub2.

Goodwin, S, JD McPherson and WR McCombie (2016). 'Coming of age: Ten years of next-generation sequencing technologies'. In: *Nat. Rev. Genet.* 17.6, pp. 333–351. DOI: 10.1038/nrg.2016.49.

Hancock, PA, CJ Hendriks, JA Tangena, H Gibson, J Hemingway, M Coleman, PW Gething, E Cameron, S Bhatt and CL Moyes (2020). 'Mapping trends in insecticide resistance phenotypes in African malaria vectors'. In: *PLoS Biol.* 18.6. Ed. by AF Read, e3000633. DOI: 10.1371/journal.pbio.3000633.

Hargreaves, K, LL Koekemoer, BD Brooke, RH Hunt, J Mthembu and M Coetzee (2000). 'Anopheles funestus resistant to pyrethroid insecticides in South Africa'. In: *Med. Vet. Entomol.* 14.2, pp. 181–189. DOI: 10.1046/j.1365-2915.2000.00234.x.

Hemingway, J, BJ Beaty, M Rowland, TW Scott and BL Sharp (2006). 'The Innovative Vector Control Consortium: improved control of mosquito-borne diseases'. In: *Trends Parasitol.* 22.7, pp. 308–312. DOI: 10.1016/j.pt.2006.05.003.

Hemingway, J et al. (2016). 'Averting a malaria disaster: Will insecticide resistance derail malaria control?' In: *The Lancet* 387.10029, pp. 1785–1788. DOI: 10.1016/S0140-6736(15)00417-1.

Holt, RA et al. (2002). 'The genome sequence of the malaria mosquito Anopheles gambiae'. In: *Science* 298.5591, pp. 129–149. DOI: 10.1126/science.1076181.

Huestis, DL et al. (2019). 'Windborne long-distance migration of malaria mosquitoes in the Sahel'. In: *Nature* 574.7778, pp. 404–408. DOI: 10.1038/s41586-019-1622-4.

Illumina (2017). *An introduction to Next-Generation Sequencing Technology*. Tech. rep. Illumina.

Kafy, HT et al. (2017). 'Impact of insecticide resistance in Anopheles arabiensis on malaria incidence and prevalence in Sudan and the costs of mitigation'. In: *Proc. Natl. Acad. Sci. U. S. A.* 114.52, E11267–E11275. DOI: 10.1073/pnas.1713814114.

Kleinschmidt, I et al. (2018). 'Implications of insecticide resistance for malaria vector control with long-lasting insecticidal nets: a WHO-coordinated, prospective, international, observational cohort study'. In: *Lancet Infect. Dis.* 18.6, pp. 640–649. DOI: 10.1016/S1473-3099(18)30172-5.

Kyrou, K, AM Hammond, R Galizi, N Kranjc, A Burt, AK Beaghton, T Nolan and A Crisanti (2018). 'A CRISPR–Cas9 gene drive targeting doublesex causes complete population suppression in caged Anopheles gambiae mosquitoes'. In: *Nat. Biotechnol.* 36.11, pp. 1062–1066. DOI: 10.1038/nbt.4245.

Lawniczak, MKN et al. (2010). 'Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences'. In: *Science* 330.6003, pp. 512–514. DOI: 10.1126/science.1195755.

Lees, R, G Praulins, R Davies, F Brown, G Parsons, A White, H Ranson, G Small and D Malone (2019). 'A testing cascade to identify repurposed insecticides for next-generation vector control tools: Screening a panel of chemistries with novel modes of action against a malaria vector'. In: *Gates Open Res.* 3, p. 1464. DOI: 10.12688/gatesopenres.12957.2.

Leffler, EM, K Bullaughey, DR Matute, WK Meyer, L Ségurel, A Venkat, P Andolfatto and M Przeworski (2012). 'Revisiting an Old Riddle: What Determines Genetic Diversity

Levels within Species?' In: *PLoS Biol.* 10.9, e1001388. DOI: 10.1371/journal.pbio.1001388.

MalariaGEN et al. (2019). 'An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples'. In: *bioRxiv*. DOI: 10.1101/824730.

Manske, M et al. (2012). 'Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing'. In: *Nature* 487.7407, pp. 375–379. DOI: 10.1038/nature11174.

Miotto, O et al. (2013). 'Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia'. In: *Nat. Genet.* 45.6, pp. 648–655. DOI: 10.1038/ng.2624.

Miotto, O et al. (2015). 'Genetic architecture of artemisinin-resistant Plasmodium falciparum'. In: *Nat. Genet.* 47.3, pp. 226–234. DOI: 10.1038/ng.3189.

Nabarro, DN and EM Tayler (1998). 'The 'roll back malaria' campaign'. In: *Science* 280.5372, pp. 2067–2068. DOI: 10.1126/science.280.5372.2067.

Neafsey, DE and SK Volkman (2017). 'Malaria Genomics in the Era of Eradication'. In: *Cold Spring Harb. Perspect. Med.* 7.8, a025544. DOI: 10.1101/cshperspect.a025544.

Neafsey, DE et al. (2014). 'Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes'. In: *Science* 347.6217, p. 1258522. DOI: 10.1126/science.1258522.

Okumu, F (2020). 'The fabric of life: What if mosquito nets were durable and widely available but insecticide-free?' In: *Malar. J.* 19.1. DOI: 10.1186/s12936-020-03321-6.

Oxborough, RM, J Kitau, R Jones, E Feston, J Matowo, FW Mosha and MW Rowland (2014). 'Long-lasting control of Anopheles arabiensis by a single spray application of micro-encapsulated pirimiphos-methyl (Actellic® 300 CS)'. In: *Malar. J.* 13.1. DOI: 10.1186/1475-2875-13-37.

Oxborough, RM et al. (2019). 'Susceptibility testing of Anopheles malaria vectors with the neonicotinoid insecticide clothianidin; Results from 16 African countries, in preparation for indoor residual spraying with new insecticide formulations'. In: *Malar. J.* 18.1. DOI: 10.1186/s12936-019-2888-6.

Pluess, B, FC Tanser, C Lengeler and BL Sharp (2010). 'Indoor residual spraying for preventing malaria'. In: *Cochrane Database Syst. Rev.* DOI: 10.1002/14651858.cd006657.pub2.

Pombi, M, AD Stump, A Della Torre and NJ Besansky (2006). 'Variation in recombination rate across the X chromosome of Anopheles gambiae'. In: *Am. J. Trop. Med. Hyg.* 75.5, pp. 901–903. DOI: 10.4269/ajtmh.2006.75.901.

Protopopoff, N et al. (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, two-by-two factorial design trial'. In: *The Lancet* 391.10130, pp. 1577–1588. DOI: 10.1016/S0140-6736(18)30427-6.

Robinson, ER, TM Walker and MJ Pallen (2013). 'Genomics and outbreak investigation: From sequence to consequence'. In: *Genome Med.* 5.4. DOI: 10.1186/gm440.

Roush, RT and GL Miller (1986). 'Considerations for Design of Insecticide Resistance Monitoring Programs'. In: *J. Econ. Entomol.* 79.2, pp. 293–298. DOI: 10.1093/jee/79.2.293.

Russell, TL, R Farlow, M Min, E Espino, A Mnzava and TR Burkot (2020). 'Capacity of National Malaria Control Programmes to implement vector surveillance: a global analysis'. In: *Malar. J.* 19.1, p. 422. DOI: 10.1186/s12936-020-03493-1.

Service, MW (1997). 'Mosquito (Diptera: Culicidae) Dispersal - The Long and Short of It'. In: *J. Med. Entomol.* 34.6, pp. 579–588. DOI: 10.1093/jmedent/34.6.579.

Sharakhova, MV, MP Hammond, NF Lobo, J Krzywinski, MF Unger, ME Hillenmeyer, RV Bruggner, E Birney and FH Collins (2007). 'Update of the Anopheles gambiae pest genome assembly'. In: *Genome Biol.* 8.1, R5. DOI: 10.1186/gb-2007-8-1-r5.

Sherrard-Smith, E et al. (2018). 'Systematic review of indoor residual spray efficacy and effectiveness against Plasmodium falciparum in Africa'. In: *Nat. Commun.* 9.1. DOI: 10.1038/s41467-018-07357-w.

South, A and IM Hastings (2018). 'Insecticide resistance evolution with mixtures and sequences: A model-based explanation'. In: *Malar. J.* 17.1. DOI: 10.1186/s12936-018-2203-y.

Sternberg, ED and MB Thomas (2018). 'Insights from agriculture for the management of insecticide resistance in disease vectors'. In: *Evol. Appl.* 11.4, pp. 404–414. DOI: 10.1111/eva.12501.

Tangena, JAA et al. (2020). 'Indoor residual spraying for malaria control in sub-Saharan Africa 1997 to 2017: An adjusted retrospective analysis'. In: *Malar. J.* 19.1. DOI: 10.1186/s12936-020-03216-6.

ten Brink, D, M Gad and F Ruiz (2018). 'Malaria innovations: pursuing value in an evolving market'. In: *Lancet Glob. Heal.* 6.2, e138–e139. DOI: 10.1016/s2214-109x(17)30495-3.

The 1000 Genomes Project Consortium (2015). 'A global reference for human genetic variation'. In: *Nature* 526.7571, pp. 68–74. DOI: 10.1038/nature15393.

The Anopheles gambiae 1000 Genomes Consortium (2017). 'Genetic diversity of the African malaria vector Anopheles gambiae'. In: *Nature* 552.7683, pp. 96–100. DOI: 10.1038/nature24995.

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'. In: *Genome Res.* 30.10, pp. 1533–1546. DOI: 10.1101/gr.262790.120.

Tiono, AB et al. (2018). 'Efficacy of Olyset Duo, a bednet containing pyriproxyfen and permethrin, versus a permethrin-only net against clinical malaria in an area with highly pyrethroid-resistant vectors in rural Burkina Faso: a cluster-randomised controlled trial'. In: *The Lancet* 392.10147, pp. 569–580. DOI: 10.1016/s0140-6736(18)31711-2.

Tu, Z and C Coates (2004). 'Mosquito transposable elements'. In: *Insect Biochem. Mol. Biol.* 34.7, pp. 631–644. DOI: 10.1016/j.ibmb.2004.03.016.

VectorBase (2019). *Anopheles gambiae PEST, AgamP4.12.* Tech. rep. VectorBase, www.vectorbase.org (Giraldo-Calderón et al. 2015).

Wesolowski, A, AR Taylor, HH Chang, R Verity, S Tessema, JA Bailey, T Alex Perkins, DE Neafsey, B Greenhouse and CO Buckee (2018). 'Mapping malaria by combining parasite genomic and epidemiologic data'. In: *BMC Med.* 16.1. DOI: 10.1186/s12916-018-1181-9.

WHO (2006). 'Indoor residual spraying. Use of indoor residual spraying for scaling up global malaria control and elimination. WHO Position Statement'. In: *Trop. Med. Int. Heal.*

WHO (2012). *Global plan for insecticide resistance management in malaria vectors*. Tech. rep. World Health Organization.

WHO (2015). *Global Technical Strategy for Malaria 2016-2030*. Tech. rep. World Health Organization.

WHO (2017). *Conditions for deployment of mosquito nets treated with a pyrethroid and piperonyl butoxide*. Tech. rep. World Health Organization.

WHO (2018). *Test procedures for insecticide resistance monitoring in malaria vector mosquitoes (Second edition)*. Tech. rep. World Health Organization.

WHO (2019). *World malaria report 2019*. Tech. rep. World Health Organization.

WHO (2020). *Prequalified Lists - Vector control products - Prequalified Products 26 August 2020*. Tech. rep. World Health Organization.

Wohl, S, SF Schaffner and PC Sabeti (2016). 'Genomic Analysis of Viral Outbreaks'. In: *Annu. Rev. Virol.* 3.1, pp. 173–195. DOI: 10.1146/annurev-virology-110615-035747.

2 Historical context: correspondence on the discovery of the *Anopheles gambiae* species complex

In this chapter I look back to a series of discoveries that were made during the first global malaria eradication campaign of the 1960s, which uncovered the fact that Anopheles gambiae is not a single mosquito species but rather a complex of multiple morphologically-identical species, with important differences in their ecology, behaviour and feeding preferences. Two entomologists, George Davidson and Hugh Paterson, were at the forefront of these discoveries. I draw on both the published literature of the time, and a collection of 45 previously unpublished letters, to tell the public and private stories of their collaboration. These events provide an introduction to our current understanding of the Anopheles gambiae complex, which includes the mosquito species responsible for the majority of malaria transmission in sub-Saharan Africa. They also mark the introduction of genetic methods into the operational surveillance of malaria vectors, and provide a valuable perspective on present day challenges in malaria vector control. The full correspondence between Davidson and Paterson is provided as a supplementary file.

Anopheles gambiae Giles

The species *Anopheles gambiae* was first described in 1902 in the second edition of Robert M. Giles' handbook on mosquitoes (Giles, 1902). As with all anopheline mosquitoes, *An. gambiae* has a life cycle that involves aquatic egg, larval and pupal stages, and an adult

stage where both males and females feed on nectar but females also require a blood meal to complete egg development. This blood feeding behaviour provides the opportunity for transmission of parasites between hosts, such as the *Plasmodium* parasites causing malaria. However, entomologists studying malaria in Africa during the early part of the 20th century discovered that, among the more than one hundred *Anopheles* mosquito species they encountered, only a handful were able to transmit human malaria, and of those, *An. gambiae* Giles was most often the dominant vector (de Meillon, 1947; de Meillon, 1950; Gillies and de Meillon, 1968).

Following the advent of chemical insecticides, there were early demonstrations during the 1940s that malaria vectors could be effectively controlled by indoor residual spraying of insecticides. Optimism spread that malaria could be eradicated by eliminating its mosquito vector, and a global malaria eradication programme (GMEP) was launched by the WHO in 1955 (Nájera et al., 2011). However, spraying campaigns met with mixed success in sub-Saharan Africa. Insecticide resistance quickly emerged at several locations where spraying had been carried out, and mosquitoes in some areas appeared to change their behaviour to avoid the insecticides. Two entomologists, George Davidson in London and Hugh Paterson in Johannesburg, were at the forefront of operational research seeking to understand these events. Learning of each other's work via preliminary reports published by the WHO, Davidson and Paterson established a long-distance collaboration.

George Davidson

George Davidson was based at the Ross Institute of Tropical Hygiene in London, and was involved during the 1940s in early trials of indoor residual spraying to counter malaria in Sierra Leone, Democratic Republic of Congo, Tanzania and Kenya. In 1954, the Western Sokoto malaria control pilot project in Northern Nigeria began indoor residual spraying of the insecticides DDT and dieldrin (Bruce-Chwatt and Archibald, 1959), but less than two years later the first reports emerged of resistance to these insecticides among anopheline mosquitoes in sprayed regions (Elliott and Ramakrishna, 1956). Eggs of resistant and susceptible mosquitoes were sent to Davidson in London, where they were used to establish

Figure 2.1. George Davidson (left), tending mosquito colonies at the London School of Hygiene and Tropical Medicine (right).

mosquito colonies and study resistance under controlled conditions. Davidson confirmed that mosquitoes from Western Sokoto were indeed resistant, particularly to dieldrin, which had been sprayed in approximately half of the study area (Davidson, 1956).

Davidson and colleagues continued to investigate the genetic basis of dieldrin resistance, using crosses between resistant and susceptible parents from colonies of mosquitoes identified as *An. gambiae* Giles originating from different locations throughout Africa. As they did so, they began to find that crosses between certain pairs of colonies always yielded male offspring that were unable to reproduce, although the female offspring were fully fertile. Recognising sterility in one sex as a hallmark of interbreeding between different species (Haldane, 1922), Davidson and colleagues began to suspect that *An. gambiae* Giles may in fact be more than one species.

Hugh Paterson

At the time that Davidson was performing crosses in London, Hugh Paterson was engaged as a WHO consultant at the East African Institute of Malaria and Vector-borne Diseases in Amani, Tanzania. *Anopheles* larvae are generally found in fresh water only, but reports

Figure 2.2. Hugh and Shirley Paterson, circa late 1980s.

had emerged from East Africa of *An. gambiae* Giles mosquitoes able to breed in estuarine salt-water environments (de Meillon, 1947; Muirhead-Thomson, 1948), and Paterson took the opportunity to investigate. He performed crosses between three colonies, finding male sterility and other evidence of reproductive incompatibility between fresh-water and salt-water-breeding parents.

After completing his tenure in Tanzania, Paterson joined Peter Mattingly, an entomologist and taxonomist from the British Museum, on a tour of several countries in Southern Africa including Swaziland, Mozambique, Zimbabwe and Mauritius. The purpose of the tour was to investigate reports alleging that *An. gambiae* mosquitoes had changed their behaviour in response to insecticide spraying campaigns, preferring to feed outdoors instead of indoors in order to avoid contact with insecticides (Mattingly, 1963). Paterson collected eggs whilst on his travels and, on his return to his permanent position at the South African Institute for Medical Research in Johannesburg, used them to establish colonies and conduct crossing experiments. These crosses, like those being performed by Davidson in London, suggested that the status of *An. gambiae* Giles as a single biological species needed to be reconsidered.

1962: Species A and B

Davidson first described the results of his crossing experiments in a WHO/Mal technical report in 1962. The report, entitled "Incipient speciation in *Anopheles gambiae* Giles", provided the first evidence for two distinct fresh-water forms of *An. gambiae*, which Davidson named "Group A" and "Group B" (Davidson and Jackson, 1962). A draft of this paper reached Paterson whilst on tour, and he wrote to Davidson from Mauritius in March 1962:

> "I have just received your recent WHO/Mal report and have read it with the greatest interest. I have also had the privilege this week of discussing the work we have been doing on gambiae with Peter Mattingly. May I say that I agree with you that your evidence is best interpreted as indicating that your two groups are separate species, especially when one takes Coronel's evidence of the presence of both at Diggi. I believe it will be shown that at Muheza, Tanganyika, both forms are present." (Letter 1)

Paterson's own work on crosses between fresh and salt water-breeding mosquitoes in Tanzania also first appeared as a WHO/Mal report (Paterson, 1962a), and Davidson replied:

> "Thank you for your letter of March 9th last. I had read the account of your work at Amani with great interest [...] I think our strains, which are all fresh-water strains, are probably more closely-related than were your fresh-water and salt-water strains. I am now trying to get eggs of Melas from Liberia, and of salt-water tolerant gambiae from Amani, to try some crosses here in London." (Letter 2)

Davidson's letter also discussed the search for morphological characteristics which could be used to distinguish the different species. Such characteristics were highly desirable because, if they could be found, they would provide a means of identifying mosquitoes in the field, rather than having to transport them to the lab and identify them via cross-mating

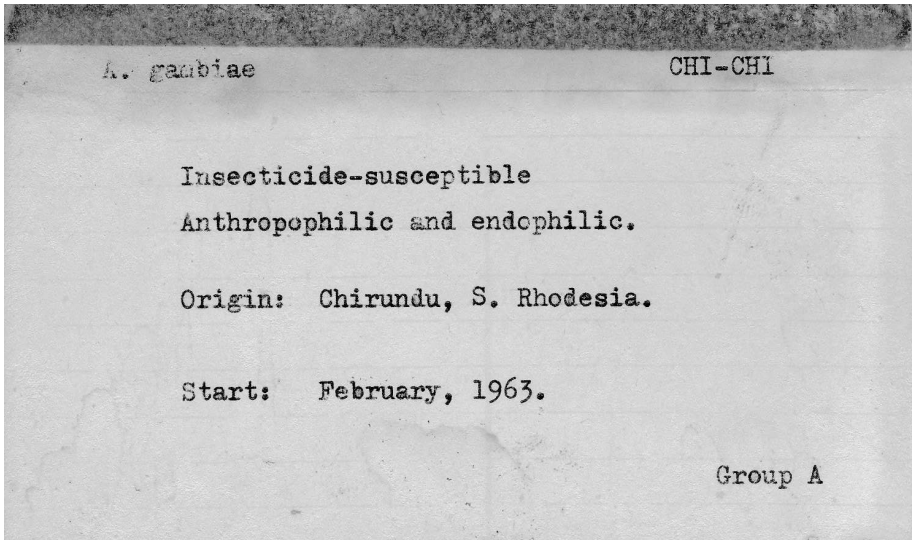

Figure 2.3. A card labelling a mosquito colony, found together with the Davidson-Paterson letters. Presumably this card labelled a colony established in London by Davidson using eggs sent by Paterson from Southern Africa.

with colonies of known types. This was highly relevant to the ongoing malaria eradication campaigns of the time, where field entomologists needed practical methods to investigate issues such as insecticide resistance and mosquito behaviour, which could differ between species. However, early studies suggesting differentiating characteristics between group A and group B (Coronel, 1962) proved to be premature, as Davidson wrote:

> "I have been looking through Miss Coronel's detailed figures on sector spot measurements and find considerable differences between strains within a group. [...] We have just identified a recently acquired gambiae strain from the Ivory Coast as belonging to group B from sector spot measurement and pupal spine characters, whereas from crossings it obviously belongs to group A. Therefore the morphological method may not always be reliable." (Letter 2)

Davidson followed up Paterson's work, performing crosses between the two freshwater groups A and B and salt-water tolerant strains from both East and West Africa, publishing the results later that year, confirming male sterility between all four forms (Davidson, 1962). For Davidson, however, the fact that sterility was restricted only to males, and therefore that there remained the possibility of gene flow between these different forms, created doubt whether they should be considered separate species.

By May 1962, Paterson had completed his tour and returned to Johannesburg. Paterson wrote to Davidson to discuss his results, and his letter illustrates one of the technical challenges that both entomologists were confronting at the time. Repeating each other's experiments required establishing and maintaining a consistent set of reference colonies between two different research institutes, but achieving that was a logistical challenge, not least because establishing colonies required posting of mosquito eggs between two continents. Paterson wrote:

> "I would like to suggest to you that all of us working on gambiae complex studies should standardize our reference colonies. I should like to see the Kisumu colony used as the "type" of group A since it is the most widely used colony of A. gambiae. I hope you will agree with this and suggest a colony as a reference colony for group B. [...] I should be very pleased if you would help us by sending us eggs of the group B colony you choose so that we may establish it here." (Letter 3)

Davidson wrote back in June to discuss the issue of reference colonies, and with further news on the search for distinguishing morphological characteristics, which continued to be fruitless (Letter 4). Paterson responded at the end of June, acknowledging successful receipt and hatching of eggs from group B sent by Davidson. He also shared findings from experiments using colonies established from eggs obtained in Mauritius:

> "I have now found evidence of both group A and group B on Mauritius. [...] This is very interesting in any case since it is one more spot where group A and group B are sympatric. Of course, what I assume to be group B may be a new group, but I think this is unlikely." (Letter 5)

The evidence that two reproductively isolated groups could be found living together at the same location (sympatric) was the key criterion for Paterson to support the elevation of these groups to species rank, because it would establish that although gene flow was possible under lab conditions, in nature the two groups remained genetically distinct.

In the final letter from 1962, Paterson wrote in July with some intriguing results:

> "I obtained some curious results with a cross I made the other day. I managed to get a couple of wild caught gambiae from Southern Rhodesia to lay some eggs. From these we got adults which we set up in both directions against Kisumu adults. The results obtained differ strikingly from results obtained from crosses between groups A & B. [...] Although one cannot base much on an isolated case like this it does suggest that there may be yet another member of the freshwater complex." (Letter 7)

Here was the first hint of a further species discovery in Southern Africa.

1963: Species C

In Paterson's initial publications on the East African salt-water *An. gambiae*, he argues that it is a distinct species, but he does not put forward any suggestions for a new species name (Paterson, 1962a; Paterson, 1962b). However, in 1962, the German entomologist F. Kuhlow also published work showing this to be a distinct species, proposing the name "Anopheles tangensis" (Kuhlow, 1962). Mattingly published a paper in the same year pointing out that "Anopheles merus", a previous synonym of "Anopheles gambiae", might be applicable (Mattingly, 1962). Paterson followed up Mattingly's suggestion, examining the remaining type specimens for *An. merus*, and must have shared his work with colleagues, because in January 1963 he received a letter from R. C. Muirhead-Thomson of the WHO Division of Malaria Eradication, who wrote:

> "Normally, I think there would be every justification to raise "salt water gambiae" to a specific rank, but as it is the implications of your findings are closely interwoven with Davidson's current work on the mating groups of the gambiae complex, work which is being continued and which will undoubtedly lead to further re-thinking on the whole A. gambiae taxonomy. [...] I would consider that it is perhaps rather premature and untimely to introduce a new specific name for any one of this complex without close consultation with the other parties concerned." (Letter 8)

Muirhead-Thomson had authority here, because in addition to his WHO role, he had previously confirmed a West African salt-water form *An. melas* to be a distinct species from fresh-water *An. gambiae* via crossing experiments (Muirhead-Thomson, 1948), as well as studying the East African salt-water form (Muirhead-Thomson, 1951). Muirhead-Thomson had sent a copy of this letter to Davidson, and in March Davidson replied:

> "Thank you for the copy of your letter to Paterson on speciation in A. gambiae. Having just read Dobzhansky's "Genetics and the Origin of Species", and White's "Animal Cytology and Evolution", and I am now in a position to make a few comments." (Letter 10)

Davidson referred to Dobzhansky's description of speciation in *Drosophila pseudoobscura* and *D. persimilis* (Dobzhansky, 1951), highlighting the parallels with *An. gambiae*:

> "Evidence of crossing in nature is "very rare" and reproductive isolation considered more or less complete. This is Dobzhansky's criterion of a species. [...] It would thus appear at first sight that we have a strong case for calling all four forms of gambiae separate species on the grounds that reproductive isolation is indicated." (Letter 10)

On the subject of naming, Davidson remained cautious:

> "However, there seems little point in calling them species if it is not possible for "the man in the field" to recognise them from morphological characters. [...] Until such time as absolute differences are forthcoming we should leave the question of specific naming in abeyance." (Letter 10)

Davidson also exchanged letters with Kuhlow in February and March, making the same points (Letter 9, Letter 11). In a further effort to rein in the taxonomists, Muirhead-Thomson wrote to Mattingly in April, again copying Davidson. Urging for a united front, he emphasized the practical impact of these decisions on the ongoing efforts towards malaria eradication:

> "While not wishing in any way to curb or restrict the rights of research workers to give free voice to their own findings and opinions, the case in point is one in which so many interests are concerned - malaria eradication in particular - that all efforts are called for to avoid confusion, dissension or disharmony." (Letter 12)

Paterson moved in 1963 from South Africa to the zoology department at the University College of Rhodesia and Nyasaland in what was then Southern Rhodesia, now Zimbabwe. He continued working throughout 1963 on both the salt-water form and on the third fresh-water form in Southern Africa. Towards the end of the year, he compiled three papers, which were published together as WHO/Mal/421. In the second these papers, Paterson laid out the evidence he had collected that "Anopheles merus" is an appropriate name for the East African salt-water *An. gambiae* (Paterson, 1963). In the third of these papers, Paterson presented his evidence for a new member of the *An. gambiae* complex, which he referred to as "Species C" (Paterson et al., 1963). This was a third fresh-water species, found in Southern Africa, with a strong preference for feeding outdoors on livestock, in contrast to the known fresh-water species A and B whose preference is to feed indoors on humans. Regarding the significance of species C, he wrote:

> "The programme, of which these studies form a part, has as its main object, the elucidation of the apparent behaviour changes which occurred in Swaziland and in the Mazoe Valley following antimalaria spraying campaigns using the insecticide BHC." (Paterson et al., 1963)

He concluded that in these areas the original vector was probably either *An. gambiae* species A or species B, and that it was effectively eliminated by the spraying campaign because of its preference for feeding indoors. Because species C preferred feeding outdoors, it evaded the insecticide and came to predominate. Thus, the previous reports of behaviour change were in fact a change in the relative abundance of different mosquito species.

1964: The species debate

Throughout 1963, Davidson was performing crosses between colonies of all the mating types so far suggested, in all completing some 200 crossings between 36 mosquito colonies. This work led up to the publication, "*Anopheles gambiae*, a complex of species" in the Bulletin of the WHO, in which he confirms "the existence of five mating-types in what was until recently considered a single species" (Davidson, 1964). The results themselves were indisputable, but their interpretation left room for debate, and this gave rise to a colourful exchange between Davidson and Paterson. In particular, Davidson reiterated a theory that the divergence between species A and B "may be occurring independently in different parts of Africa", originally put forward in Davidson and Jackson (1962). The driver behind this view appears to be the fact that both species A and B had a very broad geographical distribution, spanning much of continental Africa, although the logic is not stated.

Paterson did not agree with this theory. Davidson sent Paterson an early draft of his new paper, and Paterson responded in January 1964:

> "Perhaps I could start by explaining why I am so unenthusiastic about the possibility of A and/or B having multiple origins. [...] Sympatric speciation is enormously improbable and for it to occur independently more than once is astronomically remote. From your paper I gather that you now agree with this point, but you substitute the suggestion that sp. A (or sp. B) could have evolved independently on several occasions by geographical speciation. I feel that the improbabilities involved in the arrival at a single species by convergent evolution on several occasions are just as great as in the case I dealt with and, I am sure you will not easily find an evolutionist who will support you in this view." (Letter 15)

Paterson also argued for the practical implications for entomologists working in the field:

> "We must never forget that at present the importance of the work in which we are engaged is mainly practical. I feel very strongly therefore that every

> effort should be made to avoid confusing the applied worker with theoretical arguments." (Letter 15)

Davidson replied in February:

> "I don't think a little discussion on theoretical aspects of speciation (which you started, not I) will do any harm. Am I not entitled to express my views or have I to accept the word of Paterson as the last on the subject [...] ?" (Letter 16)

Davidson was also adamant that field workers would not make use of the knowledge of existence of multiple species until they had the tools to recognize them. Returning to the "species" question, Davidson continued:

> "It seems to me that there are so many gaps in our knowledge of this A. gambiae complex that it is too soon to be as dogmatic as you are. I would prefer to keep a more open mind on the subject and await further facts; particularly with regard to the status of the A and B forms. These seem to be so intimately mixed, and both have been recorded from areas of holoendemic malaria, that differences in their vectorial capacity can only be slight. Apparent changes in the predominance of one or the other forms has occurred over the years in Upper Volta and Western Sokoto, for example." (Letter 16)

On whether two species can have multiple origins, Davidson invoked the geographical distribution of insecticide resistance as evidence, although again the logic of his argument is not stated:

> "One final point on this A and B issue: if they have a single origin how do you account for the fact that resistance in both A and B forms is confined to West Africa, or has this no relevance?" (Letter 16)

In March, Paterson responded:

> "I must point out that you asked me to comment on your manuscript. This I did. Naturally I did not expect you to accept my views unless you were

> convinced they were correct. [...] Of course I have noted the distribution of genes determining dieldrin resistance in Africa. I have not yet attempted to explain them as I cannot do so on the available evidence." (Letter 18)

Paterson's arguments were not enough to persuade Davidson, and Davidson submitted his paper to the Bulletin of the WHO including the theory of multiple origins of species A and B. Despite this debate, the willingness to collaborate was not diminished. In addition to continuing to exchange eggs, Davidson wrote to Paterson in May with an invitation to join him in writing a chapter for a book on vector genetics being prepared by WHO. Attempting to forge a consensus, Davidson wrote:

> "It seems to me to be a golden opportunity for combining our various versions of the gambiae situation. [...] Will you agree to a conclusion on the following lines:- On present evidence A. melas, A. merus and form C are probably species and A and B are possibly species, but more information on the extent or absence of hybridization in the field is required and more cytogenetic investigation is necessary before the precise status can be given." (Letter 19)

Here "cytogenetics" refers to the study of chromosomal inversions as a means of genetically identifying species, which was new technique emerging at the time. Paterson stood firm, however, replying:

> "With regard to your proposed statement on the status of the forms, I find myself in a difficult position. Your statement starts "On present evidence...". As you know I have argued on this same evidence that the forms should be regarded as separate species on the basis of workers including Mayr, Cain and Mattingly. [...] I realize the difficulty fully. I therefore think that the best that can be done is to state your own position and to add a note that your opinion is not universally accepted." (Letter 20)

In July, Davidson attended the international congress of entomology in London, presenting his results on the *An. gambiae* complex. Davidson wrote to Paterson afterwards with the final draft of the paper (Letter 21), to which Paterson responded, perhaps somewhat wryly:

> "Many thanks for your letter and the copy of your paper. I am glad to hear that it aroused some discussion. Perhaps some interesting new suggestions came from the general evolutionists present?" (Letter 22)

There may have been some discussions in London which helped to sway Davidson. Paterson also published a short paper containing data from a study in Southern Zimbabwe at a location where species A, B and C were all found together, in which no evidence was found for any hybridisation between the forms, strengthening the argument for separate species (Paterson, 1964). In November, Davidson wrote to Paterson:

> "I enclose a copy of a paper I was invited to contribute to the coming number of the Rivista di Malariogia. In it you will see I have now come to your conclusion that all five mating types should be considered full species [...]" (Letter 23)

Davidson held on to his views on the potential for hybridisation, however, writing:

> "I think we must recognize [...] that hybridization does occur on occasion, not only between A and B, but also between melas and A (or ?B). Perhaps such hybridization occurs when high densities of the two forms overlap at certain times of the year." (Letter 23)

Concerning the paper submitted earlier in the year to the WHO, Davidson added:

> "The article I wrote in September last year ("Anopheles gambiae: A Complex of Species") [...] still lies in Geneva awaiting publication. I have now sent a postscript to this article pointing out that evidence accumulated since it was written now points to all the forms being full biological species [...]" (Letter 23)

The postscript reached Geneva in time and was published at the end of the article (Davidson, 1964). This must be one of the few examples in the scientific literature where an author puts forward a theory and then retracts it within the same publication.

Figure 2.4. Photograph of mosquito specimen collection in Zimbabwe, found together with the Davidson-Paterson letters. On the back of the photograph is written, "Lundi River (Hippo Valley Estates). Habitat of larvae A. gambiae species C. 24th September, 1967 (dry season)."

1965-1971: *Anopheles gambiae*, a complex of species

By the beginning of 1965, Davidson and Paterson were in full agreement regarding the status of the five mating types as distinct species. The focus of their correspondence then switched to preparing material for the chapter on the *An. gambiae* complex for the upcoming WHO book on vector genetics. This article was to include a map of the geographical distribution of the five species, which required pulling together all available records of the species from sites where they had been genetically identified via crossing against colonies of known species. During the second half of 1964 and throughout 1965 they exchanged letters requesting, commenting on and correcting species distribution records. Both hoped the map would be of considerable practical value to the ongoing malaria eradication efforts, and would be sufficiently complete to serve as a standard for some time. It would take a further two years to complete this work, but the chapter, including the map, was finally published in 1967 (Davidson et al., 1967). Although the map was

limited by the methods and data available at the time, it is remarkably complete and concordant with current species distribution maps (Wiebe et al., 2017). For example, the geographical limits of the coastal salt-water species were already well-established. Also, although Species A and Species B overlapped considerably, it was clear that the range of Species B extended further north in East Africa, including the Arabian peninsula. The general preference of Species B for more arid conditions was also evident.

Davidson and Paterson also continued to send each other eggs during 1965, particularly of species C, as both struggled to establish a robust colony which could be used to further study the species and type field specimens via crossing. Laboratory crosses were difficult and laborious, and the search continued for morphological characters that could separate the species, led by Mario Coluzzi in Rome. Coluzzi published initial findings in 1964 (Coluzzi, 1964) which were extended and included in Davidson et al. (1967). However, although some combinations of features could be used to differentiate species to some extent, no reliable classifications could be made, a finding corroborated by a comprehensive morphological analysis of the *An. gambiae* complex published later (Coetzee, 1989). Given these difficulties, attention shifted towards direct genetic methods of species identification. The most promising approach, cytogenetics, involved examining the banding patterns of polytene chromosomes examined under a microscope. The first cytogenetic map of *An. gambiae* had been published some years earlier (Frizzi and Holstein, 1956), and characteristic differences were ultimately found that could differentiate all five species in the *An. gambiae* complex (Coluzzi and Sabatini, 1967; Coluzzi and Sabatina, 1968; Coluzzi and Sabatina, 1969). This was a major methodological breakthrough, and Davidson wrote to Paterson in 1968 with some excitement:

> "You may not have heard of the latest development in the A. gambiae complex situation. Coluzzi can now definitely tell A from B by the X-chromosome and can also distinguish melas and merus by the X-chromosome and autosomes. I tested him on 14 of our colonies (A, B, melas, merus) and he got them all right. Now, of course, we are itching to adopt this cytogenetic technique. It will take a much shorter time and much less trouble than our present crossing

technique."

Although a definitive account of the five species had been published in 1967, the naming of the species remained unresolved. In 1968, Paterson completed a PhD thesis, which included carefully researched proposals for naming all five species, although the thesis was not published. Paterson planned a trip to England in 1970, and Davidson took this as an opportunity to resolve the naming issue, writing to Paterson in February 1970:

> "We all think it would be a good idea if we took advantage of your visit to have a little get-together on the naming of the A. gambiae complex. By all I mean Peter Mattingly, Mick Gillies, Mario Coluzzi (if we can get him here) and possibly John Reid and Professor Bertram." (Letter 39)

After some to-and-fro, the meeting was arranged, and came to a consensus, adopting the names as proposed in Paterson's thesis. However, it would take a further seven years for the resolution to be published (Mattingly, 1977). Species A became *Anopheles gambiae* sensu stricto, species B became *Anopheles arabiensis*, and species C was named *Anopheles quadriannulatus*. *Anopheles melas* was retained for the West-African salt-water breeding species, and *Anopheles merus* was confirmed for the East-African salt-water species.

Anopheles coluzzii

The story of the *Anopheles gambiae* species complex does not end there. Following the work of Davidson and Paterson, Mario Coluzzi and colleagues made a number of further discoveries, reviewed in detail by Powell et al. (2014). They used the new cytogenetic methods to perform surveys of *An. gambiae* populations across Africa, identifying a number of polymorphic chromosomal inversions, which in turn revealed further structuring within the *An. gambiae* species complex. Several distinct "chromosomal forms" were identified within *An. gambiae* sensu stricto, distinguished by characteristic combinations of chromosomal inversions, and certain heterozygous combinations were rarely observed, indicating non-random mating between genetically distinct populations (Toure et al., 1998; Coluzzi et al., 2002). In particular, several chromosomal forms were commonly found in

sympatry, suggesting a further species divisions. Molecular markers were subsequently found which demonstrated the existence of two genetically distinct populations with a broad and overlapping geographical distribution, named M and S molecular forms (della Torre et al., 2001). Despite clear genetic segregation of natural populations, mosquitoes of these two molecular forms were fully fertile when crossed in the lab, creating hesitation about whether these forms were distinct species. However, following the demonstration of genome-wide patterns of sequence divergence between M and S forms (Lawniczak et al., 2010), the M form was recognised as a distinct species in 2013 and given the name *Anopheles coluzzii* (Coetzee et al., 2013).

Conclusions

An. coluzzii, *An. gambiae* and *An. arabiensis* together account for the majority of malaria transmission in sub-Saharan Africa, and remain the subject of intense scrutiny. Because of their epidemiological importance, these three species are the primary subject of study by the *Anopheles gambiae* 1000 Genomes (Ag1000G) Project. However, the first phase of the Ag1000G project sequenced only *An. gambiae* and *An. coluzzii*, and thus these two species are the focus of this thesis. In the next chapter I describe whole-genome sequencing of 888 individual specimens of *An. gambiae* and *An. coluzzii*, and the genome-wide discovery of nucleotide polymorphisms both within and between these species.

Acknowledgments

My father, Simon Miles, completed his PhD under Hugh Paterson at the University of Western Australia. He then moved to London and worked at the London School of Hygiene and Tropical Medicine with George Davidson during the 1970s and 1980s. He moved on from entomology during the 1980s, but stayed in contact with Davidson until his death in 1997. My father inherited the collection of 45 letters on the gambiae complex discovery from Davidson. I am very grateful to my father for passing the letters on to me after I began working on *Anopheles* mosquitoes, and for sharing his memories of working with

Davidson and Paterson. I am also grateful for permission to reproduce photographs of George Davidson and Hugh and Shirley Paterson.

Sadly, Hugh Paterson passed away in 2019. I am very grateful to his daughter, Ann Paterson, for permission to include her father's letters within this thesis. I am also grateful to Maureen Coetzee for sharing some biographical information about Hugh Paterson.

References

Bruce-Chwatt, LJ and HR Archibald (1959). 'Malaria control pilot project in Western Sokoto, Northern Nigeria: report on four years' results'. In: *Proc. Sixth Int. Congr. Trop. Med. Mal.*

Coetzee, M (1989). 'Comparative morphology and multivariate analysis for the discrimination of four members of the Anopheles gambiae group in Southern Africa'. In: *Mosquito Systematics* 21, pp. 100–116.

Coetzee, M, RH Hunt, R Wilkerson, AD Torre, MB Coulibaly and NJ Besansky (2013). 'Anopheles coluzzii and anopheles amharicus, new members of the anopheles gambiae complex'. In: *Zootaxa* 3619.3, pp. 246–274. DOI: 10.11646/zootaxa.3619.3.2.

Coluzzi, M (1964). 'Morphological divergences in the Anopheles gambiae complex'. In: *Riv. Malariol.* 43, pp. 625–34.

Coluzzi, M and A Sabatina (1968). 'Cytogenetic observations on species C of the Anopheles gambiae complex'. In: *Parassitologia* 10, pp. 155–166.

Coluzzi, M and A Sabatina (1969). 'Cytogenetic observations on the salt water species, Anopheles merus and Anopheles melas, of the gambiae complex'. In: *Parassitologia* 11, pp. 177–187.

Coluzzi, M and A Sabatini (1967). 'Cytogenetic observations on species A and B of the Anopheles gambiae complex'. In: *Parassitologia* 9, pp. 73–88.

Coluzzi, M, A Sabatini, A della Torre, MA di Deco and V Petrarca (2002). 'A polytene chromosome analysis of the Anopheles gambiae species complex'. In: *Science* 298.5597, pp. 1415–1418. DOI: 10.1126/science.1077769.

Coronel, LT (1962). *Morphological variation in Anopheles gambiae Giles*. Tech. rep. WHO/Mal/328. World Health Organization.

Davidson, G (1964). 'Anopheles gambiae, a complex of species'. In: *Bull. World Health Organ.* 31.5, pp. 625–634.

Davidson, G and E Jackson (1962). *Incipient speciation in Anopheles gambiae Giles*. Tech. rep. WHO/Mal/328. World Health Organization.

Davidson, G, HE Paterson, M Coluzzi, GF Mason and DW Micks (1967). 'The Anopheles gambiae complex'. In: *Genetics of Insect Vectors of Disease*. Ed. by JW Wright and R Pal. Amsterdam: Elsevier.

Davidson, G (1956). 'Insecticide Resistance in Anopheles gambiae Giles'. In: *Nature* 178, pp. 705–706.

Davidson, G (1962). 'Anopheles gambiae complex'. In: *Nature* 196, p. 907.

de Meillon, B (1947). 'The Anophelini of the Ethiopian Geographical Region'. In: *Publ. South African Inst. Med. Res.* 49.

de Meillon, B (1950). *Species and subspecies of vectors and their bionomics*. Tech. rep. WHO/Mal/54. World Health Organization.

della Torre, A, C Fanello, M Akogbeto, J Dossou-yovo, G Favia, V Petrarca and M Coluzzi (2001). 'Molecular evidence of incipient speciation within Anopheles gambiae s.s. in West africa'. In: *Insect Mol. Biol.* 10.1, pp. 9–18. DOI: 10.1046/j.1365-2583.2001.00235.x.

Dobzhansky, T (1951). *Genetics and the origin of species*. 3rd ed. New York: Columbia Univ. Press.

Elliott, R and V Ramakrishna (1956). 'Insecticide Resistance in Anopheles gambiae Giles'. In: *Nature* 177, pp. 532–533.

Frizzi, G and M Holstein (1956). 'Etude cytogénétique d'Anopheles gambiae'. In: *Bull. World. Health. Organ.* 15.3-5, pp. 425–435.

Giles, GMJ (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*. 2nd ed. London: J. Bale, sons & Danielsson, ltd.

Gillies, MT and B de Meillon (1968). *The Anophelinae of Africa south of the Sahara (Ethiopian zoogeographical region)*. 2nd ed. Johannesburg: South African Institute for Medical Research.

Haldane, JBS (1922). 'Sex ratio and unisexual sterility in hybrid animals'. In: *J. Genet.* 12, pp. 101–109. DOI: 10.1007/BF02983075.

Kuhlow, F (1962). 'Studies on the bionomics and the morphology of the saltwater breeding Anopheles gambiae on the coast of Tanganyika'. In: *Riv. Malariol.* 41, pp. 187–197.

Lawniczak, MKN et al. (2010). 'Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences'. In: *Science* 330.6003, pp. 512–514. DOI: 10.1126/science.1195755.

Mattingly, PF (1962). *The problem of behaviour changes in Anopheles gambiae Giles.* Tech. rep. WHO/Mal/354. World Health Organization.

Mattingly, PF (1963). *Some aspects of entomological problems in malaria in Africa.* Tech. rep. WHO/Mal/389. World Health Organization.

Mattingly, PF (1977). 'Names for the Anopheles gambiae complex'. In: *Mosquito Systematics* 9.3, pp. 323–328.

Muirhead-Thomson, RC (1948). 'Studies on anopheles gambiae and A. melas in and around Lagos'. In: *Bull. Entomol. Res.* 38.4, pp. 527–558. DOI: 10.1017/S0007485300023221.

Muirhead-Thomson, RC (1951). 'Studies on Salt-Water and Fresh-Water Anopheles gambiae on the East African Coast'. In: *Bull. Entomol. Res.* 41.3, pp. 487–502. DOI: 10.1017/S0007485300027772.

Nájera, JA, M González-Silva and PL Alonso (2011). 'Some lessons for the future from the global malaria eradication programme (1955-1969)'. In: *PLoS Med.* 8.1, e1000412. DOI: 10.1371/journal.pmed.1000412.

Paterson, HE (1962a). *On the status of the East African salt water-breeding variant of Anopheles gambiae Giles.* Tech. rep. WHO/Mal/346. World Health Organization.

Paterson, HE (1962b). 'Status of the East African salt-water-breeding variant of Anopheles gambiae Giles'. In: *Nature* 195, pp. 469–470.

Paterson, HE (1963). *On the naming of the East African salt-water species of the A. gambiae complex*. Tech. rep. WHO/Mal/421. World Health Organization.

Paterson, HE (1964). 'Direct evidence for the specific distinctness of forms A, B, and C of the Anopheles gambiae complex'. In: *Riv. Malariol.* 43, pp. 191–6.

Paterson, HE, JS Paterson and GJ van Eeden (1963). *A preliminary report on a new member of the A. gambiae complex*. Tech. rep. WHO/Mal/421. World Health Organization.

