Genomic introgression events in the *Anopheles gambiae* complex.

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

Liverpool School of Tropical Medicine

January 2021

Abstract

Malaria is a disease caused by the Plasmodium parasite and is transmitted by the infectious bites of Anopheles mosquitoes. An estimated 409,000 deaths were attributed to the disease in 2019, with approximately 229 million cases worldwide. Long-lasting insecticidal nets and indoor residual spraying are two applications that have been the most effective in controlling transmission. However, the effectiveness of these public health insecticides has long been threatened by insecticide resistance in the key vector species which transmit the parasite.

The *Anopheles gambiae* 1000 genomes (Ag1000G) project is an international consortium that aims to create a databank of whole-genome sequenced Anopheles samples from across Africa. The third phase of the Ag1000G increases the number of specimens to ~3000 from 18 countries, these samples are to include *An. arabiensis* from Uganda, Malawi, Kenya and Tanzania. These data provide a unique opportunity to investigate the *An. arabiensis* genome. This is of specific interest due to the growing importance of *An. arabiensis* as it displaces *An. gambiae* populations in some regions.

In this thesis, I develop and explore three key investigations with the *An. arabiensis* genome as the focus, these are: development of ancestry informative markers, discovery of copy number variation and signals of introgression.

A panel of ancestry informative markers (AIMs), which can resolve between *An. gambiae* and *An. arabiensis* is a useful tool for many bioinformatic analyses. Here, we develop this panel and use them to inform the discovery of genome regions which may be introgressed between the two species. Our results show that the application of AIMs in this study was unable to satisfactorily identify regions of introgression between the two species. However, the panel has a broader application in other analyses which seeks to identify SNPs which are segregated between the species.

Recent findings show that detoxification genes are significantly enriched in the CNVs discovered in *An. gambiae* mosquitoes. We used the *An. arabiensis* whole genome sequence data from the Ag1000G project to apply a minimally modified version of the previously published method for CNV discovery. Though we identified many fewer CNVs in the *An. arabiensis* dataset, we did discover enrichment of CNVs which contain genes associated with insecticide detoxification.

Finally, we developed an analyses pipeline which calculates Patterson's D statistic across the genome. This pipeline is generalised and applicable to future analyses. We calculated Patterson's D statistic for all combinations of populations and outgroups that could be informative about introgression between *An. gambiae* and *An. arabiensis*. Many significant signals of introgression were observed that were both ubiquitous across all populations and in regions that contain known insecticide resistance loci. We discuss here the implications and limitations of the analyses and contextualise the importance of continuing to characterise the role

of introgression in both the general evolution and adaptation to anthropogenic pressures from insecticides. We ultimately demonstrate and provide a basis for continued analyses into the role of introgression in insecticide resistance.

Dedication

In loving memory of Mary and George Carmichael

Acknowledgments

I want to convey here that attaining my PhD and composing this thesis was the result of the efforts of so many more people than just me. I want to thank the key people here, but acknowledge that this is by no stretch of the imagination an exhaustive list of people who have helped in some way or another. My time as a University student began in 2012. Calculating the elapsed time from then to the submission of this thesis (04/01/2021), it has been 8 years, 3 months, 21 days. Certainly, I have enjoyed the benefits of student discount more than most.

The time I have spent in the pursuit of something that I find deeply interesting and impactful, simply would not have been possible if it were not for a vast number of people. To those people, I do not think I could adequately convey the extent of my gratitude. It has been the singularly most difficult, stimulating, privileged, and enriching experience of my life. I find myself somewhat lost for words in trying to describe the support I have received from my family and academic colleagues, and my gratitude for it.

My supervisory team for the PhD is larger than most. Officially, it is composed of, Professor Donnelly, Dr Weetman, Dr Lucas and Dr Sedda. Unofficially, it also includes Dr Harding, Dr Clarkson and Mr Miles of the Kwiatkowski group at the Wellcome Sanger Institute and The Big Data Institute. Academically, these are the people to whom I owe the most gratitude. Professor Donnelly and Dr Weetman provided a solid foundation and positive environment

for understanding the broader context of the research, and the standing of the research in the literature. Dr Lucas bolstered my understanding of so many things that I could not possibly list them, suffice to say I am grateful for his patience. Especially answering an uncountable number of questions that were prefixed with the phrase “I know this is a stupid question, but…” or “I know I’ve asked you this before, but…”. Dr Sedda is thanked for his enthusiastic and driven supervision of statistics. Dr Clarkson was my guide into the Ag1000G data and Python coding, and I thank him for his patience in making sure I was up to speed. Dr Harding and Mr Miles levelled up my understanding of how bioinformatic analyses should be coded and implemented. They helped me develop my understanding of how best to think about the data, the analyses, and the interactions between them. The days at the BDI with Dr Harding and Mr Miles were the most mentally exhausting, due to the sheer volume of extremely important skills they were helping me develop. I would also like to acknowledge the support staff of all the institutions, without them, academic staff would not be able to conduct their research. These people the silent work force that enable science to continue. Dr Wagstaff and Dr Roberts are sincerely thanked for their pastoral support, motivation, praise and candour in the yearly PAP reviews, over the course of my PhD.

Over the PhD I have spent a lot of time in various cities in the UK, principally, Liverpool, Cambridge, and Oxford. The Magic: the Gathering communities in these cities should be thanked for their openness and willingness to welcome new people. A lot of those people are now very good friends, and it is thanks to their hospitality that living in a new city for months at a time alone was not as lonely as

it might otherwise have been. I have honestly met some truly fascinating and generous people in this way. I would like to thank all of them, however, specifically name and thank Beth Colman, Sam Fryman, Steve Bailey, Jason Collins, Michael Yearby, James Miles, Francesdo Puglisi, Samuel H O Collis, Tom Sutton, Kal Duskryn, Kevin Donkers, Matthew Kinnear, Dominic Sanders, Josh Stimson, Luis Alverez Zucchino, Ryan Jordan, Sam Milner, Ezra Payne, Adam Thomas, Katie Roberts, Pietro Sgambati, Chris Stanley, Fabien Sivnert and Gabriel Rajib Hussaun.

Finally, my family. I am privileged to have a family that supported my pursuit of a PhD. My parents Barbara and John, my brothers Stephen and James, my Great Uncle and Godfather Jimmy and my very patient (and long suffering) partner Toni. These people have, in many ways, been the reason why I've been able to spend so much time focused on my PhD. Without betraying the privacy of family life, the financial and emotional support they provided were essential to the completion of my studies. Though they are aware of my gratitude, I want to memorialise it here that each of these people have been central to my sanity, health and development as a scientist and person over the years.

- Thank you all

Frontispiece

“Most evenings of the last 4 years” - Pictured: Sean (left) and Toni (right)

Art by Terese Neilsen (Fact or Fiction)

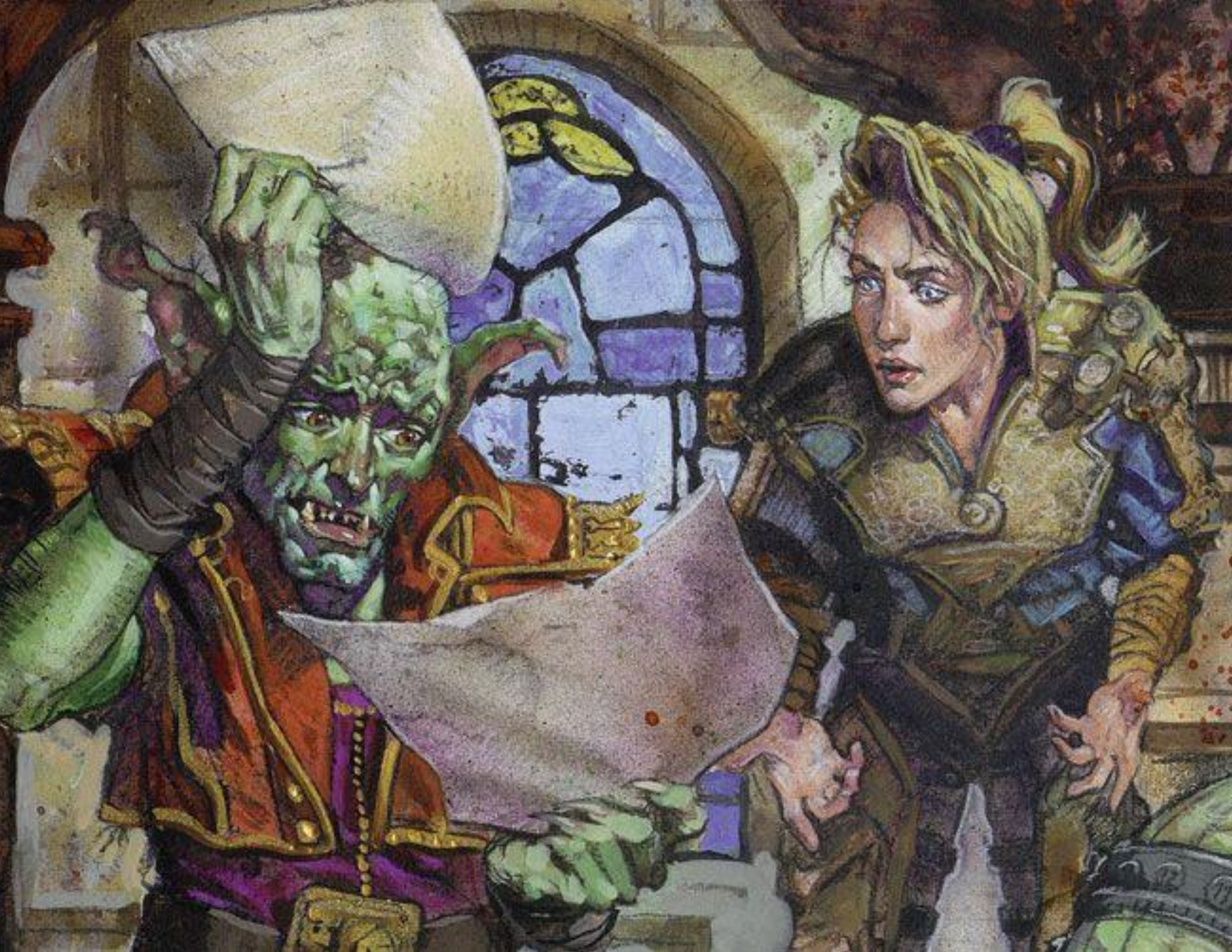


Table of Contents

Chapter I.

Introduction

Abstract

The defining conditions for speciation have traditionally been a contentious topic within evolutionary biology. However, sympatric speciation and consequently gene flow between insipient species are increasingly being recognised. *Anopheles gambiae s.s.* and *Anopheles coluzzii* are the two most closely related members of the *An. gambiae* complex and represent an ideal model for studying gene flow between malaria vector species and in determining whether their sympatric speciation has led to gene flow; and more crucially whether such gene flow has conferred a survival advantage to the malaria vectors. Indeed, many findings have suggested introgression not only between *An. gambiae s.s.* and *An. coluzzii*, but also *An. gambiae s.s.* and *Anopheles arabiensis*, a non-sister taxa relation. Such transfer of genetic information through porous species barriers is of key significance for vector control and global health research, as introgression may represent a viable route for the acquisition of selectively advantageous traits, such as insecticide resistance. In this chapter, we review the historical and contemporary evidence for introgression between *An. gambiae* complex members.

The Anopheles Genus

Of the 462 formal and 50 provisional species of Anopheles, ~70 are competent vectors for at least one of the five human malaria parasites (Hay *et al.*, 2010; Service and Townson, 2002). The taxonomical genus of *Anopheles*, belonging to the Culicidae family was recently subject to species reclassification, designating *An. gambiae* “S form” and “M form” as distinct species named *Anopheles gambiae* Giles and *An. coluzzii* Coetzee & Wilkerson, respectively - based on molecular and bionomical evidence.

Among African malaria vectors, *An. gambiae*, *An. coluzzii* and *An. arabiensis* (of the *Anopheles gambiae* complex) and *Anopheles funestus* (of the *Anopheles funestus* subgroup) are crucially significant when considering transmission of *Plasmodium* parasites to humans. Further to these species *Anopheles melas* and *Anopheles merus* (of the *Anopheles gambiae* complex) are identified as locally dominant vector species. By virtue of their significance in malaria transmission and their threat to public health, the most studied and researched vectors for the African continent are those of the *Anopheles gambiae* complex and the *Anopheles funestus* subgroup.

Public Health Context

Despite many years of research and substantial progress in malaria transmission reduction and case management, an estimated ~429,000 deaths are attributable to malaria (Bhatt *et al.*, 2015). The most important malaria control

strategies are insecticide-based interventions, such as long-lasting insecticidal nets (LLIN) and indoor residual spraying (IRS), which combined have helped to avert an estimated 450 million cases of clinical malaria, since 2000 (Bhatt *et al.*, 2015; WHO, 2016). Unfortunately, the efficacy of insecticide-based interventions has been compromised in many regions across Africa, Asia and south America, due to the emergence of insecticide resistance (WHO, 2016).

To mitigate the effects of insecticide resistance, many studies are focused on characterising the molecular landscape which underpins the insecticide resistant phenotypes. The identification of one of the underlying mechanisms for pyrethroid resistance in *Anopheles* vectors has driven efforts to develop an insecticide-synergist formulation. In the example of pyrethroids, piperonyl butoxide (PBO) – a compound with no insecticidal properties – can inhibit members of the cytochrome P450 family responsible for pyrethroid detoxification (Gleave *et al.*, 2018). Furthermore, identifying the molecular components of insecticide resistance phenotypes allows the development of diagnostic markers for resistance. Such markers allow for the detection of resistance at earlier stages than phenotypic analysis and allow malaria control programmes to monitor the spread of resistance alleles both within and across populations (Weetman and Donnelly, 2015).

The identification of resistance markers has traditionally relied upon investigating candidate genes, or quantitative trait loci- both requiring *a priori* knowledge of the genome of the study organism. As detailed by Donnelly *et al.*

(2016), development of genetic technologies and greater availability of whole genome sequence data is expected to precipitate a paradigm shift in the techniques used to leverage meaningful diagnostic information from resistant populations. Indeed, the development of genetic technologies allows the investigation of more complex genetic events that may be underlying insecticide resistance phenotypes, which may have precipitated the acquisition/development of a resistance phenotype. Here, we explore the role of introgression as one of these complex genetic events.

Evidence of Introgressive Hybridisation in Anopheles

Studies of introgression in Anophelines pre-2000 was focussed on attempting to understand the relationship between introgression and the phylogenetic nature and speciation of the *Anopheles gambiae* complex. The cause of added complexity in reconstructing the Anopheline phylogeny has been cited as relatively recent divergence of the sibling species (Besansky *et al.*, 1994). Therefore, most research focused on phylogenetic reconstructions in attempts to delineate the evolutionary relationships between the species, an effort which was encumbered by conflicting evidence, hypothesised to be caused by introgression. In this section, I discuss the key findings in Anopheline phylogenetic research before the turn of the millennium, as it relates to introgression.

Based on polytene chromosome inversions in *An. gambiae* and *An. arabiensis*, Coluzzi *et al.* (1979) hypothesised that introgressive hybridisation can

occur via fertility of female hybrids and hybrid/parental species backcrossed generations. Indeed hybridisation of *An. gambiae* and *An. arabiensis* had been identified previously (White, 1971), however, no hypothesis of introgression underpinning Anopheline chromosome inversions was proposed until the work of Coluzzi *et al.* (1979). Certainly, polytene chromosomes and amplicon sequencing-based studies show internal conflicts which are most readily reconciled by invoking pervasive introgression.

The difficulties in reconstructing a phylogeny for the *An. gambiae* complex prolonged an uncertainty surrounding the evolutionary relationships between member species. Besansky *et al.* (1994) presented evidence which challenged the generally accepted interpretation of the phylogenetic relationship between *An. gambiae* and *An. arabiensis*; being that both species occupy distal branches on the 'true' phylogenetic tree. Utilising sequences from nuclear ribosomal DNA (rDNA), mitochondrial DNA (mtDNA) and an esterase locus, Besansky and colleagues demonstrated that sequence-based methods reconstruct a tree which places *An. gambiae* and *An. arabiensis* as sister taxa. The phylogenetic reconstruction based on the esterase sequence data supported a monophyletic origin for *An. gambiae* and *An. arabiensis*, suggestive of either an introgressive event or an incorrect chromosome-based phylogeny.

Further evidence of introgression between *An. gambiae* and *An. arabiensis* was presented in a study which aimed at reconstructing the phylogeny of *Anopheles bwambae* and *An*. *melas* (Caccone *et al.*, 1996). The AT-rich control

region found within the mtDNA was sequenced in six members of the *An. gambiae*. and revealed a high congruence with tree reconstructions based on chromosomal inversions. The exception was the placement of *An. gambiae* and *An. arabiensis*, which were grouped as sister taxa. The direction of gene flow was also hypothesised to be unidirectional from *An. gambiae* to *An. arabiensis* based on the observation that the clade for *An. gambiae* and *An. arabiensis* is the predicted position for *An. gambiae*.

Garcia *et al.* (1996) outline the conflict between phylogenetic tree topologies based on inversions compared to trees reconstructed through mtDNA and rDNA intergenic spacer region sequences; the former placing *An. gambiae* and *An. merus* as sister taxa and the latter placing *An. gambiae* and *An. arabiensis* as sister taxa. Tree topologies reconstructed from inversions by Coluzzi *et al.* (1979) show strong concordance that *An. gambiae* and *An. merus* are sister taxa and *An. arabiensis* are more distantly related. However, Garcia *et al.* (1996) attempts to reconcile the apparent sister taxa relationship between *An. gambiae and An. arabiensis* when phylogenies are reconstructed from mtDNA and rDNA intergenic spacer regions. It is proposed that because there is evidence of mtDNA gene flow being more amenable in other groups, that introgression of the mtDNA between *An. gambiae An. arabiensis* may explain discordant tree topologies (Hanemaaijer *et al.*, 2018). However, this would not explain the rDNA results published by Besansky *et al.* (1994), Garcia and colleagues go on to hypothesise alternative explanations for the convoluted phylogenies, however, ultimately concluding that

regardless of the explanation - for which their study had insufficient data to investigate – there is clear evidence of introgression between *An. gambiae and An. arabiensis* – the two most medically important vectors of the time (Garcia *et al.*, 1996).

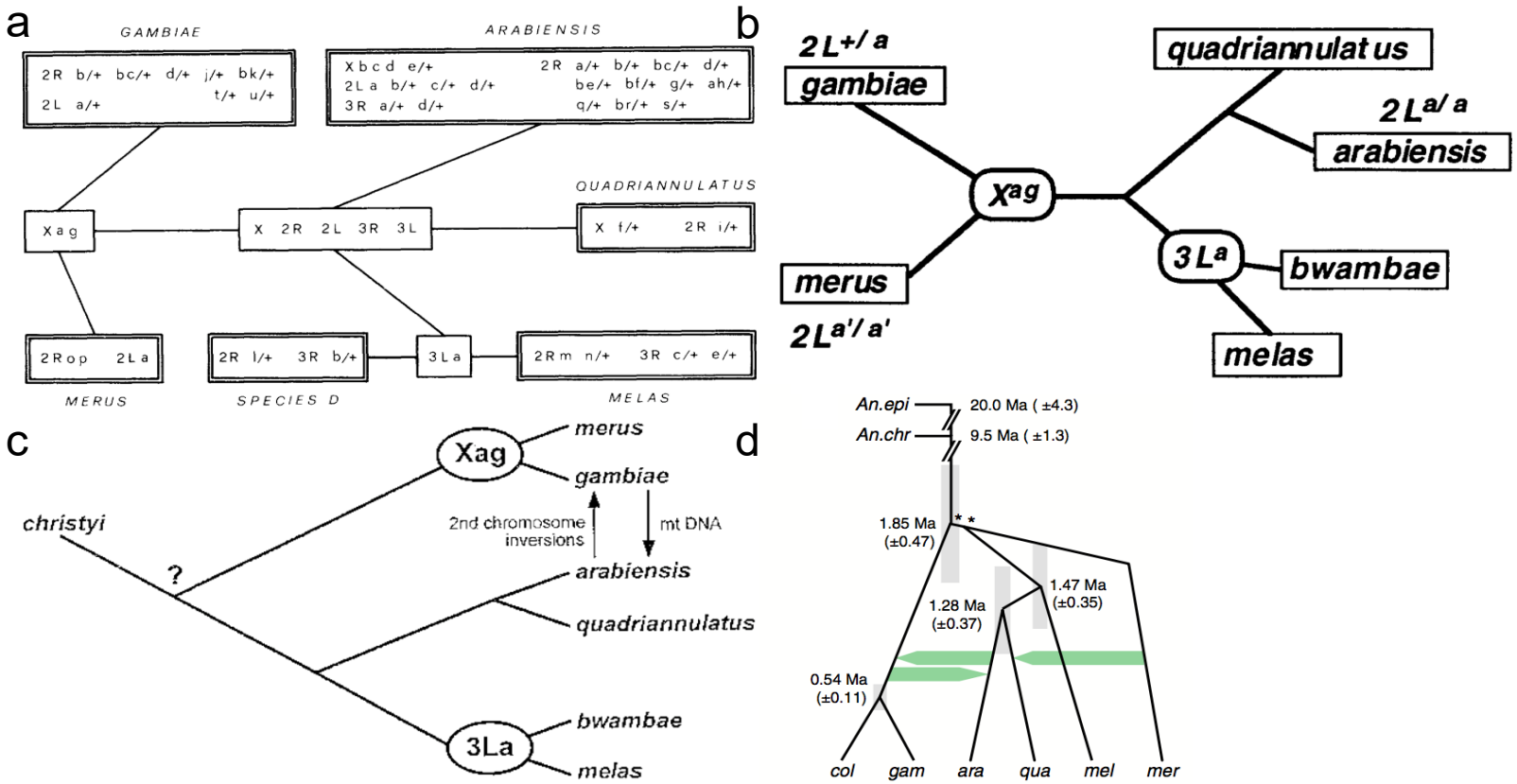


Figure 1. Phylogenic reconstructions from various studies used to illustrate the differing range of reconstructions over time. Early pylogenies observed in panels a, b and c are focussed around contextualising phylogenetic relationships in terms of inversions, such as Xag and 3La. Whereas the contemporary approach employed in panel d uses maximum-likelihood (ML) phylogenies constructed from 50-kilobase nonoverlapping windows. The reconstitution of of the phylogenies is highlighted by the polar chage that sees An. merus places distal to An gambiae in contemporary methods, where previously it was placed inthe same clade(a: Coluzzi *et al.*, 1979; b: Caccone *et al.*, 1998; c: Powell *et al.*, 1999; d: Fontaine *et al.*, 2015)

In a study similar to Garcia *et al.* (1996), Caccone and colleagues presented evidence for the monophyly of the 2La inversion in *An. gambiae* and *An. arabiensis* (Caccone *et al.*, 1998). However, Caccone *et al.* (1998) provided an interesting discussion on evidence suggesting that the 2La inversion in *An*. *merus* is of a second, independent origin. The authors discuss the possibility of introgressive hybridisation between *An. gambiae* and *An. arabiensis* and acknowledge the difficulty it provides in explaining the phylogenetic distribution of the 2La inversion.

A summary of the understanding of introgression between members of the *Anopheles gambiae* complex and how it relates to elucidating phylogenies, is provided by Powell *et al.* (1999); in which the authors collate previous results to construct Figure 1c.

The phylogenetic reconstruction of the medically important anophelines is now considered resolved and newly proposed reconstructions based on autosomal whole-genome sequence are not considered contentious (Fontaine *et al.*, 2015). The authors find that reconstruction based upon autosomal sequences supported the grouping of *An. gambiae s.s., An. coluzzii* and *An. arabiensis*, whole genome reconstruction was also in strong agreeance with this reconstruction. However, phylogenetic reconstruction based on X chromosome sequences place the *An. arabiensis* as more closely related to *An. quadriannulatus*, a vector with no role in malaria transmission. Despite concordance between autosomal and whole genome reconstructions grouping *An. gambiae s.s.*, *An. coluzzii* and *An.*

arabiensis, Fontaine *et al.* (2015) demonstrate that the historical species branching order is best explained by X chromosome reconstruction, and that pervasive introgression of autosomal loci explains why autosomal and whole genome reconstructions resolve *An. gambiae s.s.*, *An. coluzzii* and *An. arabiensis* are more closely related. Such findings support the notion that selectively advantageous haplotypes may be interspecifically shared between non-sister species.

Research concerning introgressive hybridisation in the *Anopheles gambiae s.l.* continued to reveal more of the convoluted and unexpected relationships between members of the complex. Indeed, F1 hybrids of *An. gambiae* and *An. arabiensis* with *An*. *bwambae* were identified during field collections in Uganda (Thelwell *et al.*, 2000). *An*. *bwambae* are a unique species, found only in western Uganda inhabiting a small area surrounding geothermal springs, and living in sympatry with *An. gambiae* and *An. arabiensis*. Phylogeny based on the work of Coluzzi *et al.* (1979) and then later Caccone *et al.* (1996) both support a sister taxa relationship between *An*. *bwambae* and *An*. *melas*. This presents an interesting situation wherein *An. gambiae* and *An*. *bwambae* are placed within different clades, but are still able to produce viable and fertile female F1 hybrids (Davidson and Hunt, 1973), suggesting not only that introgression between *An. gambiae* or *An. arabiensis* and *An. bwambae* is possible, but also adds to the complexity of the genetic and evolutionary relationships of *An. gambiae s.l.*

Evidence of Adaptive Introgression in the *Anopheles gambiae* Complex

Introgression has long been discussed as a potential route for the acquisition of adaptive traits, which may promote the vectorial capacity of Anophelines. Indeed, the 'knockdown resistance' (*kdr*) allele observed in *An. gambiae s.s.* in West and Central Africa was reported to be identified in *An. coluzzii*, and hypothesised to be the result of introgressive hybridisation, despite reproductive isolation between S and M 'form' being well established within the literature (Weill *et al.*, 2000; Chandre *et al.*, 1999; Favia *et al.*, 1997).

Weill *et al.* (2000) further discussion that the *kdr* allele may introgress into other members of the *Anopheles gambiae* complex, such as *An. arabiensis* and *An*. *bwambae*, which as discussed can hybridise with *An. gambiae s.s.* at low levels and shows evidence of a semipermeable species barrier (Thelwell *et al.*, 2000; Besansky *et al.*, 2003, Coluzzi *et al.*, 1979). Certainly, the discovery of *kdr* in the M 'form' of *An. gambiae* posed an opportunity for studying the genetic structure of the *Anopheles gambiae* complex and monitoring the spread of the then newly acquired adaptive allele through the M 'form' populations, across Africa. Diabate *et al.* (2003) conducted a survey in which the introgressed *kdr* allele in the M 'form' previously only found in Benin and Côte d'Ivoire, was identified in the Kou Village (VK7), Burkina Faso. The scarcity of M/S hybrids identified in collections had been previously hypothesised to be caused by partial reproductive isolation between the S and M molecular 'forms', with gene flow only occurring in particular geographical niches and/or seasons (Black and Lanzaro, 2001). Diabate *et al.*

(2003) highlights the proximity of Kou Valley to rice and cotton fields and the high density of M 'form' *An. gambiae* mixing with the S 'form' at the end of the rainy season, as an environment which may provide conducive settings for gene flow between these partially reproductively isolated chromosomal forms.

Interestingly, collection surveys revealed the Luc-Phe mutation, characteristic of the *kdr* phenotype observed in *An. gambiae* S and M 'forms' to be present in a population of *An. arabiensis* in Burkina Faso (Diabate *et al.*, 2004). However, elucidating whether the acquisition of the *kdr* allele by *An. arabiensis* was due to introgressive hybridisation or *de novo* mutation proved difficult. Diabate *et al.* (2004) present results which show fixation in nucleotide position 494 of the *kdr* allele in *An. gambiae s.s.* but reveals polymorphism at the same point in *An. arabiensis*. This, combined with authors citation of the Lue replacement mutations being the most common and the patchy distribution of *kdr*-arabiensis, suggests that the *kdr* allele in *An. arabiensis* arose because of de novo mutation (Diabate *et al.*, 2004). A single *An. arabiensis* in western Kenya was found to have the *kdr* allele, however, further analysis into the frequency of the mutation in the region and whether it likely represented an introgression event or de novo mutation, was not conducted at the time. A subsequent report by Ochomo *et al.*, (2015) showed the alleles to be in low frequency in *An. arabiensis* and high frequency in *An. gambiae*, despite being lower than previously reported.

Etang *et al.* (2009), during a study to identify the varying mutations which originated the *kdr* phenotype, identified *kdr*-originating alleles that are shared

between the S and M molecular form, suggestive of introgression. Interestingly, no hybrid individuals were identified during the 7-month collection period across seven collections sites in Cameroon, including areas of sympatry.

Retention of introgressed chromosomal regions between *An. gambiae* and *An. arabiensis* revealed that, contrary to other chromosomal elements, introgression of *kdr* into *An. arabiensis* is not affected by negative selection for the locus which contains it (Slotman *et al.*, 2005b). However, this result was achieved in laboratory strains of mosquitoes and indeed, selection against introgressed genetic elements may be modulated by field conditions.

Although *Vsgc-kdr* mutations are adaptive and confer resistance to insecticidal interventions, Mitri *et al.* (2015) show the associated fitness cost of such mutation as an increased susceptibility to plasmodium parasites. The impact of fitness costs for adaptive mutations is a consideration that must be made, as it can impact heavily on control programme strategies and indeed inform the refinement of the putative understanding of gene function.

The well-studied and characterised *ace-1*^{R} allele has also been the focus of research into the introgressive events surrounding the *An. gambiae* complex. The *ace-1*^{R} was identified in both *An. gambiae s.s.* and *An. coluzzii* (Weill *et al.*, 2004), and the consensus at the time was that S and M were incipient species. Djogbenou *et al.* (2008) therefore aimed to establish whether the G119S mutation – which underlies the *ace-1*^{R} allele – arose independently in each segregated mating chromosomal form or as through introgressive hybridisation. The authors' results

reveal that of the polymorphisms located in the exonic and upstream/downstream intronic regions of both S and M molecular forms, no identified diversity was associated with the G119S mutation. The authours discuss the extremely low probability of the same mutation arising independently but acknowledge the possibility of the G119S mutation being a shared ancesteral polymorphism, although unlikely due to this hypothesis requiring that the G119S mutation predates S and M differentiation and was maintained in both forms with no new polymorphisms arising. However, it is worth noting that two Anopheline species ariving both independently at the same mutation is not unprecendented.

During a study to identify and characterise insecticide resistance in malaria vector populations of Uganda, Mawejje *et al.* (2013) identified a small number (0.22%) of *An. gambiae s.s.* and *An. arabiensis* hybrids. This prompted consideration that introgression between these populations may be occurring. Although, Mawejje and colleagues identified that the study sequence used was insufficient for determining whether the low frequency *kdr 1014S* allele in *An. arabiensis* was the result of introgression of *de novo* mutation.

In a subsequent study, hybrid individuals identified by Mawejje *et al.* (2013) from Jinja and additional hybrid samples from Tororo were investigated. Weetman *et al.* (2014) utilising a custom array to genotype 1536 SNPs associated with insecticide resistance and utilising an F_{ST}-based approach, identified the occurrence of introgression between *An. gambiae s.s.* and *An. arabiensis*, in both the Tororo and Jinja populations. The authors also present evidence that most

hybrid individuals tested from eastern Uganda are beyond the F1 stage, suggesting that backcrossing has likely occurred. This offers a route for gene flow between the species to reach fixation, suggesting that the current ecological and phylogenetic conditions are sometimes amenable to introgression (Weetman *et al.*, 2014).

Promisingly, the previously difficult to address hypothesis of introgression and evolutionary history and the phylogenetic structure of medically important Anophelines is becoming easier. For example, Weetman and Clarkson (2015) outline the significance of newly developed genome assemblies for the multiple malaria vectors and illustrate this by comparing the relatively high alignment percentage of African vectors to the PEST reference genomes and the low alignment percentage for the major vector of south and central America. The 16 genomes consortium addressed this issue when Neafsey *et al.* (2015) published reference genomes for 16 major Anopheline vectors. The work by Neafsey *et al.* (2015) is used in this thesis and provides Anopheline outgroups for ancestry informative marker generation and Patterson's D statistic estimation.

Interestingly, Norris *et al.* (2015) present evidence that the 2L divergence island in *An. gambiae s.s.* introgressed into *An. coluzzii*, including the complement of insecticide resistance alleles found on that haplotype. This introgression event coincided with a large scale LLIN distribution programme in Mali, leading the authors to hypothesise that anthropogenic pressures changed the fitness landscape, ultimately favouring *An. gambiae s.s./An. coluzzii* hybrid (due to the

suite of resistance alleles on the 2L haplotype). This presents an interesting mechanism which if validated, could explain how introgression can occur and be maintained in a recipient species after initial hybridisation events.

Indeed, developing an understanding of the conditions which facilitate fertile hybrid formation represents a significant research goal; since the mechanisms which explain partial reproductive isolation are not fully characterised and the frequency of hybridisation rates across multiple sympatric ranges vary significantly. However, studies have identified a breakdown in assortative mating and spatial swarm segregation as a condition for hybridisation in *An. gambiae s.s.* and *An. coluzzii*, although the causes of such breakdowns remain elusive (Aboagye-Antwi *et al.*, 2015; Niang *et al.*, 2015).

Identifying adaptive introgression events need not be exclusively limited to protection against anthropogenic factors such as insecticides; indeed, identifying the acquisition of alleles capable of hindering or promoting mosquito immunity or refractoriness to malaria is of key importance. *Thioester-containing protein 1* (*TEP1*) is one of several immune response genes in anophelines which transcribes a complement-like opsonin, which can facilitate the identification and phagocytotic killing of protozoa and gram-negative bacteria (Blandin *et al.*, 2009). Mancini *et al.* (2015) discuss an interesting observation that *An. gambiae s.s.* have acquired *TEP1* through introgression from *An. coluzzii* in a hybrid zone in Guinea, with the allele not observed at fixation in either species – contrary to *TEP1* being found exclusively in *An. coluzzii* and at fixation in Mali and Burkina Faso. The authors

were circumspect about concluding that the absence of fixation indicates a lower overall fitness benefit to *An. gambiae s.s.* and acknowledge that further observations are required to understand the mechanisms at play.

Such introgressive events are not limited to either the African continent or the *Anopheles gambiae* complex. Cornel *et al.* (2003) discussed the extent of introgressive hybridisation *Culex pipiens* and *Culex quinquefasciatus* in areas of sympatry in California, USA and South Africa. Cornel and colleagues identified a near complete permeability to reproductive barriers in California species, and discussed the findings significance for potentially influencing vector control measures in response to concerns of West Nile Virus in North America (Cornel *et al.*, 2003). Similarly, the *Anopheles dirus* complex, found in Asia, has been the focus of studies aimed at elucidating the role of introgression in incipient speciation and adaptation (Walton *et al.*, 2001).

Overall, the two principle questions we seek to address are (1) What is the extent of contemporary introgression between *An. gambiae* and *An. arabiensis* in areas of sympatry and (2) Is introgression a significant route for the acquisition of selectively advantageous traits, notably insecticide resistance?

References

Aboagye-Antwi, F., Alhafez, N., Weedall, G. D., Brothwood, J., Kandola, S., Paton, D., Fofana, A., Olohan, L., Betancourth, M. P., Ekechukwu, N. E., Baeshen, R., Traore, S. F., Diabate, A. & Tripet, F. 2015. Experimental swap of Anopheles gambiae's assortative mating preferences demonstrates key role of X-chromosome divergence island in incipient sympatric speciation. *PLoS Genet,* 11**,** e1005141.

Baldini, F., Segata, N., Pompon, J., Marcenac, P., Shaw, W. R., Dabire, R. K., Diabate, A., Levashina, E. A. & Catteruccia, F. 2014. Evidence of natural Wolbachia infections in field populations of *Anopheles gambiae*. *Nat Commun,* 5**,** 3985.

Besansky, N. J., Powell, J. R., Caccone, A., Hamm, D. M., Scott, J. A. & Collins, F. H. 1994. Molecular phylogeny of the *Anopheles gambiae* complex suggests genetic introgression between principal malaria vectors. *Proc Natl Acad Sci U S A,* 91**,** 6885-8.

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briet, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E. & Gething, P. W. 2015. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature,* 526**,** 207-211.

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

Blandin, S. A., Wang-Sattler, R., Lamacchia, M., Gagneur, J., Lycett, G., Ning, Y., Levashina, E. A. & Steinmetz, L. M. 2009. Dissecting the genetic basis of resistance to malaria parasites in *Anopheles gambiae*. *Science,* 326**,** 147-50.

Caccone, A., Garcia, B. A. & Powell, J. R. 1996. Evolution of the mitochondrial DNA control region in the *Anopheles gambiae* complex. *Insect Mol Biol,* 5**,** 51-9.

Caccone, A., Min, G. S. & Powell, J. R. 1998. Multiple origins of cytologically identical chromosome inversions in the *Anopheles gambiae* complex. *Genetics,* 150**,** 807-14.

Chandre, F., Manguin, S., Brengues, C., Dossou Yovo, J., Darriet, F., Diabate, A., Carnevale, P. & Guillet, P. 1999. Current distribution of a pyrethroid resistance gene (*kdr*) in *Anopheles gambiae* complex from west Africa and further evidence for reproductive isolation of the Mopti form. *Parassitologia,* 41**,** 319-22.

Clarkson, C. S., Weetman, D., Essandoh, J., Yawson, A. E., Maslen, G., Manske, M., Field, S. G., Webster, M., Antao, T., MacInnis, B., Kwiatkowski, D. & Donnelly, M. J. 2014. Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. *Nat Commun,* 5**,** 4248.

Coluzzi, M., Sabatini, A., Petrarca, V. & Di Deco, M. A. 1979. Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. *Trans R Soc Trop Med Hyg,* 73**,** 483-97.

Cornel, A. J., McAbee, R. D., Rasgon, J., Stanich, M. A., Scott, T. W. & Coetzee, M. 2003. Differences in extent of genetic introgression between sympatric Culex pipiens and Culex quinquefasciatus (Diptera: Culicidae) in California and South Africa. *J Med Entomol,* 40**,** 36-51.

Davidson, G. & Hunt, R. H. 1973. The crossing and chromosome characteristics of a new, sixth species in the *Anopheles gambiae* complex. *Parassitologia,* 15**,** 121-8.

Diabate, A., Baldet, T., Chandre, C., Dabire, K. R., Kengne, P., Guiguemde, T. R., Simard, F., Guillet, P., Hemingway, J. & Hougard, J. M. 2003. *KDR* mutation, a genetic marker to assess events of introgression between the molecular M and S forms of *Anopheles gambiae* (Diptera: Culicidae) in the tropical savannah area of West Africa. *J Med Entomol,* 40**,** 195-8.

Diabate, A., Brengues, C., Baldet, T., Dabire, K. R., Hougard, J. M., Akogbeto, M., Kengne, P., Simard, F., Guillet, P., Hemingway, J. & Chandre, F. 2004. The spread of the Leu-Phe *kdr* mutation through *Anopheles gambiae* complex in Burkina Faso: genetic introgression and de novo phenomena. *Trop Med Int Health,* 9**,** 1267-73.

Djogbenou, L., Chandre, F., Berthomieu, A., Dabire, R., Koffi, A., Alout, H. & Weill, M. 2008. Evidence of introgression of the ace-1(R) mutation and of the ace-1 duplication in West African *Anopheles gambiae* s. s. *PLoS One,* 3**,** e2172.

Donnelly, M. J., Isaacs, A. T. & Weetman, D. 2016. Identification, Validation, and Application of Molecular Diagnostics for Insecticide Resistance in Malaria Vectors. *Trends Parasitol,* 32**,** 197-206.

Donnelly, M. J., Pinto, J., Girod, R., Besansky, N. J. & Lehmann, T. 2004. Revisiting the role of introgression vs shared ancestral polymorphisms as key processes shaping genetic diversity in the recently separated sibling species of the *Anopheles gambiae* complex. *Heredity (Edinb),* 92**,** 61-8.

Etang, J., Vicente, J. L., Nwane, P., Chouaibou, M., Morlais, I., Do Rosario, V. E., Simard, F., Awono-Ambene, P., Toto, J. C. & Pinto, J. 2009. Polymorphism of intron-1 in the voltage-gated sodium channel gene of *Anopheles gambiae* s.s. populations from Cameroon with emphasis on insecticide knockdown resistance mutations. *Mol Ecol,* 18**,** 3076-86.

Favia, G., della Torre, A., Bagayoko, M., Lanfrancotti, A., Sagnon, N., Toure, Y. T. & Coluzzi, M. 1997. Molecular identification of sympatric chromosomal forms of *Anopheles gambiae* and further evidence of their reproductive isolation. *Insect Mol Biol,* 6**,** 377-83.

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y. C., Smith, H. A., Love, R. R., Lawniczak, M. K., Slotman, M. A., Emrich,

S. J., Hahn, M. W. & Besansky, N. J. 2015. Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science,* 347**,** 1258524.

Garcia, B. A., Caccone, A., Mathiopoulos, K. D. & Powell, J. R. 1996. Inversion monophyly in African anopheline malaria vectors. *Genetics,* 143**,** 1313-20.

Gleave, K., Lissenden, N., Richardson, M., Choi, L. & Ranson, H. 2018. Piperonyl butoxide (PBO) combined with pyrethroids in insecticide-treated nets to prevent malaria in Africa. *Cochrane Database Syst Rev,* 11**,** CD012776.

Hanemaaijer, M. J., Houston, P. D., Collier, T. C., Norris, L. C., Fofana, A., Lanzaro, G. C., Cornel, A. J. & Lee, Y. 2018. Mitochondrial genomes of *Anopheles arabiensis*, *An. gambiae* and *An. coluzzii* show no clear species division. *F1000Res,* 7**,** 347.

Hay, S. I., Sinka, M. E., Okara, R. M., Kabaria, C. W., Mbithi, P. M., Tago, C. C., Benz, D., Gething, P. W., Howes, R. E., Patil, A. P., Temperley, W. H., Bangs, M. J., Chareonviriyaphap, T., Elyazar, I. R., Harbach, R. E., Hemingway, J., Manguin, S., Mbogo, C. M., Rubio-Palis, Y. & Godfray, H. C. 2010. Developing global maps of the dominant anopheles vectors of human malaria. *PLoS Med,* 7**,** e1000209.

Lee, Y., Marsden, C. D., Nieman, C. & Lanzaro, G. C. 2014. A new multiplex SNP genotyping assay for detecting hybridization and introgression between the M and S molecular forms of *Anopheles gambiae*. *Mol Ecol Resour,* 14**,** 297-305.

Mancini, E., Spinaci, M. I., Gordicho, V., Caputo, B., Pombi, M., Vicente, J. L., Dinis, J., Rodrigues, A., Petrarca, V., Weetman, D., Pinto, J. & Della Torre, A. 2015. Adaptive Potential of Hybridization among Malaria Vectors: Introgression at the Immune Locus TEP1 between Anopheles coluzzii and A. gambiae in 'Far-West' Africa. *PLoS One,* 10**,** e0127804.

Marsden, C. D., Lee, Y., Nieman, C. C., Sanford, M. R., Dinis, J., Martins, C., Rodrigues, A., Cornel, A. J. & Lanzaro, G. C. 2011. Asymmetric introgression between the M and S forms of the malaria vector, *Anopheles gambiae*, maintains divergence despite extensive hybridization. *Mol Ecol,* 20**,** 4983-94.

Mawejje, H. D., Wilding, C. S., Rippon, E. J., Hughes, A., Weetman, D. & Donnelly, M. J. 2013. Insecticide resistance monitoring of field-collected *Anopheles gambiae* s.l. populations from Jinja, eastern Uganda, identifies high levels of pyrethroid resistance. *Med Vet Entomol,* 27**,** 276-83.

Mitri, C., Markianos, K., Guelbeogo, W. M., Bischoff, E., Gneme, A., Eiglmeier, K., Holm, I., Sagnon, N., Vernick, K. D. & Riehle, M. M. 2015. The *kdr*-bearing haplotype and susceptibility to Plasmodium falciparum in *Anopheles gambiae*: genetic correlation and functional testing. *Malar J,* 14**,** 391.

Murdock, C. C., Blanford, S., Hughes, G. L., Rasgon, J. L. & Thomas, M. B. 2014. Temperature alters Plasmodium blocking by Wolbachia. *Sci Rep,* 4**,** 3932.

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arca, B., Arensburger, P., Artemov, G., Assour, L. A., Basseri, H., Berlin, A., Birren, B. W., Blandin, S. A., Brockman, A. I.,

Burkot, T. R., Burt, A., Chan, C. S., Chauve, C., Chiu, J. C., Christensen, M., Costantini, C., Davidson, V. L., Deligianni, E., Dottorini, T., Dritsou, V., Gabriel, S. B., Guelbeogo, W. M., Hall, A. B., Han, M. V., Hlaing, T., Hughes, D. S., Jenkins, A. M., Jiang, X., Jungreis, I., Kakani, E. G., Kamali, M., Kemppainen, P., Kennedy, R. C., Kirmitzoglou, I. K., Koekemoer, L. L., Laban, N., Langridge, N., Lawniczak, M. K., Lirakis, M., Lobo, N. F., Lowy, E., MacCallum, R. M., Mao, C., Maslen, G., Mbogo, C., McCarthy, J., Michel, K., Mitchell, S. N., Moore, W., Murphy, K. A., Naumenko, A. N., Nolan, T., Novoa, E. M., O'Loughlin, S., Oringanje, C., Oshaghi, M. A., Pakpour, N., Papathanos, P. A., Peery, A. N., Povelones, M., Prakash, A., Price, D. P., Rajaraman, A., Reimer, L. J., Rinker, D. C., Rokas, A., Russell, T. L., Sagnon, N., Sharakhova, M. V., Shea, T., Simao, F. A., Simard, F., Slotman, M. A., Somboon, P., Stegniy, V., Struchiner, C. J., Thomas, G. W., Tojo, M., Topalis, P., Tubio, J. M., Unger, M. F., Vontas, J., Walton, C., Wilding, C. S., Willis, J. H., Wu, Y. C., Yan, G., Zdobnov, E. M., Zhou, X., Catteruccia, F., Christophides, G. K., Collins, F. H., Cornman, R. S., *et al.* 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science,* 347**,** 1258522.

Niang, A., Epopa, P. S., Sawadogo, S. P., Maiga, H., Konate, L., Faye, O., Dabire, R. K., Tripet, F. & Diabate, A. 2015. Does extreme asymmetric dominance promote hybridization between Anopheles coluzzii and *Anopheles gambiae* s.s. in seasonal malaria mosquito communities of West Africa? *Parasit Vectors,* 8**,** 586.

Noor, M. A. & Bennett, S. M. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity (Edinb),* 103**,** 439-44.

Norris, L. C., Main, B. J., Lee, Y., Collier, T. C., Fofana, A., Cornel, A. J. & Lanzaro, G. C. 2015. Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *Proc Natl Acad Sci U S A,* 112**,** 815-20.

Ochomo, E., Subramaniam, K., Kemei, B., Rippon, E., Bayoh, N. M., Kamau, L., Atieli, F., Vulule, J. M., Ouma, C., Gimnig, J., Donnelly, M. J. & Mbogo, C. 2015. Presence of the knockdown resistance mutation, Vgsc-1014F in Anopheles gambiae and *An. arabiensis* in western Kenya. *Parasit Vectors,* 8**,** 616.

Powell, J. R., Petrarca, V., della Torre, A., Caccone, A. & Coluzzi, M. 1999. Population structure, speciation, and introgression in the *Anopheles gambiae* complex. *Parassitologia,* 41**,** 101-13.

Rosenzweig, B. K., Pease, J. B., Besansky, N. J. & Hahn, M. W. 2016. Powerful methods for detecting introgressed regions from population genomic data. *Mol Ecol,* 25**,** 2387-97.

Service, M. W. & Townson, H. 2002. The Anopheles vector. *In:* Gilles, H. M. & Warrell, D. A. (eds.) *Essential Malariolgy.* Fourth ed. London: Arnold. 59-84

Shaw, W. R., Marcenac, P., Childs, L. M., Buckee, C. O., Baldini, F., Sawadogo, S. P., Dabire, R. K., Diabate, A. & Catteruccia, F. 2016. Wolbachia

infections in natural Anopheles populations affect egg laying and negatively correlate with Plasmodium development. *Nat Commun,* 7**,** 11772.

Slotman, M., Della Torre, A. & Powell, J. R. 2005a. Female sterility in hybrids between *Anopheles gambiae* and A. arabiensis, and the causes of Haldane's rule. *Evolution,* 59**,** 1016-26.

Slotman, M. A., Della Torre, A., Calzetta, M. & Powell, J. R. 2005b. Differential introgression of chromsomal regions between *Anopheles gambiae* and *An. arabiensis*. *Am J Trop Med Hyg,* 73**,** 326-35.

Thelwell, N. J., Huisman, R. A., Harbach, R. E. & Butlin, R. K. 2000. Evidence for mitochondrial introgression between Anopheles bwambae and *Anopheles gambiae*. *Insect Mol Biol,* 9**,** 203-10.

Walton, C., Handley, J. M., Collins, F. H., Baimai, V., Harbach, R. E., Deesin, V. & Butlin, R. K. 2001. Genetic population structure and introgression in Anopheles dirus mosquitoes in South-east Asia. *Mol Ecol,* 10**,** 569-80.

