The evolution and spread of target-site resistance to pyrethroid insecticides in the African malaria vectors *Anopheles gambiae* and *Anopheles coluzzii*

Chris S. Clarkson[1], Alistair Miles[2,1], Nicholas J. Harding[2], @@TODO[1], Dominic Kwiatkowski[1,2], Martin Donnelly[3,1], and The *Anopheles gambiae* 1000 Genomes Consortium[4]

[1]Sanger @@TODO

[2]Oxford @@TODO

[3]Liverpool @@TODO

[4]MalariaGEN @@TODO

Work in progress

Abstract

Resistance to pyrethroid insecticides is a major concern for malaria vector control, because these are the only compounds approved for use in insecticide-treated bed-nets (ITNs) and are also widely used for indoor residual spraying (IRS). Pyrethroids target the voltage-gated sodium channel (VGSC), an essential component of the mosquito nervous system, but mutations in the *Vgsc* gene can disrupt the activity of these insecticides, inducing a "knock-down resistance" phenotype. Here we use Illumina whole-genome sequence data from phase 1 of the *Anopheles gambiae* 1000 Genomes

Project (Ag1000G) to provide a comprehensive account of genetic variation at the *Vgsc* locus in mosquito populations from 8 African countries. In addition to three known resistance variants that alter the protein-coding sequence of the *Vgsc* gene, we describe 19 previously unknown non-synonymous variants at appreciable frequency in one or more populations. For each variant we predict a resistance phenotype based on genetic evidence for recent selection, patterns of linkage between variants, the position of the variant within the protein structure, and experimental evidence from other species. We use analyses of haplotype structure to refine our understanding of the origins and spread of these resistance variants between species and geographical locations. These analyses identify 10 distinct lineages, each of which carries one or more resistance alleles and appears to be undergoing rapid and recent expansion in one or more populations. The most successful and widespread resistance lineage (F1) originates in West Africa and has subsequently spread to countries in Central and Southern Africa. We also reconstruct a putative ancestral haplotype for each lineage, and analyse patterns of recombination to show that lineages are unrelated and thus represent independent outbreaks of resistance. Our data demonstrate that the molecular basis of pyrethroid resistance in African malaria vectors is more complex than previously appreciated, and provide a foundation for the development of new genetic tools to inform insecticide resistance management and track the further spread of resistance.

Introduction

An estimated 663 million cases of malaria were averted in Africa between 2000 and 2015 due to public health interventions, of which 68% were prevented by insecticide-treated bednets (ITNs) and @@N% through indoor residual spraying of insecticides (IRS). However, over this same period, insecticide resistance has become increasingly prevalent in malaria vector populations. Four chemical classes of insecticides – organophosphates, carbamates, pyrethroids and organochlorines – are licensed for use in public health, but only pyrethroids are approved by the World Health Organisation (WHO) for use in ITNs. Pyrethroids are also commonly used for IRS and in agriculture, and mosquito populations are under pressure to evolve molecular mechanisms of pyrethroid resistance. There is evidence that pyrethroid resistance has a direct impact on the effectiveness of ITNs and IRS, although assessing the impact on disease prevalence is difficult and has been hampered by the fact

that pyrethroid resistance is now so pervasive that it is nearly impossible to find fully susceptible mosquito populations to serve as controls. Nevertheless, the position of the WHO remains that insecticide resistance poses a grave threat to the future of malaria control in Africa (@@REF GPIRM). Improvements are needed in our ability to monitor resistance, and gaps must be filled in our knowledge of the molecular mechanisms of resistance.

The voltage-gated sodium channel (VGSC) is the physiological target of pyrethroids and of the organochlorine DDT. The VGSC protein is integral to the insect nervous system, involved in the transmission of nerve impulses. Both pyrethroids and DDT have a similar mode of action, binding to sites within the protein channel and preventing normal nerve function, causing paralysis ("knock-down") and then death. However, amino acid substitutions at key positions within the channel can alter the interaction between the channel and the insecticide molecule, and thereby substantially increase the dosage of insecticide required for knock-down. If this tolerance exceeds the dosage present in ITNs or on indoor surfaces following IRS, these interventions may be rendered ineffective. In the African malaria vectors *Anopheles gambiae* and *Anopheles coluzzii*, three substitutions have been found in natural populations and shown to cause pyrethroid and DDT resistance. Two of these substitutions occur in codon 995[1], with the Leucine → Phenylalanine (`L995F`) substitution prevalent in West and Central Africa, and the Leucine → Serine (`L995S`) substitution found in Central and East Africa. A third variant `N1570Y` has been found in association with `L995F` in Central Africa and shown to increase resistance above `L995F` alone.

Target-site resistance to pyrethroids and DDT has also been studied in a range of other insect species, including disease vectors as well as domestic and crop pests. Because of its essential function, the VGSC protein is highly conserved across insect species, and knowledge gained from one species is relevant to another. Many resistance-associated variants have been described in these other species, and thus there are many possible amino acid substitutions that could induce a resistance phenotype in malaria vectors, other than the known variants in codons 995 and 1570. Some of these variants are within

[1]Codon numbering is given here relative to transcript @@TODO as defined in the AgamP4.@@N gene annotations. A mapping of codon numbers from @@TRANSCRIPT to *Musca domestica* @@TRANSCRIPT is given in Table 1.

the trans-membrane channel, and thus may directly interact with insecticide molecules. However, functional studies have also demonstrated that variants within internal linker domains can substantially enhance the the level of resistance, when present in combination with channel modifications. Most previous studies of *An. gambiae* and/or *An. coluzzii* have performed targeted sequencing of small regions within the gene, and there has been no comprehensive survey of variation across the entire gene in multiple populations.

Insecticide resistance monitoring in malaria vector populations now often incorporates some form of genetic assay to detect the allele present at *Vgsc* codon 995. Both alleles are present at high frequency in multiple geographical locations, and the `L995F` allele is present in both *An. gambiae* and *An. coluzzii*. The extent of mosquito migration remains an open question, however mosquitoes do travel between different locations and have the potential to spread resistance alleles from one population to another (adaptive gene flow). Hybridization between mosquito species also occurs and has the potential to transfer resistance alleles between species (adaptive introgression). Studies in West African have shown that the `L995F` allele has been transferred from *An. gambiae* into *An. coluzzii* populations. A resistance allele may also arise independently in multiple populations, either because of multiple mutational events occurring after insecticides are introduced (selection on new mutations), or because resistance alleles were already present at low frequency in mosquito populations prior to insecticide use (selection on standing variation). Previous studies have found evidence that the `L995F` allele occurs on several different genetic backgrounds, suggesting multiple origins of resistance. However, these studies have used information from only a small region of the gene, and have limited resolution to make inferences about geographical origins or history of spread. Better information about the origins and spread of resistance could improve insecticide resistance monitoring and inform strategies for insecticide resistance management.

Here we provide a detailed and comprehensive account of genetic variation within the *Vgsc* gene using data from phase 1 of the *Anopheles gambiae* 1000 Genomes Project (Ag1000G). We use genotype and haplotype data derived from whole-genome Illumina sequencing of 765 individual mosquitoes collected from natural populations in 8 African countries to survey genetic diversity and study the evolutionary and demographic history of insecticide resistance at the *Vgsc* locus. Our results reveal an unexpected diversity

of molecular mechanisms of resistance, and shed new light on the evolutionary processes underlying the rapid increase in the prevalence of resistance across multiple mosquito populations.

Results

Functional variation

To identify single nucleotide polymorphisms (SNPs) with a potentially functional role in pyrethroid resistance, we extracted SNPs from the Ag1000G phase 1 data resource that alter the amino acid sequence of the VGSC protein, and computed their allele frequencies among 9 populations defined by species and country of origin. SNPs that confer resistance are expected to increase in frequency under selective pressure, and we refined the list of potentially functional SNPs to retain only those at an appreciable frequency (>5%) in one or more populations (Table 1). The resulting list comprises 20 SNPs, including the known `L995F`, `L995S` and `N1570Y` variants, and a further 17 SNPs not previously described in these species. We reported 15 of these novel SNPs in our initial analysis of the Ag1000G phase 1 data (@@REF Ag1000G), and we extend the analyses here to incorporate two tri-allelic SNPs affecting codons 402 and 410.

The two alleles in codon 995 are clearly the main drivers of resistance at this locus. The `L995F` allele at high frequency in populations of both species from West, Central and Southern Africa, and the `L995S` allele at high frequency among *An. gambiae* populations from Central and East Africa (Table 1; @@REF Ag1000G). All haplotypes carrying L995F or L995S have evidence for strong recent positive selection (@@REF Ag1000G). Both alleles were present in populations sampled from Cameroon and Gabon, including some individuals with a hybrid `L995F/S` genotype. Within these populations, the `L995F` and `L995S` alleles were (@@TODO were not?) in Hardy-Weinberg equilibrium (P=@@), thus there does not (@@does?) appear to be selection against hybrids.

