The genetic architecture 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 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 assign a putative resistance phenotype, based on genetic evidence for recent selection, patterns of linkage between variants, and the position of the variant within the protein structure. 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 predict resistance phenotype 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 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*

[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.

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

Identification of resistance alleles

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, with 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

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)
`2,431,061 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.

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`).

from Central and East Africa (Table 1; @@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 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 @@LD), 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 trans-membrane 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

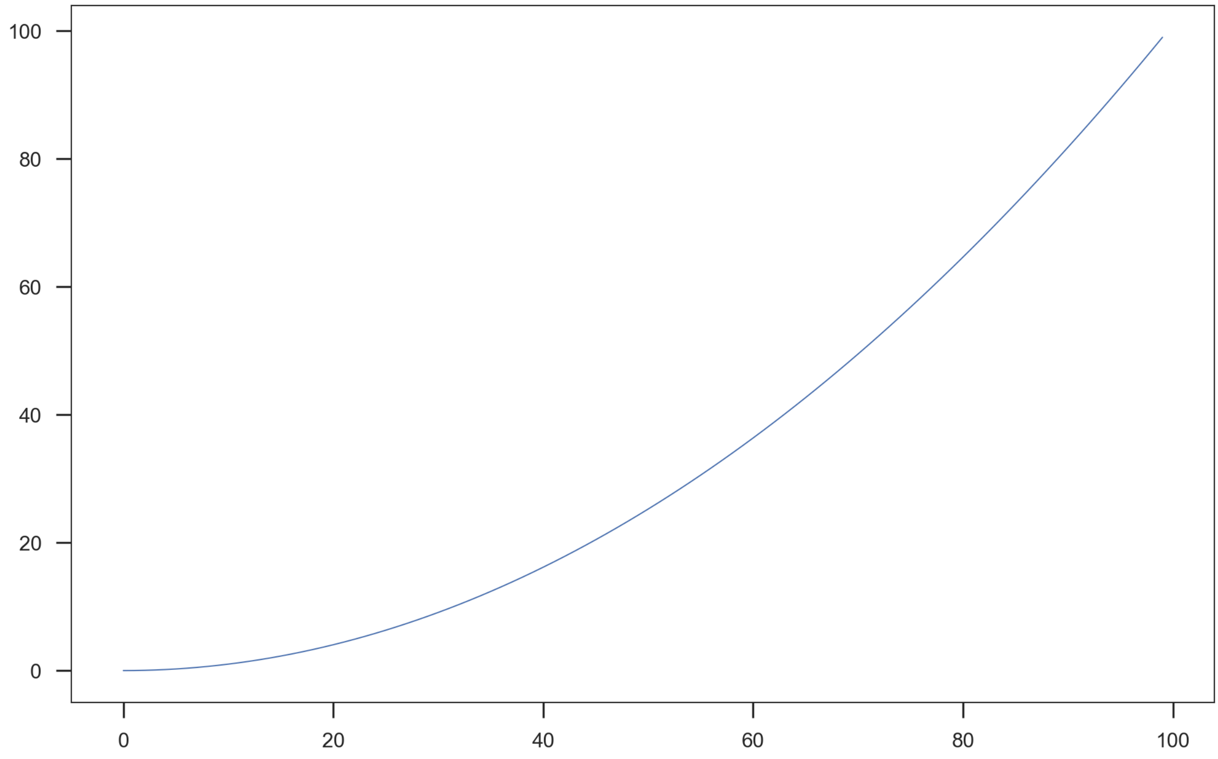

Figure 1. Placeholder for LD figure.

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 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.

Origins and spread of resistance alleles

Although it is well-known that pyrethroid resistance is becoming increasingly prevalent in malaria vector populations across Africa, it has not previously been clear whether this increase in prevalence is being driven primarily by the spread of resistance alleles from one location to another via mosquito migration, or by resistance alleles emerging independently and simultaneously in multiple locations, or by a combination of both processes. In our initial analyses of haplotype data from Ag1000G phase 1 (@@REF Ag1000G), we used a clustering approach based on genetic distance to identify 10 haplotype clusters at the *Vgsc* locus carrying a known resistance driver allele, of which five clusters carried L995F (labelled F1-F5) and a further five clusters carried L995S (labelled S1-S5). Within each cluster, haplotypes were nearly identical across the entire @@70 kbp span of the *Vgsc* gene, and therefore represent a collection of haplotypes with a very recent common ancestor. Within some of these clusters, we found haplotypes from mosquitoes collected from different geographical locations, demonstrating that adaptive gene flow has occurred between these locations. Specifically, cluster F1 contained haplotypes from Guinea, Burkina Faso, Cameroon and Angola; clusters @@ each contained haplotypes from both Cameroon and Gabon; and cluster @@ contained haplotypes from both Uganda and Kenya. The F1 cluster also contained haplotypes from both *An. gambiae* and *An. coluzzii*, demonstrating that adaptive introgression had occurred. We also used analyses of haplotype clustering on the flanks of the *Vgsc* gene to provide evidence that each of these haplotype clusters represents an independent outbreak of resistance, with the exception of clusters S4 and S5 which appear to be recently derived from the same ancestor. In this section we present several new analyses to confirm and extend our initial findings regarding the origins and spread of *Vgsc* resistance alleles.