Weetman, D. & Clarkson, C. S. 2015. Evolving the world's most dangerous animal. *Trends Parasitol,* 31**,** 39-40.

Weetman, D. & Donnelly, M. J. 2015. Evolution of insecticide resistance diagnostics in malaria vectors. *Trans R Soc Trop Med Hyg,* 109**,** 291-3.

Weetman, D., Steen, K., Rippon, E. J., Mawejje, H. D., Donnelly, M. J. & Wilding, C. S. 2014. Contemporary gene flow between wild *An. gambiae s.s.* and *An. arabiensis*. *Parasit Vectors,* 7**,** 345.

Weetman, D., Wilding, C. S., Steen, K., Pinto, J. & Donnelly, M. J. 2012. Gene flow-dependent genomic divergence between *Anopheles gambiae* M and S forms. *Mol Biol Evol,* 29**,** 279-91.

Weill, M., Chandre, F., Brengues, C., Manguin, S., Akogbeto, M., Pasteur, N., Guillet, P. & Raymond, M. 2000. The *kdr* mutation occurs in the Mopti form of *Anopheles gambiae* s.s. through introgression. *Insect Mol Biol,* 9**,** 451-5.

Weill, M., Malcolm, C., Chandre, F., Mogensen, K., Berthomieu, A., Marquine, M. & Raymond, M. 2004. The unique mutation in ace-1 giving high insecticide resistance is easily detectable in mosquito vectors. *Insect Mol Biol,* 13**,** 1-7.

Weir, B. S. & Cockerham, C. C. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution,* 38**,** 1358-1370.

Wen, D., Yu, Y., Hahn, M. W. & Nakhleh, L. 2016. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Mol Ecol,* 25**,** 2361-72.

White, G. B. 1971. Chromosomal evidence for natural interspecific hybridization by mosquitoes of the *Anopheles gambiae* complex. *Nature,* 231**,** 184-5.

WHO 2016. World Malaria Report 2016. Geneva, Switzerland.

Wilding, C. S., Weetman, D., Rippon, E. J., Steen, K., Mawejje, H. D., Barsukov, I. & Donnelly, M. J. 2015. Parallel evolution or purifying selection, not

introgression, explains similarity in the pyrethroid detoxification linked GSTE4 of *Anopheles gambiae* and *An. arabiensis*. *Mol Genet Genomics,* 290**,** 201-15.

Chapter II.

Data Quality Control and Assurance

This chapter addresses methods of preparation, cleaning and quality control analysis of whole genome sequence data and consequences of non-specific alignment.

Methods of Preparation, cleaning and quality control analysis of whole genome sequence data and consequences on non-specific alignment.

Introduction

The data used for this project forms part of the *Anopheles gambiae* 1000 genomes (*Anopheles gambiae* 1000 Genomes Consortium *et al.*, 2017). The Ag1000G project is a collaborative consortium that provides a high-resolution view of the genetic variation in African Anophelines, predominantly *An. gambiae*. The consortium uses Illumina high-throughput whole genome deep sequencing and releases data in phases. The Ag1000G project has three principal objectives:

- “Discovering natural genetic variation – We're using high-throughput sequencing of a large number of wild-caught mosquitoes sampled from across Africa to build a comprehensive catalogue of genetic variation in natural vector populations. Our primary focus is on *An. gambiae sensu strictu* and *An. coluzzii*, but we will be expanding to include *An. arabiensis* in the future.

- Describing the structure and history of vector populations – We are analysing genetic variation data to characterise key features of natural vector populations, such as patterns of diversity, linkage disequilibrium and recombination, population structure and gene flow, signals of recent selection, and demographic history.
- Connecting genetic variation and population biology with ecology and malaria epidemiology – We aim to study associations between genotype and broad phenotypes such as ecological specialisation and differences in local malaria epidemiology."

\- The Ag1000G Project

This PhD project utilises the third phase of data released by the consortium. Phase 3 contains sympatric pairs of *An. gambiae* and *An. arabiensis* which were used to interrogate the genomes for signatures of interspecific gene flow, in addition to samples without a sympatric partner.

Samples provided to the Ag1000G project are sequenced at the Wellcome Sanger Institute using the *VRPipe* sequencing workflow (Github/ Vertebrate Resequencing/ vr-pipe). Prior to the data being used for analyses a series of stages concerning data cleaning, quality control and preparation must be carried out. In this chapter, I detail the methods and results of this process for the phase 3 data, and the appropriateness of aligning *An. arabiensis* samples to the *An. gambiae* AgamP4 reference genome for introgression studies.

Data

The dataset used here consists of 833 samples. They samples are from various regions within Uganda, Tanzania and Malawi (Figure 2). The distribution breakdown of samples is given in Table 1. One sample from Uganda has no data for district and at this stage is labelled at 'Uganda Unknown'. The dataset also contains 124 crosses from Dongola and Sennar, Sudan.

Table 1. Breakdown of population sample count. Outliers are individuals samples which did not cluster typically in PCA analyses.

Population	Collection Site	Number of samples
Tanzania-MUL-arabiensis	Muleba	118
Uganda-TOR-gambiae	Tororo	112
Uganda-KAN-gambiae	Kanugu	94
Uganda-TOR-arabiensis	Tororo	76
Tanzania-TAR-arabiensis	Tarime	47
Tanzania-MOS-arabiensis	Moshi	39
Malawian-CHI-arabiensis	Chikwawa	33
Tanzania-MUH-gambiae	Muheza	32
Tanzania-MUL-gambiae	Muleba	32
Tanzania-MUH-gambiae-outliers	Muheza	4
Tanzania-MUL-gambiae-outliers	Muleba	1

Uganda-KAN-arabiensis	Kanugu	1

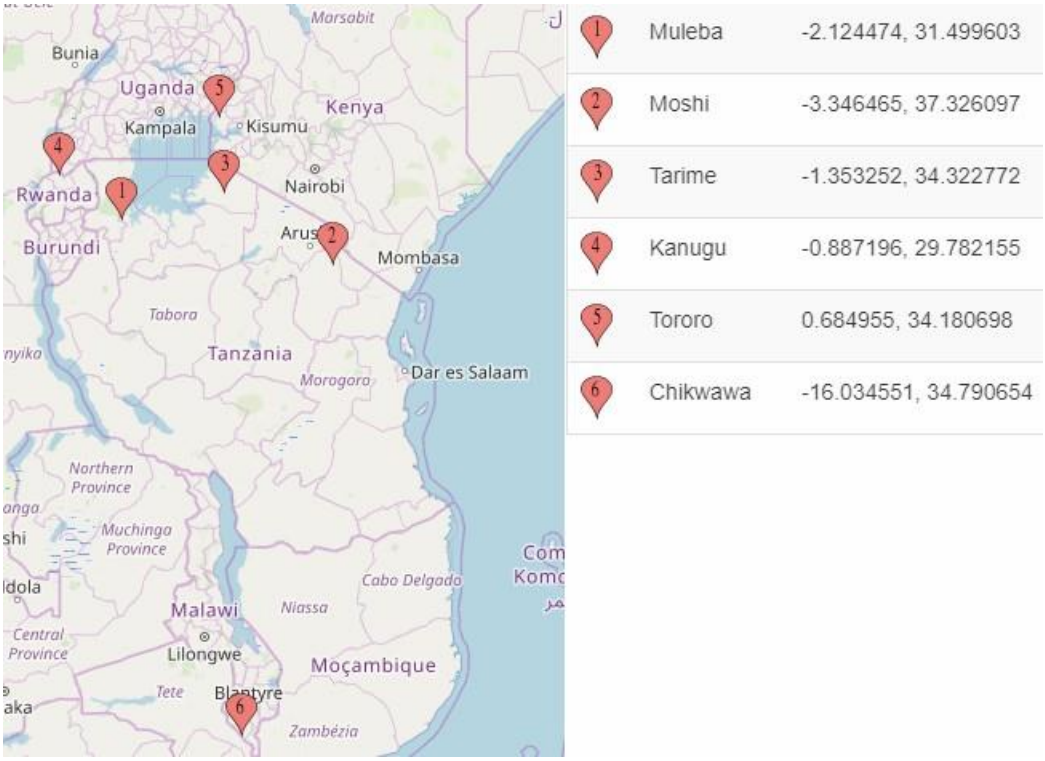


Figure 1. Origins of sample used in these analyses. For convenience, each country and district has been assigned a contraction: TZ – Tanzania, UG – Uganda, MW – Malawi, TAR – Tarime, MUH – Muheza, TOR – Tororo, MUL – Muleba, KAN – Kanugu, CHI – Chikwawa, MOS – Moshi.

Sequencing

Data used within the presented analyses were generated by the Ag1000G project data production team, following protocols previously published (*Anopheles gambiae* 1000 Genomes Consortium *et al.*, 2017). An overview of the data generation process is given in this section.

Sample sequence reads were aligned to the AgamP3 (PEST) reference genome using the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009).

Duplicate reads were identified and flagged using Picard Tools. Single nucleotide polymorphism (SNP) and insertion/deletion (indel) variants were called using the Genome Analysis Toolkit (GATK) following published best practices, using sequence alignments (McKenna *et al.*, 2010).

Meta Data Exploration

An exploratory analysis of the meta data associated with phase 3 revealed a number of discrepancies, including:

- Malawian samples being labelled as Malian
- Absent data being represented with '?' instead of NA
- Inconsistency in the reporting of collection village, ie 'Kanungu Uganda' instead of 'Kanungu'
- Errant hyphens where underscores were needed to be compatible with previous Ag1000G convention and scripts.
- No species ID for colony crosses
- Incorrect species labelling for samples from Malawi. All submitted samples are *An. arabiensis* but were labelled as *An. gambiae*.

All metadata inconsistencies and errors were addressed using a python script, which is maintained under version control to ensure that changes to metadata can be verified.

Coverage Analysis

At the time of analyses, a full reference assembly for the *An. arabiensis* genome was not available despite being in an advanced stage of development.

Therefore, we aligned all individual *An. arabiensis* samples to the *An. gambiae* reference AgamP4, as the purpose of this study is to identify regions of similarity between species and not features private to either *An. gambiae* or *An. arabiensis*. Post alignment quality control steps relating to coverage were carried out to ensure that mapping *An. arabiensis* to *An. gambiae* did not cause an unexpected divergence in key metrics that relate to the quality of sequence data, which may hinder downstream analyses. Such metrics include depth, the proportion of sites called, error rate, read count, read alignment proportion.

The sequencing depth or coverage for a given genomic position refers to the number of raw sequencing reads that align at that position. Depth is an informative measure for the quality of sequencing output and the amount of data a sample can provide to the dataset. For example, in cases where an individual has a rare SNP, high coverage is needed to be able to confidently call the rare SNP call as true, and not a sequencing error. Confidence in calls scales with the amount of coverage at a given position.

X

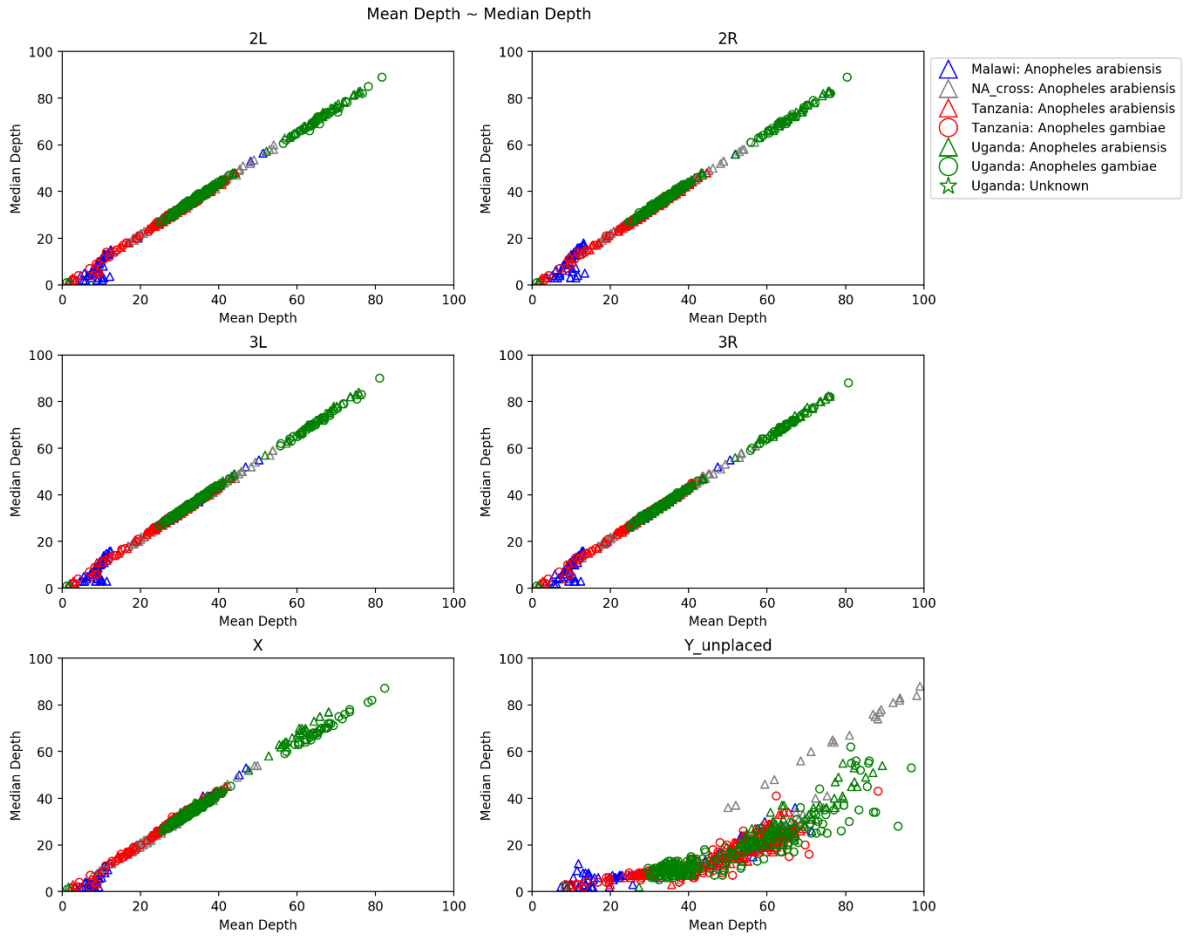


Figure 2. Mean Depth vs Median Depth for each Population by Chromosome. The depth for each sample is used in further analyses as a proxy sequencing quality. This analysis was performed to assess whether to use mean or median depth.

Figure 1 shows many samples with low mean depth, this means that the total number of called sites within a given sample is low, as a result of poor or incomplete sequencing. Figure 2 shows the relationship between mean depth and the percentage of sites called. Across all autosomes for all samples in phase 3, most of the data have percentage of sites called of >70% and a depth of >10.

Following previous conventions of the Ag1000G, samples that fall below these values were marked as failing the quality control stage. We use mean depth throughout these QC analyses, however we visualised the difference between the mean and median depth of samples to ensure a congruity between metrics, so as not to bias the samples that get filtered out. We find a high congruence between mean and median depth across the genomes of all our samples for all autosomes and the X chromosome (Figure 1). The Y/unplaced chromosome shows a skew towards a mean depth, which can result from a large number of the reads being assigned to a single region of the genome, compared to evenly distributed across the genome, as observed with the autosomes and X chromosome. The analysis steps that follow QC do not use the Y unplaced chromosome data. From this we are confident whilst we can proceed using either mean or median depth without biasing results, we use mean depth.

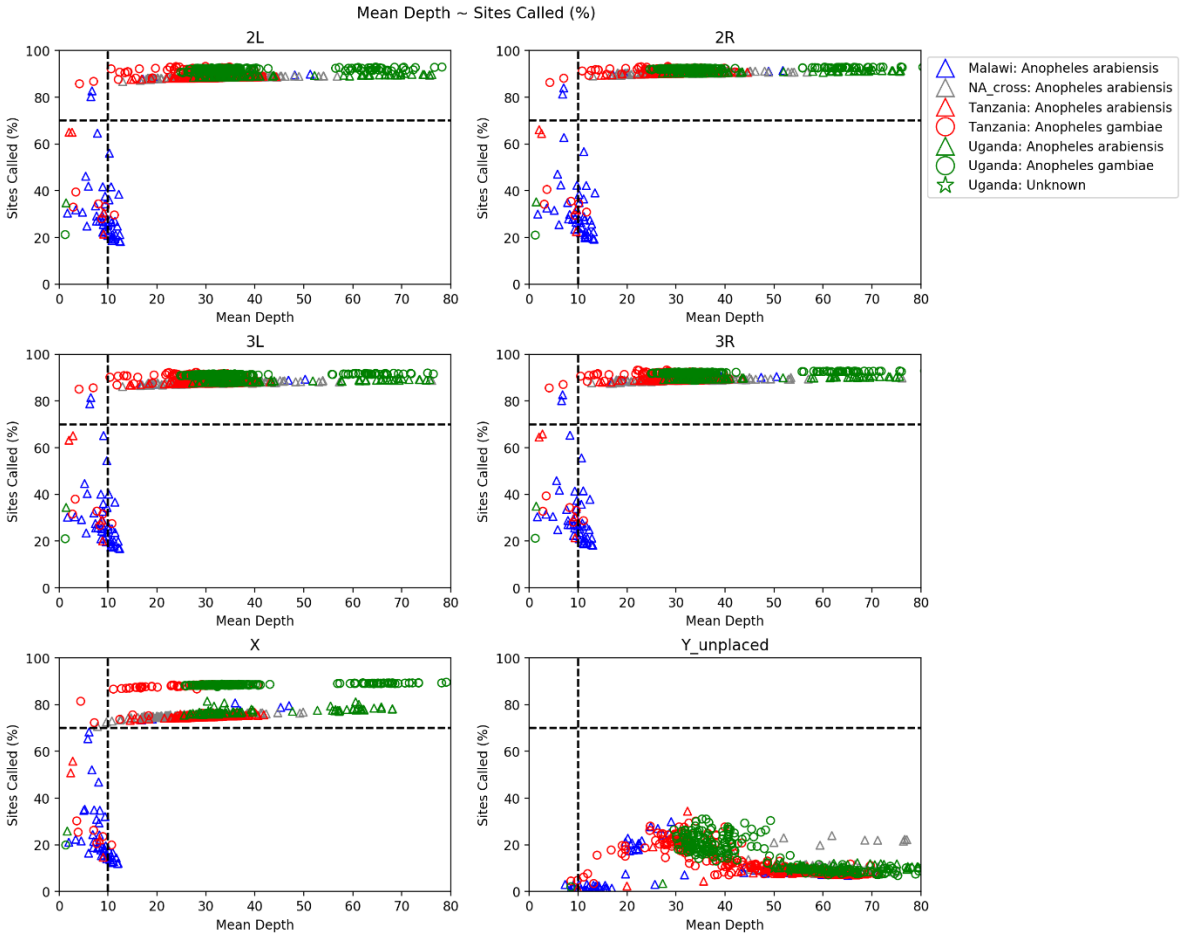


Figure 3. Mean depth vs sites called (%) for each population by chromosome. The relationship between sites called and mean depth follows the expected trend that most samples have a high proportion of sites called and a mean depth of at least 10 for the majority of samples.

To investigate a probable cause for a population-wide deficit in percentage of sites called in Malawian samples, further summary statistics were calculated. Error rate is the proportion of reads that contain errors, for example an error rate of 0.8% indicates for every 1000 reads, 8 are reported with some error by the GATK aligner. The error rate of *An. arabiensis* samples from Malawi was observed to have an error rate similar to that of *An. gambiae* and *An. arabiensis* from Tanzania (Figure 3). Given the similarity of error rate between failed Malawian samples and passed Tanzanian samples, another factor must be driving such a

low percentage of sites being called and depth. The error rate for samples that pass the 70% of sites called filter have low error rates and no trend of higher error rate for *An. arabiensis* samples over *An. gambiae* samples.

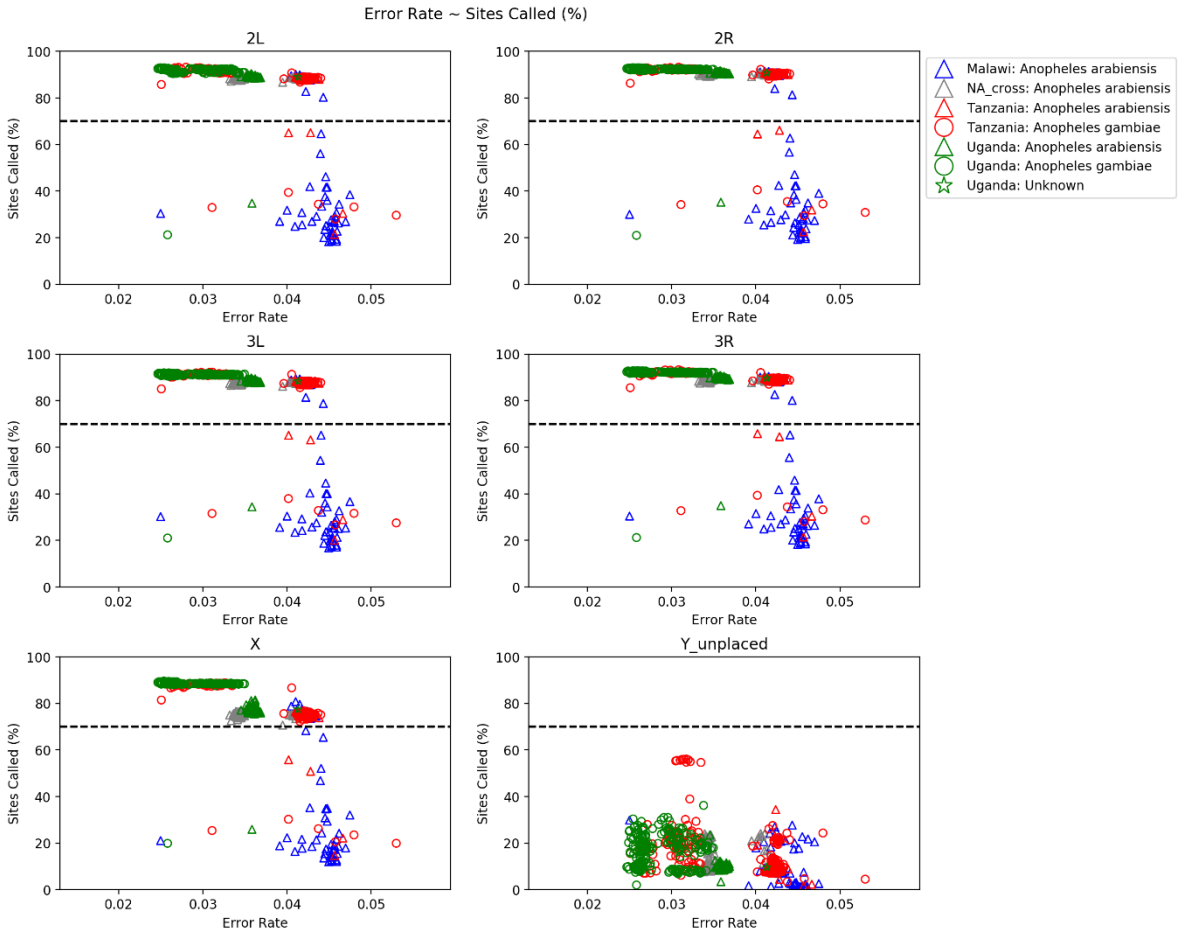


Figure 4. Error rate vs sites called (%) for each population by chromosome. The expectation is that samples have a low error rate and a high proportion of sites called. All samples have an error rate less than 0.05 except one.

The two distinct clusters of samples above the dotted line appear to be driven by population level effects, likely driven by sample preparation and not heterospecificity of the reference to which they are aligned. This is further supported by the grouping by collection location rather than species. Indeed, when visualising the raw count of reads per sample against the percentage of sites

called, no species-specific effect on read count can be observed. Further, the number of raw reads for the failed Malawian samples are similar to that of passed samples from Tanzania and Uganda for both species, again suggesting species is not the driving factor Malawian samples having <70% sites called (Figure 4).

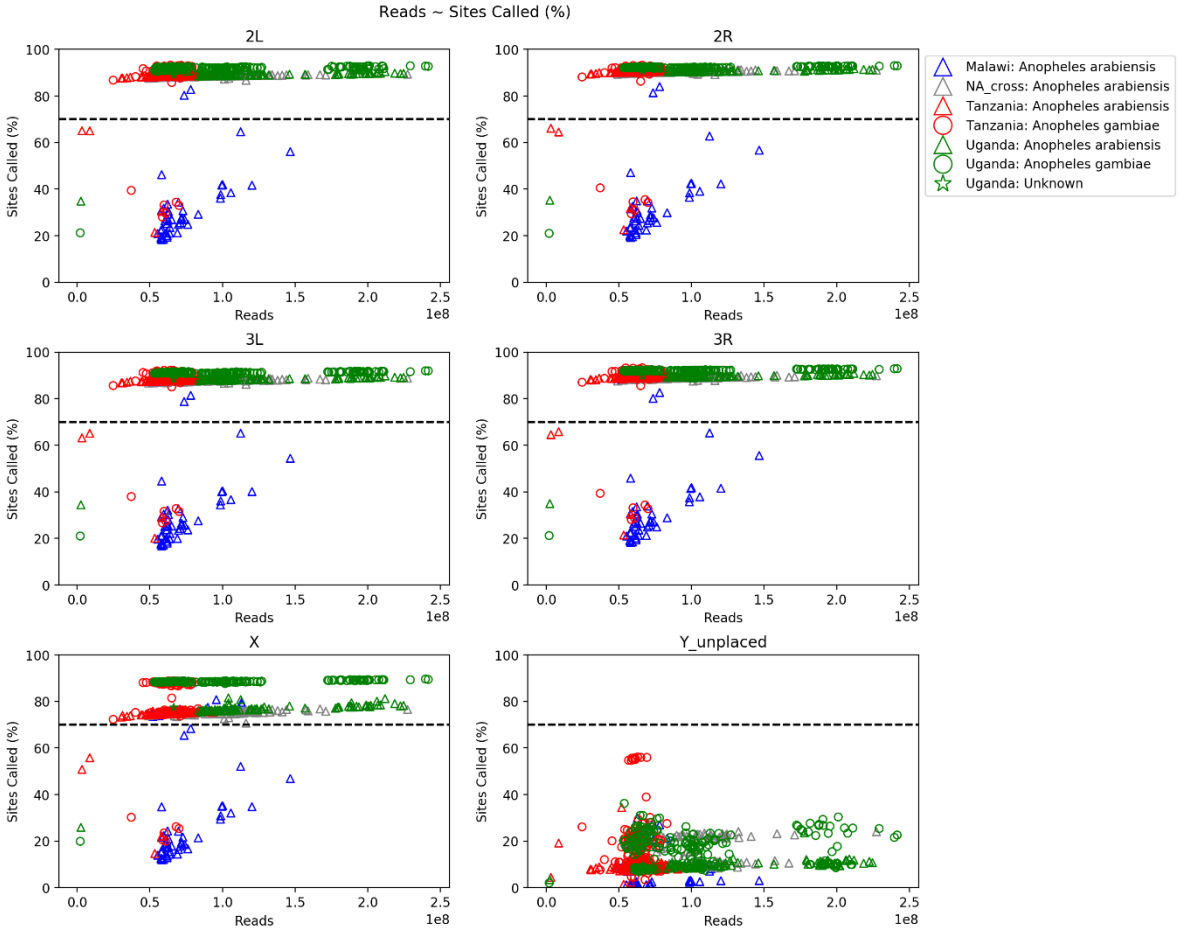


Figure 5. Number of reads vs sites called (%) for each sample by chromosome. The read count of a sample can be used a proxy for the completeness of sequence reconstruction, alongside the proportion of sites called, we observe that the majority of samples have approximately 1-2.5*10^8 reads with a sites called percentage above 70%.

Although we find that the read count for Malawian samples is similar to that of Tanzanian and Ugandan samples that pass the first defined quality control step of percentage of sites called, the proportion of these reads which align with the reference genome is drastically reduced (Figure 5). We are not able to elucidate the reason for the Malawian samples failing the QC stage, however, based on the combination of alignment statistics founds in these samples, similar read counts to passed samples, low percentage of called sites and aligned reads two hypotheses emerge; 1. That Malawian *An. arabiensis* are substantially different genetically from Tanzanian and Uganda *An. arabiensis* (which in themselves are not substantially different to their sympatric pairs), or 2. That the sample collection, storage of preparation for these samples has affected sequencing output, indeed the submission meta data documentations shows that a large number of samples from Malawi were whole genome amplified (WGA) prior to sequencing. WGA prior to sequencing can introduces artifacts caused by poor amplification quality, allelic drop-out or allelic imbalance (Borgstrom *et al.*, 2017).

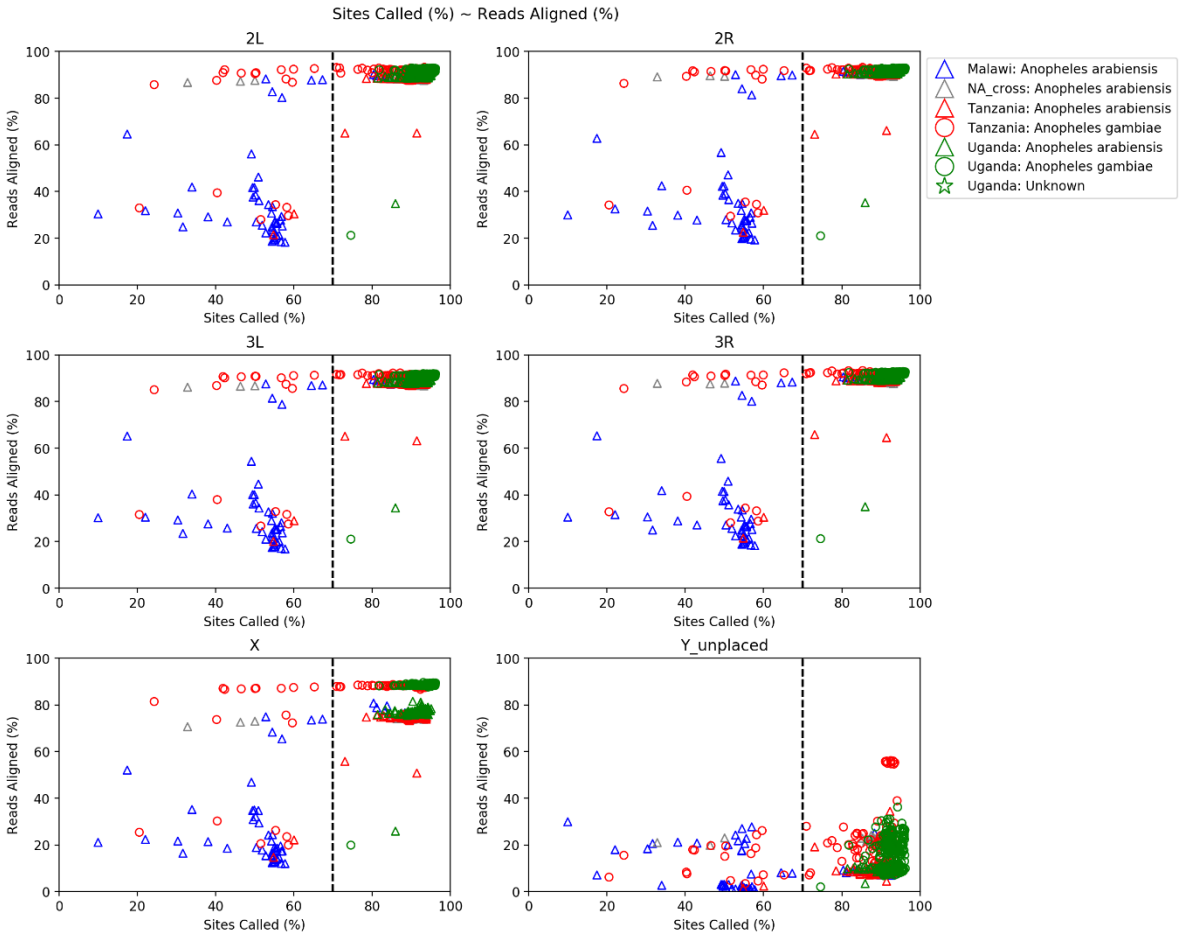


Figure 6. Sites called (%) vs reads for each sample by chromosome.

From this coverage-based analysis, we set two filters for sample inclusion, as sample must possess >10 mean depth and >70% sites called. Applying these filters removes 57 samples for the data set, the majority of which belonged to the Malawaian *An. arabiensis*, as discussed.

Contamination

Contamination from other species DNA or bacteria is a key consideration for analyses, as contaminated samples can cause poor quality genotype calls. We used the *VerifyBamID freemix* tool to assess samples for contamination confirmed with allelic imbalance plots. *Freemix* calls contamination based on heterozygosity excess. In contaminated samples the observed counts of heterozygotes at common SNP positions will exceed 2pq, under assumptions of Hardy-Weinberg Equilibrium (Equation 1). For each sample, an allelic imbalance plot was also constructed. This plot shows the depth for each genotype call for the reference and alternate allele. In contaminated samples, no clear boundary can be observed between heterozygote calls and homozygote calls in alleles (Figure 6).

Equation 1. Hardy-Weinberg Equilibrium equation.

$$p^2 + 2pq + q^2 = 1$$

p^2 = dominant homozygous frequency (AA)

$2pq$ = heterozygous frequency (Aa)

q^2 = recessive homozygous frequency (aa)

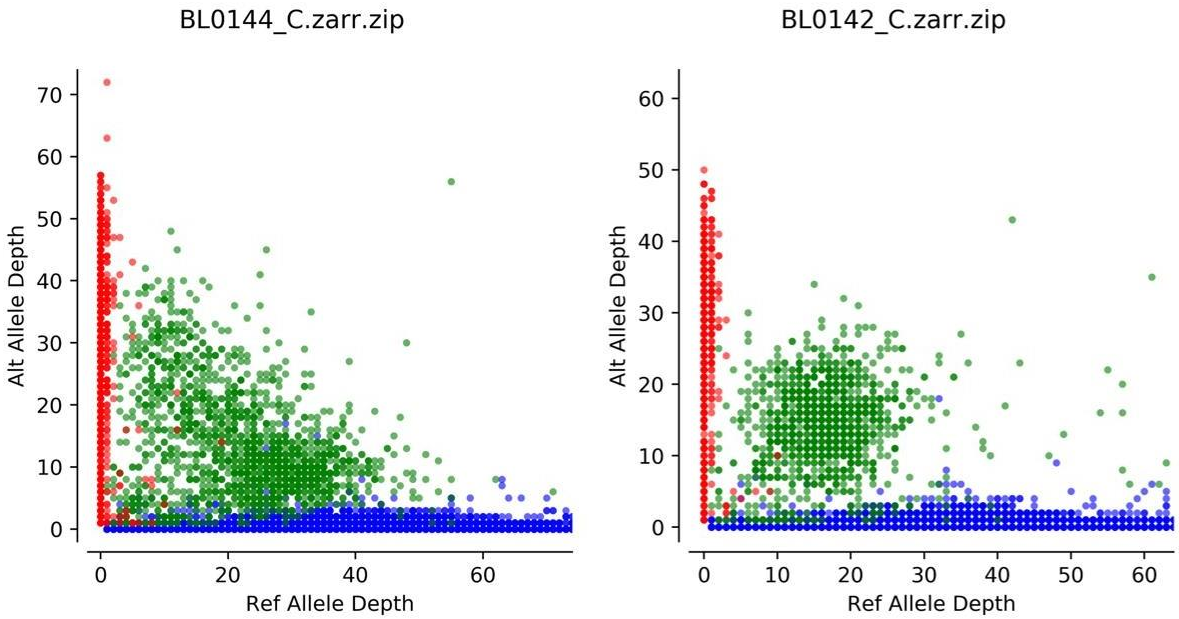


Figure 7. Allele imbalance plots used to investiagte evidence of contamination in samples (blue = homozygote reference, red = homozygote alternate, green = heterozygote). shows sample BL0144_C (left) has very little distinction between heterozygote and homozygote reference calls, suggesting contamination, whereas sample BL0142 shows clear groupings. These samples have a predicted freemix score of 6% and 1.2%, respectively.

Following the precedent set by phase 1 and 2 of the Ag1000G, we excluded all samples with a predicted *Freemix* contamination score of >3.5%, for all these samples allele balance plots showed clear evidence of contamination. 77 samples were had a *Freemix* contamination score of >3.5%, all these samples showed clear contamination on allele balance plots, as such, these samples were excluded.

Principal component analysis

To explore the population structure of phase 3 samples and verify the associated meta data, principal component analyses were performed. Only biallelic SNPs from chromosome 3L were used for these analyses, to avoid confounding signals caused by chromosomal inversions. Accessible sites for these analyses were based on Ag1000G project phase 2 accessible sites filtering. Initial PCA analyses are visualised in Figure 7.

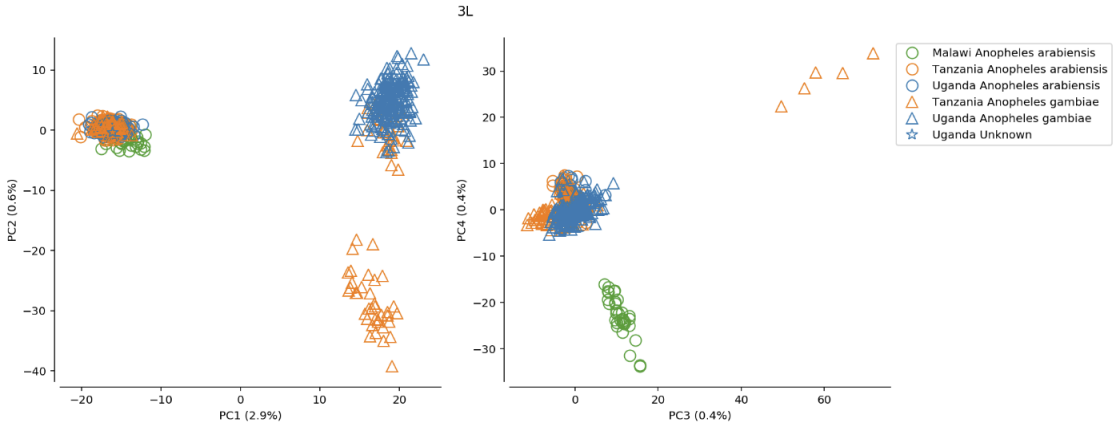


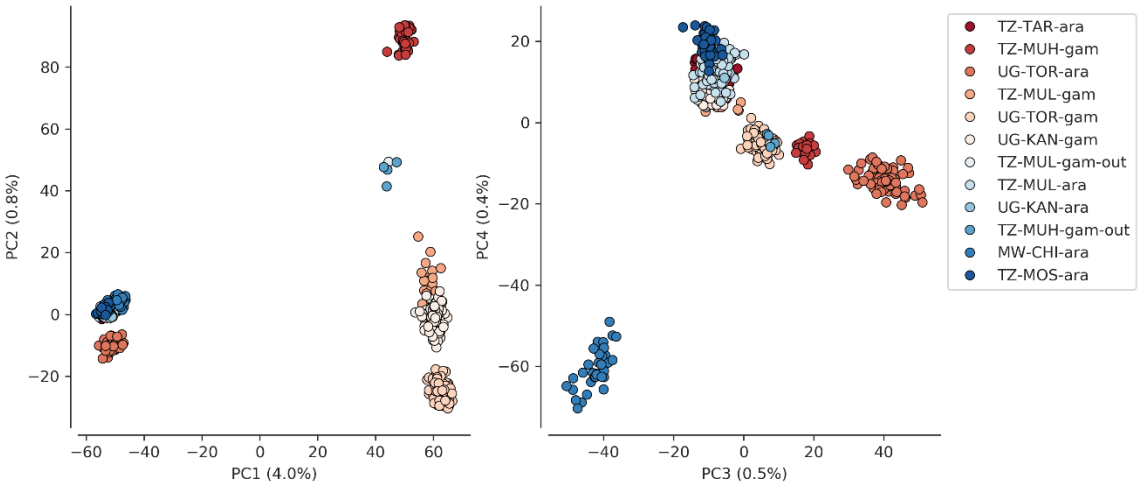
Figure 8. Initial PCA of phase 3 samples for chromosome 3L.

PC1 is driven nearly entirely by species, the exception being a cluster of putative Tanzanian *An. gambiae* samples within the *An. arabiensis* cluster. Consulting the meta data and submission documents shows that many samples from this category were missing species identification. The most parsimonious explanation for this signal is a meta data error. Upon discussion with Ag1000G senior members and confirmation with the submission data, we label these *An. gambiae* samples as *An. arabiensis* for subsequent analyses. PC2 begins to separate out geographical collection locations in the *An. gambiae* cluster of samples, pulling apart the majority of Tanzanian *An. gambiae* samples.

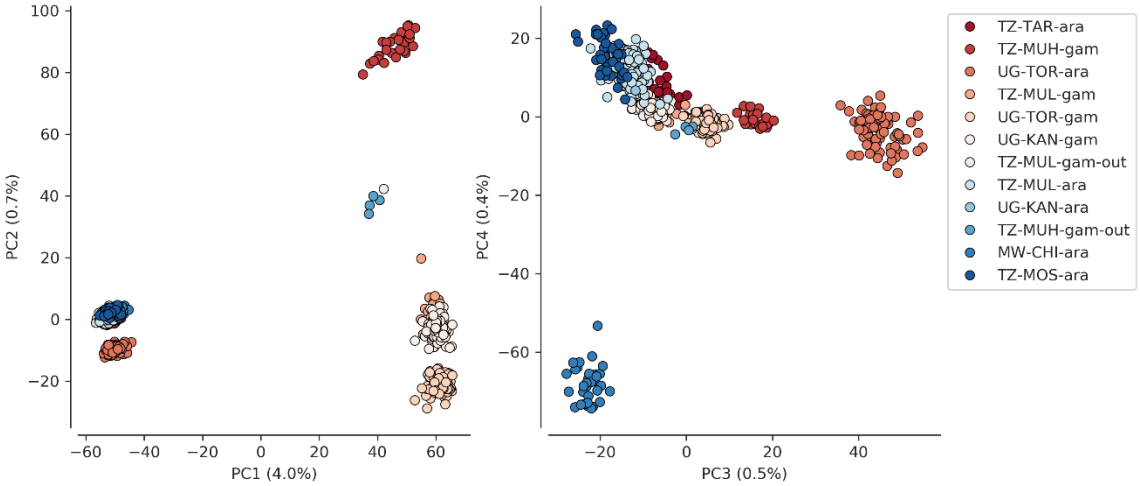
Interestingly, the *An. arabiensis* cluster does not begin to separate until PC4, where the more geographically distal Malawian *An. arabiensis* samples are separated. PC3 clusters 5 Tanzanian *An. gambiae* samples separately, these 5 samples were tagged as outliers. We used this PCA analyses to define the final populations that each sample belongs to for downstream analyses. In summary, the changes informed by this analysis are the tagging of 5 outlier samples and the relabelling of Tanzanian *An. gambiae* to *An. arabiensis* were they are found to be clustering with heterospecific samples.

Additionally, we conducted PCA analyses for all chromosomes (excluding Y) using the labelling acquired from the initial PCA (Figure 9). These analyses show a similar tight clustering of *An. arabiensis* samples; therefore, to visualise any sub-structure within the *An. arabiensis* samples, further PCA analyses were also performed exclusively on those samples (Figure 10). A geographic distribution of these locations is visualised in Figure 2.

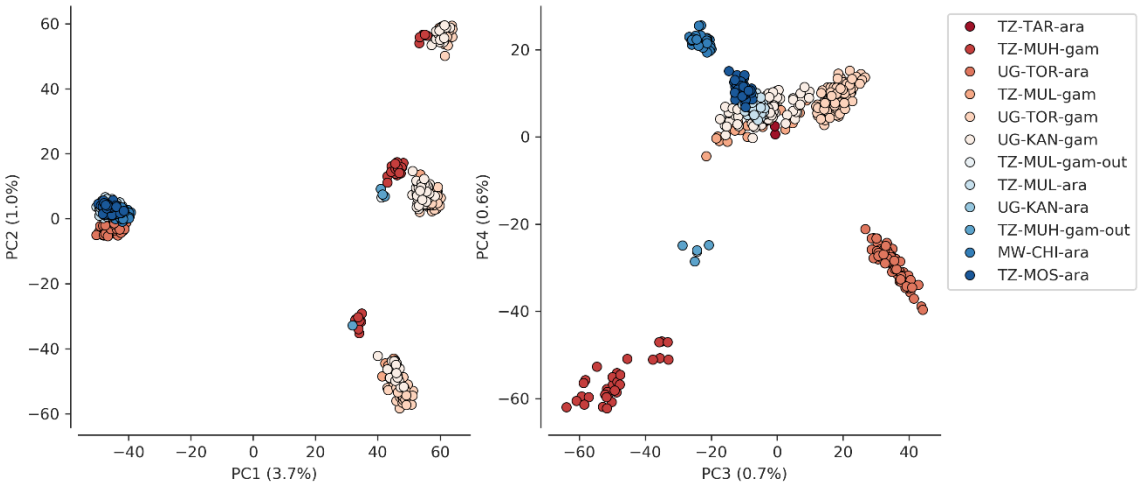
3L PCA


PC2 (0.8%)
PC1 (4.0%)
PC4 (0.4%)
PC3 (0.5%)
TZ-TAR-ara
TZ-MUH-gam
UG-TOR-ara
TZ-MUL-gam
UG-TOR-gam
UG-KAN-gam
TZ-MUL-gam-out
TZ-MUL-ara
UG-KAN-ara
TZ-MUH-gam-out
MW-CHI-ara
TZ-MOS-ara

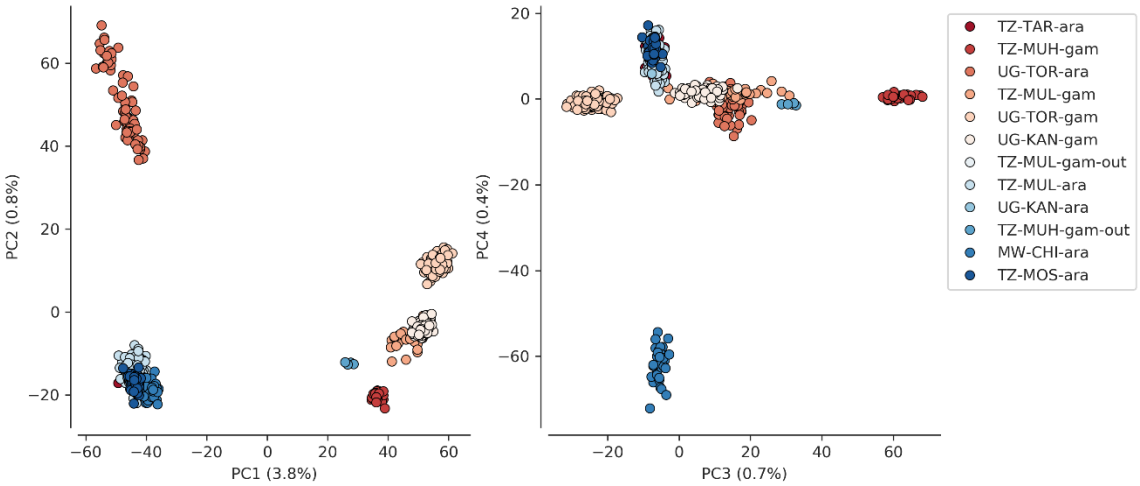
2R PCA


PC2 (0.7%)
PC1 (4.0%)
PC4 (0.4%)
PC3 (0.5%)
TZ-TAR-ara
TZ-MUH-gam
UG-TOR-ara
TZ-MUL-gam
UG-TOR-gam
UG-KAN-gam
TZ-MUL-gam-out
TZ-MUL-ara
UG-KAN-ara
TZ-MUH-gam-out
MW-CHI-ara
TZ-MOS-ara

2L PCA


PC2 (1.0%)
PC1 (3.7%)
PC4 (0.6%)
PC3 (0.7%)
TZ-TAR-ara
TZ-MUH-gam
UG-TOR-ara
TZ-MUL-gam
UG-TOR-gam
UG-KAN-gam
TZ-MUL-gam-out
TZ-MUL-ara
UG-KAN-ara
TZ-MUH-gam-out
MW-CHI-ara
TZ-MOS-ara

X PCA


PC2 (0.8%)
PC1 (3.8%)
PC4 (0.4%)
PC3 (0.7%)
TZ-TAR-ara
TZ-MUH-gam
UG-TOR-ara
TZ-MUL-gam
UG-TOR-gam
UG-KAN-gam
TZ-MUL-gam-out
TZ-MUL-ara
UG-KAN-ara
TZ-MUH-gam-out
MW-CHI-ara
TZ-MOS-ara

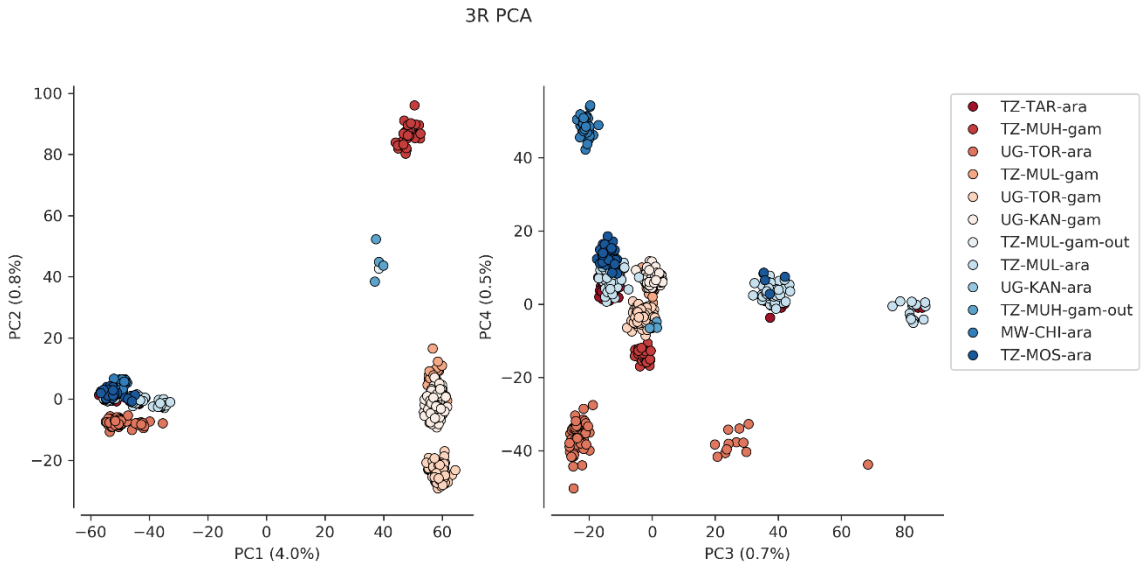
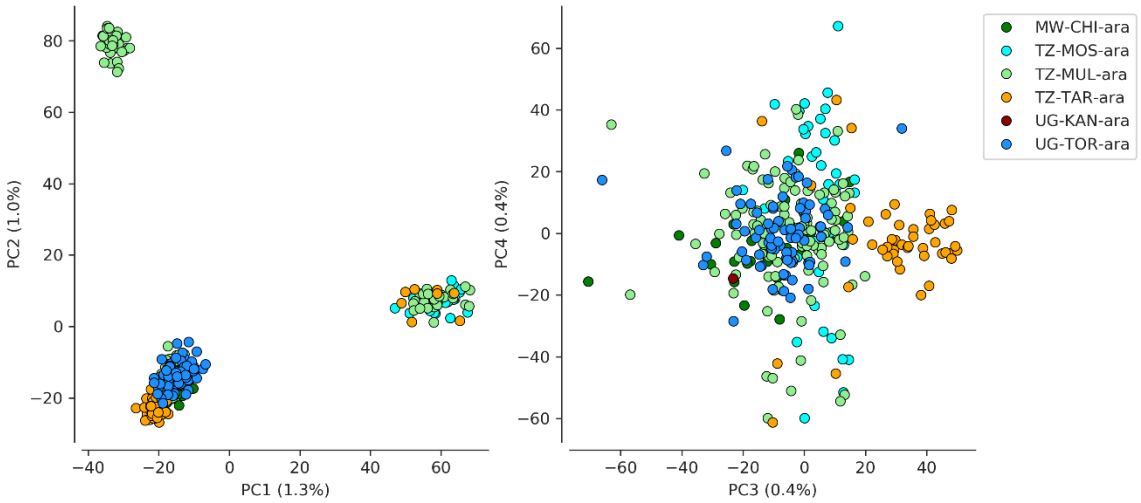


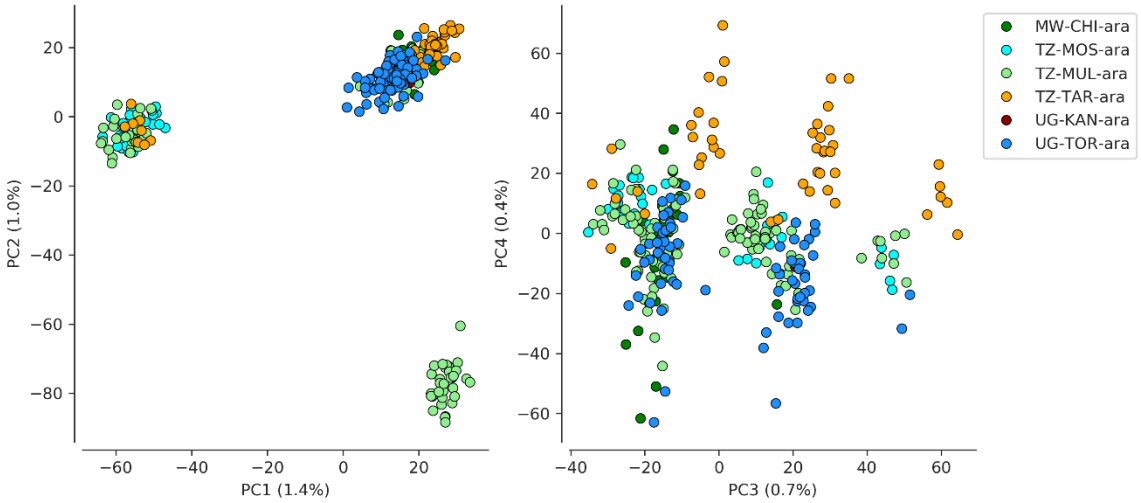
Figure 9. PCA of phase 3 samples for each chromosome at phase 2 defined accessible sites.