The `I1527T` allele is present in *An. coluzzii* from Burkina Faso at 14% frequency, and there is evidence that haplotypes carrying this allele have been positively selected (@@REF Ag1000G). Codon 1527 occurs within trans-membrane domain segment `III.S6`, immediately adjacent to a second predicted binding pocket for pyrethroid molecules, thus it is

Table 1. Non-synonymous nucleotide variation in the voltage-gated sodium channel gene. AO=Angola; BF=Burkina Faso; GN=Guinea; CM=Cameroon; GA=Gabon; UG=Uganda; KE=Kenya; GW=Guinea-Bissau; *Ac*=*An. coluzzii*; *Ag*=*An. gambiae*. All variants are at 5% frequency or above in one or more of the 9 Ag1000G phase 1 populations, with the exception of `2,400,071 G>T` which is only found in the CM*Ag* population at 0.4% frequency but is included because another mutation (`2,400,071 G>A`) is found at the same position causing the same amino acid substitution (`M490I`); and `2,431,019 T>C` (`F1920S`) which is at 4% frequency in GA*Ag* but also found in CM*Ag* and linked to `L995F`.

Variant			Population allele frequency (%)									Function	
Position[1]	*Ag*[2]	*Md*[3]	AO*Ac*	BF*Ac*	GN*Ag*	BF*Ag*	CM*Ag*	GA*Ag*	UG*Ag*	KE	GW	Domain[4]	Resistance phenotype[5]
`2,390,177 G>A`	`R254K`	`R261`	0	0	0	0	32	21	0	0	0	`IN (I.S4-I.S5)`	`L995F` enhancer (predicted)
`2,391,228 G>C`	`V402L`	`V410`	0	7	0	0	0	0	0	0	0	`TM (I.S6)`	`I1527T` enhancer (predicted)
`2,391,228 G>T`	`V402L`	`V410`	0	7	0	0	0	0	0	0	0	`TM (I.S6)`	`I1527T` enhancer (predicted)
`2,399,997 G>C`	`D466H`	`-`	0	0	0	0	7	0	0	0	0	`IN (I.S6-II.S1)`	`L995F` enhancer (predicted)
`2,400,071 G>A`	`M490I`	`M508`	0	0	0	0	0	0	0	18	0	`IN (I.S6-II.S1)`	none (predicted)
`2,400,071 G>T`	`M490I`	`M508`	0	0	0	0	0	0	0	0	0	`IN (I.S6-II.S1)`	none (predicted)
`2,416,980 C>T`	`T791M`	`T810`	0	1	13	14	0	0	0	0	0	`TM (II.S1)`	`L995F` enhancer (predicted)
`2,422,651 T>C`	`L995S`	`L1014`	0	0	0	0	15	64	100	76	0	`TM (II.S6)`	driver
`2,422,652 A>T`	`L995F`	`L1014`	86	85	100	100	53	36	0	0	0	`TM (II.S6)`	driver
`2,424,384 C>T`	`A1125V`	`K1133`	9	0	0	0	0	0	0	0	0	`IN (II.S6-III.S1)`	none (predicted)
`2,425,077 G>A`	`V1254I`	`I1262`	0	0	0	0	0	0	0	0	5	`IN (II.S6-III.S1)`	none (predicted)
`2,429,617 T>C`	`I1527T`	`I1532`	0	14	0	0	0	0	0	0	0	`TM (III.S6)`	driver (predicted)
`2,429,745 A>T`*	`N1570Y`	`N1575`	0	26	10	22	6	0	0	0	0	`IN (III.S6-IV.S1)`	`L995F` enhancer
`2,429,897 A>G`	`E1597G`	`E1602`	0	0	6	4	0	0	0	0	0	`IN (III.S6-IV.S1)`	`L995F` enhancer (predicted)
`2,429,915 A>C`	`K1603T`	`K1608`	0	5	0	0	0	0	0	0	0	`TM (IV.S1)`	`L995F` enhancer (predicted)
`2,430,424 G>T`	`A1746S`	`A1751`	0	0	11	13	0	0	0	0	0	`TM (IV.S5)`	`L995F` enhancer (predicted)
`2,430,817 G>A`	`V1853I`	`V1858`	0	0	8	5	0	0	0	0	0	`IN (IV.S6-)`	`L995F` enhancer (predicted)
`2,430,863 T>C`	`I1868T`	`I1873`	0	0	18	25	0	0	0	0	0	`IN (IV.S6-)`	`L995F` enhancer (predicted)
`2,430,880 C>T`	`P1874S`	`P1879`	0	21	0	0	0	0	0	0	0	`IN (IV.S6-)`	`L995F` enhancer (predicted)
`2,430,881 C>T`	`P1874L`	`P1879`	0	7	45	26	0	0	0	0	0	`IN (IV.S6-)`	`L995F` enhancer (predicted)
`2,431,019 T>C`	`F1920S`	`Y1925`	0	0	0	0	1	4	0	0	0	`IN (IV.S6-)`	`L995F` enhancer (predicted)
`2,431,061 C>T`	`A1934V`	`A1939`	0	12	0	0	0	0	0	0	0	`IN (IV.S6-)`	`L995F` enhancer (predicted)
`2,431,079 T>C`	`I1940T`	`I1945`	0	4	0	0	7	0	0	0	0	`IN (IV.S6-)`	`L995F` enhancer (predicted)

[1] Position relative to the AgamP3 reference sequence, chromosome arm 2L. Variants marked with an asterisk (*) failed conservative variant filters applied genome-wide in the Ag1000G phase 1 AR3 callset, but appeared sound on manual inspection of read alignments.

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

[3] Codon numbering according to *Musca domestica* EMBL accession X96668 [1].

[4] Position of the variant within the protein. `IN`=internal domain; `TM`=trans-membrane domain. The protein contains four homologous repeats (I-IV), each having six transmembrane segments (1-6). Codes in parentheses identify the specific domain, e.g., "`I.S4`" refers to trans-membrane segment 4 in repeat I, and "`IS4-IS5`" refers to the linker segment between `I.S4` and `I.S5`.

[5] Phenotype predictions are based on population genetic evidence and have not been confirmed experimentally.

plausible that I1527T could alter insecticide binding (@@REF Dong). We also found that the two variant alleles affecting codon 402, both of which induce a V402L substitution, were in strong linkage with I1527T (D'>@@N; Figure 1), and almost all haplotypes carrying I1527T also carried a V402L substitution. The most parsimonious explanation for this pattern of linkage is that the I1527T mutation occurred first, and mutations in codon 402 subsequently arose on this genetic background. Codon 402 also occurs within a transmembrane segment (I.S6), and the V402L substitution has by itself been shown experimentally to increase pyrethroid resistance in @@species and *Xenopus* oocytes (@@REFs). However, because V402L appears secondary to I1527T in our cohort, we classify I1527T as a putative resistance driver and V402L as a putative enhancer. Because of the limited geographical distribution of these alleles, we hypothesize that the I1527T+V402L combination represents a pyrethroid resistance allele that arose in West African *An. coluzzii* populations; however, the L995F allele is at higher frequency (85%) in our Burkina Faso *An. coluzzii* population, and is known to be increasing in frequency (@@REFs), therefore L995F may provide a stronger resistance phenotype and is replacing I1527T+V402L in these populations.

Of the other 16 SNPs, 13 occurred almost exclusively in combination with L995F (Figure @@; @@REF Ag1000G). These include the N1570Y allele, known to enhance pyrethroid resistance in *An. gambiae* in combination with L995F. These also include two variants in codon 1874 (P1874S, P1874L). P1874S has previously been found in a colony of the crop pest *Plutoblah blahdiblah* with a pyrethroid resistance phenotype, but has not been shown to confer resistance experimentally. 10 of these variants, including N1570Y and P1874S/L, occur within internal linker domains of the protein, and so fit the model of variants that may enhance or compensate for the driver phenotype by modifying channel gating behaviour (@@CHECK; @@REFs). The remaining 3 variants are within transmembrane domains, and so may enhance resistance by @@TODO how. Because of the tight linkage between these 13 SNPs and the L995F allele, we classify all as putative L995F enhancers, although experimental work is required to confirm a resistance phenotype.

The remaining 3 variants (M490I, A1125V, V1254I) do not occur in combination with any known resistance allele, and do not appear to be associated with haplotypes under selection (@@REF Ag1000G). A possible exception is the M490I allele found at 18% frequency in