Powell, JR, NJ Besansky, A della Torre and V Petrarca (2014). 'Mario Coluzzi (1938-2012).' In: *Malar. J.* 13.1, p. 10. DOI: 10.1186/1475-2875-13-10.

Toure, YT, V Petrarca, SF Traore, A Coulibaly, MH M., O Sankare, M Sow, MA Di Deco and M Coluzzi (1998). 'The distribution and inversion polymorphism of chromosomally recognized taxa of the Anopheles gambiae complex in Mali, West Africa'. In: *Parassitologia* 40, pp. 477–511.

Wiebe, A et al. (2017). 'Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance'. In: *Malar. J.* 16.1. DOI: 10.1186/s12936-017-1734-y.

3 The *Anopheles gambiae* 1000 Genomes Project phase 1 nucleotide variation data resource

In this chapter I describe the production and curation of a data resource of nucleotide variation obtained from whole-genome sequencing of 888 individual mosquitoes wild-caught from natural populations, and a further 80 individuals from colony crosses. Production of this data resource was a collaborative effort involving members of multiple research teams within the Ag1000G Consortium. My contribution, described in this chapter, was to analyse the raw variant calls, define and carry out quality control and validation analyses, investigate genome accessibility and define quality filters, and produce the final analysis-ready data resource. I also report analyses of these data to quantify the levels of nucleotide variation found, and to explore how genomic features such as protein-coding genes affect the genomic landscape of nucleotide variation.

Introduction

As described in chapter 1, the Ag1000G Project aims to use whole-genome deep Illumina sequencing to explore natural genetic variation among populations of malaria vectors within the *An. gambiae* complex. For logistical reasons the project was divided into three phases, and a total of 888 mosquito specimens sampled from natural populations were included in the first project phase. A further 80 mosquito specimens comprising parents and progeny of four colony crosses were also sequenced within this project phase. This

chapter is primarily methodological, describing the processes and analyses developed and used to identify, genotype and validate single nucleotide polymorphisms (SNPs) among these specimens. For completeness, I have included some brief methodological information regarding population sampling and whole-genome sequencing which were performed by members of the Ag1000G Consortium. However, the main focus of this chapter is the work I contributed to production and validation of a high quality genome-wide resource of SNP data from sequencing of these specimens. Where work was carried out by or in collaboration with other members of the Ag1000G Consortium I have indicated that within the relevant subsection.

Methods

Population sampling

Population sampling was performed by members of the Ag1000G Consortium. Below is a brief description of the cohort of mosquito specimens obtained for sequencing. A more detailed description is available in The Anopheles gambiae 1000 Genomes Consortium (2017).

A total of 888 mosquito specimens collected from natural populations were included in the Ag1000G phase 1 cohort. This included mosquitoes representing two major malaria vector species *An. gambiae* and *An. coluzzii*. Mosquitoes were sampled from 8 countries in sub-Saharan Africa representing a broad geographical range spanning the continent: Guinea-Bissau, Guinea, Burkina Faso, Cameroon, Gabon, Angola, Uganda and Kenya (Fig. 3.1). Mosquitoes had been collected prior to the initiation of the Ag1000G Project as part of previous field studies, and were collected at different times, with the earliest collections being in 2000 (Gabon) and the most recent in 2012 (Burkina Faso). Because these mosquitoes were collected in the context of different studies, a number of different collection methods were used, including light traps, pyrethrum spray catch and larval collection. In addition to the wild-caught specimens, a further 80 mosquitoes were obtained from four colony crosses, each cross comprising two parents and up to 20 progeny, where

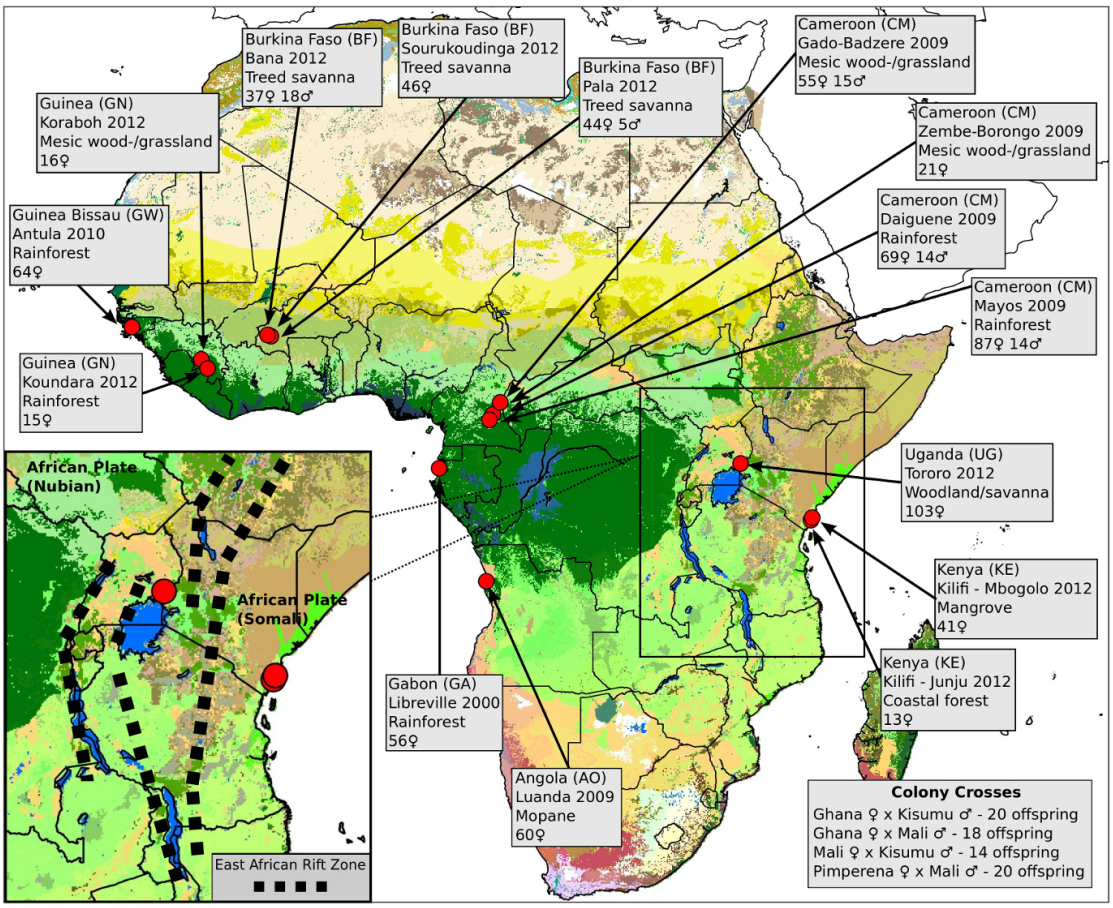

Figure 3.1. Map of sampling locations. Numbers of specimens are shown by gender. Parental colonies and numbers of offspring for colony crosses are shown inset. Numbers of samples are shown after removing samples that failed any of the quality control steps described in this chapter. Colours in the map denote ecosystem classes, see Sayre et al. (2013) Fig. 9 for a complete colour legend. This figure was produced in collaboration with Chris Clarkson.

parents were drawn from established reference colonies (Fig. 3.1 inset). The main rationale for including colony crosses was to provide a mechanism for calibrating and measuring the accuracy of variant calling methods, because colony crosses allow for the analysis of genetic inheritance and Mendelian inconsistencies between parents and offspring, which can be a useful proxy of variant calling errors (Saunders et al., 2007; Laurie et al., 2010; Pilipenko et al., 2014). DNA extraction was performed on individual mosquitoes and genomic DNA samples were shipped to the Wellcome Sanger Institute for sequencing.

Table 3.1. Depth of coverage. Coverage column shows median (interquartile range) depth coverage for all samples by country and species. Numbers of samples and coverage values are shown after removing samples that failed any of the quality control steps described in this chapter. Species status is uncertain for samples from Guinea-Bissau and Kenya, see Chapter 4 for further explanation.

Country	**Species**	**No. samples** (♀, ♂)	**Coverage**
Angola	*An. coluzzii*	60, 0	30 (22-38)
Burkina Faso	*An. coluzzii*	66, 3	33 (26-44)
Burkina Faso	*An. gambiae*	61, 20	31 (22-68)
Cameroon	*An gambiae*	232, 43	29 (19-53)
Gabon	*An. gambiae*	56, 0	31 (21-37)
Guinea	*An. gambiae*	31, 0	29 (19-39)
Uganda	*An. gambiae*	103, 0	31 (27-39)
Guinea Bissau	uncertain	46, 0	30 (24-41)
Kenya	uncertain	44, 0	31 (18-57)

Whole-genome sequencing

Whole-genome sequencing was performed by staff at the MalariaGEN Resource Center and the Wellcome Sanger Institute sample logistics, sequencing and informatics facilities. Below is a brief description of sequencing methods. A more detailed description is available in The Anopheles gambiae 1000 Genomes Consortium (2017).

Sequencing was performed on the Illumina HiSeq 2000 platform at the Wellcome Sanger Institute. Paired-end multiplex libraries were prepared using the manufacturer's protocol, with the exception that genomic DNA was fragmented using Covaris Adaptive Focused Acoustics rather than nebulization. Multiplexes comprised 12 tagged individual mosquitoes and three lanes of sequencing were generated for each multiplex to even out variations in yield between sequencing runs. Thus, after sequencing data were demultiplexed, data for each individual sample were available from three separate sequencing runs. Cluster generation and sequencing were undertaken per the manufacturer's protocol for paired-end 100 bp sequence reads with insert size in the range 100–200 bp. Target coverage was 30× per individual. The median depth of coverage obtained per individual was at least 29× for all the sample sets contributed (Table 3.1).

SNP discovery and genotyping

SNP discovery and genotyping was performed in collaboration with the MalariaGEN Resource Center data production team. Below is a brief description of SNP calling methods. A more detailed description is available in The Anopheles gambiae 1000 Genomes Consortium (2017).

SNP discovery and genotyping was performed following best practices defined for the Genome Analysis Toolkit (GATK) version 2 (McKenna et al., 2010; Depristo et al., 2011; Van der Auwera et al., 2013). Sequence reads from each sample were aligned to the AgamP3 reference genome (Holt et al., 2002; Sharakhova et al., 2007) using BWA version 0.6.2 (Li and Durbin, 2009). A BAM file for each individual was constructed by merging alignments from multiple lanes, and duplicate reads were marked using Picard version 1.96. Reads were then re-aligned around putative indels found within the alignments using GATK. SNP discovery and genotyping was performed using GATK *UnifiedGenotyper*. *UnifiedGenotyper* was run within non-overlapping 10 kb chunks, and the results were combined into a single variant call format (VCF) file for each chromosome arm using the *vcf-concat* command from vcftools version 0.1.10. SNP discovery was performed using the 888 wild-caught samples. The colony crosses samples were then genotyped at the same set of SNPs using GATK *UnifiedGenotyper*.

Sample quality control

A number of issues can arise in the process of sample and library preparation and high-throughput sequencing which result in data of insufficient quality to perform robust variant discovery and genotyping. These issues include:

- **Low yield.** The number of sequence reads generated for an individual sample is not sufficient to achieve the desired depth of coverage.
- **Low library complexity.** If the input DNA quantity is insufficient, the DNA amplification process can generate sufficient material for sequencing, but there may be significant bias, in the sense that some genome regions are overrepresented and

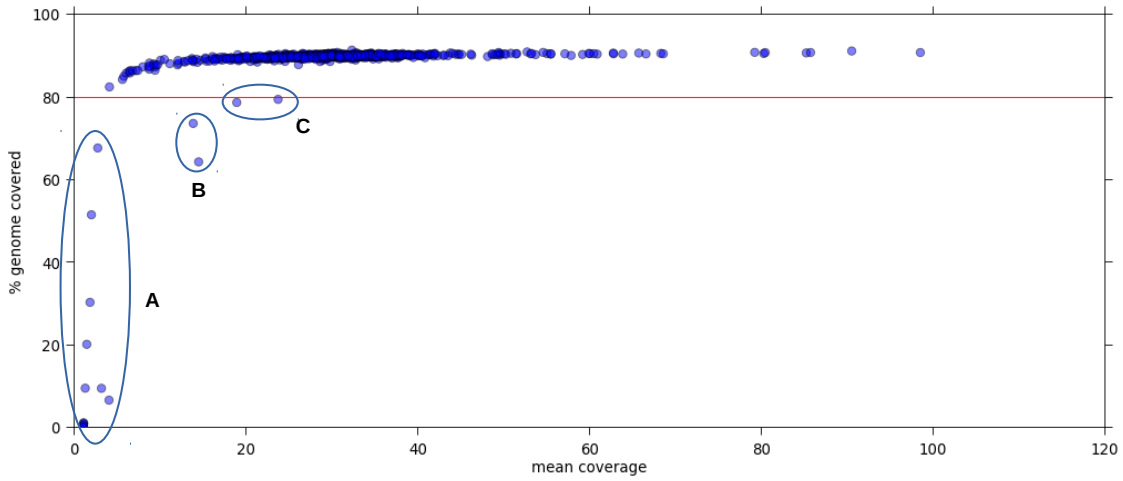

Figure 3.2. Sequencing data quality diagnostics. Each marker represents an individual sample. Mean coverage is computed as total number of sequenced bases divided by genome size. Percent genome covered is percent of nucleotides in reference genome with at least one read aligned. Highlighted groups A, B and C indicate different quality issues, described in the text.

others are underrepresented or not covered at all (Head et al., 2014).

- **Cross-contamination.** During specimen handling, sample preparation and library preparation, material and thus DNA from one sample may contaminate another. In this case the sequencing data generated will include a mixture of reads from the original sample and the contaminating sample.

- **Metadata errors.** Errors may arise during the recording and transfer of metadata about specimens and samples between information systems. This can lead to situations where the identity of one sample becomes swapped with that of another sample.

- **Unexpected taxa.** In Ag1000G phase 1 only *An. gambiae* and *An. coluzzii* mosquitoes were to be included. However, not all specimens had the conventional diagnostic species assays performed prior to selection, and thus were only identified morphologically.

In this subsection I describe analyses performed to identify and exclude samples affected by one or more of these issues.

Sequencing data quality

I analysed summary statistics generated from the sequence read alignments to identify samples with sequencing data quality issues such as low yield or low library complexity. These statistics included the number of sequence reads, mean coverage, percentage of reads mapped to the reference genome, percentage of reads marked as duplicates, and percentage of aligned bases mismatching the reference genome. I plotted various combinations of these statistics to investigate their distributions and identify outliers. A particularly informative plot was to compare the mean coverage with the percentage of genome covered by at least one read (Fig. 3.2). Most samples obtained greater than 80% of the genome covered, but a set of 13 samples had both mean coverage below 7× and less than 80% of the genome covered, indicating insufficient coverage due to low sequencing yield (Fig. 3.2 group A). There were also four samples with high mean coverage but less than 80% genome covered. Two of these samples had a high rate of duplicate reads, and thus poor genome coverage was likely due to low library complexity (Fig. 3.2 group B). The remaining two samples had a higher mismatch rate, thus poor genome coverage was likely due to an unexpected taxon with higher divergence from the reference genome (Fig. 3.2 group C). On the basis of these analyses, the percentage of genome covered appeared to be a useful metric capturing multiple quality issues, and I excluded 17 samples with less than 80% genome covered.

Cross-contamination

Cross-contamination between samples of the same species can be detected by analysing the fractions of reads supporting genotypes in each sample. Several tools have been developed to detect cross-contamination in sequencing data, such as verifyBamID (Jun et al., 2012). Previous human sequencing projects have used verifyBamID and excluded samples with an estimated contamination fraction above 2% (The 1000 Genomes Project Consortium, 2015). To analyse evidence for cross-contamination in the Ag1000G phase 1 samples, verifyBamID was run on all samples by the MalariaGEN data production team. Because this tool had been developed for and previously only applied in human sequencing studies, I performed additional analyses to confirm that the contamination predictions were reliable in

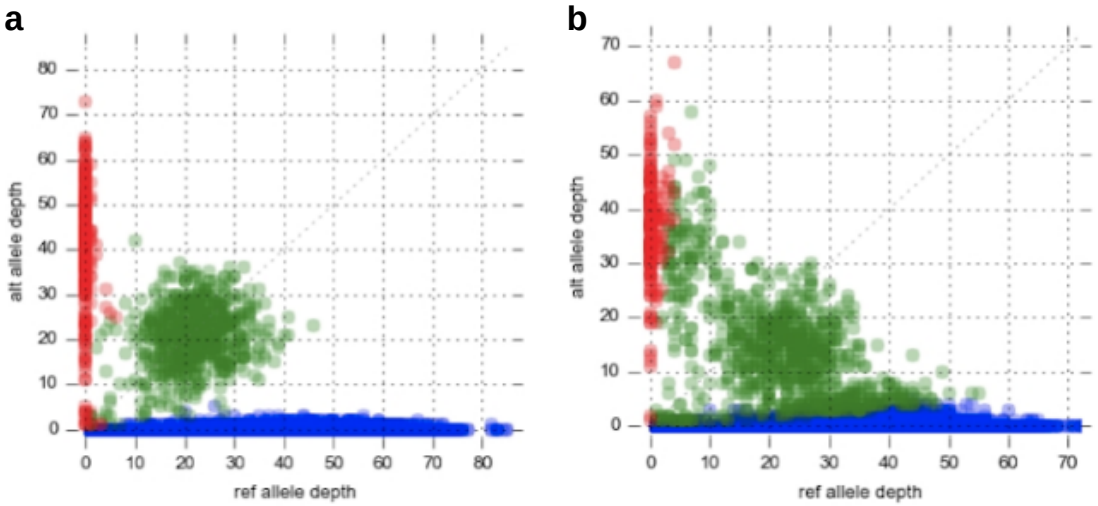

Figure 3.3. Cross-contamination diagnostics. Each plot shows an allele balance diagnostic plot for a single sample. Each marker within a plot is a genotype call (blue = homozygous reference; green = heterozygous; red = homozygous alternate). **a**, Example of sample with predicted contamination $< 2\%$. **b**, Example of sample with predicted contamination $> 4\%$.

Anopheles. I generated a set of diagnostic allele balance plots for each of the 888 mosquito samples in phase 1 which plotted data from genotypes across all autosomes, showing the genotype called, and the numbers of reads supporting the reference and alternate alleles (e.g., Fig. 3.3). I then visually inspected these plots and cross-referenced them against the contamination fraction predicted by verifyBamID. In uncontaminated samples there was clear separation between the different possible genotype calls (e.g., Fig. 3.3a) and cases of contamination were evident as a lack of a clear separation between genotype calls (e.g., Fig. 3.3b). In all cases where verifyBamID predicted contamination $> 4\%$ I confirmed a signal of contamination via allele balance plots, and conversely I found no cases of allele balance plots indicating contamination but verifyBamID predicted contamination $< 4\%$, providing confidence that verifyBamID predictions were reliable. To be conservative, I excluded all samples with verifyBamID contamination $> 2\%$ from the final dataset.

Sex metadata concordance

Both male and female mosquito specimens were contributed to Ag1000G phase 1 (Table 3.1) and the morphologically-identified gender was included in the sample metadata contributed by the Ag1000G consortium partners who collected the specimens. To confirm metadata

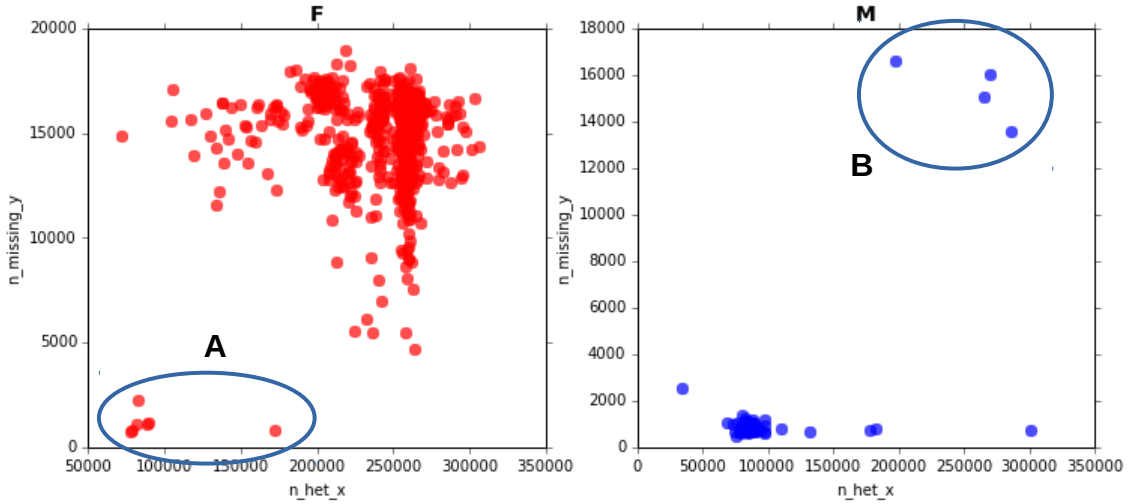

Figure 3.4. Sex metadata concordance checks. Each marker represents a sample. The left panel (labelled F) shows samples from mosquitoes reported in metadata as female, the right panel (labelled M) shows samples reported as male. *n_het_x* = No. of heterozygous genotype calls on the *X* chromosome. *n_missing_y* = No. of missing genotype calls on the *Y_unplaced* chromosome. Groups labelled A and B are described in the text.

records were correct, I compared the reported specimen gender with a gender call made directly from the sequencing data. To call gender for each sample I computed the counts of homozygous, heterozygous and missing genotype calls separately for the X and Y chromosomes. Sex determination is similar in mosquitoes to humans in that females are homogametic XX and males are heterogametic XY. Thus, in females we expect to observe heterozygous genotype calls on the X chromosome and missing genotype calls on the Y chromosome. Conversely, in males we expect only homozygous genotype calls on the X chromosome and non-missing genotype calls on the Y chromosome. Plotting these data it was evident that some samples had genotype counts not consistent with their reported sex (Fig. 3.4). This included six samples reported as female but with male genotype counts (Fig. 3.4 group A) and four samples reported as male but with female genotype counts (Fig. 3.4 group B). I excluded these ten samples from further analyses. There were also a further six samples with apparently androgynous genotype counts on sex chromosomes. Cross-checking against the contamination analysis showed these apparently androgynous samples to have predicted contamination $> 2\%$, explaining the unusual sex chromosome genotype counts.

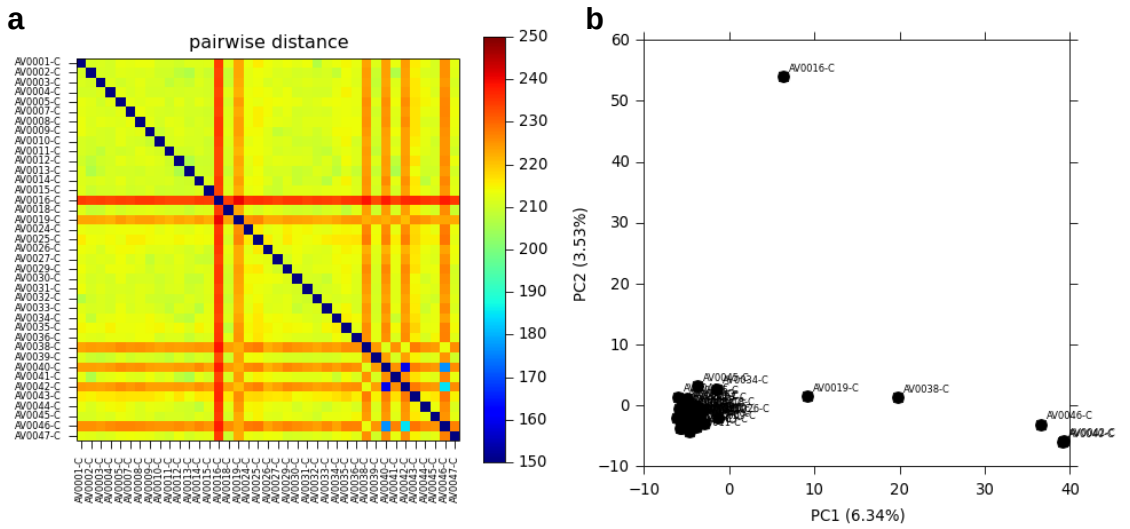

Figure 3.5. Example of population outlier analysis for Guinea *An. gambiae*. Six samples are evident as outliers, having elevated genetic distance and appearing as outliers in PCA. **a**, Pairwise genetic distance between samples, visualised as a heatmap. **b**, PCA showing first and second principal components.

Population outliers

Mosquitoes in the phase 1 cohort represented 8 countries and two species. In Burkina Faso, both *An. gambiae* and *An. coluzzii* were represented, whereas in all other countries only a single species was represented (either *An. gambiae* or *An. coluzzii*), except for Guinea-Bissau and Kenya where the species status was uncertain (Chapter 4). In some countries, specimens had been collected from multiple sites, but in all cases the sites were within a relatively small geographical distance (Fig. 3.1). Thus, in general we would expect to see little or no genetic structure among the specimens of any given species sampled from within a single country. To confirm that the country of origin reported for each specimen in the sample metadata provided by Ag1000G consortium partners was consistent with the patterns of genetic population structure present in the data, I performed an outlier analysis using both pairwise genetic distance and principal components analysis (PCA) (Patterson et al., 2006)). In this analysis I used SNPs from Chromosome 3, then grouped samples by reported country and species, then performed both PCA and pairwise distance analyses on each group in turn (e.g., Fig. 3.5). If any samples were present in a group because of incorrect geographical or species metadata, they would be expected to appear as an outlying group of samples, with elevated genetic distance from other

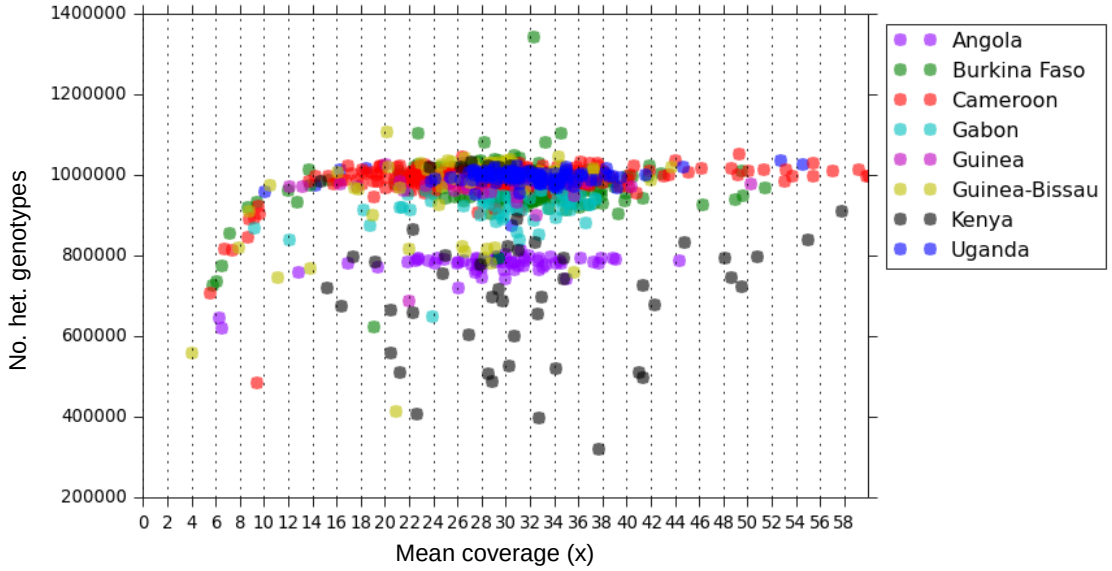

Figure 3.6. Coverage and heterozygosity. Each marker represents a sample, coloured by country of collection. No. het. genotypes = number of heterozygous genotype calls on autosomal chromosomes.

samples, and appearing as outliers in PCA. In total, I identified and excluded 55 samples as population outliers.

Minimum depth of coverage

In the analysis of sequence data quality described above I excluded samples with less than 80% of the reference genome covered by at least one read, which is a conservative way to remove samples that are highly likely to have excessive levels of missingness in their genotype calls. However, for accurate genotyping calling, particularly of heterozygous genotype calls, it is necessary to have sufficient depth of coverage. Previous studies of SNP calling from whole genome sequencing data have found that coverage of 12× is sufficient for calling heterozygous genotypes accurately (Meynert et al., 2014). I examined the total number of heterozygous genotype calls on autosomal chromosomes among the Ag1000G phase 1 samples, plotted against the genome-wide average depth of coverage, for the remaining samples not previously excluded in other sample QC analyses (Fig. 3.6). Although there are clear differences in the degree of heterozygosity in samples from different countries, which likely reflects differences in genetic diversity of the sampled populations, below 14× there is a clear trend towards lower heterozygous genotype calls. This could

indicate an effect such as a bias in the underlying genotype calling, whereby heterozygous genotypes are undercalled at lower coverage. To mitigate this I excluded samples with coverage $< 14\times$ from the final dataset.

Variant quality control

Variant calling algorithms such as GATK *UnifiedGenotyper* are typically configured by default for high sensitivity, which means that the raw variant calls they emit contain a high rate of false positive variants. Errors in variant calling can occur due to a variety of causes:

- **Sequencing errors.** With high depth of coverage, given the low error rate of Illumina sequencing, it is unlikely that sequencing errors will generate false SNP calls, because sequencing errors are generally random and unlikely to co-occur at the same genomic location in read alignments. However, in some cases sequencing errors can be systematic, such as in the vicinity of mononucleotide repeats.

- **Alignment errors.** Errors can occur in the alignment process, where reads originating from one genomic location are improperly aligned to a different genomic location. Any differences between the sequences at the two genomic locations then appear as variants in the misaligned reads. Alignment errors are particularly likely to occur in repetitive genome regions, where sequences at different locations are identical or share a high degree of homology. Several types of repetitive DNA are common throughout eukaryotic genomes:
  - **Low complexity regions**, e.g., sequences with a very high (A+T) content.
  - **Tandem repeats**, where some short sequence of nucleotides is repeated adjacently many times.
  - **Interspersed repeats**, where similar sequences are located at dispersed regions throughout the genome, commonly due to mobile genetic elements such as retrotransposons (Tu and Coates, 2004; Fernández-Medina et al., 2011).

- **Large structural variation.** Large copy number amplifications or deletions can cause errors when attempting to call smaller variants such as SNPs within the same

genomic region or in close vicinity. In particular, if a sample has a copy number amplification relative to the reference genome, sequence reads from both copies of the amplified region will be aligned to the same location in the reference genome. If there is any sequence divergence between the different copies of the amplified region, this can appear as heterozygous SNPs within the sample.

- **Sequence divergence.** If individuals contain sequence within their genomes with a higher degree of divergence from the reference genome, some or all reads may fail to be aligned at all at that location. This results in a failure to observe some alleles, and a bias towards calling true heterozygous genotypes as homozygous, also known as allelic dropout.

- **Reference genome assembly errors.** In any region where the reference genome has been incorrectly assembled, attempting to align sequences reads to that region may lead to spurious variant calls.

In this subsection I describe analyses performed to reduce the false discovery rate within a raw variant callset by designing and applying variant filters.

Genome accessibility

As a first step towards designing variant filters, I performed an analysis of genome accessibility, which determines which positions throughout the genome support unambiguous read alignments, and where there is minimal evidence for structural variation between samples. In this analysis I first computed detailed alignment statistics from read pileups for each sample at each position in the reference sequence, including the total depth of coverage, the fraction of reads aligned in a proper pair, the fraction of reads aligned ambiguously (mapping quality zero), and the root-mean-square mapping quality. In order to accelerate this computation I developed the software package pysamstats[1]. I then aggregated these data across all samples, to obtain the following genome accessibility metrics at each genome position:

[1] `https://github.com/alimanfoo/pysamstats`

- **No Coverage.** The number of samples in which there were no reads aligned at the site (0-100%).
- **Low Coverage.** The percentage of samples in which depth of coverage was less than half the modal coverage for the whole genome (0-100%).
- **High Coverage.** The percentage of samples in which the depth of coverage was more than twice the modal coverage for the whole genome (0-100%).
- **Ambiguous Alignment.** The percentage of samples in which more than 10% of aligned reads are aligned with mapping quality zero, indicating they could equally well have been aligned at another genomic location (0-100%).
- **Low Mapping Quality.** The percentage of samples in which the root-mean-square mapping quality was less than 30 (0-100%).
- **Low Read Pairing.** The percentage of samples in which less than 90% of reads were flagged as being aligned in a proper pair, meaning that the read and its mate pair were aligned in the correct orientation within a reasonable distance of each other (0-100%).

I also obtained repeat annotations for the AgamP3 reference genome for VectorBase. These comprise annotations from three algorithms: RepeatMasker, Tandem Repeat Finder (TRF) and DUST. The DUST algorithm was originally a module within BLAST (Altschul et al., 1990) and is designed to mask low complexity sequences only. The RepeatMasker software (Smit et al., 2013) is primarily designed for masking interspersed repeats matching a library of known repeat sequences (transposable elements etc.). TRF (Benson, 1999) specializes in locating tandem repeats. These provided three additional genome accessibility metrics:

- **TRF.** The genome position is within a region annotated as a tandem repeat by the tandem repeat finder software (True/False).
- **RepeatMasker.** The genome position is within a region annotated as a repeat by the RepeatMasker software (True/False).

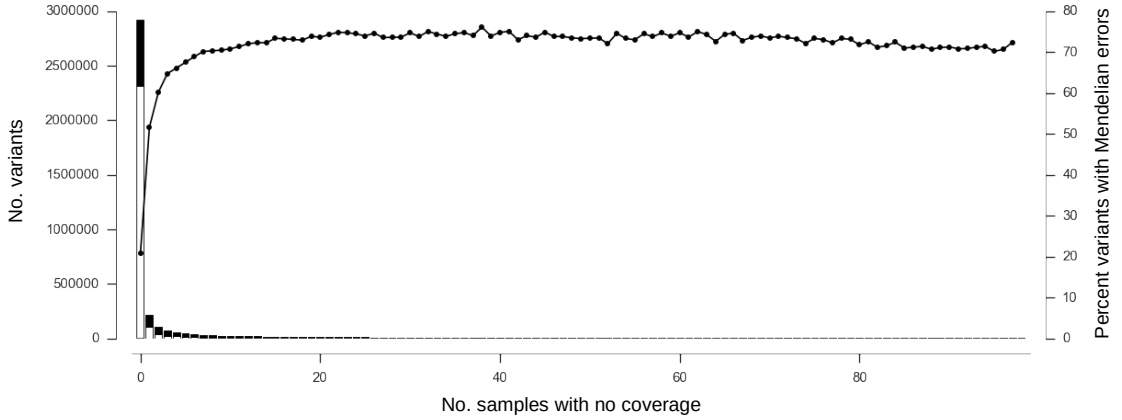

Figure 3.7. Example of using Mendelian error to calibrate thresholds for genome accessibility metrics. Plot shows the Mendelian error (ME) rate for the No Coverage metric which measures the number of samples with no coverage at a given variant site. Bars and left axis show the numbers of variants found for each value of the metric (black = variants with at least one ME; white = variants with no ME). Line and right axis show the percentage of variants found with ME for each value of the metric.

- **DUST.** The genome position is within a region marked as low complexity by the DUST software (True/False).

In previous population sequencing projects such as The 1000 Genomes Project Consortium (2015) the criteria used to determine which positions of the reference genome are considered accessible were decided theoretically, without any empirical validation. The availability of colony crosses in the Ag1000G project afforded the opportunity to select and calibrate appropriate genome accessibility criteria using Mendelian error in the crosses as an indicator of poor accessibility. The rationale here is that true variants will segregate within a cross according to Mendelian inheritance, whereas as false positive variants due to any of the above mentioned error modes would not in general be expected to segregate according to Mendelian inheritance, and thus be enriched for Mendelian errors. A very small number of Mendelian errors are expected due to *de novo* mutation, but the vast majority will be due to spurious variant calling, and hence Mendelian errors are a good indicator of poor quality variant calls (Saunders et al., 2007; Laurie et al., 2010; Pilipenko et al., 2014). As a secondary means of empirically investigating genome accessibility, I also studied the ratio of transition to transversion substitutions (Ti/Tv). Most organisms have mutation bias leading to an excess of transitions over transversions relative to that which would be

expected if substitution mutations occurred randomly (Guo et al., 2013). A low Ti/Tv ratio (approaching 0.5 expected if variants were called randomly) is an indicator of poor variant quality. For each of the genome accessibility metrics listed above, I divided the range of possible values into bins, and then for each bin computed (1) the rate of Mendelian error as the fraction of variants in one or more crosses with genotypes not consistent with Mendelian segregation (ME); and (2) the Ti/Tv rate for biallelic SNPs. I then plotted these results to identify bins within which ME was elevated and/or Ti/Tv was depressed (e.g., Fig. 3.7).

All the genome accessibility metrics displayed some structured association with Mendelian error and Ti/Tv, with higher metric values being associated with higher ME rates and lower Ti/Tv. This was strongest for the No Coverage and Ambiguous Alignment metrics. E.g., at positions where zero samples had No Coverage, the ME rate was 20%, whereas at sites where three or more samples had No Coverage, the ME rate was 60% or greater (Fig. 3.7). Because some of these accessibility metrics would be expected to be correlated with each other to some extent, such as Ambiguous Alignment and Low Mapping Quality, I repeated the above analyses for each accessibility metric, but conditional on thresholds applied to all other metrics, to look for any marginal and thus unique contribution of a single metric. In this analysis, all metrics remained informative, except for the Low Read Pairing metric. I then decided the following criteria for genome positions to be classified as accessible:

- No Coverage <= 0.2% (at most one sample with no coverage)
- Ambiguous Alignment <= 0.2% (at most one sample with 10% MQ0 alignments)
- High Coverage <= 2% (at most 15 samples with high coverage)
- Low Coverage <= 10% (at most 76 samples with low coverage)
- Low Mapping Quality <= 10% (at most 76 samples with RMS MQ < 30)
- DUST = False (position not annotated as a repeat by DUST)

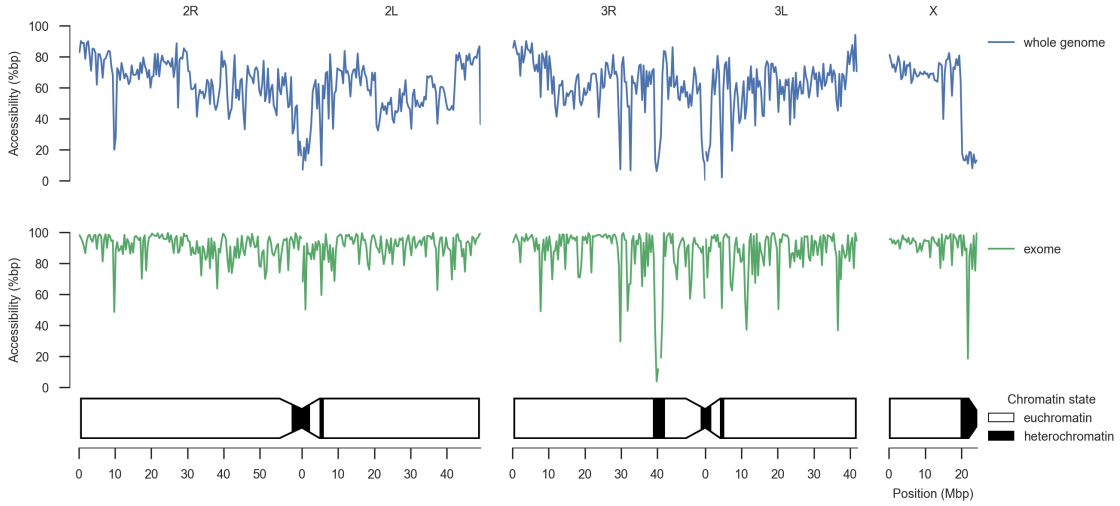

Figure 3.8. Results of genome accessibility analysis. Upper plot (blue) shows the percentage of bases classified as accessible. Lower plot (green) shows the percentage of bases within gene exons classified as accessible. Schematic of chromosomes at the bottom shows predicted chromatin state from Sharakhova et al. (2010).

After applying these criteria, 141 Mbp (61% of the reference genome) was classified as accessible, including 91% (18Mbp) of the exome and 60% (123Mbp) of non-coding positions (Fig. 3.8). There were some clear associations between broad-scale patterns of accessibility and major features of genome architecture. In particular, the pericentromeric regions of all three chromosomes, within regions predicted to be heterochromatic (Sharakhova et al., 2010), all displayed substantially lower accessibility. This is likely due to a higher content of repetitive DNA sequence. The region on chromosome arm 2L within the known polymorphic chromosomal inversion 2La (Coluzzi et al., 2002), which spans from 2L:20.5–42.2 Mb, also displayed lower accessibility. This inversion is known to be segregating in many populations, and the inverted karyotype displays higher levels of divergence with respect to the standard karyotype represented in the reference genome sequence, thus this is likely an effect of divergence.

Variant filters

The raw variant callset produced by GATK *UnifiedGenotyper* comprised a total of 95,335,499 SNPs distributed across the whole genome. I designed a set of filters to annotate those SNPs likely to be false discoveries, and thus provide a means of excluding them from

downstream analyses. I first filtered SNPs using the genome accessibility criteria defined above. I.e., a SNP was filtered if any of the accessibility criteria for the corresponding genome position were not met. I then examined a set of variant metrics produced by *UnifiedGenotyper* for each SNP, including the following metrics:

- **QD.** "Quality by depth", calculated as the evidence for an alternate allele at the site (QUAL, phred-scaled probability of no variant) divided by the total depth of coverage in samples where the alternate allele was observed.
- **FS.** Fisher's exact test for strand-bias. Phred-scaled probability of observing the given numbers of reads aligned to the forward and reverse strand if these were equally likely.
- **ReadPosRankSum.** Rank sum test for relative positioning of reference and alternate alleles within reads. Some sequencing errors tend to occur towards the ends of reads, and this metric quantifies the extent of any bias in read positioning between reference and alternate alleles.
- **HRun.** The length of any adjacent homopolymer run. Homopolymer runs are enriched for sequencing errors in Illumina sequencing data.

Taking a similar approach to the genome accessibility analysis, I used Mendelian error in the crosses and Ti/Tv ratios to investigate which variant metrics appeared to be informative with respect to variant calling errors, and where filter thresholds should be set (e.g., Fig. 3.9). Based on these analyses I filtered SNPs that failed any of the following criteria: QD < 5; FS > 60; ReadPosRankSum < -8; HRun > 4. Of 95,335,499 SNPs reported in the raw callset, 52,525,957 passed all filters. Prior to filtering, the rate of Mendelian error in genotype calls expected to be heterozygous (0/1) in the progeny of the crosses (where parental genotypes were 0/0 x 1/1 or vice versa) was between 13.0-21.7% depending on the cross. After filtering, this error rate was 0.3-0.9%.

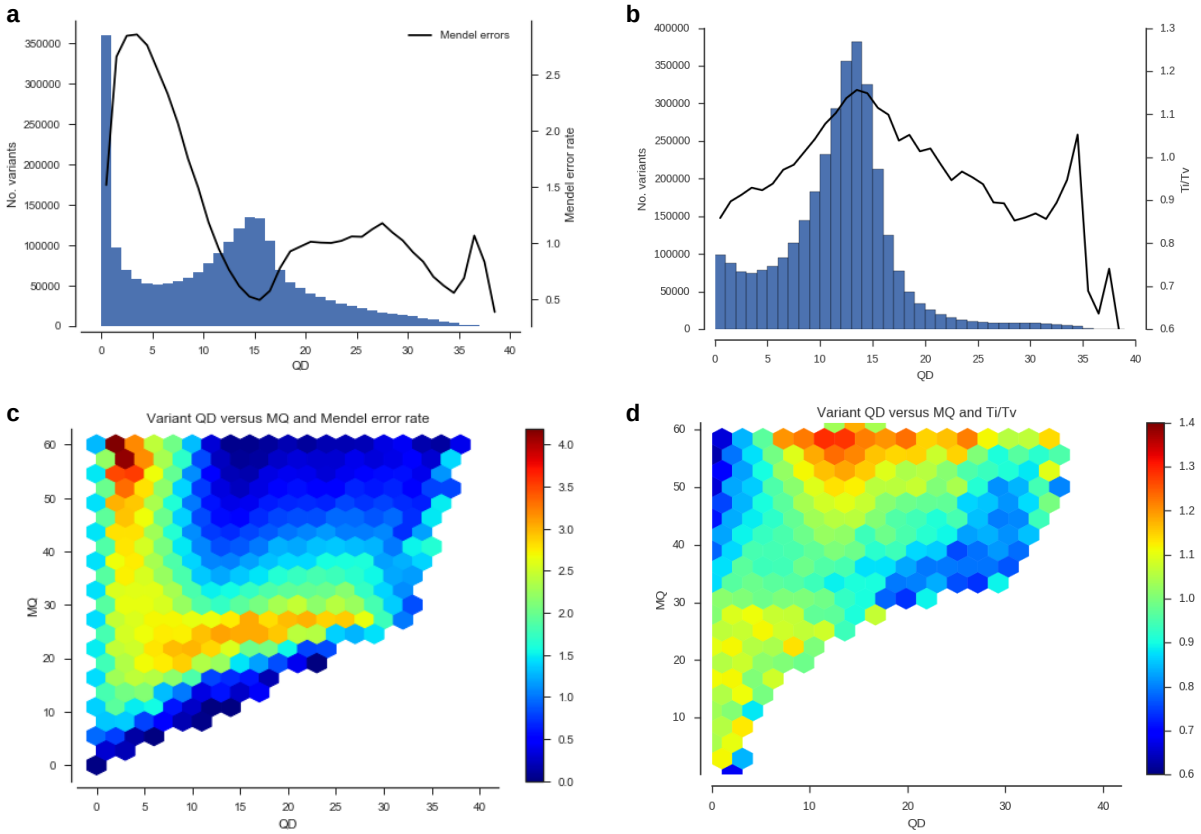

Figure 3.9. Examples of variant filter calibration analyses using Mendelian error and transition/transversion (Ti/Tv) ratio. **a**, Analysis of QD and Mendelian error in a colony cross. Histogram shows the numbers of variants segregating in the cross on chromosome arm 3R, line shows percentage of those variants with at least one Mendelian error. **b**, Analysis of QD and Ti/Tv. Histogram shows the number of variants segregating in the wild-caught samples, line shows the Ti/Tv ratio. **c**, Joint analysis of QD, MQ and Mendelian error in a colony cross. Colour denotes the percentage of SNPs with at least one Mendelian error. **d**, Joint analysis of QD, MQ and Ti/Tv. Colour denotes the Ti/Tv ratio for SNPs within each bin.