Each chromosome PCA separates species on PC1, whereas PC2 is concerned with handling different features, depending on the chromosome. For example, PC2 of 2L is likely resolving the 2La chromosomal inversion for *An. gambiae* samples. Compared to PC2 of each other chromosome, which appears to be pulling apart geographic differences in the samples. PC1 of each chromosome also fails to resolve any dimensionality within *An. arabiensis* samples. PC3 and PC4 of each chromosome separates out Malawian *An. arabiensis*, with no clear distinction between other *An. arabiensis* populations and *An. gambiae* population clusters.

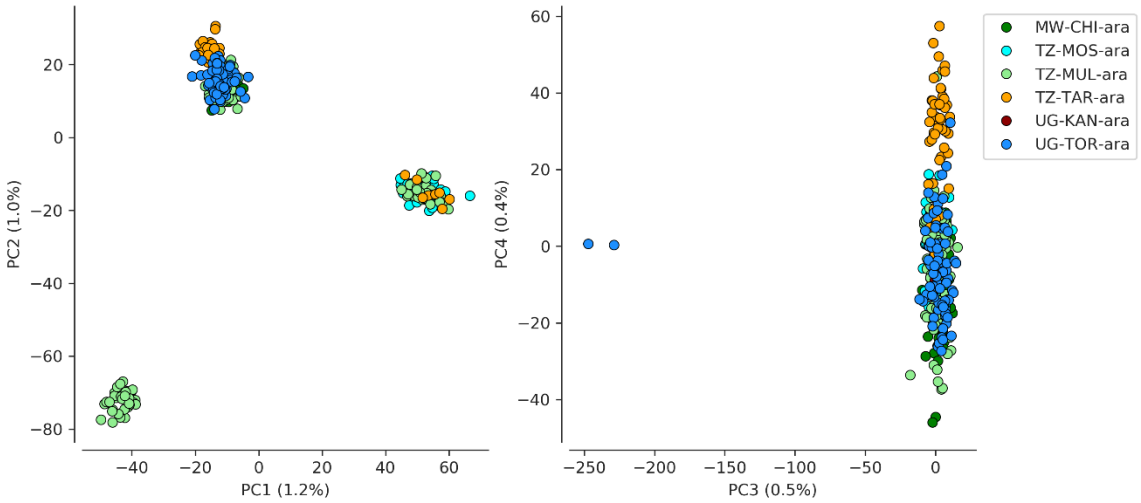
2L PCA


PC2 (1.0%)
PC1 (1.3%)
PC4 (0.4%)
PC3 (0.4%)
MW-CHI-ara
TZ-MOS-ara
TZ-MUL-ara
TZ-TAR-ara
UG-KAN-ara
UG-TOR-ara

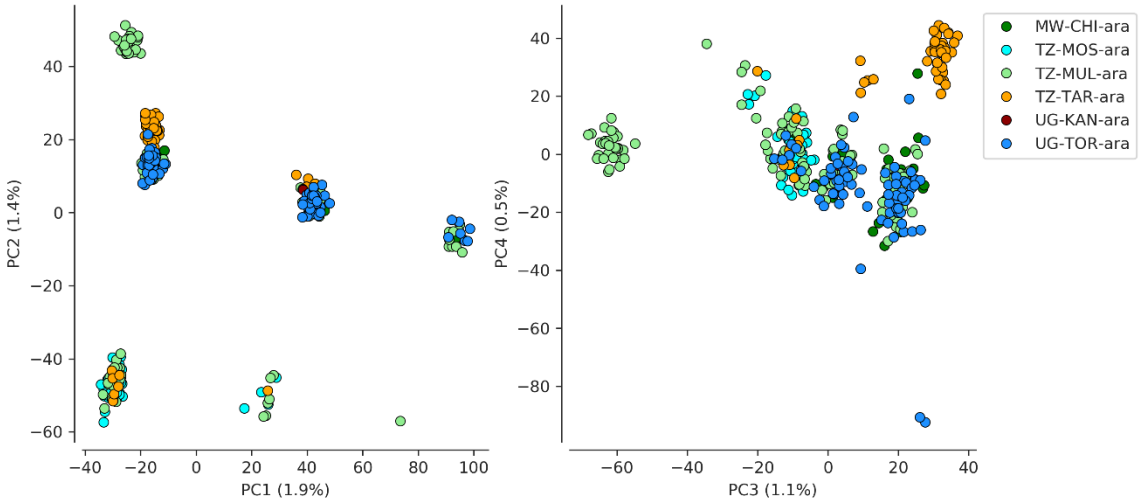
2R PCA


PC2 (1.0%)
PC1 (1.4%)
PC4 (0.4%)
PC3 (0.7%)
MW-CHI-ara
TZ-MOS-ara
TZ-MUL-ara
TZ-TAR-ara
UG-KAN-ara
UG-TOR-ara

3L PCA


PC2 (1.0%)
PC1 (1.2%)
PC4 (0.4%)
PC3 (0.5%)
MW-CHI-ara
TZ-MOS-ara
TZ-MUL-ara
TZ-TAR-ara
UG-KAN-ara
UG-TOR-ara

3R PCA


PC2 (1.4%)
PC1 (1.9%)
PC4 (0.5%)
PC3 (1.1%)
MW-CHI-ara
TZ-MOS-ara
TZ-MUL-ara
TZ-TAR-ara
UG-KAN-ara
UG-TOR-ara

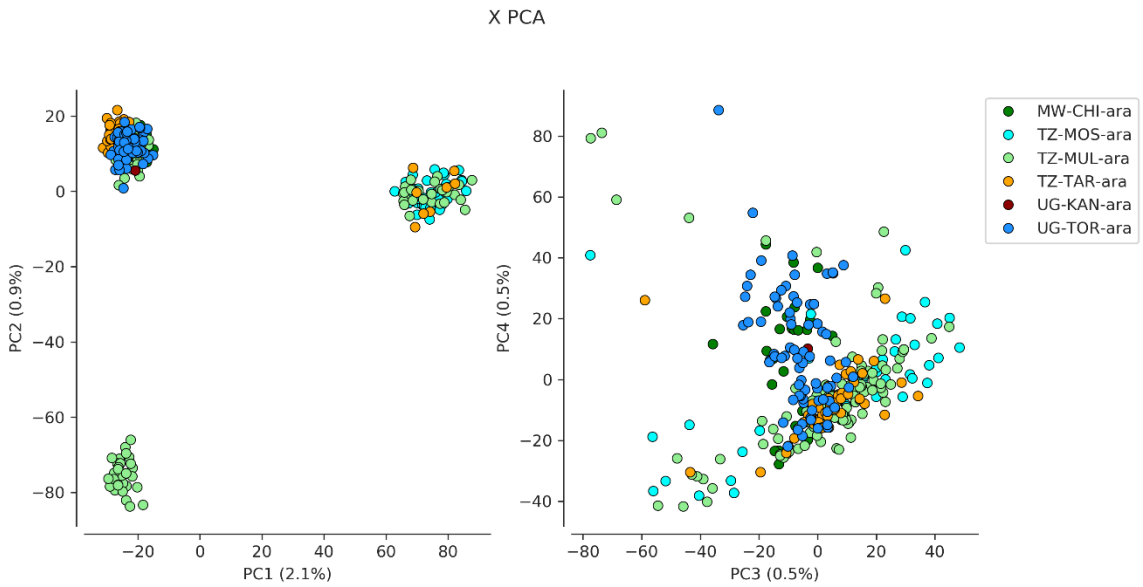


Figure 10. PCA of phase 3 *An. arabiensis* at phase 2 defined accessible sites.

Visualising the PCA of *An. arabiensis* samples only begins to show population sub-structure (Figure 10). PC1 and PC2 of chromosomes 2L, 2R, 3L and X all appear to resolve the same feature driven by population origin. However, an interesting overlap of populations is observed such that Tanzanian samples from Muleba are more distant from their conspecific counterparts than a more distal population such as Malawian *An. arabiensis* (Figure 8). This cluster of samples cluster with more closely with its conspecifics on PC3 and PC4 all chromosomes, excluding 3R. PC3 and PC4 of 2R and PC1 and PC2 of 3R appear to resolve chromosomal inversion signals, likely the 2Rb/c and 3La inversions, respectively.

Crosses

Phase 3 samples of the Ag1000G project contains 4 colony crosses of two Sudanese laboratory strains of Senn and Dongola. Whole genome sequence data

from these cross parents and progeny can be informative for mendelian error and population genetic, recombination rate, quantifying sequencing error rate and empirically deriving mutation rate. To ensure the samples were of sufficient quality and check for meta data errors, colony cross specific QC was conducted.

For each colony cross group, the expectation is that ≈50% of the member are male and ≈50% female. Supplied meta data does not indicate the sex of colony cross progeny, and only identifies the parental and maternal samples. We use two methods to identify the sex of samples. Firstly, X heterozygosity is expected to be 0 for XY males and approach 1 for XX females. Secondly, we used the ratio of the mean depth across chromosome X to 3L; under the expectation that the value for females would be approximately 2x that of the males assuming roughly uniform mean depth of chromosome 3L. These metrics were visualised against each other for each of the four colony crosses (Figure 11).

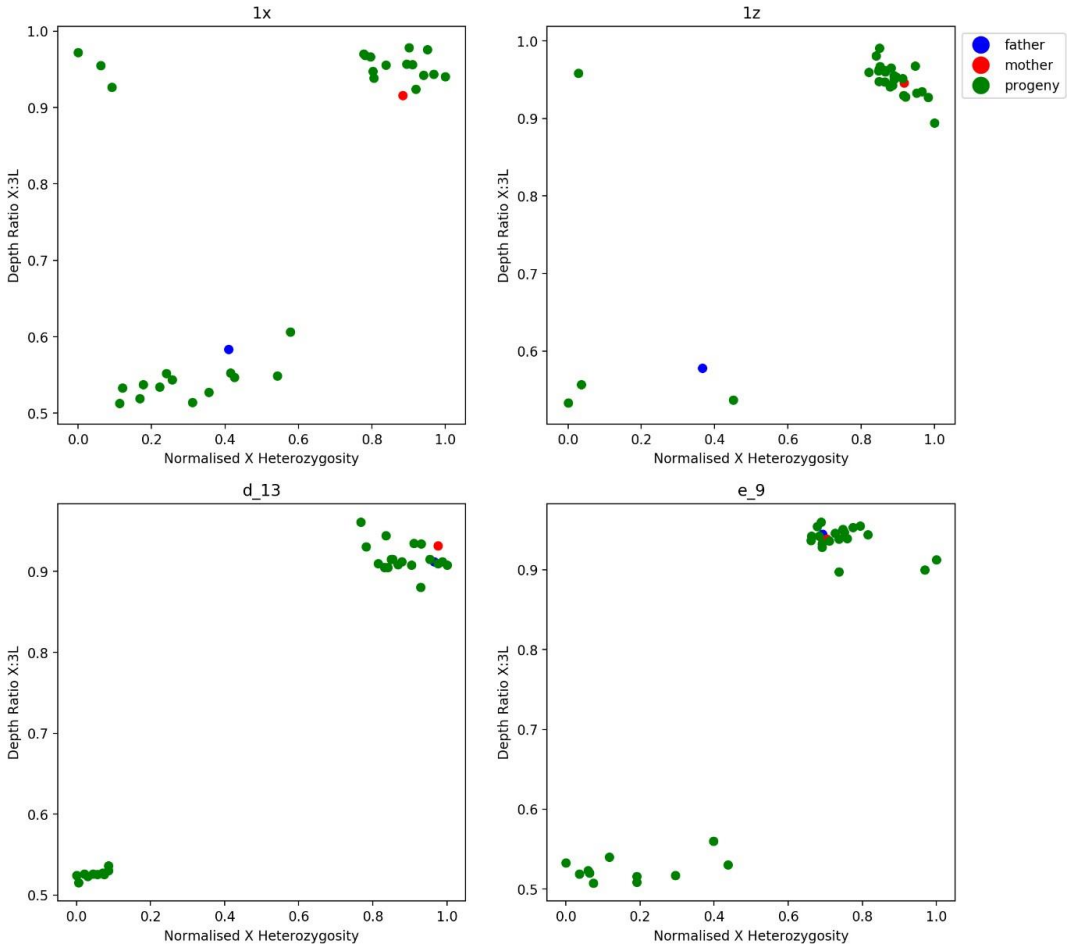


Figure 11. X heterozygosity against X:3L depth ratios as metrics for calling sex in phase 3 cross samples. In determining the sex of samples, we expect to observe male samples have no X heterozygosity and an X:3L depth ratio of roughly 0.5. Female samples have high X heterozygosity and X:3L depth ratio.

Identification of sex based on both X heterozygosity and X:3L depth ratio reveals inconsistencies in both expected patterns of signals and the supplied meta data. Crosses 1x and 1z both contain samples which have a low X heterozygosity, suggesting male and a high X:3L depth ratio, suggesting female. Additionally, the majority of samples for cross 1z are called as female. Finally, the paternal samples

for crosses d_13 and e_9 cluster with the maternal samples and female offspring. To attempt to resolve these inconsistencies, windowed heterozygosity was calculated across the genome for outlier samples and pairwise genetic distance was used to resolve the paternity and maternity of each colony cross. The method of using X:3L ratio to discriminate sex is sensitive to the mean depth chromosome 3L and X. Indeed, visualising heterozygosity across the genome seems to be an appropriate method for calling sex based solely on the presence of observed expected X heterozygosity.

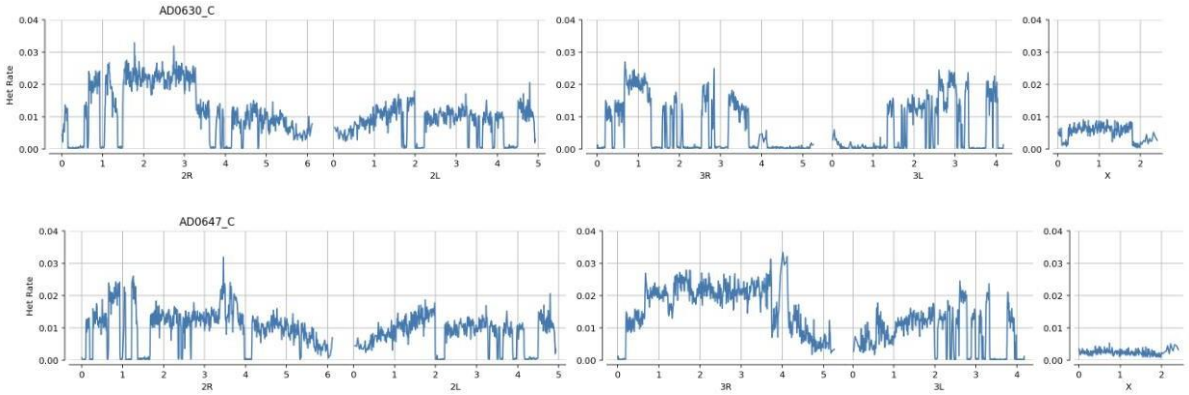


Figure 2 Representative windowed heterozygosity plots for colony cross members. Top - female, bottom - male.

Resolving dubious paternity and maternity

As observed in Figure 11, the metadata identified fathers of the sample crosses d_13 and e_9 are clustered with the putative females. Indeed, PCA analysis of the cross samples appears to show no consistent clustering (Figure 13).

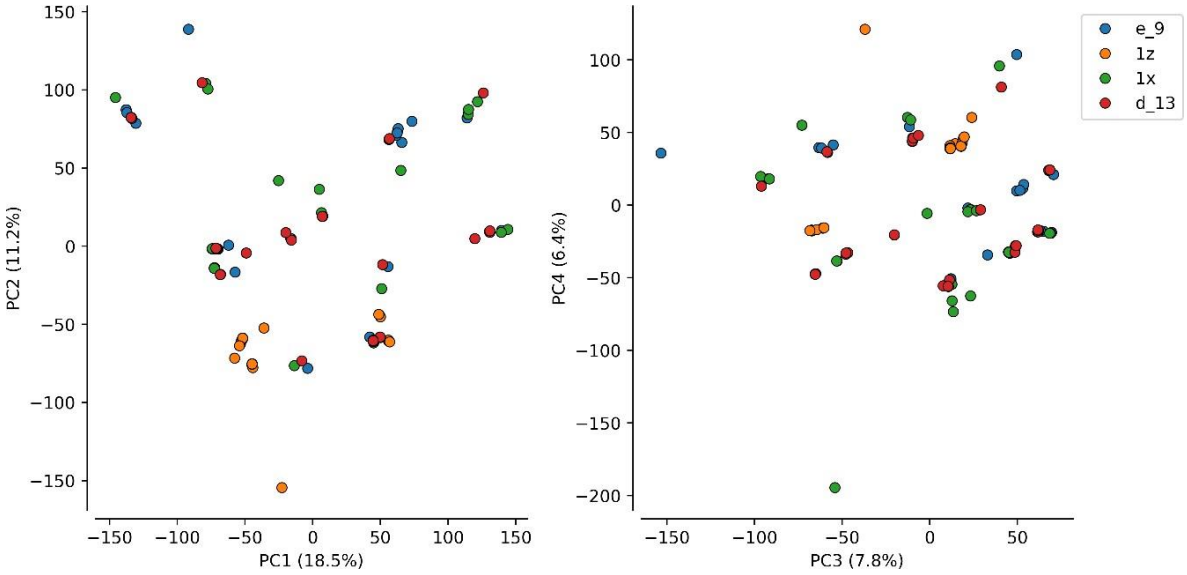


Figure 12. PCA of phase 3 cross samples for chromosome 3L by sample. No clear relationship is discernible between member of the same supposed families in this analysis.

To deconvolute the potential sample mix up and resolve paternity we performed pairwise genetic distance analysis on cross samples, using chromosome 3L. For an assumed population of siblings with both parents present, the expectation is that siblings possess a low pairwise genetic distance from each other, with a greater pairwise genetic distance from that of both parents, due to heterozygous positions between parents for a pure SENN vs DONG colony comparison. Implementing this method on Ag1000G phase 2 cross samples confirms the expected signal (Figure 14). For the phase 2 cross samples the two bottommost and leftmost samples represent the parental samples. The pairwise distance between the parents is comparatively higher than that of the distribution of pairwise distances observed between offspring.

Applying the same metric to phase 3 cross samples shows a more convoluted arrangement of putative parents and offspring (Figure 15). For the phase 3 cross samples there is no clear signal of parental samples that can be visually distinguished from pairwise genetic distance heatmaps. Cross 1z shows the least unexpected signal, showing a single sample with a higher genetic distance from that of the remaining samples. However, this signal is replicated in many other samples from cross e_9, 1x and d_13, suggesting a cross population sample mix up.

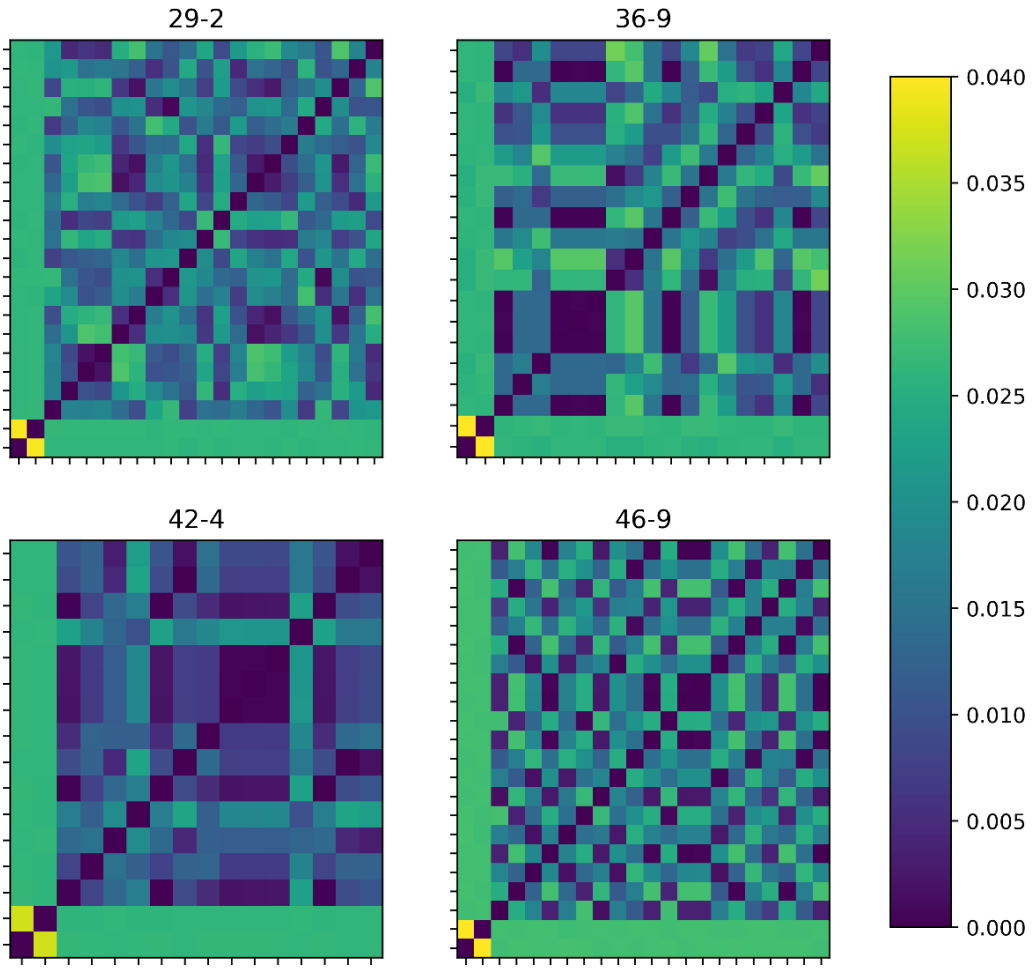


Figure14. Pairwise genetic distance of phase 2 *An. gambiae* cross samples. The pairwise genetic distance between the parents and that of the progeny is typical for all four cross populations. That is, we observe that the two parents have a genetic distance closer to each other than each of the progeny do to each other. The parental and progeny samples here are easily discerned.

To attempt to resolve sample mix up, pairwise genetic distance plots were created again with all the phase 3 cross samples combined. A hierarchical dendrogram cluster was also plotted alongside these visualisations (Figure 16). Pairwise genetic distance appears to be an inadequate way of resolving the parents and offspring for these samples. Both the combined heat maps of genetic

distance and hierarchical dendrogram clustering reveals no clear expected structure. One final method employed to resolve these samples was the genetic inconsistencies between samples.

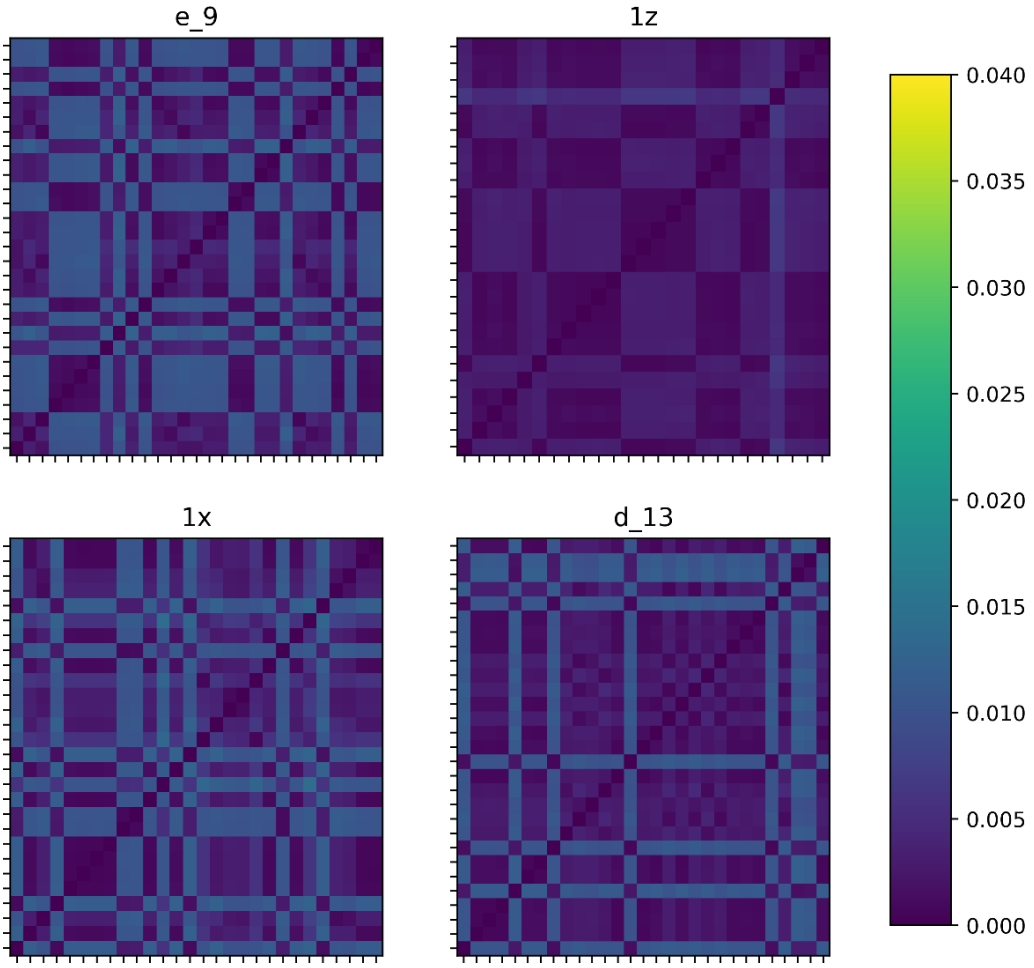


Figure 15. Pairwise genetic distance of phase 3 *An. gambiae* cross samples. In observing the pairwise genetic distance for these supposed cross population. No information as the parental or progeny samples is detectable.

An inconsistency can be viewed as a homozygote reference call versus a homozygote alternate call, for example at a given position, an offspring genotype

of 1/1 necessitates that both parents be either 0/1 heterozygotes or 1/1 homozygotes. This gives the expectation that the number of positions that show 1/1 vs 0/0 inconsistencies be low between parents and progeny, with a higher probability of finding an elevated inconsistency between parents and a low distance between siblings. Inconsistency analysis (Figure 17) does not clearly resolve parental labelling for phase 3 samples.

Attempting to resolve kinship and parental samples for phase 3 crosses reveals major sample mix up with no clear delineation between crosses. There is no clear path for in silico resolution of this mix up to rescue samples. Sample 50 meets the homozygosity constraints and the maximum distance so is a strong candidate to be a colony parent. However, no other samples are close to sample 50. Most samples have <10 discordant positions with ~20 other samples, however number 50 appears to give a unique signal (Figure 17). If parents are present within these samples, pairwise genetic distance is an insufficient method for resolving these samples, as mixed cross progeny look similar in terms of genetic distance irrespective of parents.

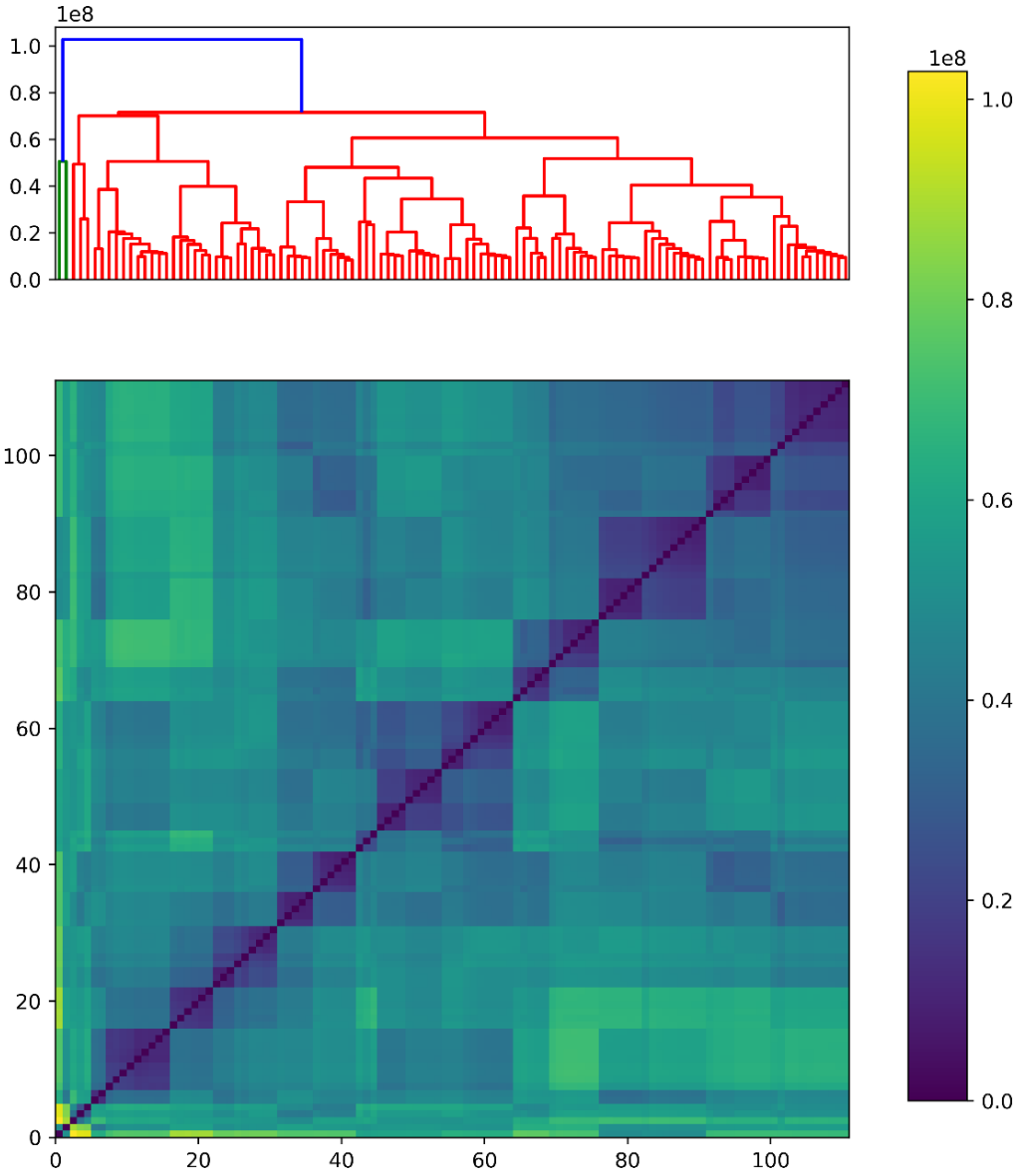


Figure 16. Dendrogram clustering and heatmap of Euclidean distance for phase 3 cross samples.

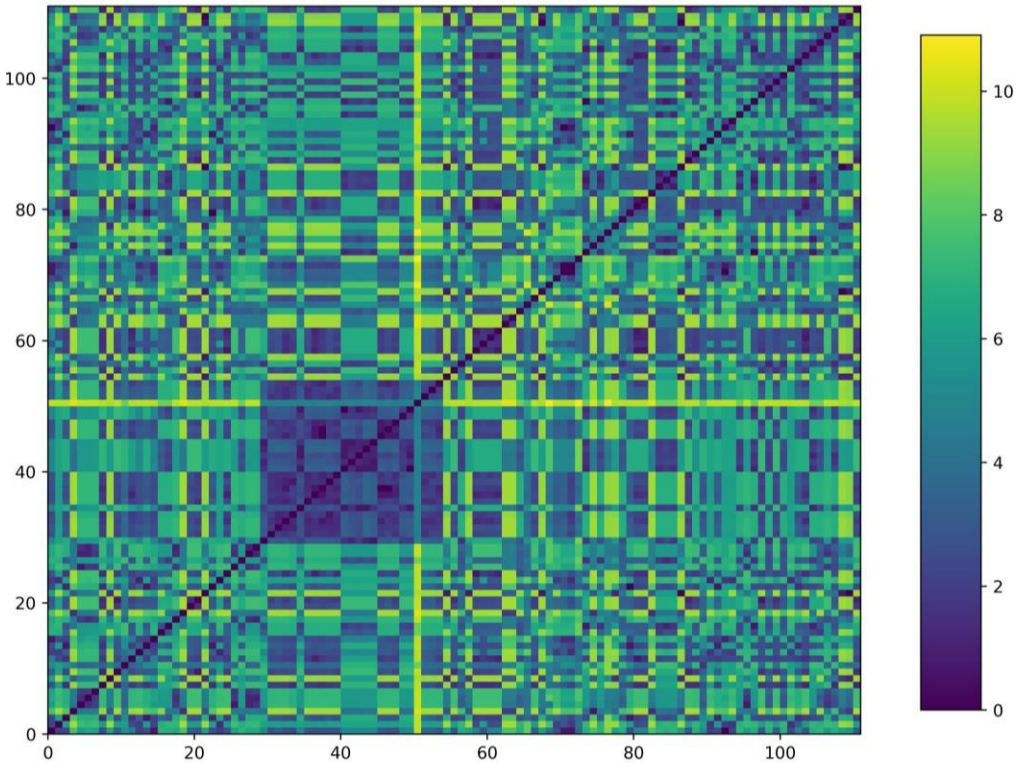


Figure 17. Euclidean distance of genotype inconsistencies for phase 3 cross samples.

References

Anopheles gambiae Genomes, C., Data Analysis, G., Partner Working, G., Sample, C.-A., Burkina, F., Cameroon, Gabon, Guinea, Guinea, B., Kenya, Uganda, Crosses, Sequencing, Data, P., Web Application, D. & Project, C. 2017. Genetic diversity of the African malaria vector *Anopheles gambiae*. Nature, 552, 96-100.

Borgstrom, E., Paterlini, M., Mold, J. E., Frisen, J. & Lundeberg, J. 2017. Comparison of whole genome amplification techniques for human single cell exome sequencing. *PLoS One,* 12**,** e0171566.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows- Wheeler transform. Bioinformatics, 25, 1754-60.

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & Depristo, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20, 1297-303.

Chapter III.

Ancestry Informative Markers

Introduction

Ancestry informative markers (AIMs) are a polymorphic set of DNA sequences that appear in varying frequencies between different populations. Within a given marker set, AIMs can provide diagnostics for different populations/species. This feature of genomes within the same species to be diagnostic across populations is driven by the polymorphic nature of DNA. The natural variation that occurs between separate niches/populations of a species have many uses. Indeed, AIMs are utilized across disciplines including genealogy, forensics, population genetics and genetic research. For example, in human genetics, AIM panels have been developed using European, African, Hispanic and American samples to determine continental origins and levels of admixture (Huckins *et al.*, 2014; Tandon *et al.*, 2011; Tian *et al.*, 2009; Wang *et al.*, 2019; Zhang *et al.*, 2014). This concept is the core of this chapter, that is, a species that has only one population which undergoes segregation into two different populations through geographic or predatory pressure – if isolated for long enough – can begin mating separate from the primary niche. In essence, these novel alleles in the nascent population are an example of the founder effect. However, those alleles can be directly used to identify from which geographical niche a given specimen belongs.

Here, we aim to identify all the regions in *Anopheles gambiae* and *Anopheles arabiensis* which are both fixed (not segregating) and different between

the species. These SNPs represent positions where *An. gambiae* and *An. arabiensis* are uniquely different due to either the requirement for genetic variation in the driving of speciation, polymorphism driven natural variation, recombination cold spots or extended hitchhiking. Given the close relationship between *An. gambiae* and *An. arabiensis*, our hypothesis is that given a panel of markers that uniquely call species based on SNPs, we can identify regions of contemporary introgression and corroborate other observed signals of introgression.

In phase 1 of the Ag1000G, ancestry informative markers were successfully used to identify introgression of a genomic region on the 2L chromosome between *An. gambiae* and *An. coluzzii* from both Burkina Faso and Angola (*Anopheles gambiae* 1000 Genomes Consortium 2019). This region encompassed the Vgsc gene, the same locus which had prior reports showing introgression in both Ghana and Mali. Therefore, there certainly exists the precedent for this method to be germane in the discussion of introgression. Additionally, the authors note that their AIMs highlighted two populations of samples with a dubious species status.

Several considerations must be made when creating a panel of AIMs in the manner. First, as we plan to use the Ag1000G project data, encompassing both *An. gambiae* and *An. arabiensis*, we need to develop our marker panel set either from a subset of samples from these species or an entirely different data set. If we were to use solely the Ag1000G dataset to identify which regions are both fixed and different between the species, when the panel is then applied to the samples to observe where discrepancies may be present, the marker set would be both 100% informative of species and trivial, as the data was gleaned from those

samples to begin with. To that end, we sourced a data set from Neafsey *et al.* (2015) which contained both *An. gambiae* and *An. arabiensis* collected in both a spatially and temporally distinct manner to the Ag1000G data.

In this chapter, we detail the methods used to create a panel of markers that are able to resolve *An. gambiae* or *An. arabiensis* species identification and discuss their application for detecting signals of introgression.

Methods

Data

To develop a panel of markers which can be used as diagnostic of species we used data from the 16 genomes project (Neafsey *et al.*, 2015). This dataset contains data from both the *An. gambiae s.s.* assembly and 12 *An. arabiensis* samples from Burkina Faso, Cameroon, Kenya (Holt *et al.*, 2002; Neafsey *et al.*, 2013; Neafsey *et al.*, 2015). These data were imputed and converted using Python 3 and Jupyter Notebooks, the script generated to conduct this imputation is supplied within Appendix A. However, in brief, to ensure consistency between these data and that of the phase 3 Ag1000G project, the 16 genomes project data had:

1. all insertion/deletion information removed. Certainly, insertion/deletion mutations have the capacity to be informative. However, because the build of the Ag1000G phase 3 data does not contain any insertion/deletion data, their inclusion would only obfuscate analyses.

2. data unified such that only positions for which an observation is present (i.e. a successful call by the BWA aligner) in both Ag1000G and 16 Genomes were retained.
3. alternate allele encoding changed to match the encoding used in the Ag1000G project. Allele encoding can take one of two primary forms produced by the aligner, either the alternate allele array is organized to be alphabetical or ordered to respect a descending order of observed frequency. In the case of 16 Genomes, the data were stored to respect the frequency of alternate alleles. For ease of analyses, we changed the 16 genomes data to resemble the format of the Ag1000G data.

AIM discovery

We used Python 3 and scikit-allele to calculate the allele counts for both *An. gambiae* and *An. arabiensis* data from Neafsey *et al.* (2015). The allele counts were then filtered for each species to retain position that were 1. Successfully called by the BWA aligner, 2. Not segregating within species and 3. Mono-allelic or bi-allelic. The remaining genomic positions from *An. gambiae s.s.* and *An. arabiensis* were then unified such that all the positions retained were present in both species' datasets. At this point, the data for *An. gambiae* and *An. arabiensis* contains only positions where both datasets contain genomic positions that are called, fixed and mono/bi-allelic. From this position, we move to identify genomic coordinates where the called alleles are different between *An. gambiae* and *An. arabiensis* datasets. This final set of positions are the ancestry informative

markers and for the dataset from which they were generated will accurately distinguish *An. gambiae s.s.* and *An. arabiensis*.

AIM application

We used the generated list of AIMs to determine regions of both *An. gambiae* and *An. arabiensis* genomes which possess discordant alleles. That is, in *An. gambiae* samples from the Ag1000G at what positions do those samples possess *An. arabiensis* alleles, and vice versa. The process of assigning a 'gambiae value' to a given position in an *An. arabiensis* genome – and vice versa - will be referred to as painting. The phase 3 data from the Ag1000G was painted using the developed AIMs for discordant alleles and a heat map style graph was generated to show the distribution of AIMs that are either concordant or discordant with that of the given species.

For each AIM position, we identify any genes which coincided with those coordinates. This was done to assess whether further introgression analyses (chapter 5) which reveals specific genes as being introgressed contain AIMs revealed in these analyses.

Results

The total number of putative AIMs identified between *An. gambiae* and *An. arabiensis* for each chromosomal arm is 57688 for 2L, 44575 for 2R, 49194 for 3L, 54027 for 3R and 359845 for X. The total distribution of the AIMs were over represented around centromeric region, however localized spikes in AIM count along the chromosomes are observed (Figure 1). Changes and polymorphisms to

the centromeres are less common than changes to that of the telomeres and mid genomic range (Choo, 1998; Mahtani and Willard, 1990). To that end it is anticipated that the centromeres show an over representation of markers that distinguish between two discrete species. However, non-centromeric regions that possess an elevated count of AIMs may suggest a highly conserved or introgressed region.

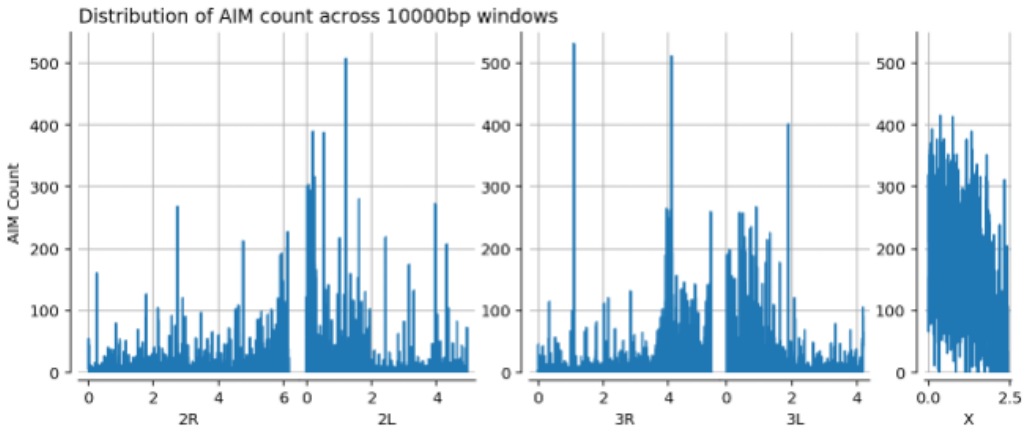


Figure 1. Total number off AIMs found within 10000 base pair windows across the genomes that are able to distinguish between *An. gambiae* and *An. arabiensis*.

The heat maps generated visualize the discordant AIMs between Ag1000G *An. gambiae* and *An. arabiensis* do not show any signals with which we can conclude introgression. We would expect to see discrete regions of discordant alleles, rather than a consistent distribution of discordant alleles across the genomes, as observed with *An. arabiensis* samples in (Figure 2). For the *An. gambiae s*amples all of the painted AIMs are *An. gambiae AIMs*, with the exception of 1 sample; this sample likely been a mislabeling during species identification, as the vast majority of its AIM position are *An. arabiensis* rather the *An. gambiae*. This

identification of species mislabelling was also observed in previous studies which used AIMs to find introgression in *An. gambiae* and *An. coluzzi* (Anopheles gambiae 1000 Genomes Consortium 2017).

All the *An. gambiae* samples rather interestingly show a strong bias toward the *An. gambiae* allele. Though this is to be predicted under the null hypothesis of no introgression, it contrasts sharply with *An. arabiensis* samples which although showing a majority of *An. arabiensis* alleles, still possess an appreciable amount *An. gambiae* alleles (Figure 2).