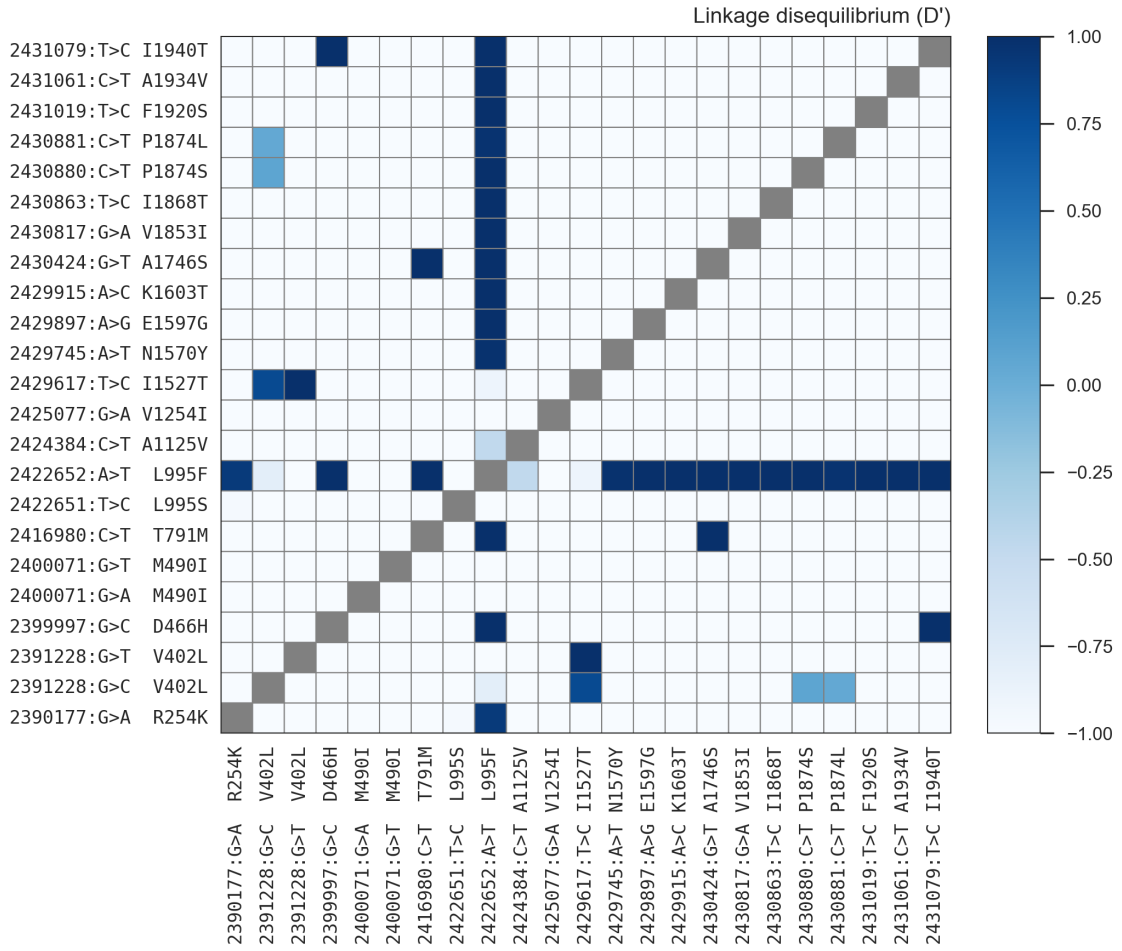

Figure 1. Linkage disequilibrium between non-synonymous variants. A value of 1 indicates that the two variants always occur in combination, and conversely a value of -1 indicates that the two variants never occur in combination. @TODO nuance this?

the Kenyan population, although the fact that this population has experienced a recent population crash makes it difficult to test for evidence of selection at this locus. All 3 variants occur in internal linker domains, and so do not fit the model of a resistance driver, although experimental work is required to rule out a resistance phenotype.

Haplotype structure

Although it is known that pyrethroid resistance is increasing in prevalence in malaria vector populations across Africa, it has not been clear whether this is being driven by the spread of resistance alleles via gene flow, or by resistance alleles emerging independently in multiple locations, or by some combination of both processes. The Ag1000G data resource provides a potentially rich source of information about the evolutionary and demographic

history of insecticide resistance in any given gene, because data are available not only for SNPs in gene coding regions, but also SNPs in introns and flanking intergenic regions, and in neighbouring genes. These additional variants can be used to analyse the genetic backgrounds (haplotypes) on which resistance alleles are found. In sexually reproducing species, DNA sequences are transmitted from parents to progeny in chunks, rearranged via recombination at each generation, and haplotypes convey information about this history of transmission and recombination, especially when haplotypes from many individuals can be compared.

In our initial analysis of the *Vgsc* (@@REF Ag1000G), we used 1710 biallelic SNPs from within the @@70 kbp *Vgsc* gene (@@N exonic, @@N intronic) to compute the number of SNP differences between all pairs of 1530 haplotypes derived from 765 wild-caught mosquitoes. This genetic distance measurement is a rough proxy for the degree of relatedness between haplotypes, in the sense that two haplotypes with a small number of SNP differences must be closely related and share a common ancestor in the recent past. This measurement cannot be used to directly estimate the time to most recent common ancestor (TMRCA) for any pair of haplotypes, however, because it does not account for the possibility of recombination events within the gene, which is increasingly likely for pairs of haplotypes that are more distantly related. Nevertheless, it provides a useful tool for exploring patterns of similarity and dissimilarity within the data. To visualise these patterns, we used the pairwise genetic distances to perform hierarchical clustering, which groups similar haplotypes together into clusters. We found that haplotypes carrying resistance alleles were grouped into 10 distinct clusters. Five of these clusters carried the `L995F` allele (labelled F1-F5), and a further five clusters carried `L995S` (labelled S1-S5). Within each cluster, haplotypes were nearly identical across all 1710 SNPs (spanning @@70 kbp), and therefore each cluster represents a collection of haplotypes with a very recent common ancestor. Within some of these clusters, we found haplotypes from mosquitoes collected from different locations. Specifically, cluster F1 contained haplotypes from Guinea, Burkina Faso, Cameroon and Angola; clusters @@ each contained haplotypes from Cameroon and Gabon; and cluster @@ contained haplotypes from Uganda and Kenya. The F1 cluster also contained haplotypes from both *An. gambiae* and *An. coluzzii* individuals. If we assume that haplotypes within each cluster share a common ancestor since the introduction

of insecticides, which is reasonable given the high degree of similarity, then each of these clusters provides evidence that resistance alleles have been spreading between geographical locations and species via adaptive gene flow. Here we present several new analyses of these haplotype data, to confirm our initial inferences regarding gene flow, and provide further details regarding the origins and movement of resistance alleles.

To provide an alternative view of the genetic similarity between haplotypes carrying resistance alleles, we used haplotype data from within the Vgsc gene region to construct median-joining networks (Figure 2). This analysis is very similar to hierarchical clustering, except that it allows for the reconstruction and placement of intermediate haplotypes that may not be observed in the data. We constructed these networks up to a maximum

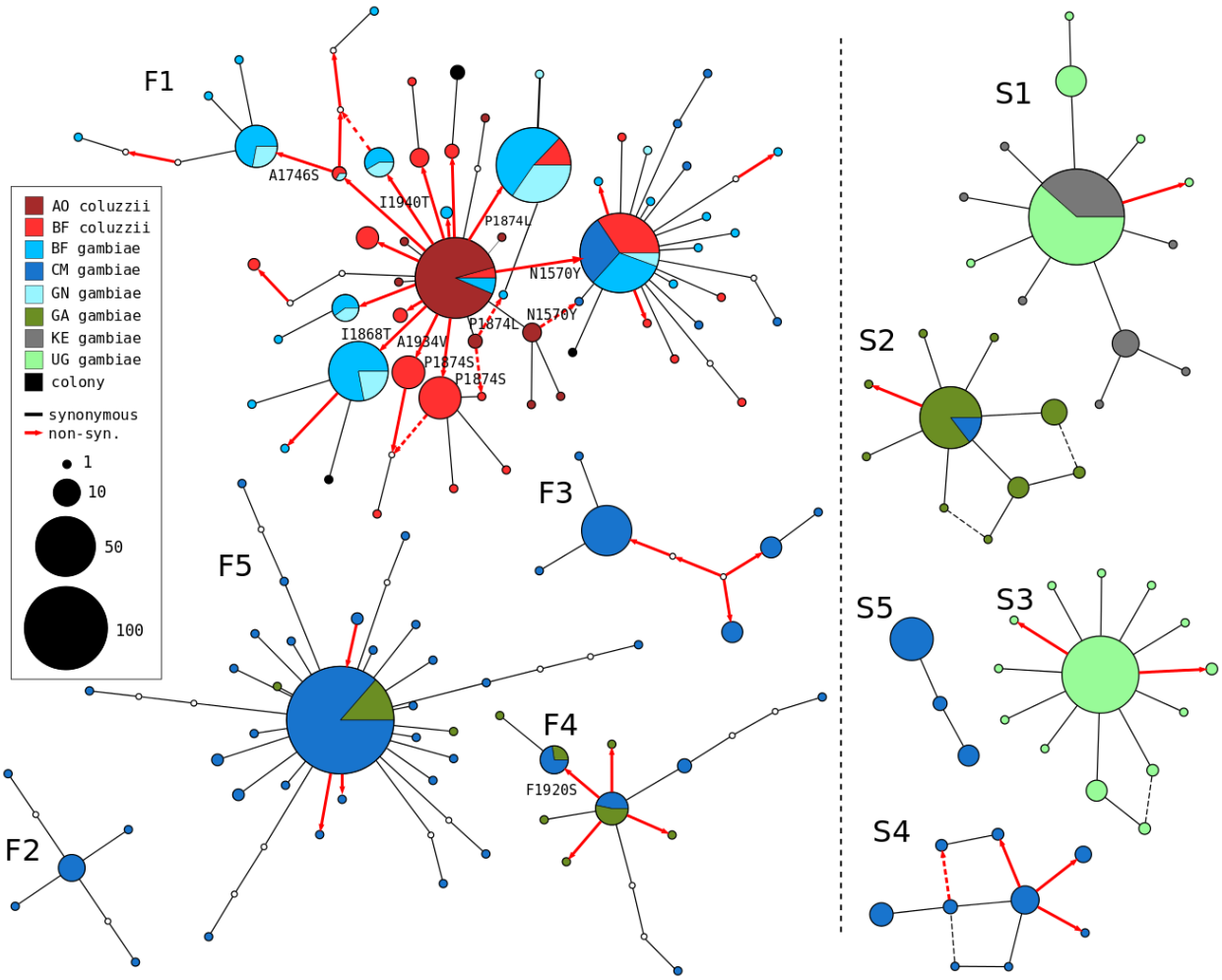

Figure 2. Haplotype networks. @@TODO redo the figure. @@TODO annotate non-syn edges in cluster F3. @@TODO mention if any clusters fixed for non-syn variants so not shown. @@TODO annotate other non-syn edges, e.g., in S4?

distance of @@2 SNP differences, to ensure that each connected component in the resulting networks represents a collection of haplotypes with a recent common ancestor, and thus which is also likely to be minimally affected by recombination within the gene. For haplotypes carrying `L995F`, the resulting network confirms the presence of five distinct clusters, with close correspondance to the clusters F1-F5 identified previously. The `L995S` network also confirms five distinct clusters, in concordance with our previous analysis.