Production of an analysis-ready genome variation dataset

I produced an analysis-ready dataset in variant call format (VCF) for each chromosome arm by applying the sample and variant QC results described above as follows. From all stages of sample QC a total of 123 samples were excluded, leaving 765 samples from wild-caught specimens passing all sample QC checks. Taking the raw variant calls as input, I first removed all non-SNP variants, then removed genotype calls for individuals failing sample QC, then removed any SNPs that were no longer variant after excluding samples, using GATK *SelectVariants*. I then added INFO annotations with genome accessibility metrics, and added FILTER annotations recording the variant QC decisions, using bcftools *annotate*. Finally, I added INFO annotations with information about functional consequences of

SNPs using SNPEFF version 4.1b (Cingolani et al., 2012). After building VCF files, I also performed a conversion of these data to the HDF5 format, which is a general-purpose binary file format for scientific data that provides a significantly more performant way to load the variation data for downstream analysis. I developed the software package vcfnp[2] for performing this VCF to HDF5 format conversion.

In order to support a broader range of population genetic analyses, the SNP genotypes in the 765 wild-caught individuals were also phased into haplotypes at the 41,476,870 biallelic sites. In order to maximise the quality of the haplotypes, the phasing pipeline made use of both read-backed and statistical phasing methods (Delaneau et al., 2013). Implementation and validation of the phasing pipeline was performed by Nick Harding from the MalariaGEN Resource Centre team, and is described in more detail in The Anopheles gambiae 1000 Genomes Consortium (2017). I supported this effort by developing an efficient algorithm for phasing the parents of the colony crosses using pedigree information, allowing the parents' haplotypes to be used as a gold standard for quantifying phasing error rates.

All the analysis-ready data files were uploaded to the Wellcome Sanger Institute public FTP site and are available at `ftp://ngs.sanger.ac.uk/production/ag1000g/phase1/`. Sequence read alignments and analysis-ready variant calls were submitted to the European Nucleotide Archive (ENA; `http://www.ebi.ac.uk/ena`) under ENA study PRJEB18691. I also collaborated with members of the MalariaGEN Web Application development team to design a web application to allow interactive exploration and browsing of these data, available at `http://www.malariagen.net/apps/ag1000g`. Briefly, this web application provides the following features: queryable tables of variants and samples; interactive plotting of variants and samples; genome browser allowing visual exploration of tracks against their physical position within a chromosome arm, including genome accessibility metrics, variant quality metrics, various population genetic metrics, and visualisation of genotype calls within individual samples.

[2] `https://github.com/alimanfoo/vcfnp`

Validation with Sanger sequencing

To provide an independent estimate for the SNP false discovery rate (FDR) and sensitivity, Sanger sequencing was performed of five loci in 58 individual mosquitoes that were also analysed using whole genome sequencing. The primers were provided by the Liverpool School of Tropical Medicine, and the Sanger sequencing was performed by the MalariaGEN Resource Center team at the Wellcome Centre for Human Genetics. The individuals used in the validation analysis included representatives of all sampling locations except Kenya. Loci were sequenced using the following sets of forward and reverse primer pairs:

- 3R:6,906,804-6,908,373
  ACTGATCGAGGTAGGGCATC/CGCATCGAGTCCAATCTTTT
- 3L:22,022,788-22,024,835
  ACCCTCGTCTTTCACCACAC/CATCCACCGGAGTACTTGCT
- 3L:39,820,960-39,822,628
  GGTAGCAGTCGCCAGTTTTT/CTTGAGCGCCACCTTAAGTC
- 2L:46,845,944-46,847,857
  GGAGTTTTACGTGGCCGTTA/CTTCAAAGTGCGGATCCATT
- 2R:28,491,415-28,493,141
  TTCAAGTAGTTGCCCGCTTT/AATCGAGCTATTGCGGAGAA

For each primer set, I aligned forward and reverse traces and called SNPs and genotypes using NovoSNP version 3.0.1 (Weckx et al., 2005). I analysed only individuals which had high quality traces for both forward and reverse reads spanning at least 90% of the distance between primers. I also excluded individuals where there was evidence for heterozygous indel variation. I defined true positives, false positives and false negatives by comparing variant calls derived from the Illumina and Sanger data for the individuals sequenced with both technologies. I defined a false positive as a SNP called from Illumina sequence data for which there was no evidence on either forward or reverse Sanger traces in any individuals. I defined a false negative as a SNP for which there was evidence in at least one individual

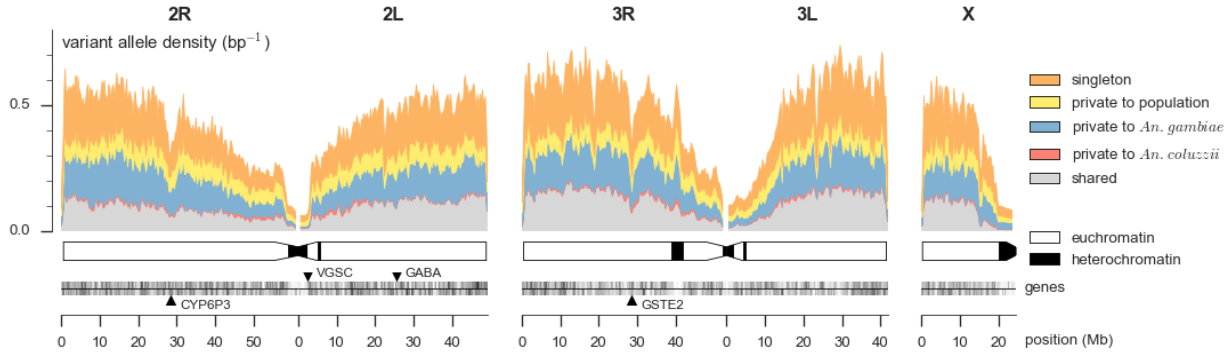

Figure 3.10. Density of variant alleles discovered across the genome. Main plot shows the density of variant alleles per accessible base, computed in 200 kb moving windows. Chromosome schematic below shows chromatin state predictions from Sharakhova et al. (2010). *Cyp6p3*, *Vgsc*, *Gaba* and *Gste2* are genes known to be involved in insecticide resistance.

on both forward and reverse Sanger traces but which was not called from Illumina sequence data. A SNP present in the Illumina calls for these individuals and in the Sanger calls was defined as a true positive. FDR was computed as the number of true positives divided by the number of true positives plus false positives. In total, there were 261 true positives and 1 false positive, an FDR of $1/262 = 0.4\%$. The false positive SNP was observed in only a single individual, as a heterozygous genotype. Sensitivity was computed as the number of true positives divided by the number of true positives plus false negatives. There were 15 false negatives at genome positions classified as accessible, a sensitivity of $261/276 = 94.6\%$. Sensitivity across the five loci ranged from 88.5%-100%. I also compared genotype calls, finding that 984/1000 (98.4%) of heterozygous genotype calls and 6881/6884 (99.96%) of homozygous genotype calls were concordant between Sanger and Illumina data.

Results

Nucleotide variation

52,525,957 SNPs were discovered in the Ag1000G phase 1 dataset passing all quality filters. Of these, 41,476,870 (79%) were biallelic and 11,049,087 (21%) were multiallelic (more than one variant allele), giving a total of 64,481,991 variant alleles. 21,783,339 variant alleles were singletons (only observed as a heterozygous genotype in a single individual), 31,228,996 were observed only in a single country and species (private to a population),

17,144,828 were found in multiple countries but observed only in *An. gambiae*, 1,252,245 were found in multiple countries but only in *An. coluzzii*, and 14,688,252 were shared between the two species.

On average one variant allele was discovered every 2.2 bases of the accessible genome, but variant alleles were not evenly distributed throughout the genome (Fig. 3.10). Variant allele density was lower towards the centromeres on all chromosome arms, and especially low within regions of pericentromeric heterochromatin where recombination is reduced (Pombi et al., 2006; Zheng et al., 1997). Diversity is also lower near centromeres in *D. melanogaster* (Langley et al., 2012) and correlated with recombination rate across a variety of species due to selection at linked sites (Charlesworth, 2012a; Elyashiv et al., 2016; Corbett-Detig et al., 2015; Burri et al., 2015; Chan et al., 2012; Spencer et al., 2006). Chromatin state was not always associated with lower variant density, however, as the region of intercalary heterochromatin on chromosome arm 3R had noticeably elevated variant density. Drops in variant allele density, and particularly the density of singleton alleles, were visible at genome regions containing the insecticide resistance genes *Cyp6p3* and *Gste2*, suggesting the presence of selective sweeps in those genome regions, investigated further in Chapter 5.

Individual samples carried between 1.6 and 2.7 million variant alleles (Fig. 3.12). There was no systematic difference between the two species, *An. gambiae* and *An. coluzzii*, in terms of the number of variant alleles discovered in each sample, indicating that the AgamP3 reference genome is not biased towards one species or the other. However, there was a clear difference in the level of variation between the two species within the centromeric region of the X chromosome, where *An. gambiae* populations had higher levels of variation than *An. coluzzii*, and thus the reference genome within that region is more representative of *An. coluzzii* (Fig. 3.13). The 2La inversion also created a notable difference between individuals both within and between populations, consistent with polymorphism and substantial sequence divergence between the two karyotypes.

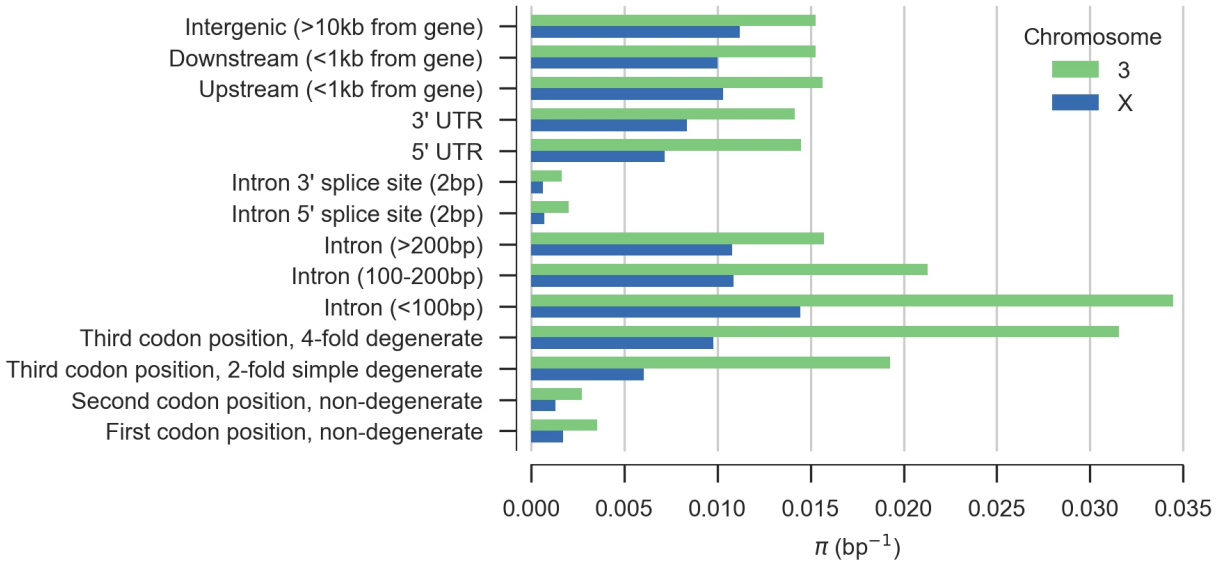

Figure 3.11. Nucleotide diversity (π) in relation to gene architecture.

Gene architecture and genetic diversity

Gene architecture is expected to exert a strong influence on the genomic landscape of nucleotide variation, because different components of protein-coding sequences are under different types and degrees of selective pressure. For example, in general, exons of protein-coding genes will be under purifying selection, because the protein is under some functional constraint (Ohta, 1996). Thus, genetic diversity within exons is expected to be lower than within introns or intergenic sequence. However, the level of constraint will depend on whether nucleotides occur in the first, second or third codon position, because third codon positions have some degeneracy and nucleotide substitutions are less likely to cause an amino acid substitution than at the first or second codon position. Although these phenomena are common to all eukaryotic genomes, there will be differences between species due to differences in their biology, genome size, mutation rate, recombination rate and effective population size. The Ag1000G data resource provides an opportunity to investigate some of these features of genome biology in the context of a eukaryotic species with large effective population sizes.

Of the 52,525,957 SNPs discovered, 4,404,390 (8.4%) occurred within a gene coding region (Table 3.2). Of the SNPs within gene coding regions, 1,809,532 (41.1%) were

non-synonymous (missense) SNPs that altered the amino acid sequence of the gene, and the remaining 2,594,858 (58.9%) were synonymous. To further investigate the impact of gene architecture on levels of nucleotide variation, I computed nucleotide diversity (π) in the phase 1 cohort within different types of coding and non-coding sites (Fig. 3.11). I computed diversity separately within the X chromosome and the autosomal Chromosome 3, because the X chromosome is hemizygous in males and thus has a lower effective population size as well as experiencing less frequent recombination than the autosomes, both of which may affect genetic diversity. The lowest diversity was found within intronic splice sites, consistent with a high degree of purifying selection, because any change in gene splicing is likely to have major functional consequences. The reduction of diversity at splice sites was strongest within the first 2 bp of the intronic sequence, but also extended up to 9 bp into the intron at the 5' end, although not at the 3' end (Fig. 3.14). Diversity was also low within exons at non-degenerate first and second codon positions, where any nucleotide substitution will cause an amino acid substitution, also consistent with strong purifying selection. In contrast, diversity was much higher at fourfold degenerate coding sites, reaching 3.2% on Chromosome 3, consistent with a lack of purifying selection. The only sequence class where diversity was higher was within short (<100 bp) introns, where π on Chromosome 3 was approaching 3.5%.

Two other results of this analysis are interesting. Firstly, on Chromosome 3 diversity was more than twice as high in short (<100 bp) introns than long (>200 bp) introns. This suggests that longer introns are in general subject to greater functional constraint, for example, due to the presence of regulatory sequences. Shorter introns might be too small to contain regulatory sequences, or might be younger and thus have had less time to evolve a regulatory function. A similar pattern is seen in other insects and arthropods (Singh et al., 2013; Martin et al., 2016; Lynch et al., 2017), and is consistent with evidence in *Drosophila* that longer non-coding sequences are more likely to contain functional elements (Casillas et al., 2007). Secondly, diversity in intergenic regions >10 kb distant from a gene was 1.5%, similar to that in longer introns, and much lower than short introns or degenerate coding positions. Diversity in intergenic regions was similar regardless of distance to a

gene, suggesting that lower diversity is not due to genetic linkage and linked purifying selection, but reflects some substantial selective constraint on intergenic sequence.

There were also notable differences between the autosomal Chromosome 3 and the X chromosome in these patterns of diversity in relation to gene architecture. Diversity was generally lower on the X chromosome across all site classes, as expected due to the fact that males carry only a single X chromosome and thus the effective population size is lower for the X chromosome than for the autosomes. However, the ratio of diversity between sex chromosomes and autosomes (X/A ratio) differed between site classes (Fig. 3.15). The X/A ratio was highest within intergenic regions >10 kb from a gene, with values between 0.63-0.76, close to the theoretical expectation of 0.75 due to reduced effective population size. In contrast, the X/A ratio was lowest (0.27-0.32) at degenerate coding positions. This suggests that genetic linkage and purifying selection at linked sites is a substantially stronger force reducing diversity on the X chromosome than the autosomes, because the X/A ratio is highest at sites furthest from genes. This is consistent with the fact that meiotic recombination occurs only in females and is less frequent on the X chromosome than autosomes where recombination occurs in both sexes. The X/A ratio can also be influenced by demographic factors, including population size fluctuations and sex biases in reproductive success and migration (Corbett-Detig et al., 2015; Webster and Wilson Sayres, 2016; Pool and Nielsen, 2007; Charlesworth, 2012b), and these may also be playing a role, although the X/A ratios were relatively stable across the different species and populations sampled in Ag1000G phase 1.

Conclusions

In this chapter I have described the discovery and genotyping of more than 50 million single nucleotide polymorphisms from deep whole genome sequencing of individual mosquitoes from natural populations of *An. gambiae* and *An. coluzzii*. This is a unique data resource, providing opportunities to study both fundamental genome biology and the evolution and demography of natural malaria vector populations. I have shown that this is a highly robust data resource, with low rates of false discovery and highly accurate genotype calls.

I have also used this resource to show that nucleotide diversity is extremely high within these mosquito species, confirming that they are among the most genetically diverse organisms studied to date (Leffler et al., 2012). In the next chapter I will use these data to investigate genetic structure among the populations sampled, and to quantify, characterise and compare genetic diversity within these populations in more detail.

Acknowledgments

I would like to thank the members of the *Anopheles gambiae* 1000 Genomes Consortium, the MalariaGEN Resource Centre team, and the staff at the Wellcome Sanger Institute sample logistics, sequencing and informatics facilities, for their contributions towards the production of the Ag1000G phase 1 data resource.

Supplemental figures

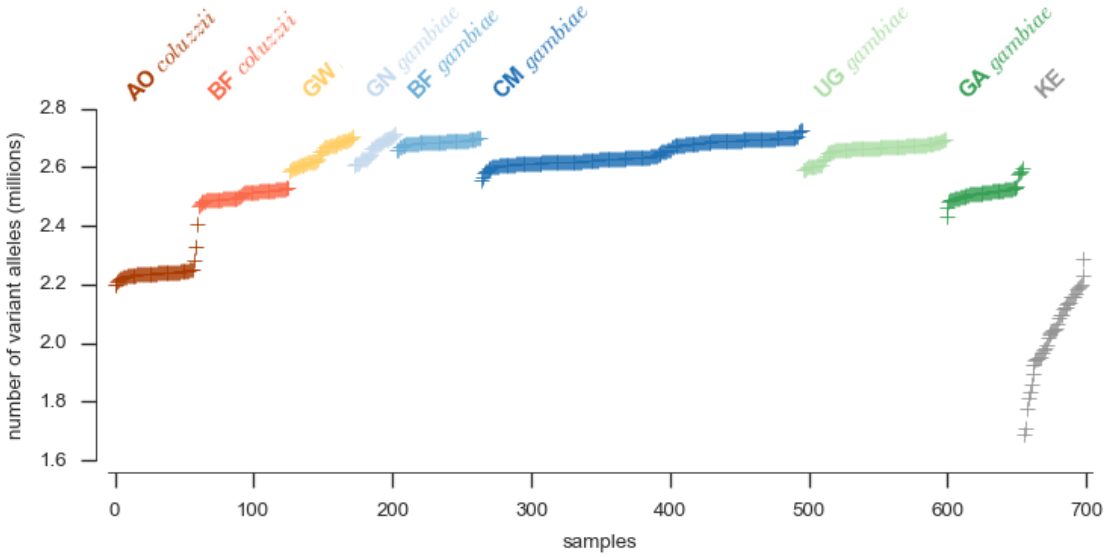

Figure 3.12. Number of variant alleles by sample. Each marker represents an individual sample, coloured by country and species. Number of variant alleles is computed over the whole genome.

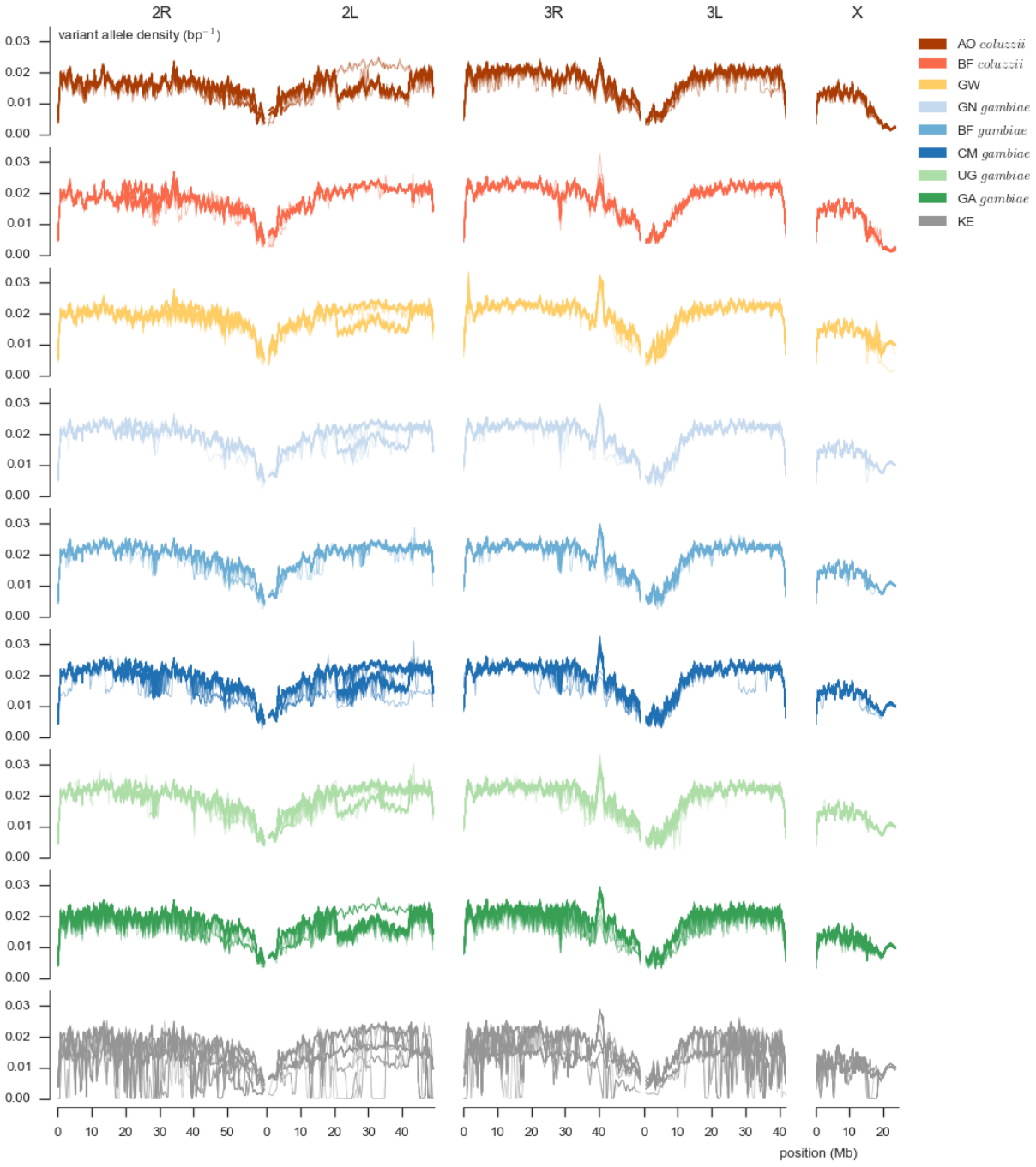

Figure 3.13. Variant allele density by sample. Each individual line plots variant allele density (number of variant alleles per accessible base) for a single sample in 200 kb windows across the genome. Individuals are grouped by country and species to facilitate comparison within and between populations.

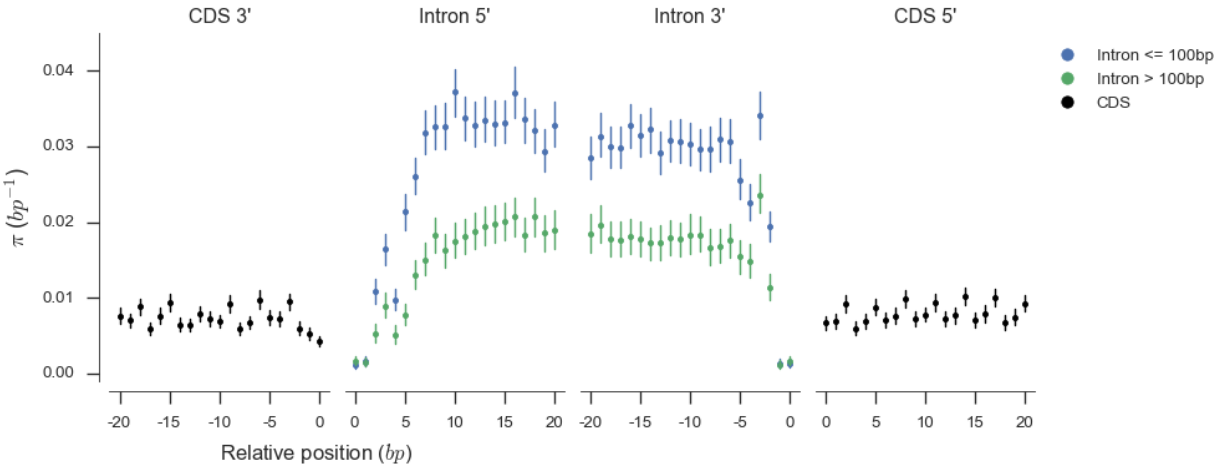

Figure 3.14. Nucleotide diversity (π) in relation to intron/exon boundaries. Plots show diversity averaged across all genes on chromosome arm 3R, error bars show results from bootstrapping over sites.

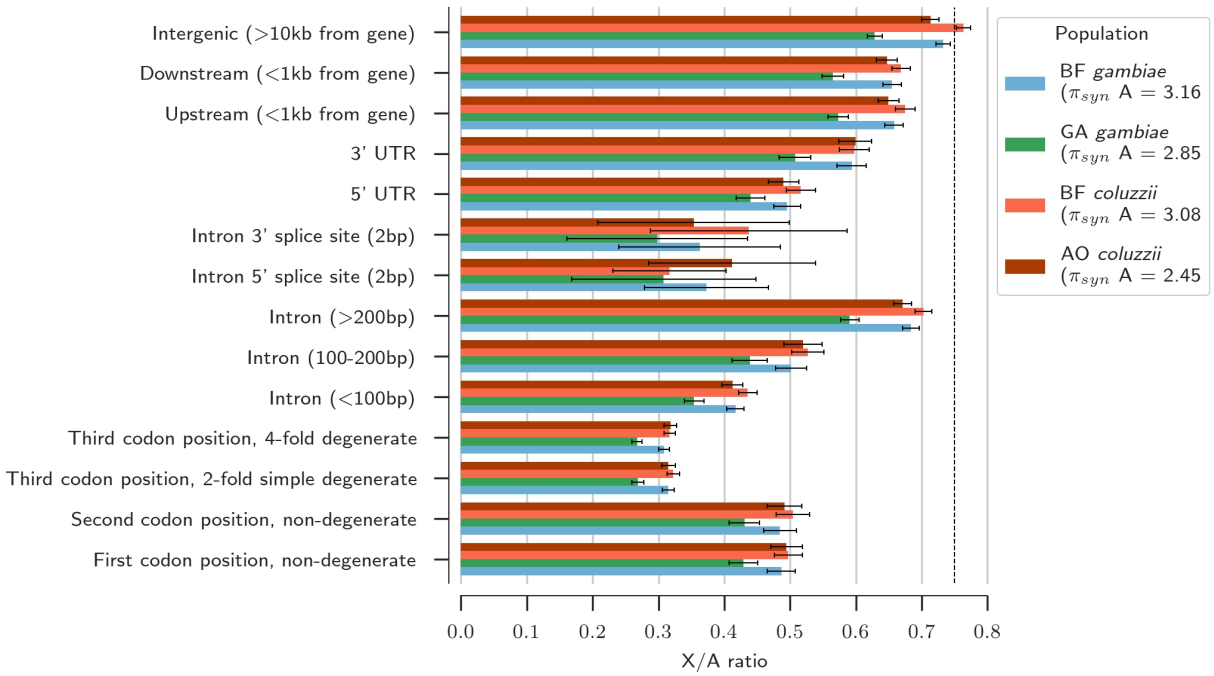

Figure 3.15. Ratios of nucleotide diversity between the sex chromosome (X) and the autosomes (represented by Chromosome 3). Error bars computed by bootstrapping over sites. The theoretical expectation for a diploid species with one sex hemizygous for the X chromosome is 0.75 (shown as dashed line) due to differences in effective population sizes.

Supplemental tables

Table 3.2. Number of SNPs discovered, grouped by sequence ontology annotation class as determined by SNPEFF (Cingolani et al., 2012).

Annotation	No. SNPs
intergenic_region	21,932,312
intron_variant	11,040,938
upstream_gene_variant	8,960,291
downstream_gene_variant	5,012,563
synonymous_variant	2,594,858
missense_variant	1,809,532
3_prime_UTR_variant	480,546
5_prime_UTR_variant	325,007
intragenic_variant	280,084
splice_region_variant	186,102
5_prime_UTR_premature_start_codon_gain_variant	60,784
stop_gained	15,744
splice_donor_variant	3,946
splice_acceptor_variant	3,117
stop_retained_variant	2,472
stop_lost	1,546
start_lost	1,444
initiator_codon_variant	257
non_canonical_start_codon	10

References

Altschul, SF, W Gish, W Miller, EW Myers and DJ Lipman (1990). 'Basic local alignment search tool'. In: *J. Mol. Biol.* 215.3, pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Benson, G (1999). 'Tandem repeats finder: A program to analyze DNA sequences'. In: *Nucleic Acids Res.* 27.2, pp. 573–580. DOI: 10.1093/nar/27.2.573.

Burri, R et al. (2015). 'Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers.' In: *Genome Res.* 25.11, pp. 1656–1665. DOI: 10.1101/gr.196485.115.

Casillas, S, A Barbadilla and CM Bergman (2007). 'Purifying Selection Maintains Highly Conserved Noncoding Sequences in Drosophila'. In: *Mol. Biol. Evol.* 24.10, pp. 2222–2234. DOI: 10.1093/molbev/msm150.

Chan, AH, PA Jenkins and YS Song (2012). 'Genome-wide fine-scale recombination rate variation in Drosophila melanogaster.' In: *PLoS Genet.* 8.12. Ed. by G McVean, e1003090. DOI: 10.1371/journal.pgen.1003090.

Charlesworth, B (2012a). 'The Effects of Deleterious Mutations on Evolution at Linked Sites'. In: *Genetics* 190.1, pp. 5–22. DOI: 10.1534/genetics.111.134288.

Charlesworth, B (2012b). 'The Role of Background Selection in Shaping Patterns of Molecular Evolution and Variation: Evidence from Variability on the Drosophila X Chromosome'. In: *Genetics* 191.1, pp. 233–246. DOI: 10.1534/genetics.111.138073.

Cingolani, P, A Platts, LL Wang, M Coon, T Nguyen, L Wang, SJ Land, X Lu and DM Ruden (2012). 'A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff'. In: *Fly* 6.2, pp. 80–92. DOI: 10.4161/fly.19695.

Coluzzi, M, A Sabatini, A della Torre, MA di Deco and V Petrarca (2002). 'A polytene chromosome analysis of the Anopheles gambiae species complex'. In: *Science* 298.5597, pp. 1415–1418. DOI: 10.1126/science.1077769.

Corbett-Detig, RB, DL Hartl and TB Sackton (2015). 'Natural Selection Constrains Neutral Diversity across A Wide Range of Species'. In: *PLOS Biol.* 13.4. Ed. by NH Barton, e1002112. DOI: 10.1371/journal.pbio.1002112.

Delaneau, O, B Howie, AJ Cox, JF Zagury and J Marchini (2013). 'Haplotype estimation using sequencing reads.' English. In: *Am. J. Hum. Genet.* 93.4, pp. 687–696. DOI: 10.1016/j.ajhg.2013.09.002.

Depristo, MA et al. (2011). 'A framework for variation discovery and genotyping using next-generation DNA sequencing data'. In: *Nat. Genet.* 43.5, pp. 491–498. DOI: 10.1038/ng.806.

Elyashiv, E, S Sattath, TT Hu, A Strutsovsky, G McVicker, P Andolfatto, G Coop and G Sella (2016). 'A Genomic Map of the Effects of Linked Selection in Drosophila.' In: *PLoS Genet.* 12.8. Ed. by NH Barton, e1006130. DOI: 10.1371/journal.pgen.1006130.

Fernández-Medina, RD, CJ Struchiner and JM Ribeiro (2011). 'Novel transposable elements from Anopheles gambiae'. In: *BMC Genomics* 12.1. DOI: 10.1186/1471-2164-12-260.

Guo, Y, F Ye, Q Sheng, T Clark and DC Samuels (2013). 'Three-stage quality control strategies for DNA re-sequencing data'. In: *Brief. Bioinform.* 15.6, pp. 879–889. DOI: 10.1093/bib/bbt069.

Head, SR, H Kiyomi Komori, SA LaMere, T Whisenant, F Van Nieuwerburgh, DR Salomon and P Ordoukhanian (2014). 'Library construction for next-generation sequencing: Overviews and challenges'. In: *Biotechniques* 56.2, pp. 61–77. DOI: 10.2144/000114133.

Holt, RA et al. (2002). 'The genome sequence of the malaria mosquito Anopheles gambiae'. In: *Science* 298.5591, pp. 129–149. DOI: 10.1126/science.1076181.

Jun, G, M Flickinger, KN Hetrick, JM Romm, KF Doheny, GR Abecasis, M Boehnke and HM Kang (2012). 'Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data'. In: *Am. J. Hum. Genet.* 91.5, pp. 839–848. DOI: 10.1016/j.ajhg.2012.09.004.

Langley, CH et al. (2012). 'Genomic Variation in Natural Populations of Drosophila melanogaster'. In: *Genetics* 192.2, pp. 533–598. DOI: 10.1534/genetics.112.142018.

Laurie, CC et al. (2010). 'Quality control and quality assurance in genotypic data for genome-wide association studies'. In: *Genet. Epidemiol.* 34.6, pp. 591–602. DOI: 10.1002/gepi.20516.

Leffler, EM, K Bullaughey, DR Matute, WK Meyer, L Ségurel, A Venkat, P Andolfatto and M Przeworski (2012). 'Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species?' In: *PLoS Biol.* 10.9, e1001388. DOI: 10.1371/journal.pbio.1001388.

Li, H and R Durbin (2009). 'Fast and accurate short read alignment with Burrows-Wheeler transform.' In: *Bioinformatics* 25.14, pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

Lynch, M, R Gutenkunst, M Ackerman, K Spitze, Z Ye, T Maruki and Z Jia (2017). 'Population Genomics of Daphnia pulex'. In: *Genetics* 206.1, pp. 315–332. DOI: 10.1534/genetics.116.190611.

Martin, SH, M Möst, WJ Palmer, C Salazar, WO McMillan, FM Jiggins and CD Jiggins (2016). 'Natural selection and genetic diversity in the butterfly Heliconius melpomene'. In: *Genetics* 203.1, pp. 525–541. DOI: 10.1534/genetics.115.183285.

McKenna, A et al. (2010). 'The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data'. In: *Genome Res.* 20.9, pp. 1297–1303. DOI: 10.1101/gr.107524.110.

Meynert, AM, M Ansari, DR FitzPatrick and MS Taylor (2014). 'Variant detection sensitivity and biases in whole genome and exome sequencing.' In: *BMC Bioinf.* 15.1, p. 247. DOI: 10.1186/1471-2105-15-247.

Ohta, T (1996). 'The neutral theory is dead. The current significance and standing of neutral and nearly neutral theories'. In: *BioEssays* 18.8, pp. 673–677. DOI: 10.1002/bies.950180811.

Patterson, N, AL Price and D Reich (2006). 'Population structure and eigenanalysis'. In: *PLoS Genet.* 2.12, e190. DOI: 10.1371/journal.pgen.0020190.

Pilipenko, VV, H He, BG Kurowski, ES Alexander, X Zhang, L Ding, TB Mersha, L Kottyan, DW Fardo and LJ Martin (2014). 'Using Mendelian inheritance errors as quality control criteria in whole genome sequencing data set'. In: *BMC Proc.* 8.S1, S21. DOI: 10.1186/1753-6561-8-S1-S21.

Pombi, M, AD Stump, A Della Torre and NJ Besansky (2006). 'Variation in recombination rate across the X chromosome of Anopheles gambiae'. In: *Am. J. Trop. Med. Hyg.* 75.5, pp. 901–903. DOI: 10.4269/ajtmh.2006.75.901.

Pool, JE and R Nielsen (2007). 'Population size changes reshape genomic patterns of diversity'. In: *Evolution* 61.12, pp. 3001–3006. DOI: 10.1111/j.1558-5646.2007.00238.x.

Saunders, IW, J Brohede and GN Hannan (2007). 'Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference'. In: *Genomics* 90.3, pp. 291–296. DOI: 10.1016/j.ygeno.2007.05.011.

Sayre, R et al. (2013). 'A new map of standardized terrestrial ecosystems of Africa'. In: *African Geogr. Rev.*

Sharakhova, MV, MP Hammond, NF Lobo, J Krzywinski, MF Unger, ME Hillenmeyer, RV Bruggner, E Birney and FH Collins (2007). 'Update of the Anopheles gambiae pest genome assembly'. In: *Genome Biol.* 8.1, R5. DOI: 10.1186/gb-2007-8-1-r5.

Sharakhova, MV, P George, IV Brusentsova, SC Leman, JA Bailey, CD Smith and IV Sharakhov (2010). 'Genome mapping and characterization of the Anopheles gambiae heterochromatin.' In: *BMC Genomics* 11.1, p. 459. DOI: 10.1186/1471-2164-11-459.

Singh, ND, JD Jensen, AG Clark and CF Aquadro (2013). 'Inferences of demography and selection in an African population of Drosophila melanogaster.' In: *Genetics* 193.1, pp. 215–28. DOI: 10.1534/genetics.112.145318.

Smit, AFA, R Hubley and P Green (2013). *RepeatMasker Open-4.0*.

Spencer, CCA, P Deloukas, S Hunt, J Mullikin, S Myers, B Silverman, P Donnelly, D Bentley and G McVean (2006). 'The Influence of Recombination on Human Genetic Diversity'. In: *PLoS Genet.* 2.9. Ed. by JD Wall, e148. DOI: 10.1371/journal.pgen.0020148.

The 1000 Genomes Project Consortium (2015). 'A global reference for human genetic variation'. In: *Nature* 526.7571, pp. 68–74. DOI: 10.1038/nature15393.

The Anopheles gambiae 1000 Genomes Consortium (2017). 'Genetic diversity of the African malaria vector Anopheles gambiae'. In: *Nature* 552.7683, pp. 96–100. DOI: 10.1038/nature24995.

Tu, Z and C Coates (2004). 'Mosquito transposable elements'. In: *Insect Biochem. Mol. Biol.* 34.7, pp. 631–644. DOI: 10.1016/j.ibmb.2004.03.016.

Van der Auwera, GA et al. (2013). 'From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline'. In: *Curr. Protoc. Bioinforma.* 43.1. DOI: 10.1002/0471250953.bi1110s43.

Webster, TH and MA Wilson Sayres (2016). 'Genomic signatures of sex-biased demography: progress and prospects'. In: *Curr. Opin. Genet. Dev.* 41, pp. 62–71. DOI: 10.1016/j.gde.2016.08.002.

Weckx, S, J Del-Favero, R Rademakers, L Claes, M Cruts, P De Jonghe, C Van Broeckhoven and P De Rijk (2005). 'novoSNP, a novel computational tool for sequence variation discovery.' In: *Genome Res.* 15.3, pp. 436–442. DOI: 10.1101/gr.2754005.

Zheng, L, AJ Cornel, R Wang, H Erfle, H Voss, W Ansorge, FC Kafatos and FH Collins (1997). 'Quantitative trait loci for refractoriness of Anopheles gambiae to Plasmodium cynomolgi B.' In: *Science* 276.5311, pp. 425–428. DOI: 10.1126/science.276.5311.425.

4 Population structure and genetic diversity

In this chapter I investigate genetic population structure and diversity among the 765 wild-caught mosquitoes sequenced in Ag1000G phase 1, using the genome-wide data resource on nucleotide variation described in the previous chapter. I explore evidence for genetic differentiation between populations from different species and geographical locations, and investigate heterogeneity in rates of gene flow between these populations. I quantify and characterise genetic diversity within these populations, and provide evidence for contrasting population size histories. I also explore the impact of nucleotide variation within these populations on the availability of potential gene drive targets. These analyses provide new insights into the complex demography of An. gambiae and An. coluzzii populations, and a firm foundation from which to explore the evolution of insecticide resistance in subsequent chapters.

Introduction

In the previous chapter I described the *Anopheles gambiae* 1000 Genomes Project (Ag1000G) phase 1 data resource, which comprises genome variation data from 765 individual mosquitoes sampled from 8 countries spanning sub-Saharan Africa, and includes representation of both *An. gambiae* and *An. coluzzii*. The availability of genomic data from multiple species and geographical locations provides an opportunity to investigate many facets of their population biology and demography. In this chapter I describe analyses of the Ag1000G phase 1 data investigating genetic population structure among the mosquito populations sampled, and characterising levels of genetic diversity within and differentiation between populations. These analyses are interesting because they allow us to begin to

build a picture of the underlying demography of these populations, including variation in population size over time and space, and in the degree of connectivity and hence gene flow between populations. Exploring these heterogeneities is particularly relevant because *An. gambiae* and *An. coluzzii* both have an extremely broad geographical and ecological range (della Torre et al., 2001; Tene Fossog et al., 2015; Wiebe et al., 2017). *An. coluzzii* is found from the West Coast throughout West and Central Africa. The range of *An. gambiae* overlaps that of *An. coluzzii* and extends across the Great Rift to the East coast, stretching down to South Africa. Both species' ranges span the equator and encompass a remarkably diverse range of environments, including coastal, savanna, sahel and rainforest. Human population density and land use, and the history and current coverage of malaria vector control interventions, also vary substantially throughout this range (Binswanger-Mkhize and Savastano, 2017; WHO, 2019). While we do not have the sampling resolution to attempt to correlate any of these individual variables with genetic features of mosquito populations, we can begin to highlight major variations between populations and generate hypotheses for further investigation.

These analyses of population structure and diversity were carried out as part of a broader investigation of population history and demography performed by the Ag1000G Consortium Analysis Working Group. In this chapter I focus on the analyses that I led and performed individually. However, to provide some additional context I also mention some analyses in which I worked together with other Consortium members, and indicate the contributions of others in the relevant sections.

Results

The influence of genome architecture on population structure

Investigating genetic population structure means studying the extent to which individual mosquitoes are genetically related to each other, and identifying groups of individuals which are more or less related. The Ag1000G phase 1 resource comprises data on more than 52 million single nucleotide polymorphisms (SNPs) distributed throughout the genome, and

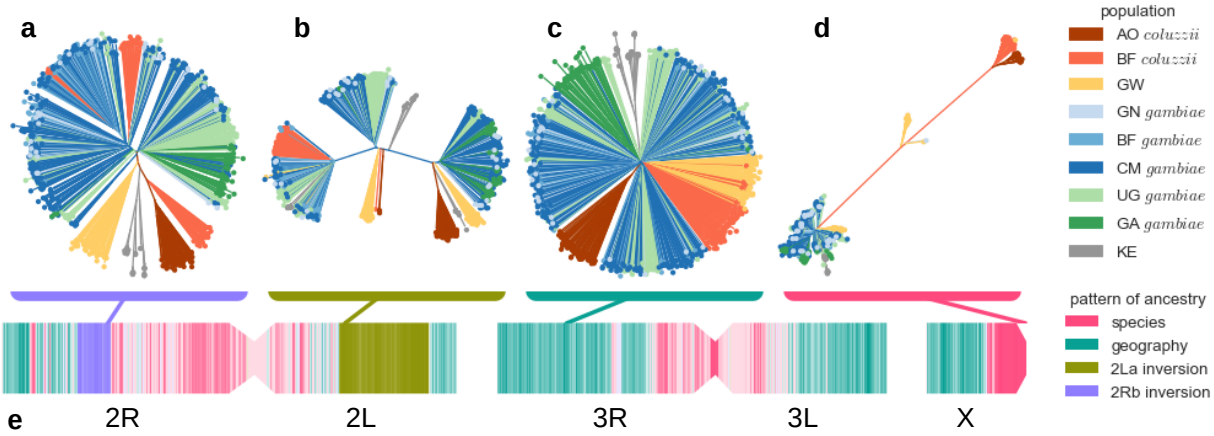

Figure 4.1. Variation across the genome in patterns of relatedness between individual mosquitoes. Panels **a-d** each show a neighbour-joining tree computed from pairwise genetic distances within a specific 200 kb genomic window, illustrating the four most common patterns of relatedness found throughout the genome. Each node in the tree represents a single individual, coloured by population. **a**, Example of tree within the 2Rb inversion. **b**, Example of tree within the 2La inversion. **c**, Example of tree from euchromatic region of Chromosome 3. **d**, Example of tree pericentromeric region of the X chromosome. **e**, Painting of the genome illustrating where the four major patterns of relatedness are found. Each 200 kb window is painted with a colour to indicate which of the four major patterns of relatedness it is most strongly correlated with.

thus provides an extremely rich and high-resolution source of information with which to investigate population structure. However, a complication arises because previous genomic studies of *An. gambiae* and *An. coluzzii* have shown that different regions of the genome can convey different information about how individuals are related to each other (Fontaine et al., 2014). In particular, several studies have found that differentiation between the two species *An. gambiae* and *An. coluzzii* is particularly high within certain genome regions but lower or almost absent elsewhere (Turner et al., 2005; White et al., 2010; Weetman et al., 2011; Cruickshank and Hahn, 2014). There are also a number of large chromosomal inversions that are polymorphic within both of these species (della Torre et al., 2001; Coluzzi et al., 2002). Chromosomal inversions cause a greatly reduced rate of recombination between different karyotypes (Stump et al., 2007), which in turn will affect patterns of relatedness between individuals. Any analysis of population structure that fails to take these heterogeneities into account will struggle to present a clear picture.

To explore the relationship between genetic population structure and genome architecture, I divided the genome into non-overlapping 200 kb windows, and then computed pairwise genetic distance between individuals within each window separately. For each window I

computed a neighbour-joining tree and visualised each of the resulting trees as an unrooted dendrogram. Several qualitatively different tree topologies were evident in different genome regions (e.g., Fig. 4.1a-d). For example, within the pericentromeric region of the X chromosome, trees showed extremely strong clustering by species (e.g., Fig. 4.1d). In contrast, trees from euchromatic regions of Chromosome 3 showed some clustering by geographical location, but no clustering by species at all (e.g., Fig. 4.1c). Trees within the 2Rb and 2La inversions also had unique topologies, consistent with clustering by inversion karyotype (e.g., Fig. 4.1a,b).

To analyse these variations in tree topology systematically, I computed the Pearson correlation coefficient between genetic distance matrices from all pairs of genomic windows. I then performed dimensionality reduction on the resulting correlation matrix via multidimensional scaling, to identify common patterns of relatedness found in multiple genome windows. The first three principal coordinates (PCs) from this analysis displayed a strong association with specific patterns of relatedness and genome regions. To visualise these results, I devised a transformation from these first three PCs to different colours representing the different patterns of relatedness, and used these to paint the associated genome windows (Fig. 4.1e). The first PC identified the common pattern of relatedness found throughout the 2La inversion. The third PC identified the pattern of relatedness found within the 2Rb inversion. The second PC identified the contrast between the highly species-driven patterns of relatedness found generally in pericentromeric regions, particularly of the X chromosome, and the more geographically-driven patterns found in euchromatic regions of X chromosome and Chromosome 3. There were also a minority of genome windows which did not display a strong correlation any of the four major patterns of relatedness, shown in Fig 4.1e as paler colours. These included the windows spanning the insecticide resistance gene *Vgsc* which is found near to the centromere of chromosome arm 2L, and which is known to have experienced adaptive introgression between *An. gambiae* and *An. coluzzii* (Clarkson et al., 2014; Norris et al., 2015). This provides a clue that windows affected by strong positive selection may display unusual patterns of relatedness due to adaptive gene flow between countries and/or species, explored further in Chapter 6.