To provide an alternative view of the genetic similarity between haplotypes carrying known or predicted resistance driver alleles, we used haplotype data across all 1710 (@@CHECK) SNPs within the *Vgsc* open reading frame to construct median-joining networks (Figure 2). We constructed these networks up to a maximum distance of @@4 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 L995F, the resulting network confirms the presence of five distinct clusters of closely-related haplotypes, with close correspondance to the clusters F1-F5 identified previously (@@REF Ag1000G). The haplotype network for cluster F1 brings into sharp relief the explosive evolution of secondary amino acid substitutions within the F1 lineage, providing further evidence that they play a functional role in enhancing the L995F resistance phenotype. The distribution of secondary variants within the F1 network also allows us to infer multiple introgression events between the two species. @@TODO details of which variants are found in both species.

The L995S network also confirms five distinct clusters, in concordance with our previous analysis. The contrast between the haplotype networks for the L995F and L995S alleles is striking, because of the near-total absence of any non-synonymous variants within the L995S networks. As we reported previously, there is a highly significant enrichment for non-synonymous variation among haplotypes carrying L995F relative to haplotypes with L995S. This difference is the basis for our prediction that all of the secondary non-

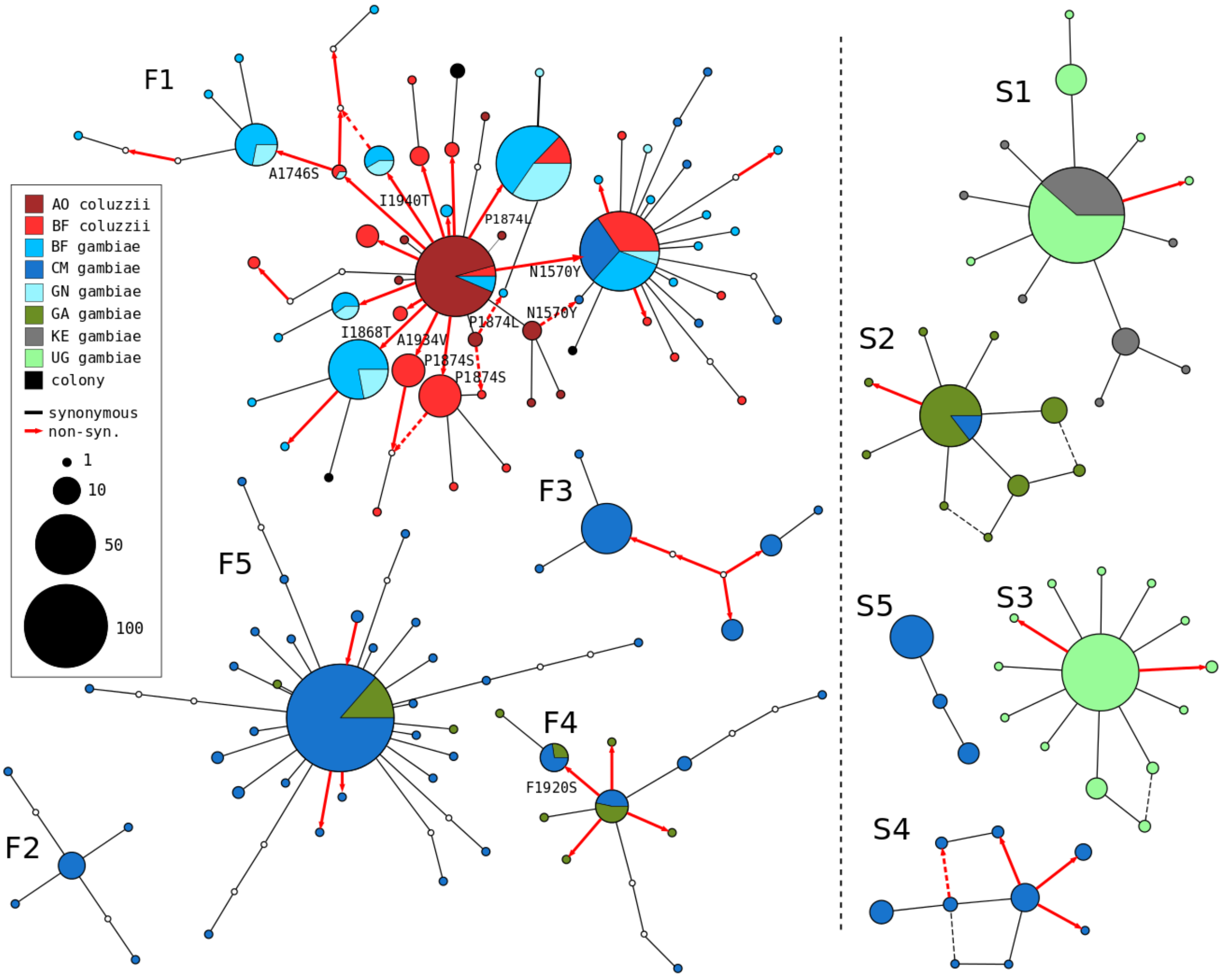

Figure 2. Haplotype networks. @@TODO redo the figure. @@TODO annotate non-syn edges in cluster F3. @@TODO mention if any clusters fixed for non-syn variants so not shown. @@TODO annotate other non-syn edges, e.g., in S4?