The haplotype networks bring into sharp relief the explosive evolution of amino acid substitutions secondary to the `L995F` allele. Within the F1 network, nodes carrying non-synonymous variants radiate out from a central node carrying only `L995F`, indicating that the central node represents the ancestral haplotype carrying `L995F` alone which initially came under selection, and these secondary variants have arisen subsequently as new mutations. Many of the nodes carrying secondary variants are large, consistent with positive selection and a functional role for these secondary variants as enhancers of the `L995F` resistance phenotype. The F1 network also allows us to infer multiple introgression events between the two species. The central (ancestral) node comprises haplotypes from both species, as do nodes carrying the `N1570Y`, `P1874L`, and @@TODO one more variant@@. This structure is consistent with an initial introgression of the ancestral F1 haplotype, followed by introgression of haplotypes carrying secondary mutations. The contrast between the haplotype networks for the `L995F` and `L995S` alleles is striking because of the near-total absence of non-synonymous variation within the `L995S` networks. As we reported previously, this difference is highly significant – the ratio of non-synonymous to synonymous nucleotide diversity (@@piN/piS) is @@N times higher among haplotypes carrying `L995F` relative to haplotypes carrying `L995S` (@@Test; P=@@) (@@REF Ag1000G). Some secondary variants are present within the `L995S` networks, but all are at low frequency, and thus may be neutral or mildly deleterious variants that are hitch-hiking on selective sweeps for the `L995S` allele.

While the haplotype clustering and network analyses provide evidence for the spread of resistance alleles via adaptive gene flow, and for the secondary evolution of `L995F` enhancer alleles, they have several limitations. Within haplotype clusters where gene flow has occurred, they have poor resolution to infer the origin and direction of gene flow. This is because the analyses only leverage information about genetic distance within the *Vgsc*

gene, and for very recent events, insufficient time has elapsed for informative mutations to accumulate within this relatively small genome region. Also, the fact that we observe five distinct clusters for each of the codon 995 alleles suggests that each cluster is in some sense independent from the others, and thus gene flow is not required for resistance to emerge in multiple geographical locations. However, the threshold for the genetic distance at which we have chosen to divide haplotypes into different networks or clusters is to a certain extent arbitrary, and based on an intuitive sense of how much variation could have accumulated among the descendants of a single resistant ancestor since the onset of selective pressure. We also need to clarify what we mean by "independent", as there are several possible scenarios under which resistance could evolve in multiple populations in the absence of gene flow. Finally, analyses of genetic distance within a fixed genome region can be confounded by recombination events occurring within that region. For example, a recombination event within the *Vgsc* gene upstream of codon 995 could cause us to split a collection of haplotypes into two clusters, even though they are ancestrally related within the region downstream of the recombination event. In the next sub-sections we provide some conceptual foundations to help clarify these ambiguities, and use analyses of haplotype sharing from the genome regions flanking the *Vgsc* gene to provide finer resolution to diagnose recent gene flow events.

Insecticide resistance outbreaks

To provide an aid to further interpretation of the genetic data, and relating them to the challenges of insecticide resistance management, we introduce the concept of an **insecticide resistance outbreak**. Informally, we define a resistance outbreak by analogy with the epidemiological concept of an outbreak, as a rapid increase in the prevalence of insecticide resistance among mosquitoes at a particular place and time. Note that this does not imply that the overall abundance of mosquitoes is increase, just that the relative frequency of resistance within mosquito populations is increasing. We also require that all occurrences of insecticide resistance within the same outbreak are connected by a chain of transmission of resistance alleles from parent to progeny mosquitoes, and thus can be traced back to a single resistant common ancestor. A resistance outbreak can be **localised**, meaning that it affects a small group of mosquitoes of a single species from a limited

geographical area. Alternatively, a resistance outbreak may be **spreading**, meaning that resistance alleles have been transmitted since the introduction of insecticides by inter-breeding of mosquitoes of different species and/or originating from different geographical locations.

Our goal for the *Vgsc* gene can now be restated, which is to perform an insecticide resistance outbreak analysis. We would like to diagnose how many separate outbreaks have occurred, which outbreaks are localised, and which are spreading. For spreading outbreaks, we would like to reconstruct the path of transmission of resistance alleles between mosquito populations, and to provide information on the probable source. We would, of course, also like to identify the primary and secondary genetic factors that are driving each outbreak. Stated in this way, it is easier to discuss how this information is potentially relevant to insecticide resistance management, and to frame key epidemiological questions. For example, we would like to begin to build a picture of where and when local conditions have favoured the evolution of insecticide resistance, and whether those conditions are relatively patchy (and hence outbreaks are mainly localised) or whether conditions are

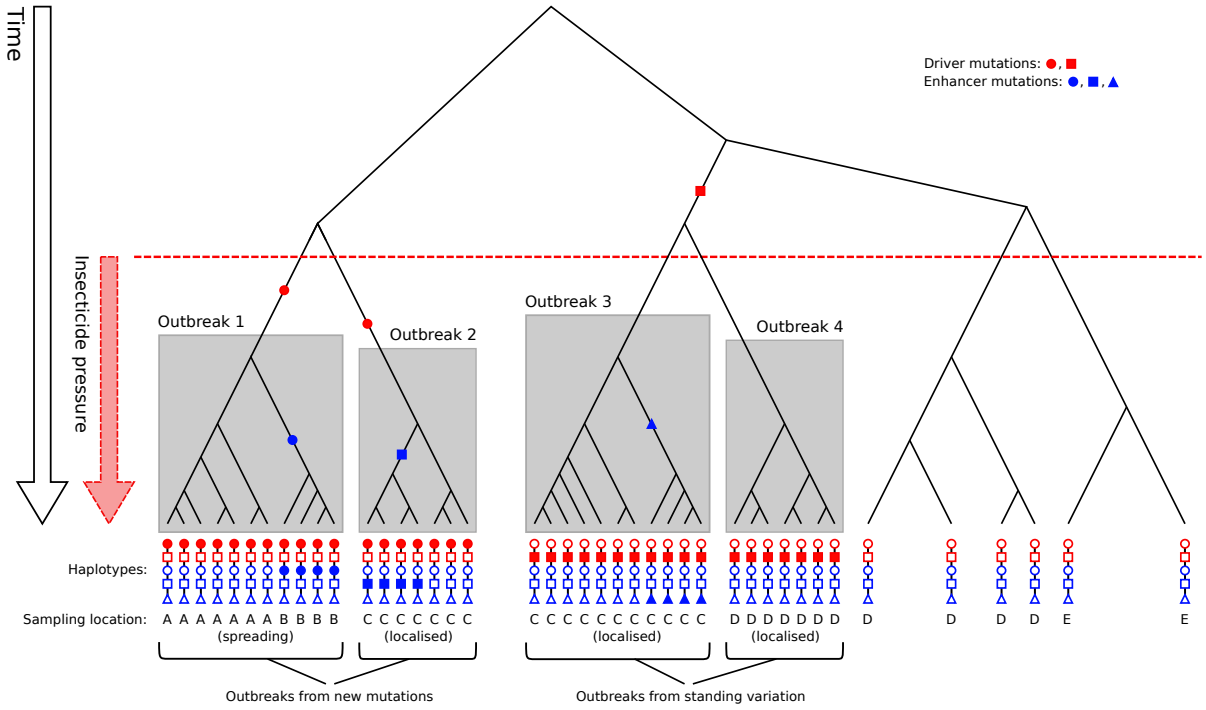

Figure 3. Illustration of insecticide resistance outbreaks. @@TODO explanation.

consistent over broad areas (and hence can support a spreading outbreak). We would also like to know which mosquito populations are sufficiently connected to enable outbreak spread, and if there is any consistent pattern to the direction of spread. This information could be relevant to discussions about how resources for insecticide resistance management might be targeted, what strategies are appropriate in which settings, and where and when insecticide resistance management needs to be coordinated between different countries and/or at different levels of administration.