It is still not clear why we observe such a stark contrast between the pericentromeric and euchromatic regions of the X chromosome, and to a lesser extent Chromosome 3. One factor that undoubtedly plays some part is the reduction in the rate of recombination towards the centromeres (Cruickshank and Hahn, 2014). Reduced recombination means that selection at linked sites will play a stronger role. This includes purifying selection, which acts to reduce genetic diversity and could accelerate the fixation of different alleles between the two species. In the previous chapter we saw clearly that the level of nucleotide variation was much reduced towards pericentromeric genome regions. This could also include divergent selection acting on speciation genes (Wolf and Ellegren, 2017). The genes involved in maintaining reproductive isolation between *An. gambiae* and *An. coluzzii* still remain a mystery, but if one or more key genes were located towards the pericentromeric region of the X chromosome, then they would be expected to have a substantial effect on the surrounding genome region. Experimental studies have found a key role for the pericentromeric region of the X chromosome in determining assortative mating behaviour in these species (Aboagye-Antwi et al., 2015). *An. gambiae* and *An. coluzzii* hybrids are fully fertile, and the two species are known to undergo some hybridisation in natural populations, the degree of which may vary over both space and time (Weetman et al., 2011; Lee et al., 2013). However, genome regions linked to speciation genes would display very little if any evidence for gene flow between the species due to selection against hybrids, whereas unlinked regions would be much less constrained and alleles could move more freely. In any case, resolving the causes of these contrasting patterns of relatedness is beyond the scope of this thesis, and remains an area of active research and debate. For the present purposes, I continued my investigation of population structure, diversity and differentiation using only the euchromatic regions of Chromosome 3, because this is unaffected by large polymorphic inversions or regions of reduced recombination.

Population structure

To further explore patterns of genetic population structure I used SNPs from euchromatic regions of Chromosome 3 to perform a principal components analysis (PCA) (Patterson

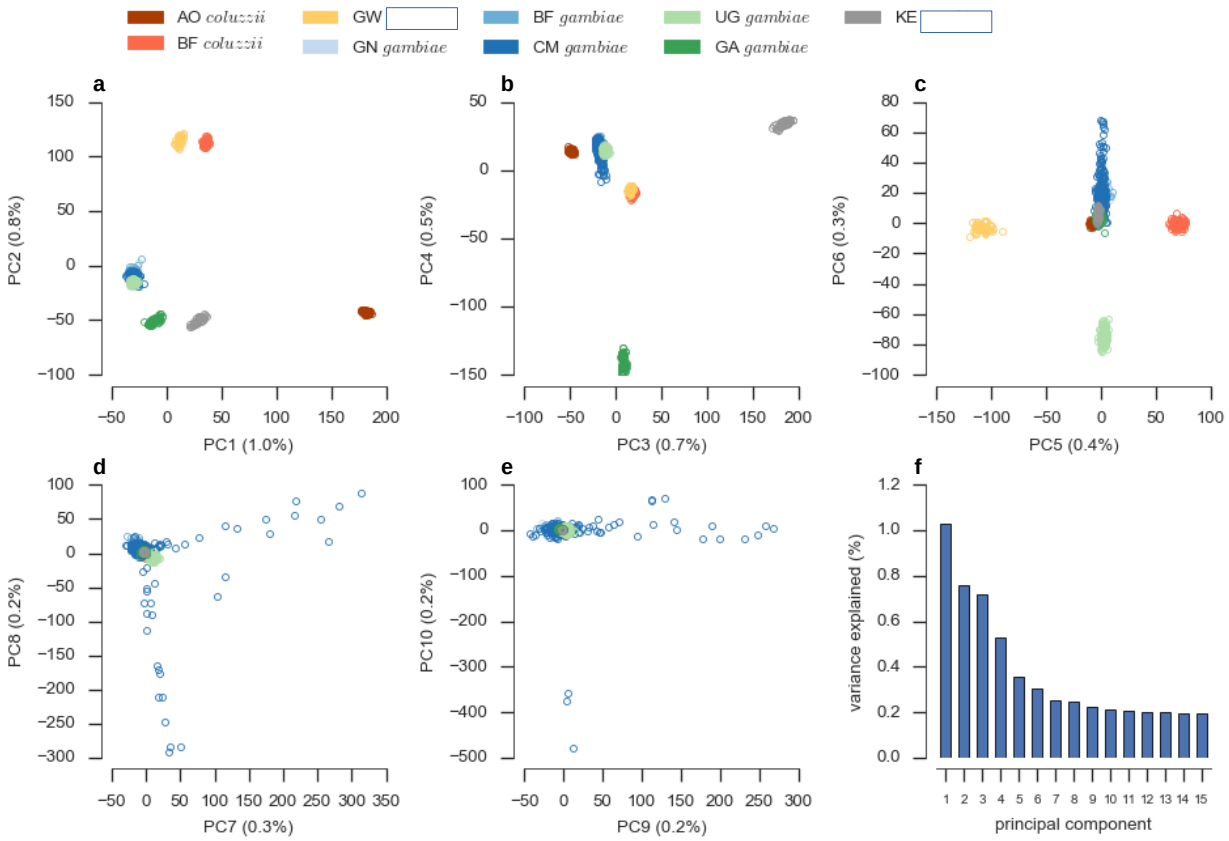

Figure 4.2. Principal components analysis of SNP genotypes in the 765 wild-caught mosquitoes in the Ag1000G phase 1 cohort. **a**–**e**, Scatter plots show principal components (PCs) 1–10, each marker is an individual mosquito. **f**, Bar plot shows the variance explained by the first 15 PCs.

et al., 2006) (Fig. 4.2). This analysis is a form of dimensionality reduction, and condenses the information present in the hundreds of thousands of SNPs into a smaller number of principal components (PCs) that explain the maximum amount of variance within the data. When applied to genetic variation data, each principal component usually identifies two or more groupings of individuals, such that genetic relatedness is higher within groups than between them. PCA in fact has a direct genealogical interpretation, and under certain conditions each principal component will capture a historical divergence between populations (McVean, 2009). In practice the interpretation of PCA is not always so simple, because demographic complexities such as admixture events and technical factors such as variations sample size will affect the shape of the result. Nevertheless, PCA allows us to identify genetically distinct populations, by examining how samples group together within the highest PCs.

The first four PCs were clearly elevated above lower PCs in the amount of variance explained, and between them revealed six distinct groupings of individuals (Fig. 4.2). These six groups highlighted both geographical and species divisions. Within *An. gambiae*, individuals from Gabon formed a distinct group, and the remaining individuals from Guinea, Burkina Faso, Cameroon and Uganda grouped together. Within *An. coluzzii*, individuals from Burkina Faso and Angola each formed a distinct group. The two remaining groups comprised individuals from Guinea-Bissau and Kenya respectively. The species status of these two groups was uncertain, because individuals from Guinea-Bissau displayed a mixture of species genotypes according to conventional molecular assays, and those assays were not performed on the Kenyan samples. I return to the question of species assignment for these groups below. PC5 revealed a further split within *An. gambiae* between Uganda and the remaining individuals. PC6 emphasized the distinction between Burkina Faso *An. coluzzii* and Guinea-Bissau, which were visibly distinct also in higher components but to a lesser degree. PCs 7–10 all displayed some structuring among the Cameroon *An. gambiae*, with individuals from the southern-most collection site spreading out away from the main group of individuals collected from the other more northerly sites.

Population differentiation

To further investigate genetic population structure, I computed the pairwise average F_{ST} between all pairs of populations defined by species and country (Fig. 4.3a). The F_{ST} statistic summarises the extent to which allele frequencies differ between two groups of individuals, and provides information about the degree of differentiation between the two sampled populations (Rousset, 1997; Holsinger and Weir, 2009; Bhatia et al., 2013). In general, higher F_{ST} values indicate a lower rate of gene flow and greater genetic drift between two populations. I also computed the rate of doubleton (f_2) variant sharing between the same pairs of populations (Fig. 4.3b). Doubleton variants are those where the alternate allele is only observed twice. In general, doubleton variants are likely to be enriched for recent mutations, because it takes longer time for alleles to reach higher frequencies. Thus patterns of sharing of doubleton variants between individuals are more

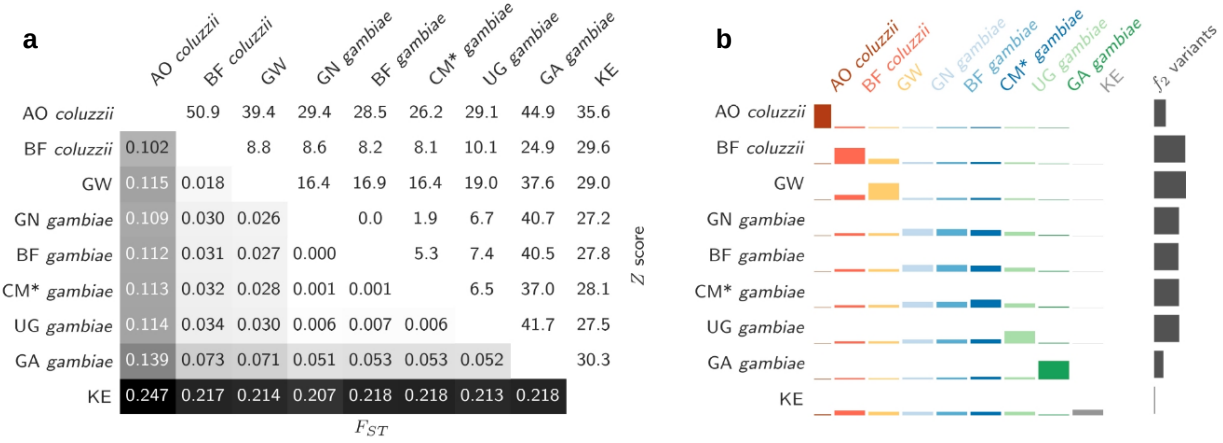

Figure 4.3. Measures of genetic differentiation between pairs of populations defined by species and country of collection, using SNPs from euchromatic regions of Chromosome 3. **a**, Pairwise average F_{ST}. The lower triangle shows F_{ST} values, the upper triangle shows Z scores computed using the block-jackknife process described by Bhatia et al. (2013). **b**, Doubleton (f_2) variant sharing between populations. Each bar shows the number of doubleton variants shared between two populations (or found within a single population on the diagonal). Bar heights are normalised within each row to show the relative rate of doubleton sharing with other populations. Grey bars at the right margin show the total numbers of doubleton variants found for each population.

indicative of recent patterns of relatedness (The 1000 Genomes Project Consortium, 2012).

These analyses confirmed relatively strong differentiation between Gabon and other *An. gambiae* populations, with $F_{ST} > 0.051$. In contrast, F_{ST} values between the other four *An. gambiae* populations were an order of magnitude smaller, at most 0.007. The degree of differentiation between Gabon and other *An. gambiae* populations was greater than that found between the two species within Burkina Faso, where both species were sampled at the same location ($F_{ST} = 0.031$). The higher F_{ST} values involving Gabon cannot simply be a reflection of greater geographical distance, because the distance between Uganda and *An. gambiae* populations to the west is greater, yet F_{ST} values are much lower. These results indicate that the Gabon individuals are separated by other *An. gambiae* populations to the north by some significant barrier to gene flow. The obvious candidate for such a barrier is the equatorial rainforest, which represents a major ecological discontinuity. Similarly, *An. coluzzii* from Angola were highly differentiated from *An. coluzzii* from Burkina Faso ($F_{ST} = 0.102$). Again, the level of differentiation was greater than that between the two species within Burkina Faso. Angola is the southern-most sampling location in the Ag1000G phase 1 cohort, and provides further support for a north/south ecogeographical

barrier between mosquito populations, affecting both species.

A surprising result was the near-total lack of differentiation between the *An. gambiae* populations from Guinea, Burkina Faso and Cameroon, with F_{ST} reaching at most 0.001 and barely achieving statistical significance (Fig. 4.3a). This result was also supported by the doubleton analysis, with near-equal rates of doubleton sharing both within and between these populations (Fig. 4.3b). The lack of differentiation indicates high rates of gene flow, but this is surprising because the physical distances are considerable, with 2,170 km separating sampling sites in Cameroon and Burkina Faso, and 640 km separating sampling sites in Burkina Faso from Guinea. Previously, accepted wisdom was that anopheline mosquitoes disperse at most 5 km during their lifetime (Service, 1997), and therefore we would expect to see some geographical isolation by distance between these locations (Rousset, 1997). This is being challenged by recent observations of anopheline mosquitoes including *An. coluzzii* engaging in what appears to be purposeful wind-assisted long-distance migration (Huestis et al., 2019). However, the evidence is conflicting, as studies in the same geographical region indicate that *An. gambiae* is more likely to engage in migratory behaviour, whereas *An. coluzzii* appears to cope with highly seasonal availability of breeding habitat by aestivation (Dao et al., 2014). Unfortunately, Ag1000G phase 1 includes only limited geographical sampling of *An. coluzzii*, and so it is not possible to make a comparison between the two species at this stage. Subsequent project phases will expand sampling of both species and provide richer opportunities to investigate gene flow questions.

The Kenyan population displayed a high level of differentiation ($F_{ST} > 0.207$) with all other populations, with similar F_{ST} values against populations of both species. The Kenyan samples were expected to be genetically isolated to a certain degree, being sampled on the coast at the most easterly location among the collection sites, and being the only samples collected to the east of the Rift Valley, which presents a natural geographical barrier to gene flow. However, the magnitude of this difference was surprising, being more than twice that found between most other population pairs. This could indicate extremely low rates of gene flow, but could also be an effect of strong genetic drift within the Kenyan

population, which would be consistent with analyses of genetic diversity presented below. The fact that the Kenyan population was equally differentiated from both *An. gambiae* and *An. coluzzii* was also surprising, given that the Kenyan samples were believed to be *An. gambiae*, explored further below in the sub-section on gene flow between species.

Genetic diversity within populations

In the previous chapter I performed an initial analysis of genetic diversity, computing nucleotide diversity (π) within the Ag1000G phase 1 cohort as a whole. This analysis of course ignored the fact that different populations within a species may exhibit different levels of genetic diversity, due to differences in their demographic history such as contractions or expansions in population size. To investigate population differences in genetic diversity I performed four further analyses within each of nine populations defined by country and species (Fig. 4.4). The first of these analyses computed π, which summarises the fraction of nucleotide differences between pairs of chromosomes within a population (Fig. 4.4a). The second analyses computed Tajima's D, which summarises the distribution of allele frequencies within a population (Fig. 4.4b). The third analysis computed the full site frequency spectrum (SFS) for each population, which provides information about how many SNPs are observed at different allele frequencies (Fig. 4.4c). The final analysis examined the decay of linkage disequilibrium, which summarised the degree to which genotypes are correlated at pairs of SNPs at different physical distances from each other (Fig. 4.4d).

These analyses revealed strong contrasts between populations in both the magnitude and architecture of genetic diversity. All the populations to the north of the equatorial rainforest, from Guinea-Bissau in the West to Uganda in the East, including both species, displayed very similar results, with the highest genome-wide average $\pi = 1.5\%$, Tajima's $D < -1.5$, SFS with a strong excess of SNPs at low minor allele frequency, and LD decaying to background levels within 2 kb. These characteristics indicate large effective population size and are also consistent with a major population expansion at some point within their history. In both Gabon *An. gambiae* and Angola *An. coluzzii*, π was lower, Tajima's D was approaching zero, SFS were more balanced and LD decay was > 10 kb.

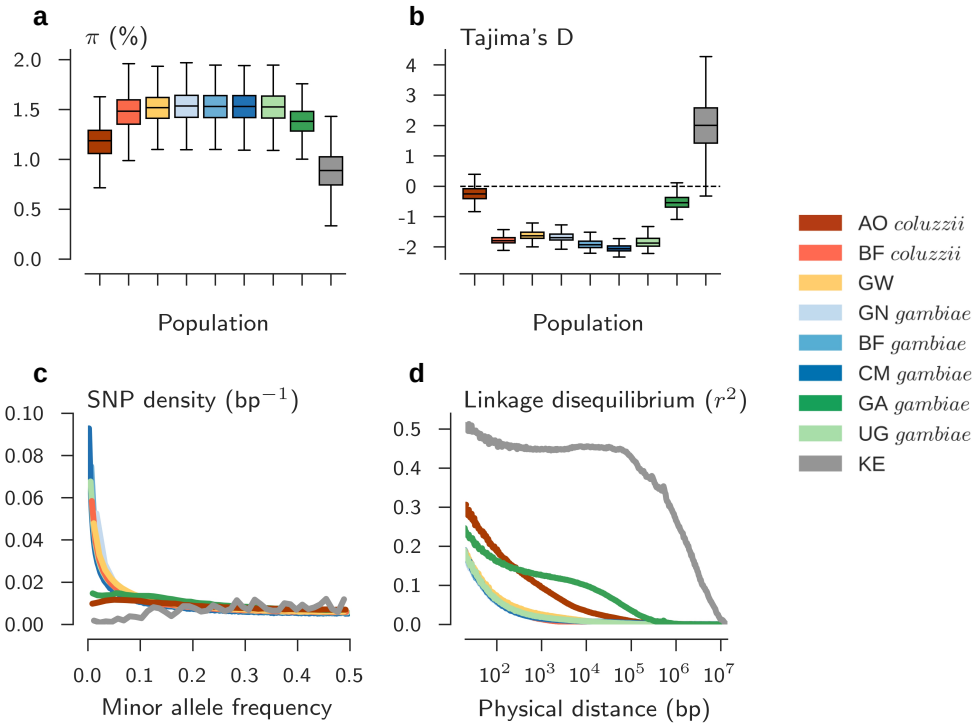

Figure 4.4. Genetic diversity within populations, computed using SNP genotypes from Chromosome 3. **a**, Nucleotide diversity, computed for each population in 20 kb windows. **b**, Tajima's D, computed in 20 kb windows. **c**, Folded site frequency spectra, normalised such that a population with stable population size is expected to show equal SNP density across all minor allele frequencies. **d**, Decay of linkage disequilibrium, averaged over pairs of randomly sampled SNPs at different physical distances.

These characteristics are consistent with an effective population size that is both smaller and has been more stable over time. The Kenyan population displayed the most extreme patterns of diversity, with the lowest π, Tajima's $D > 2$, SFS with a deficit of SNPs at lower minor allele frequencies, and LD not decaying to background until > 10 Mb. These characteristics indicate a strong and recent reduction in effective population size. Further examination of levels of heterozygosity within individuals revealed that Kenyan mosquitoes displayed long runs of homozygosity, in some cases affecting almost entire chromosome arms, a pattern not observed in mosquitoes from any other population (Fig. 4.6). Such patterns of homozygosity are usually only observed after multiple generations of inbreeding, similar to patterns found in domesticated animals (e.g. Purfield et al., 2012) and mosquitoes maintained in lab colonies for several years.

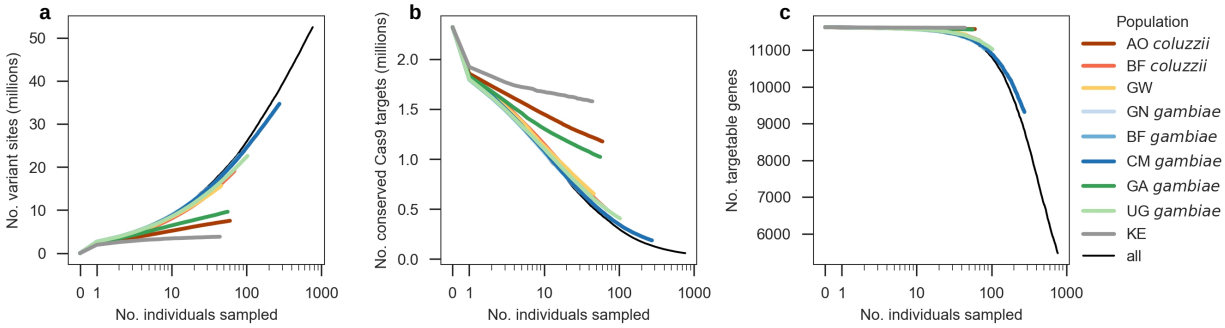

Figure 4.5. Genetic diversity and its impact on the availability of conserved Cas9 targets for the design of gene drives. **a**, Numbers of sites with nucleotide variation across the whole genome. **b**, Numbers of Cas9 targets with no nucleotide variation. **c**, Numbers of genes containing at least one conserved Cas9 target site.

Genetic variation within Cas9 gene drive targets

The high level of nucleotide diversity within many of the mosquito populations sampled in Ag1000G phase 1 is interesting in its own right, but also has practical consequences for the development of new vector control tools based on gene drives. Gene drives are a promising future technology for mosquito control, whereby an engineered genetic element is integrated into the mosquito genome that has the capability to be transmitted to progeny with super-Mendelian inheritance (Burt, 2003). Because of this biased inheritance, a gene drive will propagate itself within a mosquito population, even if it is in some way deleterious to its carrier. For example, gene drives have been engineered and proven effective under lab conditions which cause population suppression by biasing the sex ratio among offspring to be almost all male (Kyrou et al., 2018). Current gene drives make use of CRISPR/Cas9 homing endonucleases, which can be engineered to target almost any gene, because they include a guide RNA which matches a 21 bp sequence in the target genome. The guide RNA can be designed to match any sequence of 18 nucleotides, then must terminate with the protospacer adjacent motif (PAM). However, any natural genetic variation in the target sequence within the mosquito population could reduce efficacy of a gene drive, because it will affect binding of the Cas9 guide RNA.

I explored the impact of nucleotide variation on the availability of highly conserved Cas9 target sequences, in collaboration with Ag1000G Consortium members Mara Lawniczak and Krzysztof Kozak. I was particularly interested to explore the impact of sample size and

of population differences in levels of genetic diversity on the ascertainment of viable Cas9 gene drive target sites. I performed an analysis whereby I identified all 21 bp sequences in the reference genome terminating in a PAM motif, within which no nucleotide variation was found within a given Ag1000G phase 1 population. I repeated this multiple times with successively greater down-sampling of individuals from each population. I then plotted for each population the number of variant sites (Fig. 4.5a), the number of conserved Cas9 targets (Fig. 4.5b) and the number of genes containing at least one viable Cas9 target (Fig. 4.5c), and how these metrics varied with increasing sample size. These analyses showed that the higher levels of nucleotide diversity among the populations sampled to the north of the equatorial rainforest had a dramatically greater impact on the reduction of available Cas9 targets, compared with the more southerly populations from Gabon and Angola, and the Kenyan population. At a sample size of 50, more than 1.5 million Cas9 targets remained conserved in Kenya, and more than 1 million were conserved in Angola and Gabon, whereas less than 0.5 million were conserved in any of the remaining populations. Within the whole cohort, less than 3% of Cas9 targets were fully conserved at the nucleotide level. Clearly some *An. gambiae* and *An. coluzzii* populations will be more naturally resistant to gene drives than others due to higher diversity, and will require more intensive sampling in order to survey existing variation and design gene drives to accommodate this.

Gene flow between species

A long-standing question in the field is the extent to which gene flow occurs via hybridisation between species within the *An. gambiae* complex. I investigated genome-wide evidence for gene flow between species using data from Ag1000G phase 1, combined with data from *An. arabiensis* and outgroup species from Neafsey et al. (2014). I used the f_3 and f_4 statistics which test for evidence of admixture, as defined by Patterson et al. (2012). The f_3 statistic tests if allele frequencies in a population with suspected admixture are on average intermediate between two donor populations. The f_4 statistic, also known as the ABBA-BABA test or Patterson's D, tests for unbalanced allele sharing in a four

population tree. In this subsection to be concise when describing different f_4 tests I use population abbreviations of the form {country code}-{species}, e.g., "BF-gam" is Burkina Faso *An. gambiae*, "AO-col" is Angola *An. coluzzii*, "GW" is Guinea-Bissau, etc., and "O" denotes the outgroup species *An. christyi*.

Hybrids between *An. gambiae* and *An. coluzzii* are fully fertile and several studies have found evidence suggesting that hybridisation occurs at a rate that varies over both time and space (Lee et al., 2013; Weetman et al., 2011). In Ag1000G phase 1, the samples from Guinea-Bissau displayed a mixture of species genotypes according to conventional molecular diagnostics, including individuals with an apparent hybrid genotype. Similar results have been obtained from previous studies in the Far West region, and have been interpreted as evidence for a localised breakdown in reproductive isolation between *An. gambiae* and *An. coluzzii* (Oliveira et al., 2008; Marsden et al., 2011; Weetman et al., 2011; Gordicho et al., 2014; Vicente et al., 2017). Ag1000G Consortium member Giordano Botta investigated this further using data from Ag1000G phase 1, showing that Guinea-Bissau individuals had a mixture of species genotypes across ancestry-informative markers on all chromosome arms (The Anopheles gambiae 1000 Genomes Consortium, 2017). There are several possible explanations for these data, however, including contemporary or historical admixture between species, and retention of ancestral variation. The results from PCA suggest that the Guinea-Bissau individuals are not the result of ongoing hybridisation between species, because all individuals form a single group distinct from groups of each species from nearby countries (Fig. 4.2), rather than being spread out between them as would be expected under recent admixture. Consistent with this, results of f_3 tests with Guinea-Bissau as putatively admixed and Burkina Faso *An. gambiae* and *An. coluzzii* as donor populations were significantly positive, indicating allele frequencies in Guinea-Bissau were not intermediate and arguing against a simple admixture scenario. The Guinea-Bissau population was most closely related to the Burkina Faso *An. coluzzii* population in PCA, F_{ST} and doubleton sharing analyses (Figs. 4.2, 4.3). f_4 tests of the form D(GW, BF-col; GN-gam; O) were balanced and did not provide any significant evidence for admixture between Guinea-Bissau and Guinea *An. gambiae* on any chromosome, except for the

pericentromeric region of the X chromosome (Table 4.2; Fig. 4.7). Taken together, these results show that the origins of the Guinea-Bissau population and its relationships with *An. gambiae* and *An. coluzzii* populations are more complex than previously appreciated, and caution against over-interpretation of data from conventional species diagnostics.

A surprising result from the AIM analysis was that the Kenyan individuals also had a mixture of *An. gambiae*, *An. coluzzii* and heterozygous genotypes, similar to Guinea-Bissau (The Anopheles gambiae 1000 Genomes Consortium, 2017). The range of *An. coluzzii* is not known to extend to the east of the Rift Valley, and so the Kenyan population cannot be the result of any recent admixture between species. f_3 tests with Kenya as putatively admixed and all other pairs of populations as putative donors did not support evidence for admixture. I also computed f_4 statistics for the Kenyan population, using *An. gambiae* and *An. coluzzii* from Burkina Faso as potential sibling populations and *An. Christyi* as the outgroup. These analyses gave conflicting results depending on which chromosome arm was analysed, with 3R indicating that Kenya was sister with *An. coluzzii* (*D(KE, BF-col; BF-gam, O)*; $Z = 0.1$), 3L indicating Kenya was a cryptic taxon ancestral to both species (D(BF-gam, BF-col; KE, O); Z=-0.7) and the X chromosome not providing clear support for any one hypothesis. Thus, the origins and species status of the Kenyan population are also uncertain. These results reinforce the fact that reliance on a single diagnostic marker to identify species is very limiting and may be missing important signals of introgression or cryptic taxa.

Among the remaining populations, the AIMs supported a clear assignment of *An. gambiae* or *An. coluzzii* species status, in agreement with conventional molecular diagnostics (The Anopheles gambiae 1000 Genomes Consortium, 2017). However, this does not rule out low level gene flow between species, and among these populations there was clear evidence for a greater degree of gene flow in some populations relative to others. All f_4 tests with *An. coluzzii* from Angola and Burkina Faso as sister populations provided significant evidence for admixture between Angola and *An. gambiae*, with the strongest signal obtained for Gabon (e.g., D(AO-col, BF-col; GA-gam, O); $Z = 14.0$). Conversely, all tests with two *An. gambiae* populations as sisters provided evidence for admixture between Gabon *An.*

gambiae and *An. coluzzii*, with the strongest signal obtained for Angola (e.g., D(GA-gam, GN-gam; AO-col, O); $Z = 11.1$). These results show that gene flow between the two species has been stronger in southerly populations than northerly populations, and that it has been bidirectional.

Hybridisation between *An. gambiae* and *An. arabiensis* is also known to occur in nature but thought to be less common (White, 1971; Coluzzi et al., 1979; della Torre et al., 1997). I used f_4 tests of the form D(X, Y; A, O), where X and Y were all pairs of Ag1000G phase 1 populations, and A was *An. arabiensis* from Neafsey et al. (2014), to investigate whether gene flow with *An. arabiensis* is greater for some populations relative to others. These tests were significant in some cases (e.g., D(BF-gam, BF-col; A, O); $Z = 8.0$; D(UG-gam, BF-gam; A, O); $Z = 6.9$) but not others, allowing for a partial ordering of populations by relative degree of introgression . This partial ordering was GW $<$ {AO-col, BF-col} $<$ {GN-gam, BF-gam, CM-gam, GA-gam} $<$ UG-gam $<$ KE, indicating that introgression with *An. arabiensis* is higher in *An. gambiae* relative to *An. coluzzii*, and higher in easterly populations relative to westerly populations.

Conclusions

In this chapter I have showed that there are clear patterns of genetic structure among the populations sampled in the Ag1000G phase 1 cohort. These patterns of population structure are also associated with differences in the magnitude and architecture of genetic diversity within populations. Taken together, these results point to complex and varied demographic histories, and multiple biological and geographical factors affecting population size and rates of gene flow between populations and species. Within both species there is a strong north/south divide, with high differentiation between populations on either side of the equatorial rainforest. There is higher genetic diversity and evidence for a major population expansion in northerly but not southerly populations. Within *An. gambiae* there appears to be an extremely high rate of gene flow between the northerly populations, despite substantial physical distance, raising important questions about the rate and range of migration within this species. There are also unusual populations in both the far West

and far East, where individuals carry an apparent mixture of species ancestry, but this cannot be explained by a simple model of recent admixture between species. We also see an extreme contrast between populations on either side of the East African Rift, with evidence for a recent and substantial population bottleneck in the Kenyan population. Clearly there are limits to the extent to which these findings can be generalised, because geographical sampling is limited particularly within *An. coluzzii*, and much remains to be learned about both of these species throughout their complete geographical ranges. However, these results show that there are important heterogeneities within these species, which are of basic biological interest, but also relevant to malaria vector population surveillance, because they will affect factors such as the efficacy of gene drives for population suppression, or the speed, direction and extent to which insecticide resistance mutations spread. In the next chapter I investigate each of these populations in more detail, focusing on their evolutionary history and searching for genome regions evolving rapidly due to positive selection for insecticide resistance.

Methods

In order to facilitate the population genetic analyses described in this chapter on the relatively large Ag1000G SNP variation data, I developed a new software package named scikit-allel[1] for the Python programming language. scikit-allel aims to support a range of standard data manipulations and statistical functions that are useful for exploratory analysis of large-scale genome variation data, with implementations that are sufficiently performant and scalable such that those analyses can be run interactively, allowing rapid iteration and hypothesis generation. To achieve this, scikit-allel builds on several high-performance general-purpose scientific computing libraries, including NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), Cython (Behnel et al., 2011), Pandas (McKinney, 2010) and Matplotlib (Hunter, 2007). The sub-sections below provide further details of specific analyses. Unless otherwise stated, analyses were performed using scikit-allel and custom Python notebooks.

[1] https://github.com/cggh/scikit-allel

Genetic distance analyses

To investigate variations in patterns of relatedness between individuals across the genome, I divided the genome into 1,418 contiguous non-overlapping windows, where each window contained 100 kb of accessible positions (equally-accessible windows). Within each window I computed the city-block distance between all pairs of individuals, producing a 765×765 distance matrix. I used these distance matrices to compute and plot a neighbour-joining tree for each window via the ape package version 3.5 for R (Popescu et al., 2012). To compare these distance matrices systematically, I computed the Pearson correlation coefficient between all pairs of distance matrices, producing a 1418×1418 correlation matrix. I converted this to a distance matrix by computing $\sqrt{(1 - r^2)}$ where r is the correlation coefficient, and performed a singular value decomposition (SVD) on this matrix. I then plotted the first ten components from this transformation against position in the genome. To provide a visual representation of this transformation in Fig. 4.1b I painted the genome as follows. Within the 2La inversion I painted with one colour, using the value from the first SVD component rescaled between 0 and 1 as the alpha value. Within the 2Rb inversion I painted with a second colour, using the value from the third SVD component rescaled between 0 and 1 as the alpha value. Within the remainder of the genome I painted with either a third or a fourth colour, depending on whether the value of the second SVD component was positive or negative. Positive and negative values were separately rescaled between 0 and 1 to provide alpha values.

Principal components analysis

SNPs for inclusion in principal components analysis were chosen by selecting biallelic variants from within the regions 3R:1-37 Mbp and 3L:15-41 Mbp. Only variants with minor allele frequency >= 1% were retained and each chromosome arm was randomly down-sampled to 100,000 variants. I then pruned to remove SNPs in linkage disequilibrium, excluding SNPs above an r^2 threshold of 0.01 in moving windows of 500 SNPs with a step size of 250 SNPs. SNPs from both chromosome arms were then concatenated, and PCA was run following methods described in Patterson et al. (2006).

Population differentiation analyses

Average F_{ST} was computed between all pairs of 9 populations defined by country of origin and species, except for Guinea-Bissau and Kenya which were each treated as a single population with uncertain species status. Hudson's F_{ST} estimator was used, and the ratio of averages computed following Bhatia et al. (2013). Only SNPs within the regions 3R:1-37 Mbp and 3L:15-41 Mbp and which were segregating in both populations were used. Standard error for each average was computed using a block-jackknife with block size 10,000 SNPs.

The 765 individuals from natural populations were grouped into 9 populations as described above for F_{ST} analysis. Each population was then randomly down-sampled to the size of the smallest population (Guinea *An. gambiae*, N=31). After down-sampling, I ascertained SNPs from Chromosome 3 with an alternate allele count of 2 (doubletons). For each population, I identified the set of doubletons with at least one allele originating from an individual in that population. I then computed the fraction of those doubletons shared with each other population including itself.

Genetic diversity analyses

To compute a genome-wide average value for nucleotide diversity (π) I divided the genome into non-overlapping contiguous windows where each window contained 20 kb accessible positions. For each SNP (including multiallelic SNPs) I used allele counts to compute the mean number of allelic differences between all pairs of individuals within each population. A value of π was computed for each window for each population by summing the mean pair-wise differences over all SNPs in the window and dividing by the number of accessible positions.

From SNP validation experiments I estimated an FDR of 0.4% and a sensitivity of 94.6% (Chapter 3). To explore the impact of error rates on estimates of diversity, I computed conservative lower bounds on π and Watterson's θ for each population under an assumption of 1% FDR and 100% sensitivity by randomly sampling 99% of SNPs in the dataset without replacement. I computed conservative upper bounds under an assumption of 0% FDR and

94% sensitivity by randomly sampling 106% of SNPs from the dataset with replacement. Values for chromosome arm 3R excluding pericentromeric regions are given in Table 4.1. I report unadjusted values in the results section above.

For each population, a site frequency spectrum (SFS) was computed using allele counts in SNPs from Chromosome 3 excluding pericentromeric regions. Folded SFS were multiplied by the scaling factor $k(n-k)/n$ where k is the minor allele count and n is the number of chromosomes, to facilitate comparison with theoretical SFS for a population with constant size (expected to have constant scaled frequency for all values of k).

SNPs from Chromosome 3 excluding pericentromeric regions were used for LD decay analyses. For each population, the genotype correlation coefficient r^2 (Rogers and Huff, 2009) was computed for randomly sampled pairs of SNPs at a range of physical distances. Only biallelic SNPs with a minor allele frequency greater than 10% within the population were used. Values were corrected for sample size by subtracting $1/n$ where n is the number of sampled chromosomes.

Admixture tests

The f_3 and f_4 tests were performed using a block size of 100,000 SNPs to estimate standard error via a block jackknife procedure (Patterson et al., 2012). Z scores reported in the results are computed by dividing the test statistic by the estimated standard error, and thus indicate the number of standard deviations from zero.

Supplemental figures

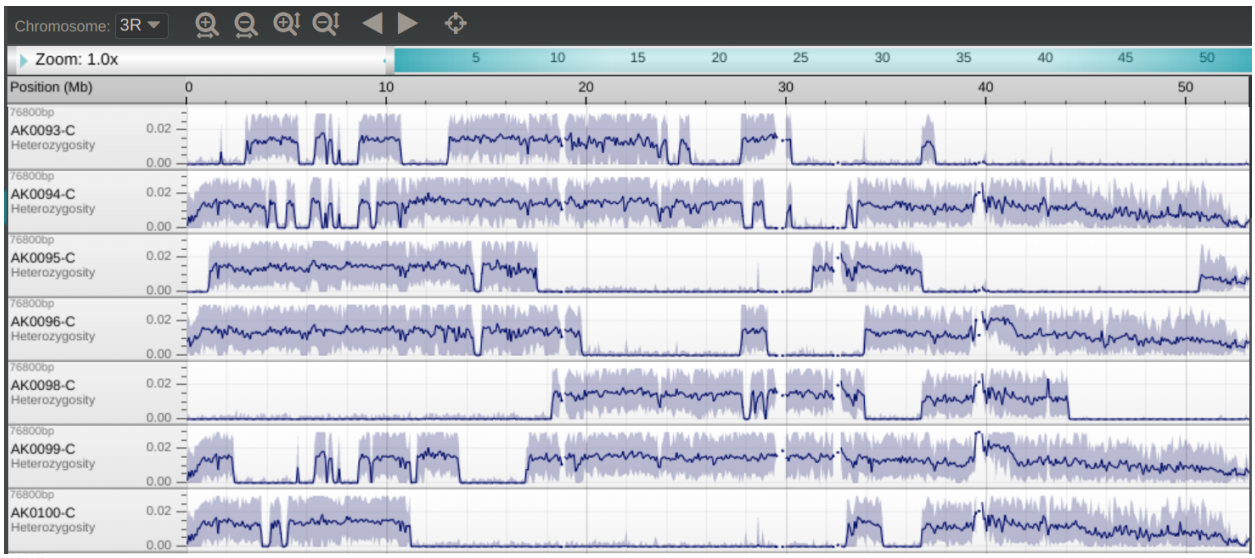

Figure 4.6. Screenshot from the Panoptes Web application showing runs of homozygosity on chromosome arm 3R in a selection of seven individuals from Kenya.

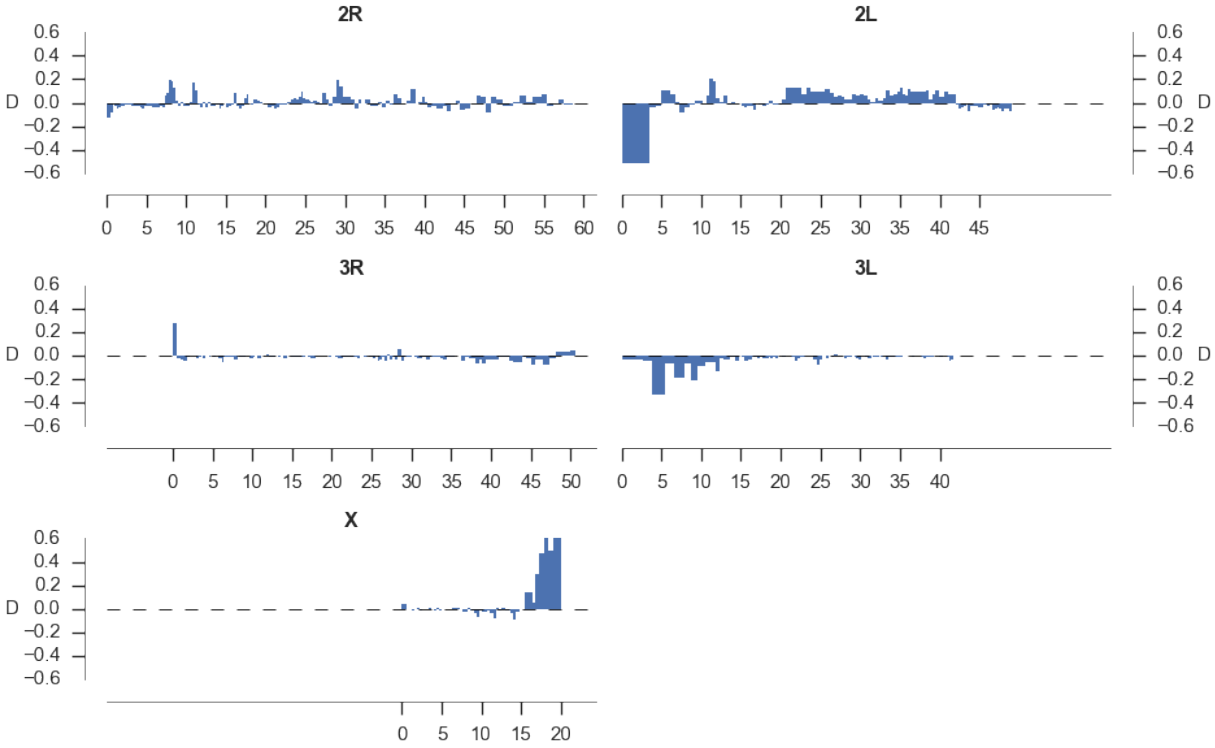

Figure 4.7. Results of the f_4 test D(GW, BF-col; GN-gam, O) for admixture between Guinea-Bissau (GW) and Guinea *An. gambiae* (GN-gam). Values for each block of 100,000 SNPs are plotted over the genome, to assess whether any particular genome regions provide evidence for admixture. The only region providing strong support for admixture is the pericentromeric region of the X chromosome.

Supplemental tables

Table 4.1. Genome-wide average values for π and θ_W with upper and lower bounds computed as described in Methods.

Population	π (%)	θ_W (%)
AO coluzzii	1.16 (1.14-1.23)	1.24 (1.23-1.32)
BF coluzzii	1.45 (1.44-1.54)	3.11 (3.08-3.31)
GW	1.49 (1.48-1.59)	2.80 (2.78-2.98)
GN gambiae	1.50 (1.49-1.60)	2.82 (2.79-3.00)
BF gambiae	1.50 (1.48-1.59)	3.58 (3.54-3.80)
CM gambiae	1.50 (1.49-1.59)	4.60 (4.56-4.90)
GA gambiae	1.35 (1.34-1.44)	1.61 (1.60-1.71)
UG gambiae	1.50 (1.48-1.59)	3.51 (3.48-3.74)
KE	0.87 (0.86-0.93)	0.55 (0.55-0.59)

Table 4.2. Results of the f_4 test D(GW, BF-col; GN-gam, O) for admixture between Guinea-Bissau (GW) and Guinea *An. gambiae* (GN-gam).

Chromosome	*D*	*SE*	*Z*
2R	0.007	0.0048	1.4
2L	0.021	0.0093	2.3
3R	-0.011	0.0039	-2.7
3L	-0.029	0.0067	-4.4
X	0.078	0.0361	2.2

References

Aboagye-Antwi, F et al. (2015). 'Experimental Swap of Anopheles gambiae's Assortative Mating Preferences Demonstrates Key Role of X-Chromosome Divergence Island in Incipient Sympatric Speciation.' In: *PLoS Genet.* 11.4. Ed. by BA Payseur, e1005141. DOI: 10.1371/journal.pgen.1005141.

Behnel, S, R Bradshaw, C Citro, L Dalcin, DS Seljebotn and K Smith (2011). 'Cython: The Best of Both Worlds'. In: *Computing in Science & Engineering* 13.2, pp. 31–39. DOI: 10.1109/mcse.2010.118.

Bhatia, G, N Patterson, S Sankararaman and AL Price (2013). 'Estimating and interpreting FST: the impact of rare variants.' In: *Genome Res.* 23.9, pp. 1514–1521. DOI: 10.1101/gr.154831.113.

Binswanger-Mkhize, HP and S Savastano (2017). 'Agricultural intensification: The status in six African countries'. In: *Food Policy* 67, pp. 26–40. DOI: 10.1016/j.foodpol.2016.09.021.

Burt, A (2003). 'Site-specific selfish genes as tools for the control and genetic engineering of natural populations'. In: *Proc. R. Soc. B Biol. Sci.* 270.1518, pp. 921–928. DOI: 10.1098/rspb.2002.2319.

Clarkson, CS et al. (2014). 'Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation.' In: *Nat. Commun.* 5.1, p. 4248. DOI: 10.1038/ncomms5248.

Coluzzi, M, A Sabatini, V Petrarca and MA Di Deco (1979). 'Chromosomal differentiation and adaptation to human environments in the anopheles gambiae complex'. In: *Trans. R. Soc. Trop. Med. Hyg.* 73.5, pp. 483–497. DOI: 10.1016/0035-9203(79)90036-1.

Coluzzi, M, A Sabatini, A della Torre, MA di Deco and V Petrarca (2002). 'A polytene chromosome analysis of the Anopheles gambiae species complex'. In: *Science* 298.5597, pp. 1415–1418. DOI: 10.1126/science.1077769.

Cruickshank, TE and MW Hahn (2014). 'Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow'. In: *Mol. Ecol.* 23.13, pp. 3133–3157. DOI: 10.1111/mec.12796.

Dao, A, AS Yaro, M Diallo, S Timbiné, DL Huestis, Y Kassogué, AI Traoré, ZL Sanogo, D Samaké and T Lehmann (2014). 'Signatures of aestivation and migration in Sahelian malaria mosquito populations.' In: *Nature* 516.7531, pp. 387–390. DOI: 10.1038/nature13987.

della Torre, A, C Fanello, M Akogbeto, J Dossou-yovo, G Favia, V Petrarca and M Coluzzi (2001). 'Molecular evidence of incipient speciation within Anopheles gambiae s.s. in West africa'. In: *Insect Mol. Biol.* 10.1, pp. 9–18. DOI: 10.1046/j.1365-2583.2001.00235.x.

della Torre, A, L Merzagora, JR Powell and M Coluzzi (1997). 'Selective introgression of paracentric inversions between two sibling species of the Anopheles gambiae complex'. In: *Genetics* 146.1, pp. 239–244.

Fontaine, MC et al. (2014). 'Extensive introgression in a malaria vector species complex revealed by phylogenomics.' In: *Science* 347.6217, p. 1258524. DOI: 10.1126/science.1258524.