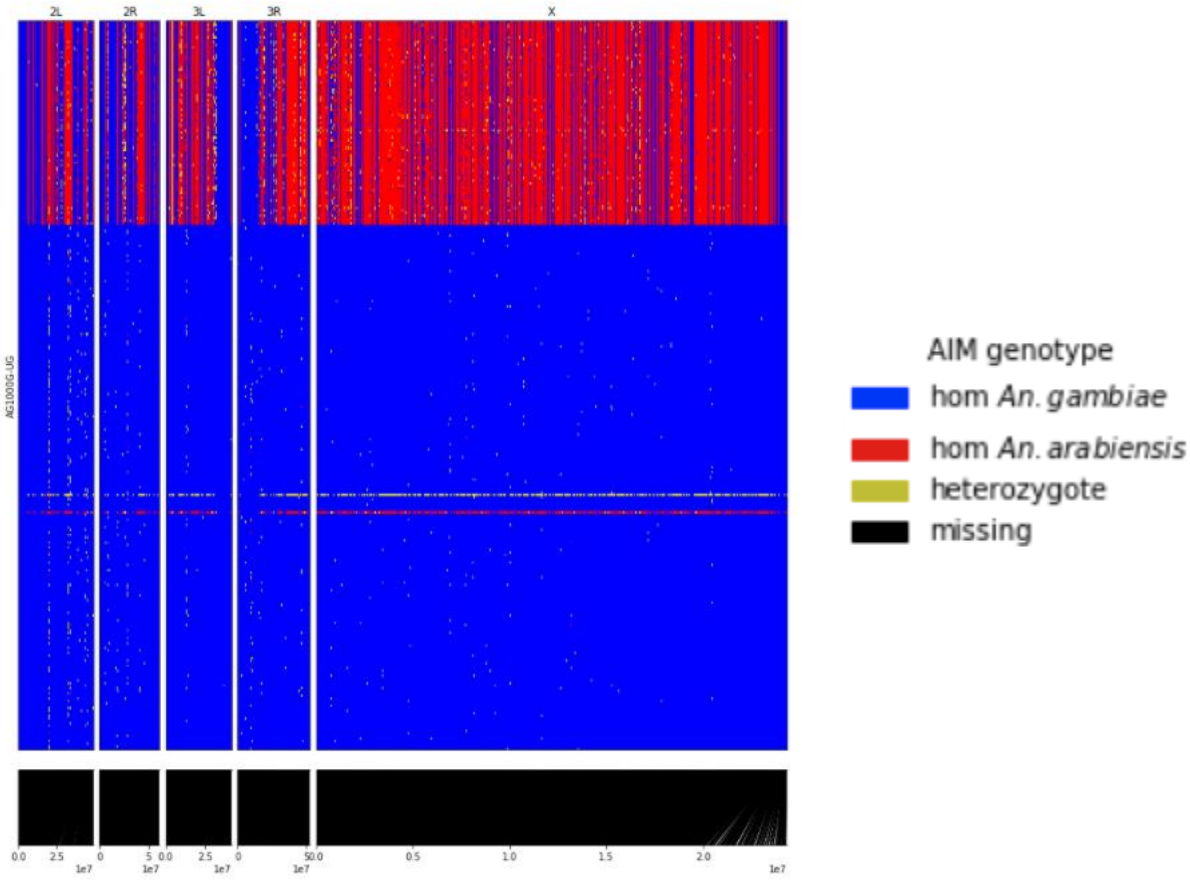


Figure 2. Heat map showing the distribution of all discovered AIMs between *An. gambiae* and *An. arabiensis* colored by whether that position for each sample is homozygous for the *An. gambiae*, homozygous *An. arabiensis* allele, heterozygous, or missing.

Discussion

The overall objective of the work contained within this chapter was to develop a set of ancestry informative markers that can be used in conjunction with introgression analyses detailed in a later chapter, which utilizes Patterson's D statistic to estimate whether admixture is likely given a WGS data for a set of populations. Though the aim of this chapter has been achieved there exists limitations to the AIM based approach which need to be considered.

The first key limitation of the approach is that the method is highly sensitive to the dataset used to generate the panel of markers. In this case, we used publicly available data from Neafsey *et al.* (2015) as it represented the most readily available and complete source of both test genomes and in a format that was computationally economical to handle and convert. A foreseeable issue with using this dataset for this purpose lies with the fact the *An. gambiae* and *An. arabiensis* samples that comprised the foundation of the study conducted in Neafsey *et al.* (2015) were from West Africa. This contrasts with the East African origin of phase 3 Ag1000G data. Genetic information present in the Eastern populations with respect to admixture may be completely absent in Western populations. As such, this results in missing information where in the Western populations, *An. gambiae* and *An. arabiensis* may be fixed and identical rather than fixed and different.

Further, an issue that limits the AIM based approach which is adjacent to that of the limitation previously discussed is the time of speciation. Since any admixture that occurred prior to species divergence, which is not required for speciation, can be viewed as providing alleles that will show as simply the shared

ancestral allele. Therefore, this method only can detect admixture events that postdate the speciation event. Ultimately, this method does serve to aid in detecting the principal interest of this thesis – contemporary introgression events – however, since later chapters present the results of methods that are able to detect admixture that predates speciation, there exists a confounding issue that can hamper comparisons which seek to validate signals of introgression identified via Patterson's D – for example – with the AIMs found discovered through these analyses (Zheng and Janke, 2018).

Finally, in drawing parallels between the use of AIMs in phase 2 of the Ag1000G and the analyses presented here the authors note that coastal Kenyan samples were found to carry alleles from both *An. gambiae* and *An. coluzzii* on all of the chromosomes (*Anopheles gambiae* 1000 Genomes Consortium *et al.*, 2019). Since the geographical range of *An. coluzzii* is not thought to extend beyond the East African rift, the results we unexpected prompting the authors to opine on the possible explanations for such an observation (*Anopheles gambiae* 1000 Genomes Consortium *et al.*, 2017). Their suggestions include historical admixture and retention of ancestral variation. Ultimately, the conclusion was that – at least – for *An. gambiae*/*An. coluzzii* comparisons, one must be cautioned "against using any single maker to infer species ancestry or recent hybridization". Despite the divergence between *An. gambiae* and *An. coluzzii* being much more recent and heightening the concern of shared ancestral polymorphism, similar care and caution are advised with the AIMs developed within this chapter for comparisons between *An. gambiae* and *An. arabiensis*.

Despite the host of caveats and limitation these AIMs have utility for the intended purpose of supplementing further introgression analyses. The intent of further work is to develop a Patterson's D statistic across the genome of *An. gambiae* and *An. arabiensis* to detect signals of admixture. These AIMs will be assessed for how well discordant alleles align within those putative regions of admixture.

References

Anopheles gambiae Genomes, C., Data analysis, g., Partner working, g., Sample, c.-A., Burkina, F., Cameroon, Gabon, Guinea, Guinea, B., Kenya, Uganda, Crosses, Sequencing, data, p., Web application, d. & Project, c. 2017. Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature,* 552**,** 96-100.

Choo, K. H. 1998. Why is the centromere so cold? *Genome Res,* 8**,** 81-2.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M., Wides, R., Salzberg, S. L., Loftus, B., Yandell, M., Majoros, W. H., Rusch, D. B., Lai, Z., Kraft, C. L., Abril, J. F., Anthouard, V., Arensburger, P., Atkinson, P. W., Baden, H., de Berardinis, V., Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chaturverdi, K., Christophides, G. K., Chrystal, M. A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C. A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R., Hillenmeyer, M. E., Hladun, S. L., Hogan, J. R., Hong, Y. S., Hoover, J.,

Jaillon, O., Ke, Z., Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J. J., Lobo, N. F., Lopez, J. R., Malek, J. A., McIntosh, T. C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S. D., O'Brochta, D. A., Pfannkoch, C., Qi, R., Regier, M. A., Remington, K., Shao, H., Sharakhova, M. V., Sitter, C. D., Shetty, J., Smith, T. J., Strong, R., Sun, J., Thomasova, D., Ton, L. Q., Topalis, P., Tu, Z., Unger, M. F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K. J., Wortman, J. R., Wu, M., Yao, A., Zdobnov, E. M., Zhang, H., Zhao, Q., *et al.* 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science,* 298**,** 129-49.

Huckins, L. M., Boraska, V., Franklin, C. S., Floyd, J. A., Southam, L., Gcan, Wtccc, Sullivan, P. F., Bulik, C. M., Collier, D. A., Tyler-Smith, C., Zeggini, E., Tachmazidou, I., Gcan & Wtccc 2014. Using ancestry-informative markers to identify fine structure across 15 populations of European origin. *Eur J Hum Genet,* 22**,** 1190-200.

Mahtani, M. M. & Willard, H. F. 1990. Pulsed-field gel analysis of alpha-satellite DNA at the human X chromosome centromere: high-frequency polymorphisms and array size estimate. *Genomics,* 7**,** 607-13.

Neafsey, D. E., Christophides, G. K., Collins, F. H., Emrich, S. J., Fontaine, M. C., Gelbart, W., Hahn, M. W., Howell, P. I., Kafatos, F. C., Lawson, D., Muskavitch, M. A., Waterhouse, R. M., Williams, L. J. & Besansky, N. J. 2013. The evolution of the Anopheles 16 genomes project. *G3 (Bethesda),* 3**,** 1191-4.

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arca, B., Arensburger, P., Artemov, G.,

Assour, L. A., Basseri, H., Berlin, A., Birren, B. W., Blandin, S. A., Brockman, A. I., Burkot, T. R., Burt, A., Chan, C. S., Chauve, C., Chiu, J. C., Christensen, M., Costantini, C., Davidson, V. L., Deligianni, E., Dottorini, T., Dritsou, V., Gabriel, S. B., Guelbeogo, W. M., Hall, A. B., Han, M. V., Hlaing, T., Hughes, D. S., Jenkins, A. M., Jiang, X., Jungreis, I., Kakani, E. G., Kamali, M., Kemppainen, P., Kennedy, R. C., Kirmitzoglou, I. K., Koekemoer, L. L., Laban, N., Langridge, N., Lawniczak, M. K., Lirakis, M., Lobo, N. F., Lowy, E., MacCallum, R. M., Mao, C., Maslen, G., Mbogo, C., McCarthy, J., Michel, K., Mitchell, S. N., Moore, W., Murphy, K. A., Naumenko, A. N., Nolan, T., Novoa, E. M., O'Loughlin, S., Oringanje, C., Oshaghi, M. A., Pakpour, N., Papathanos, P. A., Peery, A. N., Povelones, M., Prakash, A., Price, D. P., Rajaraman, A., Reimer, L. J., Rinker, D. C., Rokas, A., Russell, T. L., Sagnon, N., Sharakhova, M. V., Shea, T., Simao, F. A., Simard, F., Slotman, M. A., Somboon, P., Stegniy, V., Struchiner, C. J., Thomas, G. W., Tojo, M., Topalis, P., Tubio, J. M., Unger, M. F., Vontas, J., Walton, C., Wilding, C. S., Willis, J. H., Wu, Y. C., Yan, G., Zdobnov, E. M., Zhou, X., Catteruccia, F., Christophides, G. K., Collins, F. H., Cornman, R. S., *et al.* 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science,* 347**,** 1258522.

Tandon, A., Patterson, N. & Reich, D. 2011. Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays. *Genet Epidemiol,* 35**,** 80-3.

Tian, C., Kosoy, R., Nassir, R., Lee, A., Villoslada, P., Klareskog, L., Hammarstrom, L., Garchon, H. J., Pulver, A. E., Ransom, M., Gregersen, P. K. &

Seldin, M. F. 2009. European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. *Mol Med,* 15**,** 371-83.

Wang, L. J., Zhang, C. W., Su, S. C., Chen, H. H., Chiu, Y. C., Lai, Z., Bouamar, H., Ramirez, A. G., Cigarroa, F. G., Sun, L. Z. & Chen, Y. 2019. An ancestry informative marker panel design for individual ancestry estimation of Hispanic population using whole exome sequencing data. *BMC Genomics,* 20**,** 1007.

Zhang, X., Mu, W., Liu, C. & Zhang, W. 2014. Ancestry-informative markers for African Americans based on the Affymetrix Pan-African genotyping array. PeerJ, 2, e660.

Zheng, Y. & Janke, A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. BMC Bioinformatics, 19, 10.

Chapter IV.

Copy Number Variation

This chapter has been prepared for submission as a manuscript

Abstract

Resistance to public health insecticides in mosquitoes is a major concern for disease control and prevention across multiple continents, yet the genetic mechanisms of this resistance are poorly understood. Copy number variants (CNVs) are a form of genetic polymorphism that can influence gene expression, coding sequence and zygosity, and represent an understudied area of research in Anopheles mosquitoes. We investigate the presence and role of CNVs in the malaria vector *An. arabiensis*, using whole-genome sequence data from phase 3 of the *An. gambiae* 1000 genomes project. In a population of *An. arabiensis from Uganda,* we found CNVs encompassing a total of 60 genes, of which 3 were in gene families previously associated with metabolic insecticide resistance, including one in the glutathione S transferase *Gstd3*. We developed a PCR assay to detect th*is* CNV in further samples and found it to be at a frequency of 15.79% in samples from Uganda collected in 2012. We found no association between the presence of the *Gstd3* CNV and resistance to permethrin, leaving the possibility of a role in resistance to other insecticides. Ultimately, further research is needed both in *An. gambiae* and *An. arabiensis* to understand the roles of CNVs in resistance.

Introduction

Malaria control has been a major component in public health strategies for many years and has gained increased funding focus since its inclusion in both the Millennium Development Goals, and the subsequent Sustainable Development Goals. These are targets set by the UN and aim to catalyse activities with a view to wiping out poverty, fighting inequality and tackling climate change (UN, 2020). Despite heightened efforts and substantial progress in malaria management, an estimated ~409,000 annual deaths are still attributable to malaria and its associated pathologies (WHO, 2020).

The most important components of malaria control strategies are insecticide-based interventions, such as long-lasting insecticidal nets (LLIN) and indoor residual spraying (IRS), which together have helped avert ~450 million cases of clinical malaria, since 2000 (Bhatt *et al.*, 2015; WHO, 2016). Unfortunately, the efficacy of insecticide-based interventions is in danger of being compromised in many regions across Africa, Asia and South America, due to the emergence of insecticide resistance (Kafy *et al.*, 2017; Kigozi *et al.*, 2012; WHO, 2018).

To mitigate the effects of insecticide resistance, many studies are focused on characterising the molecular mechanisms which cause it (Corbel *et al.*, 2019). Elucidating these mechanisms allows for targeted interventions and strategic deployment of in-country resources to help maximise the reduction in malaria cases. Such strategies include the rotation of active ingredients (AI) to prevent any one resistance mechanism from reaching high frequencies, replacement of AI to

which resistance has become widespread, deployment of synergists (e.g. piperonyl butoxide (PBO)) that inhibit existing resistance mechanisms or increased IRS targeting.

Given that both LLINs and IRS are deployed within homes, they are the best suited to controlling the mosquitoes which feed and rest indoors, such as *An. gambiae* and *An. funestus* - the two vector species in which insecticide resistance is best understood (Alegana *et al.*, 2016; Helinski *et al.*, 2015; Kabbale *et al.*, 2013; Lynd *et al.*, 2019; Okia *et al.*, 2018; Oxborough *et al.*, 2019). In contrast, mosquitoes with less endophilic tendencies are less effectively controlled by these methods. For example, *An. arabiensis* feed on both humans and animals indoors and outdoors. Therefore, efforts to further understand and control malaria vectors need to include consideration for outdoor transmission and the species which exhibit such behaviours, particularly in view of evidence that shows resistance in such species (Mawejje *et al.*, 2013; Isaacs *et al.*, 2018; Ismail *et al.*, 2018).

The use of LLINs and IRS has been effective indoors, however, this alone has not been able to eradicate malaria. In some places, incidence is still high despite *An. gambiae* catch rates being very low, suggesting that these interventions have been successful in tackling indoor transmitters, but that malaria continues to be transmitted outdoors (Sherrard-Smith *et al.*, 2019). This residual transmission may be due to the persistence of outdoor biting species, suggesting that further intervention is necessary to control these species (Killeen *et al.*, 2017) Residual transmission is the term given such malaria transmission that persists once LLINs and IRS have attained universal coverage (Killeen, 2014). A study

evaluating the effect of a pirimiphos-methyl (Actellic 300CS) IRS campaign in Kenya found that the number of *An. funestus* collected per house post-intervention was significantly reduced compared to pre-intervention surveys, whereas no significant change was detected in *An. arabiensis* (Abong'o *et al.*, 2020).

Following an entomological survey and LLIN distribution campaign in Uganda in 2017, 6-, 12- and 18-month surveys showed an overall decrease in *An. gambiae* population number whilst *An. arabiensis* population counts were unchanged (Lynd 2020, personal communication). The increased relative proportion of *An. arabiensis* may result in an increased epidemiological importance of this vector. Since insecticides will play an important role in control of outdoor biting mosquitoes, there is an urgent need to understand the genetic mechanisms of resistance in these species.

Copy number variation (CNV) is a form of structural mutation consisting of a change in the number of copies of a unit of genetic code. CNVs result from either deletion or duplication of genetic material, potentially altering the expression and/or structural conformation of coding sequences (Perry, 2008). Amplification, an increase in copy number of a genetic region, can cause increased expression of a given protein if the CNV covers the whole coding portion of that protein's gene sequence (Zhou *et al.*, 2017). Further, duplication or deletion of an incomplete section of a gene sequence can cause major structural conformation changes in the translated protein (Vaszko *et al.*, 2016).

Heterologous gene duplication describes a case of gene amplification with allelic variation between the copies on the same chromatid. For example, the

G119S point mutation in acetocholinesterase-1 (*Ace1*) changes the target site of both carbamate and organophosphate insecticides. This causes a marked resistance phenotype, which is strongly associated with a fitness cost, caused by the overall reduced efficacy of *Ace1* - an essential enzyme involved in nerve signal transmission (Alout *et al.*, 2008; Weetman *et al.*, 2018). Homozygous *Ace1*-119S mosquitoes survive in the presence of insecticidal pressure despite the increased fitness cost over heterozygotes; however, homozygotes are outcompeted in the absence of such pressure (Djogbenou *et al.*, 2015). Interestingly, all resistant *Ace1 alleles* appear to be duplicated, and this may be part of a heterologous or homologous duplication (Weetman et al., 2015; Grau Bove et al., 2020)

Heterozygotes bear a lesser fitness cost, but at a decreased resistance capacity (Assogba *et al.*, 2015). Heterologous duplication of the *Ace1* gene, which combines the mutant and wild-type alleles on the same haplotype, therefore, at least in part, mitigates the fitness cost of the resistant genotype (Assogba *et al.*, 2015). This 'permanent heterozygosity' at the locus has occurred multiple times in distinct species (Assogba *et al.*, 2015; Berticat *et al.*, 2002; Bourguet *et al.*, 2004; Weill *et al.*, 2004a; Weill *et al.*, 2004b).

An. arabiensis is an increasingly important vector species and more work is needed to characterise the mechanisms and components of the unexplained variance in resistance phenotypes. Recent work on *An. gambiae* whole-genome sequence data showed widespread gene duplication at loci associated with insecticide detoxification (Lucas *et al.*, 2019). Using the resources of *An. gambiae 1000 genomes* (Ag1000G) project *(The Anopheles gambiae 1000 Genomes*

Consortium, 2017) phase 3, we investigated whether this was the case in *An. arabiensis*.

Results

Using Ugandan *An. arabiensis* samples from the Ag1000G project, we identified 272 distinct CNVs, with 60 genes covered by at least one CNV. The overall distribution of the 272 CNVs by chromosome is: 2L – 52, 2R – 54, 3L – 62, 3R – 53, X – 51. Information about the full 272 CNVs can be found in the Additional Materials. Of the 60 genes found within CNV regions, 52 of the genes had no annotation. Three CNVs were identified as including genes tagged as potential metabolic detoxification genes (Table 1). Each of these CNVs contained a single detoxification gene.

Gene ontology (GO) enrichment analysis revealed no significantly overrepresented GO terms found within the genes of CNV regions after multiple correction to a Q-value threshold of 0.05. Simulations which randomised the sets of genes included within CNV regions indicated that CNVs may be marginally enriched for detoxification gene families. Only 432 out of 10,000 (i.e. <5%) simulations produced as many as three detoxification genes in CNV regions.

Table 2. Summary of 8 CNVs which contained a known detox gene. % frequency is calculated as the proportion of individuals in the population (the total number of *An. arabiensis* samples) carrying the CNV.

CNV name	Gene name	Freq. (%)	Chromosome
CNV_2R0003	*Cyp325b1* (AGAP002210)	7.89	2R
CNV_2R0035	*Gstd3* (AGAP004382)	15.79	2R
CNV_2R0018	*Vg* (AGAP004203)	100	2R
CNV_2R0033	*Abca7* (AGAP001523)	5.26	2R
CNV_2R0044	*Or26* (AGAP004354)	22.37	2R
CNV_3L0090	*Clipe7* (AGAP011786)	15.79	3L
CNV_X0028	*RpS10* (AGAP000739)	26.32	X
CNV_X0048	*Gstt1* (AGAP000761)	11.84	X

All 3 metabolic detoxification genes found within CNVs were from gene families that are implicated in resistance to insecticides. *Cyp325b1* belonging to the cytochrome P450 family and *Gstt1* and *Gstd3* belonging to the glutathione S-transferase family. None of these three genes were found duplicated in a previous study in *An. gambiae* and *An. coluzzii* (Lucas *et al.*, 2020). Of the CNVs we identified with genes tagged as detoxification genes, the one containing *Gstd3* was the most frequent (15.79%). We therefore used the reads aligning over and around the breakpoints of this CNV to precisely identify its genomic extent (Figure 3) and used this information to design a PCR assay to screen for the presence of this CNV and test for a possible association with insecticide resistance.

We tested 252 *An. arabiensis* samples from a 2012 collection of samples collected in Tororo, Uganda and tested for resistance against permethrin (Mawejje *et al.*, 2013; Wilding *et al.*, 2015). We found no significant difference between

resistance and susceptible samples in relation to their CNV state ($P = 0.48$) (Figure 1).

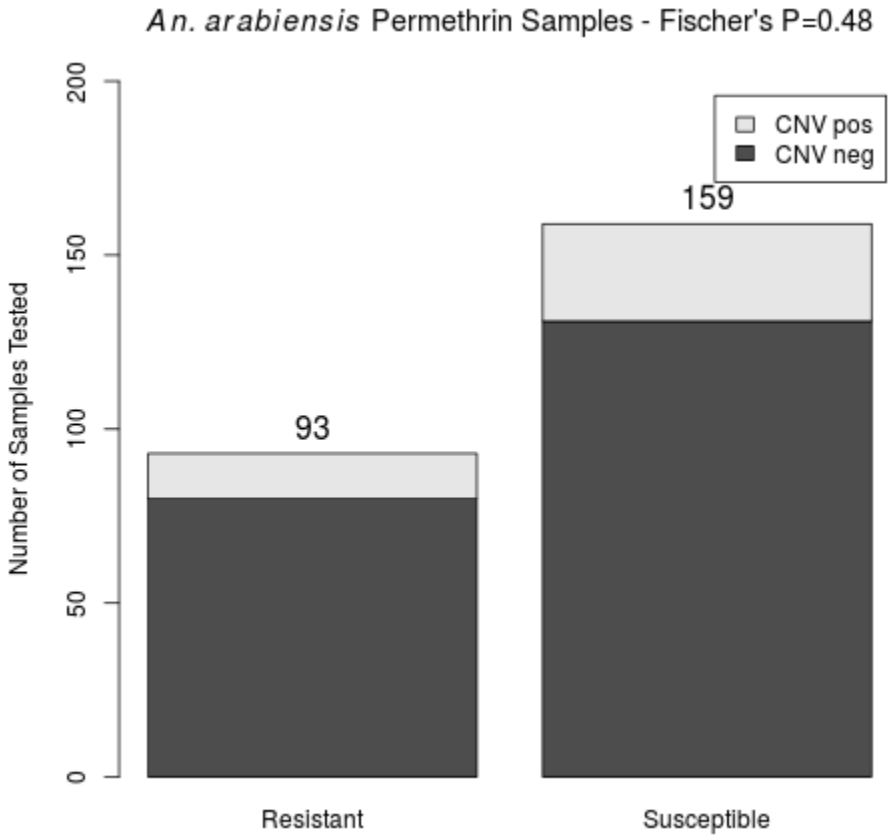


Figure 1. Association testing of *An. arabiensis* phenotyped against Permethrin. 93 resistant and 159 susceptible samples were genotyped for the Gstd3-containing CNV. No significant difference was found between samples that were CNV positive compared to CNV -negative samples in terms of resistance/susceptibility to Permethrin.

Discussion

Understanding the role of CNVs in insecticide resistance in medically important vector species may be key to developing strategies to mitigate the mortality and morbidity of the disease they transmit. Certainly, this is pertinent when considering residual transmission in regions with universal coverage of

LLINs and IRS. Previous work characterising the CNVs in *An. gambiae* identified 316 CNVs in Ugandan *An. gambiae*, with 6 of these containing detoxification genes, using the same a priori list of detoxification genes as used in this study (Lucas et al., 2019). We identified 272 CNVs in *An. arabiensis* (Tororo, Uganda), with 3 containing genes linked to detoxification. GO terms associated with cytochrome P450s were found to be significantly overrepresented in a combined study of *An. gambiae and An. coluzzii* CNVs, whereas we saw no significant GO term enrichment in CNVs in *An. arabiensis*. However, when *An. gambiae* data were rerun to only include samples from Uganda, the results fall below accepted significance levels, leaving no over-represented GO-terms in *An. gambiae*. Therefore, while our results contrast with those found in *An. gambiae/An. coluzzii* as a whole, this appears to be driven primarily by the wider range of populations that were represented in the Ag1000G phase 2 analyses (Lucas *et al.*, 2019; The *Anopheles gambiae* 1000 Genomes Consortium, 2017). In particular, a large cluster of P450s around *Cyp6p3* were found in CNVs from West African populations of *An. coluzzii*, which may have been largely responsible for the significant enrichment of P450-related GO terms.

Nevertheless, it is notable that none of the CNVs identified in the present study included any of the commonly reported detox genes, and that there was no overlap between the list of metabolic detoxification genes found in CNVs in this study and in Lucas *et al.* (2020). Despite the marginally significant enrichment in metabolic detoxification genes that we found, our results as a whole suggest that selective pressure on resistance-associated CNVs may be lower in *An. arabiensis*

than An. gambiae/An. coluzzii. We had anticipated to observe significantly fewer CNVs in *An. arabiensis* than in *An. gambiae* due to the lack of an *An. arabiensis* specific reference genome, potentially caused by low percentage of homology differences between the reference assemblies. The *An. arabiensis* samples used here at the time of analyses were aligned to the AgamP4 PEST reference genome assembly, based on *An. gambiae*. However, we observed similar results in QC of Ugandan *An. arabiensis* compared to Ugandan *An. gambiae*, suggesting this is likely not a factor.

The role of *Gstd3* in conveying a resistance phenotype to public health insecticides is not well described. Indeed, the inclusion of *Gstd3* as a candidate in insecticide resistance studies appears to be driven by its repeated appearance in microarray and qPCR data that shows it as over-expressed in resistant populations (Isaacs *et al.*, 2018; Tchigossou *et al.*, 2018). The identification of a CNV that duplicates the exonic regions of this locus may be of significant importance to public health research concerning insecticide resistance. However, in our experiments using permethrin resistant *An. arabiensis* from Jinja (2011) for the *Gstd3*-containing CNV, we were not able to identify an association between the presence of the CNV and resistance to permethrin. This may be because such CNV does not convey functional resistance due to the truncation of the second intronic region or an insecticide other than permethrin is conferred resistance by *Gstd3*. Further work is needed to characterise the function of *Gstd3*-containing CNVs in conferring resistance to insecticide, the distribution of the CNV and to examine the other two CNVs containing deoxification genes identified in this study.

Methods

Samples and whole-genome sequencing

For CNV discovery, we analysed data from 75 individual wild-caught, female *An. arabiensis* from Tororo, Uganda, collected in 2012 and sequenced as part of phase 3 of the Ag1000G project (The *Anopheles gambiae* 1000 Genomes Consortium, 2017). Specimens were sequenced using the Illumina HiSeq platform with a target of 30X coverage for 100-bp paired-end reads. Details of sampling and sequencing strategies of the Ag1000G have been previously published (The *Anopheles gambiae* 1000 Genomes Consortium 2017). For CNV validation by PCR and investigation of metabolic CNV resistance association we used samples from Jinja, Uganda; these samples have been previously described by Mawejje *et al.* (2013).

Coverage calculation and normalisation

The CNV-detection methods described in this study have been previously published (Lucas *et al.*, 2019). We used the *pysam* package in Python to identify aligned reads in non-overlapping 300 base-pair windows across the genome. To account for the inherent variation in coverage due to nucleotide composition, coverage of each window was normalised by taking into account the GC content of the region. The GC content of each 300bp window was computed with reference to the AgamP4 *An. gambiae* reference genome and used to normalise each window by the mean coverage over all autosomal windows containing the same

GC content. All *An. arabiensis* samples used here were aligned to this reference. For this normalisation, we removed windows with low accessibility (< 90% accessible bases (Miles *et al.*, 2016), based on the accessibility map for phase 2 as a map for phase 3 was unavailable). Among other criteria, this accessibility map removes sites that have unusually high or low coverage. Normalised coverage calculations were doubled to allow for genome regions with a typical diploid copy number to have an expected coverage of 2.

Before CNV discovery, we removed windows based on two filters. First, we excluded windows in which >2% of aligned reads had a mapping quality of 0. Secondly, we filtered windows where the GC content is rarely represented by the reference sequence, i.e. fewer than 100 with the same GC content. The first filter removes windows that contain unreliable coverage information because the reads aligning to them could belong elsewhere in the genome. The second removes windows for which coverage normalisation would be unreliable.

Copy-number variation discovery

We used a Gaussian Hidden Markov Model (HMM) to estimate the likely copy number state (CNS) of each window within each sample. This model takes the normalised coverage data as its input and identifies regions of elevated normalised coverage. Full details of the model employed here have been previously published (Lucas *et al.*, 2019). We called CNVs based on 5 contiguous windows of an amplified CNS (>2).

CNV filtering

After creating a list of initial CNV calls for each sample, we computed a measure of confidence for each CNV by calculating the relative likelihood of the predicted CNS against a null model in which the CNV was absent (CNS = 2). We then removed CNV calls with a likelihood ratio of <1000. To turn individual-level CNV calls into population-level data, it was necessary to determine when CNV calls in different individuals represented the same CNV allele. A CNV in different individuals was considered to be identical when the start and end points were no more than one window apart. CNVs found in only a few individuals are less likely to be of adaptive importance and more likely to be false positives than more common CNVs. We therefore imposed a further filter on the CNV calls by removing CNVs found in fewer than 4 of individuals.

Gene duplication and enrichment

We determined the genes contained in each CNV by comparing the start and end points of the CNV to the start and end point of all the genes listed in the AgamP4.2 gene annotations. The start and end points of a CNV were given to be the median of all the start and end points that were matched to it. Genes that were covered by less than 50% of filtered windows were excluded. If all filtered windows inside a given gene were also within the bounds of a given CNV, we classified that gene as being within the CNV. We used the list of genes as given in Lucas *et al.*, which was compiled by searching the AgamP4 genome annotation file for the terms “P450,” “glutathione S-transferase,” and “carboxylesterase”. The *topGO* R package was used to perform GO term enrichment analyses. The false discovery rates were calculated using the *fdrtool* R package.

Investigating metabolic CNVs

We characterised the *Gstd3* CNV in more detail to determine exactly where the CNV starts and ends, allowing diagnostics for the CNV both in-silico and in vitro by PCR. Tandem duplications can be identified in sequencing data through distinctive “face-away reads”, that is read pairs that map facing away from each other at the start and end points of the duplication (Figure 3). Having identified the approximate start and end points of the CNV using the change in coverage detected by the HMM we looked for face-away reads in this genomic region that were found in samples in which the CNV was present according to the HMM model. Similarly, the precise start and end point of the CNV (“breakpoints”) was determined by identifying soft-clipped reads that mapped in these same areas and were again found in samples in which the CNV was present. The breakpoint reads and the face-away reads were used to reconfirm previous analyses by checking all samples for these diagnostic reads. Samples with at least two of these reads were considered as having the duplication. There was a perfect overlap between the HMM-based calls and the diagnostic read-based calls.

To determine the exact genetic sequence around the CNV breakpoint, we took the soft-clipped bases from the reads mapping at the breakpoint and BLASTed them against the *An. gambiae* (taxid: 7165) RefSeq Represented genomes database. This revealed a 17 base pair insertion within the breakpoint. The recreated sequence around the CNV breakpoint was then used to design primers using Primer-BLAST. These primers were designed to detect the presence of the breakpoint and inserted region of the CNV (Figure 2).

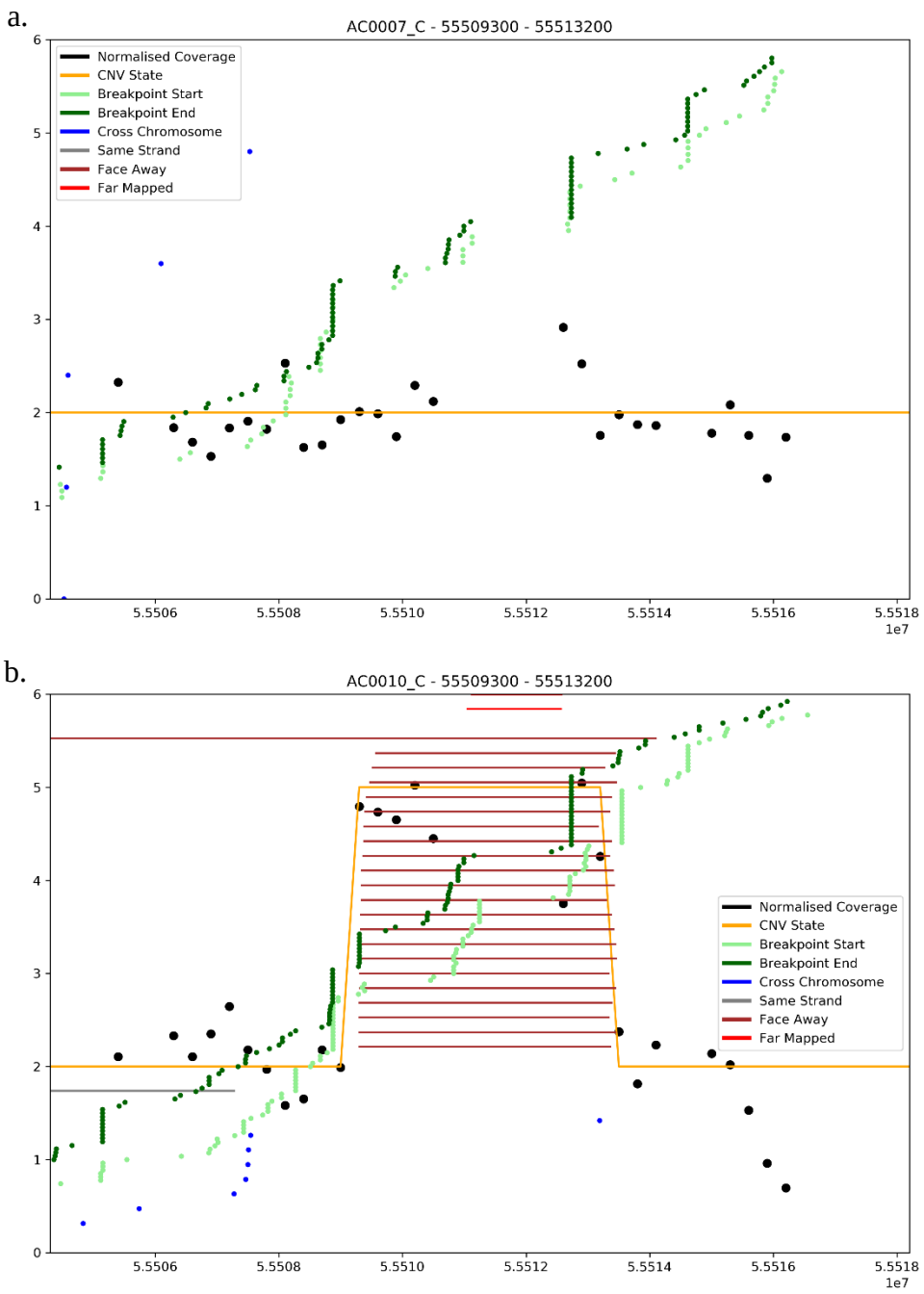


Figure 2. Example samples for Gstd3-containing CNV. The y-axis represents normalised coverage denoted primarily by the black dots which the HMM (orange line) uses to predict CNV state. Discordant reads have been overlayed over this representing face-away reads, far-mapped reads, same strand reads, cross chromosome reads and the detected breakpoints. a. Sample positive for the Gstd3-containing CNV. The

normalised coverage is elevated in the duplicated region and the HMM CNV state predicts an elevation of copy number in that region. Additionally, there are breakpoint start and ends that align with this region, further supported by face away reads surrounding the breakpoints. b. Sample not containing the CNV and the elements mentioned for panel a.

Genotyping of *Gstd3* by PCR

We validated the CNV-detection primers by basic PCR on the 6 of the 75 samples used to identify the CNV in silico, 3 containing the CNV and 3 not containing the CNV. This was done to validate that the PCR primer were working. Further, we performed the same PCR on samples from across Uganda which have not been whole genome sequenced and for which the CNV status for the *Gstd3* signal was unknown. The samples we used for this have been previously described by Mawejje *et al.*, (2013) and are 252 specimens from Uganda.

DNA was extracted from individual mosquitoes identified phenotypically as *Anopheles gambiae s.l.* using nexttec Biotechnologie GmbH extraction plates according to manufacturer's instructions and used as template for following PCR reactions. Mosquitoes were subsequently identified to species level by SINE (Santolamazza *et al.*, 2008), or by melt-curve based species identification (Chabi *et al.*, 2019).

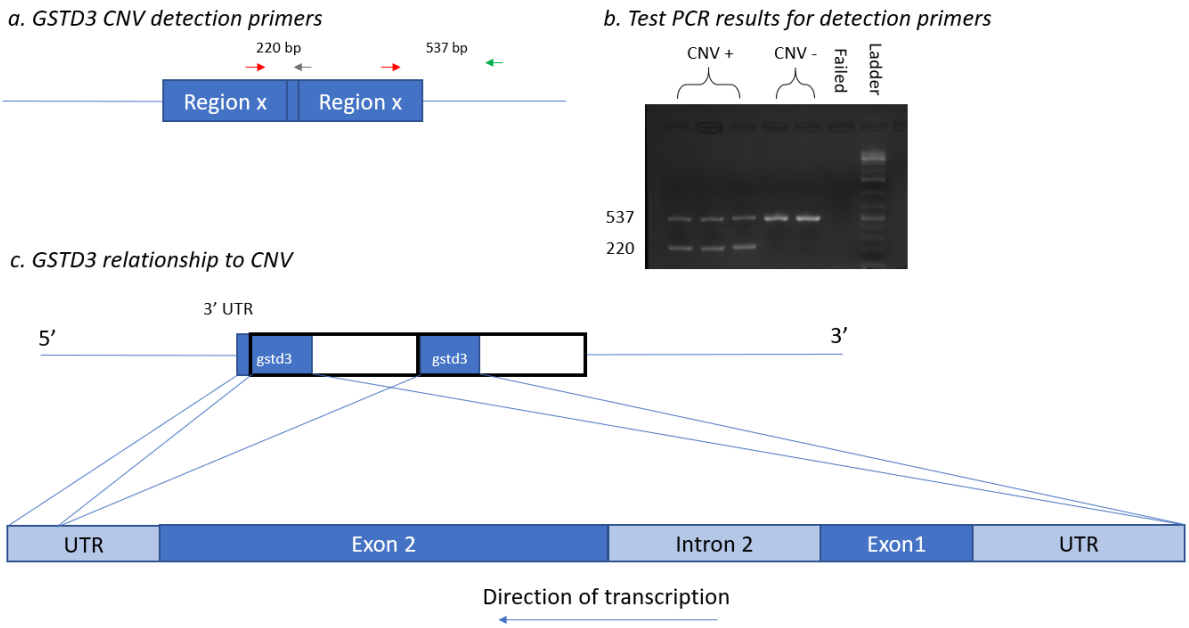


Figure 3 Gstd3 CNV structure and PCR assay design. a. Schematic relationship of the CNV region, inserted sequence and primers used to test for CNV presence. The red arrows denote the same forward primer that anneals to the same sequence and amplifies a control band and CNV band with two different reverse primers, where the control primer always anneals and the CNV primer only anneals when the CNV is present. b. Result of primer testing on 3 samples found to have the CNV in silico and 3 found to not have the CNV. The control band is present both in the presence and absence of the CNV, and the CNV band in only present in the CNV+ samples. c. Schematic relationship of the Gstd3 gene to the CNV. Gstd3 is transcribed on the strand complementary to the reference, leaving both exon regions, the intro region and the 5' UTR within the bound of the CNV. Only a small portion of the 3' UTR is not contained within the CNV.

PCR was carried out to detect the presence or absence of the CNV. Primers CNV1F, CNV1R and CNV1R-control were used to amplify a 220 bp fragment denoting the presence of the CNV and a 537 bp control fragment. The control band is designed to always be present even if the CNV is absent. The primer sequences

are CNV1F ACGACTCCCAAATGACGTGT with CNV1R GTTCCCGCACATTAAGGGAT for a product length of 220 with CNV1-Control CCCGAGTCCGAGAAATACCG for an additional product length of 537.

PCR was carried out using 1µl of DNA in a 15µl reaction volume with a final concentration of 1X Buffer, 0.33 mM dNTP's (Sigma dNTP-100), 0.33 µM of primer CNV1-F and 0.2 µM of primers CNV1-R and CNV1R-control (Integrated DNA technologies), Taq DNA polymerase 0.033 U/µl (DreamTaq Green DNA Polymerase (ThermoFisher Scientific)). Reaction conditions were 95°C for 3 min, 40 cycles of 95°C for 30 sec, 60°C for 30 sec, 72°C for 30 sec; and a final extension step of 72°C for 7 min. PCR reactions were performed in ABI GeneAmp PCR system 2700 or MJ Research PTC-200 DNA Engine thermal cyclers, with bands visualised on a 2% TAE agarose gel.

Statistics

Statistical analysis was performed in R (R Core Team 2015). Contingency tables were analyzed with the Fisher's exact and Chi-squared tests.

Acknowledgements

Author contributions: S.T., E.R.L., and M.J.D. designed the study. S.T. and E.R.L. carried out the analysis. The Ag1000G Consortium undertook collection, preparation, sequencing, and primary analysis of the samples. ST wrote the manuscript. ST and ERL performed bioinformatic analyses. ST and AL performed the laboratory work. MD participated in the study design and coordination and helped to draft the manuscript All authors read and approved the final manuscript.

References

Alegana, V. A., Kigozi, S. P., Nankabirwa, J., Arinaitwe, E., Kigozi, R., Mawejje, H., Kilama, M., Ruktanonchai, N. W., Ruktanonchai, C. W., Drakeley, C., Lindsay, S. W., Greenhouse, B., Kamya, M. R., Smith, D. L., Atkinson, P. M., Dorsey, G. & Tatem, A. J. 2016. Spatio-temporal analysis of malaria vector density from baseline through intervention in a high transmission setting. *Parasit Vectors,* 9**,** 637.

Assogba, B. S., Djogbenou, L. S., Milesi, P., Berthomieu, A., Perez, J., Ayala, D., Chandre, F., Makoutode, M., Labbe, P. & Weill, M. 2015. An ace-1 gene duplication resorbs the fitness cost associated with resistance in *Anopheles gambiae*, the main malaria mosquito. *Sci Rep,* 5**,** 14529.

Berticat, C., Boquien, G., Raymond, M. & Chevillon, C. 2002. Insecticide resistance genes induce a mating competition cost in Culex pipiens mosquitoes. *Genet Res,* 79**,** 41-7.

Bourguet, D., Guillemaud, T., Chevillon, C. & Raymond, M. 2004. Fitness costs of insecticide resistance in natural breeding sites of the mosquito Culex pipiens. *Evolution,* 58**,** 128-35.

Helinski, M. E., Nuwa, A., Protopopoff, N., Feldman, M., Ojuka, P., Oguttu, D. W., Abeku, T. A. & Meek, S. 2015. Entomological surveillance following a long-lasting insecticidal net universal coverage campaign in Midwestern Uganda. *Parasit Vectors,* 8**,** 458.

Isaacs, A. T., Mawejje, H. D., Tomlinson, S., Rigden, D. J. & Donnelly, M. J. 2018. Genome-wide transcriptional analyses in Anopheles mosquitoes reveal an unexpected association between salivary gland gene expression and insecticide resistance. *BMC Genomics,* 19**,** 225.

Ismail, B. A., Kafy, H. T., Sulieman, J. E., Subramaniam, K., Thomas, B., Mnzava, A., Abu Kassim, N. F., Ahmad, A. H., Knox, T. B., Kleinschmidt, I. & Donnelly, M. J. 2018. Temporal and spatial trends in insecticide resistance in *Anopheles arabiensis* in Sudan: outcomes from an evaluation of implications of insecticide resistance for malaria vector control. *Parasit Vectors,* 11**,** 122.

Kabbale, F. G., Akol, A. M., Kaddu, J. B. & Onapa, A. W. 2013. Biting patterns and seasonality of *Anopheles gambiae* sensu lato and Anopheles funestus mosquitoes in Kamuli District, Uganda. *Parasit Vectors,* 6**,** 340.

Kafy, H. T., Ismail, B. A., Mnzava, A. P., Lines, J., Abdin, M. S. E., Eltaher, J. S., Banaga, A. O., West, P., Bradley, J., Cook, J., Thomas, B., Subramaniam, K., Hemingway, J., Knox, T. B., Malik, E. M., Yukich, J. O., Donnelly, M. J. & Kleinschmidt, I. 2017. Impact of insecticide resistance in *Anopheles arabiensis* on malaria incidence and prevalence in Sudan and the costs of mitigation. *Proc Natl Acad Sci U S A,* 114**,** E11267-E11275.

Kigozi, R., Baxi, S. M., Gasasira, A., Sserwanga, A., Kakeeto, S., Nasr, S., Rubahika, D., Dissanayake, G., Kamya, M. R., Filler, S. & Dorsey, G. 2012. Indoor residual spraying of insecticide and malaria morbidity in a high transmission intensity area of Uganda. *PLoS One,* 7**,** e42857.

Killeen, G. F. 2014. Characterizing, controlling and eliminating residual malaria tran Killeen, G. F., Kiware, S. S., Okumu, F. O., Sinka, M. E., Moyes, C. L., Massey, N. C., Gething, P. W., Marshall, J. M., Chaccour, C. J. & Tusting, L. S. 2017. Going beyond personal protection against mosquito bites to eliminate malaria transmission: population suppression of malaria vectors that exploit both human and animal blood. *BMJ Glob Health,* 2**,** e000198.

smission. *Malar J,* 13**,** 330.

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

Lynd, A., Gonahasa, S., Staedke, S. G., Oruni, A., Maiteki-Sebuguzi, C., Dorsey, G., Opigo, J., Yeka, A., Katureebe, A., Kyohere, M., Hemingway, J., Kamya, M. R. & Donnelly, M. J. 2019. LLIN Evaluation in Uganda Project (LLINEUP): a cross-sectional survey of species diversity and insecticide resistance in 48 districts of Uganda. *Parasit Vectors,* 12**,** 94.

Mawejje, H. D., Wilding, C. S., Rippon, E. J., Hughes, A., Weetman, D. & Donnelly, M. J. 2013. Insecticide resistance monitoring of field-collected *Anopheles gambiae* s.l. populations from Jinja, eastern Uganda, identifies high levels of pyrethroid resistance. *Med Vet Entomol,* 27**,** 276-83.

Okia, M., Hoel, D. F., Kirunda, J., Rwakimari, J. B., Mpeka, B., Ambayo, D., Price, A., Oguttu, D. W., Okui, A. P. & Govere, J. 2018. Insecticide resistance status of the malaria mosquitoes: *Anopheles gambiae* and Anopheles funestus in eastern and northern Uganda. *Malar J,* 17**,** 157.

Oxborough, R. M., Seyoum, A., Yihdego, Y., Dabire, R., Gnanguenon, V., Wat'senga, F., Agossa, F. R., Yohannes, G., Coleman, S., Samdi, L. M., Diop, A., Faye, O., Magesa, S., Manjurano, A., Okia, M., Alyko, E., Masendu, H., Baber, I., Sovi, A., Rakotoson, J. D., Varela, K., Abong'o, B., Lucas, B., Fornadel, C. & Dengela, D. 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. *Malar J,* 18**,** 264.

The *Anopheles gambiae* 1000 Genomes Consortium 2017. Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature,* 552**,** 96-100.

Weill, M., Berthomieu, A., Berticat, C., Lutfalla, G., Negre, V., Pasteur, N., Philips, A., Leonetti, J. P., Fort, P. & Raymond, M. 2004a. Insecticide resistance: a silent base prediction. *Curr Biol,* 14**,** R552-3.

Weill, M., Malcolm, C., Chandre, F., Mogensen, K., Berthomieu, A., Marquine, M. & Raymond, M. 2004b. The unique mutation in ace-1 giving high insecticide resistance is easily detectable in mosquito vectors. *Insect Mol Biol,* 13**,** 1-7.

Weetman, D., Mitchell, S. N., Wilding, C. S., Birks, D. P., Yawson, A. E., Essandoh, J., Mawejje, H. D., Djogbenou, L. S., Steen, K., Rippon, E. J., Clarkson,

C. S., Field, S. G., Rigden, D. J. & Donnelly, M. J. 2015. Contemporary evolution of resistance at the major insecticide target site gene Ace-1 by mutation and copy number variation in the malaria mosquito Anopheles gambiae. *Mol Ecol,* 24**,** 2656-72.

WHO 2018. Global report on insecticide resistance in malaria vectors: 2010–2016.

Wilding, C. S., Weetman, D., Rippon, E. J., Steen, K., Mawejje, H. D., Barsukov, I. & Donnelly, M. J. 2015. Parallel evolution or purifying selection, not introgression, explains similarity in the pyrethroid detoxification linked GSTE4 of *Anopheles gambiae* and *An. arabiensis*. *Mol Genet Genomics,* 290**,** 201-15.

Chapter V.

Introgression

Chapter Overview

In this chapter I outline that malaria vector control faces two problems:

- Emergence of resistance
- Residual transmission by vectors about which we know little.

My work within this chapter addresses these by examining resistance in the residual vector *Anopheles arabiensis* and particularly whether introgression could be a major source of adaptive variation. This chapter is also being prepared as a manuscript for submission.