For clarity, we also define the concept of an insecticide resistance outbreak formally in terms of coalescent theory, as a collection of lineages (1) sharing a resistance driver allele by descent, (2) coalescing more recently than the onset of insecticide pressure, and (3) having increased in frequency because of positive selection due to insecticides. This definition is illustrated for four hypothetical outbreaks in Figure 3. Because mosquitoes are sexually recombining, genealogical trees vary along the genome, and so we define resistance outbreaks with respect to a specific gene locus, which for the present study is codon 995 within the *Vgsc* gene. Note that separate outbreaks may be driven by the same resistance allele, and this can occur if multiple mutational events occur after the introduction of insecticides (Figure 3, outbreaks 1 and 2), or if a resistance allele is present in mosquito populations as standing variation prior to insecticide use (Figure 3, outbreaks 3 and 4). Here we are primarily concerned with whether outbreaks are localised or spreading, because this has immediate epidemiological relevance. We do not attempt to infer whether separate outbreaks with the same driver allele arose via standing variation or new mutations, however this is an interesting biological question to address in future studies. As a technical note, there is a simple correspondance with terminology conventionally used in the population genetics literature to describe selective sweeps. At a given gene locus, a hard selective sweep gives rise to a single resistance outbreak, and a soft selective sweep gives rise to multiple resistance outbreaks.

Outbreak analysis from haplotype age

As described above, haplotype data from genome regions both within and flanking the *Vgsc* gene provide a higher resolution for reconstructing recent historical events. To leverage this information, we used a heuristic approach to estimate the time to most recent

common ancestor (TMRCA) or "age" for each pair of haplotypes in our dataset, centering the analysis on *Vgsc* codon 995. For each pair of haplotypes, we estimated the length of the region shared identical by descent (IBD), and the number of mutations that have accumulated since the most recent common ancestor. We then combined these two pieces of information to produce a point estimate for the haplotype age (Methods). We studied the overall distribution of pairwise haplotype ages (Figure 4), and used hierarchical clustering to construct a dendrogram and visualise the overall age structure (Figure 5). We caution that although the estimated ages are in units of generations, these estimates have not been calibrated, and there is substantial uncertainty regarding both the mutation and recombination rate parameters. The ages therefore should not be interpreted as reliable absolute values, but they can be compared to each other to investigate the relative age of different events.

A key feature of the overall age distribution is that it is bimodal, with a minor mode of haplotypes coalescing recently, and a major mode coalescing further in the past (Figure 4). This is expected at an insecticide resistance locus experiencing one or more resistance

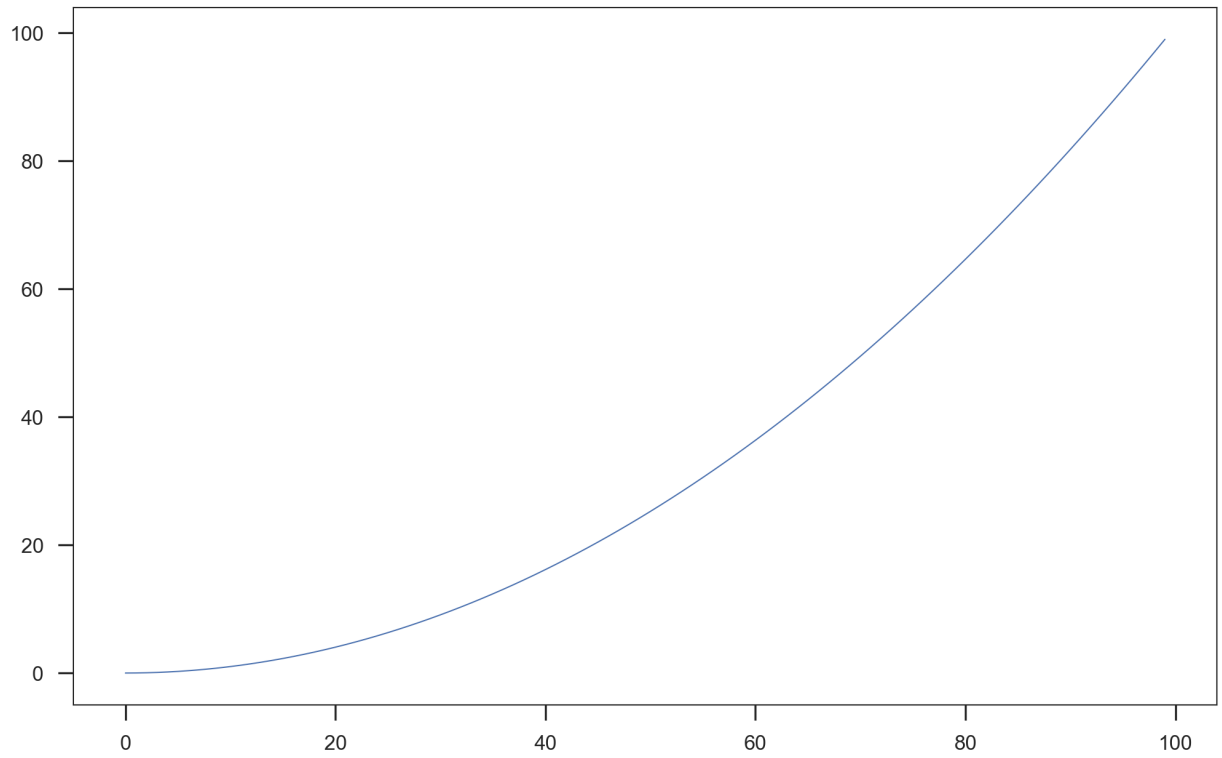

Figure 4. Haplotype age distribution. @@TODO real figure.

outbreaks. Within each outbreak, all haplotypes share a very recent common ancestor, but between outbreaks and among haplotypes without any resistance allele, haplotypes are more distantly related, and the distribution of ages is influenced by mosquito population size and other demographic factors. In particular, mosquito populations generally have a large effective population size (@@REF Ag1000G), and so in the absence of selection, haplotypes are expected to coalesce slowly. The bimodal age distribution is not due to geographical population structure, because the same bimodality is observed within several populations. We take the midpoint between these two modes as an estimate for the earliest time of onset of selective pressure due to insecticides, and thus for the maximum age of a resistance outbreak. To identify haplotype clusters representing putative resistance outbreaks, we then cut the haplotype dendrogram at this maximum outbreak age (Figure 5). Comparing this to previous analyses of haplotype structure based on genetic distance, we find clusters F1-F5 and S1-S3 recapitulated with close correspondence, and S4 and S5 merged into a single cluster. We label a new cluster "L@@" representing an outbreak driven by the `I1527T` allele in combination with one or the other `V402L` allele. We also label a cluster "L@@" capturing a set of haplotypes from Kenya carrying the `M490I` variant, although the fact that these haplotypes all share a recent common ancestor may be a reflection of the unusual demography of the Kenyan population which has experienced

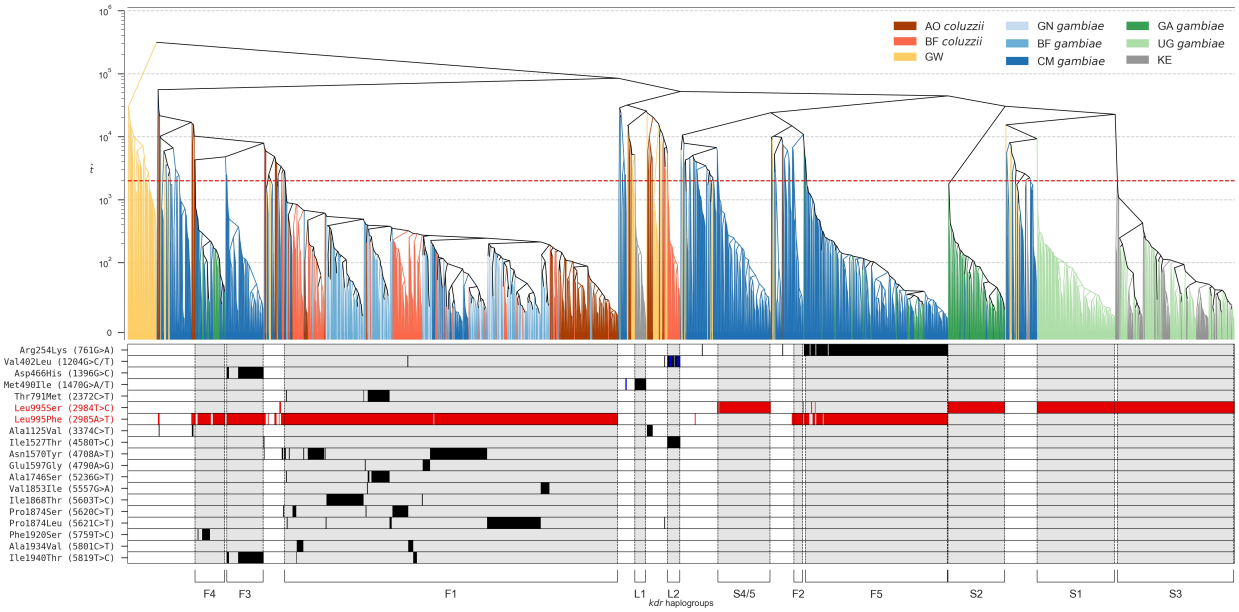

Figure 5. Clustering of haplotypes by age. @@TODO bigger font. @@TODO change "kdr haplogroups" to something else. @@TODO yticks to show number of haplotypes.

a severe population crash (@@REF) and not be due to recent selection for insecticide resistance. As in earlier analyses, clusters F1, F4, F5 and S3 all include haplotypes sampled from multiple geographical locations, and thus represent spreading outbreaks. Clusters F2, F3, S1, S2, S4/5 and L1 include only haplotypes from a single sampling location, and thus appear to represent localised outbreaks.