Gordicho, V et al. (2014). 'First report of an exophilic Anopheles arabiensis population in Bissau City, Guinea-Bissau: recent introduction or sampling bias?' In: *Malar. J.* 13.1, p. 423. DOI: 10.1186/1475-2875-13-423.

Harris, CR et al. (2020). 'Array programming with NumPy'. In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2. eprint: 2006.10256.

Holsinger, KE and BS Weir (2009). 'Genetics in geographically structured populations: Defining, estimating and interpreting FST'. In: *Nat. Rev. Genet.* 10.9, pp. 639–650. DOI: 10.1038/nrg2611.

Huestis, DL et al. (2019). 'Windborne long-distance migration of malaria mosquitoes in the Sahel'. In: *Nature* 574.7778, pp. 404–408. DOI: 10.1038/s41586-019-1622-4.

Hunter, JD (2007). 'Matplotlib: A 2D graphics environment'. In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Kyrou, K, AM Hammond, R Galizi, N Kranjc, A Burt, AK Beaghton, T Nolan and A Crisanti (2018). 'A CRISPR–Cas9 gene drive targeting doublesex causes complete population suppression in caged Anopheles gambiae mosquitoes'. In: *Nat. Biotechnol.* 36.11, pp. 1062–1066. DOI: 10.1038/nbt.4245.

Lee, Y, CD Marsden, LC Norris, TC Collier, BJ Main, A Fofana, AJ Cornel and GC Lanzaro (2013). 'Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, Anopheles gambiae.' In: *Proc. Natl. Acad. Sci. U. S. A.* 110.49, pp. 19854–19859. DOI: 10.1073/pnas.1316851110.

Marsden, CD, Y Lee, CC Nieman, MR Sanford, J Dinis, C Martins, A Rodrigues, AJ Cornel and GC Lanzaro (2011). 'Asymmetric introgression between the M and S forms of the malaria vector, Anopheles gambiae, maintains divergence despite extensive hybridization'. In: *Mol. Ecol.* 20.23, pp. 4983–4994. DOI: 10.1111/j.1365-294X.2011.05339.x.

McKinney, W (2010). 'Data Structures for Statistical Computing in Python'. In: *Proceedings of the 9th Python in Science Conference*. Ed. by S van der Walt and J Millman. SciPy, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

McVean, G (2009). 'A genealogical interpretation of principal components analysis.' In: *PLoS Genet.* 5.10. Ed. by M Przeworski, e1000686. DOI: 10.1371/journal.pgen.1000686.

Neafsey, DE et al. (2014). 'Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes'. In: *Science* 347.6217, p. 1258522. DOI: 10.1126/science.1258522.

Norris, LC, BJ Main, Y Lee, TC Collier, A Fofana, AJ Cornel and GC Lanzaro (2015). 'Adaptive introgression in an African malaria mosquito coincident with the increased

usage of insecticide-treated bed nets'. In: *Proc. Natl. Acad. Sci.* 112.3, pp. 815–820. DOI: 10.1073/pnas.1418892112.

Oliveira, E, P Salgueiro, K Palsson, JL Vicente, AP Arez, TG Jaenson, A Caccone and J Pinto (2008). 'High Levels of Hybridization between Molecular Forms of Anopheles gambiae from Guinea Bissau'. In: *J. Med. Entomol.* 45.6, pp. 1057–1063. DOI: 10.1093/jmedent/45.6.1057.

Patterson, N, P Moorjani, Y Luo, S Mallick, N Rohland, Y Zhan, T Genschoreck, T Webster and D Reich (2012). 'Ancient admixture in human history.' In: *Genetics* 192.3, pp. 1065–1093. DOI: 10.1534/genetics.112.145037.

Patterson, N, AL Price and D Reich (2006). 'Population structure and eigenanalysis'. In: *PLoS Genet.* 2.12, e190. DOI: 10.1371/journal.pgen.0020190.

Popescu, AA, KT Huber and E Paradis (2012). 'ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R'. In: *Bioinformatics* 28.11, pp. 1536–1537. DOI: 10.1093/bioinformatics/bts184.

Purfield, DC, DP Berry, S McParland and DG Bradley (2012). 'Runs of homozygosity and population history in cattle.' In: *BMC Genet.* 13.1, p. 70. DOI: 10.1186/1471-2156-13-70.

Rogers, AR and C Huff (2009). 'Linkage disequilibrium between loci with unknown phase'. In: *Genetics* 182.3, pp. 839–844. DOI: 10.1534/genetics.108.093153.

Rousset, F (1997). 'Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance'. In: *Genetics* 145.4, pp. 1219–1228.

Service, MW (1997). 'Mosquito (Diptera: Culicidae) Dispersal - The Long and Short of It'. In: *J. Med. Entomol.* 34.6, pp. 579–588. DOI: 10.1093/jmedent/34.6.579.

Stump, AD, M Pombi, L Goeddel, JMC Ribeiro, JA Wilder, A della Torre and NJ Besansky (2007). 'Genetic exchange in 2La inversion heterokaryotypes of Anopheles gambiae.' In: *Insect Mol. Biol.* 16.6, pp. 703–9. DOI: 10.1111/j.1365-2583.2007.00764.x.

Tene Fossog, B et al. (2015). 'Habitat segregation and ecological character displacement in cryptic African malaria mosquitoes'. In: *Evol. Appl.* 8.4, pp. 326–345. DOI: 10.1111/eva.12242.

The 1000 Genomes Project Consortium (2012). 'An integrated map of genetic variation from 1,092 human genomes'. In: *Nature* 491.7422, pp. 56–65. DOI: 10.1038/nature11632.

The Anopheles gambiae 1000 Genomes Consortium (2017). 'Genetic diversity of the African malaria vector Anopheles gambiae'. In: *Nature* 552.7683, pp. 96–100. DOI: 10.1038/nature24995.

Turner, TL, MW Hahn and SV Nuzhdin (2005). 'Genomic islands of speciation in Anopheles gambiae'. In: *PLoS Biol.* 3.9. Ed. by N Barton, e285. DOI: 10.1371/journal.pbio.0030285.

Vicente, J et al. (2017). 'Massive introgression drives species radiation at the range limit of Anopheles gambiae'. In: *Sci. Rep.* 7.1. DOI: 10.1038/srep46451.

Virtanen, P et al. (2020). 'SciPy 1.0: fundamental algorithms for scientific computing in Python'. In: *Nat. Methods* 17.3, pp. 261–272. DOI: 10.1038/s41592-019-0686-2. eprint: 1907.10121.

Weetman, D, CS Wilding, K Steen, J Pinto and MJ Donnelly (2011). 'Gene flow-dependent genomic divergence between Anopheles gambiae M and S forms.' In: *Mol. Biol. Evol.* 29.1, pp. 279–291. DOI: 10.1093/molbev/msr199.

White, BJ, C Cheng, F Simard, C Costantini and NJ Besansky (2010). 'Genetic association of physically unlinked islands of genomic divergence in incipient species of Anopheles gambiae'. In: *Mol. Ecol.* 19.5, pp. 925–939. DOI: 10.1111/j.1365-294X.2010.04531.x.

White, GB (1971). 'Chromosomal evidence for natural interspecific hybridization by mosquitoes of the anopheles gambiae complex'. In: *Nature* 231.5299, pp. 184–185. DOI: 10.1038/231184a0.

WHO (2019). *World malaria report 2019*. Tech. rep. World Health Organization.

Wiebe, A et al. (2017). 'Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance'. In: *Malar. J.* 16.1. DOI: 10.1186/s12936-017-1734-y.

Wolf, JBW and H Ellegren (2017). 'Making sense of genomic islands of differentiation in light of speciation.' In: *Nature reviews. Genetics* 18 (2), pp. 87–100. DOI: 10.1038/nrg.2016.133.

5 Recent positive selection

In this chapter I continue analysing the Ag1000G phase 1 data resource by performing genome-wide scans for signals of recent positive selection. I apply three selection scan methods to each of the mosquito populations in the Ag1000G phase 1 cohort, and develop a systematic approach to identifying and mapping selection signals within these scans. I use this approach to build a catalog of 289 putative selection signals, including signals at five loci containing genes that have a functionally-validated role in insecticide resistance in An. gambiae and/or An. coluzzii. I describe a web application that provides a means for exploring these selection signals, and illustrate its use by discovering selection signals at a diacylglycerol kinase gene with a potentially novel role in resistance to organophosphate insecticides. Selection for insecticide resistance has clearly been the major driver of recent selection in these species, and the capability for further adaptation remains a significant threat to the control of malaria vectors in sub-Saharan Africa. The data and methods described in this chapter provide a means to discover which genes are driving this adaptive response.

Introduction

As described in Chapter 1, malaria vector control programmes have massively expanded since the turn of the millennium, with a heavy reliance on insecticide-based interventions (Cibulskis et al., 2016; Bhatt et al., 2015; WHO, 2019). Human population distributions and patterns of land use have also changed dramatically, including rapid urbanisation (Awumbila, 2017; OECD and Sahel and West Africa Club, 2020) and the expansion and intensification of agriculture, where insecticides are also used extensively (Otsuka

and Place, 2014; Binswanger-Mkhize and Savastano, 2017; Sternberg and Thomas, 2018). Thus, the environment in which natural malaria vector populations seek food and suitable habitats throughout their multi-stage life cycle has radically altered in recent decades, generating new and complex selection pressures. At the heart of this transformation is the massive scale-up of long-lasting insecticidal bednet (LLIN) distribution (WHO, 2005; RBM, 2008; WHO, 2017a; Bhatt et al., 2015; Okumu, 2020). The proportion of the population at risk sleeping under an LLIN has increased from less than 2% to more than 50% of the population by 2015 (Cibulskis et al., 2016; Bhatt et al., 2015), although coverage varies substantially between countries (WHO, 2019). All LLINs are treated with a pyrethroid insecticide which serves both to repel mosquitoes and to kill mosquitoes that come into physical contact with the net (WHO, 2020; Okumu, 2020). LLINs target malaria vector species that are highly anthropophilic, in particular *An. gambiae* and *An. coluzzii*. Unsurprisingly, pyrethroid resistance has become widespread in these species and increased in intensity over the time period of LLIN scale-up (Hemingway et al., 2016; Hancock et al., 2020).

Several molecular mechanisms of pyrethroid resistance are known in malaria vectors (Hemingway et al., 2016). Yet there remain substantial gaps in our knowledge of the genetic changes that have occurred in natural *An. gambiae* and *An. coluzzii* populations in response to pyrethroid selection pressure. For example, adaptations affecting cytochrome P450 (CYP) enzymes are known to have occurred in some mosquito populations, inducing pyrethroid resistance by increasing the metabolism of insecticide molecules within the mosquito (Ranson and Lissenden, 2016; Hemingway et al., 2016). This form of metabolic resistance is perceived as a significant threat to the efficacy of LLINs (Churcher et al., 2016; WHO, 2017b). There are currently 108 CYP genes annotated in the *An. gambiae* reference genome (Giraldo-Calderón et al., 2014; VectorBase, 2019), of which several have been shown to have the potential to metabolise pyrethroids, but it is still not clear which CYP genes are the primary drivers of adaptation to pyrethroids in natural mosquito populations (Mohammed et al., 2017). It is also not clear which populations currently carry CYP-mediated pyrethroid resistance adaptations, and whether the same CYP genes are

involved across multiple populations or not. Furthermore, other classes of enzyme such as glutathione S-transferases also have the potential to metabolise insecticides, but their role in the evolution of pyrethroid resistance is unclear (Adolfi et al., 2019). Additionally, entirely new molecular mechanisms of pyrethroid resistance have recently been discovered (Ingham et al., 2019), opening up the possibility of a much larger adaptive landscape than previously appreciated.

Malaria vectors may also encounter a range of other insecticides during their lifetime, either because of malaria vector control interventions or agricultural use. Indoor residual spraying of insecticides (IRS) coverage has not reached the same level as LLINs, but has nevertheless had a measurable impact on malaria transmission (Bhatt et al., 2015), with 10.1% of the population at risk protected by IRS in 2010, although this has fallen back to 4.5% in 2019 (WHO, 2019). There is even greater spatial heterogeneity in IRS coverage than for LLINs, with IRS generally reserved for use in high transmission regions (WHO, 2019). Part of the reduction in IRS coverage in recent years can be attributed to the introduction of more expensive insecticides. In 2010, most IRS programmes used pyrethroids, but by 2018 many had switched to use organophosphates because of pyrethroid resistance, although pyrethroids, carbamates and organochlorines all remain in use (WHO, 2019). All insecticides used in public health either have been or continue to be used in agriculture, and a major open question remains whether agricultural pesticide use is driving selection for insecticide resistance in malaria vectors (Georghiou, 1990; Nkya et al., 2013; Philbert et al., 2014; Reid and McKenzie, 2016). As IRS programmes continue to switch away from pyrethroids to use other insecticides, it is important to understand which resistance adaptations exist in malaria vector populations, either because of prior use in agriculture or because of prior public health use of insecticides with a similar mode of action (Fouet et al., 2020).

Given the variety and heterogeneity of these new selection pressures, evolution is likely to be occurring at multiple loci throughout the genomes of malaria vector species, many of which may be unknown. The availability of data from the Ag1000G project on genetic variation in natural malaria vectors provides a unique opportunity to study the full genomic

landscape of recent selection, to discover new adaptations to insecticide resistance, and to compare the genomic profile of adaptation between different species and populations. A number of statistical methods have been developed for performing genome-wide selection scans using high quality whole genome variation data from individuals sampled from natural populations (Oleksyk et al., 2010; Haasl and Payseur, 2015; Vatsiou et al., 2016; Pavlidis and Alachiotis, 2017; Booker et al., 2017). These methods work in different ways, but all leverage the fact that recent positive selection leaves a characteristic signature at affected loci, which can be detected against a genomic background where the majority of genes have not experienced recent positive selection. For example, H12 (Garud et al., 2015) detects a localised decrease in haplotype diversity, IHS (Voight et al., 2006) and XPEHH (Sabeti et al., 2007) detect a localised increase in haplotype sharing, either within or between populations respectively, and PBS (Yi et al., 2010; Crawford et al., 2017) detects a localised increase in genetic differentiation between populations. These methods are not perfect, having varying power to detect different types of selection under different population demographic scenarios (Haasl and Payseur, 2015; Vatsiou et al., 2016; Pavlidis and Alachiotis, 2017; Booker et al., 2017). They may also correctly detect a signal of selection but fail to provide enough precision to narrow down the target of selection to a single gene. Nevertheless, genome-wide selection scans can provide valuable information about genomic regions under selection, within which candidate genes can be identified and prioritised for further study.

In this chapter I use data from Ag1000G phase 1 to explore the genomic landscape of recent positive selection in *An. gambiae* and *An. coluzzii* populations from multiple countries. I integrate results from multiple genome-wide selection scans in mosquito populations representing different species and geographical locations. I also describe an online resource where all selection signals can be searched and browsed, and investigate the strongest selection signals to identify candidate genes driving novel forms of insecticide resistance. The analyses described in this chapter were performed in the context of a broader collaboration with the Ag1000G Analysis Working group, and particularly with my colleague Nick Harding from the MalariaGEN Resource Centre team. Here I focus on those

analyses that I devised and performed, but include some results from analyses performed jointly for additional context, and indicate joint contributions within the relevant sections.

Results

Genome-wide selection scans

Genome-wide selection scans were performed using nucleotide variation data from Ag1000G phase 1, which comprises genotypes in 765 individuals at 41,476,870 biallelic SNPs, phased into haplotypes as described in Chapter 3. Three selection scan methods were chosen because of their power to detect recent positive selection: H12 (Garud et al., 2015), IHS (Voight et al., 2006) and XPEHH (Sabeti et al., 2007). The Ag1000G phase 1 resource includes data on nine mosquito populations, with two *An. coluzzii* populations, five *An. gambiae* populations, and two further populations of uncertain species status, described in Chapter 4. However, the Kenyan population exhibited extremely low levels of genetic diversity across the whole genome when compared with other populations, and in exploratory analyses it was evident that this low diversity was associated with increased noise in genome-wide selection scans. The Kenyan population was therefore excluded from further selection analyses. Population structure analyses also revealed evidence for structure among the Cameroon *An. gambiae* mosquitoes, associated with the different collection sites. Only the Cameroon mosquitoes from the savannah collection sites were therefore included. Thus, eight populations were analysed, with sample size ranging from N=31 (Guinea *An. gambiae*) to N=103 (Uganda *An. gambiae*).

H12 and IHS genome-wide selection scans were computed for each population, and XPEHH scans were computed for selected pairs of populations. The XPEHH method is designed to identify genome locations where selection is acting in some populations but not others, and therefore combinations of populations were chosen for these scans to allow for comparisons between species and between geographically distant locations. All together, this comprised a total of 40 genome-wide selection scans. To facilitate the rapid computation of these selection scans on the relatively large Ag1000G data resource,

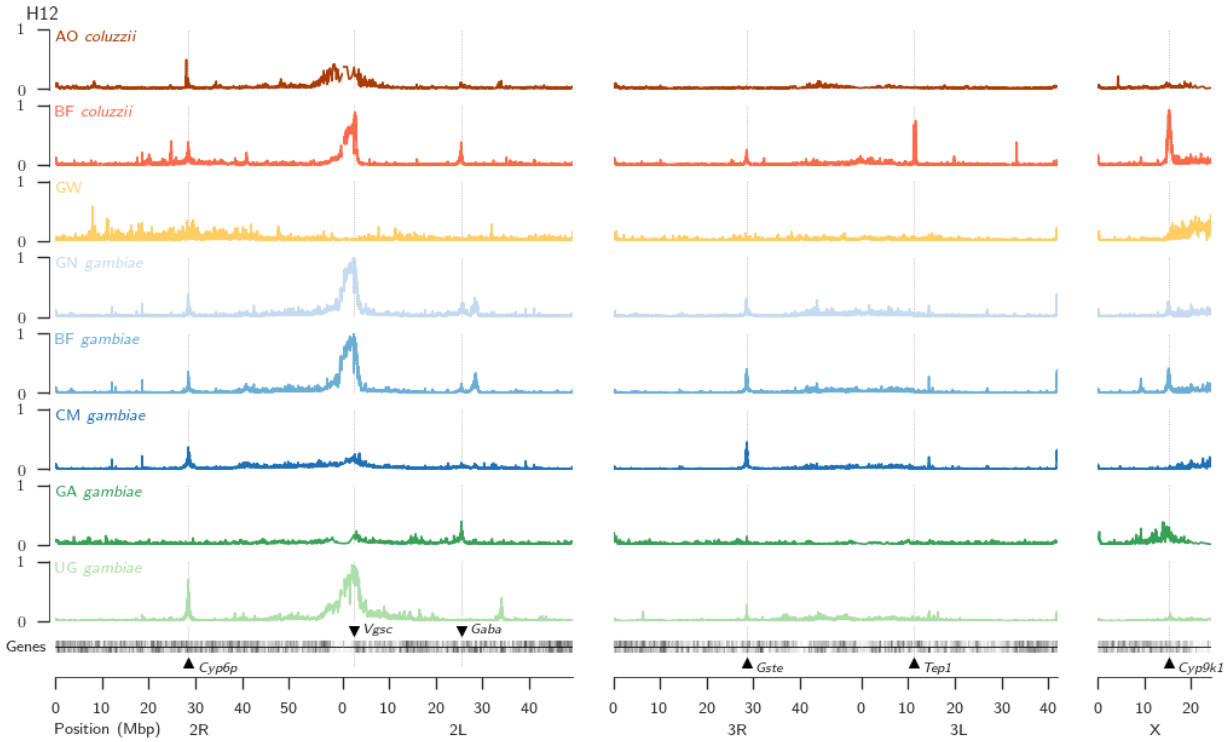

Figure 5.1. H12 selection scans. Each track shows results from H12 selection scan in a single population. AO = Angola; BF = Burkina Faso; GW = Guinea-Bissau; GN = Guinea; CM = Cameroon; GA = Gabon; UG = Uganda. H12 values range between 0 and 1, where a value of zero indicates high haplotype diversity within a genomic window, and a value of one indicates low haplotype diversity (at most two distinct haplotypes observed within the sampled individuals). Validated insecticide resistance genes (*Vgsc*, *Gaba*, *Cyp9k1*) or gene clusters (*Cyp6p*, *Gste*) are labelled at the bottom. *Tep1* is an immune system gene previously found to be under selection in *An. coluzzii* (White et al., 2010).

I reimplemented the H12, IHS and XPEHH methods in the scikit-allel software package[1], making use of general purpose high-performance scientific computing libraries for the Python programming language. I calibrated and ran the H12 scans, and the IHS and XPEHH scans were run by Nick Harding. Each of these scans produced a test statistic for each segregating SNP or for each of a set of genomic windows, where higher absolute values of the statistic indicate stronger evidence for recent positive selection.

Selection signals at known insecticide resistance loci

To provide an initial view of these data, I plotted the results of the H12 scans over the genome (Fig. 5.1). There were a number of clear peaks within these scans that were replicated across multiple populations, including at the following five loci containing genes

[1] https://github.com/cggh/scikit-allel

that have been functionally validated as playing a role in insecticide resistance in *An. gambiae* and/or *An. coluzzii*:

- **2L:2.4 Mb** - This locus contains the voltage-gated sodium channel gene (*Vgsc*; `AGAP004707`) which encodes a nervous system protein that is the binding target of DDT and pyrethroid insecticides (Dong et al., 2014). Amino acid substitutions in this gene confer resistance to DDT and pyrethroids in *An. gambiae* (Martinez-Torres et al., 1998; Ranson et al., 2000b; Davies et al., 2007; Lynd et al., 2010; Jones et al., 2012; Wang et al., 2015).

- **3R:28.6 Mb** - This locus contains a cluster of eight glutathione S-transferase genes, which encode enzymes involved in detoxification of xenobiotic substances. In *An. gambiae* this locus was initially discovered as a major locus of DDT resistance (Prapanthadara et al., 1993; Ranson et al., 2000a; Ranson et al., 2001; Ding et al., 2003). An amino acid substitution in one of the genes in this cluster, *Gste2* (`AGAP009194`) was subsequently shown to confer elevated DDT resistance (Mitchell et al., 2014). Increased expression of *Gste2* has also been shown experimentally to confer resistance to DDT and organophosphates (Adolfi et al., 2019).

- **2R:28.5 Mb** - This locus contains a cluster of ten genes, nine of which encode cytochrome P450 enzymes. This includes *Cyp6p3* (`AGAP002865`) which is associated with pyrethroid resistance in *An. gambiae* and is capable of metabolising multiple pyrethroid insecticides (Müller et al., 2008). Increased expression of *Cyp6p3* has also been shown experimentally to confer resistance to both pyrethroids and carbamates in *An. gambiae* (Adolfi et al., 2019).

- **X:15.2 Mb** - This locus contains the cytochrome P450 gene *Cyp9k1* (`AGAP000818`). A selection signal has previously been found at this locus in *An. coluzzii* in Mali (Main et al., 2015). Evolution of *Cyp9k1* has also been observed in *An. coluzzii* on Bioko Island in response to combined use of pyrethroids in IRS and LLIN programmes (Vontas et al., 2018). *Cyp9k1* metabolises the pyrethroid deltamethrin, but also pyriproxyfen, a non-pyrethroid insecticide (Vontas et al., 2018).

- **2L:25.4 Mb** - This locus contains the *Gaba* gene (`AGAP006028`) and is also known as the resistance to dieldrin (*Rdl*) locus. This gene encodes another component of the nervous system, the gamma-aminobutyric acid receptor subunit, which is the binding target for the insecticide dieldrin. Dieldrin use ceased in the 1970s, but resistance has remained persistent in *Anopheles* for decades afterwards (Du et al., 2005). Amino acid substitutions are known in *An. gambiae* and *An. coluzzii* to confer dieldrin resistance (Du et al., 2005; Lawniczak et al., 2010).

Although these five loci contain genes with a known role in insecticide resistance, the fact that there are strong signals of selection in multiple populations in the Ag1000g phase 1 cohort provides valuable confirmation that these loci are indeed playing an important role in adaptation to insecticide pressure in natural malaria vector populations. Because we have a strong prior expectation for selection at these loci, they also provide us with valuable positive controls, allowing us to study the character of true selection signals in more detail. This in turn can guide the design and calibration of algorithms for discovering signals of selection for insecticide resistance at novel loci.

Selection signal discovery and mapping

To learn more about the character of signals of selection for insecticide resistance, I studied the selection scan statistics in more detail at the five known insecticide resistance loci listed above. A common feature of selection signals at these loci was a clear peak architecture, with a maximum (peak) value close to the target gene, and values of the selection statistic decaying to background levels both upstream and downstream of the gene. In the H12 selection scans in particular, values on either side of the target gene appeared to decay asymptotically, and it occurred to me to model this decay by fitting exponential functions to each flank of a selection signal via least-squares regression. This exponential peak model provided a good fit to the selection signals at many of the known insecticide resistance loci (e.g., Fig. 5.2a-g). I also tested several other peak models, and found that a Gaussian peak model provided a better fit in a minority of cases (e.g., Fig. 5.2h-l). At the *Vgsc* locus the peaks were highly skewed when plotted against physical genome coordinates, but this

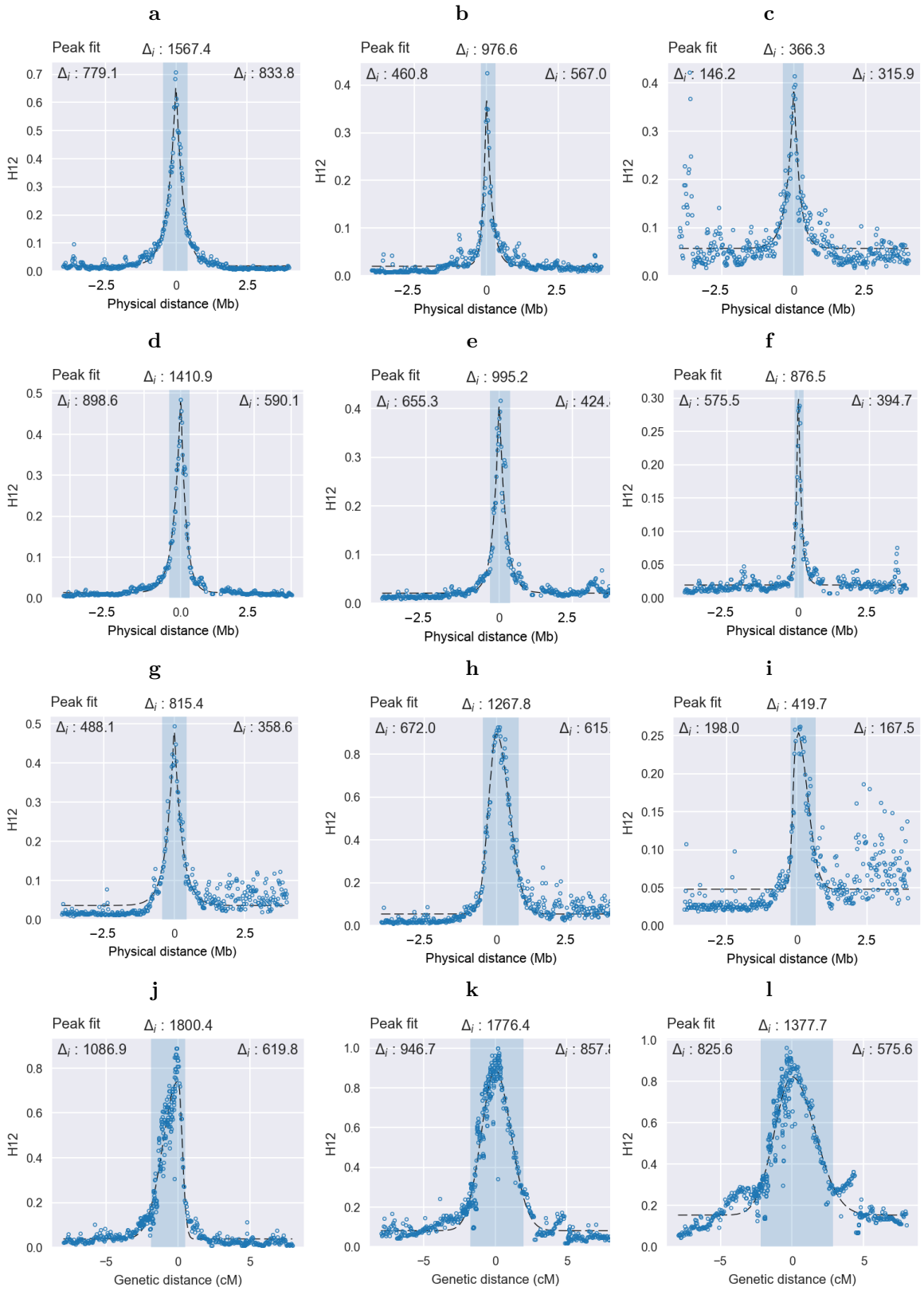

Figure 5.2. Examples of peak models fitted to H12 selection scans at known insecticide resistance loci. Dashed lines show model fit, markers show data. **a**, *Cyp6p*, Uganda *An. gambiae*. **b**, *Cyp6p*, Burkina Faso *An. gambiae*. **c**, *Cyp6p*, Burkina Faso *An. coluzzii*. **d**, *Gste*, Cameroon *An. gambiae*. **e**, *Gste*, Burkina Faso *An. gambiae*. **f**, *Gste*, Uganda *An. gambiae*. **g**, *Cyp9k1*, Burkina Faso *An. gambiae*. **h**, *Cpy9k1*, Burkina Faso *An. coluzzii*. *i*, *Cyp9k1, Guinea An. gambiae*. **j**, *Vgsc*, Burkina Faso *An. coluzzii*. **k**, *Vgsc*, Burkina Faso *An. gambiae*. **l**, *Vgsc*, Uganda *An. gambiae*.

locus lies on the border of pericentromeric heterochromatin, and this skew could be almost entirely corrected by assuming the heterochromatin recombination rate is four times lower than euchromatin. I devised a way to quantify the support for these peak models by also fitting a null constant model at the same locus and computing the difference in Akaike Information Criterion (Δ_i) between the peak and null models. In general, when comparing two models, $\Delta_i > 10$ is considered sufficient evidence to reject the model with higher AIC (Burnham and Anderson, 2002), and at many of the known insecticide resistance loci I observed $\Delta_i > 1000$, indicating strong support for the exponential peak model.

Previous studies performing genome-wide selection scans, such as Garud et al. (2015), have identified peaks by first choosing a fixed significance threshold, and then defining peaks to be contiguous genome regions where selection scan values exceed this threshold. Such an approach, however, can have several shortcomings, illustrated in the Ag1000G data in Fig. 5.3. Firstly, determining an appropriate significance threshold is not straightforward, and requires knowledge of the demographic history of the population being studied, which may be unknown or difficult to infer with any certainty. Secondly, different types of noise within the data can lead to incorrect inferences, such as mistakenly inferring a signal from an isolated high statistic value, or mistakenly splitting a peak into two signals due to an isolated low statistic value. Thirdly, from inspecting another known insecticide resistance gene, *Ace1* (`AGAP001356`), it was also clear that signals might be small in magnitude despite having a clear peak architecture, and so fall below a fixed significance threshold. Examining the Ag1000G selection scans at known insecticide resistance loci suggests how these issues could be overcome by leveraging the information provided by the shape of peaks at true selection signals. At a locus under recent positive selection, genetic hitchhiking of neutral variants on either flank of the locus is expected due to linkage disequilibrium, and the probability of linkage is an exponential function of the genetic distance between them (Maynard Smith and Haigh, 1974). Wiener and Pong-Wong (2011) showed that this theory could be used as a tool for mapping selection signals, by fitting exponential functions to heterozygosity data on either flank of a locus under positive selection via least squares regression. The primary benefit of this type of approach to modelling support for

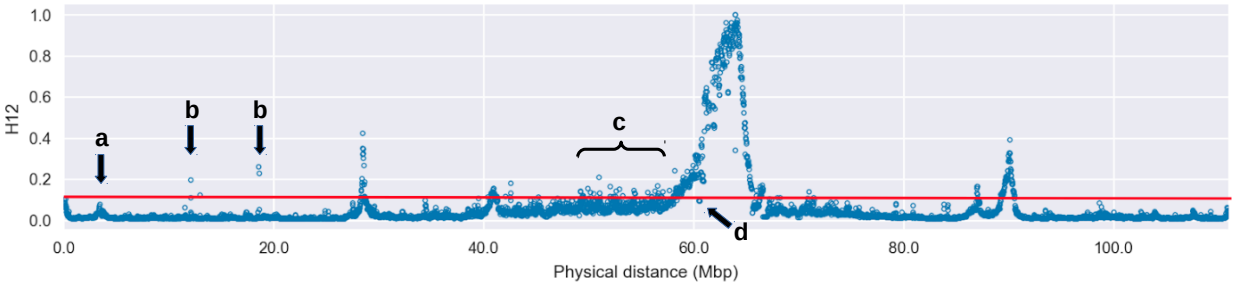

Figure 5.3. Illustration of potential problems with peak identification using consecutive windows exceeding a fixed threshold, as applied by Garud et al. (2015). Blue markers show H12 values from selection scan in Burkina Faso *An. gambiae*. Red line shows an arbitrary significance threshold, chosen to illustrate potential problems in these data (these problems are associated with the use of a fixed threshold and can arise regardless of the actual threshold value chosen). **a**, This locus contains the known insecticide resistance gene *Ace1*, and a small peak is visible in the selection scan, but the magnitude of the peak falls below the threshold and so would not be discovered. **b**, Isolated high values of the selection statistic are probably false positives due to noise, such signals are never observed at known insecticide resistance genes. **c**, Some genome regions can be inherently noisier than others, i.e., background levels of the selection statistic are higher. **d**, Isolated low values of the selection statistic could falsely lead us to break up a single peak into two peaks, such as at this selection signal at the *Vgsc* insecticide resistance gene.

a selection signal is that it integrates evidence not just at the locus under selection, but from the flanking regions as well.

Developing this idea further, I devised an algorithm to systematically identify selection signals within each of the Ag1000G selection scans, quantify their support, map the most likely focus of selection, and provide some indication of the relative uncertainty surrounding the mapped focus. The algorithm uses non-linear least squares regression to fit peak and null models to the selection scan values at regular 20 kb intervals throughout the genome. After performing a complete scan through the genome, the peak model with the highest Δ_i is called as a signal. To deal with cases where true selection signals occur in close proximity and peaks may be partially overlapping, the algorithm then subtracts the fitted peak values from the data, and refits peaks in adjacent regions. This process is then applied iteratively, until no more peak models with Δ_i above a given level are found. After applying this algorithm to all Ag1000G selection scans, I found all peaks overlapping known insecticide resistance genes had $\Delta_i > 90$, and so only called signals above this level. This generated a catalog of 289 selection signals (Figs. 5.5, 5.6, 5.7, 5.8, 5.9), which I then ranked by Δ_i.

Signal discovery and mapping performance

To provide some assessment of the performance of the signal discovery algorithm, I took an empirical approach, making use of the five known insecticide resistance genes described above, in addition to the *Ace1* gene where variants are known to confer resistance to carbamate and organophosphate insecticides (Weill et al., 2003; Weill et al., 2004; Djogbénou et al., 2008).

To estimate sensitivity I analysed the *Vgsc*, *Gaba*, *Ace1* and *Gste2* genes, where SNPs conferring insecticide resistance are known from previous studies. For each of these genes I ascertained the populations in which a known resistance variant was at 5% allele frequency or greater, and used these as a truth set. I then determined a true positive to be a population in which a selection signal was found overlapping the gene, and a false negative to be a population in which no overlapping selection signal was found. Across all four loci and three selection scan methods, the overall sensitivity was 100% (Table 5.1). Sensitivity was highest in the H12 and XPEHH scans (both 79%) and lowest in the IHS scans (37%).

To estimate signal mapping accuracy I used all six known insecticide resistance genes. For each selection signal overlapping one of these genes, I computed the mapping error (ME) as the distance from the fitted peak centre to the gene center, then summarised these values across multiple scans and loci by computing the median and $25 - 75^{th}$ percentiles (Table 5.2). Across all six loci, median mapping error was lowest in the H12 scans (78 kb) and highest in the IHS scans (258 kb). To estimate a confidence interval for the focus of selection within each peak I computed the region within which peak models obtained Δ_i within 95% of the best peak model. This confidence interval was correlated with the mapping error in the H12 and XPEHH signals but not IHS (Fig. 5.10). This confidence included the target gene in 33–63 % of cases depending on scanning method, and increased to 56–68 % if extended by 50 kb. Clearly there is room for improvement here in estimating a confidence interval for the focus of the selection signal, but at least for the H12 and XPEHH scans there is an indication that some uncertainty is being captured.

A Web application for exploring selection signals

Of the 289 selection signals identified, 229 did not overlap any of the known insecticide resistance loci described above. To facilitate an exploration of these potentially novel selection signals, and an investigation of the genes that might be driving them, I developed a prototype Web application, available at `https://malariagen.github.io/agam-selection-atlas/`. Each selection signal can span multiple genes, and thus it can be time-consuming to explore signals and generate hypotheses about the most likely candidate genes under selection. The purpose of the Web application is to accelerate this investigation process, by providing a visual interface to the data, with tools for comparing selection signals across populations and selection scans, and for zooming into individual signals and inspecting genes. The Web application provides several entry points to the data, including:

- **All selection signals**. A table of selection signals, ranked by the degree of supporting evidence (Δ_i). For each signal, the table provides the selection statistic, the population in which the signal was found, the focal region, and whether the signal overlaps any known loci.

- **Signals by population**. One page for each population, with an interactive genome plot for each chromosome arm providing a means to view and browse the selection signals, and an associated table of selection signals.

- **Signals by chromosome arm**. One page for each chromosome arm, with an interactive genome plot showing signals from all populations, and an associated table of signals (Figs. 5.5, 5.6, 5.7, 5.8 and 5.9 are screenshots from these pages).

- **Signals by scanning statistic**. One page for each scanning method used (H12, XPEHH, IHS), with a table of associated selection signals from all populations and chromosome arms.

- **Known loci**. A set of pages for functionally-validated insecticide resistance loci where there is a strong expectation for observing selection signals in at least some populations.

- **Insecticide resistance candidate genes**. Pages providing catalogues of *a priori* candidate insecticide resistance genes based on gene function, organised into four lists: metabolic, target-site, behavioural and cuticular.

From these entry points, a user can access a dedicated **signal page** for each individual selection signal. Each signal page includes an interactive genome plot for the region containing and surrounding the peak, showing the selection statistic values, and the fitted peak model (Fig. 5.4 panels a-d are screenshots of this feature from four signal pages). This plot can be zoomed and panned to allow inspection of the genes nearest to the peak centre. The set of genes overlapping the inferred focal region are also listed below the plot. There is also a table of overlapping selection signals, which can be useful to assess the degree of replication across both populations and selection methods. Finally, there are a collection of diagnostic plots, and a model fit report, providing detailed information to allow assessment of how well the peak model fitted the data. If a user wants to investigate a particular gene of interested, then can click through to a dedicated **gene page**. Each gene page as summary information about the gene, with links through to VectorBase, and a table of all overlapping and adjacent selection signals.

Discovery of a novel candidate insecticide resistance gene

To illustrate the potential of these selection scans to identify novel candidate insecticide resistance genes, I provide an example of selection signals at a novel locus on the X chromosome (Fig. 5.4). Selection signals were found at this locus in both species from Burkina Faso, and were replicated in both populations across all three selection scan methods. Selection peaks were also relatively narrow, and only a single gene overlapped the peak focus in all H12 and XPEHH scans. This gene was `AGAP000519`, a diacylglycerol kinase. This gene is intriguing because its function does not obviously fit any of the established categories of insecticide resistance gene. I performed a literature search and could not find any previous studies associating diacylglycerol kinase enzymes with insecticide resistance in any insect species. I did, however, find two lines of evidence suggesting a potential role in adaptation to malaria vector control interventions.

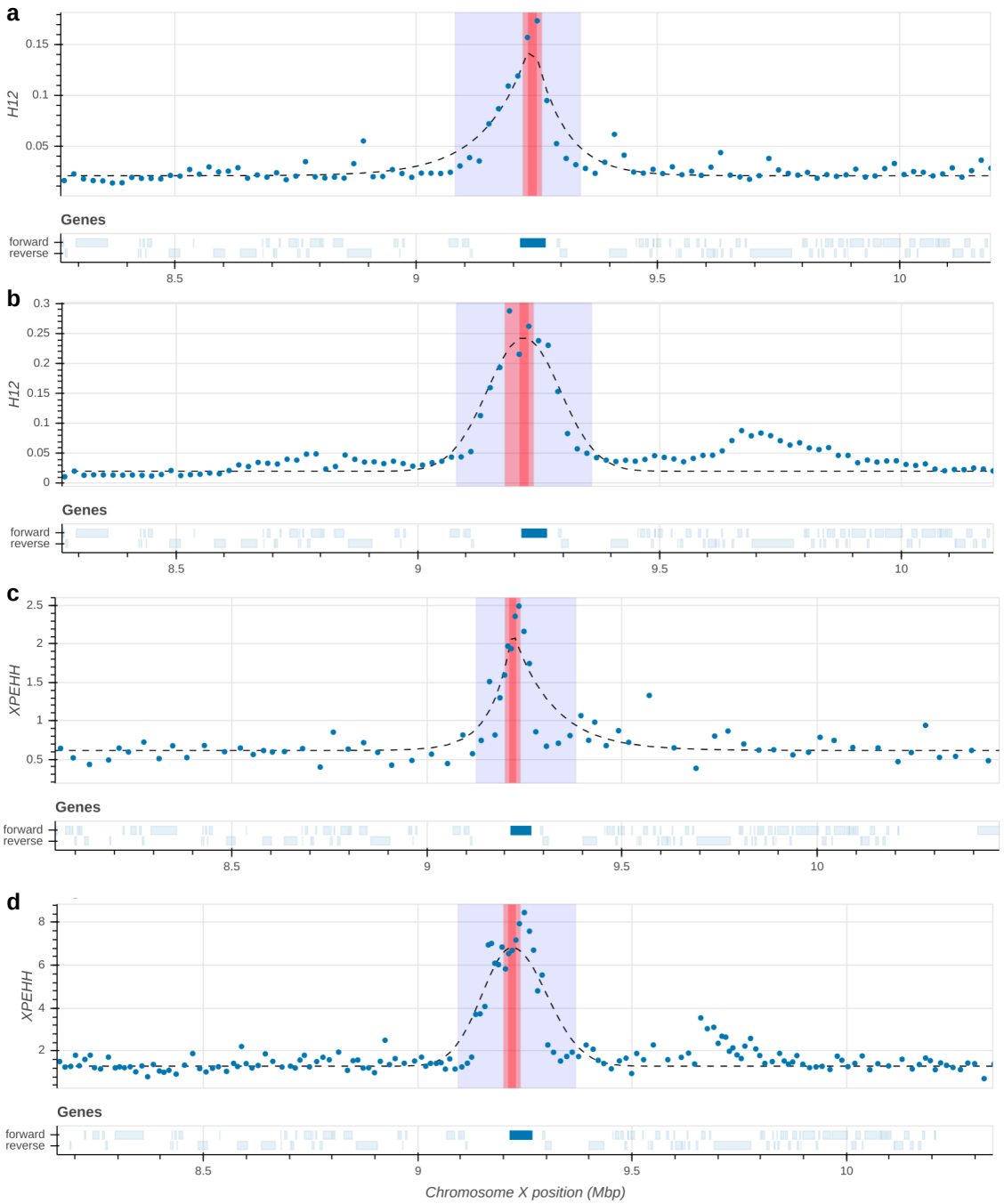

Figure 5.4. Selection signals at a novel locus on the X chromosome. The gene shown in blue in all subplots is `AGAP000519`, a diacylglycerol kinase. **a**, Burkina Faso *An. coluzzii*, H12. **b**, Burkina Faso *An. gambiae*, H12. **c**, Burkina Faso *An. coluzzii*, XPEHH versus Guinea-Bissau. **d**, Burkina Faso *An. gambiae*, XPEHH versus Guinea-Bissau.

In *C. elegans*, the diacylglycerol kinase DGK-1 is involved in regulating release of the neurotransmitter acetylcholine at synaptic junctions (Miller et al., 1999). This discovery was made by using genetic screens for mutants with resistance to aldicarb, a carbamate insecticide. Carbamates bind to acetylcholinesterase and prevent its normal function in

inactivating acetylcholine, causing a toxic accumulation of acetylcholine at synapses. Mutations in the upstream regulatory pathway that reduce the normal amount of acetylcholine present in synapses can thus alter sensitivity to carbamates. Miller et al. (1999) found mutations in several genes that confer aldicarb resistance by this means. They also found that DGK-1 is a component of this regulatory machinery, and normally acts to negatively regulate synaptic transmission. A loss of function in DGK-1 results in hypersensitivity to aldicarb, in addition to hyperactive locomotion. If synaptic transmission is regulated via a similar pathway in *Anopheles* mosquitoes, then this suggests a hypothesis that a gain of function mutation in a diacylglycerol kinase gene could confer some degree of resistance to carbamate and organophosphate insecticides, which share the same mode of action. Carbamates have been used for indoor residual spraying (IRS) fairly extensively, and organophosphates are increasingly being used, thus any novel mechanism of resistance to these insecticides in malaria vectors would be of high importance.

There is also a second hypothesis, concerning a possible role in behavioural adaptation. `AGAP000519` is a one-to-one homolog of the *D. melanogaster* gene *rdgA*. In *D. melanogaster*, *rdgA* is an eye-specific diacylglycerol kinase involved in vision via the phototransduction signalling cascade (Masai et al., 1993). *rdgA* loss-of-function mutants have constitutive activation of light-sensitive ion channels, causing retinal degeneration during development (Raghu et al., 2000). These observations have demonstrated that *rdgA* is required for light signalling response termination via the phosphoinositide cycle (Katz and Minke, 2009). Thus, any modulation of the activity of *rdgA* has the potential to affect sensitivity to light. In *An. gambiae*, genes in this visual signalling pathway, including `AGAP000519`, are expressed rhythmically under circadian clock control (Rund et al., 2011). Expression peaks just before dusk, suggesting a role in tuning the *An. gambiae* visual system ahead of nocturnal activities such as flight and biting. The increase in coverage of insecticide-treated bednets has long been hypothesised to create a selective pressure for individuals that initiate host-seeking behaviour earlier in the night, but no genetic changes have so far been found affecting this behaviour.

Diacylglycerol is a signalling molecule used in a variety of other roles, and there may

be other causes for selection at this locus in *An. gambiae* and *An. coluzzii*. Nevertheless, the presence of clear selection signals at this locus in both species, the high degree of confidence regarding the target gene, and the presence of compelling and plausible links to adaptation to vector control interventions, strongly advise for the further study of this gene. Furthermore, although the two hypotheses presented above are different, the underlying molecular pathways are the same on both cases, with diacylglycerol kinase enzymes serving to modulate the excitability of neurons. This in turn argues for a deeper study of this pathway in malaria vectors.