Code used throughout these analyses can be found at the public GitHub page: https://github.com/SeanTomlinson30/phd-ops

N.B. The samples collection, processing and sequencing discussed in the methods section of this chapter was carried out by the Ag1000G team members prior to my commencement of PhD studies at LSTM. All downstream computations and analyses were conducted within the purview of my PhD studies.

Introduction

Malaria transmission reduction relies heavily on insecticides, indeed the success in reducing disease incidence at the beginning of the 21st century is largely attributed to the effectiveness of long-lasting insecticidal nets and indoor residual spraying. Malaria eradication has been given enormous political capital as a corollary to efforts concerning raising the standards of global health; in 2016, US$4.3 billion was spent on malaria worldwide (Haakenstad *et al.*, 2019). Between 2000 and 2015 *Plasmodium falciparum* infection rates have dropped by 50% and the number of clinical cases has reduced by 40%, with a residual ~450,000 deaths per year are still attributed to malaria (Bhatt *et al.*, 2015).

The WHO global report on insecticide resistance in malaria vectors show that between 2010 and 2016, resistance to all 4 insecticide classes is distributed in all major vector species across Africa, the Americas, South-East Asia, the Eastern Mediterranean and the Western Pacific (WHO, 2018). Further, the WHO outline that of the 83 countries where malaria is endemic, all 73 countries that provided data for the 2010 – 2018 period showed resistance to at least 1 of the 4 public health insecticide classes (with 26 countries showing resistance to all 4 classes). The same analyses for the 2010 – 2017 period showed 5 fewer countries having resistance to at least 1 class (WHO, 2019).

The *Anopheles gambiae* 1000 genomes (Ag1000G) project is an international consortium using whole genome sequencing to develop a comprehensive source of genetic variation data in natural populations of Anopheles mosquitoes (*Anopheles gambiae* 1000 Genomes *Consortium et al.*,

2017). Phase 1 and 2 of the projects focused on the principal vectors *An. gambiae s.s.*, and *An. coluzzii* sequencing 1142 wild-caught specimens from across 13 countries and 234 individuals from 11 laboratory colony crosses. Phase 3 of the project aims to increase the number of specimens to ~3000 from 18 countries, these samples are to include *An. arabiensis* using the same preparation and sequencing pipeline.

These data provide a unique opportunity to study genetic introgression between *An. gambiae* and *An. arabiensis*. Data from both species have been exposed to the sample pipeline and are provided in the format for comparative analyses. Further, the reference genome assembly is the same (AgamP4 PEST), meaning that although analyses are unable to identify any signal unique to *An. arabiensis*, genetic regions that are shared between the two species are identifiable.

The capacity for introgression to occur between members of the *Anopheles gambiae* complex is directly relevant to public health. Resistance mechanisms to insecticides used for malaria vector control are genetically driven, highly advantageous and selected for by the environmental ubiquity of insecticides. Introgression has previously been shown to be a driver of both the evolution of mosquito species and to be involved in the spread of insecticide resistant alleles between populations and species (Bernardini *et al.*, 2019; Crawford *et al.*, 2015; Grau-Bove *et al.*, 2020; Hanemaaijer *et al.*, 2019; Norris *et al.*, 2015; Vicente *et al.*, 2017)

. In this chapter, I outline the methods and results of analyses that investigated the *An. gambiae* and *An. arabiensis* genomes for signals of introgression in East African samples from phase 3 of the Ag1000G project.

The use of long-lasting insecticidal nets and indoor residual spraying interventions have historically been employed to aid in the reduction of malaria. These tools are intended to provide lethal doses of insecticide to mosquitoes when they seek to feed on humans, rest on walls inside houses or enter other structures treated with similar chemical compositions such as livestock shelters. Despite these measures, and in addition to biological/biochemical insecticide resistance, malaria vectors can exhibit behavioural preferences that make them more suitable for a given environment (Killeen, 2014). Such behaviours include:

- Contact avoidance with insecticide treated surfaces inside houses.
- Feeding on humans during unprotected hours of the day.
- Feed on animals thereby avoiding insecticidal measures aimed at protecting humans.
- Resting outdoors out of range of insecticides.

Indeed, residual transmission is all forms of transmission that can persist through the attainment of universal coverage of LLINs and IRS, using insecticides to which the local vector populations are fully susceptible. Residual transmission in mosquitoes is a mechanism of insecticide subversion that ultimately threatens to thwart malaria elimination attempts.

In regions of *An. gambiae* and *An. arabiensis* sympatry with heavy use of insecticides through LLINs or IRS, *An. gambiae* mosquitoes risk exposure and

death whilst seeking a human bloodmeal. Whereas, although *An. arabiensis* blood feed on humans, they also feed on livestock. This feeding preference may explain data showing an overall decrease of *An. gambiae* and its replacement by *An. arabiensis*. In this sense, *An. arabiensis* is in a 'passive competition' with *An. gambiae*, utilizing their feeding preferences to avoid detrimental environments.

Following an entomological survey and LLIN distribution campaign in Uganda in 2017, 6-, 12- and 18-month subsequent surveys showed an overall decrease in *An. gambiae* abundance (Lynd 2020, unpublished). Further, a study evaluating the effect of Pirimiphos-Methyl (Actellic 300CS) IRS campaign in Kenya found that number of *An. funestus* identifies post-intervention was significantly reduced compared to pre-intervention surveys; whereas, no significant change in the mean number *An. arabiensis* was identified (Abong'o *et al.*, 2020).

Overall, there is growing interest in characterizing *An. arabiensis*, not only in terms of their feeding/hunting and breeding preferences, but also the genetic components that may exist which allow *An. arabiensis* to perform as a competent vector for Plasmodium infections in the presence of interventions that disproportionately act upon other local vector species. The resultant increased relative proportion of *An. arabiensis* may result in an increased epidemiological importance of this vector. This highlights the need to further characterise the genetic components of resistance and for targeted control interventions, part of which could include an insecticide resistance management strategy.

Both pre- and post-zygotic isolation mechanisms are present between *An. gambiae* and *An. arabiensis*. Wing-beat frequencies, swarming behaviour and

odour are thought to be involved in species recognition, but the precise mechanisms for prezygotic isolation remain an active area of research (Bernardini *et al.*, 2019). Haldane's rule makes the prediction that under a post-zygotic isolation model F1 hybrid males are sterile, whilst the homogametic sex (females) is fertile. Despite these isolation mechanisms, hybrids are detected in very low frequencies, typically reported between 0.02-0.76% where *An. gambiae* and *An. arabiensis* are sympatric. However, these frequencies may underestimate the true value due to ineffective identification of hybrids beyond F1 (Mawejje *et al.*, 2013; Temu *et al.*, 1997; Toure *et al.*, 1998; Weetman *et al.*, 2014).

Despite very distinct feeding, resting, and mating behaviours between key vector species, genetic divergence between them is relatively low. The impact of a very recent speciation event lends itself to such low divergence between species and leads to atypical scenarios, such as porous barriers to reproductive isolation. *An. gambiae* and *An. coluzzii* are the two most closely related species of the *An. gambiae* complex, with very recent speciation. Their designation into two separate species taxa was described by Coetzee *et al.* (2013). Prior to this, these species were termed 'molecular forms' of the same species with reports of incipient speciation, being found to be sharing alleles through introgression. Such evidence revealed the potentially porous nature of the reproductive barriers in Anophelines.

Indeed, introgression of the insecticide resistance mutation (Vgsc-1014F) located within a genomic island of divergence separating *An. gambiae* and *An. coluzzii* has been documented (Weill *et al.*, 2000; White *et al.*, 2010; Clarkson *et al.*, 2014). This gene flow led to the total homogenisation of the whole genomic

island with no detectable effect on the reproductive isolation between *An. gambiae* and *An. coluzzii*. The authors use these findings to highlight “how resilience of genomes to massive introgression can permit rapid adaptive response to anthropogenic selection and that even extreme prominence of genomic islands of divergence can be an unreliable indicator of importance in speciation” (Clarkson *et al.*, 2014). The elimination of divergence in the island of speciation and discovery of further regions introduced friction to the theory that the 2L and X genomic islands are the driver of incipient speciation. As such, the role and importance of genomic islands in speciation is spurious, however the presence and impact of interspecific gene flow is not (Aboagye-Antwi *et al.*, 2015).

An. arabiensis is a much more genetically distant species from *An. gambiae* than *An. coluzzii*. However, there still exists evidence of introgression and wild-caught hybrids, suggesting that despite clear divergence between *An. gambiae* and *An. arabiensis*, reproductive isolation between these species is porous. Such findings are directly relevant to public health when concerning loci that confound the effort of insecticide-based interventions to reduce malaria burden (Weetman et al., 2014). We investigate the potential for contemporary introgression to be driving insecticide resistance in *An. arabiensis* populations.

Methods

Data Collection

The *An. arabiensis* data used for these analyses were provided by the Ag1000G project phase 3. All specimens were collected in eastern Africa. The

dataset consisted of 583 wild-caught mosquitoes, with a population breakdown detailed in Table 3.

Table 3. Population breakdown of phase 3 samples used in introgression analyses. For the purposes of brevity and formatting these populations will take the shortened forms of a two-letter country code followed by three letters denoting the region and three letters denoting the species, i.e. Tanzania-Muleba-gambiae becomes TZ-MUL-ara.

Country	Region	Species	Count	Shortcode
Malawi	Chikwawa	arabiensis	33	MW-CHI-ara
Tanzania	Moshi	arabiensis	39	TZ-MOS-ara
Tanzania	Muheza	gambiae	32	TZ-MUH-gam
Tanzania	Muleba	arabiensis	118	TZ-MUL-ara
Tanzania	Muleba	gambiae	31	TZ-MUL-gam
Tanzania	Tarime	arabiensis	47	TZ-TAR-ara
Uganda	Kanungu	arabiensis	1	UG-KAN-ara
Uganda	Kanungu	gambiae	94	UG-KAN-gam
Uganda	Tororo	arabiensis	76	UG-TOR-ara
Uganda	Tororo	gambiae	112	UG-TOR-gam

As described in Chapter 2, all samples used throughout the analyses presented here underwent an additional quality control step. Briefly, this control step was used to apply the following filters to the data and exclude samples that

did not meet these requirements. Samples with erroneous/missing metadata that could not be corrected or located were excluded from analyses.

The Patterson's D statistic (ABBABABA) is a test for introgression that uses four populations, 3 in-groups and 1 outgroup. The test takes into account ancestral ("A") and derived ("B") alleles, predicting that under a model of incomplete lineage sorting, two patterns of SNPs – ABBA and BABA – will be balanced (Durand et al., 2011; Green et al., 2010; Kulathinal et al., 2009). An excess of either SNP pattern is considered to be indicative of introgression (Figure 2).

The tests assume a tree structure of (((P1,P2),P3),O) (where P1 = population 1 etc.) and using a haploid sequence (H2, H3, etc) for each population make a count of site where H2 and H3 shared a derived allele B and H1 has the ancestral state A, given by the outgroup haplotype – these are ABBA sites, BABA sites are the converse where H1 and H3 are tested against the presence/absence of the ancestral allele in H2. Under a model of no introgression a1:1 ratio of ABBA and BABA sites is expected. The D statistic represents the deviation from this balance, given by D = [sum(ABBA) – sum(BABA)] / [sum(ABBA) + sum(BABA)]. Outgroup data were obtained from (Neafsey *et al.*, 2013). We created a conversion script that prepared these data to the same format as phase 3 data (Appendix B). Transformations to the data consisted of removing indels, unifying the chromosome data and unifying the alternate allele encoding. Reads within the data contained positions where the aligner called either an insertion or a deletion, these features can confound analyses, so the data were limited to a single base per genomic coordinate. For each set of data some of the genome has no read

data associated with it, to ensure uniformity between the data we only retained position where both data sets had information for a given coordinate. The methods of encoding alternate alleles can either be ordered based on frequency or alphabetically from the remaining non-reference bases, the phase 3 data used alphabetical encoding and data from Fontaine *et al.,* (2015) used frequency encoding, we changed this to reflect the structure of the data from the Ag1000G project. The species outgroups obtained were *An. christyi*, *An. epiroticus*, *An. melas*, *An. merus* and *An. quadriannulatus.*

DNA extraction for each sample was performed using the Qiagen DNEasy Blood and Tissue Kit (Qiagen Science). Samples were sequenced using the Illumina HiSeq 200 platform with paired-end libraries, with a target coverage of 30X per specimen, sequencing was performed at the Wellcome Sanger Institute, UK. Details concerning the methods of sampling and sequencing have also been published previously (reference miles et al 2017 and 2019 ag1000g paper). Variant calling was performed with the Burrows-Wheeler Alignment tool *bwa 0.6.2* (Li and Durbin, 2009) and the *Genome Analyses ToolKit GATK 2.7.4 UnifiedGenotyper* module (Van der Auwera *et al.*, 2013). The reference genome used for this study was the PEST AgamP4 genome assembly, using gene annotations AgamP4.3 (Holt *et al.*, 2002; Sharakhova *et al.*, 2007).

Data Preparation

All analyses were performed using Python 3 and the language-associated pipeline tool Snakemake (Figure 1). We computed allele counts for each population of the phase 3 data and the Fontaine outgroup data using the *scikit-*

allel 1.2.1 library. Each computed allele counts array was saved to disc for future computation. For each possible combination of popA, popB, popC and popD where A and B represent two *An. gambiae* populations, C represents *An. arabiensis* and D represents an outgroup, allele counts were loaded and each position that was represented in all populations were retained. All non-biallelic positions were also dropped. From these remaining data, frequencies of the major alleles were computed and recorded in a table that consisted of N rows, given by the number of retained genomic positions and 4 columns given by the current populations defining A, B C and D for a given comparison.

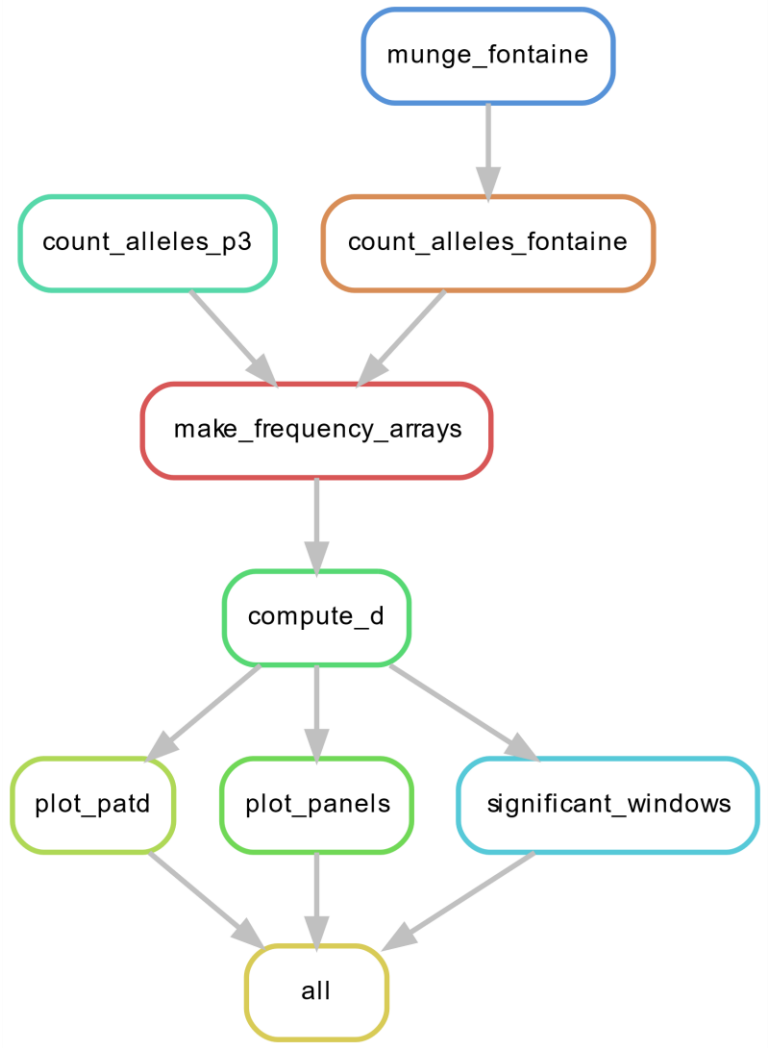


Figure 1. Overview of snakemake pipeline used for Patterson's D estimation. Each bubble represents a 'rule' in the snakemake workflow. These rules determine what actions can be performed given the data available and calculate what operations to perform if a given piece of data is absent. The 'all' rule represents the end of the analysis pipeline where all computations have taken place. Working backwards from this, the 'all' rule relies on the presence of the data output from the 'plot_patd', 'plot_panels' and 'significant_windows' rules. Working up this directed acyclic graph in the same fashion reveals that these three rules are take the output of the 'compute_d' rule, which in turn makes its calculations based on the frequency arrays produced by its parent rule. Above this rule is an origin rule that does not rely on any other rule output, but rather the given ata from the user. The beginning rule 'munge_fontaine' refers to the manipulation of data into the desired format, where it meets a parallel rule 'count_alleles_p3' as 'count alleles_fontaine' which creates the necessary alleles counts for making the frequency arrays and subsequent statistical calculations.

Patterson's D Statistic Calculation

Using a window size of 10,000 base-pairs we computed Patterson's D across each chromosome for each comparison (Figure 2). The statistics calculated for each window were the average D value, the standard error, Z-score and the value of the statistic in each block and the value of the statistic from block-jackknife resampling.

Using these statistics, we implemented a script to identify where windows identify a significant increase in Patterson's D at $p<0.001$, based on normalized Z-scores using the percent point function of the *scipy* package. The values of the Patterson's D statistic were plotted across the genome using *matplotlib.*

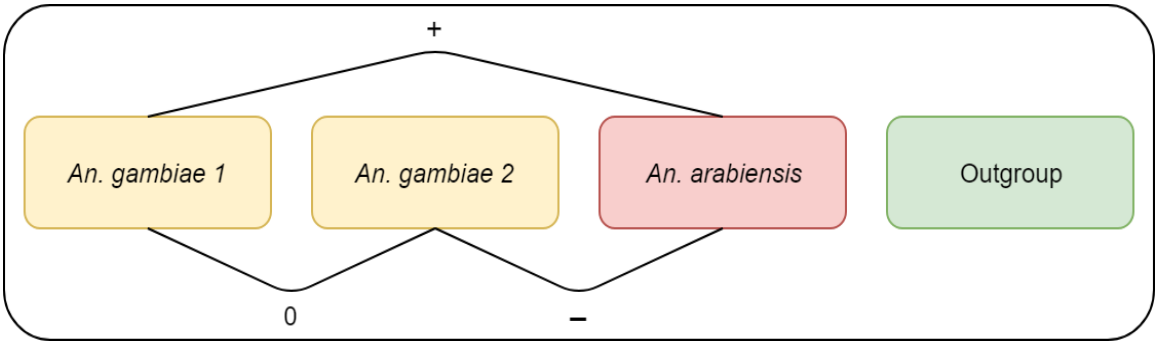


Figure 2. Summary of the Patterson's D test schema used in these analyses. Patterson's D statistic (also called ABBA-BABA) uses allele frequencies to test four given populations A, B, C and D. The test assumes the ((A, B), C), D phylogeny where groups A, B and C have possible gene flow, with D being an unintrogressed outgroup. When the D statistic is > 0 an excess of allele frequencies similarities is observed between group A and C. This could take the form of A > C introgression or C > A introgression (Figure 1). An excess of allele frequency similarities between B and C is indicated by D < 0. When D ≈ 0 the assumed phylogeny is balanced, and no signals of introgression are observed. Patterson's D statistic is not able to inform which direction (i.e. species) a given introgression signal originates.

Investigation of Significant Patterson's D Signals

To ensure stringency, we discarded all isolated significant windows, that is, we only marked a signal for further analyses if there were at least 2 significant windows in contiguous order across the genome. This approach reduces the number of type-1 errors (retaining false signals), with the risk of increasing type-2 errors (discarding true signals). The genomic coordinates of the start and end of the consecutive windows were recorded for each chromosome. The coordinates of consecutive significant Patterson's D were used to identify the genes contained within each region using the GenomeFeaturesFile obtained from VectorBase (Holt *et al.*, 2002; Sharakhova *et al.*, 2007). Visualisation of the connections between each significant Patterson's D statistic was performed using both the Python *networkx* library and the Ruby based *Circos* package.

2Rc Investigation

The inversion 2Rc is a polymorphic inversion spanning 26780000 - 31450000. The cytochrome P450 cluster is located within this range. Therefore, to investigate the potential introgression of this cluster, we needed to further characterise this inversion. To investigate the properties of the region encompassing the 2Rc inversion we calculated G_{min} across the region. G_{min} is a method of identifying genomic regions experiencing introgression in a secondary contact model. G_{min} is defined as the ratio of the minimum intra-population number

of nucleotide differences in a genomic window to the average number of between-population differences (Geneva *et al.*, 2015). Considering the coordinates 26000000 – 32500000 for each combination of popA and popB, where population A represents an *An. gambiae* population an B represents an *An. arabiensis* population. The values of G_{min} were plotted across these windows using matplotlib.

Results

Patterson's D Estimation

Examining the differences between the results of each comparison for all the possible outgroups (*An. christyi, An. epiroticus, An. melas, An. merus* and *An. quadriannulatus*) obtained from Neafsey *et al.* (2013), we observed a negligible difference in Patterson's D estimates and overall trends relating the number of significant windows discovered during the analysis stage of this work (Figure 3). From here we chose to proceed with a single set of outgroup comparisons for ease of displaying results and discussing their significance *An. christyi* was chosen from the set of outgroups.

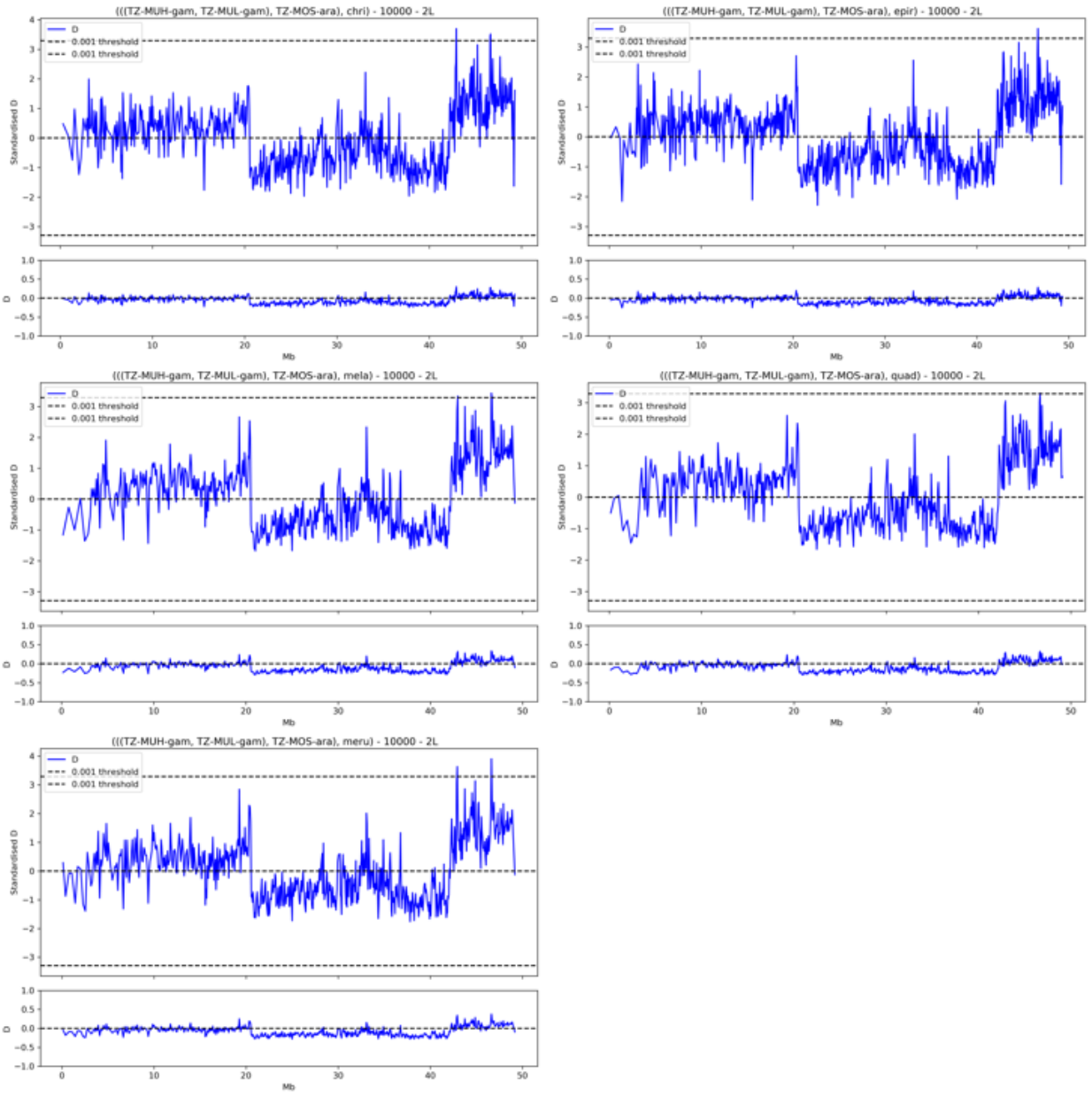


Figure 3. Comparison of the effects of differing outgroup species for Patterson's D statistic. This panel is an illustrative representation of the negligible difference in the five options for outgroup selection needed for the Patterson's D test. Each graph includes the raw D statistic and the standardised D statistic which is calculated to impose a $p<0.001$ threshold upon the data allowing a determination of significance.

Associating Genes with Significant Windows

Appendix B contains tables which represents a complete summary of the significant signals of Patterson's D statistic identifying throughout these analyses. Each signal of significant Patterson's D statistic identifies on chromosome arm 2L encompassed the same list of genes. These included a DNA topoisomerase, an exchange factor associated with neuronal function, hairy and enhancer-of-split proteins and 5 leucine-rich immune proteins (Appendix A, Table 4). Chromosome arm 2R had the greatest number of significant signals (hits) of Patterson's D statistics. Of note from the listed genes encompassed by these signals are the inclusion of known detoxification genes. These being a range of genes from the *cyp6* cytochrome P450 family and *COEAE60* carboxylesterase alpha esterase. Of the genes found contained within signals on chromosome arm 3L, none were identified as being associated with insecticide resistance or detoxification (Appendix A, Table 4). The genes are associated with cell maintenance and development pathways. Genes associated with insecticide detoxification were identified on chromosome arm 3R, all belonging to the glutathione-S-transferase Epsilon class of genes (Appendix A, Table 5). Finally, no genes on chromosome arm X were found that are known to be associated with insecticide resistance (Appendix A, Table 8).

Visualizing the Extent of Patterson's D Statistic

Using the data provided within Tables 1—5 of Appendix A, we plotted the start and end coordinate of each significant range of signals identified. Each line on these graphs represents a single entry in Appendix B Tables 2—6.

For chromosome arm 2L, the range of each significant signal are all overlapping, with a genomic extent of 46576227— 46698004 with an estimated midpoint centred around 46640000. The 2L chromosome features the 2La inversion spanning the genomic coordinates 20524058 — 42165532, when considering these signals, it will be key to consider that these signals are located within the break points of the 2La inversion. This is because polymorphic inversions in a population gives rise to signals on Patterson's D that are based on the inversion pushing a different set of alleles to the other end of the break point. Ultimately, this creates a nonsensical signal that can be countered by either ensuring the inversion is homomorphic, all samples have the same inversion status or working around the inversion in analysis synthesis. Chromosome arm 2R is the most signal-rich chromosome in these analyses. The data contain 4 principal regions of non-contiguous signal, yet all the signals are localized to be found within a single 1.7Mb window.

Signals identified on chromosome arm 3L consist of 4 non-contiguous regions of the genome. Of note, three of these signals were identified in only one A, B, C and D comparison set. Whereas the signal spanning the 5.3 – 5.6 Mb is represented in 6 total comparisons.

The largest extent of non-contiguous signals identified are found within chromosome arm 3R, with four unique regions having associated signals. These range from 10.1-10.15Mb with 4 supporting comparisons.

Signals identified on chromosome arm X, much like the signals found in chromosome arm 2L have a range which all signals are overlapping with a

genomic extent of 17.8Mb and 18.15Mb. A total of 3 comparisons identified a significant signal with an estimated midpoint of 18.025Mb.

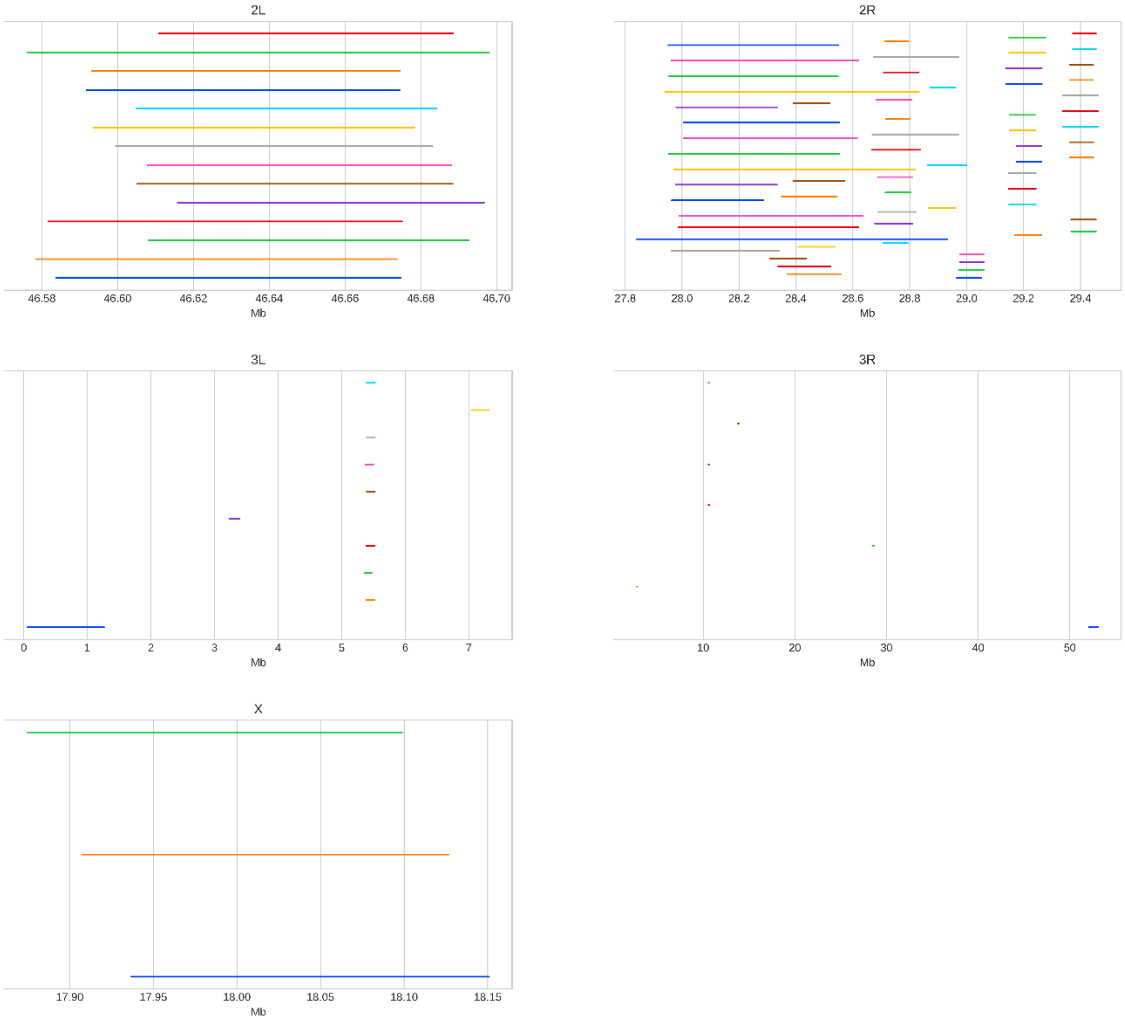


Figure 4. Panel figure showing the ranges of significant signals of introgression identified using Patterson's D statistic. The colours do not convey any data and are presented to help visually distinguish the extent of each range. Each line shows where along a given chromosome a significant signal of introgression was identified. The graphs are presented here to give a visual representation of both the clustering of significant signals on chromosomes such as 2L, 2R and X, compared chromosomes 3L and 3R which has signal spread across multiple discontinuous regions of the chromosome.

Visualising the Relationship of Patterson's D Signals

To give an overview of the relationships of Patterson's D statistic, two types of graphs were produced. First, we created a simple circular network graph that connects each node (population) with a line if there is a signal that supports a significant signal found between those two populations. These network graphs highlight the relationships of Patterson's D signal varies significantly between chromosome arms. Secondly, we created a circular genome plot with lines representing the section of the genome that gave rise to the associated significant signal. The difference in these figures is that the first demonstrates the number of connections to each population made by Patterson's D, whereas the second demonstrate the region to which signals belong.

For chromosome arm 2L, all populations have a connection to at least 3 other nodes, except for TZ-TAR-ara, which is only connected to two Ugandan *An. gambiae* nodes. The relationships of signals on chromosome arm 2R are very homogenous, each node is connected to at least three other nodes, with all the signals associated within the same region of the genomes. Chromosome 3L have signals of Patterson's D that converge onto a single population TZ-MUL-gam, with the exception of UG-KAN-gam which first connects to TZ-MUL-ara before then connecting to the principal TZ-MUL-gambiae node. All signals on 3L are localised to the first 7Mb even distributed along that extent. In a similar fashion data from chromosome 3R show populations converging to a single principal TZ-MUL-ara node. However, in this instance we can observe that only one population directly connects to this node, the remaining populations connect to secondary and tertiary

nodes before meeting the principal node. All nodes for chromosome X converge on TZ-MUL-gam, these being two Tanzanian and one Malawian *An. arabiensis* node.

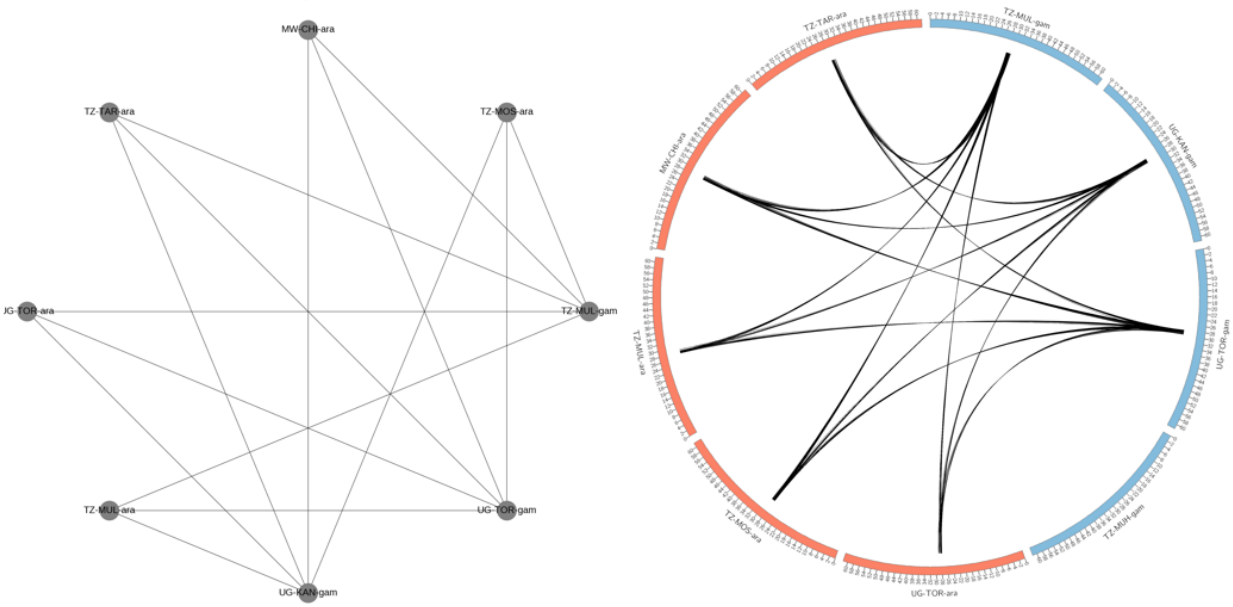


Figure 5 Network work and network genome graphs of significant signals found between *An. gambiae* and *An. arabiensis* on chromosome arm 2R. The network graph (left) is used to illustrate the number of and connections of significant signal of introgression. The circos network graph (right) is similar in that it conveys the same information; however, it also adds the dimension of where on the chromosome the significant signal is located. These graphs show near total ubiquity of the introgression signal on chromosome 2R with the Tanzania Muheza *An. gambiae* population not showing this signal. It also shows that the location of that ubiquitous signal is the same across all significant observations.

Investigation of the 2Rc region by G_{min}

Each combination of *An. gambiae*/*An. arabiensis* species pairs was used to calculate the G_{min} statistic across 10000 base pair sliding windows, across the range 26Mb – 31Mb of chromosome 2R. We observed that for 15 samples the

observed G_{min} value averages approximately 0.75 across the range, 0.5 for 3 samples. Moreover, three population pairs – Tanzania-Muheza *An. gambiae*/Malawi-Chikawa *An. arabiensis*, Tanzanian Muheza *An. gambiae*/Tanzania Moshi *An. arabiensis*, Tanzanian Muheza *An. gambiae*/Tanzania Muleba *An. arabiensis* – showed a significant drop in G_{min} between 28Mb and 29.3Mb (Figure 6). This region coincides with the signal of introgression identified by Patterson's D statistic. These findings were contrary to our expectation that the G_{min} statistic would mirror the signals identified in the population pairs through Patterson D.

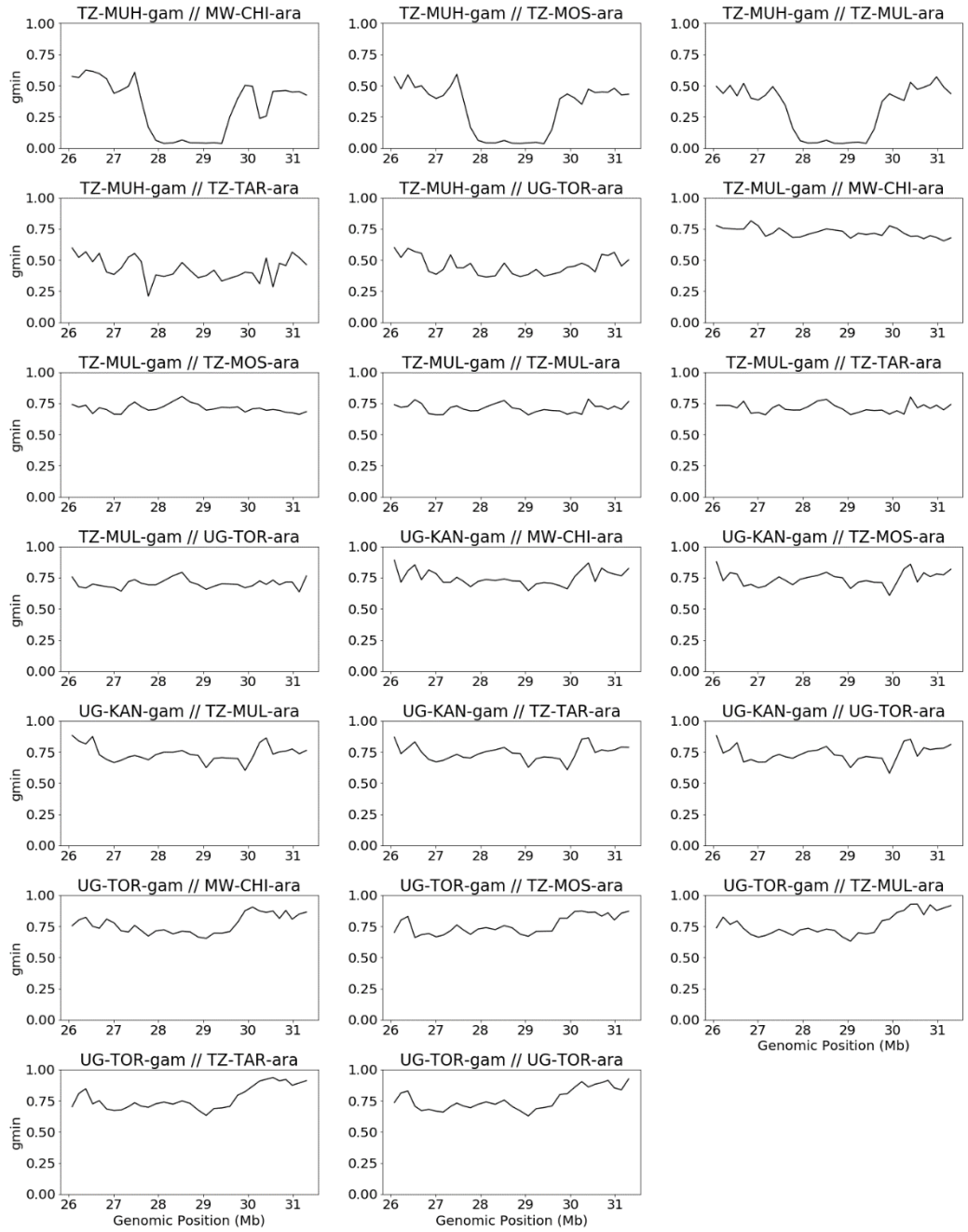


Figure 6. Panel figure showing the calculated G_{min} statistic across 10000 base pair windows, to encompass the 2R range 26 - 31 Mb. Each possible comparison of popA and popB is presented where popA represents the *An. gambiae* populations and *An. arabiensis* is given by popB. Akin to the adjacent F_{ST} statistic a G_{min} value of 0 indicates panmixis/introgression/share alleles within that window, where a value of 1 indicates complete and total isolation.

Discussion

The characterisation of the genetic landscape of both the Plasmodium parasite and its anopheline vector species is paramount to global health and the socioeconomical development of nations where malaria is endemic. Vector research as it relates to malaria concerns many areas, however, arguably the largest component is that of insecticide resistance. In the introduction to this chapter, we outlined the critical role that insecticides play in the control and mitigation of malaria. Resistance to such public health insecticide threatens to reverse the considerable success in the control and ablation of malaria transmission over the last 20 years. Here, we discuss the results from the analyses which sought to identify the presence of introgression between *An. gambiae* and *An. arabiensis* and assess the potential role insecticides resistance has in those introgression events.

Indeed, several introgression events were identified in the present analyses. The genomic range 2R 28000000 - 29000000 was the range of a near ubiquitous signal of introgression. It was identified that every *An. gambiae*/*An. arabiensis* population pair shared a significant signal of introgression within this range with the exception of Tanzania Muheza *An. gambiae* population. This finding is significant for the fact that this genomic range contains members of the cytochrome P450 gene family. This gene family is known to play a central role in insecticide resistance and is well established to be a highly polymorphic region undergoing segregation across multiple populations/species (Edi *et al.*, 2014;

Ibrahim *et al.*, 2015; Namountougou *et al.*, 2012; Nwane *et al.*, 2013; Park and Brown, 2002; Safi *et al.*, 2019; Silva Martins *et al.*, 2019; Weetman *et al.*, 2010).

The single population that did not possess this signal of introgression was the Tanzanian Muheza *An. gambiae* population. The reasons for this are not clear, however, this population is the only coastal one. This may be a representation of the fact that this dataset does not contain Tanzanian Muheza *An. arabiensis* samples, though possible this is unlikely the case, since the Tanzanian Tarime *An. arabiensis* population also does not have its *An. gambiae* counterpart and still shows the signal of introgression. However, it may be a feature of the fact that this is the only coastal population in our analyses. Certainly, geographical barriers to gene flow have been established in the literature; only such geographical feature is relevant to malaria in East Africa is the Rift Valley (Lehmann *et al.*, 2000).

Other introgression signals were identified amongst the other chromosomes. Chromosome 2L is an important chromosome in Anophelines and is strongly influenced by the presence of the 2La inversion across a large portion of its range. The significance of inversions is explored later in this discussion. Notwithstanding inversions, the genes encompassed by signals of introgression between *An. gambiae* and *An. arabiensis* were both ubiquitous and identical (Appendix D Figure 6). That is, all species pair comparisons (except for Tanzanian Muheza *An. gambiae*) contained the same list of genes that within the introgressed region (Appendix A, Table 2). None of these listed genes were found in the literature to be known associates of insecticide resistance in mosquitoes. However, mutations in DNA topoisomerases have been associated with resistance to

campthothecin, a plant alkaloid with applications in both chemotherapy and pest control (Zhang *et al.*, 2013). Further, glutathione-S-transferases are well documented in the literature to be associated with insecticide resistance (Edi *et al.*, 2014; Mitchell *et al.*, 2014). Indeed, the increase in pyrethroid resistance in the *An. coluzzii* population of Burkina Faso is linked to the increased expression of several gene family, including GSTs, overall, GSTE2, GSTE5, GSTM1, GSTMS3 and GSTS1-2 are all associated with insecticide resistance in this species (Toe *et al.*, 2015). Similarly, resistance in *An. gambiae* populations has been associated and or implicated with GSTD3, GSTE2 and GSTS1-2 (Isaacs *et al.*, 2018; Nardini *et al.*, 2017; Yahouedo *et al.*, 2017). Genes identified on autosome 3L and chromosome X were not identified to be associated with resistance in any capacity and more readily represented genes associated with cell maintenance and signalling (Appendix A, Table 6 and 8).

Inversion polymorphisms are a key part of the Anopheles genome and key features that must be considered in all genetic analyses. Certainly, when ascertaining the presence of introgression between species, inversion can confound results. For example, if a given region possesses an inversion polymorphism and is undergoing segregation in either of the populations or is fixed in only one of the populations, it would appear to the Patterson's D statistic that the calculated allele counts across the region are substantially different from the outgroup and report as a D statistic distal from zero. Supplementing these analyses with inversion karyotype data would serve to alleviate this confounding issue, since an informed comparison can be made between samples that share

compatible inversion statuses. However, at the time of analyses and writing the inversion karyotype data was not available for phase 3 for the Ag1000G. We explored methods of *in silico* calling of inversion status. Principal component analyses were previously used in analyses to karyotype samples where similar samples already possessed karyotype data (Grau-Bove *et al.*, 2020). Where samples with a known karyotype clustered with samples of an unknown karyotype cluster together, they were assigned to have the same karyotype. This method whilst utilised successfully in the literature by Grau-Bove *et al.* (2020), unfortunately was not applicable here since there were no karyotype data for *An. arabiensis* available for comparisons at time of analyses. Secondly, at the time of analyses Love *et al.* (2019) published methodology that can karyotype inversions completely *in silico* and agnostic of the need for samples to compare against. This method relies on tagging SNPs where biallelic genotypes are associated with inversion genotypes. The extent of the inversion 2Rc encompasses the introgression signal observed on chromosome 2R. Concerns were raised as the potential confounding effect of this inversion on the signal of introgression. However, in discussion with the primary author of Love *et al.* (2019) it was determined that the 2Rc inversion is absent in both *An. arabiensis* and *An. gambiae* in Eastern populations and has only been observed in Western populations. Therefore, it is assumed that the signal of introgression on chromosome 2R is unaffected by the 2Rc inversion. Further, no other signal of introgression associated with insecticide resistance is found within regions known to possess polymorphic chromosomal inversions.

Patterson's D statistic is not able to convey information regarding the timing or direction of any detected introgression. Further, the detection of introgression within a given locus need not represent a single event to that locus and may have be subject to multiple events gene flow events. Whether or not introgression detected in these analyses represent contemporary or historical gene flow is not determined by the analyses themselves and is largely based on analysing the surrounding data and making hypothesis which can be further investigated if supposed to be contemporary introgression. Contemporary introgression events are important to characterise since they represent a substantive change to the genetic background of the species and only increases in importance when it concerns insecticide resistance. Furthermore, evidence suggesting that residual transmission is occurring in sub-Saharan Africa lends credence to the suggestions that the targeting of a specific vector in one region can create the expansion of an alternate dominant vector species in the same niche. For example, following an entomological survey and LLIN distribution campaign in Uganda in 2017, 6-, 12- and 18-month surveys showed an overall decrease in *An. gambiae* population number whilst *An. arabiensis* population counts were unchanged (Lynd 2020, personal communication). The increased relative proportion of *An. arabiensis* may result in an increased epidemiological importance of this vector, leading to residual transmission. This was also observed in a study evaluating the effect of a Pirimiphos-Methyl (Actellic 300CS) IRS campaign in Kenya, which found the number of *An. funestus* collected per house identified post-intervention was significantly reduced compared to pre-intervention surveys, whereas no significant

change was detected in *An. arabiensis* (Abong'o *et al.*, 2020). Therefore, our hypothesis is that any identified *contemporary* introgression is likely to be in the direction that favours the species undergoing establishment in the local niche following the reduction of the competing species due to anthropogenic pressure. Since any introgression into the species undergoing the pressure is likely to cause an increased fitness to the local environment, by virtue of the fact that the species being establish is more apparently more fit.

Methods exist for determining the nature of both the direction and timing of introgression events. However, haplotypes are a required feature of the dataset and at the time of analyses, the Ag1000G had not produced haplotype phasing data for phase 3 which would encompass the *An. arabiensis* and *An. gambiae* samples used here. This certainly represents a limitation of the analyses presented here. However, the present results serve to indicate that further investigation is required once haplotype phase data becomes available, since we have identified introgression signals between *An. gambiae* and *An. arabiensis*, it becomes crucial for evaluating to what extent these finding impact the understanding of insecticide resistance and subsequently, the management of it.