We then studied the distribution of haplotype ages within each spreading outbreak, to attempt to reconstruct information about the historical path of transmission of resistance alleles between locations. To do this, we grouped the haplotypes within each spreading outbreak by sampling location, and compared the distribution of haplotype ages both within and between locations. To aid in interpreting these data, we define three possible spreading scenarios, being: (1) a directional spread from one population to another; (2) spread from an unsampled population into the sampled populations; and (3) a complex scenario involving multiple gene flow events. In Figure 6 we illustrate the expected genealogy and haplotype age distribution under each of these scenarios.

The clearest result was obtained for outbreak F1 (Figure 7). Within this outbreak, haplotypes from Cameroon and Angola are significantly younger than haplotypes from Burkina Faso and Guinea. The age distributions are consistent with an outbreak originating in West Africa and subsequently spreading towards Cameroon and separately towards

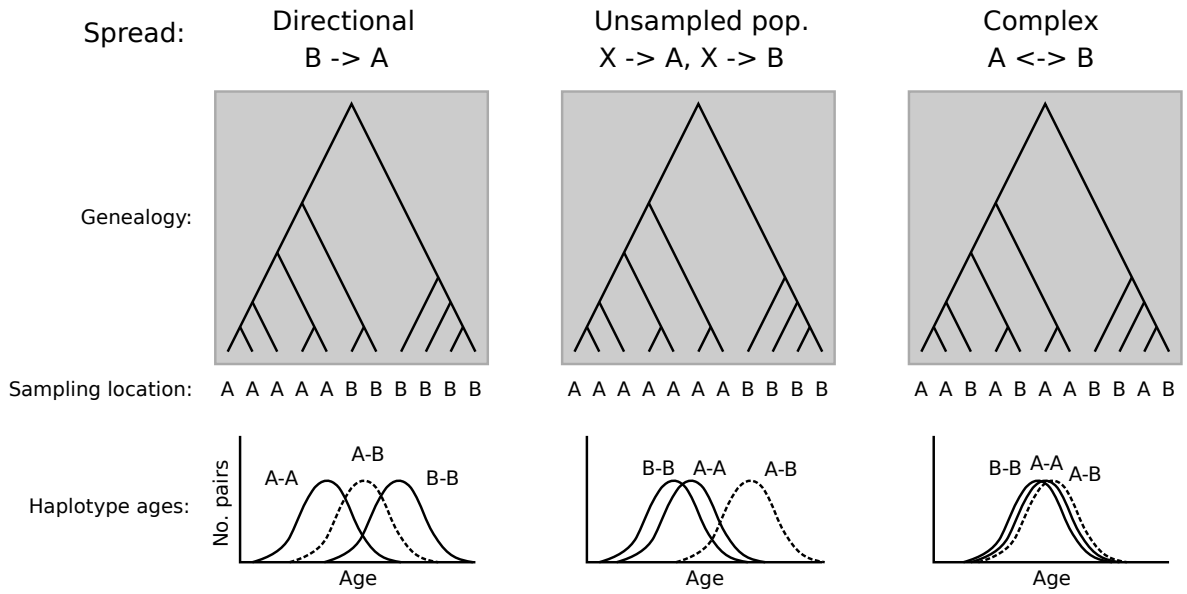

Figure 6. Inferring history of spread from haplotype ages. @@TODO explain.

Angola. We were surprised that the age distributions for *An. gambiae* and *An. coluzzii* from Burkina Faso are very similar, despite the fact that previous studies have shown that introgression has occurred from *An. gambiae* into *An. coluzzii*. This may indicate that the initial introgression event happened during the early phases of the outbreak, but is also consistent with a complex history of multiple gene flow events between the species.

Outbreaks F4, F5 and S2 each involve haplotypes from both Cameroon and Gabon. Interpreting the age distributions for these outbreaks is difficult, because mosquitoes from Gabon were collected at a much earlier time point (2000) than mosquitoes from Cameroon (20@@). If our haplotype age estimates were well-calibrated, and we also had reliable estimates for the number of mosquito generations per year, then we might be able to adjust for this time difference, however we are not able to do so presently. An interesting feature of these outbreaks, however, is that we would expect haplotypes from Gabon to appear older due to the time of sampling, which is observed for outbreak S2 but not for F4 or F5. Indeed, S2 is at a high frequency among all Gabon haplotypes and a low frequency among Cameroon haplotypes, whereas the reverse is true for F4 and F5. These data suggest that F4 and F5 have spread from Cameroon towards Gabon, while S2 has spread in the opposite direction. A lot can happen in mosquito populations in @@N years, however, and these conclusions remain highly speculative pending further sampling from both locations.

For outbreak S3 involving haplotypes from Uganda and Kenya, the age distributions

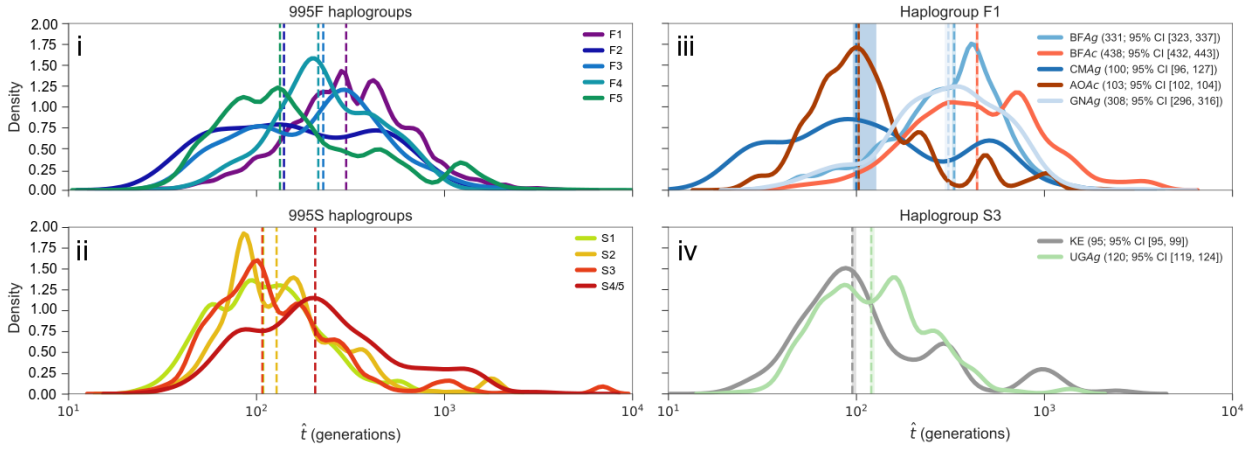

Figure 7. Haplotype age distributions. @@TODO rethink what goes in here, also if this needs to be here or can go to supplementary.

do not suggest any clear direction of gene flow. This could reflect multiple gene flow events in either or both directions. However, another outbreak (S1) is localised in Uganda and represented within the Ugandan population at roughly equal frequency with S3. If transmission was occurring from Uganda towards Kenya, we might expect both outbreaks to have spread to Kenya. Thus the localisation of S1 suggests S3 has spread into Uganda from Kenya or another location. Again, this conclusion remains tentative and requires confirmation via further sampling.

To summarise these conclusions in a concise way, we have depicted the distribution and spread of resistance outbreaks via the map shown in Figure 8. We have plotted haplotypes from each sampling location as a pie chart. The overall size of each pie chart represents the number of haplotypes sampled, and coloured wedges within each pie represent the frequency of each resistance outbreak within the population. Coloured arrows are used to depict our inferences regarding the transmission paths for spreading outbreaks. Our conclusions regarding direction of spread for outbreaks F4, F5, S2 and S3 are tentative, and we indicate this with a question mark. Because of the relatively sparse geographical representation within the Ag1000G phase 1 dataset, and the fact that collections were

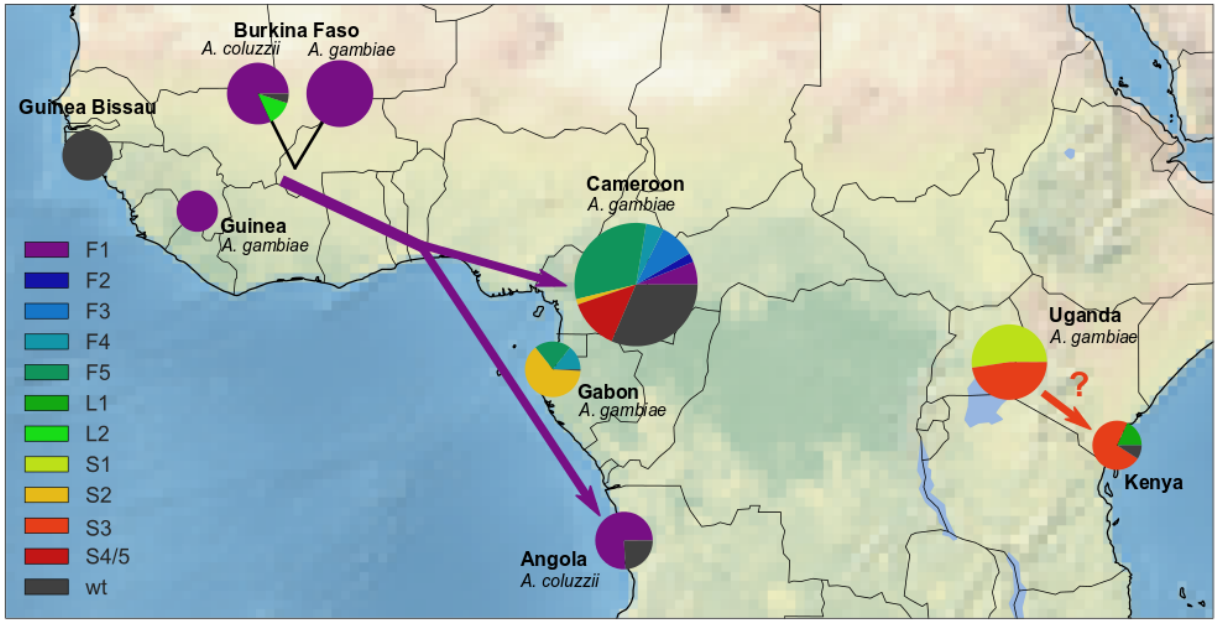

Figure 8. Geographical distribution of resistance outbreaks. @@TODO arrows for Gabon <-> Cameroon. @@TODO change arrow for Kenya -> Uganda. @@TODO add source area for F1.