Conclusions

In this chapter I have explored the genomic landscape of recent selection within the mosquito populations sampled in the first phase of the Ag1000G project. There are strong signals of selection at multiple loci with clear links to insecticide resistance, demonstrating that malaria vector control interventions have driven a strong adaptive response. It is also clear that this adaptive response is both polygenic and heterogeneous over both species and geography, with different genes under selection in different combinations of populations. Furthermore, a substantial number of selection signals are present at genes not previously known to be involved in insecticide resistance, where signals are of comparable shape and statistical support to those found at validated insecticide resistance genes. This strongly suggests that the molecular landscape of insecticide resistance in *An. gambiae* and *An. coluzzii* is not fully understood and there are major gaps in our current knowledge. The Ag1000G data provide a unique opportunity to fill some of these gaps, generating new candidate genes under selection for further investigation and validation through lab and field work.

Methods

Genome-wide selection scans

Selection scans using the H12 statistic were performed following methods described in Garud et al. (2015) as implemented in scikit-allel version 0.21.1. H12 was computed in non-overlapping windows over the genome, where each window contained a fixed number of SNPs. The extent of linkage disequilibrium was different in different populations (Chapter 4) and so I calibrated the window size independently in each population. To calibrate the window sizes I ran the H12 scans with a range of different window sizes, and chose the smallest window size for which the mean value of H1 over all windows was below 0.01. XPEHH scans were computed following methods described in Sabeti et al. (2007) as implemented in scikit-allel version 0.21.1. For each population comparison, SNPs with a minor allele frequency greater than 5% in the union of both populations were used. XPEHH scores were normalised within each chromosome (2, 3, X). IHS scans were computed following methods described in Voight et al. (2006) as implemented in scikit-allel version 0.21.1. For each population, SNPs with minor allele frequency above 5% were used. IHS scores were normalised within each chromosome (2, 3, X).

Signal discovery and mapping

Signal discovery was performed for each genome-wide selection scan by fitting peak models using non-linear least squares regression, via the lmfit package version 0.9.7. The exponential peak model had the following parameters: *centre* (genomic position where the peak is centred), *amplitude* (maximum value of the peak), *decay* (exponential rate at which values decay on each flank), *skew* (variation in rate of decay between left and right flanks), *baseline* (constant noise term). The Gaussian peak model had the same parameters, except for a *sigma* (width) parameter instead of a *decay* parameter. The constant (null) model had a single *baseline* parameter. The peak-finding algorithm first stepped through the genome in increments of 20 kb, fitting both peak models and the null model, centered at each step, storing the resulting model fits. For each peak model, Δ_i was computed as the

difference between the peak and null model AIC values. After the first complete pass through a chromosome, the peak model with the highest Δ_i was located and called as a selection signal. The fitted peak model was then subtracted from the selection scan values, and peak models refitted within the surrounding region, to improve fitting of nearby peaks where flanks overlap. Selection scan data were converted to genetic distance prior to signal discovery and mapping, assuming a recombination rate of 0.5 cM/Mb in heterochromatin and 2.0 cM/Mb elsewhere. All source code files are available from `https://github.com/malariagen/agam-selection-atlas`.

Web application development

The Web application was developed using the Sphinx package version 1.6.5. Python scripts were written which generated a Sphinx documentation site in markdown format from the input data on selection scans and fitted selection signals. Sphinx was then used to generate a static HTML site, and this was deployed to GitHub pages. The dynamic plotting components (interactive genome plots) were implemented using Bokeh version 0.12.13 and exported as JavaScript to be embedded in the HTML.

Supplemental figures

Figure 5.5. Selection signals discovered on chromosome arm 2R. Each horizontal bar represents a selection signal. The full width of the bar shows the span of the peak, where the fitted peak model is above 20% of the peak amplitude. The red region within each bar shows the peak focus, representing the peak center plus a region of uncertainty.

Figure 5.6. Selection signals discovered on chromosome arm 2L. See Fig. 5.5 for figure legend.

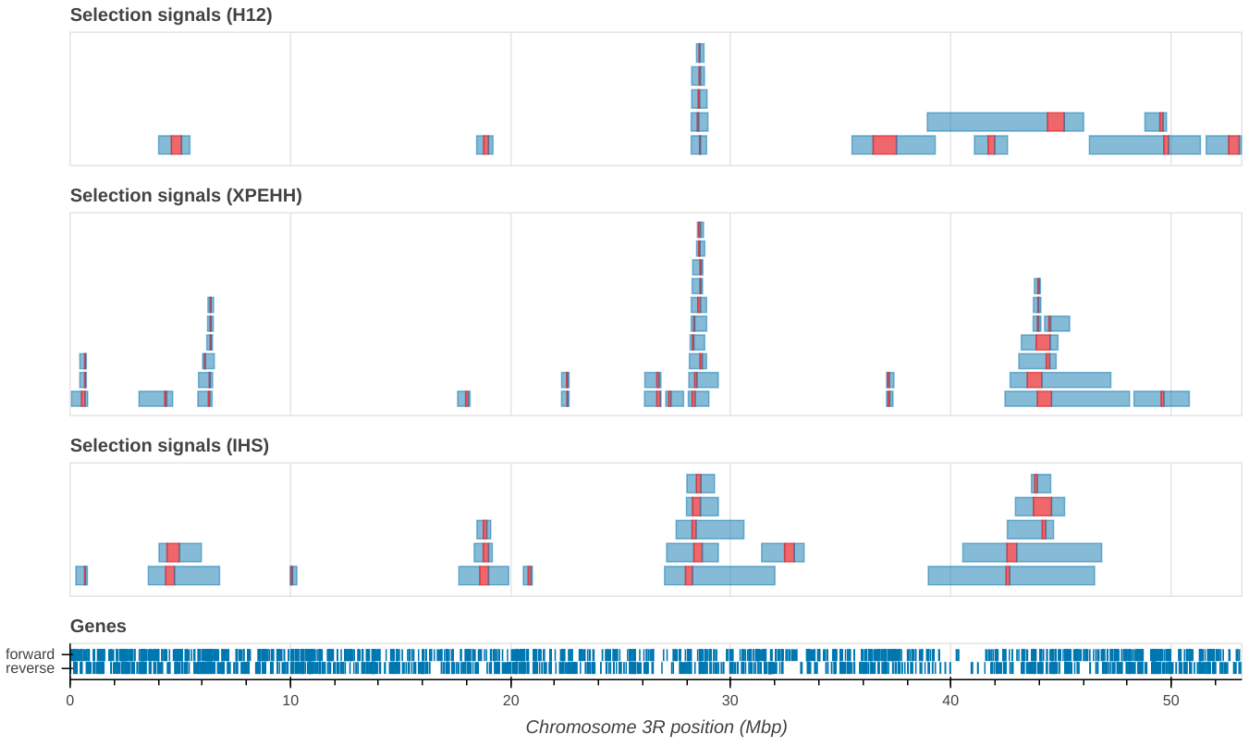

Figure 5.7. Selection signals discovered on chromosome arm 3R. See Fig. 5.5 for figure legend.

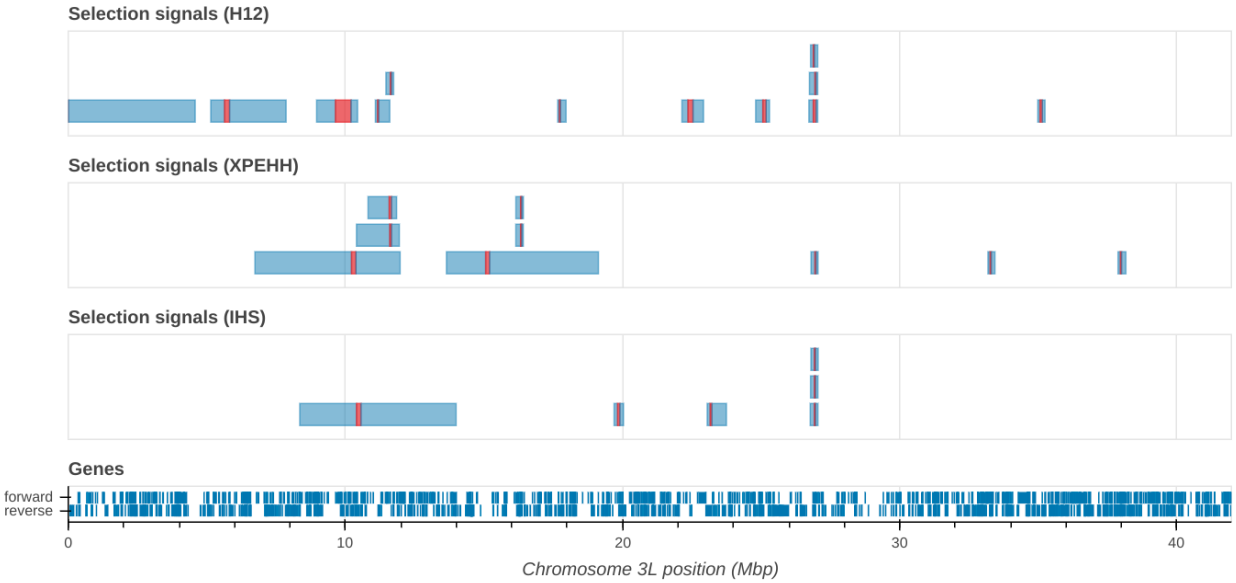

Figure 5.8. Selection signals discovered on chromosome arm 3L. See Fig. 5.5 for figure legend.

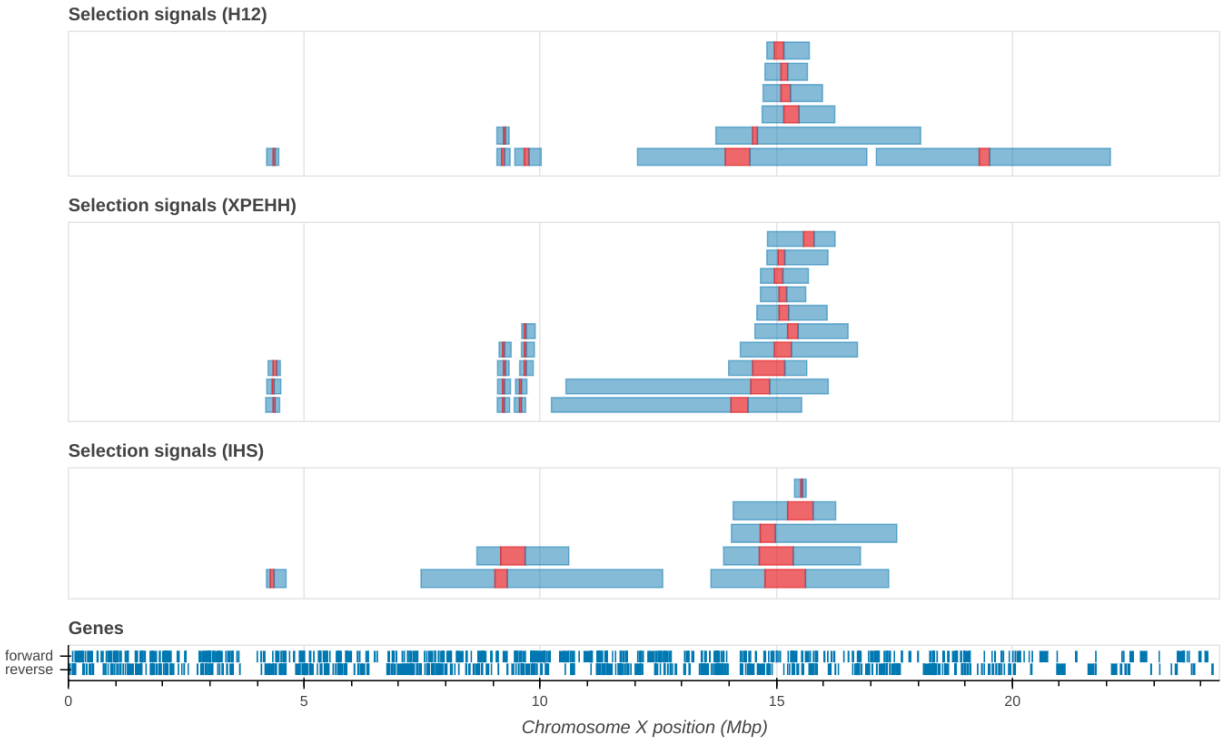

Figure 5.9. Selection signals discovered on the X chromosome. See Fig. 5.5 for figure legend.

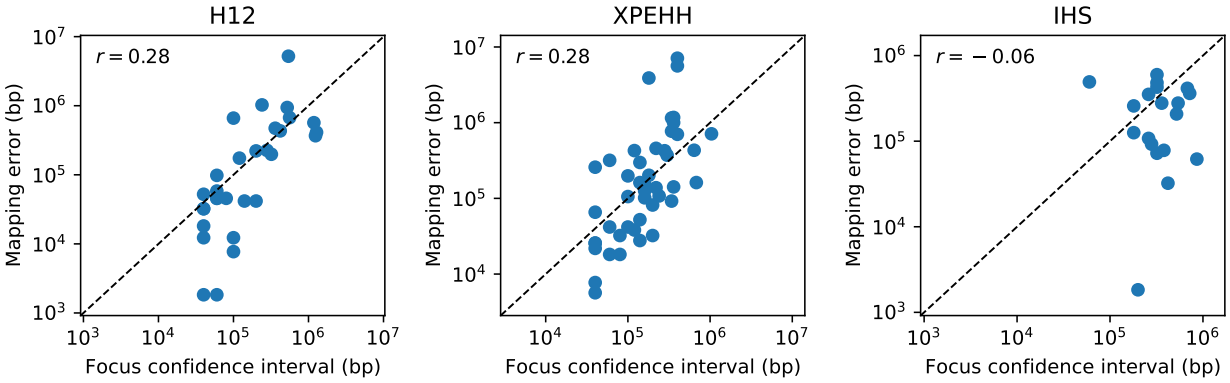

Figure 5.10. Analysis of mapping error and estimated signal focus confidence intervals at known insecticide resistance genes. Each marker represents a selection signal. Each plot compares the confidence interval for the focus of each selection interval, estimated by comparing peak model fits within the vicinity of a signal, and the mapping error, computed as the distance from the fitted peak focus to the known resistance gene. r is the Pearson correlation coefficient.

Supplemental tables

Table 5.1. Analysis of sensitivity of selection signal discovery using genes with known insecticide resistance variants.

Gene	Statistic	No. true	No. positive	No. true positive	Sensitivity (%)
Vgsc	H12	7	7	7	100
Vgsc	XPEHH	7	6	6	86
Vgsc	IHS	7	0	0	0
Vgsc	all	7	7	7	100
Gaba	H12	6	5	5	83
Gaba	XPEHH	6	5	5	83
Gaba	IHS	6	2	2	33
Gaba	all	6	6	6	100
Ace1	H12	2	0	0	0
Ace1	XPEHH	2	1	1	50
Ace1	IHS	2	2	2	100
Ace1	all	2	2	2	100
Gste2	H12	4	5	3	75
Gste2	XPEHH	4	5	3	75
Gste2	IHS	4	5	3	75
Gste2	all	4	6	4	100
all	H12	19	17	15	79
all	IHS	19	9	7	37
all	XPEHH	19	17	15	79
all	all	19	21	19	100

Table 5.2. Analysis of mapping error using selection signals at known insecticide resistance genes. N = no. of selection signals found overlapping the gene. $ME_n = n^{th}$ percentile of mapping error (distance from signal center to gene) in all selection signals spanning the gene. Within CI = percentage of signals where the gene was within the estimated focus confidence interval.

Gene	Statistic	N	ME_{25}	ME_{50}	ME_{75}	Within CI (%)	Within CI + 50 kb (%)
Ace1	IHS	2	244	280	316	50	50
Ace1	XPEHH	1	92	92	92	100	100
Cyp6p3	H12	7	12	32	42	43	86
Cyp6p3	IHS	6	77	100	348	67	67
Cyp6p3	XPEHH	8	31	42	174	62	62
Cyp9k1	H12	6	81	210	552	33	50
Cyp9k1	IHS	4	224	320	392	75	75
Cyp9k1	XPEHH	10	139	162	394	30	40
Gaba	H12	5	46	46	174	0	60
Gaba	IHS	2	198	270	342	50	100
Gaba	XPEHH	8	26	86	133	12	88
Gste2	H12	5	2	18	58	60	80
Gste2	IHS	5	78	258	278	60	60
Gste2	XPEHH	10	26	42	243	30	60
Vgsc	H12	7	422	472	620	43	43
Vgsc	XPEHH	8	757	1162	4326	25	25
all	H12	30	35	78	427	37	63
all	IHS	19	85	258	388	63	68
all	XPEHH	45	42	142	428	33	56

References

Adolfi, A, B Poulton, A Anthousi, S Macilwee, H Ranson and GJ Lycett (2019). 'Functional genetic validation of key genes conferring insecticide resistance in the major African malaria vector, Anopheles gambiae'. In: *Proc. Natl. Acad. Sci. U. S. A.* 116.51, pp. 25764–25772. DOI: 10.1073/pnas.1914633116.

Awumbila, M (2017). *Drivers of Migration and Urbanization in Africa: Key Trends and Issues*. Tech. rep. United Nations.

Bhatt, S et al. (2015). 'The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015'. In: *Nature* 526.7572, pp. 207–211. DOI: 10.1038/nature15535.

Binswanger-Mkhize, HP and S Savastano (2017). 'Agricultural intensification: The status in six African countries'. In: *Food Policy* 67, pp. 26–40. DOI: 10.1016/j.foodpol.2016.09.021.

Booker, TR, BC Jackson and PD Keightley (2017). 'Detecting positive selection in the genome'. In: *BMC Biol.* 15.1. DOI: 10.1186/s12915-017-0434-y.

Burnham, KP and DR Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer.

Churcher, TS, N Lissenden, JT Griffin, E Worrall and H Ranson (2016). 'The impact of pyrethroid resistance on the efficacy and effectiveness of bednets for malaria control in Africa'. In: *Elife* 5. DOI: 10.7554/eLife.16090.

Cibulskis, RE et al. (2016). 'Malaria: Global progress 2000 - 2015 and future challenges'. In: *Infect. Dis. Poverty* 5.1. DOI: 10.1186/s40249-016-0151-8.

Crawford, JE et al. (2017). 'Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans'. In: *Am. J. Hum. Genet.* 101.5, pp. 752–767. DOI: 10.1016/j.ajhg.2017.09.023.

Davies, T, LM Field, P Usherwood and MS Williamson (2007). 'A comparative study of voltage-gated sodium channels in the Insecta: Implications for pyrethroid resistance in Anopheline and other Neopteran species'. In: *Insect Mol. Biol.* 16.3, pp. 361–375. DOI: 10.1111/j.1365-2583.2007.00733.x.

Ding, Y, F Ortelli, LC Rossiter, J Hemingway and H Ranson (2003). 'The Anopheles gambiae glutathione transferase supergene family: Annotation, phylogeny and expression profiles'. In: *BMC Genomics* 4.1, p. 35. DOI: 10.1186/1471-2164-4-35.

Djogbénou, L, F Chandre, A Berthomieu, R Dabiré, A Koffi, H Alout and M Weill (2008). 'Evidence of introgression of the ace-1R mutation and of the ace-1 duplication in West African Anopheles gambiae s. s'. In: *PLoS One* 3.5. Ed. by DA Carter, e2172. DOI: 10.1371/journal.pone.0002172.

Dong, K, Y Du, F Rinkevich, Y Nomura, P Xu, L Wang, K Silver and BS Zhorov (2014). 'Molecular biology of insect sodium channels and pyrethroid resistance.' In: *Insect Biochem. Mol. Biol.* 50, pp. 1–17. DOI: 10.1016/j.ibmb.2014.03.012.

Du, W, TS Awolola, P Howell, LL Koekemoer, BD Brooke, MQ Benedict, M Coetzee and L Zheng (2005). 'Independent mutations in the Rdl locus confer dieldrin resistance to Anopheles gambiae and An. arabiensis.' In: *Insect Mol. Biol.* 14.2, pp. 179–183. DOI: 10.1111/j.1365-2583.2005.00544.x.

Fouet, C, AF Ashu, MM Ambadiang, WT Tchapga, CS Wondji and C Kamdem (2020). 'Resistance of Anopheles gambiae to the new insecticide clothianidin associated with unrestricted use of agricultural neonicotinoids in Yaounde, Cameroon'. In: *bioRxiv*. DOI: 10.1101/2020.08.06.239509.

Garud, NR, PW Messer, EO Buzbas and DA Petrov (2015). 'Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps'. In: *PLoS Genet.* 11.2. Ed. by GP Copenhaver, e1005004. DOI: 10.1371/journal.pgen.1005004. eprint: 1303.0906.

Georghiou, GP (1990). 'The Effect of Agrochemicals on Vector Populations'. In: *Pesticide Resistance in Arthropods*. Ed. by RT Roush and BE Tabashnik. New York: Springer, pp. 183–202. DOI: 10.1007/978-1-4684-6429-0_7.

Giraldo-Calderón, GI et al. (2014). 'VectorBase: An updated Bioinformatics Resource for invertebrate vectors and other organisms related with human diseases'. In: *Nucleic Acids Res.* 43.D1, pp. D707–D713. DOI: 10.1093/nar/gku1117.

Haasl, RJ and BA Payseur (2015). 'Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication'. In: *Mol. Ecol.* 25.1, pp. 5–23. DOI: 10.1111/mec.13339.

Hancock, PA, CJ Hendriks, JA Tangena, H Gibson, J Hemingway, M Coleman, PW Gething, E Cameron, S Bhatt and CL Moyes (2020). 'Mapping trends in insecticide resistance phenotypes in African malaria vectors'. In: *PLoS Biol.* 18.6. Ed. by AF Read, e3000633. DOI: 10.1371/journal.pbio.3000633.

Hemingway, J et al. (2016). 'Averting a malaria disaster: Will insecticide resistance derail malaria control?' In: *The Lancet* 387.10029, pp. 1785–1788. DOI: 10.1016/S0140-6736(15)00417-1.

Ingham, VA, A Anthousi, V Douris, NJ Harding, G Lycett, M Morris, J Vontas and H Ranson (2019). 'A sensory appendage protein protects malaria vectors from pyrethroids'. In: *Nature* 577.7790, pp. 376–380. DOI: 10.1038/s41586-019-1864-1.

Jones, CM, M Liyanapathirana, FR Agossa, D Weetman, H Ranson, MJ Donnelly and CS Wilding (2012). 'Footprints of positive selection associated with a mutation (N1575Y) in the voltage-gated sodium channel of Anopheles gambiae.' In: *Proc. Natl. Acad. Sci. U. S. A.* 109.17, pp. 6614–6619. DOI: 10.1073/pnas.1201475109.

Katz, B and B Minke (2009). 'Drosophila photoreceptors and signaling mechanisms'. In: *Front. Cell. Neurosci.* 3.JUN. DOI: 10.3389/neuro.03.002.2009.

Lawniczak, MKN et al. (2010). 'Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences'. In: *Science* 330.6003, pp. 512–514. DOI: 10.1126/science.1195755.

Lynd, A, D Weetman, S Barbosa, A Egyir Yawson, S Mitchell, J Pinto, I Hastings and MJ Donnelly (2010). 'Field, genetic, and modeling approaches show strong positive selection acting upon an insecticide resistance mutation in Anopheles gambiae s.s.' In: *Mol. Biol. Evol.* 27.5, pp. 1117–1125. DOI: 10.1093/molbev/msq002.

Main, BJ, Y Lee, TC Collier, LC Norris, K Brisco, A Fofana, AJ Cornel and GC Lanzaro (2015). 'Complex genome evolution in Anopheles coluzzii associated with increased insecticide usage in Mali'. In: *Mol. Ecol.* 24.20, pp. 5145–5157. DOI: 10.1111/mec.13382.

Martinez-Torres, D, F Chandre, MS Williamson, F Darriet, JB Bergé, AL Devonshire, P Guillet, N Pasteur and D Pauron (1998). 'Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector Anopheles gambiae s.s.' In: *Insect Mol. Biol.* 7.2, pp. 179–84.

Masai, I, A Okazaki, T Hosoya and Y Hotta (1993). 'Drosophila retinal degeneration A gene encodes an eye-specific diacylglycerol kinase with cysteine-rich zinc-finger motifs and ankyrin repeats'. In: *Proc. Natl. Acad. Sci. U. S. A.* 90.23, pp. 11157–11161. DOI: 10.1073/pnas.90.23.11157.

Maynard Smith, J and J Haigh (1974). 'The hitch-hiking effect of a favourable gene'. In: *Genet. Res. (Camb).* 89.5-6, pp. 391–403. DOI: 10.1017/S0016672308009579.

Miller, KG, MD Emerson and JB Rand (1999). 'G(o)α and diacylglycerol kinase negatively regulate the G(q)α pathway in C. elegans'. In: *Neuron* 24.2, pp. 323–333. DOI: 10.1016/S0896-6273(00)80847-8.

Mitchell, SN et al. (2014). 'Metabolic and target-site mechanisms combine to confer strong DDT resistance in Anopheles gambiae.' In: *PLoS One* 9.3. Ed. by K Michel, e92662. DOI: 10.1371/journal.pone.0092662.

Mohammed, BR, M S, Malang, S Kawe, RIS Agbede and RD Finn (2017). 'Cytochrome P450s in Anopheles gambiae (Diptera: Culicidae) and Insecticide Resistance in Africa: A Mini Review'. In: *Entomol. Ornithol. Herpetol. Curr. Res.* 06.03. DOI: 10.4172/2161-0983.1000200.

Müller, P et al. (2008). 'Field-caught permethrin-resistant Anopheles gambiae overexpress CYP6P3, a P450 that metabolises pyrethroids.' In: *PLoS Genet.* 4.11. Ed. by DL Stern, e1000286. DOI: 10.1371/journal.pgen.1000286.

Nkya, TE, I Akhouayri, W Kisinza and JP David (2013). 'Impact of environment on mosquito response to pyrethroid insecticides: Facts, evidences and prospects'. In: *Insect Biochem. Mol. Biol.* 43.4, pp. 407–416. DOI: 10.1016/j.ibmb.2012.10.006.

OECD and Sahel and West Africa Club (2020). *Africa's Urbanisation Dynamics 2020*. Paris: Organisation for Economic Cooperation and Development (OECD). DOI: 10.1787/b6bccb81-en.

Okumu, F (2020). 'The fabric of life: What if mosquito nets were durable and widely available but insecticide-free?' In: *Malar. J.* 19.1. DOI: 10.1186/s12936-020-03321-6.

Oleksyk, TK, MW Smith and SJ O'Brien (2010). 'Genome-wide scans for footprints of natural selection'. In: *Philos. Trans. R. Soc. B Biol. Sci.* 365.1537, pp. 185–205. DOI: 10.1098/rstb.2009.0219.

Otsuka, K and F Place (2014). *Changes in Land Tenure and Agricultural Intensification in Sub-Saharan Africa*. Tech. rep. UNU-WIDER. DOI: 10.35188/unu-wider/2014/772-1.

Pavlidis, P and N Alachiotis (2017). 'A survey of methods and tools to detect recent and strong positive selection'. In: *J. Biol. Res.* 24.1, p. 7. DOI: 10.1186/s40709-017-0064-0.

Philbert, A, SL Lyantagaye and G Nkwengulila (2014). 'A Review of Agricultural Pesticides Use and the Selection for Resistance to Insecticides in Malaria Vectors'. In: *Adv. Entomol.* 02.03, pp. 120–128. DOI: 10.4236/ae.2014.23019.

Prapanthadara, LA, J Hemingway and AJ Ketterman (1993). 'Partial Purification and Characterization of Glutathione S-Transferases Involved in DDT Resistance from the Mosquito Anopheles gambiae'. In: *Pestic. Biochem. Physiol.* 47.2, pp. 119–133. DOI: 10.1006/pest.1993.1070.

Raghu, P, K Usher, S Jonas, S Chyb, A Polyanovsky and RC Hardie (2000). 'Constitutive activity of the light-sensitive channels TRP and TRPL in the Drosophila diacylglycerol kinase mutant, rdgA'. In: *Neuron* 26.1, pp. 169–179. DOI: 10.1016/S0896-6273(00)81147-2.

Ranson, H, B Jensen, X Wang, L Prapanthadara, J Hemingway and FH Collins (2000a). 'Genetic mapping of two loci affecting DDT resistance in the malaria vector Anopheles gambiae'. In: *Insect Mol. Biol.* 9.5, pp. 499–507. DOI: 10.1046/j.1365-2583.2000.00214.x.

Ranson, H, L Rossiter, F Ortelli, B Jensen, X Wang, CW Roth, FH Collins and J Hemingway (2001). 'Identification of a novel class of insect glutathione S-transferases involved in resistance to DDT in the malaria vector Anopheles gambiae'. In: *Biochem. J.* 359.2, pp. 295–304. DOI: 10.1042/0264-6021:3590295.

Ranson, H, B Jensen, JM Vulule, X Wang, J Hemingway and FH Collins (2000b). 'Identification of a point mutation in the voltage-gated sodium channel gene of Kenyan Anopheles gambiae associated with resistance to DDT and pyrethroids.' In: *Insect Mol. Biol.* 9.5, pp. 491–7. DOI: 10.1046/j.1365-2583.2000.00209.x.

Ranson, H and N Lissenden (2016). 'Insecticide Resistance in African Anopheles Mosquitoes: A Worsening Situation that Needs Urgent Action to Maintain Malaria Control'. In: *Trends Parasitol.* 32.3, pp. 187–196. DOI: 10.1016/j.pt.2015.11.010.

RBM (2008). *The global malaria action plan for a malaria-free world.* Tech. rep. Roll Back Malaria Partnership.

Reid, MC and FE McKenzie (2016). 'The contribution of agricultural insecticide use to increasing insecticide resistance in African malaria vectors'. In: *Malar. J.* 15.1. DOI: 10.1186/s12936-016-1162-4.

Rund, SS, TY Hou, SM Ward, FH Collins and GE Duffield (2011). 'Genome-wide profiling of diel and circadian gene expression in the malaria vector Anopheles gambiae'. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.32, E421–E430. DOI: 10.1073/pnas.1100584108.

Sabeti, PC et al. (2007). 'Genome-wide detection and characterization of positive selection in human populations'. In: *Nature* 449.7164, pp. 913–918. DOI: 10.1038/nature06250.

Sternberg, ED and MB Thomas (2018). 'Insights from agriculture for the management of insecticide resistance in disease vectors'. In: *Evol. Appl.* 11.4, pp. 404–414. DOI: 10.1111/eva.12501.

Vatsiou, AI, E Bazin and OE Gaggiotti (2016). 'Detection of selective sweeps in structured populations: A comparison of recent methods'. In: *Mol. Ecol.* 25.1, pp. 89–103. DOI: 10.1111/mec.13360.

VectorBase (2019). *Anopheles gambiae PEST, AgamP4.12*. Tech. rep. VectorBase, www.vectorbase.org (Giraldo-Calderón et al. 2015).

Voight, BF, S Kudaravalli, X Wen and JK Pritchard (2006). 'A map of recent positive selection in the human genome'. In: *PLoS Biol.* 4.3. Ed. by L Hurst, e72. DOI: 10.1371/journal.pbio.0040072.

Vontas, J et al. (2018). 'Rapid selection of a pyrethroid metabolic enzyme CYP9K1 by operational malaria control activities'. In: *Proc. Natl. Acad. Sci. U. S. A.* 115.18, pp. 4619–4624. DOI: 10.1073/pnas.1719663115.

Wang, L, Y Nomura, Y Du, N Liu, BS Zhorov and K Dong (2015). 'A mutation in the intracellular loop III/IV of mosquito sodium channel synergizes the effect of mutations in helix IIS6 on pyrethroid resistance'. In: *Mol. Pharmacol.* 87.3, pp. 421–429. DOI: 10.1124/mol.114.094730.

Weill, M, C Malcolm, F Chandre, K Mogensen, A Berthomieu, M Marquine and M Raymond (2004). 'The unique mutation in ace-1 giving high insecticide resistance is easily

detectable in mosquito vectors'. In: *Insect Mol. Biol.* 13.1, pp. 1–7. DOI: 10.1111/j.1365-2583.2004.00452.x.

Weill, M, G Luffalla, K Mogensen, F Chandre, A Berthomieu, C Berticat, N Pasteur, A Philips, P Fort and M Raymond (2003). 'Insecticide resistance in mosquito vectors'. In: *Nature* 423.6936, pp. 136–137. DOI: 10.1038/423136b.

White, BJ, MKN Lawniczak, C Cheng, MB Coulibaly, MD Wilson, N Sagnon, C Costantini, F Simard, GK Christophides and NJ Besansky (2010). 'Adaptive divergence between incipient species of Anopheles gambiae increases resistance to Plasmodium.' In: *Proc. Natl. Acad. Sci. U. S. A.* 108.1, pp. 244–249. DOI: 10.1073/pnas.1013648108.

WHO (2005). *Scaling up insecticide-treated netting programmes in Africa - A strategic framework for coordinated national action*. Tech. rep. World Health Organization.

WHO (2017a). *Achieving and maintaining universal coverage with long-lasting insecticidal nets for malaria control*. Tech. rep. World Health Organization.

WHO (2017b). *Conditions for deployment of mosquito nets treated with a pyrethroid and piperonyl butoxide*. Tech. rep. World Health Organization.

WHO (2019). *World malaria report 2019*. Tech. rep. World Health Organization.

WHO (2020). *Prequalified Lists - Vector control products - Prequalified Products 26 August 2020*. Tech. rep. World Health Organization.

Wiener, P and R Pong-Wong (2011). 'A Regression-Based Approach to Selection Mapping'. In: *J. Hered.* 102.3, pp. 294–305. DOI: 10.1093/jhered/esr014.

Yi, X et al. (2010). 'Sequencing of 50 human exomes reveals adaptation to high altitude'. In: *Science* 329.5987, pp. 75–78. DOI: 10.1126/science.1190371.

6 The evolution and spread of target-site resistance to pyrethroid insecticides

Resistance to pyrethroid insecticides is a serious challenge for malaria vector control in Africa, because pyrethroids remain a vital component of all long-lasting insecticidal bed-nets currently available for public health use. Target-site resistance to pyrethroids involves nucleotide substitutions in the voltage-gated sodium channel gene (Vgsc), which encodes an essential component of the insect nervous system and the binding target for pyrethroid molecules. In this chapter, I use genome variation data from phase one of the Ag1000G project to study the molecular evolution and geographical spread of pyrethroid target-site resistance among nine mosquito populations. I describe non-synonymous nucleotide variation throughout the entire gene coding sequence, including both known and novel polymorphisms, and report population allele frequencies and patterns of linkage disequilibrium. I then analyse the genetic backgrounds on which resistance alleles are found, to look for evidence of the geographical spread of resistance via contemporary gene flow, and to confirm positive selection for resistance alleles. I conclude by identifying a smaller subset of marker SNPs which could be used to track the further spread of resistance via low-cost high-throughput genetic assays. These analyses show that the molecular basis of pyrethroid target-site resistance is substantially more complex and diverse than previously appreciated in these species, and demonstrate that long-distance gene flow between countries and adaptive introgression between species are both playing a major part in the rising prevalence of resistance alleles.

Introduction

Pyrethroids in malaria vector control

Pyrethroids are a class of synthetic insecticides, based on a natural compound pyrethrin found in the flowers of pyrethrum (Chrysanthemum) plants (Elliott, 1989). Pyrethroids suitable for commercial use in agriculture and public health were discovered in the 1970s, and fall into two major groups based on chemical structure, either type I (e.g., permethrin (Elliott et al., 1973)) or type II (e.g., deltamethrin (Elliott et al., 1974)). These compounds have a high toxicity to insects but relatively low toxicity to mammals, and are photostable but do not accumulate to contaminate the environment unlike insecticides used previously such as DDT or dieldrin. Pyrethroids were approved for use in public health by the WHO pesticide evaluation scheme (WHOPES) in 1978 (Quélennec, 1988). Landmark studies during the 1980s and 1990s showed that the use of pyrethroid-treated bed-nets caused a significant reduction in malaria prevalence (Carnevale and Gay, 2019). Advances in net manufacturing during that period allowed the development of long-lasting insecticidal nets (LLINs), which retain insecticidal activity for up to 3 years without re-treatment. Pyrethroid LLINs have become the cornerstone of efforts to control malaria in Africa, with more than 100 million nets distributed in Africa each year since 2013 (Bhatt et al., 2015; AMP, 2020).

Pyrethroid resistance in the *Anopheles gambiae* complex

The first reports of pyrethroid resistance in *An. gambiae* originated from Côte d'Ivoire (Elissa et al., 1993) and Kenya (Vulule et al., 1994). Subsequently, a study of six countries found pyrethroid resistance in Burkina Faso, Côte d'Ivoire and Benin (Chandre et al., 1999). In parallel with the major scale-up of LLIN distributions from 2000 onwards, malaria control programmes have routinely performed bioassays to monitor pyrethroid resistance (WHO, 2018b; WHO, 2017). As those data have accumulated, it has become clear that both the prevalence and intensity of pyrethroid resistance have increased (Ranson et al., 2011; Hemingway et al., 2016; WHO, 2012; WHO, 2018a; Implications of Insecticide Resistance Consortium, 2018). Geostatistical modelling of bioassay data has supported

this, showing a dramatic increase in the prevalence of resistance across sub-Saharan Africa over the period 2005–2017, although the picture remains complex, with considerable spatial heterogeneity (Hancock et al., 2020).

The pyrethroid mode of action

Pyrethroid molecules interact with the voltage-gated sodium channel (VGSC), an essential membrane protein which propagates nerve impulses via action potentials (Dong et al., 2014). In all insects, the VGSC protein comprises four homologous domains (I-IV), each of which has six transmembrane segments (S1-S6), which together form a gated pore that is sensitive to changes in membrane potential. Under normal function, the channel opens during the rising phase of an action potential, allowing sodium ions to flow into the cell, then closes shortly afterwards, to allow re-polarisation of the membrane. The structure of the protein creates some rotational symmetry, and it is believed that pyrethroids can bind to either of two analogous sites within the pore, referred to as PyR1 and PyR2 (Du et al., 2013). Pyrethroid binding alters gating behaviour, enhancing activation and inhibiting inactivation, causing the channel to remain open, which at the cellular level causes continuous firing of nerve impulses (Dong et al., 2014). The VGSC protein of *An. gambiae* comprises 2139 amino acids and has a high degree of homology with other insects, sharing the same overall topology (Davies et al., 2007).

The molecular basis of pyrethroid target-site resistance in *An. gambiae*

Any variations within the VGSC protein which alter the action of pyrethroids are known collectively as pyrethroid target-site resistance. Prior to the present study, the molecular basis of pyrethroid target-site resistance in *An. gambiae* appeared relatively straightforward. An `L995F` substitution, initially found in West Africa (Martinez-Torres et al., 1998), and an `L995S` substitution, initially found in East Africa (Ranson et al., 2000), have both been shown to confer resistance to pyrethroids and DDT[1]. A third substitution, `N1570Y`, was subsequently found in West and Central Africa, occurring exclusively in combination with

[1]Codon numbering is given relative to *An. gambiae* transcript `AGAP004707-RA` in gene annotation set AgamP4.4. A mapping to *Musca domestica* codon numbers is given in Table 6.1.

L995F (Jones et al., 2012). *In vitro*, VGSC double mutants carrying either L995F or L995S together with N1570Y are substantially more resistant to pyrethroids than L995F or L995S alone (Wang et al., 2015). In other insects, the molecular basis of pyrethroid target-site resistance is more varied. For example, in the invasive mosquito species *Aedes aegypti*, 11 amino acid substitutions have been found in natural populations and associated with pyrethroid resistance (Du et al., 2016; Haddi et al., 2017). Across all arthropods, more than 50 sodium channel SNPs or combinations of SNPs have been associated with pyrethroid resistance (Dong et al., 2014). Since discovery of the L995F and L995S substitutions, studies in *An. gambiae* have mostly focused on typing those polymorphisms, or sequencing a small region of the gene, and thus the full coding sequence has not been fully surveyed for variation in natural populations.

The spread of pyrethroid target-site resistance in *An. gambiae* and *An. coluzzii*

The L995F and L995S alleles have now been observed in multiple countries in both West and East Africa, with L995F being widespread particularly in West Africa (WHO, 2018a). These alleles have also been found in both *An. gambiae* and *An. coluzzii* (Clarkson et al., 2014; Norris et al., 2015; Djouaka et al., 2018). In West Africa, L995F has spread from *An. gambiae* to *An. coluzzii* (Clarkson et al., 2014; Norris et al., 2015). However, for each of these alleles, it is not clear whether it is spreading between countries via contemporary gene flow, or whether there are multiple geographical origins of resistance. Pinto et al. (2007) genotyped Vgsc codon 995 and performed partial sequencing of the upstream intron in *An. gambiae* from 15 African countries, finding two common haplotypes carrying L995F, and a further two haplotypes associated with L995S. Some of these haplotypes where shared between countries, but there were only four informative SNPs within the 438 bp region sequenced, and so resolution was not sufficient to infer gene flow with confidence.

Scope of this chapter

In this chapter I analyse data from phase 1 of the Ag1000G project to investigate the molecular evolution and geographical spread of target-site resistance to pyrethroids. I report non-synonymous SNPs within the *Vgsc* gene that occur at appreciable frequency within one or more populations. I also investigate the linkage between these SNPs, to determine whether some alleles occur in combination with others, and thus may have a synergistic effect. I then use data on non-coding SNPs both within the gene introns and in the flanking intergenic sequences to investigate the haplotypes on which resistance alleles occur. I use analyses of genetic relatedness between haplotypes to make inferences about gene flow events between populations from different geographical locations and species. The work described in this chapter was performed as part of a broader collaboration within the Ag1000G Consortium analysing insecticide resistance genes, particularly with MalariaGEN Resource Center team colleague Chris Clarkson. My contributions, described in this chapter, were to investigate the population-genetic aspects of resistance. Any work carried out jointly is noted in the relevant section.

Results

Non-synonymous SNPs within the *Vgsc* gene

There were 63 non-synonymous SNPs within the *Vgsc* gene (`AGAP004707`) that were polymorphic within the Ag1000G phase 1 dataset. For these SNPs, I computed the frequency of each non-synonymous variant allele within each of nine populations defined by species and country of origin. If an allele has a resistance phenotype and has been under positive selection, then it will occur at some appreciable frequency in one or more populations. I therefore filtered the SNPs to remove any rare variants, retaining only those SNPs with a variant allele occurring at a frequency above 4% in at least one population. This resulted in a set of 21 SNPs, including two multiallelic SNPs (Table 6.1).

The two known pyrethroid resistance variants in codon 995 were at the highest overall frequency in the cohort. `L995F` was at high frequency in both *An. coluzzii* populations,

Table 6.1. Non-synonymous SNPs in the *Vgsc* gene. AO=Angola; BF=Burkina Faso; GN=Guinea; CM=Cameroon; GA=Gabon; UG=Uganda; KE=Kenya; GW=Guinea-Bissau; *Ac*=*An. coluzzii*; *Ag*=*An. gambiae*. Species status of specimens from Kenya and Guinea-Bissau is uncertain. All variants are at 4% frequency or above in one or more of the 9 Ag1000G phase 1 populations, with the exception of `2,400,071 G>T` which is only found in the CM*Ag* population at 0.4% frequency but is included because another mutation (`2,400,071 G>A`) is found at the same position causing the same amino acid substitution (`M490I`).

Variant			Population allele frequency (%)								
Position[1]	*Ag*[2]	*Md*[3]	AO*Ac*	BF*Ac*	GN*Ag*	BF*Ag*	CM*Ag*	GA*Ag*	UG*Ag*	KE	GW
`2,390,177 G>A`	`R254K`	`R261`	0	0	0	0	32	21	0	0	0
`2,391,228 G>C`	`V402L`	`V410`	0	7	0	0	0	0	0	0	0
`2,391,228 G>T`	`V402L`	`V410`	0	7	0	0	0	0	0	0	0
`2,399,997 G>C`	`D466H`	`-`	0	0	0	0	7	0	0	0	0
`2,400,071 G>A`	`M490I`	`M508`	0	0	0	0	0	0	0	18	0
`2,400,071 G>T`	`M490I`	`M508`	0	0	0	0	0	0	0	0	0
`2,416,980 C>T`	`T791M`	`T810`	0	1	13	14	0	0	0	0	0
`2,422,651 T>C`	`L995S`	`L1014`	0	0	0	0	15	64	100	76	0
`2,422,652 A>T`	`L995F`	`L1014`	86	85	100	100	53	36	0	0	0
`2,424,384 C>T`	`A1125V`	`K1133`	9	0	0	0	0	0	0	0	0
`2,425,077 G>A`	`V1254I`	`I1262`	0	0	0	0	0	0	0	0	5
`2,429,617 T>C`	`I1527T`	`I1532`	0	14	0	0	0	0	0	0	0
`2,429,745 A>T`	`N1570Y`	`N1575`	0	26	10	22	6	0	0	0	0
`2,429,897 A>G`	`E1597G`	`E1602`	0	0	6	4	0	0	0	0	0
`2,429,915 A>C`	`K1603T`	`K1608`	0	5	0	0	0	0	0	0	0
`2,430,424 G>T`	`A1746S`	`A1751`	0	0	11	13	0	0	0	0	0
`2,430,817 G>A`	`V1853I`	`V1858`	0	0	8	5	0	0	0	0	0
`2,430,863 T>C`	`I1868T`	`I1873`	0	0	18	25	0	0	0	0	0
`2,430,880 C>T`	`P1874S`	`P1879`	0	21	0	0	0	0	0	0	0
`2,430,881 C>T`	`P1874L`	`P1879`	0	7	45	26	0	0	0	0	0
`2,431,019 T>C`	`F1920S`	`Y1925`	0	0	0	0	1	4	0	0	0
`2,431,061 C>T`	`A1934V`	`A1939`	0	12	0	0	0	0	0	0	0
`2,431,079 T>C`	`I1940T`	`I1945`	0	4	0	0	7	0	0	0	0

[1] Position relative to the AgamP3 reference sequence, chromosome arm 2L.

[2] Codon numbering according to *Anopheles gambiae* transcript AGAP004707-RA in gene annotations AgamP4.4.

[3] Codon numbering according to *Musca domestica* EMBL accession X96668.

at fixation in both West African *An. gambiae* populations, and also present in the two central African *An. gambiae* populations. `L995S` was at high frequency in the Kenyan population, fixed in the Uganda *An. gambiae* population, and also present in both central African *An. gambiae* populations. The combined frequency of `L995F` and `L995S` mutations in countries where both variants co-occurred was 68% in Cameroon and 100% in Gabon. The other SNP with a known resistance phenotype in *An. gambiae*, `N1570Y`, was present in *An. gambiae* samples from Guinea and Cameroon and both species from Burkina Faso.