Given the near total ubiquity of an introgression signal on chromosome 2R encompassing the CYP450 gene cluster, it was anticipated that the G_{min} statistic calculated across those windows would also capture this introgression. Though it appears this is the case for three species pair comparisons, the majority of comparisons across the windows in the region remain stable. This may be due to the unsuitability of the G_{min} statistic in this setting. Indeed, the authors who initially

described the statistic for use where secondary contact has occurred (Geneva *et al.*, 2015). That is, when sympatry of species is restored after some amount of time undergoing evolution in allopatry. Under a model of isolation, only a one ancestral lineage of each population is anticipated to remain, meaning that the ratio between the number of nucleotide differences between haplotypes sampled from different populations, and to the mean number of inter-population differences approaches unity. Only in cases where very recent divergence has occurred, is G_{min} expected to be less than unity. This is because not all coalescence events will occur between the ancestral lineages within a population before coalescence can occur between lineages of different populations. Ultimately, this may indicate that the introgression signals that are identified by Patterson's D statistic, but not indicated by the G_{min} statistic represent historical introgression events. Whereas, for the species pairs that are suggested as having undergone introgression by both Patterson's D and G_{min} statistics may represent contemporary introgression.

Fontaine et al. (2015) utilised reference genome assemblies to reconstruct and resolve the phylogenetic relationships and introgression in mosquito sibling species. They generated a species branching order 6 of the most medically important malaria vectors and showed that lineages leading to these principal vectors were the first to diverge. Their study was able to identify major widespread introgression between *An. coluzzii* > *An. arabiensis*, *An. arabiensis* > *An. gambiae* and *An. meras* > *An. quadriannulatus*. The extent of the introgression observed was so vast that only a small percentage of the whole genome had not crossed the species barrier. The majority of these genomic regions were located on the X

chromosome. One inference from Fontaine et al. (2015) is that vectorial capacity-enhancing traits may be gained by introgression.

The tree topologies generated strongly supported *An. arabiensis* being a sister species to *An. gambiae* and *An. coluzzii*. Indeed, in the Patterson's D statistical testing carried out by Fontaine et al. (2015) pervasive introgression across all autosomes between An. arabiensis and the ancestor of An. gambiae and An. coluzzii was observed. The authors note that scale and recent timing of the introgression impeded their ability to detect old introgression events. This is contrasted against conclusions drawn here where because of the localised regions of introgression and the specific focus on insecticide resistance we argued that introgression is likely to favour the local niche following the reduction of the competing species due to anthropogenic pressure.

We find that there exists evidence of introgression between *An. gambiae* and *An. arabiensis* and such introgression includes regions known to be associated with insecticide resistance. The analyses contained within this chapter were subject to limitations, discussed here; however, they represent both the value of and first step in characterising the role introgression plays in the evolution of vectors relevant to global health. With further maturation of the data and the public release of the fully curated phase 3 dataset, the questions addressed and raised by thesis can be further investigated.

Notwithstanding the limitations and various caveats to the results which have been discussed, we observed significant signals of introgression between *An. gambiae* and *An. arabiensis*. Given the evidence in the literature of the

increasing prevalence of the *An. arabiensis* population and studies focussing on residual transmission, our view is that future analyses should first and foremost aim to determine the direction and relative age of the introgression. These are two questions that were not answered due to the absence of haplotype phasing data during our analyses. Determining that these signals are both recent and donated to the *An. arabiensis* genome from *An. gambiae* would corroborate the hypothesis that increase use of insecticides that preferentially target *An. gambiae/An. coluzzi* over *An. arabiensis* can lead not only to residual transmission but the establishment of a new dominant vector species with resistance loci novel to that species. The implications of this on vector control could represent a significant development on the understanding of evolution to anthropogenic-mediated pressures.

References

Aboagye-Antwi, F., Alhafez, N., Weedall, G. D., Brothwood, J., Kandola, S., Paton, D., Fofana, A., Olohan, L., Betancourth, M. P., Ekechukwu, N. E., Baeshen, R., Traore, S. F., Diabate, A. & Tripet, F. 2015. Experimental swap of Anopheles gambiae's assortative mating preferences demonstrates key role of X-chromosome divergence island in incipient sympatric speciation. *PLoS Genet,* 11**,** e1005141

Abong'o, B., Gimnig, J. E., Torr, S. J., Longman, B., Omoke, D., Muchoki, M., Ter Kuile, F., Ochomo, E., Munga, S., Samuels, A. M., Njagi, K., Maas, J., Perry, R. T., Fornadel, C., Donnelly, M. J. & Oxborough, R. M. 2020. Impact of indoor residual spraying with pirimiphos-methyl (Actellic 300CS) on entomological

indicators of transmission and malaria case burden in Migori County, western Kenya. *Sci Rep,* 10**,** 4518.

Anopheles gambiae Genomes, C., Data analysis, g., Partner working, g., Sample, c.-A., Burkina, F., Cameroon, Gabon, Guinea, Guinea, B., Kenya, Uganda, Crosses, Sequencing, data, p., Web application, d. & Project, c. 2017. Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature,* 552**,** 96-100.

Bernardini, F., Kriezis, A., Galizi, R., Nolan, T. & Crisanti, A. 2019. Introgression of a synthetic sex ratio distortion system from *Anopheles gambiae* into *Anopheles arabiensis*. *Sci Rep,* 9**,** 5158.

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briet, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E. & Gething, P. W. 2015. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature,* 526**,** 207-211.

Clarkson, C. S., Weetman, D., Essandoh, J., Yawson, A. E., Maslen, G., Manske, M., Field, S. G., Webster, M., Antao, T., MacInnis, B., Kwiatkowski, D. & Donnelly, M. J. 2014. Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. *Nat Commun,* 5**,** 4248.

Coetzee, M., Hunt, R. H., Wilkerson, R., Della Torre, A., Coulibaly, M. B. & Besansky, N. J. 2013. Anopheles coluzzii and Anopheles amharicus, new members of the *Anopheles gambiae* complex. *Zootaxa,* 3619**,** 246-74.

Crawford, J. E., Riehle, M. M., Guelbeogo, W. M., Gneme, A., Sagnon, N., Vernick, K. D., Nielsen, R. & Lazzaro, B. P. 2015. Reticulate Speciation and Barriers to Introgression in the Anopheles gambiae Species Complex. *Genome Biol Evol,* 7**,** 3116-31.

Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. 2011. Testing for ancient admixture between closely related populations. *Mol Biol Evol,* 28**,** 2239-52.

Edi, C. V., Djogbenou, L., Jenkins, A. M., Regna, K., Muskavitch, M. A., Poupardin, R., Jones, C. M., Essandoh, J., Ketoh, G. K., Paine, M. J., Koudou, B. G., Donnelly, M. J., Ranson, H. & Weetman, D. 2014. CYP6 P450 enzymes and ACE-1 duplication produce extreme and multiple insecticide resistance in the malaria mosquito *Anopheles gambiae*. *PLoS Genet,* 10**,** e1004236.

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y. C., Smith, H. A., Love, R. R., Lawniczak, M. K., Slotman, M. A., Emrich, S. J., Hahn, M. W. & Besansky, N. J. 2015. Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science,* 347**,** 1258524.

Geneva, A. J., Muirhead, C. A., Kingan, S. B. & Garrigan, D. 2015. A new method to scan genomes for introgression in a secondary contact model. *PLoS One,* 10**,** e0118621.

Grau-Bove, X., Tomlinson, S., O'Reilly, A. O., Harding, N. J., Miles, A., Kwiatkowski, D., Donnelly, M. J., Weetman, D. & *Anopheles gambiae* Genomes, C. 2020. Evolution of the Insecticide Target Rdl in African Anopheles Is Driven by Interspecific and Interkaryotypic Introgression. *Mol Biol Evol,* 37**,** 2900-2917.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. & Paabo, S. 2010. A draft sequence of the Neandertal genome. *Science,* 328**,** 710-722.

Haakenstad, A., Harle, A. C., Tsakalos, G., Micah, A. E., Tao, T., Anjomshoa, M., Cohen, J., Fullman, N., Hay, S. I., Mestrovic, T., Mohammed, S., Mousavi, S. M., Nixon, M. R., Pigott, D., Tran, K., Murray, C. J. L. & Dieleman, J. L. 2019. Tracking spending on malaria by source in 106 countries, 2000-16: an economic modelling study. *Lancet Infect Dis,* 19**,** 703-716.

Hanemaaijer, M. J., Higgins, H., Eralp, I., Yamasaki, Y., Becker, N., Kirstein, O. D., Lanzaro, G. C. & Lee, Y. 2019. Introgression between Anopheles gambiae and Anopheles coluzzii in Burkina Faso and its associations with kdr resistance and Plasmodium infection. *Malar J,* 18**,** 127.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M., Wides, R., Salzberg, S. L., Loftus, B., Yandell, M., Majoros, W. H., Rusch, D. B., Lai, Z., Kraft, C. L., Abril, J. F., Anthouard, V., Arensburger, P., Atkinson, P. W., Baden, H., de Berardinis, V., Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chaturverdi, K., Christophides, G. K., Chrystal, M. A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C. A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R., Hillenmeyer, M. E., Hladun, S. L., Hogan, J. R., Hong, Y. S., Hoover, J., Jaillon, O., Ke, Z., Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J. J., Lobo, N. F., Lopez, J. R., Malek, J. A., McIntosh, T. C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S. D., O'Brochta, D. A., Pfannkoch, C., Qi, R., Regier, M. A., Remington, K., Shao, H., Sharakhova, M. V., Sitter, C. D., Shetty, J., Smith, T. J., Strong, R., Sun, J., Thomasova, D., Ton, L. Q., Topalis, P., Tu, Z., Unger, M. F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K. J., Wortman, J. R., Wu, M., Yao, A., Zdobnov, E. M., Zhang, H., Zhao, Q., *et al.* 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science,* 298**,** 129-49.

Ibrahim, S. S., Riveron, J. M., Bibby, J., Irving, H., Yunta, C., Paine, M. J. & Wondji, C. S. 2015. Allelic Variation of Cytochrome P450s Drives Resistance to Bednet Insecticides in a Major Malaria Vector. *PLoS Genet,* 11**,** e1005618.

Isaacs, A. T., Mawejje, H. D., Tomlinson, S., Rigden, D. J. & Donnelly, M. J. 2018. Genome-wide transcriptional analyses in Anopheles mosquitoes reveal an unexpected association between salivary gland gene expression and insecticide resistance. *BMC Genomics,* 19**,** 225.

Killeen, G. F. 2014. Characterizing, controlling and eliminating residual malaria transmission. *Malar J,* 13**,** 330.

Kulathinal, R. J., Stevison, L. S. & Noor, M. A. 2009. The genomics of speciation in Drosophila: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genet,* 5**,** e1000550.

Lehmann, T., Blackston, C. R., Besansky, N. J., Escalante, A. A., Collins, F. H. & Hawley, W. A. 2000. The Rift Valley complex as a barrier to gene flow for Anopheles gambiae in Kenya: the mtDNA perspective. *J Hered,* 91**,** 165-8.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

Love, R. R., Redmond, S. N., Pombi, M., Caputo, B., Petrarca, V., Della Torre, A., *Anopheles gambiae* Genomes, C. & Besansky, N. J. 2019. In Silico Karyotyping of Chromosomally Polymorphic Malaria Mosquitoes in the *Anopheles gambiae* Complex. *G3 (Bethesda),* 9**,** 3249-3262.

Mawejje, H. D., Wilding, C. S., Rippon, E. J., Hughes, A., Weetman, D. & Donnelly, M. J. 2013. Insecticide resistance monitoring of field-collected

Anopheles gambiae s.l. populations from Jinja, eastern Uganda, identifies high levels of pyrethroid resistance. *Med Vet Entomol,* 27**,** 276-83.

Mitchell, S. N., Rigden, D. J., Dowd, A. J., Lu, F., Wilding, C. S., Weetman, D., Dadzie, S., Jenkins, A. M., Regna, K., Boko, P., Djogbenou, L., Muskavitch, M. A., Ranson, H., Paine, M. J., Mayans, O. & Donnelly, M. J. 2014. Metabolic and target-site mechanisms combine to confer strong DDT resistance in *Anopheles gambiae*. *PLoS One,* 9**,** e92662.

Namountougou, M., Simard, F., Baldet, T., Diabate, A., Ouedraogo, J. B., Martin, T. & Dabire, R. K. 2012. Multiple insecticide resistance in *Anopheles gambiae* s.l. populations from Burkina Faso, West Africa. *PLoS One,* 7**,** e48412.

Nardini, L., Hunt, R. H., Dahan-Moss, Y. L., Christie, N., Christian, R. N., Coetzee, M. & Koekemoer, L. L. 2017. Malaria vectors in the Democratic Republic of the Congo: the mechanisms that confer insecticide resistance in *Anopheles gambiae* and Anopheles funestus. *Malar J,* 16**,** 448.

Neafsey, D. E., Christophides, G. K., Collins, F. H., Emrich, S. J., Fontaine, M. C., Gelbart, W., Hahn, M. W., Howell, P. I., Kafatos, F. C., Lawson, D., Muskavitch, M. A., Waterhouse, R. M., Williams, L. J. & Besansky, N. J. 2013. The evolution of the Anopheles 16 genomes project. *G3 (Bethesda),* 3**,** 1191-4.

Norris, L. C., Main, B. J., Lee, Y., Collier, T. C., Fofana, A., Cornel, A. J. & Lanzaro, G. C. 2015. Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *Proc Natl Acad Sci U S A,* 112**,** 815-20.

Nwane, P., Etang, J., Chouasmall yi, U. M., Toto, J. C., Koffi, A., Mimpfoundi, R. & Simard, F. 2013. Multiple insecticide resistance mechanisms in *Anopheles gambiae* s.l. populations from Cameroon, Central Africa. *Parasit Vectors,* 6**,** 41.

Park, S. & Brown, T. M. 2002. Linkage of genes for sodium channel and cytochrome P450 (CYP6B10) in Heliothis virescens. *Pest Manag Sci,* 58**,** 209-12.

Safi, N. H. Z., Ahmadi, A. A., Nahzat, S., Warusavithana, S., Safi, N., Valadan, R., Shemshadian, A., Sharifi, M., Enayati, A. & Hemingway, J. 2019. Status of insecticide resistance and its biochemical and molecular mechanisms in Anopheles stephensi (Diptera: Culicidae) from Afghanistan. *Malar J,* 18**,** 249.

Sharakhova, M. V., Hammond, M. P., Lobo, N. F., Krzywinski, J., Unger, M. F., Hillenmeyer, M. E., Bruggner, R. V., Birney, E. & Collins, F. H. 2007. Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biol,* 8**,** R5.

Silva Martins, W. F., Wilding, C. S., Isaacs, A. T., Rippon, E. J., Megy, K. & Donnelly, M. J. 2019. Transcriptomic analysis of insecticide resistance in the lymphatic filariasis vector Culex quinquefasciatus. *Sci Rep,* 9**,** 11406.

Staedke, S. G., Kamya, M. R., Dorsey, G., Maiteki-Sebuguzi, C., Gonahasa, S., Yeka, A., Lynd, A., Opigo, J., Hemingway, J. & Donnelly, M. J. 2019. LLIN Evaluation in Uganda Project (LLINEUP) - Impact of long-lasting insecticidal nets with, and without, piperonyl butoxide on malaria indicators in Uganda: study protocol for a cluster-randomised trial. *Trials,* 20**,** 321.

Temu, E. A., Hunt, R. H., Coetzee, M., Minjas, J. N. & Shiff, C. J. 1997. Detection of hybrids in natural populations of the *Anopheles gambiae* complex by the rDNA-based, PCR method. *Ann Trop Med Parasitol,* 91**,** 963-5.

Toe, K. H., N'Fale, S., Dabire, R. K., Ranson, H. & Jones, C. M. 2015. The recent escalation in strength of pyrethroid resistance in Anopheles coluzzi in West Africa is linked to increased expression of multiple gene families. *BMC Genomics,* 16**,** 146.

Toure, Y. T., Petrarca, V., Traore, S. F., Coulibaly, A., Maiga, H. M., Sankare, O., Sow, M., Di Deco, M. A. & Coluzzi, M. 1998. The distribution and inversion polymorphism of chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa. *Parassitologia,* 40**,** 477-511.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. & DePristo, M. A. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics,* 43**,** 11 10 1-11 10 33.

Vicente, J. L., Clarkson, C. S., Caputo, B., Gomes, B., Pombi, M., Sousa, C. A., Antao, T., Dinis, J., Botta, G., Mancini, E., Petrarca, V., Mead, D., Drury, E., Stalker, J., Miles, A., Kwiatkowski, D. P., Donnelly, M. J., Rodrigues, A., Torre, A. D., Weetman, D. & Pinto, J. 2017. Massive introgression drives species radiation at the range limit of Anopheles gambiae. *Sci Rep,* 7**,** 46451.

Weetman, D., Steen, K., Rippon, E. J., Mawejje, H. D., Donnelly, M. J. & Wilding, C. S. 2014. Contemporary gene flow between wild *An. gambiae s.s.* and *An. arabiensis*. *Parasit Vectors,* 7**,** 345.

Weetman, D., Wilding, C. S., Steen, K., Morgan, J. C., Simard, F. & Donnelly, M. J. 2010. Association mapping of insecticide resistance in wild *Anopheles gambiae* populations: major variants identified in a low-linkage disequilbrium genome. *PLoS One,* 5**,** e13140.

Weill, M., Chandre, F., Brengues, C., Manguin, S., Akogbeto, M., Pasteur, N., Guillet, P. & Raymond, M. 2000. The *kdr* mutation occurs in the Mopti form of *Anopheles gambiae* s.s. through introgression. *Insect Mol Biol,* 9**,** 451-5.

White, B. J., Cheng, C., Simard, F., Costantini, C. & Besansky, N. J. 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Mol Ecol,* 19**,** 925-39.

WHO 2018. Global report on insecticide resistance in malaria vectors: 2010–2016.

WHO 2019. World Malaria Report.

Yahouedo, G. A., Chandre, F., Rossignol, M., Ginibre, C., Balabanidou, V., Mendez, N. G. A., Pigeon, O., Vontas, J. & Cornelie, S. 2017. Contributions of cuticle permeability and enzyme detoxification to pyrethroid resistance in the major malaria vector *Anopheles gambiae*. *Sci Rep,* 7**,** 11091.

Zhang, L., Ma, D., Zhang, Y., He, W., Yang, J., Li, C. & Jiang, H. 2013. Characterization of DNA topoisomerase-1 in Spodoptera exigua for toxicity evaluation of camptothecin and hydoxy-camptothecin. *PLoS One,* 8**,** e56458.

Chapter VI.

Final Conclusions

The costs associated with whole genome sequencing, the storage of data and analysis of that data is decreasing year on year. The application of WGS-based technologies expands into many areas of industry, research, health and business and according to some commentators, is set to be the foundation of a turning point, much akin to the industrial revolution. Indeed, focussing on just health, understanding and characterising the unique genetic landscape of each individual has the ability to drive granular and bespoke health care interventions.

Mosquitoes are of central importance and focus to global health. Culex, Aedes and Anopheles mosquitoes are responsible for transmitting a plethora of diseases. Of these, malaria causes the most annual deaths, indeed, is the disease that causes the highest mortality in world. Much of the success in controlling malaria transmission relies of avoidance and control of the mosquito populations. Public health insecticides in LLINs and IRS are the primary methods for the control of mosquitoes. Resistance to these insecticides is decidedly present and threatening to reverse the decreasing trend in malaria cases. Although the genetic components of resistance have been studied and characterised for many years, in this thesis, we further the exploration of two areas in malaria research that are crucial and comparatively understudied – *An. arabiensis* and Introgression.

Chapter II

In this chapter, we carried out vital quality control analyses on the data, moving on to correct any errors in the metadata and verify the population and species labelling via principal component analyses. At the time of writing this thesis, phase 3 of the Ag1000G project was not published and the *An. arabiensis* data set had not been fully curated when we received them. Therefore, it was key that we ensured the data were in the correct state before analyses were performed. Though these samples have been fully sequenced, and that data is immutable, the alignment of the data was pushed through the Sanger VR Pipe analyses pipeline, so that the development of analyses scripts could begin. Therefore, visualised metrics of each samples that would quantify the quality of the alignment data which were received. Indeed, only 288 samples were excluded based on either poor performance or belonging to the faulty cross data. Analyses of the population structures of the samples allowed us to confirm the meta data and correct obvious errors in labelling. Finally, following extended analyses in attempting to deconvolute parent/progeny sample crosses, we excluded these samples entirely. In discussion with member of the Ag1000G, it was concluded that the cross samples were mixed up beyond saving and would require investigation, beyond the scope of this thesis.

Chapter III

In this chapter, we developed a method for identifying regions of the genome between two species that can be used as diagnostic markers for the two species. These ancestry informative markers were developed using *An. gambiae*

s.s. and *An. arabiensis* samples from The 16 Genomes Project Neafsey *et al.* (2015). Using these samples meant developing data impute scripts to handle and convert the data into a suitable format. These scripts have been used recently in a publication by Grau-Bové *et al.* (2020).

Once we identified markers that resolve species between *An. gambiae* and *An. arabiensis*, they were applied over the phase 3 dataset. The objective of this analyses was to observe any regions of discordancy. That is, in *An. arabiensis* samples, “do we observe any regions of the genome that are tagged with *An. gambiae* alleles?”, and vice versa. Such a signal was not clearly observed, and the method of using AIMs for this dataset was not fruitful for providing information about potential regions of introgression between the species.

Despite this, the marker panel is a useful tool for QC diagnostics and other analyses. Since the generation of the marker panel, it has been used in several analyses by the Donnelly-Weetman group at the Liverpool School of Tropical Medicine. This includes the methods of data handling, analyses and the final panel itself.

Chapter IV

Copy number variation mutations are an important mutation evolution and disease. Their role in insecticide resistance in *An. gambiae* has only recently been suggested by Lucas *et al.* (2020). Given the availability of the data, we felt it prudent to explore the *An. arabiensis* genome for CNVs linked with insecticide resistance for two key reasons. 1. Understanding the genetic landscape of the loci that convey resistance phenotypes is the backbone to effective interventions and

management strategies. 2. Given that we are also researching the extent of introgression in *An. arabiensis* and the potential link of such introgression to insecticide resistance, it provides a more complete story as whether a CNV have been introgressed or whether an introgressed loci has since been subject to CNV mutation.

We identified CNV which contained *Gstd3*, a gene belonging to the Glutathione-S-transferase family, which has been long implicated in insecticide resistance. Certainly, a review of the literature revealed that although Gstd3 itself is not solidly implicated in conferring a resistance phenotype, it has been implicated on microarray and qPCR data.

After discovery of the CNV containing the Gstd3 locus, we developed an assay to detect the CNV in samples with an unknown CNV state. We found that the CNV is present at frequency of 15.79% in a set of samples collected in Uganda, 2012. We also tested previously phenotyped samples, for which their susceptibility or resistance to permethrin was already established. Unfortunately, we were unable to identify an association between the presence of the *Gstd3* CNV and resistance to permethrin. Ultimately, this leaves the possibility that the CNV conveys resistance to another insecticide. To echo the conclusion of the manuscript in Chapter IV, further research is needed both in *An. gambiae* and *An. arabiensis* to concretely understand the roles of CNVs in resistance.

Chapter V

Introgression in the *An. gambiae* sensu lato has been long documented as occurring in the sibling species pair *An. gambiae* s.s. and *An. coluzzii*. Early efforts to understand introgression were instigated by the dubious and unresolved natured of the inversions and phylogeny of the species complex. Later, introgression was discovered to play a role in the transfer of advantageous alleles, in addition to general evolution.

An. arabiensis is becoming an increasingly more important vector as it displaces *An. gambiae* s.s. and the discussions surrounding residual transmission focusses on the species as the responsible agent for persistence transmission, despite universal LLIN coverage and intensive IRS campaigns. We therefore find it important to characterise the genetic features of *An. arabiensis* that are driving resistance.

In evaluating the *An. arabiensis* data set from phase 3 of the Ag1000G, we found multiple regions of the genomes showing significant signal of introgression, with some signal pervasive across all study populations (except one). These results clearly show that introgression *An. gambiae* ⇆ *An. arabiensis* has occurred and has occurred in regions previously associated with insecticides resistance.

Limitations were faced with these analyses, despite the discovery of introgression in insecticide resistance associated loci. However, we feel that the results here represent of vital first step and foundation for further characterising the role of introgression in Anophelines. Additionally, maturation and publication of the fully curated phase 3 dataset will allow the limitations of this study to be

addressed. Indeed, the scripts and code produced as part of this thesis will go towards significantly speeding up those subsequent analyses.

Those subsequent analyses should focus on addressing the haplotype-based questions that have been left unanswered. Primarily those are, the direction and relative ago the introgression events. Identifying that these signals are recent and transferred *An. gambiae* s.s. → *An. arabiensis* would confirm the conjecture that the interventions causing an increase in the population of *An. arabiensis* can cause both the establishment of a new local dominant vector species and a novel resistance locus in that species. It is the implication of this finding that the analyses provided here enable the finding of and may represent a significant development of the understanding of evolution to anthropogenic-mediated pressures.

Conclusion

The research conducted within this thesis is demonstrable of the wide scope that genomic and bioinformatic analyses can occupy in addressing questions pertinent to public health. The characterisation of *An. arabiensis* is becoming increasingly more important as is displaces *An. gambiae* s.s. population in sub-Saharan Africa. The role of introgression in the evolution and adaptation of mosquito species – including *An. arabiensis* – is not a new area of research. However, it does represent an area burgeoning with unanswered questions that relate directly to insecticide resistance. The increase in availability of whole-genome sequencing and the efforts of the Ag1000G allow the interrogation of genetic material to address these questions. Here, we demonstrate that there exist

CNVs that contain the putative detoxification gene *Gstd3* and regions of pervasive introgression that contain the cytochrome P450 gene cluster. The implication of these genetic features is an area of research that needs to be developed beyond this thesis.

Publication 1.

Open source 3D printable replacement parts for the WHO insecticide susceptibility bioassay system.

Parasites and Vectors 2019/11

PMID: 31727146

DOI: 10.1186/s13071-019-3789-9

Tomlinson *et al. Parasites Vectors* (2019) 12:539
https://doi.org/10.1186/s13071-019-3789-9

Parasites & Vectors

RESEARCH Open Access

Open source 3D printable replacement parts for the WHO insecticide susceptibility bioassay system

Sean Tomlinson[1*], Henrietta Carrington Yates[1], Ambrose Oruni[1,2], Harun Njoroge[1,3], David Weetman[1], Martin J. Donnelly[1] and Arjen E Van't Hof[1]

Abstract

Background: Malaria vector control and research rely heavily on monitoring mosquito populations for the development of resistance to public health insecticides. One standard method for determining resistance in adult mosquito populations is the World Health Organization test (WHO bioassay). The WHO bioassay kit consists of several acrylic pieces that are assembled into a unit. Parts of the kit commonly break, reducing the capacity of insectaries to carry out resistance profiling. Since there is at present only a single supplier for the test kits, replacement parts can be hard to procure in a timely fashion.

Methods: Using computer-aided design software and widely available polylactic acid (PLA) filament as a printing material, we 3D designed and printed replacement parts for the WHO bioassay system. We conducted a comparison experiment between original WHO bioassay kits and 3D printed kits to assess congruence between results. The comparison experiment was performed on two Kenyan laboratory strains of *Anopheles gambiae* (*s.s.*), Kilifi and Mbita. Student's t-tests were used to assess significant differences between tube types. Finally, we exposed the PLA filament to common solutions used with the bioassay kit.

Results: We were able to design and print functional replacements for each piece of the WHO bioassay kit. Replacement parts are functionally identical to and interchangeable with original WHO bioassay parts. We note no significant difference in mortality results obtained from PLA printed tubes and WHO acrylic tubes. Additionally, we observed no degradation of PLA in response to prolonged exposure times of commonly used cleaning solutions.

Conclusions: Our designs can be used to produce replacement parts for the WHO bioassay kit in any facility with a 3D printer, which are becoming increasingly widespread. 3D printing technologies can affordably and rapidly address equipment shortages and be used to develop bespoke equipment in laboratories.

Keywords: 3D printing, WHO bioassay, Mosquito profiling, Insecticide resistance

Background

Malaria remains a critical public health problem across sub-Saharan Africa, with vector control—a vital part of efforts to control and eradicate malaria-relying heavily on efficacious insecticides [1]. Widespread and emerging resistance poses a significant threat to public health and is reflected by increased efforts to understand and characterize the distribution of resistant mosquito populations and associated genetic variants across endemic regions of Africa [2, 3].

The World Health Organization insecticide susceptibility test (WHO bioassay) is a standard method implemented to assess resistance in adult mosquito populations. During this test, mosquitoes are held in one of two tubes (Fig. 1a), either lined with untreated paper (control) or insecticide-impregnated paper (exposure) held in place with spring clips (Fig. 1c). Both tubes are

*Correspondence: sean.tomlinson@lstmed.ac.uk
[1] Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool, UK
Full list of author information is available at the end of the article

separated by a slide unit (Fig. 1e) and slide (Fig. 1f), while the ends of the tubes are capped with a screen mesh (Fig. 1b) and a screw cap (Fig. 1a). Mosquitoes are held in the insecticide tube for one hour, and the percentage mortality of exposed mosquitoes 24 hours post-exposure is a measurement for insecticide susceptibility [4]. A single experimental unit for the WHO bioassay kit is comprised of two mesh screens, two screw caps, two tubes, four spring clips, one slide unit, one slide (Fig. 1).

Certain parts of the WHO bioassay kit are more liable to become worn, damaged or lost, causing a reduced capacity of insectaries to conduct bioassays. Most notably, in our experience, the mesh screen can become easily lost or damaged during cleaning. The slide unit is subject to friction from the slide and when combined with the gradual weakening of the chemical bond through repeated uses and washes, frequently splits. spring clips are often lost during washing procedures. Test kit distribution is coordinated by the Vector Control Research Unit, Universiti Sains Malaysia, Penang, Malaysia. A single kit costs US$78 at the time of publication. Long shipping times and associated costs mean that replacing lost or damaged parts can become economically or logistically unviable. To address these problems, we used computer-aided design software to produce 3D printable versions of the parts that comprise the WHO bioassay kit.

Accurate, reliable and affordable (US$200–1000) 3D printing technologies are now commercially available. The most common 3D printer form utilizes a Cartesian axis system to control the deposition of molten plastic filament onto a print surface, in a process called fused filament fabrication (FFF). Many different plastics and materials can be used for 3D printing, such as polylactic acid (PLA), acrylonitrile butadiene styrene (ABS), nylon, polyethylene terephthalate glycol (PET-G) and polycarbonate. PLA is an easy plastic to 3D print, is widely available and is suitable for use in most laboratory plastic equipment. Indeed, 3D printing technologies are increasingly being used in research settings [5]. The glass transition temperature of PLA is 60–65 °C with a melting temperature of ~180 °C, meaning in cold or low-temperature settings PLA is thermally stable.

Here, we present 3D printable replacement parts for the WHO bioassay kit which print without the need for tools or glue, and which interface with existing WHO bioassay parts. We discuss the design challenges, modifications from existing WHO bioassay kits and files needed to print replacement parts for the WHO bioassay kit.

Methods

Designing 3D models

We used SketchUp (Trimble Inc., Sunnyvale CA, USA, free for a personal license, US$55/year for Education Licence) and OpenSCAD (Marius Kintel, Openscad.org, Toronto, Canada, free) to create the 3D model files in the stereolithography (STL) format needed to enable 3D printing of parts. Some parts were technically difficult or impossible to directly replicate using current FFF 3D-printing. In these cases, we modified the existing design to allow printing, while retaining the same physical function.

Support material is plastic printed alongside the desired part to prevent necessary plastic overhangs from dropping below their intended position. This support material is printed in such a way that it is easily detached from the finished piece; however, its inclusion leads to longer print times and higher plastic consumption. Around the circumference of the tube, two rims are present to provide a positive stop for when the tubes are fully inserted into the slide unit. On the original WHO bioassay tube, these rims are squared on the edges, replicating this feature would require support material during printing. To reduce print time, plastic consumption and potential interference with tube threads, the outer geometry of the rim was changed to triangular. This geometry can be printed without any lower support while retaining the function of the original part.

The slide unit has an internal section into which the slide sits. This geometry is complex; indeed, the original part is manufactured in two halves and chemically bonded together. The concept for this project required that the entire system be 3D printable, to increase accessibility and use. To be practically printable, this part needed adapting for 3D printing. Like the WHO bioassay slide unit, we created two halves and developed a method of bonding the pieces together. We designed a sliding clip method of joining two screw flanges of the slide unit. Two halves of the slide unit are printed with the addition of arrow-like notches on each side; these interface with a sliding lock clip that mechanically locks the two halves together and creates a gap for the gate to slide through (Fig. 1g, h and i).

On the inside of the slide unit are two friction nodules (Fig. 1h) that retain the slide in either the closed or open position, preventing the slide from falling out of the slide unit during handling. To address this, we designed the whole slide unit to include front clips that retain the friction nodules. These changes now necessitate some assembly of the slide unit once printed. However, the slide unit has been designed to allow hand-assembly without the need for tools. Despite the changes to this part of the WHO bioassay kit and the increase in physical size, the mechanical function remains the same.

The mesh screen used at the end of the tubes is manufactured from a flexible material that allows it to have no border. In our prototyping, we found that printed mesh

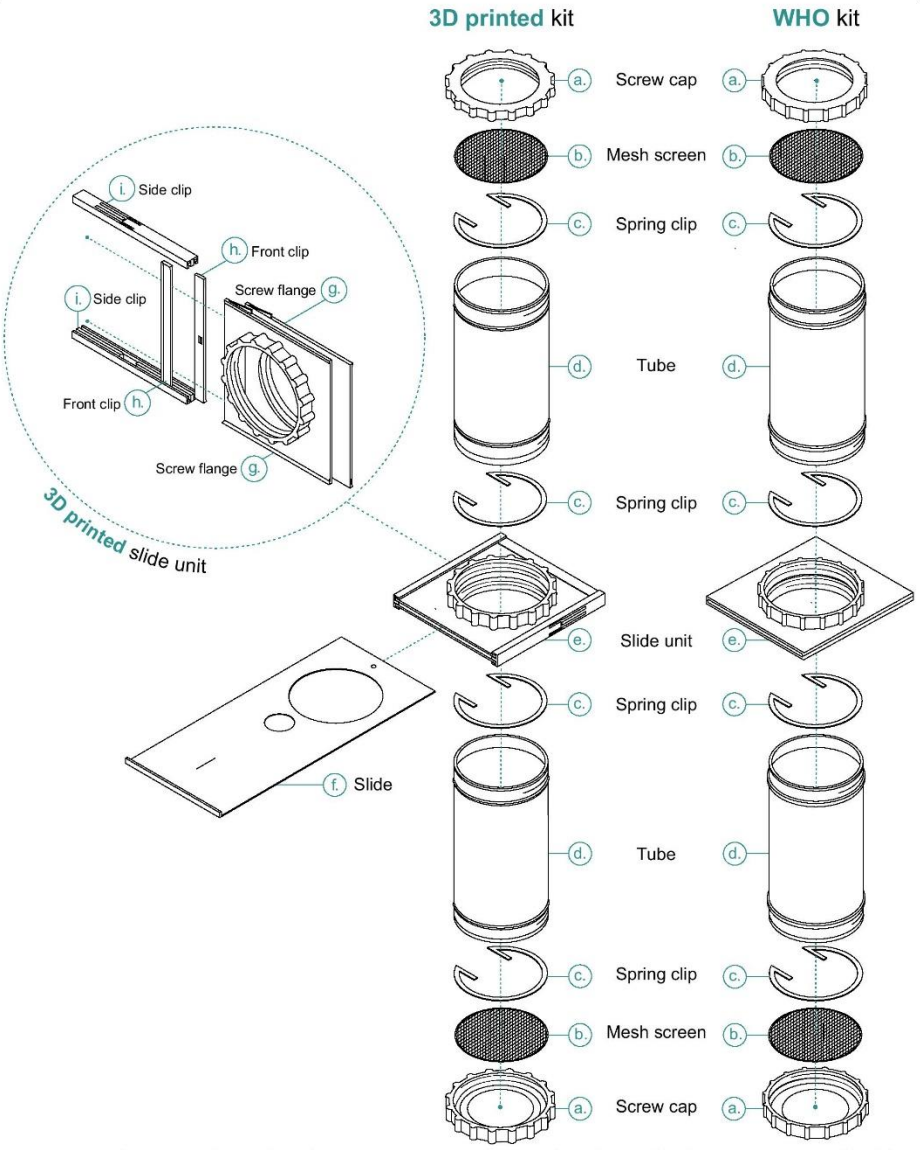


Fig. 1 World Health Organisation bioassay kit and equivalent 3D printed parts. Cutaway shows the assembly of the 3D printed version of the slide unit. For illustration purposes, the WHO bioassay slide has been omitted but is visually identical to f

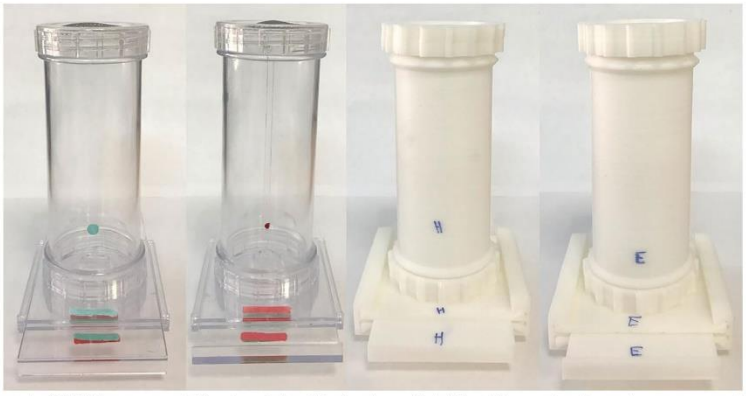


Fig. 2 Photograph of WHO bioassay parts (left) and equivalent 3D printed parts (right). E and H are used to denote the exposure and holding sides, respectively, whereas on the WHO bioassay parts red and green dots/lines are used

screens were too weak to be handled when printed without a border. Therefore, a 3 mm border was added to the CAD version of the mesh; this does not extend past the lip of the screw cap, retaining the same function as the original.

3D printing

3D printing was carried out on an Original Prusa i3 MK3 and an MK3S (Prusa Research, Prague, Czechia, £699) modified with a BuildTak print surface (https://www.buildtak.eu/), using white 1.75 mm PLA filament (ZIRO3D, Shenzhen, China, £14.99/kg). Designed CAD models were exported as STL files. STL files must be converted to machine instructions following the G-code standard to be processed by 3D printers. This conversion process is called slicing. The STL model files were sliced using Cura 3.3.1 (Ultimaker, Utrecht, The Netherlands) with the following key slicer settings: 100% infill, two shells/perimeters, 0.15 mm layer height.

Reliably and efficiently 3D printing transparent objects is technically difficult with commercially available 3D printers and typically results in a cloudy translucent finish. During prototyping, we identified that bright white filament, though not transparent, provides enough contrast for mosquitoes to be easily counted while viewing through the mesh screen. Commercially available WHO bioassay tubes use a green and red dot to denote both the holding and exposure side of the bioassay kit, respectively. We used a permanent marker to label the corresponding printed parts with an 'E' (exposure) and 'H' (holding) (Fig. 2).

Bioassay testing

To ensure that the 3D printed tubes performed in a similar manner to the acrylic tubes we performed susceptibility testing using standard 4% WHO diagnostic dose of dichlorodiphenyltrichloroethane (DDT) insecticide on two Kenyan laboratory strains of *Anopheles gambiae* (*s.s.*) (Kilifi and Mbita). DDT was chosen to match specimen availability. Kilifi and Mbita samples were freely available at the time of experiments. These strains are partially resistant to DDT, meaning we expect to see mortality less than 100%, allowing clear observation of the effect of PLA on mosquito mortality, if any individual tube had a mortality of 100%.

Batches of ~25 3–5-day-old female mosquitoes, were exposed in each tube. The number of replicate exposures was as follows: Kilifi WHO tubes ($n=10$), Kilifi 3D tubes ($n=11$), Mbita WHO tubes ($n=7$), Mbita 3D tubes ($n=7$). Exposures were carried out in tandem for both 3D printed and WHO bioassay kits. The total number of exposures were performed over separate days to allow adult females to reach the correct age for exposure. Each exposure tube was paired with a corresponding (3D printed or WHO) control tube. WHO guidelines require at least four replicates; specimen availability allowed us to exceed this minimum. Percentage mortality was recorded after a holding period of 24-hours. We used a Student's t-test to compare the mean between standard WHO tubes and 3D printed tubes. All graphic visualizations and statistical analyses were performed using R, data and R code for analyses and figure generation are provided (Additional file 1: Table S1; Additional file 2: Text S1).

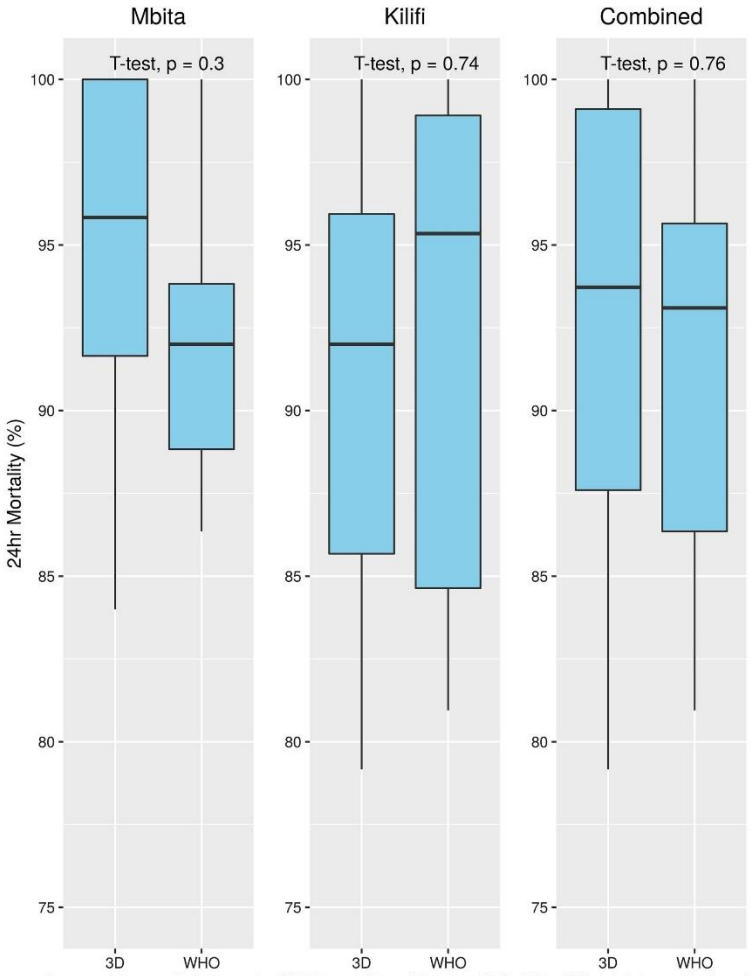


Fig. 3 Bioassay results comparing mortality between the WHO bioassay kit and 3D printed kit for Mbita, Kilifi and their combined results. A Student's t-test was used to compare means for each test. Kilifi WHO tubes ($n = 10$); Kilifi 3D tubes ($n = 11$), Mbita WHO tubes ($n = 7$), Mbita 3D tubes ($n = 7$). Each tube contained ~25 3–5-day-old female mosquitoes

An additional experiment to assess insecticide retention and absorption into PLA was performed. The Kisumu strain of mosquito are susceptible to DDT, therefore we exposed ~25 3–5-day-old female Kisumu mosquitoes to cleaned WHO and 3D printed tubes that previously held 4% DDT for a standard exposure. This experiment was performed in triplicate for both WHO bioassay tubes and 3D printed tubes. Mortality was recorded at 1-hour and mosquitoes were moved to a paper holding cup. The mosquitoes were fed on sugar and the 24-hour mortality was recorded.

Mosquito rearing was conducted at the Liverpool School of Tropical Medicine insectaries, following standard operating procedures. The Mbita strain was collected at Mbita Point, Kenya in 1999, and has been maintained

as a laboratory strain since this time. The Kilifi strain was collected in Kilifi County, Kenya in 2012. The colony is maintained by both the Liverpool School of Tropical Medicine and Kenya Medical Research Institute.

PLA reactivity with bioassay solutions

To assess whether the PLA would interact with solutions that are commonly used during the bioassay protocol, we exposed printed PLA parts to 4 different solutions to observe any degradation of the plastic. (i) Cotton pads soaked with 10% sucrose solution, typically used to feed mosquitoes during the recovery period, were placed on six mesh screens for seven days. Cotton pads were soaked daily with fresh 10% sucrose solution to replace evaporated solution. (ii) Four slides were submerged in 3% Rely+On Virkon (Lanxess, Cologne, Germany) for five days. (iii) Four slides were submerged in 5% Decon 90 (Decon Laboratories Ltd., Hove, England) for five days. (iv) Six screw caps were submerged in 70% ethanol for five days.

Results

3D printing

Printed parts interface as expected with current WHO bioassay parts, allowing any configuration of 3D printed and WHO parts to be assembled together. The printed kits assembled easily without the need for additional tools. CAD and STL files produced are available at https://github.com/SeanTomlinson30/3D-Printable-WHO-Bioassay-Parts.

Bioassay testing

Bioassays with 4% DDT using the Mbita and Kilifi strains showed no significant difference mortality after 24 hours, for measurements between 3D printed and WHO bioassay kits (Fig. 3). We observed that mosquitoes can sugar feed through the 3D printed mesh screens. We did not observe any mortality in control tubes for either 3D printed of WHO bioassay tubes. We did not observe any visual signs of residue or insecticide retention on the ridged surface of the PLA after washing the parts. All Kisumu mosquitoes exposed to cleaned WHO bioassay and 3D printed tubes survived the 1-hour exospore and all survived to 24-hours post-exposure.

PLA reactivity with bioassay solutions

After exposure to 10% sucrose, 70% ethanol, 3% Rely+On Virkon (Lanxess) and 5% Decon 90 (Decon Laboratories Ltd.), we observed no signs of degradation of the PLA strength, tensibility, surface color or size.

Discussion

We have developed, and provide here, printable versions of all pieces that compose a WHO bioassay kit. We see the primary use case for these parts as a replacement library for missing and damaged parts of an original WHO bioassay kit. Bioassay data for DDT exposure indicate no significant difference between 3D printed and WHO bioassay kits; although, other insecticides/strain combinations may react differently when interacting with 3D printed materials. We acknowledge the need to further test and validate 3D printed alternatives to the WHO bioassay kit, extending the tested insecticides to include pyrethroids, organophosphates and carbamates, the four main classes of public health insecticide.

Anecdotally, in our insectaries, we find that the most in-demand 3D printed replacement parts are the slide unit and mesh screen, with tubes being the most durable parts and least likely to be needed. The design challenges of 3D printing the WHO bioassay kit necessitated some changes to the geometry of individual parts. Most notably, to retain all functionality, the 3D printable slide unit had to be printed as six individual pieces that are assembled. In addition to showing no functional differences during operation and manual handling, because the 3D printed slide unit does not use chemical bonding, it is more durable to general wear and less likely to become damaged, in terms of splitting. Though, we do note that when using PLA as a 3D printing material, operators must be cognizant of the effect of hot temperatures causing material deformation.

Conclusions

We present files that allow printing of all parts of the WHO bioassay kit. To achieve this, we replicated existing parts in CAD software, modifying and adapting the designs where necessary to permit 3D printing. The printed parts work with standard WHO bioassay kits and in the case of full 3D printed kits, produce results not significantly different from standard WHO bioassay kits. 3D printing in laboratory environments has become more achievable thanks to the continued reduction in costs and developments in 3D printing technologies. Through the distribution of the 3D printable laboratory equipment, researchers can maintain testing capacity, reduce costs and adapt apparatus for bespoke purposes.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13071-019-3789-9.

Additional file 1: Table S1. Bioassay mortality data for WHO vs 3D printed kits with both Mbita and Kilifi stains.

Additional file 2: Text S1. Figure and statistics generation script.

Abbreviations

ABS: acrylonitrile butadiene styrene; CAD: computer-aided design; DDT: dichlorodiphenyltrichloroethane; FFF: fused filament fabrication; PET-G:

polyethylene terephthalate glycol; PLA: polylactic acid; STL: stereolithography; WHO: World Health Organisation.

Acknowledgments
We are grateful to Manuela Bernardi, who illustrated Fig. 1. We also would like to thank Giorgio Praulins for providing example WHO bioassay parts for the CAD modelling process.

Authors' contributions
ST, AVH and MJD designed the study. ST and AVH designed the CAD models and 3D printed the parts. DW, HN and AO helped design the assays. AO carried out the assays. HCY performed initial prototype testing. ST and HCY performed the absorption assay ST wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the Medical Research Council United Kingdom (MR/P02520X/1), and the National Institute of Allergy and Infectious Diseases ([NIAID] R01-AI116811). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAID or National Institutes of Health (NIH). ST is supported by an MRC doctoral training programme studentship (1855159).