not synchronized but span several years, we cannot be precise about the geographical origins of these resistance outbreaks. Even for outbreak F1 where we have clear evidence of spread from West Africa towards Central and Southern Africa, we have only sampled mosquitoes from Guinea and Burkina Faso, and the true source of the outbreak may not be either of these countries. We indicate this uncertainty regarding the outbreak source as a coloured area with a dashed border. This representation is imperfect, as is our knowledge regarding the sources and transmission paths of these outbreaks, but we hope this depiction may at least serve to stimulate further sampling, analysis and discussion, with the aim of improving our knowledge of resistance outbreaks for *Vgsc* as well as other insecticide resistance genes.

Design of genetic assays for outbreak surveillance

The insecticide resistance outbreaks we have identified here are undoubtedly ongoing, affecting many more mosquito populations than we have sampled in Ag1000G phase 1, and continuing to spread. In addition, other outbreaks may be occurring in populations that we have not sampled, or in populations we have sampled but since the sampling date. Whole genome sequencing of individual mosquitoes clearly provides data of sufficient resolution to detect resistance outbreaks and provide ongoing outbreak surveillance. The cost of whole genome sequencing continues to fall, with the present cost being approximately 100 GBP to obtain 30X coverage of an individual genome. Mobile sequencing technology is also developing rapidly, and may be a realistic prospect for mosquito population surveillance within a few years. There is an interim period, however, during which it may be more practical to develop targeted genetic assays for outbreak surveillance that could scale to tens of thousands of mosquitoes at a fraction of the cost of whole genome sequencing. For example, SNP genotyping using mass spectrometry and amplicon sequencing are two available technologies that could be applied now at scale and at modest cost.

To facilitate the development of targeted genetic assays for *Vgsc* insecticide resistance outbreak surveillance, we have produced two supplementary data tables. In Supplementary Table 1 we provide a list of all SNPs discovered in this study within the *Vgsc* gene and in the @@20 kbp upstream and downstream flanking intergenic regions. Both amplicon sequencing and genotyping by mass spectrometry require the design of PCR primers to

amplify the targeted genome region. To aid in primer design, for each SNP we provide the flanking sequence for @@200 bp upstream and downstream of the SNP position, including information about any polymorphisms within these flanking regions. Not all SNPs are informative for detecting whether an individual mosquito carries a haplotype from a resistance outbreak, and we provide some summary statistics for each SNP to aid in the selection of the most informative SNPs. For each SNP we report the allele frequencies within each of the outbreaks identified here, as well as for populations of susceptible haplotypes. We also provide the overall variance in allele frequencies, the information gain, and the Gini impurity for each SNP. Note that recombination events are more likely at increasing distances upstream and downstream of the resistance variants under selection, and thus the most informative SNPs are found closest to the resistance variants within the gene (@@REF Figure @@). However, SNPs with some information gain are available throughout the gene and in flanking regions.

We suggest that the design of a genetic assay proceed by (1) performing an initial round of filtering to remove SNPs which are not informative (e.g., low information gain); (2) performing a round of primer design to remove SNPs for which primers are unlikely to

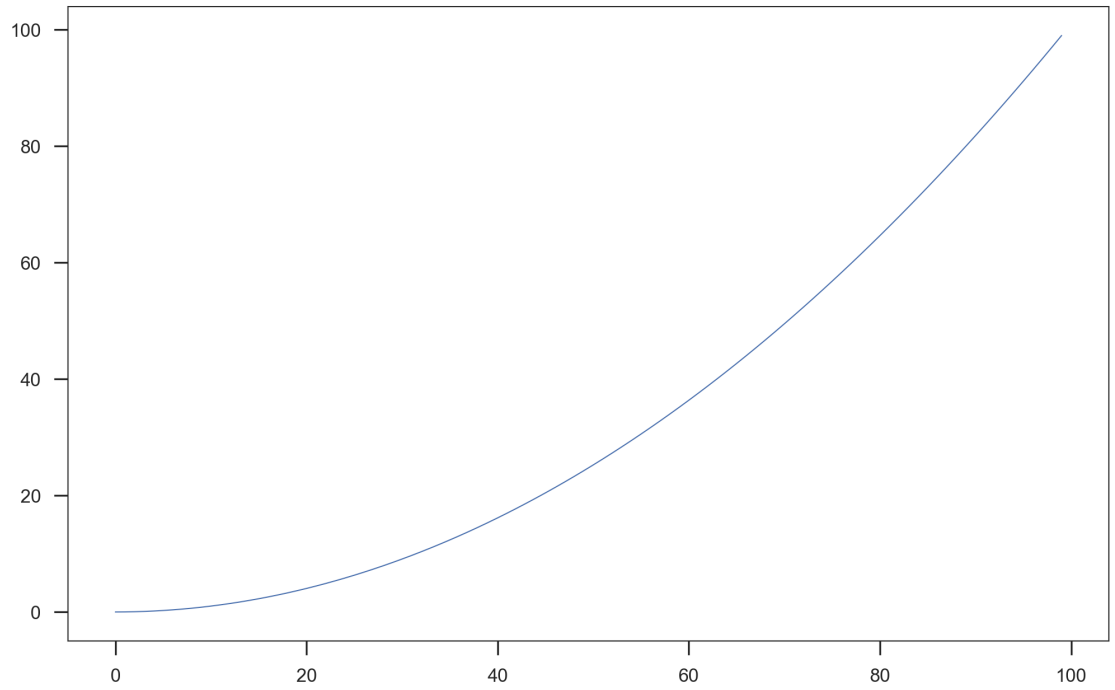

Figure 9. Information gain. @@TODO make the figure. @@TODO also include cross-validation scores for trees versus number of features, or separate figure?

be successful; (3) performing a full analysis of the remaining SNPs to select a subset that is sufficient to classify all outbreaks identified here; (4) finalise primer designs for the chosen panel of SNPs. A possible methodology for step 3 would be to use an algorithm such as ID3 to build a decision tree. To aid in the development of a classification algorithm, in Supplementary Table 2 we provide our classification for each of the 1530 haplotypes sampled here, along with the alleles carried by each haplotype for each of the SNPs included in Supplementary Table 1. To test the methodology, we constructed decision trees using the ID3 algorithm and using all available SNPs as input features (i.e., assuming primers could be designed in all cases). In Figure @@REF we show the cross-validation scores obtained for trees constructed at increasing numbers of SNPs. This analysis suggests that it should be possible to construct a reasonable classifier using @@N SNPs or less.

Recombination

As mentioned earlier, analyses of haplotype structure based on genetic distance within the fixed window of the *Vgsc* gene could be affected if recombination events occurred within the gene. Our analyses of haplotype age should be less affected by recombination, because they explicitly take recombination into account, estimating the positions at which recombination events have occurred to interrupt regions shared IBD between pairs of haplotypes. However, these analyses were based on a heuristic method for estimating recombination breakpoints, and there are several potential sources of error. To study the evidence for recombination within the genome region spanning the *Vgsc* gene, and provide some additional confirmation that our inferences regarding insecticide resistance outbreaks have not been affected by recombination or other sources of error, we performed an additional analysis of genetic distance between haplotypes. We first constructed a putative ancestral haplotype for each of the outbreaks we identified, by starting from the codon 995 position and separately moving upstream and downstream, assuming the major allele at each SNP bifurcation point represents the ancestral haplotype. We then computed the genetic distance (D_{XY}) between each of our sampled haplotypes and each of the inferred ancestral outbreak haplotypes, computing the distance in @@ overlapping windows of @@ bp across a 2 Mbp region spanning the *Vgsc* gene. The results for outbreaks F1-F5 are plotted in Figure 10, and outbreaks S1-S4/5 are shown in Figure 11. In these

plots we expect that all haplotypes from a given outbreak should share very close genetic similarity ($D_{XY} \approx 0$) with each other and with the ancestral haplotype for that outbreak within the *Vgsc* gene itself, with an increasing number of haplotypes recombining away from the ancestral outbreak haplotype as we move away from the gene in either the upstream or downstream direction. Conversely, haplotypes from one outbreak should not share any close genetic similarity ($D_{XY} > 0$) with the inferred ancestral haplotype from a different outbreak, either within the *Vgsc* gene or in flanking regions.