None of the remaining SNPs have a known phenotype in *An. gambiae* or *An. coluzzii*. Among these, the following SNPs are of particular interest, based on observations in other studies:

- `P1874L/S` - `P1874L` was the most common allele after the known resistance alleles, being found at 45% and 26% frequency in *An. gambiae* from Burkina Faso and Cameroon respectively. Both `P1874L` and `P1874S` were present in *An. coluzzii* from Burkina Faso, and the combined allele frequency of codon 1874 variants was 28%. `P1874L` was also present in sequencing data reported by Jones et al. (2012) in samples of both *An. gambiae* and *An. coluzzii* from Burkina Faso. `P1874S` has not been reported elsewhere in mosquitoes, but was found in pyrethroid-resistant field strains of the diamond-back moth *Plutella xylostella* (Sonoda, 2010).

- `I1527T` - Previously reported by Jones et al. (2012) at low frequency in *An. coluzzii* from Burkina Faso, this allele was at 14% frequency in the Ag1000G *An. coluzzii* from Burkina Faso. Recently, Collins et al. (2019) reported `I1527T` in *An. gambiae* from Guinea and found it was associated with resistance to permethrin. `I1527T` has also been recently found in *Aedes albopictus* (Auteri et al., 2018).

- `V402L` - The SNP at `2L:2,391,228` had two variant alleles which both conferred `V402L`, and which were each at 7% frequency in Burkina Faso *An. coluzzii*, but not found elsewhere. Mutations in codon 402 have been shown to cause pyrethroid resistance in multiple insect species, and the pyrethroid resistance phenotype has also been confirmed and characterized in vitro (Dong et al., 2014). Recently, `V402L`

has been found in *Aedes aegypti* where it confers high levels of resistance both type I and type II pyrethroids (Haddi et al., 2017; Villanueva-Segura et al., 2019).

Associations between non-synonymous SNPs

In several insect species, two or more non-synonymous variants in the *Vgsc* gene have been observed occurring together on the same haplotype. For example, `N1570Y` is only found in *An. gambiae* in combination with `L995F` (Jones et al., 2012). Similarly, in the German cockroach *Blatella germanica*, the combinations `L993F+E434K`, `L993F+C764R` and `L993F+E434K+C764R` have been found (Tan et al., 2002). In the diamond back moth, the alleles `A1101T+P1879S` are found in combination (Sonoda, 2010). In *Aedes aegypti*, the combination `V410L+F1534C` has been observed (Haddi et al., 2017), as well as each allele separately. The occurrence of alleles in combination is interesting because these alleles may act synergistically to enhance a pyrethroid resistance phenotype. For example, in *An. gambiae* the `L995F+N1570Y` combination is approximately ten times more resistant to permethrin than `L995F` alone (Wang et al., 2015). Similarly, in the German cockroach, `L993F+E434K` and `L993F+C764R` are each approximately 20 times more resistant to deltamethrin than `L993F` alone (Tan et al., 2002), and the triple mutant `L993F+E434K+C764R` is 100 times more resistant.

To investigate associations between non-synonymous variants within the Ag1000G phase 1 dataset, I calculated the D' linkage disequilibrium (LD) statistic between all pairs of variant alleles (Lewontin, 1964). I then combined these data with allele frequencies, to identify three possible types of association:

- **Complete association** - Two alleles are only ever found in combination ($D' \approx 1$, equal allele frequencies).
- **Subordinate association** - A primary allele, sometimes found alone, and a secondary allele, only ever found in combination with the primary allele ($D' \approx 1$, with one allele at higher frequency than the other).
- **Partial association** - Two alleles, each observed alone and in combination ($-1 <$

$D' < 1$).

The computed LD values are shown together with overall allele frequencies within the Ag1000G phase 1 cohort in Fig. 6.1. Fourteen variant alleles were in subordinate association with `L995F` ($D' > 0.91$) including `N1570Y`, `P1874S` and `P1874L`. There were also two triple-mutant combinations, `L995F+D466H+I1940T` and `L995F+T791M+A1746S`. In contrast, there were no non-synonymous variant alleles in any kind of association with `L995S`. Thus, substantial molecular evolution appears to be occurring on haplotype backgrounds carrying `L995F`. Each of the two `V402L` alleles was in subordinate association with `I1527T`, but at the amino acid level, `V402L` and `I1527T` were in near-complete association, indicating a strong mutual dependence between them.

Geographical spread of pyrethroid resistance alleles

To investigate evidence for geographical spread of pyrethroid resistance alleles, I analysed patterns of genetic similarity between the 1530 haplotypes in the Ag1000G phase 1 dataset. The Ag1000G haplotypes were phased across the whole genome, and incorporate data on non-synonymous SNPs, both within the *Vgsc* introns and in the flanking intergenic regions, which provide high resolution to identify different degrees of similarity between haplotypes. I firstly analysed the region spanning the *Vgsc* gene itself, which spans 73.5 kb and contains a total of 1,710 biallelic SNPs (1,607 intronic, 103 exonic) phased in the Ag1000G dataset. I performed a hierarchical clustering of these haplotypes, based on the number of SNP differences between them, and constructed a dendrogram, shown in Fig. 6.2. The likely presence of recombination events within the genomic interval bounded by the *Vgsc* gene, and the use of a simple distance metric and clustering method, means that this dendrogram cannot be interpreted as a complete phylogeny. However, there were some clear patterns within these data. In particular, there were several large clusters of near-identical haplotypes, each of which was strongly associated with either `L995F` or `L995S`. To provide a means of navigating these data, I cut the dendrogram at a maximum distance of 12 SNPs. I then labelled the five largest clusters containing the `L995F` allele as F1-F5, and the five largest clusters containing the `L995S` allele as S1-S5. The clusters

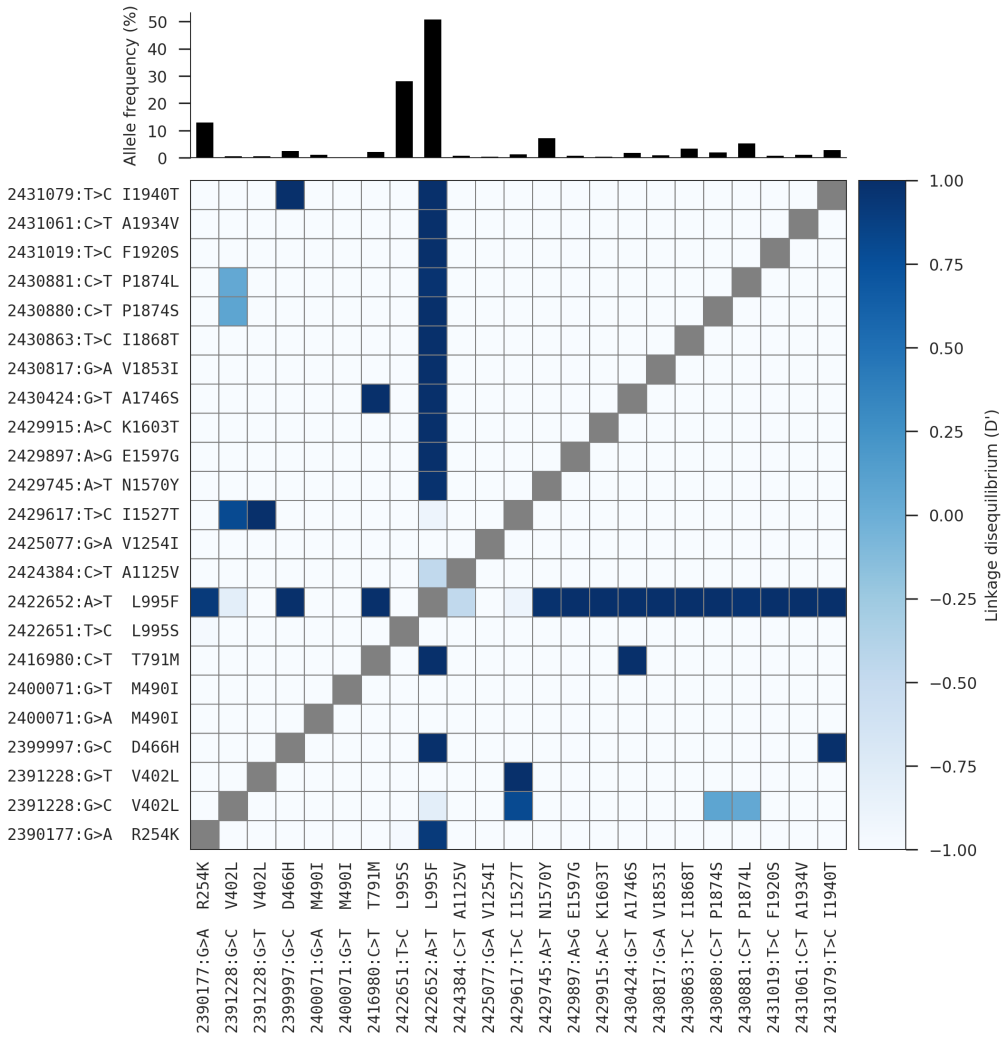

Figure 6.1. Linkage disequilibrium (D') between non-synonymous variant alleles. A value of 1 indicates that two alleles are in perfect linkage, meaning that one of the alleles is only ever found in combination with the other. Conversely, a value of -1 indicates that two alleles are never found in combination with each other. The bar plot at the top shows the frequency of each allele within the Ag1000G phase 1 cohort. See Table 6.1 for population allele frequencies.

F1-F5 together accounted for 96% of haplotypes carrying the `L995F` allele, and the clusters S1-S5 together accounted for 99% of haplotypes carrying the `L995S` allele.

Five of these haplotype clusters carrying resistance alleles contained haplotypes from different countries and/or species. Cluster F1 contained haplotypes from Burkina Faso, Guinea, Cameroon and Angola, and from both *An. gambiae* and *An. coluzzii*. Clusters F4, F5 and S2 each contained haplotypes from both Cameroon and Gabon *An. gambiae*. Cluster S3 contained haplotypes from both Uganda *An. gambiae* and Kenya. To confirm that this degree of haplotype similarity between different populations is unusual, and

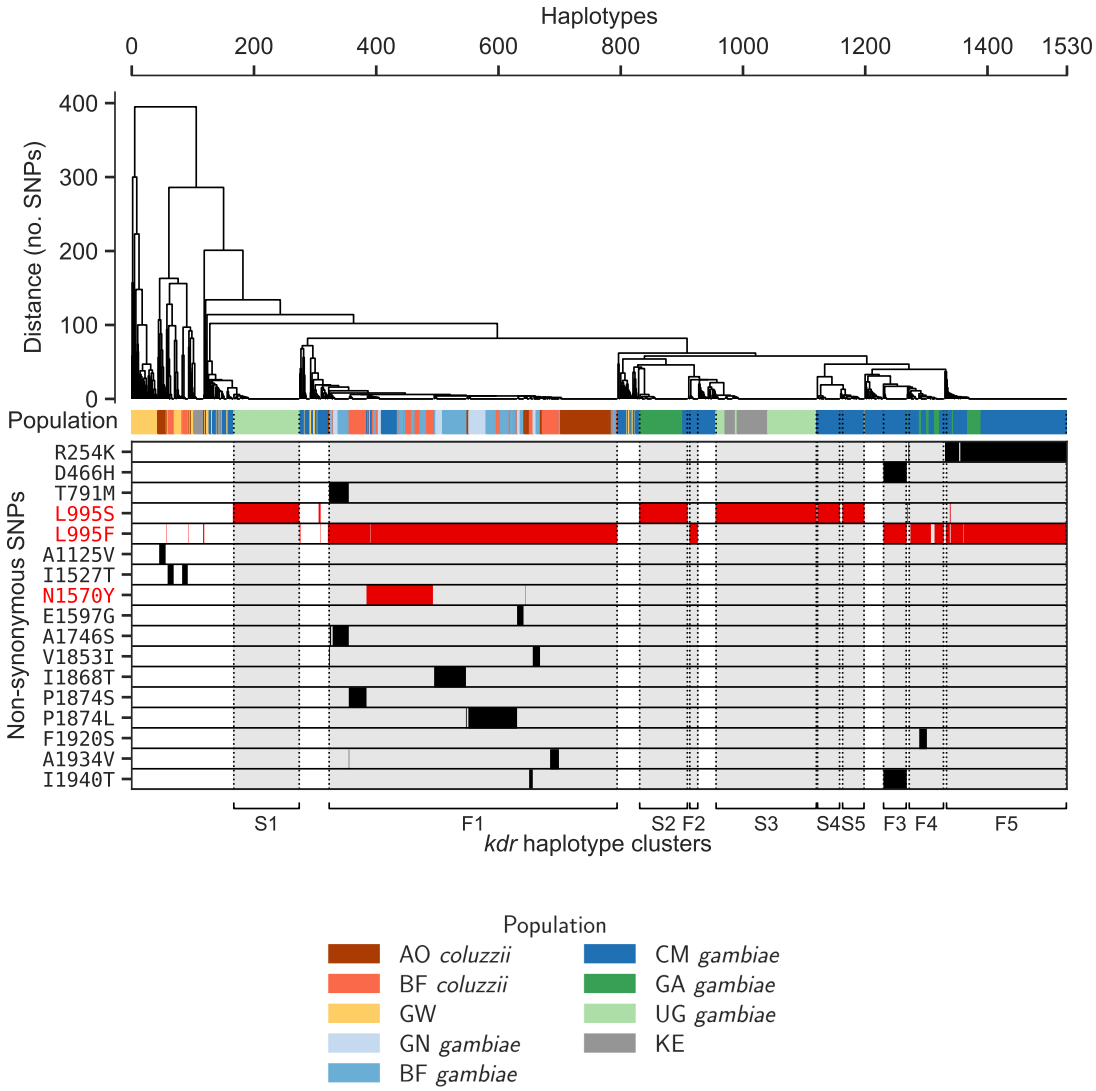

Figure 6.2. Clustering of haplotypes within the *Vgsc* gene. The upper panel shows a dendrogram obtained by hierarchical clustering of haplotypes from wild-caught individuals. The colour bar below shows the population of origin for each haplotype. The lower panel shows alleles carried by each haplotype at non-synonymous SNPs (white=reference allele, black=alternate allele, red=previously known resistance allele). At the lower margin, F1–F5 label clusters carrying the `L995F` allele, and S1–S5 label clusters carrying the `L995S` allele.

likely represents the result of contemporary gene flow driven by selection for resistance alleles in the *Vgsc* gene, I analysed patterns of haplotype sharing between populations across the whole of chromosome arm 2L (Fig. 6.7). For all pairs of populations which were both found within the same *Vgsc* haplotype cluster, I also observed a clear signal of between-population haplotype homozygosity at the *Vgsc* gene, which was not present elsewhere on the chromosome arm. Conversely, for pairs of populations never found in the

same *Vgsc* haplotype cluster, levels of between-population haplotype homozygosity were similarly low across the whole chromosome arm. Taken together, these results provide strong support for adaptive gene flow of pyrethroid resistance alleles between countries and species, in some cases separated by large geographical distances.

Positive selection for pyrethroid resistance alleles

To investigate evidence for positive selection on non-synonymous alleles, I performed an analysis of extended haplotype homozygosity (EHH) (Sabeti et al., 2002). Haplotypes under recent positive selection will have increased rapidly in frequency, thus have had less time to be broken down by recombination, and should on average have longer regions of haplotype homozygosity relative to haplotypes carrying wild-type alleles. I defined a core region spanning *Vgsc* codon 995, and an additional 6 kb of flanking sequence, which was the minimum required to find core haplotypes corresponding to the haplotype clusters F1-F5 and S1-S5 identified via the clustering analysis described above. I found 18 distinct haplotypes within this core region that were at a frequency above 1% within the cohort. These included core haplotypes corresponding to each of the 10 haplotype clusters carrying `L995F` or `L995S` alleles, as well as a core haplotype carrying `I1527T` which I labelled as L1 (due to it carrying the wild-type leucine at codon 995). I also found a core haplotype corresponding to a group of haplotypes from Kenya carrying a `M490I` substitution, which I labelled L2. All the other core haplotypes were labelled as wild-type (wt). I then computed EHH decay for each core haplotype up to 1 Mb upstream and downstream of the core locus (Fig. 6.3).

Haplotypes carrying either `L995F` or `L995S` all experienced slower decay of EHH relative to wild-type haplotypes, indicating positive selection for these resistance alleles. Previous studies have reported evidence for different rates of EHH decay between `L995F` and `L995S` haplotypes, suggesting a difference in the timing and/or strength of selection. I found no systematic difference in the extent of haplotype homozygosity when comparing F1-F5 (carrying `L995F`) against S1-S5 (carrying `L995S`) (Fig. 6.8). There were, however, some differences between core haplotypes carrying the same allele. For example, haplotypes

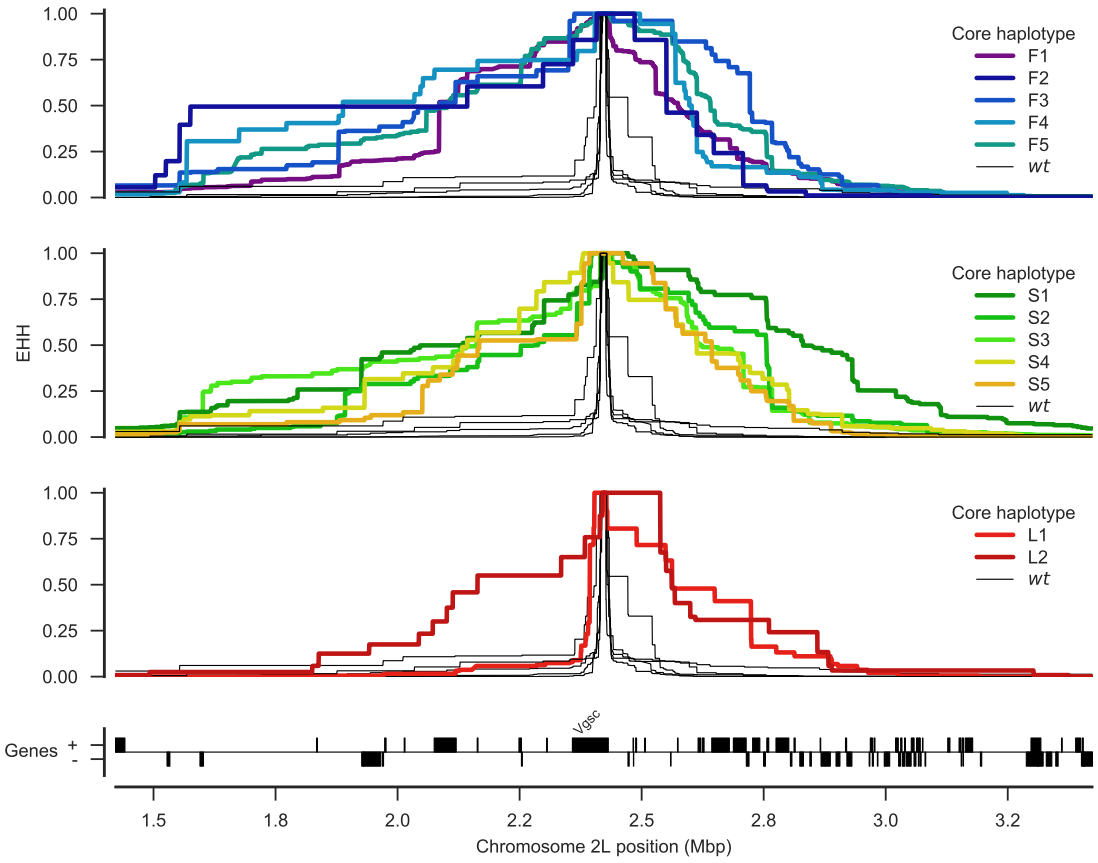

Figure 6.3. Evidence for positive selection on haplotypes carrying known or putative resistance alleles. Each panel plots the decay of extended haplotype homozygosity (EHH) for a set of core haplotypes centred on *Vgsc* codon 995. Core haplotypes F1–F5 carry the `L995F` allele, S1–S5 carry the `L995S` allele, L1 carries the `I1527T` allele, L2 carries the `M490I` allele. Wild-type (wt) haplotypes do not carry known or putative resistance alleles. A slower decay of EHH relative to wild-type haplotypes implies positive selection (each panel plots the same collection of wild-type haplotypes).

were significantly longer for S1 (median 1.091 cM, 95% bootstrap CI [1.076 - 1.091]) versus other core haplotypes carrying `L995S` (e.g., S2 median 0.699 cM, 95% bootstrap CI [0.696 - 0.705]). Longer shared haplotypes indicate a more recent common ancestor, and thus some of these core haplotypes may have experienced more recent and/or more intense selection than others.

The L1 haplotypes carrying `I1527T+V402L` exhibited a slow decay of EHH on the downstream flank of the gene, similar to haplotypes carrying `L995F` or `L995S`, indicating that this combination of alleles has experienced positive selection. EHH decay on the upstream flank was faster, similar to wild-type haplotypes, but there were two separate SNPs conferring `V402L` within this group of haplotypes, and a faster EHH decay on this

flank is consistent with recombination events bringing `V402L` alleles from different genetic backgrounds together with a haplotype carrying `I1527T`. The L2 haplotype carrying `M490I` exhibited EHH decay on both flanks comparable to haplotypes carrying known resistance alleles. This could indicate positive selection on the `M490I` allele, but these haplotypes are derived from a Kenyan mosquito population where there is evidence for a severe recent bottleneck as described in Chapter 4. There were not enough wild-type haplotypes from Kenya with which to compare, thus this signal could also be due to the extreme demographic history of this population.

Genetic surveillance of pyrethroid resistance

Entomological monitoring programmes supporting malaria vector control in Africa do in some cases genotype mosquitoes for known resistance alleles in *Vgsc* codon 995, and use those results as an indicator for the presence of pyrethroid resistance, alongside results from resistance bioassays. They typically do not, however, sequence the gene or genotype any other polymorphisms within the gene. Thus, if there are other polymorphisms within the gene that cause or enhance pyrethroid resistance, these will not be detected. Also, if a codon 995 resistance allele is observed, there is no way to know whether the allele is on a genetic background that has also been observed in other mosquito populations, and thus whether resistance alleles are emerging locally or spreading from elsewhere. Whole-genome sequencing (WGS) of individual mosquitoes clearly provides data of sufficient resolution to answer these questions, and could be used to provide ongoing resistance surveillance. The cost of WGS continues to fall and could feasibly be used for deep genomic surveillance of mosquitoes from a network of sentinel sites across multiple countries. However, to achieve higher spatial and temporal coverage of mosquito populations within countries, it would currently be necessary to also develop targeted genetic assays for resistance surveillance. Technologies such as amplicon sequencing could scale to tens of thousands of mosquitoes at low cost, and could be implemented in local laboratories.

To explore the feasibility of designing amplicon sequencing assays for tracking the spread of pyrethroid resistance, I investigated the minimum number of SNPs that would be

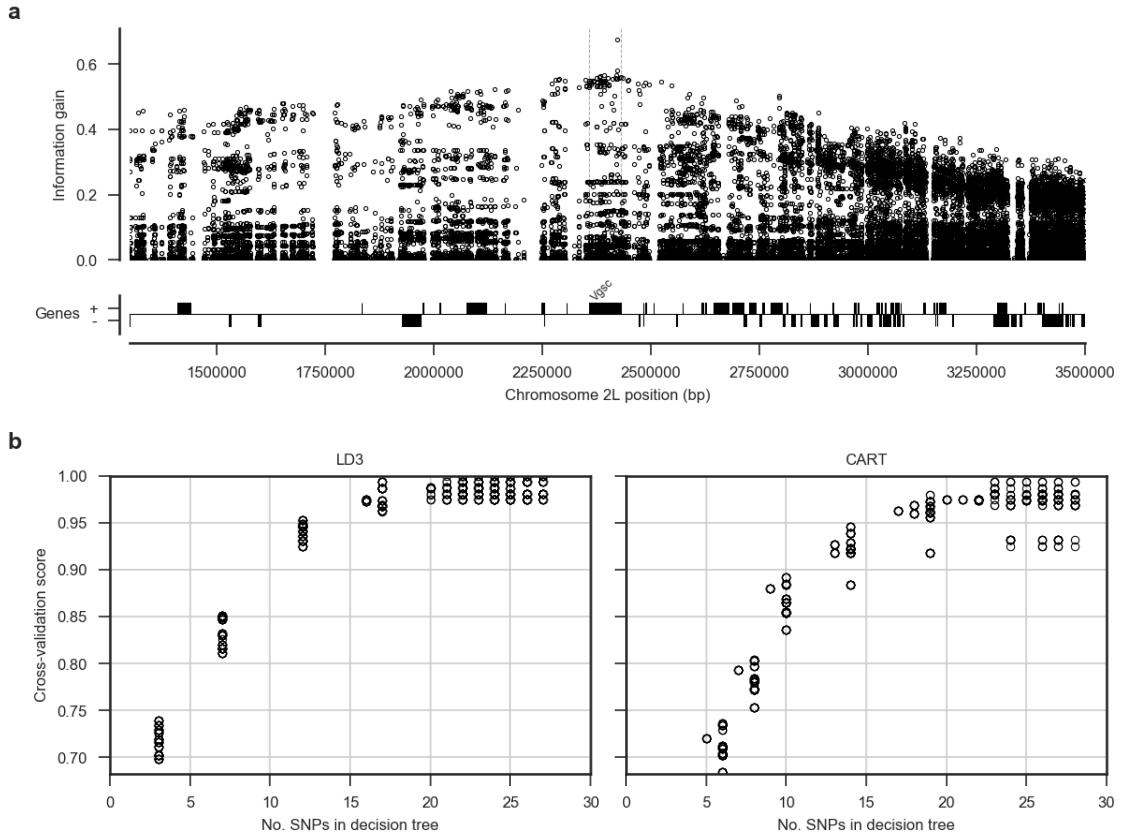

Figure 6.4. Ascertainment of informative SNPs for tracking resistance haplotypes. **a**, Each data point represents a single SNP. The information gain value for each SNP provides an indication of how informative the SNP is likely to be if used as part of a genetic assay for testing whether a mosquito carries a resistance haplotype, and if so, which haplotype group it belongs to. **b**, Number of SNPs required to accurately predict which group a resistance haplotype belongs to. Each data point represents a single decision tree. Decision trees were constructed using either the LD3 (left) or CART (right) algorithm for comparison. Accuracy was evaluated using 10-fold stratified cross-validation.

required to differentiate the different resistance haplotype clusters discovered above. I used the Ag1000G haplotype data to construct decision trees that could classify which of the haplotype clusters a given haplotype belongs to (Fig. 6.4). For each of two different tree-construction algorithms, I constructed decisions trees with a limit on the depth of the tree, repeating the process for progressively larger depths. I then scored the classification performance of each tree via 10-fold stratified cross-validation. These analyses suggested that it should be possible to classify haplotypes with $> 95\%$ accuracy using a decision tree with 20 SNPs or less. In practice, more SNPs would be needed to provide some redundancy, and also to type non-synonymous polymorphisms in addition to identifying the genetic background. However, it is still likely to be well within the number of SNPs that could be

assayed in a single amplicon sequencing multiplex. Thus, it should be feasible to produce low-cost, high-throughput genetic assays for tracking the spread of pyrethroid resistance. If combined with a limited amount of whole-genome sequencing at sentinel sites across a broad geographical range, this could also allow targeted assays to be continuously updated to track newly emerging resistance outbreaks.

Conclusions

Molecular evolution of pyrethroid target-site resistance is clearly far more complex than previously appreciated. Secondary evolution is occurring on haplotypes carrying the primary `L995F` allele, leading to double and triple mutant haplotypes. The phenotype of these combined mutants needs to be urgently characterised, because *Vgsc* combination mutants in other insect species have far higher levels of pyrethroid resistance than single mutants. The operational significance of these mutants for malaria vector control also needs to be urgently characterised, to investigate any effect on LLIN efficacy. Resistance alleles are also clearly spreading between countries, in some cases reaching countries separated by up to 4,000 km. This includes alleles spreading between countries on both sides of the equatorial rainforest, and on both sides of the East African Rift, which was surprising given the relatively high levels of genome-wide differentiation between these populations observed in Chapter 4. Thus, the *Vgsc* gene reveals how mosquito populations are connected across the continent, and how resistance alleles can spread even where rates of gene flow might otherwise appear low. In the next and final chapter I discuss the implications of these and other findings from this thesis for the future of malaria vector control, and review both the opportunities and challenges involved in developing genomic surveillance systems for malaria vectors.

Methods

Ascertainment of non-synonymous SNPs within the *Vgsc gene*

I extracted all single nucleotide polymorphisms (SNPs) within the *Vgsc* gene from the Ag1000G phase 1 data resource, and annotated each SNP according to its effect on the protein coding sequence. Within the gene, a total of 63 non-synonymous SNPs were found, of which 51 passed and 12 failed genome-wide quality filters (Chapter 3). One of the SNPs that failed quality filters was the known `N1570Y` resistance variant, but genome-wide SNP filters were generally conservative, erring on the side of minimising the false discovery rate at the expense of some sensitivity. Because any non-synonymous SNP in the *Vgsc* gene could potentially cause an important phenotypic effect, I performed a manual examination of genome accessibility metrics within all coding regions of the gene (Fig. 6.5). All the non-synonymous SNPs within the gene that failed quality filters were close to the thresholds set, and there was no evidence of structural variation, alignment ambiguity or other accessibility issues in any of the gene exons. I therefore included all 63 non-synonymous SNPs in subsequent analyses.

The Vgsc gene is also known to undergo splicing variation in multiple insect species including *An. gambiae*, and this is believed to be an important factor enabling insects to achieve phenotypic variation in different tissues (Dong et al., 2014; Davies et al., 2007). For *An. gambiae*, at the time this analysis was performed, three transcripts were present in the AgamP4.4 gene annotations provided by VectorBase, which derived primarily from automated gene predictions (Curwen et al., 2004). However, the canonical study of the *Vgsc* gene in *An. gambiae* is Davies et al. (2007), where substantially greater splice variation was reported. A difficulty of the data from Davies et al. (2007) is that there are no complete transcript sequences, only separate cDNA sequences from the first and second halves of the transcript. To allow for all possibilities, I constructed a set of transcripts from all possible combinations of the first and second half cDNA sequences (Fig. 6.6). Some of these transcripts included optional exons or exon regions not represented in the VectorBase transcripts. I then reran SNP effect predictions for all SNPs discovered within the Vgsc

gene with each of these alternative transcripts, to determine whether any non-synonymous SNPs had occurred within an exon or exon segment not present in the canonical AgamP4.4 `AGAP004707-RA` transcript. Three additional non-synonymous variants were found in these other transcripts, one in each of exons 2j, 3 and 20d, but none of these occurred at above 3% allele frequency in any population, and so were not analysed further.

Additional phasing

The haplotype data from Ag1000G phase 1 only include biallelic SNPs passing all quality filters, and thus did not include `N1570Y` due to failing genome-wide filters, nor the multiallelic SNPs at positions 2L:2,391,228 and 2L:2,400,071. To incorporate these additional SNPs into the analysis, I phased these SNPs onto the Ag1000G haplotype scaffold using MVNcall version 1.0.

Haplotype clustering and gene flow analyses

Haplotypes spanning the *Vgsc* gene (`AGAP004707`, 2L:2,358,158–2,431,617) were extracted from the genome-wide dataset. I computed the Hamming distance between all pairs of haplotypes, then performed hierarchical clustering of haplotypes and visualised the results as a dendrogram. These analyses were performed using SciPy version 0.16.1 (Virtanen et al., 2020). To identify haplotype clusters, I cut the dendrogram at a distance of 12 SNPs and studied the largest clusters carrying `L1014F/S` resistance alleles. A non-zero cutting distance allows for a small number of SNP differences within each cluster, which is expected as new mutations begin to occur on haplotypes that are increasing in frequency under positive selection. I used the complete linkage clustering method to generate the dendrogram and clusters reported here, but I repeated the analysis using the single and average linkage methods, finding highly concordant results.

To investigate whether recombination events within the *Vgsc* gene might have affected the clustering of haplotypes carrying resistance alleles, I performed an additional analysis using data from regions flanking the gene. I repeated the clustering analysis separately on non-overlapping windows upstream and downstream of the *Vgsc* gene, using the same

window size (1,718 SNPs) as used for the clustering analysis within the gene. I then compared the resulting haplotype clusters by computing the set intersection between all pairs of clusters across all pairs of windows. If recombination events happened, this would affect the haplotype clustering in some windows but not others, causing some discordance between clustering in adjacent windows. In the window immediately upstream of the gene I found clusters with at least 93% intersection with the clusters identified within the gene, except for clusters S4 and S5 which merged into a single cluster. In the window immediately downstream of the gene I found clusters with at least 93% intersection, again except for S4 and S5 which merged into a single cluster, and except for F1 which split into a major cluster carrying 402/465 (86%) of the F1 cluster and a minor cluster carrying 60/465 haplotypes. The most parsimonious explanation for the S4/S5 clustering together on both flanks of the gene is that these are both derived from the same haplotype, but experienced a gene conversion or sequence of crossover events within the *Vgsc* gene to bring some portion of a different haplotype into the original genetic background. The region of differentiation between S4 and S5 is limited to a 13 kb region upstream of codon 995, and these two clusters share haplotype homozygosity spanning codon 995.

To confirm evidence of resistance allele gene flow between populations, I performed a between-population analysis of haplotype homozygosity, implemented via a custom Python script. Briefly, this analysis extracts haplotypes from each of two populations in moving windows of 2000 SNPs across chromosome arm 2L, and computes the fraction of haplotype pairs that are identical (haplotype homozygosity), considering only pairs with one haplotype from each population. A value of 0 indicates no haplotype pairs are identical between populations, and a value of 1 indicates all haplotype pairs are identical between populations.

Positive selection

Core haplotypes were defined on a 6,078 bp region spanning *Vgsc* codon 995, from chromosome arm 2L position 2,420,443 and ending at position 2,426,521. This region was chosen as it was the smallest region sufficient to differentiate between the ten haplotype clusters

carrying either of the known resistance alleles L1014F or L1014S. Extended haplotype homozygosity (EHH) was computed for all core haplotypes as described in Sabeti et al. (2002) using scikit-allel version 1.1.9, excluding non-synonymous and singleton SNPs.

Decision tree analyses

To explore the feasibility of identifying a small subset of SNPs that would be sufficient to identify each of the genetic backgrounds carrying known or putative resistance alleles, I started with an input data set of all SNPs within the *Vgsc* gene or in the flanking regions 20 kbp upstream and downstream of the gene. Each of the 1530 haplotypes in the Ag1000G phase 1 cohort was labelled according to which core haplotype it carried, combining all core haplotypes not carrying known or putative resistance alleles together as a single "wild-type" group. Decision tree classifiers were then constructed using scikit-learn version 0.19.0 for a range of maximum depths, repeating the tree construction process 10 times for each maximum depth with a different initial random state. The classification accuracy of each tree was evaluated using stratified 10-fold cross-validation.

Supplemental figures

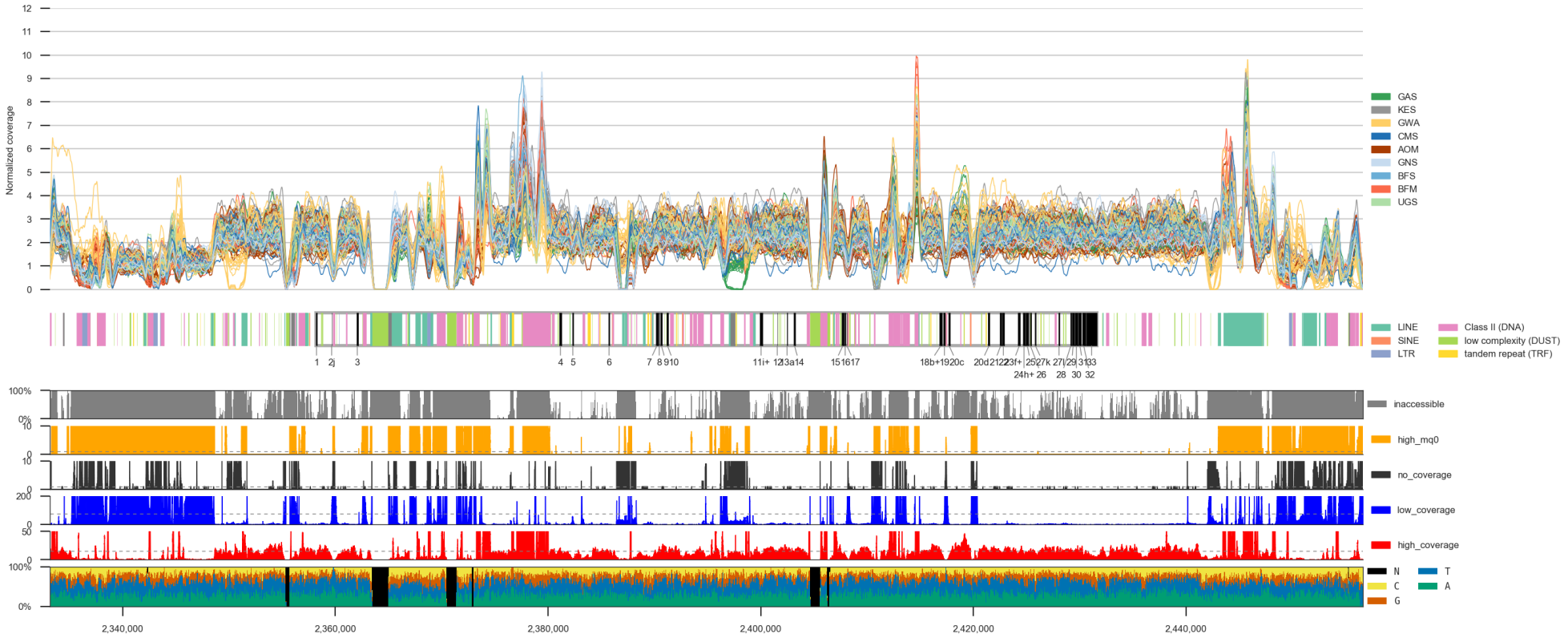

Figure 6.5. Accessibility of the *Vgsc* gene. Upper plot shows coverage traces for all individuals in the Ag1000G phase 1 cohort, coloured by population, normalised by genome-wide coverage. Schematic below shows locations of *Vgsc* gene exons (black) and different classes of repetitive or low complexity sequences. Tracks at the bottom show plots of various genome accessibility metrics.

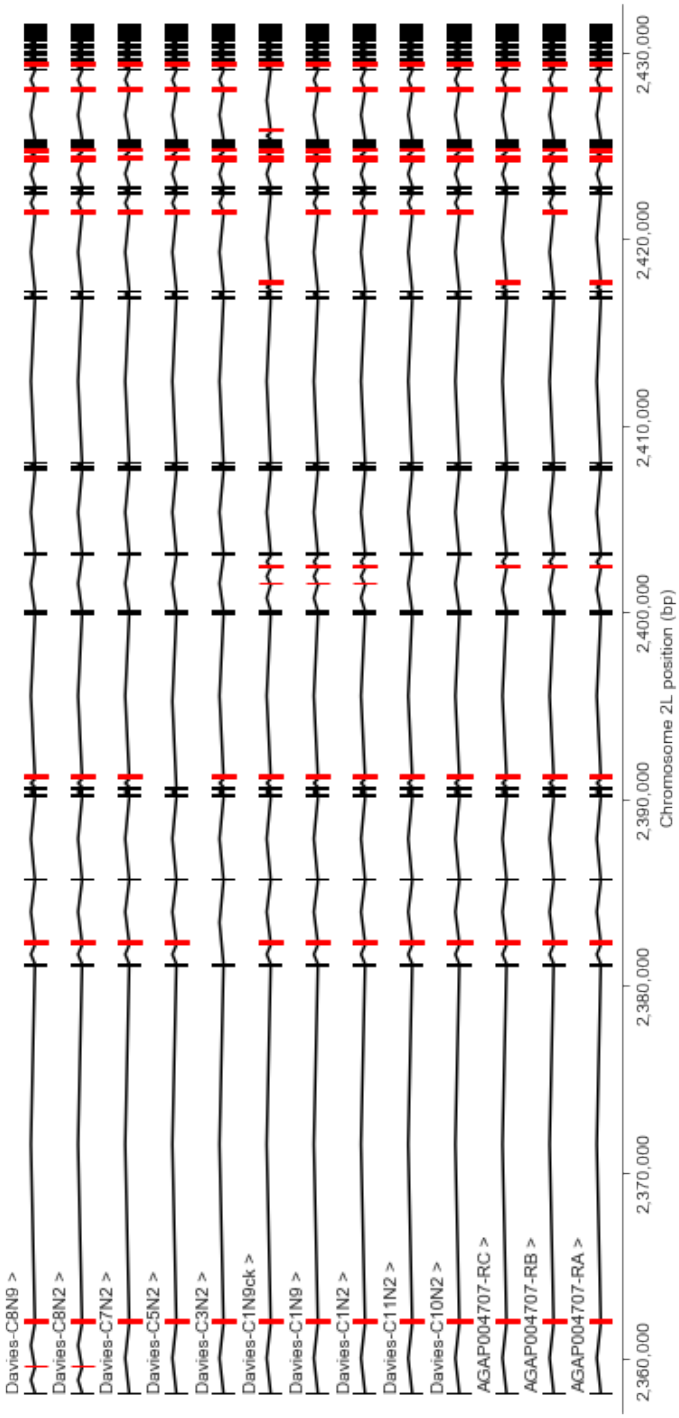

Figure 6.6. Alternative transcripts for the *Vgsc* gene, inferred from cDNA sequences reported by Davies et al. (2007).

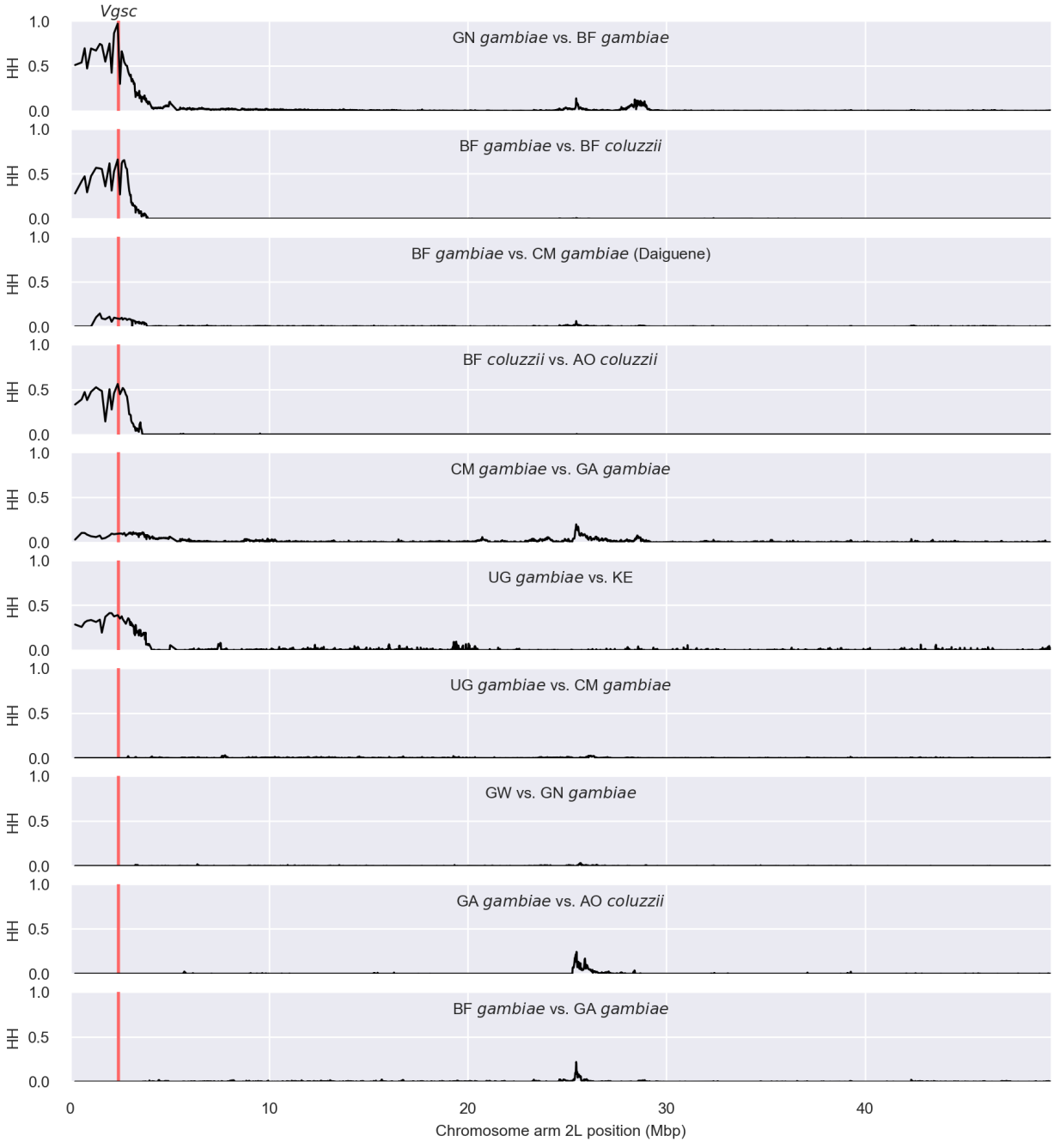

Figure 6.7. Chromosome-wide scans for between-population haplotype homozygosity (HH). Each track plots the fraction of between-population haplotype pairs that are identical within moving windows of 2,000 SNPs. A value of 1 indicates all haplotype pairs are identical. A value of 0 indicates no haplotype pairs are identical.