Availability of data and materials
The CAD and STL files produced are in the supplementary materials and are available at https://github.com/SeanTomlinson30/3D-Printable-WHO-Bioassay-Parts. Bioassay data and analyses scripts are provided in Additional files 1 and 2.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool, UK. [2] College of Veterinary Medicine, Animal Resources & Biosecurity, Makerere University, Kampala, Uganda. [3] Department of Biochemistry, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya.

Received: 12 September 2019 Accepted: 4 November 2019
Published online: 14 November 2019

References

1. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. Nature. 2015;526:207–11.
2. The Anopheles gambiae 1000 Genomes Consortium. Genetic diversity of the African malaria vector *Anopheles gambiae*. Nature. 2017;552:96–100.
3. Lucas ER, Miles A, Harding NJ, Clarkson CS, Lawniczak MKN, Kwiatkowski DP, et al. Whole-genome sequencing reveals high complexity of copy number variation at insecticide resistance loci in malaria mosquitoes. Genome Res. 2019;29:1250–61.
4. WHO. Test procedures for insecticide resistance monitoring in malaria vector mosquitoes. Geneva: World Health Organisation; 2016.
5. Witmer K, Sherrard-Smith E, Straschil U, Tunnicliff M, Baum J, Delves M. An inexpensive open source 3D-printed membrane feeder for human malaria transmission studies. Malar J. 2018;17:282.

Publisher's Note

Publication 2.

Malaria Data by District: An open-source web application for increasing access to malaria information [version 2; peer review: 2 approved]

Wellcome Open Research 2019/10

SOFTWARE TOOL ARTICLE

REVISED Malaria Data by District: An open-source web application for increasing access to malaria information [version 2; peer review: 2 approved]

Sean Tomlinson [1,2*], Andy South [1], Joshua Longbottom [1,2*]

[1]Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool, L3 5QA, UK
[2]Centre for Health Informatics, Computing and Statistics, Lancaster University, Lancaster, LA1 4YW, UK

[*] Equal contributors

v2 **First published:** 09 Oct 2019, **4**:151
https://doi.org/10.12688/wellcomeopenres.15495.1
Latest published: 02 Dec 2019, **4**:151
https://doi.org/10.12688/wellcomeopenres.15495.2

Abstract

Preventable diseases still cause huge mortality in low- and middle-income countries. Research in spatial epidemiology and earth observation is helping academics to understand and prioritise how mortality could be reduced and generates spatial data that are used at a global and national level, to inform disease control policy. These data could also inform operational decision making at a more local level, for example to help officials target efforts at a local/regional level. To be usable for local decision-making, data needs to be presented in a way that is relevant to and understandable by local decision makers. We demonstrate an approach and prototype web application to make spatial outputs from disease modelling more useful for local decision making. Key to our approach is: (1) we focus on a handful of important data layers to maintain simplicity; (2) data are summarised at scales relevant to decision making (administrative units); (3) the application has the ability to rank and compare administrative units; (4) open-source code that can be modified and re-used by others, to target specific user-needs. Our prototype application allows visualisation of a handful of key layers from the Malaria Atlas Project. Data can be summarised by administrative unit for any malaria endemic African country, ranked and compared; e.g. to answer questions such as, 'does the district with the highest malaria prevalence also have the lowest coverage of insecticide treated nets?'. The application is developed in R and the code is open-source. It would be relatively easy for others to change the source code to incorporate different data layers, administrative boundaries or other data visualisations. We suggest such open-source web application development can facilitate the use of data for public health decision making in low resource settings.

Keywords

Malaria, Open-access, Shiny, Application, R, Data accessibility

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 02 Dec 2019	✓ report	✓
version 1 09 Oct 2019	? report	? report

1 **Peter Macharia**, Kenya Medical Research Institute - Wellcome Trust Research Programme, Nairobi, Kenya

2 **Caroline A. Lynch**, London School of Hygiene and Tropical Medicine, London, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Sean Tomlinson (sean.tomlinson@lstmed.ac.uk), Andy South (andy.south@lstmed.ac.uk)

Author roles: Tomlinson S: Conceptualization, Methodology, Project Administration, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **South A**: Conceptualization, Methodology, Project Administration, Software, Writing – Review & Editing; **Longbottom J**: Conceptualization, Methodology, Project Administration, Software, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was a runner-up in the 2019 Wellcome Data Re-use Prize in Malaria and is eligible for publication on Wellcome Open Research. The work itself is not associated with a funded project. ST acknowledges additional funding from the Medical Research Council (Award no. 1855159). JL acknowledges additional funding from the Medical Research Council (Award no. 1964851).

How to cite this article: Tomlinson S, South A and Longbottom J. **Malaria Data by District: An open-source web application for increasing access to malaria information [version 2; peer review: 2 approved]** Wellcome Open Research 2019, **4**:151 https://doi.org/10.12688/wellcomeopenres.15495.2

First published: 09 Oct 2019, **4**:151 https://doi.org/10.12688/wellcomeopenres.15495.1

REVISED **Amendments from Version 1**

We thank the reviewers for their useful comments. We have added two paragraphs to the discussion (paragraphs 4 and 5), to expand on the limitations of our prototype as raised by the reviewers, and to discuss how we see that open-source approaches can address these issues in the future. Briefly, the main points addressed are the temporal resolution of the data utilised and its suitability for guiding contemporary policies; the spatial resolution to which we aggregate pixel-level estimates, and the possibility of expanding upon available administrative boundary areal data. Our expanded discussion highlights that the submitted work is a prototype and that the authors have received funding to expand upon the functionality of this app through the development of reusable R software building blocks. We have responded to each of the reviewer's points individually in our full response, which can be viewed alongside the full text for this manuscript.

Any further responses from the reviewers can be found at the end of the article

Introduction

In recent years, the mapping of diseases has improved considerably in extent, resolution and accuracy (Kraemer *et al.*, 2016). Increasingly, data and related spatial outputs are being made publicly available (Briand *et al.*, 2018; Flueckiger *et al.*, 2015). However, the full potential of associated modelled outputs will only be realised if data are accessed and used to inform local decision making. Recent reviews have suggested that data repositories are mainly targeted toward researchers rather than decision makers and that there is a need to improve indicator data use in low- and middle-income countries (Briand *et al.*, 2018; Omumbo *et al.*, 2013). We describe the development of an open-source web application, MaDD (Malaria Data by District) (Tomlinson *et al.*, 2019), that enables disease distribution data to be more accessible at a local level.

The Malaria Atlas Project (MAP) is an international consortium which provides geographical information on diverse aspects of malaria epidemiology (Hay & Snow, 2006). The open-access data generated by MAP have the potential to influence policy at the national and subnational level (Hay & Snow, 2006; Moyes *et al.*, 2013). The project includes sophisticated interpolation models that allow inference of malaria prevalence, as detailed in national and regional indicator surveys, at non-sampled locations (Giorgi *et al.*, 2018; Hay & Snow, 2006). Getting contemporary estimates of malaria metrics to policy makers is essential, but barriers to acceptance exist, notably for modelled predictions; these include the complexity of the statistics described within output reports, and the description of assumptions made during the modelling process (Whitty, 2015). Additional barriers include the sheer wealth of data available, making it difficult to find and choose data surfaces despite central repositories that may be easily navigable. These factors have contributed towards a general lack of modelled outputs being used by local-level implementation programmes in Africa (Omumbo *et al.*, 2013).

Most modelled MAP data are provided as spatial estimates, presented as 5 × 5 km gridded surfaces, for example, estimates of *Plasmodium falciparum* prevalence and mortality, estimates of indoor residual spraying coverage and estimates of dominant vector species distributions and abundance (Bhatt *et al.*, 2015; Gething *et al.*, 2016; Sinka *et al.*, 2016). Though data generated at this spatial resolution provides a visual indication of subnational disparities, it is not immediately clear how these data may be used directly in operational decision-making. For modelled data to be utilised by operational staff at a local level, there is a requirement for additional tools and the ability to convert such data into operationally useful metrics at the level of administrative units (Knight *et al.*, 2016; Omumbo *et al.*, 2013; Whitty, 2015).

Data curated by MAP can already be accessed via online interactive maps (Malaria Atlas Project, 2019), an online country profiles tool and the *malariaAtlas* R package (Pfeffer *et al.*, 2018). These are powerful tools enabling access to MAP generated data that do include data summaries by administrative units. However, because of the wealth of data and functionality it is not straightforward to find and use these tools to perform district-level comparisons. Here, we present an application that allows rapid generation and comparison of summary statistics for a select suite of malaria indicator variables at the sub-national administrative level. MaDD is open-source and coded in R, so it can easily be modified to address local needs (R Core Team, 2019). This is a step towards developing tools for local decision makers to inform questions such as, "where should we prioritise the targeting of IRS rounds this season?".

Methods

Development background

Malaria and malaria-associated data from MAP are curated from a wide variety of sources, including national control programmes, national survey data, satellite imagery, published and grey literature (Moyes *et al.*, 2013). Presently, MAP provide 93 data layers relating to malaria and associated metrics (Pfeffer *et al.*, 2018). The data are mostly stored as gridded surfaces at a resolution of 1 or 5 km. To ensure that MaDD was easy to use and that users are not overwhelmed with the diverse range of MAP data, we refined the list of input data surfaces down to four impactful malariometric variables, these are:

1. Malaria incidence: all-age incidence rate (clinical cases per 1,000 population per annum) of *Plasmodium falciparum* malaria (Bhatt *et al.*, 2015).

2. Malaria prevalence in children: age-standardised parasite rate for *Plasmodium falciparum* malaria for children two to ten years of age (PfPR2–10) (Bhatt *et al.*, 2015).

3. Insecticide treated net coverage: proportion of the population who were protected by ITNs (Bhatt *et al.*, 2015).

4. Travel time to nearest city: minutes (Weiss *et al.*, 2018).

The data surfaces were chosen to illustrate the distribution of malaria incidence in the context of bed net distributions and proximity to cities. Other data layers could be added relatively easily by small modifications to the open source code. Surfaces were obtained from MAP utilising the 'malariaAtlas' R package

(version 0.0.3) (Pfeffer *et al.*, 2018). These surfaces were then aggregated using first-level administrative boundaries for each country in sub-Saharan Africa as provided by the Food and Agriculture Organization of the United Nations, and the 'raster' package (version 2.8-19) (FAO, 2008; Hijmans, 2019). Boundaries for display were simplified to make the user interface quick and responsive to user choices (i.e. reduction of polygon features and aggregation of spatial resolution), using the 'rmapshaper' package (version 0.4.1) (Teucher & Russell, 2018). Due to the computational time required to run zonal aggregation, this process was pre-computed, and the output data are stored with the application source code. We provide this preprocessing code within our open-source repository, so that other data layers from MAP or elsewhere can also be presented within the same framework.

Implementation

We used R (version 3.5.0) and the Shiny (version 1.2.0) framework to develop an online user-interface which allows users to interact with African MAP data, providing summary statistics by subnational administrative units (Chang *et al.*, 2018; R Core Team, 2019). Shiny allows interactive web applications to be created with R code. This enables the creation and modification of web applications by researchers without specific knowledge of application scripting languages and workflows, something that traditionally required a dedicated web developer. During development we made several choices regarding the user interface (UI) to ensure that the tool is user-friendly. The key goal that drove UI development was ease of use. The aim was to minimise the total number of inputs that the user was required to consider and maximise the space allocated for visualising the spatial surfaces and output report statistics. The UI for MaDD went through several iterations during development to reflect growing functionality and target audience feedback. Anticipated target audiences for MaDD are detailed in Table 1.

Operation

MaDD (Tomlinson *et al.*, 2019) can be publicly accessed at https://seantomlinson30.shinyapps.io/shiny-map-prize/, with any modern web browser (e.g. Firefox, Chrome, Safari or Edge). Though we will make efforts to ensure MaDD remains available on a public platform, this relies upon continued free server hosting being available. Interfacing with MaDD on a hosted platform allows computation and memory requirements to be handled by the hosting server, reducing the technical requirements for users. MaDD can also be run from the source code locally using R and RStudio. All source code for MaDD can be found under version control at https://github.com/SeanTomlinson30/SHINY-map-prize. The minimum system requirements of R and RStudio may change and can be found at https://www.r-project.org/ and https://www.rstudio.com/, respectively.

Use case

Users can first select a country of interest. This selection is made from a list of malaria-endemic sub-Saharan African countries with modelled metrics, as provided in the scrollable "Country" field in the left-hand side panel of the MaDD UI (Figure 1, red box). This country selection triggers an update of the map visible within the "Map" tab on the right-hand side of the application (Figure 1, blue box). The "Map" tab is visible by default and dynamically updates to reflect the data surfaces the user has selected for comparison, as indicated by the "Data to show and compare" selection box within the input panel. Within this map, the user can hover the cursor over polygons (boundaries representing each administrative unit) to determine the name of the visualised administrative unit.

In the input section on the left, users can filter administrative divisions to compare; by default, all first-level administrative divisions for the country are selected. The mean values for all surfaces by administrative unit can be compared and ranked within the "Table" tab. This presents an interactive table (Figure 2) that includes the priority (rank) for each variable and the ability to reorder the rows based on any column. Priority is set such that 1 indicates the "greatest need" i.e. highest malaria prevalence, lowest proportional coverage of bed nets and most remote district. This allows the user, for example, to order the rows based on malaria prevalence and then quickly see whether the administrative units with the greatest need also have other underperforming metrics; for example, a low proportional coverage of bed nets. There are likely to be other complicating factors influencing the interaction and causality across metrics, but this table allows users to determine where broad patterns are not as expected.

The "Generate Report" button present at the bottom of the left-hand side panel is used to generate summary statistics and country-specific choropleth maps, which are then shown in the "Output Report" tab (Figure 3). These maps are also provided as a rendered RMarkdown document (saved locally as a temporary file on the user's machine). Once the report is generated, a "Download Report" button becomes available, on the left-hand side panel, which enables the user to save the report as a locally viewable HTML file.

Discussion

We present, MaDD (Tomlinson *et al.*, 2019) a prototype user interface demonstrating how MAP and other data can be

Table 1. Anticipated main interests for target audiences.

Target audience	Anticipated main interests
High-level policymakers	Prioritisation of the most burdensome administrative areas in which to roll out an intervention/control programme.
Donors	Maximising existing data accessibility to meet malaria eradication targets.
Technical health agencies (e.g. World Health Organization)	Monitoring of administrative-level trends across a suite of indicators, as a means of identifying operation efficacy.

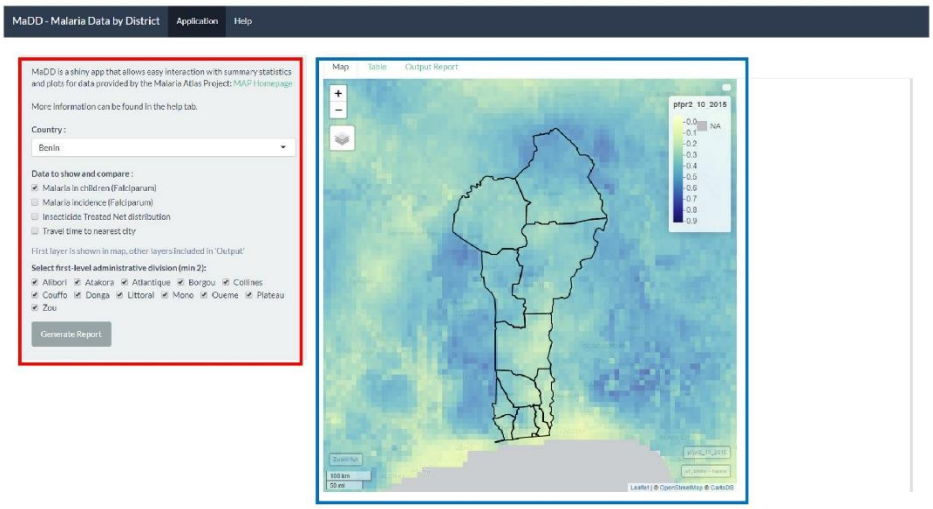


Figure 1. Screenshot of the Malaria Data by District application homepage. The main application page is split into a user input field (red box), which allows user to select the country, variable and districts of choice. The selected country is interactively plotted on the right side of the main page (blue box) and visualises the variables selected by the user.

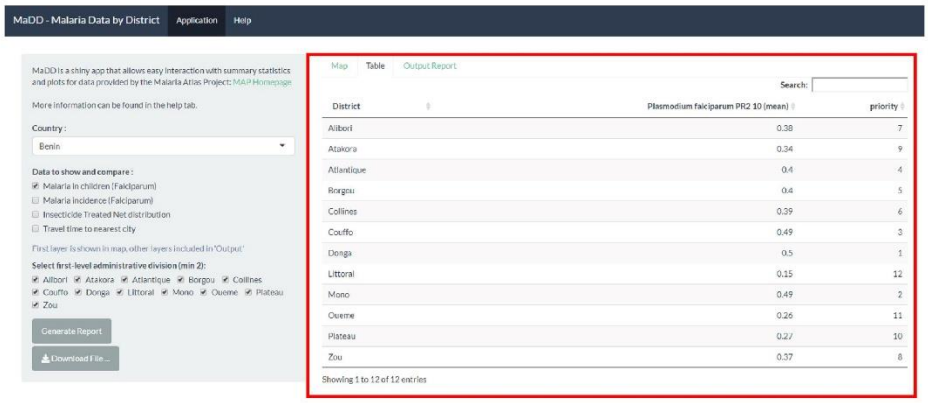


Figure 2. Screenshot of the Malaria Data by District application interactive table. Upon generation of the output file, an interactive table is populated (red box), which allows users to rank order data by column and search for variable names and statistics.

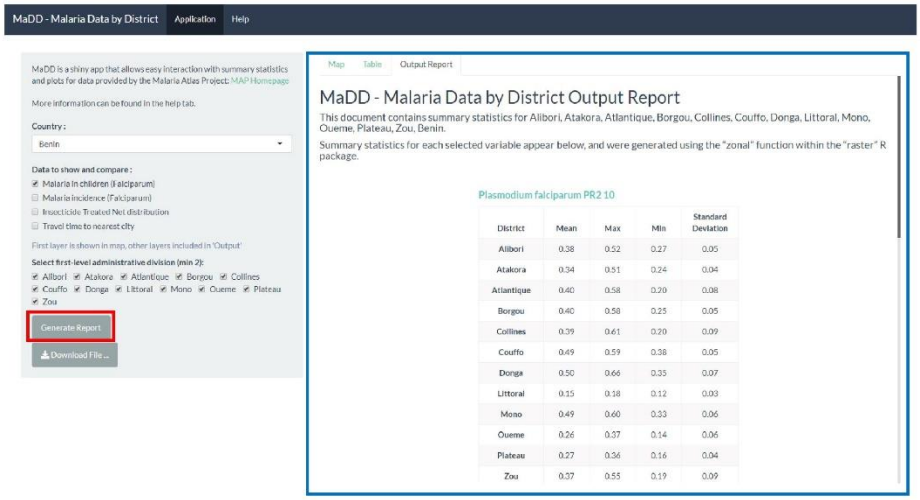


Figure 3. Screenshot of the Malaria Data by District application output report. Upon generation of the report (blue box), a download button becomes available (red box) allowing users to save the output file locally as an HTML file.

made more accessible to local decision makers. MaDD permits easy access to MAP data at the level of administrative units, through administrative division aggregation, interactive tables and plots. MaDD has been developed with open-source software and can be easily edited to include additional data surfaces, calculations, visualisations, or expanded to focus on specific geographic regions.

One of the biggest challenges in developing successful digital tools is in getting them adopted and used (Knight *et al.*, 2016). This has been recognised for digital projects in the development sector, prompting a set of Principles for Digital Development. These include recommendations that projects be designed with the user, understand the existing ecosystem, be collaborative and use open standards, open data, and open-source. Using open-source code (and particularly R given its wide adoption within a broad user community beyond computer scientists) has the advantage that tools can be adapted to fit local needs.

MaDD was developed to show summary statistics and plots for four impactful variables relating to malaria control; however, the framework we have developed is receptive to any of the surfaces generated by MAP. Though MaDD was developed for the African continent, future work could expand to include additional countries of interest, such as those with endemic malaria transmission in Asia or South America. Presenting publicly available data in an easily interactable and navigable way has the potential to increase public engagement and awareness of malaria trends to those concerned.

There are limitations to our application and approach. We aimed to demonstrate, with a prototype application, that open-source R software can be used to create useful and usable tools, improving access to data for malaria or other public health issues. However, we advise caution as this approach, and the application particularly, are not without risks. Firstly, focusing on the prototype application itself. We suggest that potential users critically assess whether the application and the data behind it, are appropriate to inform the questions they wish to address. For example, we would be comfortable with the application being used to give background on the malaria situation between administrative regions, but we would not be comfortable with operational decisions being made purely upon the rankings in one of the tables. One issue is the timeliness of the data layers. The process of data collection, modeling, layer creation and provision often takes years, such that publically available data may not be considered contemporary. Nevertheless, those data may be the best currently available and may still be useful to inform decision making in combination with other information sources.

On the issue of using other information sources, the prototype application deliberately had restricted options to promote usability. We recognise that the four data layers and single set of administrative boundaries we chose will not satisfy all potential users. There is a tension between usability and flexibility, i.e. making applications more flexible can make them more difficult to use and increases hosting requirements. Our aim was to show that highly usable applications can be created with open-source code. To move beyond the prototype, dialogue with

users is required to establish what data layers and functionality are required and the open-source code can be modified accordingly. Such modification would not require a large software company or us but could be achieved by local R developers or data scientists. Since this work, we have won funding for a small project (afrimapr) to develop reusable R software building blocks, to make it easier to build applications like this with different data sources and functionalities. We will be working to promote the local development of open-source digital tools using R. These could, for example, use administrative boundaries available from national sources, which may be more recent or go to finer administrative levels. This will prompt other questions, such as the spatial resolution at which different global data layers can provide useful information.

We hoped that providing this demonstration and open-source code would allow others to create similar applications using data to inform local decision making for disease control. We were pleased to be recently contacted by the Data Integration team at USAID President's Malaria Initiative, who indicated they are integrating our application into their own malaria information platform (Okoko, personal communications). They are developing analytical tools to inform the geographic allocation of malaria resources by in-country staff and national malaria control programmes in 27 countries and have plans to adapt our code to their own needs. We look forward to this and other open tools being developed to improve the use of data for local decision making in public health.

Data availability

The malariometric datasets analysed during this study are available from the Malaria Atlas Project database at: http://www.map.ox.ac.uk/explorer/, and the *malariaAtlas* R package (Pfeffer *et al.*, 2018).

Software availability

MaDD application available at: https://seantomlinson30.shinyapps.io/shiny-map-prize/.

Source code available from: https://github.com/SeanTomlinson30/SHINY-map-prize.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.3466884 (Tomlinson *et al.*, 2019).

License: MIT License.

Author contributions

All authors conceived the project, created and maintain the application, and were responsible for data curation, investigation and methodology development. JL and ST wrote the first draft of the manuscript, and all authors reviewed and approved the final manuscript.

Acknowledgements

The authors would like to thank Professor Martin J Donnelly who provided feedback on functionality and UI during development. The authors are very grateful to Mr James D G Miles for testing MaDD near the end of development for bugs and unintended behaviour. This work was produced for the Wellcome Data Re-use Prize, for which winning applications would receive a monetary prize. The authors formed a team that won a runner's up prize of £5000. We thank Lungi Okoko for permission to reference our personal communications in this article.

References

Bhatt S, Weiss DJ, Cameron E, *et al.*: **The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015.** *Nature.* 2015; **526**(7572): 207–211.
PubMed Abstract | Publisher Full Text | Free Full Text

Briand D, Roux E, Desconnets JC, *et al.*: **From global action against malaria to local issues: state of the art and perspectives of web platforms dealing with malaria information.** *Malar J.* 2018; **17**(1): 122.
PubMed Abstract | Publisher Full Text | Free Full Text

Chang W, Cheng J, Allaire J, *et al.*: **shiny: Web Application Framework for R.** R package version 1.2.0. [Online]. 2018.
Reference Source

Flueckiger RM, Nikolay B, Gelderblom HC, *et al.*: **Integrating data and resources on neglected tropical diseases for better planning: the NTD mapping tool (NTDmap.org).** *PLoS Negl Trop Dis.* 2015; **9**(2): e0003400.
PubMed Abstract | Publisher Full Text | Free Full Text

Food and Agriculture Organization of the United Nations: **The Global Administrative Unit Layers (GAUL): Technical Aspects**. Rome: Food and Agriculture Organization of the United Nations, EC-FAO Food Security Programme (ESTG), 2008.
Reference Source

Gething PW, Casey DC, Weiss DJ, *et al.*: **Mapping *Plasmodium falciparum* Mortality in Africa between 1990 and 2015.** *N Engl J Med.* 2016; **375**(25): 2435–2445.
PubMed Abstract | Publisher Full Text | Free Full Text

Giorgi E, Diggle PJ, Snow RW, *et al.*: **Geostatistical Methods for Disease Mapping and Visualisation Using Data from Spatio-temporally Referenced Prevalence Surveys.** *Int Stat Rev.* 2018; **86**(3): 571–597.
Publisher Full Text

Hay SI, Snow RW: **The malaria Atlas Project: developing global maps of malaria risk.** *PLoS Med.* 2006; **3**(12): e473.
PubMed Abstract | Publisher Full Text | Free Full Text

Hijmans RJ: **raster: Geographic Data Analysis and Modeling**. R package version 2.8-19. [Online]. 2019.
Reference Source

Knight GM, Dharan NJ, Fox GJ, *et al.*: **Bridging the gap between evidence and policy for infectious diseases: How models can aid public health decision-making.** *Int J Infect Dis.* 2016; **42**: 17–23.
PubMed Abstract | Publisher Full Text | Free Full Text

Kraemer MUG, Hay SI, Pigott DM, *et al.*: **Progress and Challenges in Infectious Disease Cartography.** *Trends Parasitol.* 2016; **32**(1): 19–29.
PubMed Abstract | Publisher Full Text

Malaria Atlas Project: **Interactive map and data explorer [Online]**. 2019.
Reference Source

Moyes CL, Temperley WH, Henry AJ, *et al.*: **Providing open access data online to advance malaria research and control.** *Malar J.* 2013; **12**: 161.
PubMed Abstract | Publisher Full Text | Free Full Text

Omumbo JA, Noor AM, Fall IS, *et al.*: **How well are malaria maps used to design**

and finance malaria control in Africa? *PLoS One.* 2013; **8**(1): e53198.
PubMed Abstract | Publisher Full Text | Free Full Text

Pfeffer DA, Lucas TCD, May D, *et al.*: ***malariaAtlas*: an R interface to global malariometric data hosted by the Malaria Atlas Project.** *Malar J.* 2018; **17**(1): 352.
PubMed Abstract | Publisher Full Text | Free Full Text

R Core Team: **R: A language and environment for statistical computing**. *R Foundation for Statistical Computing*, Vienna, Austria. 2019.
Reference Source

Sinka ME, Golding N, Massey NC, *et al.*: **Modelling the relative abundance of the primary African vectors of malaria before and after the implementation of indoor, insecticide-based vector control.** *Malar J.* 2016; **15**: 142.
PubMed Abstract | Publisher Full Text | Free Full Text

Teucher A, Russell K: **rmapshaper: Client for 'mapshaper' for 'Geospatial' Operations**. R package version 0.4.1. 2018.
Reference Source

Tomlinson S, Longbottom J, South A: **SeanTomlinson30/SHINY-map-prize: MaDD (Version v1).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3466884

Weiss DJ, Nelson A, Gibson HS, *et al.*: **A global map of travel time to cities to assess inequalities in accessibility in 2015.** *Nature.* 2018; **553**(7688): 333–336.
PubMed Abstract | Publisher Full Text

Whitty CJ: **What makes an academic paper useful for health policy?** *BMC Med.* 2015; **13**: 301.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 16 December 2019

https://doi.org/10.21956/wellcomeopenres.17109.r37248

✔ **Caroline A. Lynch**
Faculty of Epidemiology & Population Health, London School of Hygiene and Tropical Medicine, London, UK

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Malaria programme design and evaluation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 03 December 2019

https://doi.org/10.21956/wellcomeopenres.17109.r37247

✔ **Peter Macharia**
Population Health Unit, Kenya Medical Research Institute - Wellcome Trust Research Programme, Nairobi, Kenya

I have no further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Disease mapping and spatial analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 07 November 2019

https://doi.org/10.21956/wellcomeopenres.16953.r36714

?

Caroline A. Lynch
Faculty of Epidemiology & Population Health, London School of Hygiene and Tropical Medicine, London, UK

The authors describe an open-source, easily accessible application that has great potential for use by policy-makers or National Malaria Control Programmes deciding on how best to target interventions to ensure greatest coverage relative to burden of disease.

My main comments are around the conclusions of the tool and its performance. The authors conclude that MaDD can present publicly available data in an easily interactable and navigable way. This is true. However, they also assert that this tool could inform local decision-makers asking questions such as 'where should we prioritise the targeting of IRS rounds this season'. This highlights the fact that the limitations of the application are not fully discussed.

Those limitations include timeliness of MAP revisions and incorporation of new prevalence data into models and ITN coverage data for example, relative to timing of decision-making at local levels. Is it true that a national policy maker could rely on this application to make year-on-year decisions as to where to target interventions and if not, then what would be required to allow that to happen?

Another limitation related to above is the confidence with which estimates can be made at lower administrative levels. For example, how useful are the maps in DRC if they are only available at provincial level? If policy makers add in layers of lower administrative levels is there enough prevalence data to have the confidence to predict down to those levels?

The authors correctly state that one of the biggest challenges in developing successful digital tools is getting them adopted and used. This tool is a streamlined simple tool which is really appealing, but the simplicity of it may mask the complexity of malaria epidemiology at sub-national levels and lead to a lack of confidence by users (e.g. National control programmes) of the data presented to some extent.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Malaria programme design and evaluation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 29 Nov 2019

Sean Tomlinson, Liverpool School of Tropical Medicine, Liverpool, UK

We thank Caroline Lynch for her insightful and practical comments that we have responded to below.

"The authors describe an open-source, easily accessible application that has great potential for use by policy-makers or National Malaria Control Programmes deciding on how best to target interventions to ensure greatest coverage relative to burden of disease."

"My main comments are around the conclusions of the tool and its performance. The authors conclude that MaDD can present publicly available data in an easily interactable and navigable way. This is true. However, they also assert that this tool could inform local decision-makers asking questions such as 'where should we prioritise the targeting of IRS rounds this season'. This highlights the fact that the limitations of the application are not fully discussed."

Response: We agree that improved discussion of the limitations of the application and approach would be useful for the reader. We already assert that the application is a prototype and a step towards useful tools for local decision-makers, but we have added two paragraphs to the discussion expanding on the limitations of this prototype.

"Those limitations include timeliness of MAP revisions and incorporation of new prevalence data into models and ITN coverage data for example, relative to timing of decision-making at local levels. Is it true that a national policy maker could rely on this application to make year-on-year decisions as to where to target interventions and if not, then what would be required to allow that to happen?"

Response: This is an important point that limitations of decision-supporting applications and the data behind them need to be considered carefully. The timeliness of data is particularly relevant. We have presented a prototype and our aim is to show that useful decision-supporting applications can be built with open-source software. There is indeed a risk that applications could provide poor advice to decision-makers because of these issues. However, we suggest that such applications could make it easier to give decision-makers access to the currently-best-available data (than for-example commercial software applications or paper reports both of which could take longer than modifying an open-source application). We cover this in the new discussion paragraph 4.

"Another limitation related to above is the confidence with which estimates can be made at lower administrative levels. For example, how useful are the maps in DRC if they are only available at provincial level? If policy makers add in layers of lower administrative levels is there enough prevalence data to have the confidence to predict down to those levels?"

Response: We also agree that care needs to be taken with the spatial resolution that data are aggregated to. In the future, we would like to see more guidance on the confidence that can be had in using such global, modeled datasets at a local level. We cover this in the new discussion paragraph 5.

"The authors correctly state that one of the biggest challenges in developing successful digital tools is getting them adopted and used. This tool is a streamlined simple tool which is really appealing, but the simplicity of it may mask the complexity of malaria epidemiology at sub-national levels and lead to a lack of confidence by users (e.g. National control programmes) of the data presented to some extent."

Response: We agree that there is a risk that local users may not trust applications such as this if they feel that it is over-simplified and does not represent the situation on the ground that they see. Our aim was to provide a prototype demonstrating that streamlined applications can be made. The beauty of open-source software is that if local users decide that we have gone too far and taken too much out, it is relatively straightforward to add additional data or functionality. Such modification would not require paying a northern hemisphere software company (or us) more money, but instead could be done by local data-scientists or R developers. We cover this in the new discussion paragraph 5.

Competing Interests: No competing interests were disclosed.

Reviewer Report 10 October 2019

https://doi.org/10.21956/wellcomeopenres.16953.r36713

?

Peter Macharia
Population Health Unit, Kenya Medical Research Institute - Wellcome Trust Research Programme, Nairobi, Kenya

The authors describe an application that has a big potential, especially for policy makers to make informed decisions by making public health-related datasets/maps easily accessible via an online platform. My main comments include allowing for more flexibility for the end-user (allow the user to pre-load a shapefile of choice or present more options) and incorporation of more public health-related datasets.

1. While refining MAP data to four products makes the app easily navigable, what of other users who might be interested in other products and not necessarily the four that were chosen for MaDD?

2. Are there plans to extend the app and have more products? There are more spatial products in the public domain that would benefit from being repackaged at a subnational scale for district managers. For example, the population distribution maps, gridded surface showing travel time to different tiers of healthcare. This is because the end users of the app are mainly policy makers as opposed to researchers.

3. The subnational boundaries are dynamic and vary by size, shape and number within a country. It would be important for the policy makers to be able to load different sub-national units within the app or add functionality to the app to load boundaries from other sources (GADM, HDX, DIVA-GIS and others).

4. It is also often the case that a country has different administrative boundaries at level 2, 3 and 4 and users would be interested at lower scale variation if they are district managers. For example, under the current implementation of MaDD, Kenya is shown by its 8 former provinces while the current structure is 47 sub national units (counties) representing 47 county governments which are further broken down to sub counties and wards.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Disease mapping and spatial analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 29 Nov 2019

Sean Tomlinson, Liverpool School of Tropical Medicine, Liverpool, UK

We thank Peter Macharia for his detailed and helpful comments that we have responded to below.

Response to 1: This is a good point. We recognise that the four data layers we chose will not satisfy all potential users. There is a tension between usability and flexibility. Our aim was to show that highly usable applications can be created with open-source code. Our application is a prototype. Dialog with users is required to establish where additional data layers are required and then the open-source code can be modified to incorporate them. We include this text in two paragraphs we have added to the discussion.

Response to 2: In short, yes, there are plans to extend this work. Specifically, we have won funding to develop reusable R software building blocks to make it easier to build applications like this with different data sources and functionalities. We added mention of this to the discussion paragraph 5.

Response to 3: We agree that applications should be able to use boundaries from different sources. In this prototype, we used a single set of boundaries to maintain high usability. As we provide the source code it would be relatively easy for developers to use different boundaries. A primary aim of our upcoming project, to develop reusable R building blocks for spatial data in Africa, is to make it easier to use different boundary data sets.

Response to 4: An excellent point that sub-national boundaries are constantly changing and that it is difficult for global public data providers to keep up-to-date. Countries often have finer scale sub-national and administrative boundaries. When designing applications for national-level decision making it should be possible to use these other national data sets. Flexible application design could allow default pre-loaded boundaries to be used with the option to load other boundaries if they are available. We added mention of this to the discussion paragraph 5.

Competing Interests: The authors declare no competing interests

Other Publications

This section contains a list of all the publication with which I have been involved over the course of my MRC DTP for which I am a contributing author.

Evolution of the Insecticide Target Rdl in African Anopheles Is Driven by Interspecific and Interkaryotypic Introgression

Mol Biol Evol 2020/05

PMID: 32449755

DOI: 10.1093/molbev/msaa128

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.

Parasites & Vectors 2018/10

PMID: 30348209

DOI: 10.1186/s13071-018-3101-4

Genome-wide transcriptional analyses in Anopheles mosquitoes reveal an unexpected association between salivary gland gene expression and insecticide resistance.

BMC Genomics 2018/03

PMID: 29587635

DOI: 10.1186/s12864-018-4605-1

Appendix A

The tables contained within this section represent a complete summary of the significant signals of Patterson's D statistic identifying throughout these analyses. The data in each table represent a specific chromosomal arm. Each row within a table has information as to the specific A, B, C and D populations used, and the range for which the significant signal encompasses.

An aggregated average of the Patterson's D statistic is given for each region. Also included for each entry are the genes that are contained within each region, 'missing' denotes that no name for a given locus was provided in the gene features file, where a '.' denotes a complete absence of data for the region in gene features file. A given A, B, C and D comparison may appear more than once for a given chromosomal arm. This represents where we identified multiple non-contiguous stretches of genomic sequence with an associated significant Patterson's D statistic.

Table 4. Significant regions of Patterson's D statistic found between *An. gambiae* and *An. arabiensis* for chromosome 2L.

A	B	C	D	start	stop	d_average	genes
TZ-MUL-gam	TZ-MUH-gam	UG-TOR-ara	chri	46583750	46674780	-0.27891	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
TZ-MUL-gam	TZ-MUH-gam	TZ-MOS-ara	chri	46578462	46673797	-0.2791	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
TZ-MUL-gam	TZ-MUH-gam	TZ-MUL-ara	chri	46608230	46692650	-0.26666	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
TZ-MUL-gam	TZ-MUH-gam	MW-CHI-ara	chri	46581635	46675096	-0.2854	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)',

							'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
UG-KAN-gam	TZ-MUH-gam	UG-TOR-ara	chri	46615850	46696691	-0.28234	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
UG-KAN-gam	TZ-MUH-gam	TZ-MOS-ara	chri	46605081	46688479	-0.30818	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
UG-KAN-gam	TZ-MUH-gam	TZ-MUL-ara	chri	46607942	46688024	-0.2999	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
UG-KAN-gam	TZ-MUH-gam	MW-CHI-ara	chri	46599498	46683101	-0.31682	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
UG-KAN-gam	TZ-MUH-gam	TZ-TAR-ara	chri	46593657	46678262	-0.31103	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
UG-TOR-gam	TZ-MUH-gam	UG-TOR-ara	chri	46604871	46684185	-0.31711	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
UG-TOR-gam	TZ-MUH-gam	TZ-MOS-ara	chri	46591801	46674435	-0.31829	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
UG-TOR-gam	TZ-MUH-gam	TZ-MUL-ara	chri	46593147	46674494	-0.31302	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
UG-TOR-gam	TZ-MUH-gam	MW-CHI-ara	chri	46576227	46698004	-0.28723	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']
UG-TOR-gam	TZ-MUH-gam	TZ-TAR-ara	chri	46610912	46688571	-0.32745	['DNA topoisomerase 2-associated protein PAT1', 'LRIM10 leucine-rich immune protein (Short)', 'neuronal guanine nucleotide exchange factor', 'LRIM7 leucine-rich immune protein (Short)', 'LRIM8B leucine-rich immune protein (Short)', 'hairy and enhancer of split related with YRPW motif', 'missing', 'LRIM8A leucine-rich immune protein (Short)', 'LRIM9 leucine-rich immune protein (Short)']

Table 5. Significant regions of Patterson's D statistic found between *An. gambiae* and *An. arabiensis* for chromosome 2R.

A	B	C	D	start	stop	d_average	genes
TZ-MUL-gam	UG-KAN-gam	MW-CHI-ara	chri	28965596	29052195	0.146109	['CYP6Z4 cytochrome P450', 'mRpS23 28S ribosomal protein S232C mitochondrial', 'RNA-binding protein Nova', 'heterogeneous nuclear ribonucleoprotein F-H', 'mRNA (2-O-methyladenosine-N6-)-methyltransferase']
TZ-MUL-gam	UG-KAN-gam	TZ-TAR-ara	chri	28370472	28557313	0.108417	['COEAE6O carboxylesterase alpha esterase', 'CYP6P5 cytochrome P450', 'solute carrier family 8 (sodium-calcium exchanger)', 'CYP6P1 cytochrome P450', 'CYP6P15P cytochrome P450', 'CYP6P3 cytochrome P450', 'lipase', 'CYP6P2 cytochrome P450', 'CYP6AD1 cytochrome P450', 'Sodium-potassium-transporting ATPase subunit alpha', 'CYP6P4 cytochrome P450', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'missing']
TZ-MUL-gam	UG-KAN-gam	TZ-TAR-ara	chri	28972902	29060308	0.160914	['CYP6Z4 cytochrome P450', 'mRpS23 28S ribosomal protein S232C mitochondrial', 'RNA-binding protein Nova', 'heterogeneous nuclear ribonucleoprotein F-H', 'mRNA (2-O-methyladenosine-N6-)-methyltransferase']
TZ-MUL-gam	UG-TOR-gam	UG-TOR-ara	chri	28337593	28521773	0.11246	['COEAE6O carboxylesterase alpha esterase', 'CYP6P5 cytochrome P450', 'solute carrier family 8 (sodium-calcium exchanger)', 'CYP6P1 cytochrome P450', 'CYP6P15P cytochrome P450', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'CYP6AD1 cytochrome P450', 'Sodium-potassium-transporting ATPase subunit alpha', 'CYP6P4 cytochrome P450', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'missing']
TZ-MUL-gam	UG-TOR-gam	UG-TOR-ara	chri	28976116	29060996	0.180595	['CYP6Z4 cytochrome P450', 'mRpS23 28S ribosomal protein S232C mitochondrial', 'RNA-binding protein Nova', 'heterogeneous nuclear ribonucleoprotein F-H', 'mRNA (2-O-methyladenosine-N6-)-methyltransferase']
TZ-MUL-gam	UG-TOR-gam	TZ-MUL-ara	chri	28309783	28436679	0.113519	['solute carrier family 8 (sodium-calcium exchanger)', 'missing', 'Sodium-potassium-transporting ATPase subunit alpha']
TZ-MUL-gam	UG-TOR-gam	UG-TOR-ara	chri	28976116	29060996	0.180595	['CYP6Z4 cytochrome P450', 'mRpS23 28S ribosomal protein S232C mitochondrial', 'RNA-binding protein Nova', 'heterogeneous nuclear ribonucleoprotein F-H', 'mRNA (2-O-methyladenosine-N6-)-methyltransferase']
TZ-MUL-gam	TZ-MUH-gam	UG-TOR-ara	chri	27963741	28341220	-0.44728	['GPRTAK1 putative tachykinin receptor 1', 'PPO1 prophenoloxidase 1', 'endoribonuclease Dicer', 'Sodium-potassium-transporting ATPase subunit alpha', 'Trunk', 'Niemann-Pick Type C-2', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'ribosomal RNA small subunit methyltransferase H', 'Niemann-Pick C2 protein', 'pyridoxal phosphate phosphatase PHOSPHO2', 'protein-tyrosine phosphatase', 'SPN-E ATP-dependent RNA helicase spindle-E', 'missing', 'alpha-tocopherol transfer protein-like protein', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor']
TZ-MUL-gam	TZ-MUH-gam	UG-TOR-ara	chri	28406404	28537569	-0.44661	['COEAE6O carboxylesterase alpha esterase', 'CYP6P5 cytochrome P450', 'solute carrier family 8 (sodium-calcium exchanger)', 'CYP6P1 cytochrome P450', 'CYP6P15P cytochrome P450', 'CYP6P3 cytochrome P450', 'lipase', 'CYP6P2 cytochrome P450', 'CYP6AD1 cytochrome P450', 'CYP6P4 cytochrome P450', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'missing']