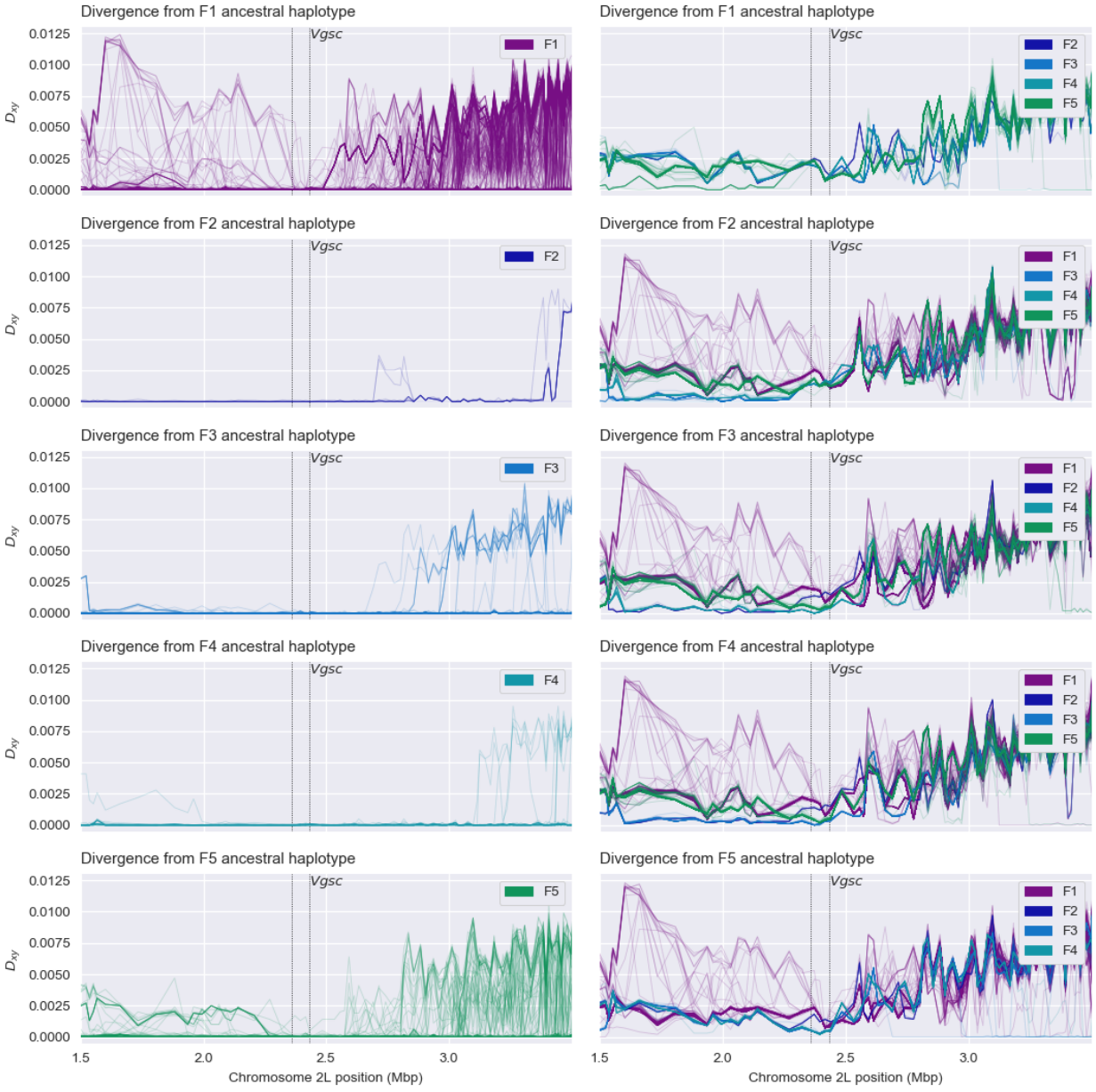

Figure 10. Recombination and ancestral haplotypes for L995F. @@TODO legend

The results for all outbreaks are largely consistent with this expectation. For this analysis we treated S4/5 as a single outbreak, as indicated by the haplotype age analysis, and we can gain some insight into why these two were split into separate clusters in earlier analyses. All haplotypes in the S4/5 outbreak share close similarity with the ancestral haplotype on both flanks of the *Vgsc* gene, but there is a short region of within the gene where a subset of haplotypes are diverged. This region of divergence accounts for the S4/S5 split in earlier analyses. @@TODO explain @@TODO also note relatively low divergence among F2, F3, F4 on upstream flank and explain

Discussion

@@TODO Discuss accessibility, have we missed any functional variation?

@@TODO Discuss weaknesses, caveats and potential improvements to method for estimating haplotype age.

@@TODO What are the implications for insecticide resistance management? Realistically how could this information be used?

@@TODO What about DDT? If prior selection for DDT resistance, how might this complicate the picture? Do we see any evidence for multiple phases of selection?

@@TODO Speculate on why `L995F` but not `L995S` has evolved secondary variation.

Legacy

@@TODO describe how we put together the map (Figure 8).

Recombination and independent outbreaks of resistance

@@TODO

Discussion

@@TODO

Methods

@@TODO

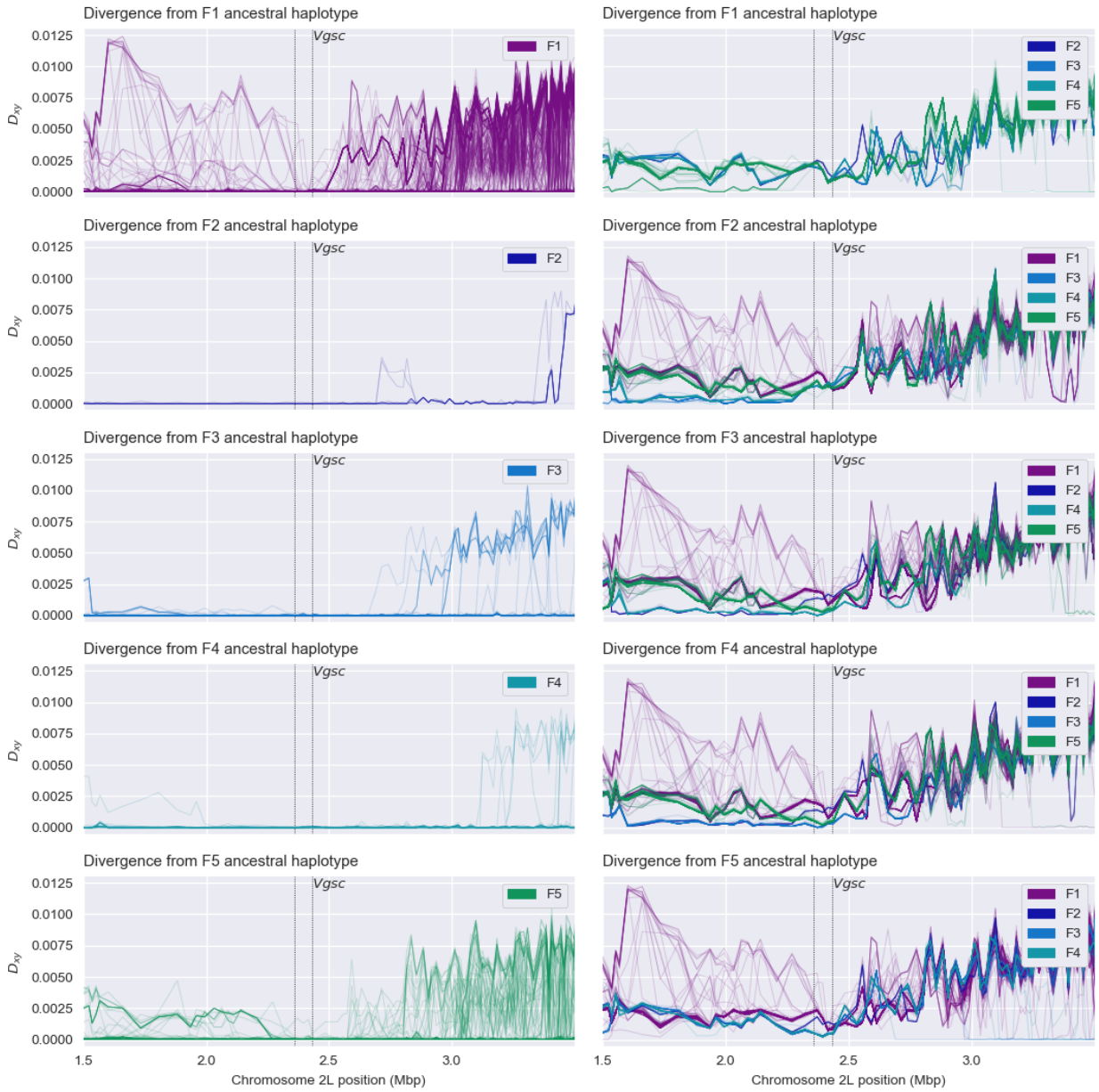

Figure 11. Recombination and ancestral haplotypes for L995S. @@TODO legend

References

[1] Martin S Williamson et al. 'Identification of mutations in the houseflypara-type sodium channel gene associated with knockdown resistance (kdr) to pyrethroid insecticides'. In: *Molecular and General Genetics MGG* 252.1 (1996), pp. 51–60.