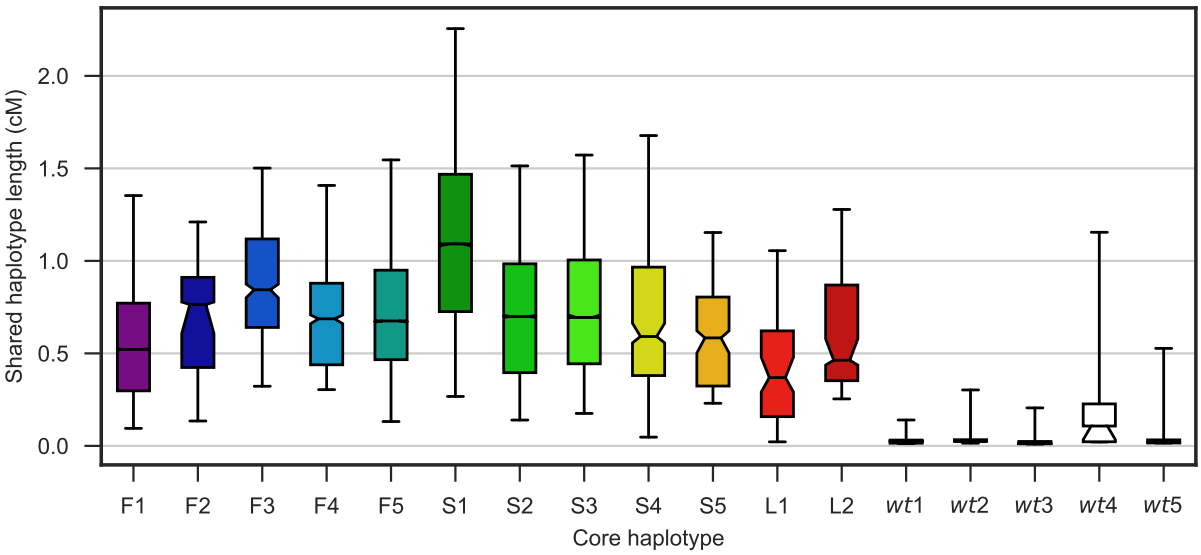

Figure 6.8. Analysis of shared haplotype lengths. Each bar shows the distribution of shared haplotype lengths between all pairs of haplotypes with the same core haplotype. For each pair of haplotypes, the shared haplotype length was computed as the region extending upstream and downstream from the core locus (*Vgsc* codon 995) over which haplotypes are identical at all non-singleton variants. The *Vgsc* gene sits on the border of pericentromeric heterochromatin and euchromatin, and I assumed different recombination rates in upstream and downstream regions. The shared haplotype length is expressed in centiMorgans (cM) assuming a constant recombination rate of 2.0 cM/Mb on the downstream (euchromatin) flank and 0.6 cM/Mb on the upstream (heterochromatin) flank. Bars show the inter-quartile range, fliers show the $5 - 95^{th}$ percentiles, horizontal black line shows the median, notch in bar shows the 95% bootstrap confidence interval for the median. Haplotypes F1–5 each carry the `L995F` resistance allele. Haplotypes S1–5 each carry the `L995S` resistance allele. Haplotype L1 carries the `I1527T` allele. Haplotype L2 carries the `M490I` allele. Wild-type (wt) haplotypes do not carry any known or putative resistance alleles.

References

AMP (2020). *Net Mapping Project, Current ITN Global Delivery Quarterly Report, Q2 2020*. Tech. rep. Alliance for Malaria Prevention.

Auteri, M, F La Russa, V Blanda and A Torina (2018). 'Insecticide Resistance Associated with kdr Mutations in Aedes albopictus: An Update on Worldwide Evidences'. In: *Biomed Res. Int.* 2018, pp. 1–10. DOI: 10.1155/2018/3098575.

Bhatt, S et al. (2015). 'The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015'. In: *Nature* 526.7572, pp. 207–211. DOI: 10.1038/nature15535.

Carnevale, P and F Gay (2019). 'Insecticide-treated mosquito nets'. In: *Malaria Control and Elimination*. Ed. by F Ariey, F Gay and R Ménard. New York: Springer, pp. 221–232. DOI: 10.1007/978-1-4939-9550-9_16.

Chandre, F, F Darrier, L Manga, M Akogbeto, O Faye, J Mouchet and P Guillet (1999). 'Status of pyrethroid resistance in Anopheles gambiae sensu lato'. In: *Bull. World Health Organ.* 77.3, pp. 230–234.

Clarkson, CS et al. (2014). 'Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation.' In: *Nat. Commun.* 5.1, p. 4248. DOI: 10.1038/ncomms5248.

Collins, E, NM Vaselli, M Sylla, AH Beavogui, J Orsborne, G Lawrence, RE Wiegand, SR Irish, T Walker and LA Messenger (2019). 'The relationship between insecticide resistance, mosquito age and malaria prevalence in Anopheles gambiae s.l. from Guinea'. In: *Sci. Rep.* 9.1. DOI: 10.1038/s41598-019-45261-5.

Curwen, V, E Eyras, TD Andrews, L Clarke, E Mongin, SM Searle and M Clamp (2004). 'The Ensembl automatic gene annotation system'. In: *Genome Res.* 14.5, pp. 942–950. DOI: 10.1101/gr.1858004.

Davies, T, LM Field, P Usherwood and MS Williamson (2007). 'A comparative study of voltage-gated sodium channels in the Insecta: Implications for pyrethroid resistance in Anopheline and other Neopteran species'. In: *Insect Mol. Biol.* 16.3, pp. 361–375. DOI: 10.1111/j.1365-2583.2007.00733.x.

Djouaka, R, I Djègbè, R Akoton, G Tchigossou, KM Ahadji-Dabla, SM Atoyebi, R Adéoti, F Zeukeng and GK Ketoh (2018). 'First report of the presence of L1014S Knockdown-resistance mutation in Anopheles gambiae s.s and Anopheles coluzzii from Togo, West Africa'. In: *Wellcome Open Res.* 3, p. 30. DOI: 10.12688/wellcomeopenres.13888.1.

Dong, K, Y Du, F Rinkevich, Y Nomura, P Xu, L Wang, K Silver and BS Zhorov (2014). 'Molecular biology of insect sodium channels and pyrethroid resistance.' In: *Insect Biochem. Mol. Biol.* 50, pp. 1–17. DOI: 10.1016/j.ibmb.2014.03.012.

Du, Y, Y Nomura, G Satar, Z Hu, R Nauen, SY He, BS Zhorov and K Dong (2013). 'Molecular evidence for dual pyrethroid-receptor sites on a mosquito sodium channel'. In: *Proc. Natl. Acad. Sci. U. S. A.* 110.29, pp. 11785–11790. DOI: 10.1073/pnas.1305118110.

Du, Y, Y Nomura, BS Zhorov and K Dong (2016). 'Sodium channel mutations and pyrethroid resistance in Aedes aegypti'. In: *Insects* 7.4, p. 60. DOI: 10.3390/insects7040060.

Elissa, N, J Mouchet, F Riviere, JY Meunier and K Yao (1993). 'Resistance of Anopheles gambiae s.s. to pyrethroids in Cote d'Ivoire'. In: *Ann. Soc. belge Med. trop.* 73, pp. 291–294.

Elliott, M, AW Farnham, NF Janes, PH Needham and DA Pulman (1974). 'Synthetic insecticide with a new order of activity'. In: *Nature* 248.5450, pp. 710–711. DOI: 10.1038/248710a0.

Elliott, M, AW Farnham, NF Janes, PH Needham, DA Pulman and JH Stevenson (1973). 'A Photostable Pyrethroid'. In: *Nature* 246.5429, pp. 169–170. DOI: 10.1038/246169a0.

Elliott, M (1989). 'The pyrethroids: Early discovery, recent advances and the future'. In: *Pestic. Sci.* 27.4, pp. 337–351. DOI: 10.1002/ps.2780270403.

Haddi, K, HV Tomé, Y Du, WR Valbon, Y Nomura, GF Martins, K Dong and EE Oliveira (2017). 'Detection of a new pyrethroid resistance mutation (V410L) in the sodium channel of Aedes aegypti: A potential challenge for mosquito control'. In: *Sci. Rep.* 7.1, pp. 1–9. DOI: 10.1038/srep46549.

Hancock, PA, CJ Hendriks, JA Tangena, H Gibson, J Hemingway, M Coleman, PW Gething, E Cameron, S Bhatt and CL Moyes (2020). 'Mapping trends in insecticide resistance phenotypes in African malaria vectors'. In: *PLoS Biol.* 18.6. Ed. by AF Read, e3000633. DOI: 10.1371/journal.pbio.3000633.

Hemingway, J et al. (2016). 'Averting a malaria disaster: Will insecticide resistance derail malaria control?' In: *The Lancet* 387.10029, pp. 1785–1788. DOI: 10.1016/S0140-6736(15)00417-1.

Implications of Insecticide Resistance Consortium (2018). 'Implications of insecticide resistance for malaria vector control with long-lasting insecticidal nets: Trends in pyrethroid resistance during a WHO-coordinated multi-country prospective study'. In: *Parasites and Vectors* 11.1, p. 550. DOI: 10.1186/s13071-018-3101-4.

Jones, CM, M Liyanapathirana, FR Agossa, D Weetman, H Ranson, MJ Donnelly and CS Wilding (2012). 'Footprints of positive selection associated with a mutation (N1575Y) in the voltage-gated sodium channel of Anopheles gambiae.' In: *Proc. Natl. Acad. Sci. U. S. A.* 109.17, pp. 6614–6619. DOI: 10.1073/pnas.1201475109.

Lewontin, RC (1964). 'The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models.' In: *Genetics* 49, pp. 49–67.

Martinez-Torres, D, F Chandre, MS Williamson, F Darriet, JB Bergé, AL Devonshire, P Guillet, N Pasteur and D Pauron (1998). 'Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector Anopheles gambiae s.s.' In: *Insect Mol. Biol.* 7.2, pp. 179–84.

Norris, LC, BJ Main, Y Lee, TC Collier, A Fofana, AJ Cornel and GC Lanzaro (2015). 'Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets'. In: *Proc. Natl. Acad. Sci.* 112.3, pp. 815–820. DOI: 10.1073/pnas.1418892112.

Pinto, J et al. (2007). 'Multiple origins of knockdown resistance mutations in the afrotropical mosquito vector Anopheles gambiae'. In: *PLoS One* 2.11. Ed. by N Ahmed, p. 1243. DOI: 10.1371/journal.pone.0001243.

Quélennec, G (1988). 'Pyrethroids in the WHO Pesticide Evaluation scheme (WHOPES)'. In: *Parasitol. Today* 4.7, S15–S17. DOI: 10.1016/0169-4758(88)90082-8.

Ranson, H, B Jensen, JM Vulule, X Wang, J Hemingway and FH Collins (2000). 'Identification of a point mutation in the voltage-gated sodium channel gene of Kenyan Anopheles gambiae associated with resistance to DDT and pyrethroids.' In: *Insect Mol. Biol.* 9.5, pp. 491–7. DOI: 10.1046/j.1365-2583.2000.00209.x.

Ranson, H, R N'Guessan, J Lines, N Moiroux, Z Nkuni and V Corbel (2011). 'Pyrethroid resistance in African anopheline mosquitoes: What are the implications for malaria control?' In: *Trends Parasitol.* 27.2, pp. 91–98. DOI: 10.1016/j.pt.2010.08.004.

Sabeti, PC et al. (2002). 'Detecting recent positive selection in the human genome from haplotype structure.' In: *Nature* 419.6909, pp. 832–837. DOI: 10.1038/nature01140.

Sonoda, S (2010). 'Molecular analysis of pyrethroid resistance conferred by target insensitivity and increased metabolic detoxification in Plutella xylostella'. In: *Pest Manag. Sci.* 66.5, pp. 572–575. DOI: 10.1002/ps.1918.

Tan, J, Z Liu, TD Tsai, SM Valles, AL Goldin and K Dong (2002). 'Novel sodium channel gene mutations in Blattella germanica reduce the sensitivity of expressed channels to deltamethrin.' In: *Insect Biochem. Mol. Biol.* 32.4, pp. 445–454. DOI: 10.1016/s0965-1748(01)00122-9.

Villanueva-Segura, OK, KA Ontiveros-Zapata, B Lopez-Monroy, G Ponce-Garcia, SM Gutierrez-Rodriguez, JA Davila-Barboza, EDJ Mora-Jasso, AE Flores and D Severson (2019). 'Distribution and Frequency of the kdr Mutation V410L in Natural Populations of Aedes aegypti (L.) (Diptera: Culicidae) from Eastern and Southern Mexico'. In: *J. Med. Entomol.* 57.1. Ed. by D Severson, pp. 218–223. DOI: 10.1093/jme/tjz148.

Virtanen, P et al. (2020). 'SciPy 1.0: fundamental algorithms for scientific computing in Python'. In: *Nat. Methods* 17.3, pp. 261–272. DOI: 10.1038/s41592-019-0686-2. eprint: 1907.10121.

Vulule, JM, RF Beach, FK Atieli, JM Roberts, DL Mount and RW Mwangi (1994). 'Reduced susceptibility of Anopheles gambiae to permethrin associated with the use of permethrin-impregnated bednets and curtains in Kenya'. In: *Med. Vet. Entomol.* 8.1, pp. 71–75. DOI: 10.1111/j.1365-2915.1994.tb00389.x.

Wang, L, Y Nomura, Y Du, N Liu, BS Zhorov and K Dong (2015). 'A mutation in the intracellular loop III/IV of mosquito sodium channel synergizes the effect of mutations in helix IIS6 on pyrethroid resistance'. In: *Mol. Pharmacol.* 87.3, pp. 421–429. DOI: 10.1124/mol.114.094730.

WHO (2012). *Global plan for insecticide resistance management in malaria vectors*. Tech. rep. World Health Organization.

WHO (2017). *Framework for a national plan for monitoring and management of insecticide resistance in malaria vectors*. Tech. rep. World Health Organization.

WHO (2018a). *Global report on insecticide resistance in malaria vectors: 2010–2016*. Tech. rep. World Health Organization.

WHO (2018b). *Test procedures for insecticide resistance monitoring in malaria vector mosquitoes (Second edition)*. Tech. rep. World Health Organization.

7 Discussion: Towards genomic surveillance systems for malaria vectors

Genomic surveillance in the time of a pandemic

As of Saturday 19 December 2020, the county of Berkshire in which I live has been placed under Tier 4 COVID-19 restrictions, alongside counties across the South and East of England. These new restrictions have been imposed in an effort to limit the spread of the newly discovered SARS-CoV-2 lineage B.1.1.7, which carries an unusual number of mutations in the spike protein, and appears to be out-competing other virus strains because of increased transmissibility (Rambaut et al., 2020; Davies et al., 2020). International borders have closed, people have been cut off from friends and family days before Christmas, and further restrictions seem inevitable given rising cases and hospital admissions in many areas. At the same time, another unusual SARS-CoV-2 lineage 501Y.V2 has been discovered in South Africa, which is phylogenetically unrelated to B.1.1.7 but shares a common `N501Y` mutation in the receptor binding domain of the spike protein, and also carries an unusually high number of other mutations within the spike protein (Tegally et al., 2020). The full implications of these discoveries have yet to unfold, but they have demonstrated beyond question the value of genomic surveillance systems, which were established in the UK and South Africa early in the COVID-19 pandemic (The COVID-19 Genomics UK (COG-UK) consortium, 2020a; Msomi et al., 2020). These surveillance systems have been regularly sequencing a subset of viral samples from infected individuals, sharing data openly with scientists both nationally and internationally. The UK currently leads the world in the scale of its genomic surveillance operation, having sequenced 137,540 SARS-CoV-2 genomes

up to 12th December (The COVID-19 Genomics UK (COG-UK) consortium, 2020b).

Although the SARS-CoV-2 virus and *Anopheles* mosquitoes are fundamentally different forms of life, they share in common the fact that they are both experiencing new selective forces, which are driving their evolution in a way that has major consequences for human health. In the case of SARS-CoV-2, passage from an animal reservoir to human hosts has created a selective pressure to adapt infection pathways and immune evasion mechanisms to their new host's biology. In the case of malaria vectors, our efforts to eradicate malaria in sub-Saharan Africa through large-scale vector control programmes have created a strong selective pressure to evolve insecticide resistance. However, unlike SARS-CoV-2, we do not have genomic surveillance systems in place for malaria vectors. This means that, as new forms of insecticide resistance emerge and spread within malaria vector populations, we have no means to detect these events, nor to design and coordinate any kind of effective response or mitigation.

Next-generation malaria vector control needs next-generation surveillance

Sequencing the genomes of *Anopheles* mosquitoes from the field and accurately detecting genetic variation presents a number of challenges. In contrast to pathogens like SARS-CoV-2, the *Anopheles* genome is orders of magnitude larger and has greater complexity. In contrast to human populations where large-scale genome sequencing efforts have also been undertaken, *Anopheles* mosquito populations are an order of magnitude more genetically diverse. In this thesis I have shown that these challenges can be addressed, and that it is possible to robustly call nucleotide variation using Illumina whole-genome sequencing of *An. gambiae* and *An. coluzzii* mosquitoes collected from natural populations, discovering more than 52 million single nucleotide polymorphisms in these species. I have used these data to confirm that malaria vectors are among the most genetically diverse species on Earth. Much of this variation is shared between species and geographically distant populations, but there is also population structure, with both the equatorial rainforest and the East

African Rift appearing to play a major role in partitioning populations within both species. There are strong signals of recent selection in most of the mosquito populations sampled, affecting multiple insecticide resistance genes, and leading us towards new loci that may harbour novel forms of insecticide resistance or other adaptations to malaria vector control interventions. In spite of population structure, alleles conferring resistance to pyrethroid insecticides have found a way to spread between species and across large geographical distances, revealing hidden connections between mosquito populations, and demonstrating that managing insecticide resistance cannot be a local concern but will require international coordination.

The work described in this thesis was carried out in the context of the *Anopheles gambiae* 1000 Genomes Project, which is laying the foundations for a new genomic approach to malaria vector surveillance. It comes at a time when malaria vector control undergoing its most significant change in 20 years. In an attempt to mitigate the impact of pyrethroid resistance, malaria control programmes have begun using and rotating "next-generation" indoor residual spraying (IRS) formulations, which use the organophosphate pyrimiphos methyl and the neonicotinoid clothianidin (Oxborough et al., 2014; Oxborough et al., 2019; WHO, 2019). "Next-generation" long-lasting insecticidal bednets (LLINs) incorporating the synergist piperonyl butoxide (PBO) have been shown to be effective, and large scale procurements are planned in multiple countries (Protopopoff et al., 2018; Staedke et al., 2020). Other next-generation LLIN products combining a pyrethroid with a second insecticide are also approved and will surely be deployed at scale (Bayili et al., 2017; Tiono et al., 2018). Malaria vectors are thus beginning to experience a variety of new selective pressures as they encounter these new insecticides and synergists. We know from experiences of the first global malaria eradication campaign of the 1950s, and of the Roll Back Malaria campaign of the last two decades, that the clock is now ticking, and an evolutionary response will follow (Elliott and Ramakrishna, 1956; Hancock et al., 2020). Now, more than ever, we need a next-generation of surveillance systems for malaria vectors, so that we can observe new evolutionary events as they occur, and can change course early, rather than waiting until resistance is entrenched, and the efficacy of these new control

tools is irreversibly undermined.

A roadmap for malaria vector genomic surveillance systems

What, then, is the roadmap for translation of genome sequencing technologies into operational malaria vector surveillance systems? A comprehensive answer to that question is beyond the scope of this thesis, but I would like to highlight the following necessary elements.

Expanded genome variation data resources

First and foremost, we need to expand our knowledge of genetic variation within natural populations of all the major malaria vector species in sub-Saharan Africa. The first phase of the Ag1000G project sequenced *An. gambiae* and *An. coluzzii* mosquitoes from eight countries, but representation of *An. coluzzii* was limited to only two countries, and there was no representation of *An. arabiensis*, the third major vector species in the *Anopheles gambiae* complex. The second phase of the Ag1000G project is now nearing completion, and has increased sampling to 1,142 mosquitoes from 13 countries, including *An. coluzzii* from five countries (The Anopheles gambiae 1000 Genomes Consortium, 2020). The third phase of the Ag1000G project is in progress, and will shortly release nucleotide variation data for 2,784 wild-caught mosquitoes from 18 countries, including *An. arabiensis* from four countries. Beyond the *Anopheles gambiae* complex, *An. funestus* is also a major malaria vector in many parts of Africa. Studies have begun surveying natural genetic variation in this species, and have found several novel insecticide resistance adaptations that are increasing in frequency and spreading geographically (Weedall et al., 2020). A new project has been established within MalariaGEN to survey genetic variation in *An. funestus* and is sequencing mosquitoes from a broad geographical range, expecting to release data in the next year.

In this thesis I have investigated single nucleotide polymorphisms, but we also need to expand our data resources to include other types of genetic variation. In particular, copy number variation (CNV) has long been suspected to play a major role in the evolution of

insecticide resistance (Devonshire and Field, 1991; Hemingway et al., 1998; Bass and Field, 2011). In the second phase of the Ag1000G project we have expanded our analysis to include CNVs, finding CNV hotspots at several loci containing genes involved in metabolic resistance to insecticides (Lucas et al., 2019). Short insertion/deletion (indel) polymorphisms are also likely to be abundant in *Anopheles* genomes and could have important functional consequences (Montgomery et al., 2013). Indels could also be valuable markers of recent evolutionary and demographic events, given their higher mutation rates (Redmond et al., 2018). Unfortunately, indels remain difficult to discover and genotype with accuracy, and further work is needed to improve indel calling and filtering methods for *Anopheles*.

Contemporary time series

Second, we need to bring these genomic data resources up to date, by sequencing mosquitoes collected within the last year, and by establishing partnerships in order to regularly collect and sequence mosquitoes from sentinel sites. By establishing a time series of genomic data from multiple locations, we would gain the ability to observe significant changes as they occur, and to provide early warning of new evolutionary events of public health relevance, such as the emergence of a new form of insecticide resistance. We would also be able to learn more about the dynamics of malaria vector populations, including both annual and seasonal fluctuations in population size due to natural environmental factors, as well as the demographic impact of vector control interventions. Such data could help us learn more about which vector control interventions are more effective, and work towards better estimates of contemporary effective population size (Hui and Burt, 2015). It could also help to resolve fundamental behavioural questions such as whether some mosquitoes undergo aestivation (Dao et al., 2014) or intentional wind-assisted long-distance migration (Huestis et al., 2019). These questions are not only of academic interest, but are central to the planning of future vector control programmes using gene drives (North et al., 2019).

Optimised and standardised protocols for a faster response

Third, we need to substantially reduce the time taken to go from mosquito collection to genomic insights, so that information can be delivered in a timely manner. Returning to the COVID-19 analogy, the preliminary report on the B.1.1.7 lineage was published on 18th December 2020, and draws on sequence data from samples collected up to 30th November (Rambaut et al., 2020). In comparison, if a new form of insecticide resistance was spreading in malaria vector populations, we would not currently be able to match anything like this turnaround time of less than three weeks from samples collection to analysed sequence data. There are bottlenecks at all stages of the process, including study approval, sample collection, sample shipping, DNA extraction, library preparation, whole-genome sequencing, variant calling, data curation, and analysis. In particular, many aspects of population-genomic data analysis remain something of an art for non-model species like *Anopheles* mosquitoes that are sexually recombining with large and diverse genomes, requiring specialised training and time to set up and perform correctly. We need to establish standardised analytical protocols and well-engineered supporting software tools for a core suite of malaria vector genomic surveillance analyses, such as scans for genes under selection, or analyses of insecticide resistance outbreaks. The work I have done to develop robust and performant analytical software packages like scikit-allel[1] for the scientific Python ecosystem is a small step in this direction, but there is much more to do.

Decentralised sequencing and analytical capacity

Fourth, the need for increased geographical and temporal coverage, and for faster turnaround times from samples to data, means that sequencing and analytical capabilities need to be decentralised and developed within multiple public health laboratories and institutions, particularly those on the African continent. This need is now widely recognised, and efforts to develop these capabilities have been accelerated by the COVID-19 pandemic, which will hopefully benefit other diseases as well. In October 2020, The Africa Centres for Disease Control and Prevention (CDC) received a $100 million investment for pathogen

[1] `https://github.com/cggh/scikit-allel`

genomics research and development, through the Africa Pathogen Genomics Initiative (PGI) (Makoni, 2020). This initiative will support the development of next-generation sequencing capability in multiple public health laboratories, and will hopefully trigger further investments from both domestic and international funders.

A key question is whether pathogen genomics initiatives can be expanded to include capacity for malaria vector sequencing. A particular challenge for malaria vectors relative to pathogens is the size of the genome. Whole-genome sequencing (WGS) of malaria vectors has the potential to reveal a wealth of new insights, as demonstrated by this thesis, and could serve as a valuable surveillance tool because genes and variants of public health relevance do not need to be known *a priori*. However, performing WGS at high throughput and low cost requires additional up-front investment in infrastructure and personnel, which may be hard to achieve in the near term outside of specialised sequencing centres. Targeted sequencing of amplicons in known insecticide resistance genes could offer a viable alternative, which is more amenable to implementation within a decentralised lab network. I have shown that tracking of important insecticide resistance variants can be done with a relatively small number of markers, demonstrating the potential value of this approach. Some WGS will still be required, however, to discover new genes and variants of concern, and the challenge will be to design, validate and deploy new assays with sufficient speed to track new events.

In addition to lab capacity, bioinformatics and data analysis capacity is also a critical resource. Initiatives such as H3ABioNet (Kumuthini et al., 2019), providing general bioinformatics support and capacity development for genomics, are essential. These need to be expanded to include analytical capacity for genomic epidemiology of disease pathogens and vectors.

Data sharing and cooperation

Fifth, as multiple sequencing centres come online and begin generating genomic surveillance data, networks will be needed to coordinate these efforts and to integrate the data in order to provide a coherent transnational view. The Ag1000G project was carried out

within the context of MalariaGEN, a data-sharing network intended to provide an equitable framework for multi-institution collaboration on large-scale genomic epidemiology studies. MalariaGEN, H3Africa and others have demonstrated the potential for international cooperation in genomics (Mulder et al., 2018). A key challenge when establishing genomic surveillance networks is to find a balance between the needs of surveillance and academic research. Effective surveillance requires data to be shared quickly and openly in order to coordinate a rapid response to emerging threats. Academic research also benefits from open data, but requires attribution and protection for those who generate data and wish to publish deeper analyses in peer-reviewed journals. In emerging situations, the line between surveillance and research can become blurred, and clear policies on data sharing and publication are needed. Statements on data sharing in public health emergencies by Wellcome and WHO are positive developments and will hopefully provide a platform to establish appropriate data sharing frameworks for malaria vector genomic surveillance (WHO, 2015; Dye et al., 2016; Wellcome, 2016; Wellcome, 2020).

Further technology and analytical methods development

A key insight from this thesis, and from the broader programme of work carried out by the Ag1000G project, is that current genome sequencing technology and analytical methods are mature and ready for translation into operational genomic surveillance systems for malaria vectors. In other words, implementation of malaria vector genomic surveillance systems should not wait for further research and development. Nevertheless, further development of both sequencing technology and analytical methods could enhance the capabilities of genomic surveillance systems in several ways. There are many avenues for potential development, but I would like to highlight two areas in particular.

Sequencing technology itself is still a rapidly developing field, with long-read sequencing technologies from Pacific Biosciences and Oxford NanoPore reaching maturity and demonstrating the capability to sequence DNA from smaller organisms such as mosquitoes (Kingan et al., 2019; Ghurye et al., 2019; Zamyatin et al., 2020). Although long-read sequencing is currently being applied primarily for the construction of improved reference genomes,

as costs fall it will also become applicable to population-scale sequencing and analysis of natural genome variation. At that point it could become a valuable surveillance tool, because it will allow for much improved discovery and genotyping of structural variation, which we know plays a major role in insecticide resistance (Lucas et al., 2019). There are also new sequencing technologies based on high-throughput Illumina short read sequencing but using linked-reads to allow reconstruction of long-range haplotypes (Mostovoy et al., 2016). Accurate haplotypes are beneficial for surveillance because they increase power to perform a variety of demographic and evolutionary inferences, including identifying genes under recent selection, and investigating the origins and spread of insecticide resistance outbreaks. In both cases, they key to making these sequencing technologies accessible for malaria vector surveillance will be to reduce costs sufficiently to allow sequencing of samples at an informative spatial and temporal scale.

On the analytical side, based on my experience of working on the Ag1000G project, my impression is that the field of population genomics is still young, and there is much room for improvement in the available statistical and machine learning methods for demographic and evolutionary inference. In some ways, genomic surveillance of infectious disease pathogens and vectors shares something in common with the field of conservation genetics, because important decisions may be influenced by analyses of genomic data, and therefore inferences must be robust (McMahon et al., 2014; Supple and Shapiro, 2018). A particular area for development is the analysis of the emergence and spread of new insecticide resistance variants. Here there are strong parallels with the analysis of outbreaks of an infectious pathogen, and many of the questions we might like to ask are common, such as where and when did the outbreak begin, are there multiple simultaneous outbreaks occurring, where and how are outbreaks spreading between different locations, and are some variants spreading more rapidly than others. In the case of pathogen outbreak investigation, analytical methods based on phylogenetic inference are reasonably well-developed (De Maio et al., 2015; Grubaugh et al., 2018). Standard phylogenetic methods cannot be applied directly to insecticide resistance outbreaks, however, because *Anopheles* mosquitoes are sexually recombining, and thus there is no single phylogeny for any given pair of

haplotypes, rather there is a sequence of phylogenies that varies along the genome. In my analysis of the *Vgsc* gene in Chapter 6, I used relatively simple heuristic methods based on fixed genomic windows to show that there is a clear structuring among the haplotypes carrying pyrethroid resistance alleles, and that resistance is spreading between countries and species. Developing these analyses further to answer other questions, such the origins, timing and direction of spread of resistance variants, will require methods that can infer a phylogeny accurately at a specific locus of interest within a recombining genome, such as an insecticide resistance variant. Several new methods have recently been developed in this direction (Kelleher et al., 2019; Speidel et al., 2019). In particular, it may be valuable to leverage information about recombination when inferring phylogenies, because recombination events may provide greater resolution to resolve recent genealogical relationships, whereas mutations may take longer to accumulate (Albers and McVean, 2020; Mathieson and McVean, 2014). To build a coherent statistical framework for the investigation of insecticide resistance outbreaks, a synthesis is required that brings together inference methods from phylogeography and phylodynamics and applies them to genealogies inferred at a recombining insecticide resistance locus.

Investment in entomological surveillance capacity

As noted in the introduction, current malaria vector surveillance programmes are not sufficiently resourced, and there is a lack of trained personnel and facilities (Russell et al., 2020). No amount of technological innovation can compensate for this. Molecular surveillance methods cannot stand by themselves, but rather add value to and complement existing entomological surveillance activities, which collect and publish data on the bionomics and phenotype of malaria vectors and other relevant variables such as intervention coverage. Significant investment is thus essential to develop a strong foundation of entomological surveillance capability. It is also increasingly important to standardise surveillance methods. Standardisation is particularly important for the measurement of insecticide resistance phenotypes, and especially for new insecticides and synergists, where rapid confirmation of any genetic signatures of emerging resistance will be needed. Finally, we will need an

improved capability to make predictions about operational outcomes based on surveillance data, with sufficient confidence to justify changes in vector control policy. This is a complex challenge and an important area for further research and policy development (Churcher et al., 2016; Sherrard-Smith et al., 2018; South and Hastings, 2018; Lines, 2019), explored further below.

Bridging the gap from surveillance to impact

I would like to conclude by briefly discussing the broader context, and some of the other challenges that will need to be addressed before genomic surveillance systems for malaria vectors could contribute to achieving a real impact on improved public health. In the emerging field of genomic surveillance, the term "actionable information" is often used to articulate the idea that genomic data should generate insights that policy-makers and public health agencies can act on. However, in general, genomic data will only become actionable when integrated with many other sources of data, via mathematical or computational models that predict the outcome of a given course of action with sufficient accuracy. To restate, making informed decisions requires (1) multiple data streams, and (2) models to make predictions from data.

To illustrate this point, consider again the example of the novel SARS-CoV-2 lineage B.1.1.7. This lineage was detected via genomic surveillance, which showed that it carried an unusual number of mutations in functionally-relevant genomic locations, and also showed that it appeared to be replacing other lineages, suggesting it carried a selective advantage (Rambaut et al., 2020). However, the genomic data did not tell us why the lineage was expanding, or whether new public health measures would be required. Providing answers to these questions requires fitting mathematical transmission models to other data including hospital admissions, hospital and ICU bed occupancy, deaths, PCR prevalence and seroprevalence, in addition to genetic data on strain frequency (Davies et al., 2020). Deciding an appropriate course of action then requires consideration of many other factors, including logistics, economics, political and social impact, etc.

Adapting this example to malaria vectors, consider the scenario where a next-generation

PBO LLIN product is deployed at scale in a particular country. Now assume that a new genetic variant emerges in a major malaria vector species which restores full pyrethroid resistance by somehow subverting the action of the PBO synergist or utilising a different mechanism of resistance. Genomic surveillance would detect this variant, because it would rise rapidly in frequency and begin spreading to multiple locations. However, these data would not tell us to what extent the efficacy of the new LLINs was reduced due to the new variant, nor what effect this would have on disease prevalence. To do that, genetic data would need to be integrated with data from insecticide resistance bioassays and other entomological variables, in addition to any and all available epidemiological data. Even then, no fully developed models exist for predicting efficacy from these data, and some form of experimental trials would be needed to link genotype to phenotype in order to parameterize models. Finally, given the long lead times for LLIN procurement, it remains questionable whether sufficient evidence could be generated in time for an effective course correction to be made.

Consider also a scenario where a country has an IRS programme targeting regions of high transmission, which is rotating annually between three next-generation IRS products. The efficacy of IRS rotation relies on the management of resistance allele frequencies, where each insecticide is used for a period of time that allows resistance alleles to the other insecticides to fall but does not allow resistance alleles to the insecticide in use to reach fixation. Genomic surveillance could track the frequencies of all known alleles conferring resistance to the insecticides used in the rotation, and provide information regarding whether any alleles have reached fixation. Observing fixation of a resistance allele would indicate in theory that a particular insecticide should be removed from the rotation, and might also suggest that the duration of the rotation should be altered. However, insecticide resistance in malaria vectors is polygenic, and so observing fixation of one resistance allele might not justify removing an entire product from use if other resistance alleles remain at lower frequencies, and the product remains effective. Thus, data on changes in efficacy in response to genetic changes would also be needed to reach an informed decision. Determining whether the rotation strategy was well-designed would

also require modelling of multiple alleles, which would require estimation of parameters such as selection coefficients. Furthermore, financial constraints might limit choices, or there may simply not be any other suitable products available to fill a gap in the rotation.

Improving malaria vector control through improved surveillance is thus a multifaceted challenge, where genome sequencing can play an important role, but where a holistic approach is needed. The entomologist Peter Mattingly wrote in 1963, during the first global malaria eradication campaign:

> "Every eradication campaign is, at all stages, a piece of operational research and our ignorance is such that we must be prepared in all cases to learn as we go along. The usefulness of the entomologist in the final phases of an eradication campaign must largely depend on the amount they have been able to learn during the preceding phases and the extent to which they have been willing and able to peer ahead into the future."

Despite the passage of nearly sixty years, these words remain relevant today. They remind us that the value of new technologies such as genome sequencing resides in their ability to confront us anew with our own ignorance, and to help us learn as we act in this next phase of the journey towards a world without malaria.

References

Albers, PK and G McVean (2020). 'Dating genomic variants and shared ancestry in population-scale sequencing data.' In: *PLoS Biol.* 18.1 (1). Ed. by NH Barton, e3000586. DOI: 10.1371/journal.pbio.3000586.

Bass, C and LM Field (2011). 'Gene amplification and insecticide resistance.' In: *Pest Manag. Sci.* 67.8 (8), pp. 886–890. DOI: 10.1002/ps.2189.

Bayili, K, S N'do, M Namountougou, R Sanou, A Ouattara, RK Dabiré, AG Ouédraogo, D Malone and A Diabaté (2017). 'Evaluation of efficacy of Interceptor® G2, a long-lasting insecticide net coated with a mixture of chlorfenapyr and alpha-cypermethrin, against

pyrethroid resistant Anopheles gambiae s.l. in Burkina Faso'. In: *Malar. J.* 16.1. DOI: 10.1186/s12936-017-1846-4.

Churcher, TS, N Lissenden, JT Griffin, E Worrall and H Ranson (2016). 'The impact of pyrethroid resistance on the efficacy and effectiveness of bednets for malaria control in Africa'. In: *Elife* 5. DOI: 10.7554/eLife.16090.

Dao, A, AS Yaro, M Diallo, S Timbiné, DL Huestis, Y Kassogué, AI Traoré, ZL Sanogo, D Samaké and T Lehmann (2014). 'Signatures of aestivation and migration in Sahelian malaria mosquito populations.' In: *Nature* 516.7531, pp. 387–390. DOI: 10.1038/nature13987.

Davies, N et al. (2020). *Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England*. Tech. rep. Centre for Mathematical Modelling of Infectious Diseases (CMMID).

De Maio, N, CH Wu, KM O'Reilly and D Wilson (2015). 'New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation.' In: *PLoS Genet.* 11.8 (8). Ed. by JK Pritchard, e1005421. DOI: 10.1371/journal.pgen.1005421.

Devonshire, AL and LM Field (1991). 'Gene amplification and insecticide resistance.' In: *Annual review of entomology* 36.1, pp. 1–21. DOI: 10.1146/annurev.en.36.010191.000245.

Dye, C, K Bartolomeos, V Moorthy and MP Kieny (2016). 'Data sharing in public health emergencies: a call to researchers.' In: *Bull. World Health Organ.* 94.3 (3), p. 158. DOI: 10.2471/BLT.16.170860.

Elliott, R and V Ramakrishna (1956). 'Insecticide Resistance in Anopheles gambiae Giles'. In: *Nature* 177, pp. 532–533.

Ghurye, J, S Koren, ST Small, S Redmond, P Howell, AM Phillippy and NJ Besansky (2019). 'A chromosome-scale assembly of the major African malaria vector Anopheles funestus.' In: *GigaScience* 8.6 (6). DOI: 10.1093/gigascience/giz063.

Grubaugh, ND, JT Ladner, P Lemey, OG Pybus, A Rambaut, EC Holmes and KG Andersen (2018). 'Tracking virus outbreaks in the twenty-first century.' In: *Nature Microbiology* 4.1 (1), pp. 10–19. DOI: 10.1038/s41564-018-0296-2.

Hancock, PA, CJ Hendriks, JA Tangena, H Gibson, J Hemingway, M Coleman, PW Gething, E Cameron, S Bhatt and CL Moyes (2020). 'Mapping trends in insecticide resistance phenotypes in African malaria vectors'. In: *PLoS Biol.* 18.6. Ed. by AF Read, e3000633. DOI: 10.1371/journal.pbio.3000633.

Hemingway, J, N Hawkes, L Prapanthadara, KG Jayawardenal and H Ranson (1998). 'The role of gene splicing, gene amplification and regulation in mosquito insecticide resistance.' In: *Philos. Trans. R. Soc. B Biol. Sci.* 353.1376 (1376). Ed. by I Denholm, JA Pickett and AL Devonshire, pp. 1695–1699. DOI: 10.1098/rstb.1998.0320.

Huestis, DL et al. (2019). 'Windborne long-distance migration of malaria mosquitoes in the Sahel'. In: *Nature* 574.7778, pp. 404–408. DOI: 10.1038/s41586-019-1622-4.

Hui, TYJ and A Burt (2015). 'Estimating effective population size from temporally spaced samples with a novel, efficient maximum-likelihood algorithm.' In: *Genetics* 200.1 (1), pp. 285–293. DOI: 10.1534/genetics.115.174904.

Kelleher, J, Y Wong, AW Wohns, C Fadil, PK Albers and G McVean (2019). 'Inferring whole-genome histories in large population datasets.' In: *Nat. Genet.* 51.9 (9), pp. 1330–1338. DOI: 10.1038/s41588-019-0483-y.

Kingan, SB, H Heaton, J Cudini, CC Lambert, P Baybayan, BD Galvin, R Durbin, J Korlach and MKN Lawniczak (2019). 'A High-Quality De novo Genome Assembly from a Single Mosquito Using PacBio Sequencing.' In: *Genes* 10.1 (1), p. 62. DOI: 10.3390/genes10010062.

Kumuthini, J et al. (2019). 'The H3ABioNet helpdesk: an online bioinformatics resource, enhancing Africa's capacity for genomics research.' In: *BMC Bioinf.* 20.1 (1), p. 741. DOI: 10.1186/s12859-019-3322-3.

Lines, J (2019). 'Malaria nets shape up for resistance'. In: *Nature Microbiology* 5.1, pp. 6–7. DOI: 10.1038/s41564-019-0646-8.

Lucas, ER, A Miles, NJ Harding, CS Clarkson, MKN Lawniczak, DP Kwiatkowski, D Weetman, MJ Donnelly and A gambiae 1000 Genomes Consortium (2019). 'Whole-genome sequencing reveals high complexity of copy number variation at insecticide

resistance loci in malaria mosquitoes.' In: *Genome Res.* 29.8 (8), pp. 1250–1261. DOI: 10.1101/gr.245795.118.

Makoni, M (2020). 'Africa's $100-million Pathogen Genomics Initiative'. In: *The Lancet Microbe* 1.8, e318. DOI: 10.1016/S2666-5247(20)30206-8.

Mathieson, I and G McVean (2014). 'Demography and the age of rare variants.' In: *PLoS Genet.* 10.8 (8). Ed. by J Novembre, e1004528. DOI: 10.1371/journal.pgen.1004528.

McMahon, BJ, EC Teeling and J Höglund (2014). 'How and why should we implement genomics into conservation?' In: *Evol. Appl.* 7.9 (9), pp. 999–1007. DOI: 10.1111/eva.12193.

Montgomery, SB et al. (2013). 'The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes.' In: *Genome Res.* 23.5 (5), pp. 749–761. DOI: 10.1101/gr.148718.112.

Mostovoy, Y et al. (2016). 'A hybrid approach for de novo human genome sequence assembly and phasing.' In: *Nat. Methods* 13.7 (7), pp. 587–590. DOI: 10.1038/nmeth.3865.

Msomi, N et al. (2020). 'A genomics network established to respond rapidly to public health threats in South Africa'. In: *The Lancet Microbe* 1.6, e229–e230. DOI: 10.1016/s2666-5247(20)30116-6.

Mulder, N, A Abimiku, SN Adebamowo, J de Vries, A Matimba, P Olowoyo, M Ramsay, M Skelton and DJ Stein (2018). 'H3Africa: current perspectives.' In: *Pharmacogenomics and personalized medicine* 11, pp. 59–66. DOI: 10.2147/PGPM.S141546.

North, AR, A Burt and HCJ Godfray (2019). 'Modelling the potential of genetic control of malaria mosquitoes at national scale.' In: *BMC Biol.* 17.1 (1), p. 26. DOI: 10.1186/s12915-019-0645-5.

Oxborough, RM, J Kitau, R Jones, E Feston, J Matowo, FW Mosha and MW Rowland (2014). 'Long-lasting control of Anopheles arabiensis by a single spray application of micro-encapsulated pirimiphos-methyl (Actellic® 300 CS)'. In: *Malar. J.* 13.1. DOI: 10.1186/1475-2875-13-37.

Oxborough, RM et al. (2019). 'Susceptibility testing of Anopheles malaria vectors with the neonicotinoid insecticide clothianidin; Results from 16 African countries, in preparation

for indoor residual spraying with new insecticide formulations'. In: *Malar. J.* 18.1. DOI: 10.1186/s12936-019-2888-6.

Protopopoff, N et al. (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, two-by-two factorial design trial'. In: *The Lancet* 391.10130, pp. 1577–1588. DOI: 10.1016/S0140-6736(18)30427-6.

Rambaut, A, N Loman, O Pybus, W Barclay, J Barrett, A Carabelli, T Connor, T Peacock, DL Robertson and E Volz (2020). *Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations*. Tech. rep. The COVID-19 Genomics UK (COG-UK) consortium.

Redmond, SN, BM MacInnis, S Bopp, AK Bei, D Ndiaye, DL Hartl, DF Wirth, SK Volkman and DE Neafsey (2018). 'De Novo Mutations Resolve Disease Transmission Pathways in Clonal Malaria.' In: *Mol. Biol. Evol.* 35.7 (7). Ed. by D Falush, pp. 1678–1689. DOI: 10.1093/molbev/msy059.

Russell, TL, R Farlow, M Min, E Espino, A Mnzava and TR Burkot (2020). 'Capacity of National Malaria Control Programmes to implement vector surveillance: a global analysis'. In: *Malar. J.* 19.1, p. 422. DOI: 10.1186/s12936-020-03493-1.

Sherrard-Smith, E et al. (2018). 'Systematic review of indoor residual spray efficacy and effectiveness against Plasmodium falciparum in Africa'. In: *Nat. Commun.* 9.1. DOI: 10.1038/s41467-018-07357-w.

South, A and IM Hastings (2018). 'Insecticide resistance evolution with mixtures and sequences: A model-based explanation'. In: *Malar. J.* 17.1. DOI: 10.1186/s12936-018-2203-y.

Speidel, L, M Forest, S Shi and SR Myers (2019). 'A method for genome-wide genealogy estimation for thousands of samples.' In: *Nat. Genet.* 51.9 (9), pp. 1321–1329. DOI: 10.1038/s41588-019-0484-x.

Staedke, SG et al. (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.' In: *The Lancet* 395.10232 (10232), pp. 1292–1303. DOI: 10.1016/S0140-6736(20)30214-2.

Supple, MA and B Shapiro (2018). 'Conservation of biodiversity in the genomics era.' In: *Genome Biol.* 19.1 (1), p. 131. DOI: 10.1186/s13059-018-1520-3.

Tegally, H et al. (2020). 'Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa'. In: *medRxiv*. DOI: 10.1101/2020.12.21.20248640.

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'. In: *Genome Res.* 30.10, pp. 1533–1546. DOI: 10.1101/gr.262790.120.

The COVID-19 Genomics UK (COG-UK) consortium (2020a). 'An integrated national scale SARS-CoV-2 genomic surveillance network'. In: *The Lancet Microbe* 1.3, e99–e100. DOI: 10.1016/s2666-5247(20)30054-9.

The COVID-19 Genomics UK (COG-UK) consortium (2020b). *Summary report: COG-UK geographic coverage of SARS-CoV-2 sample sequencing*. Tech. rep. The COVID-19 Genomics UK (COG-UK) consortium.

Tiono, AB et al. (2018). 'Efficacy of Olyset Duo, a bednet containing pyriproxyfen and permethrin, versus a permethrin-only net against clinical malaria in an area with highly pyrethroid-resistant vectors in rural Burkina Faso: a cluster-randomised controlled trial'. In: *The Lancet* 392.10147, pp. 569–580. DOI: 10.1016/s0140-6736(18)31711-2.

Weedall, GD, JM Riveron, J Hearn, H Irving, C Kamdem, C Fouet, BJ White and CS Wondji (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'. In: *PLoS Genet.* 16.6 (6). Ed. by RH ffrench-Constant, e1008822. DOI: 10.1371/journal.pgen.1008822.

Wellcome (2016). *Statement on data sharing in public health emergencies*.

Wellcome (2020). *Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak*.

WHO (2015). *Developing global norms for sharing data and results during public health emergencies*. Tech. rep. World Health Organization.

WHO (2019). *World malaria report 2019*. Tech. rep. World Health Organization.

Zamyatin, A, P Avdeyev, J Liang, A Sharma, C Chen, V Lukyanchikova, N Alexeev, Z Tu, MA Alekseyev and IV Sharakhov (2020). 'Chromosome-level genome assemblies of the malaria vectors Anopheles coluzzii and Anopheles arabiensis'. In: *bioRxiv*. DOI: 10.1101/2020.09.29.318477.