TZ-MUL-gam	TZ-MUH-gam	UG-TOR-ara	chri	28706574	28793439	-0.44539	['.']
TZ-MUL-gam	TZ-MUH-gam	TZ-MOS-ara	chri	27840309	28932363	-0.53834	['CLIPA15 CLIP-domain serine protease', 'PPO1 prophenoloxidase 1', 'Ras-related protein R-Ras2', 'rRNA small subunit pseudouridine methyltransferase Nep1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'adenosine deaminase2C tRNA-specific 22C TAD2 homolog', 'CYP6P1 cytochrome P450', 'single-strand selective monofunctional uracil DNA glycosylase', 'lipase', 'GPRNNA2 putative GPCR class a orphan receptor 2', 'V-type H+-transporting ATPase subunit B', '45 kDa calcium-binding protein', 'pyridoxal phosphate phosphatase PHOSPHO2', 'COP9 signalosome complex subunit 5', 'SPN-E ATP-dependent RNA helicase spindle-E', ', ', 'GPRNNA3 putative GPCR class a orphan receptor 3', 'GPRTAK1 putative tachykinin receptor 1', '3-oxoacyl-(acyl-carrier-protein) synthase II', 'endoribonuclease Dicer', 'ERO1-like protein alpha', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'CLIPD6 CLIP-domain serine protease', 'COEAE6O carboxylesterase alpha esterase', 'DNA excision repair protein ERCC-4', 'protein HEXIM1-2', 'Cystatin-like protein', 'CYP6P15P cytochrome P450', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'CYP6AA1 cytochrome P450', 'Sodium-potassium-transporting ATPase subunit alpha', 'cathepsin F', 'CLIPD4 CLIP-domain serine protease', 'CYP6AA2 cytochrome P450', 'GPRNPR1 putative neuropeptide receptor 1', 'Tetratricopeptide repeat protein 30 homolog']
TZ-MUL-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29170325	29262995	-0.4854	['CTLMA9 C-type lectin (CTL) - mannose binding', 'Control protein HCTL029', 'missing', 'Salivary C-type lectin']
TZ-MUL-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29367589	29454546	-0.48072	['missing']
TZ-MUL-gam	TZ-MUH-gam	TZ-MUL-ara	chri	27986208	28619534	-0.50227	['PPO1 prophenoloxidase 1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'adenosine deaminase2C tRNA-specific 22C TAD2 homolog', 'CYP6P1 cytochrome P450', 'single-strand selective monofunctional uracil DNA glycosylase', 'lipase', 'pyridoxal phosphate phosphatase PHOSPHO2', 'COP9 signalosome complex subunit 5', 'SPN-E ATP-dependent RNA helicase spindle-E', 'GPRTAK1 putative tachykinin receptor 1', 'endoribonuclease Dicer', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'COEAE6O carboxylesterase alpha esterase', 'protein HEXIM1-2', 'Cystatin-like protein', 'CYP6P15P cytochrome P450', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'Sodium-potassium-transporting ATPase subunit alpha', 'cathepsin F', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'GPRNPR1 putative neuropeptide receptor 1', 'Tetratricopeptide repeat protein 30 homolog']
TZ-MUL-gam	TZ-MUH-gam	TZ-MUL-ara	chri	28677978	28810154	-0.48458	['GPRNNA2 putative GPCR class a orphan receptor 2', 'missing']
TZ-MUL-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29367589	29454546	-0.48072	['missing']
TZ-MUL-gam	TZ-MUH-gam	MW-CHI-ara	chri	27990949	28634602	-0.50152	['PPO1 prophenoloxidase 1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'adenosine deaminase2C tRNA-specific 22C TAD2 homolog', 'CYP6P1 cytochrome P450', 'single-strand selective monofunctional uracil DNA glycosylase', 'lipase', 'pyridoxal phosphate phosphatase PHOSPHO2', 'COP9 signalosome complex subunit 5', 'SPN-E ATP-dependent RNA helicase spindle-E', 'GPRTAK1 putative tachykinin receptor 1', 'endoribonuclease Dicer', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'COEAE6O carboxylesterase alpha esterase', 'protein HEXIM1-2', 'Cystatin-like protein', 'CYP6P15P cytochrome P450', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'Sodium-potassium-transporting ATPase subunit alpha', 'cathepsin F', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'GPRNPR1 putative neuropeptide receptor 1', 'Tetratricopeptide repeat protein 30 homolog']
TZ-MUL-gam	TZ-MUH-gam	MW-CHI-ara	chri	28689515	28821156	-0.52966	['GPRNNA2 putative GPCR class a orphan receptor 2', '.']
TZ-MUL-gam	TZ-MUH-gam	MW-CHI-ara	chri	28867514	28960491	-0.44256	['missing', 'NADH dehydrogenase (ubiquinone) 1 subcomplex unknown 2', 'Lipid storage droplets surface-binding protein 1', 'GPRMGL4 putative metabotropic glutamate receptor 4']
TZ-MUL-gam	TZ-MUH-gam	MW-CHI-ara	chri	29148406	29243033	-0.47901	['Control protein HCTL029', 'splicing factor U2AF 65 kDa subunit', 'Salivary C-type lectin', 'CTLMA9 C-type lectin (CTL) - mannose binding', 'missing']
TZ-MUL-gam	TZ-MUH-gam	TZ-TAR-ara	chri	27963803	28285700	-0.46868	['GPRTAK1 putative tachykinin receptor 1', 'PPO1 prophenoloxidase 1', 'endoribonuclease Dicer', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'WD repeat-containing protein 85', 'CLIPD1 protein', 'ribosomal RNA small subunit methyltransferase H', 'pyridoxal phosphate phosphatase PHOSPHO2', 'protein-tyrosine phosphatase', 'SPN-E ATP-dependent RNA helicase spindle-E', 'missing', 'alpha-tocopherol transfer protein-like protein', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor']
TZ-MUL-gam	TZ-MUH-gam	TZ-TAR-ara	chri	28350014	28543316	-0.46045	['COEAE6O carboxylesterase alpha esterase', 'CYP6P5 cytochrome P450', 'solute carrier family 8 (sodium-calcium exchanger)', 'CYP6P1 cytochrome P450', 'CYP6P15P cytochrome P450', 'CYP6P3 cytochrome P450', 'lipase', 'CYP6P2 cytochrome P450', 'CYP6AD1 cytochrome P450', 'Sodium-potassium-transporting ATPase subunit alpha', 'CYP6P4 cytochrome P450', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'missing']
TZ-MUL-gam	TZ-MUH-gam	TZ-TAR-ara	chri	28714833	28802551	-0.44965	['.']
TZ-MUL-gam	TZ-MUH-gam	MW-CHI-ara	chri	29148406	29243033	-0.47901	['Control protein HCTL029', 'splicing factor U2AF 65 kDa subunit', 'Salivary C-type lectin', 'CTLMA9 C-type lectin (CTL) - mannose binding', 'missing']
UG-KAN-gam	TZ-MUH-gam	UG-TOR-ara	chri	27977194	28334074	-0.46899	['GPRTAK1 putative tachykinin receptor 1', 'PPO1 prophenoloxidase 1', 'endoribonuclease Dicer', 'Sodium-potassium-transporting ATPase subunit alpha', 'Trunk', 'Niemann-Pick Type C-2', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'ribosomal RNA small subunit methyltransferase H', 'Niemann-Pick C2 protein', 'pyridoxal phosphate phosphatase PHOSPHO2', 'protein-tyrosine phosphatase', 'SPN-E ATP-dependent RNA helicase spindle-E', 'missing', 'alpha-tocopherol transfer protein-like protein', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor']
UG-KAN-gam	TZ-MUH-gam	UG-TOR-ara	chri	28391675	28571664	-0.47619	['COEAE6O carboxylesterase alpha esterase', 'CYP6P5 cytochrome P450', 'solute carrier family 8 (sodium-calcium exchanger)', 'CYP6P1 cytochrome P450', 'adenosine deaminase2C tRNA-specific 22C TAD2 homolog', 'protein HEXIM1-2', 'CYP6P15P cytochrome P450', 'CYP6P3 cytochrome P450', 'lipase', 'single-strand selective monofunctional uracil DNA glycosylase', 'CYP6P2 cytochrome P450', 'CYP6AD1 cytochrome P450', 'CYP6P4 cytochrome P450', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'missing']
UG-KAN-gam	TZ-MUH-gam	UG-TOR-ara	chri	28687626	28809466	-0.49159	['GPRNNA2 putative GPCR class a orphan receptor 2', '.']
TZ-MUL-gam	TZ-MUH-gam	MW-CHI-ara	chri	29148406	29243033	-0.47901	['Control protein HCTL029', 'splicing factor U2AF 65 kDa subunit', 'Salivary C-type lectin', 'CTLMA9 C-type lectin (CTL) - mannose binding', 'missing']
UG-KAN-gam	TZ-MUH-gam	TZ-MOS-ara	chri	27972182	28819363	-0.5786	['PPO1 prophenoloxidase 1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'adenosine deaminase2C tRNA-specific 22C TAD2 homolog', 'CYP6P1 cytochrome P450', 'single-strand selective monofunctional uracil DNA glycosylase', 'lipase', 'GPRNNA2 putative GPCR class a orphan receptor 2', 'V-type H+-transporting ATPase subunit B', 'pyridoxal phosphate phosphatase PHOSPHO2', 'COP9 signalosome complex subunit 5', 'SPN-E ATP-dependent RNA helicase spindle-E', ', ', 'GPRTAK1 putative tachykinin receptor 1', 'endoribonuclease Dicer', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'COEAE6O carboxylesterase alpha esterase', 'protein HEXIM1-2', 'Cystatin-like protein', 'CYP6P15P cytochrome P450', 'WD repeat-containing protein 85', 'CLIPD1 protein', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'Sodium-potassium-transporting ATPase subunit alpha', 'cathepsin F', 'CYP6AA1

							cytochrome P450', 'CYP6AA2 cytochrome P450', 'GPRNPR1 putative neuropeptide receptor 1', 'Tetratricopeptide repeat protein 30 homolog']
UG-KAN-gam	TZ-MUH-gam	TZ-MOS-ara	chri	28863734	29000367	-0.49662	['NADH dehydrogenase (ubiquinone) 1 subcomplex unknown 2', 'CYP6Z4 cytochrome P450', 'Lipid storage droplets surface-binding protein 1', 'mRpS23 28S ribosomal protein S232C mitochondrial', 'RNA-binding protein Nova', 'heterogeneous nuclear ribonucleoprotein F-H', 'mRNA (2-O-methyladenosine-N6-)-methyltransferase', 'missing', 'GPRMGL4 putative metabotropic glutamate receptor 4']
UG-KAN-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29175628	29262422	-0.50504	['CTLMA9 C-type lectin (CTL) - mannose binding', 'Control protein HCTL029', 'missing', 'Salivary C-type lectin']
UG-KAN-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29362649	29446074	-0.54433	['missing']
UG-KAN-gam	TZ-MUH-gam	TZ-MUL-ara	chri	27953933	28552301	-0.52898	['PPO1 prophenoloxidase 1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'CYP6P1 cytochrome P450', 'lipase', 'pyridoxal phosphate phosphatase PHOSPHO2', 'SPN-E ATP-dependent RNA helicase spindle-E', 'GPRTAK1 putative tachykinin receptor 1', 'endoribonuclease Dicer', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'COEAE6O carboxylesterase alpha esterase', 'CYP6P15P cytochrome P450', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'Sodium-potassium-transporting ATPase subunit alpha', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450']
UG-KAN-gam	TZ-MUH-gam	TZ-MUL-ara	chri	28666626	28837315	-0.5146	['GPRNNA2 putative GPCR class a orphan receptor 2', 'V-type H+-transporting ATPase subunit B', 'GPRNNA3 putative GPCR class a orphan receptor 3', '.', 'missing']
UG-KAN-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29175628	29262422	-0.50504	['CTLMA9 C-type lectin (CTL) - mannose binding', 'Control protein HCTL029', 'missing', 'Salivary C-type lectin']
UG-KAN-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29362649	29446074	-0.54433	['missing']
UG-KAN-gam	TZ-MUH-gam	MW-CHI-ara	chri	28005936	28614652	-0.52131	['PPO1 prophenoloxidase 1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'adenosine deaminase2C tRNA-specific 22C TAD2 homolog', 'CYP6P1 cytochrome P450', 'single-strand selective monofunctional uracil DNA glycosylase', 'lipase', 'pyridoxal phosphate phosphatase PHOSPHO2', 'COP9 signalosome complex subunit 5', 'SPN-E ATP-dependent RNA helicase spindle-E', 'GPRTAK1 putative tachykinin receptor 1', 'endoribonuclease Dicer', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'COEAE6O carboxylesterase alpha esterase', 'protein HEXIM1-2', 'Cystatin-like protein', 'CYP6P15P cytochrome P450', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'Sodium-potassium-transporting ATPase subunit alpha', 'cathepsin F', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'Tetratricopeptide repeat protein 30 homolog']
UG-KAN-gam	TZ-MUH-gam	MW-CHI-ara	chri	28669062	28971303	-0.49323	['NADH dehydrogenase (ubiquinone) 1 subcomplex unknown 2', 'Lipid storage droplets surface-binding protein 1', 'GPRNNA2 putative GPCR class a orphan receptor 2', 'V-type H+-transporting ATPase subunit B', 'GPRNNA3 putative GPCR class a orphan receptor 3', '.', 'missing', 'GPRMGL4 putative metabotropic glutamate receptor 4']
UG-KAN-gam	TZ-MUH-gam	MW-CHI-ara	chri	29150976	29240446	-0.50015	['Control protein HCTL029', 'splicing factor U2AF 65 kDa subunit', 'Salivary C-type lectin', 'CTLMA9 C-type lectin (CTL) - mannose binding', 'missing']
UG-KAN-gam	TZ-MUH-gam	MW-CHI-ara	chri	29338424	29461953	-0.47175	['missing']
UG-KAN-gam	TZ-MUH-gam	TZ-TAR-ara	chri	28005441	28552690	-0.48456	['PPO1 prophenoloxidase 1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'CYP6P1 cytochrome P450', 'lipase', 'pyridoxal phosphate phosphatase PHOSPHO2', 'SPN-E ATP-dependent RNA helicase spindle-E', 'GPRTAK1 putative tachykinin receptor 1', 'endoribonuclease Dicer', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'COEAE6O carboxylesterase alpha esterase', 'CYP6P15P cytochrome P450', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'Sodium-potassium-transporting ATPase subunit alpha', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450']
UG-KAN-gam	TZ-MUH-gam	TZ-TAR-ara	chri	28715850	28798798	-0.49728	['.']
UG-KAN-gam	TZ-MUH-gam	MW-CHI-ara	chri	29150976	29240446	-0.50015	['Control protein HCTL029', 'splicing factor U2AF 65 kDa subunit', 'Salivary C-type lectin', 'CTLMA9 C-type lectin (CTL) - mannose binding', 'missing']
UG-KAN-gam	TZ-MUH-gam	MW-CHI-ara	chri	29338424	29461953	-0.47175	['missing']
UG-TOR-gam	TZ-MUH-gam	UG-TOR-ara	chri	27979616	28333820	-0.46636	['GPRTAK1 putative tachykinin receptor 1', 'PPO1 prophenoloxidase 1', 'endoribonuclease Dicer', 'Sodium-potassium-transporting ATPase subunit alpha', 'Trunk', 'Niemann-Pick Type C-2', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'ribosomal RNA small subunit methyltransferase H', 'Niemann-Pick C2 protein', 'pyridoxal phosphate phosphatase PHOSPHO2', 'protein-tyrosine phosphatase', 'SPN-E ATP-dependent RNA helicase spindle-E', 'missing', 'alpha-tocopherol transfer protein-like protein', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor']
UG-TOR-gam	TZ-MUH-gam	UG-TOR-ara	chri	28391019	28519033	-0.50883	['COEAE6O carboxylesterase alpha esterase', 'CYP6P5 cytochrome P450', 'solute carrier family 8 (sodium-calcium exchanger)', 'CYP6P1 cytochrome P450', 'CYP6P15P cytochrome P450', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'CYP6AD1 cytochrome P450', 'CYP6P4 cytochrome P450', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'missing']
UG-TOR-gam	TZ-MUH-gam	UG-TOR-ara	chri	28682469	28806422	-0.47563	['.', 'missing']
UG-KAN-gam	TZ-MUH-gam	MW-CHI-ara	chri	29338424	29461953	-0.47175	['missing']
UG-TOR-gam	TZ-MUH-gam	TZ-MOS-ara	chri	27941552	28832246	-0.57396	['PPO1 prophenoloxidase 1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'adenosine deaminase2C tRNA-specific 22C TAD2 homolog', 'CYP6P1 cytochrome P450', 'single-strand selective monofunctional uracil DNA glycosylase', 'lipase', 'GPRNNA2 putative GPCR class a orphan receptor 2', 'V-type H+-transporting ATPase subunit B', 'pyridoxal phosphate phosphatase PHOSPHO2', 'COP9 signalosome complex subunit 5', 'SPN-E ATP-dependent RNA helicase spindle-E', '.', 'GPRNNA3 putative GPCR class a orphan receptor 3', 'GPRTAK1 putative tachykinin receptor 1', 'endoribonuclease Dicer', 'ERO1-like protein alpha', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'COEAE6O carboxylesterase alpha esterase', 'protein HEXIM1-2', 'Cystatin-like protein', 'CYP6P15P cytochrome P450', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'Sodium-potassium-transporting ATPase subunit alpha', 'cathepsin F', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'GPRNPR1 putative neuropeptide receptor 1', 'Tetratricopeptide repeat protein 30 homolog']
UG-TOR-gam	TZ-MUH-gam	TZ-MOS-ara	chri	28872069	28960526	-0.52309	['missing', 'NADH dehydrogenase (ubiquinone) 1 subcomplex unknown 2', 'Lipid storage droplets surface-binding protein 1', 'GPRMGL4 putative metabotropic glutamate receptor 4']

UG-TOR-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29138460	29264087	-0.50876	['Control protein HCTL029', 'splicing factor U2AF 65 kDa subunit', 'Salivary C-type lectin', 'CTLMA9 C-type lectin (CTL) - mannose binding', 'missing']
UG-TOR-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29362605	29444567	-0.52452	['missing']
UG-TOR-gam	TZ-MUH-gam	TZ-MUL-ara	chri	27954546	28547424	-0.53194	['PPO1 prophenoloxidase 1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'CYP6P1 cytochrome P450', 'lipase', 'pyridoxal phosphate phosphatase PHOSPHO2', 'SPN-E ATP-dependent RNA helicase spindle-E', 'GPRTAK1 putative tachykinin receptor 1', 'endoribonuclease Dicer', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'COEAE6O carboxylesterase alpha esterase', 'CYP6P15P cytochrome P450', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'Sodium-potassium-transporting ATPase subunit alpha', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450']
UG-TOR-gam	TZ-MUH-gam	TZ-MUL-ara	chri	28708907	28831662	-0.52699	['GPRNNA2 putative GPCR class a orphan receptor 2', 'GPRNNA3 putative GPCR class a orphan receptor 3', '.']
UG-TOR-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29138460	29264087	-0.50876	['Control protein HCTL029', 'splicing factor U2AF 65 kDa subunit', 'Salivary C-type lectin', 'CTLMA9 C-type lectin (CTL) - mannose binding', 'missing']
UG-TOR-gam	TZ-MUH-gam	TZ-MOS-ara	chri	29362605	29444567	-0.52452	['missing']
UG-TOR-gam	TZ-MUH-gam	MW-CHI-ara	chri	27962270	28619555	-0.51852	['PPO1 prophenoloxidase 1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'adenosine deaminase2C tRNA-specific 22C TAD2 homolog', 'CYP6P1 cytochrome P450', 'single-strand selective monofunctional uracil DNA glycosylase', 'lipase', 'pyridoxal phosphate phosphatase PHOSPHO2', 'COP9 signalosome complex subunit 5', 'SPN-E ATP-dependent RNA helicase spindle-E', 'GPRTAK1 putative tachykinin receptor 1', 'endoribonuclease Dicer', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'COEAE6O carboxylesterase alpha esterase', 'protein HEXIM1-2', 'Cystatin-like protein', 'CYP6P15P cytochrome P450', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'Sodium-potassium-transporting ATPase subunit alpha', 'cathepsin F', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450', 'GPRNPR1 putative neuropeptide receptor 1', 'Tetratricopeptide repeat protein 30 homolog']
UG-TOR-gam	TZ-MUH-gam	MW-CHI-ara	chri	28674339	28971776	-0.49001	['NADH dehydrogenase (ubiquinone) 1 subcomplex unknown 2', 'Lipid storage droplets surface-binding protein 1', 'GPRNNA2 putative GPCR class a orphan receptor 2', 'GPRNNA3 putative GPCR class a orphan receptor 3', '.', 'missing', 'GPRMGL4 putative metabotropic glutamate receptor 4']
UG-TOR-gam	TZ-MUH-gam	MW-CHI-ara	chri	29148835	29277000	-0.4943	['Control protein HCTL029', 'galactokinase', 'splicing factor U2AF 65 kDa subunit', 'Salivary C-type lectin', 'CTLMA9 C-type lectin (CTL) - mannose binding', 'missing']
UG-TOR-gam	TZ-MUH-gam	MW-CHI-ara	chri	29373370	29454364	-0.46976	['missing']
UG-TOR-gam	TZ-MUH-gam	TZ-TAR-ara	chri	27949989	28549842	-0.48457	['PPO1 prophenoloxidase 1', 'CYP6AD1 cytochrome P450', 'protein-tyrosine phosphatase', 'CYP6P4 cytochrome P450', 'CYP6P5 cytochrome P450', 'C-1-tetrahydrofolate synthase2C mitochondrial precursor', 'Trunk', 'CYP6P1 cytochrome P450', 'lipase', 'pyridoxal phosphate phosphatase PHOSPHO2', 'SPN-E ATP-dependent RNA helicase spindle-E', 'GPRTAK1 putative tachykinin receptor 1', 'endoribonuclease Dicer', 'ERO1-like protein alpha', 'Niemann-Pick Type C-2', 'solute carrier family 8 (sodium-calcium exchanger)', 'Med7 Mediator of RNA polymerase II transcription subunit 7', 'Niemann-Pick C2 protein', 'missing', 'alpha-tocopherol transfer protein-like protein', 'COEAE6O carboxylesterase alpha esterase', 'CYP6P15P cytochrome P450', 'CLIPD1 protein', 'WD repeat-containing protein 85', 'CYP6P3 cytochrome P450', 'CYP6P2 cytochrome P450', 'ribosomal RNA small subunit methyltransferase H', 'Sodium-potassium-transporting ATPase subunit alpha', 'CYP6AA1 cytochrome P450', 'CYP6AA2 cytochrome P450']
UG-TOR-gam	TZ-MUH-gam	TZ-TAR-ara	chri	28712722	28795263	-0.51203	['.']
UG-TOR-gam	TZ-MUH-gam	MW-CHI-ara	chri	29148835	29277000	-0.4943	['Control protein HCTL029', 'galactokinase', 'splicing factor U2AF 65 kDa subunit', 'Salivary C-type lectin', 'CTLMA9 C-type lectin (CTL) - mannose binding', 'missing']
UG-TOR-gam	TZ-MUH-gam	MW-CHI-ara	chri	29373370	29454364	-0.46976	['missing']

Table 6. Significant regions of Patterson's D statistic found between *An. gambiae* and *An. arabiensis* for chromosome 3L.

A	B	C	D	start	stop	d_average	genes
TZ-MUL-gam	UG-KAN-gam	UG-TOR-ara	chri	62826	1266174	0.136433	['dehydrogenase-reductase SDR family member 11 precursor', 'serine-threonine-protein kinase ATR', 'Groucho', 'keren', 'heat shock protein 110kDa', 'guanine nucleotide-binding protein subunit beta-5', 'DNA-directed RNA polymerase II subunit RPB3', 'SH3-binding domain kinase', 'aquaporin', 'snRNA-activating protein complex subunit 1', 'eukaryotic peptide chain release factor subunit', 'nardilysin', 'missing']
TZ-MUL-gam	UG-KAN-gam	UG-TOR-ara	chri	5388267	5519732	0.131719	['GPRMAC1 putative muscarinic acetylcholine receptor 1', 'missing']
TZ-MUL-gam	UG-KAN-gam	TZ-MOS-ara	chri	5357311	5473747	0.123365	['GPRMAC1 putative muscarinic acetylcholine receptor 1', 'missing']

TZ-MUL-gam	UG-KAN-gam	UG-TOR-ara	chri	5388267	5519732	0.131719	['GPRMAC1 putative muscarinic acetylcholine receptor 1', 'missing']
TZ-MUL-gam	UG-KAN-gam	TZ-MUL-ara	chri	3233865	3393966	-0.07568	['DnaJ homolog subfamily C member 11', 'adenylate cyclase 2', 'missing', 'dynein heavy chain 92C axonemal']
TZ-MUL-gam	UG-KAN-gam	TZ-MUL-ara	chri	5396600	5523207	0.119357	['GPRMAC1 putative muscarinic acetylcholine receptor 1', 'missing']
TZ-MUL-gam	UG-KAN-gam	MW-CHI-ara	chri	5374546	5501504	0.108069	['GPRMAC1 putative muscarinic acetylcholine receptor 1', 'missing']
TZ-MUL-gam	UG-KAN-gam	TZ-MUL-ara	chri	5396600	5523207	0.119357	['GPRMAC1 putative muscarinic acetylcholine receptor 1', 'missing']
UG-KAN-gam	UG-TOR-gam	TZ-MUL-ara	chri	7042046	7317705	-0.09786	['NHLRC2 protein', 'elongation factor 1-beta', 'female reproductive tract protease GLEANR_896', 'mitochondrial import inner membrane translocase subunit Tim8 B', 'eupolytin', 'DnaJ (Hsp40) homolog subfamily C', 'glycine transporter', 'Early trypsin', 'Clathrin light chain', 'missing', 'Insulin-like peptide 1', 'Insulin-like peptide 2 precursor', 'Insulin-like peptide 3 precursor']
TZ-MUL-gam	UG-KAN-gam	TZ-MUL-ara	chri	5396600	5523207	0.119357	['GPRMAC1 putative muscarinic acetylcholine receptor 1', 'missing']

Table 7. Significant regions of Patterson's D statistic found between *An. gambiae* and *An. arabiensis* for chromosome 3R.

A	B	C	D	start	stop	d_average	genes
TZ-MUL-gam	UG-KAN-gam	TZ-TAR-ara	chri	52107122	53091457	0.146995	['YY1-associated factor 2', 'autophagy-related protein 7', 'V-type H+-transporting ATPase subunit D', 'mitochondrial glutamate carrier 1', 'WD repeat and SOF domain-containing protein 1', 'missing', 'wingless-type MMTV integration site family2C member 5', 'nucleolar complex protein 3', 'guanine nucleotide exchange factor VAV', 'Pleckstrin homology domain containing2C family F (with FYVE domain) member 2', 'F-box and leucine-rich repeat protein 14']
TZ-MUL-gam	UG-TOR-gam	TZ-TAR-ara	chri	2734375	2812347	0.101282	['xanthine dehydrogenase', 'missing', 'xanthine dehydrogenase-oxidase', 'ABCC12 ATP-binding cassette transporter (ABC transporter) family C member 12']
UG-KAN-gam	UG-TOR-gam	TZ-MOS-ara	chri	28505376	28623877	-0.06398	['GSTE5 glutathione S-transferase epsilon class 5', 'palmitoyltransferase ZDHHC24', 'Indanol dehydrogenase', 'solute carrier family 39 (zinc transporter)2C member 9', 'GSTE7 glutathione S-transferase epsilon class 7', 'GSTE4 glutathione S-transferase epsilon class 4', 'GSTE6 glutathione S-transferase epsilon class 6', 'GSTE1 glutathione S-transferase epsilon class 1', 'GSTE8 glutathione S-transferase epsilon class 8', 'collagen type IV alpha', 'GSTE3 glutathione S-transferase epsilon class 3', 'missing', 'GSTE2 glutathione S-transferase epsilon class 2']
UG-KAN-gam	TZ-MUH-gam	TZ-MUL-ara	chri	10565780	10667224	-0.29526	['Ras-related protein Rab-14', 'WW domain-containing oxidoreductase', 'RNA-binding protein 39', 'UTP6 U3 small nucleolar RNA-associated protein 6 homolog', 'peptidyl-tRNA hydrolase ICT1', 'SWI-SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1', 'pre-rRNA-processing protein TSR3', 'IAP5 inhibitor of apoptosis 5', 'N-acetyltransferase 11', 'missing']
UG-KAN-gam	TZ-MUH-gam	TZ-TAR-ara	chri	10561905	10665391	-0.28925	['Ras-related protein Rab-14', 'WW domain-containing oxidoreductase', 'RNA-binding protein 39', 'UTP6 U3 small nucleolar RNA-associated protein 6 homolog', 'peptidyl-tRNA hydrolase ICT1', 'SWI-SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1', 'pre-rRNA-processing protein TSR3', 'IAP5 inhibitor of apoptosis 5', 'N-acetyltransferase 11', 'missing']
UG-TOR-gam	TZ-MUH-gam	UG-TOR-ara	chri	13771912	13871217	-0.30325	['APG7A autophagy related gene', 'Pigeon protein', 'protein SDA1', 'alpha-aminoadipic semialdehyde synthase', 'zinc finger CCCH domain-containing protein 15', 'GPRMGL2 putative metabotropic glutamate receptor 2', 'missing']
UG-TOR-gam	TZ-MUH-gam	TZ-MUL-ara	chri	10565519	10663369	-0.33012	['Ras-related protein Rab-14', 'WW domain-containing oxidoreductase', 'RNA-binding protein 39', 'UTP6 U3 small nucleolar RNA-associated protein 6 homolog', 'peptidyl-tRNA hydrolase ICT1', 'SWI-SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1', 'pre-rRNA-processing protein TSR3', 'IAP5 inhibitor of apoptosis 5', 'N-acetyltransferase 11', 'missing']

Table 8. Significant regions of Patterson's D statistic found between *An. gambiae* and *An. arabiensis* for chromosome X.

A	B	C	D	start	stop	d_average	genes
TZ-MUL-gam	UG-KAN-gam	TZ-MOS-ara	chri	17936617	18150774	0.263826	['platelet-activating factor acetylhydrolase IB subunit beta-gamma', 'CTL7 C-type lectin (CTL)', 'missing']
TZ-MUL-gam	UG-KAN-gam	MW-CHI-ara	chri	17907431	18126629	0.266283	['platelet-activating factor acetylhydrolase IB subunit beta-gamma', 'missing']
TZ-MUL-gam	UG-KAN-gam	TZ-TAR-ara	chri	17874704	18098795	0.273468	['missing']

Appendix B

Appendix B contains the Python 3 code used in the conversion of the data obtained from Neafsey *et al.* (2013). The need for this conversion script is borne from the inherent differences in the methods used to store and generate the data of the 16 Genomes Project and the Ag1000G. Since all analyses downstream of this conversion and built on the foundational assumption that this conversion script is correct, it has been included here for evaluation. Similarly, Appendix C contain Python 3 code used to verify the transformations and operation performed upon the original data. This additional script containing the unit testing of the functions is also included here as it served to provide confidence in the conversion script that served as the part of the basis for the all the downstream analyses.

NB: This conversion script was also used in the conversion of the 16 Genomes Data in the analyses of the Grau-Bove *et al.* (2020), a study for which I am credited as a co-author.

```
%run ../73-arabiensis-qc/20180822-module-setup.ipynb

chroms = ['2L', '2R', '3L', '3R', 'X']
species = ['mela', 'meru', 'quad', 'epir', 'chri']

def remove_indels(zarrh, chrom, species):
"""This removes indels from the Fontaine data, which would hinder AIM finding"""

### Loading in the root of the data structures
group = zarr.open_group(os.path.join(
    '/kwiat/vector/ag1000g/release/phase1.AR3/agc/UnifiedGenotyper',
    '{species}_ref_ug_vqsr_cnvrt_sort.zarr/'.format(species=species)),mode='r')

### Load data
pos = allel.SortedIndex(group['{chrom}/variants/POS'.format(chrom=chrom)])
gt = group['{chrom}/calldata/GT'.format(chrom=chrom)]
ref = group['{chrom}/variants/REF'.format(chrom=chrom)]
alt = group['{chrom}/variants/ALT'.format(chrom=chrom)]

### Utility
mylen = np.vectorize(len)

### ALT
count_alt = mylen(alt)
alt_count = np.sum(count_alt, axis=1)

# This array returns true at each position if the alt is fixed for the reference
# or has only one alternate allele.
is_ref_or_snp = alt_count <= 1

### REF
### Let's pull out the positions where it's a single base
length_of_ref = mylen(ref) <= 1

### return the logical and of these 2 arrays

# this removes the second instance of a duplicated position
is_not_dup = np.concatenate([[True], np.diff(pos) > 0])

# this is the final bool array to filter the data with, and results in positions that are not multiallelic
# not containing indel and not containing duplicated positions
bool_mask = length_of_ref & is_ref_or_snp & is_not_dup

# save the results to zarr, dont use a dict, way too large.
chrom_g = zarrh.require_group(chrom)
sp_g = chrom_g.require_group(species)

sp_g.array("REF", data=np.compress(bool_mask, ref, axis=0), dtype='|S1')
sp_g.array("ALT", data=np.compress(bool_mask, alt, axis=0), dtype='|S1')
sp_g.array("POS", data=pos.compress(bool_mask, axis=0))
sp_g.array("GT", data=np.compress(bool_mask, gt, axis=0))
sp_g.array("indel_rm_filter", data=bool_mask)

print(chrom, species, 'DONE')
```

```
# create zarr group for indel cleared data
zf = zarr.open_group(
    "/kwiat/vector/observatory/analysis/20181128-fontain-indels-rm",
    mode='a')

# clear the indels
for chrom in chroms:
    for sp in species:
        remove_indels(zf, chrom, sp)

# create zarr group for unified data
compress_zarr = zarr.open_group(
    "/kwiat/vector/observatory/analysis/20181128-fontaine-indels-rm-ref-rm",
    mode="a")

def remove_ref_empties(zarrh, chrom, species):
    """This removes positions where the REF is N, this is different to uncalled positions and is something
    that can cause headaches"""

    # Load in the data
    gt = allel.GenotypeChunkedArray(zf[chrom][species]['GT'])
    ref = zf[chrom][species]['REF'].astype("<U1")[:]
    alt = zf[chrom][species]['ALT'][:]
    pos = allel.SortedIndex(zf[chrom][species]['POS'])
    ac = allel.AlleleCountsChunkedArray(zf[chrom][species]['AC'])

    bad_sites = ref != 'N'

    gt = gt.compress(bad_sites, axis=0)
    ref = ref.compress(bad_sites, axis=0)
    alt = alt.compress(bad_sites, axis=0)
    pos = pos.compress(bad_sites, axis=0)
    ac = ac.compress(bad_sites, axis=0)

    chrom_g = zarrh.require_group(chrom)
    sp_g = chrom_g.require_group(species)

    sp_g.array("GT", data=gt)
    sp_g.array("REF", data=ref, dtype='|S1')
    sp_g.array("ALT", data=alt, dtype='|S1')
    sp_g.array("POS", data=pos)
    sp_g.array("AC", data=ac)

    print(chrom, species, 'DONE')

# unify the data
for chrom in chroms:
    for sp in species:
        remove_ref_empties(compress_zarr, chrom, sp)

# create zarr group for fixed alternate array
fix_zarr = zarr.open_group(
    "/kwiat/vector/observatory/analysis/20181203-fontaine-indels-rm-ref-rm-alt-fixed",
    mode="a")
```

```
def fix_alternate(zarrh, chrom, species):
    """This make it so that the ALT between species are in identical order"""
    # Load in the unified data
    gt = allel.GenotypeChunkedArray(compress_zarr[chrom][species]['GT'])
    ref = compress_zarr[chrom][species]['REF'].astype("<U1")[:]
    alt = compress_zarr[chrom][species]['ALT'][:]

    # Make a dict of desired ALT sequences
    alleles = set(["A", "C", "G", "T"])
    new_alt = np.empty((ref.shape[0], 3), dtype="|S1")

    alternate_dict = {}
    for base in alleles:
        alternate_dict[base] = sorted(list(alleles.symmetric_difference(base)))

    # Assign a new alts array
    for base in ['A', 'C', 'G', 'T']:
        new_alt[np.where(ref[:] == base)] = alternate_dict[base]
    new_alt = new_alt.astype('>U1')

    # @sean: This was the error. We hadn't concatenated
    # the ref to the new alts. Oops.
    na = np.hstack([ref[:, None], new_alt])
    assert na.shape[1] == 4, "wrong shape"

    # Map alleles with _new_ new alt
    allele_mapping = allel.create_allele_mapping(ref, alt, na)

    new_gt = gt.map_alleles(allele_mapping)

    # Check that missing counts are the same. This is a good check,
    # if we screw up we will add missing alleles
    missing_a = gt.count_missing(axis=1)
    missing_b = new_gt.count_missing(axis=1)
    assert np.array_equal(missing_a[:], missing_b[:]), "Bad mapping"

    # save alt fixed data
    chrom_g = fix_zarr.require_group(chrom)
    sp_g = chrom_g.require_group(species)

    sp_g.array("REF", data=ref, dtype='|S1')
    sp_g.array("ALT", data=new_alt, dtype='|S1')
    sp_g.array("POS", data=compress_zarr[chrom][species]["POS"])
    sp_g.array("GT", data=new_gt)
    sp_g.array("AC", data=new_gt.count_alleles(max_allele=3))

    print(chrom, species, 'DONE')

for chrom in chroms:
    for sp in species:
        fix_alternate(fix_zarr, chrom, sp)
```

```
%run ../73-arabiensis-qc/20180822-module-setup.ipynb

chroms = ['2L', '2R', '3L', '3R', 'X']

def get_bases_ix(gt, alt, ref, pos, loc):
    """Gets the bases at stated positions"""
    ix = pos.locate_keys(loc)
    gt = gt.compress(ix, axis=0)
    alt = alt.compress(ix, axis=0)
    ref = ref.compress(ix, axis=0)

    assert gt.shape[0] == alt.shape[0] == ref.shape[0]

    ac = gt.count_alleles(max_allele=3)
    allele = ac.max_allele()[:]
    results = np.empty((allele.shape[0]), dtype='object')

    for count, i in enumerate(allele):
        if i == 0:
            results[count] = ref[count]
        elif i == 1:
            results[count] = alt[count][0]
        elif i == 2:
            results[count] = alt[count][1]
        elif i == 3:
            results[count] = alt[count][2]
        else:
            print('error')

    return(results.astype('>U1'))


aim = zarr.open_group(
    "/kwiat/vector/observatory/analysis/20181205-fontain-aims", mode="r")


# Validation

def three_arr_is_equal(arr1, arr2, arr3):
    """takes 3 arrays and returns true if all are equal"""
    if np.array_equal(arr1, arr2):
        return(np.array_equal(arr1, arr3))
    else:
        print('All three not equal')
```

```
# ALT Flip

alt_fix = zarr.open_group(
    '/kwiat/vector/observatory/analysis/20181203-fontaine-indels-rm-union-alt-fixed/',
    mode='r')

alt_flip_gam_bases = get_bases_ix(
    gt = allel.GenotypeChunkedArray(alt_fix['2L/gam/GT']),
    alt = alt_fix['2L/gam/ALT'][:],
    ref = alt_fix['2L/gam/REF'][:],
    pos = allel.SortedIndex(alt_fix['2L/gam/POS']),
    loc = aim['2L/POS'])

alt_flip_arab_bases = get_bases_ix(
    gt = allel.GenotypeChunkedArray(alt_fix['2L/arab/GT']),
    alt = alt_fix['2L/arab/ALT'][:],
    ref = alt_fix['2L/arab/REF'][:],
    pos = allel.SortedIndex(alt_fix['2L/arab/POS']),
    loc = aim['2L/POS'])


assert alt_flip_gam_bases.shape[0] == alt_flip_arab_bases.shape[0] # returns true if okay
assert ~np.any(alt_flip_gam_bases == alt_flip_arab_bases) # returns true if okay
assert np.array_equal(alt_flip_gam_bases, aim['2L/gam_bases'])
assert np.array_equal(alt_flip_arab_bases, aim['2L/arab_bases'])

# Union

union = zarr.open_group(
    '/kwiat/vector/observatory/analysis/20181128-fontaine-indels-rm-union/',
    mode='r')

union_gam_bases = get_bases_ix(
    gt = allel.GenotypeChunkedArray(union['2L/gam/GT']),
    alt = union['2L/gam/ALT'][:],
    ref = union['2L/gam/REF'][:],
    pos = allel.SortedIndex(union['2L/gam/POS']),
    loc = aim['2L/POS'])

union_arab_bases = get_bases_ix(
    gt = allel.GenotypeChunkedArray(union['2L/arab/GT']),
    alt = union['2L/arab/ALT'][:],
    ref = union['2L/arab/REF'][:],
    pos = allel.SortedIndex(union['2L/arab/POS']),
    loc = aim['2L/POS'])

assert union_gam_bases.shape[0] == union_arab_bases.shape[0] # returns true if okay
assert ~np.any(union_gam_bases == union_arab_bases) # returns true if okay
assert np.array_equal(union_gam_bases, aim['2L/gam_bases'])
assert np.array_equal(union_arab_bases, aim['2L/arab_bases'])
```

```
# Indel_rm

rm_indel = zarr.open_group(
    '/kwiat/vector/observatory/analysis/20181128-fontain-indels-rm/',
    mode='r')

rm_indel_gam_bases = get_bases_ix(
    gt = allel.GenotypeChunkedArray(rm_indel['2L/gam/GT']),
    alt = rm_indel['2L/gam/ALT'][:],
    ref = rm_indel['2L/gam/REF'][:],
    pos = allel.SortedIndex(rm_indel['2L/gam/POS']),
    loc = aim['2L/POS'])

rm_indel_arab_bases = get_bases_ix(
    gt = allel.GenotypeChunkedArray(rm_indel['2L/arab/GT']),
    alt = rm_indel['2L/arab/ALT'][:],
    ref = rm_indel['2L/arab/REF'][:],
    pos = allel.SortedIndex(rm_indel['2L/arab/POS']),
    loc = aim['2L/POS'])

assert rm_indel_gam_bases.shape[0] == rm_indel_arab_bases.shape[0] # returns true if okay
assert ~np.any(rm_indel_gam_bases == rm_indel_arab_bases) # returns true if okay
assert np.array_equal(rm_indel_gam_bases, aim['2L/gam_bases'])
assert np.array_equal(rm_indel_arab_bases, aim['2L/arab_bases'])

# Let's just check that there's concordance between all three data sets
assert three_arr_is_equal(rm_indel_gam_bases, union_gam_bases, alt_flip_gam_bases)
assert three_arr_is_equal(rm_indel_arab_bases, union_arab_bases, alt_flip_arab_bases)
```

Appendix D

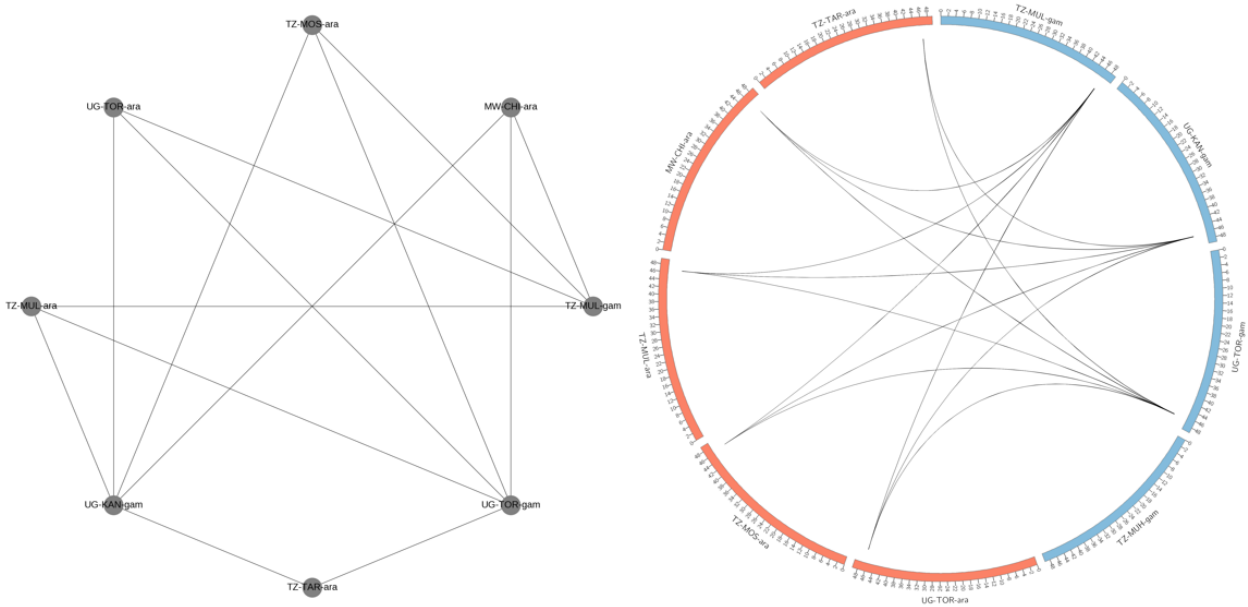


Figure 1 Network work and network genome graphs of significant signals found between An. gambaie and *An. arabiensis* on chromosome arm 2L.

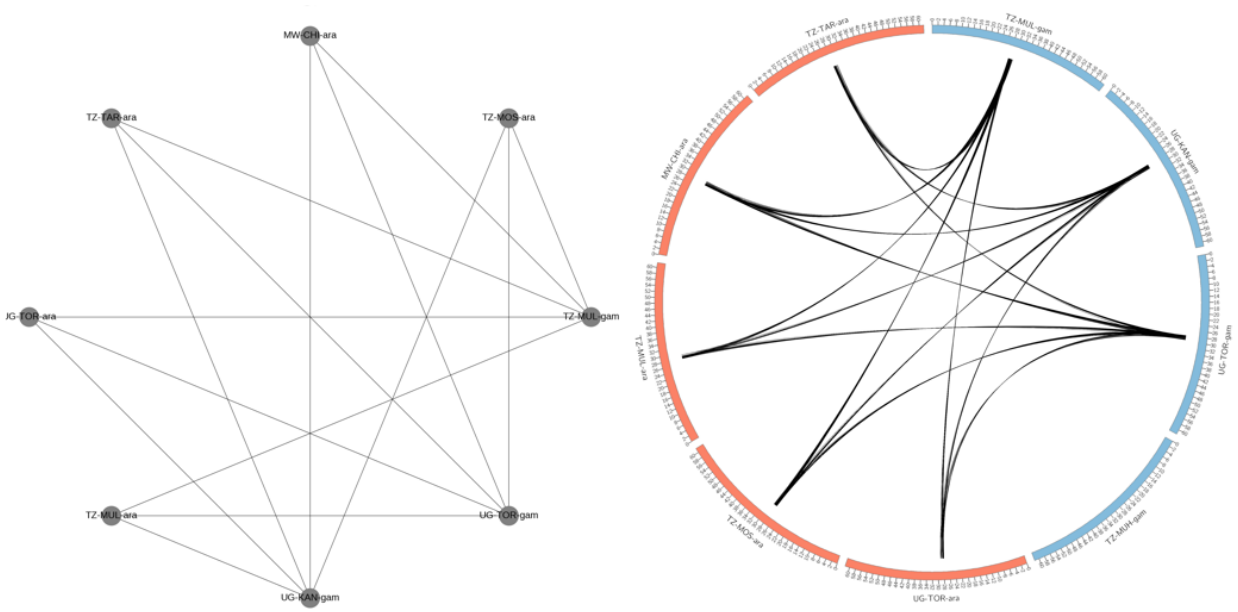


Figure 2 Network work and network genome graphs of significant signals found between An. gambaie and *An. arabiensis* on chromosome arm 2R.

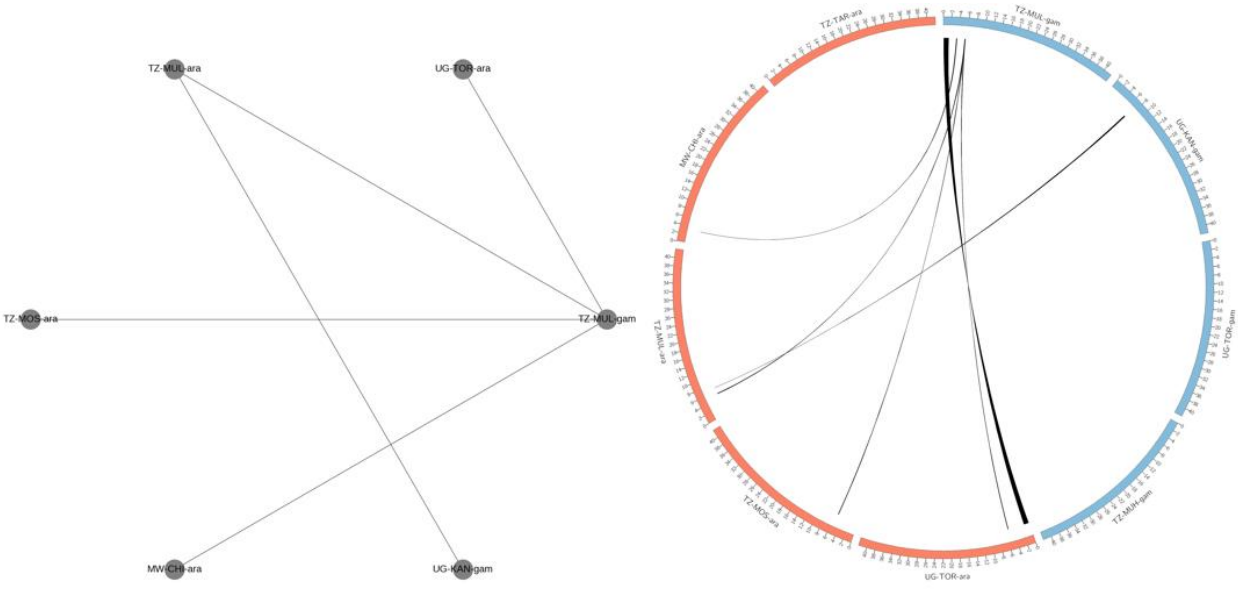


Figure 3 Network work and network genome graphs of significant signals found between An. gambaie and *An. arabiensis* on chromosome arm 3L.

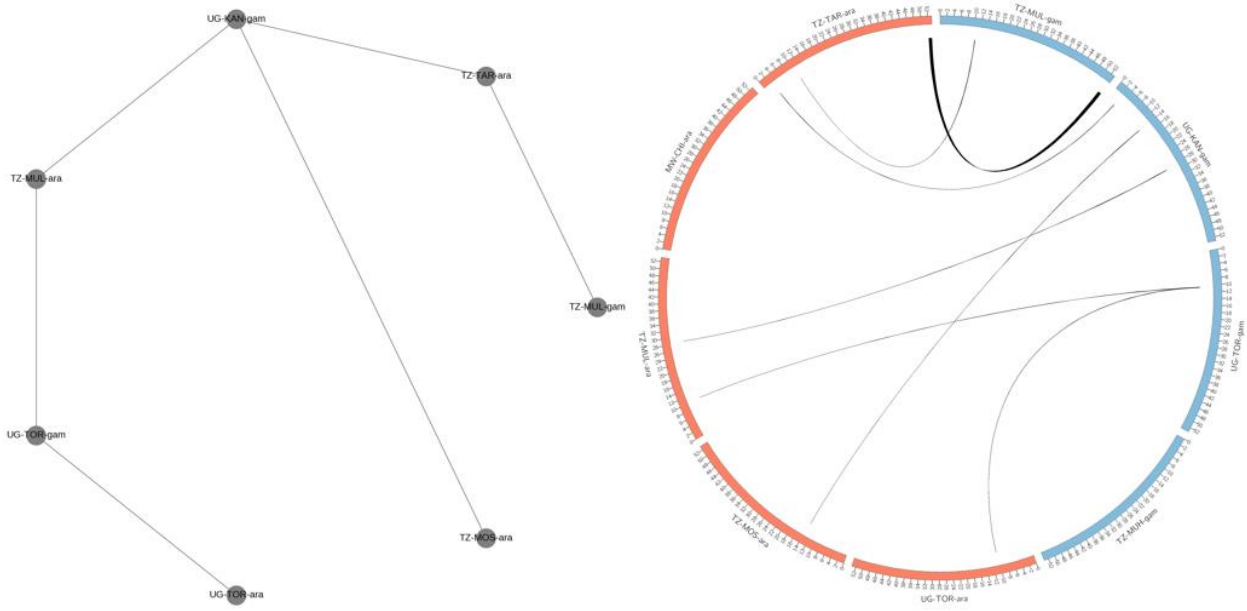


Figure 4 Network work and network genome graphs of significant signals found between An. gambaie and *An. arabiensis* on chromosome arm 3R.

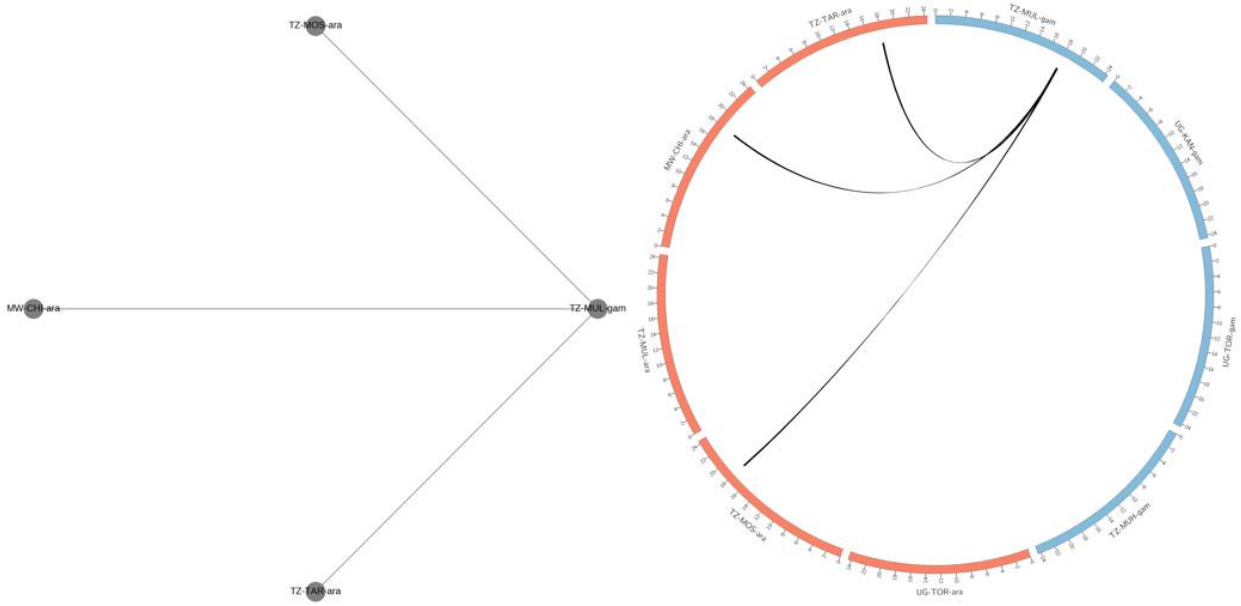


Figure 5 Network work and network genome graphs of significant signals found between An. gambaie and *An. arabiensis* on chromosome arm X.