synonymous variants found at appreciable frequency among L995F haplotypes are likely to enhance the resistance phenotype, rather than being deleterious or neutral variants that are hitch-hiking on selective sweeps for the driver allele, because hitch-hiking should be similar for both L995F and L995S.

A limitation of both the hierarchical clustering and network analyses is that they only leverage information from within the *Vgsc* gene. Because they are relying on estimates of genetic distance from within a relatively small genome region, they have limited resolution to infer very recent events, because insufficient time has elapsed for mutations to occur. This means that, within each of the lineages we have identified where adaptive gene flow has occurred, the analyses provide little information about the direction or relative timing of gene flow events. To improve our resolution to infer recent events, we used information about the extent of haplotype sharing on both flanks of the gene. For each pair of haplotypes, we used a heuristic approach to estimate the length of the region shared identical by descent (IBD) between two haplotypes, extending both upstream and downstream of the gene, and to estimate the number of mutations that have accumulated within IBD regions since the most recent common ancestor (MRCA). We then combined the information about IBD length and number of mutations to estimate the time to MRCA or "age" for each pair of haplotypes. We used hierarchical clustering to visualise the overall age structure (Figure 3), and also analysed the distribution of ages within and between different collections of haplotypes (Figure @@REF). We caution that these analyses are relatively crude, and could be improved in a number of ways. However, estimating haplotype age and local genealogy is a very active research area, and we use a heuristic approach here to provide an initial view into the data. We also caution that although the estimated ages are in units of generations, these estimates have not been calibrated, and so cannot be taken as accurate absolute values. They can, however, be compared with each other, to explore the relative age of different events.

A key feature of the overall distribution of haplotype ages is that it is bimodal, with a minor mode of haplotypes coalescing recently, and a major mode coalescing further in the

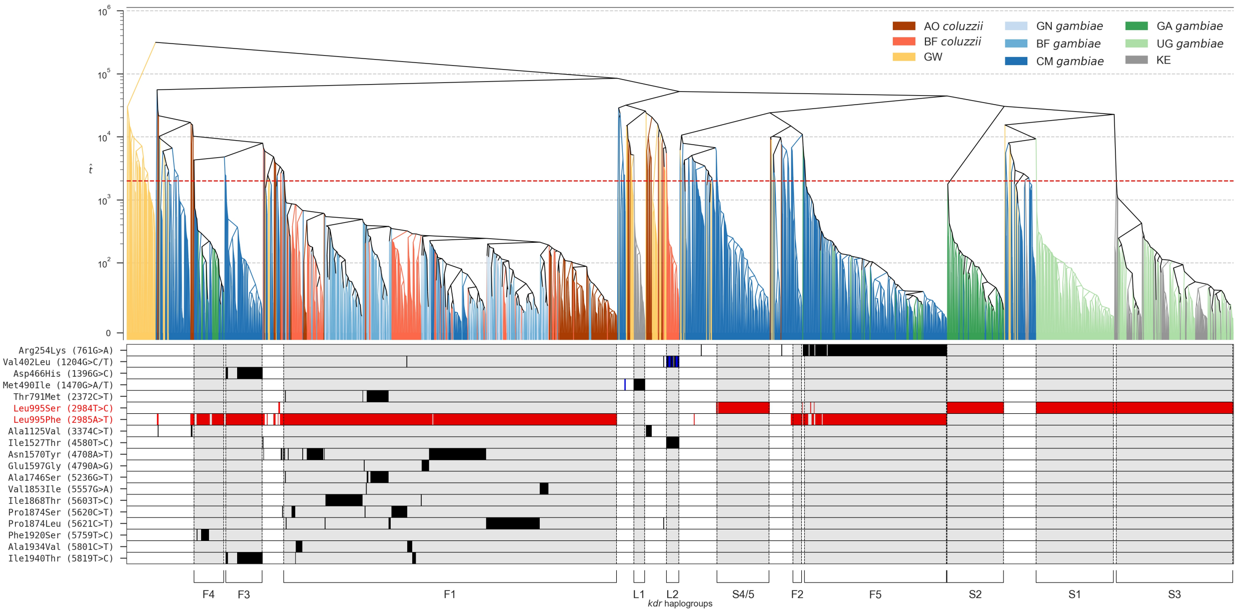

Figure 3. Clustering of haplotypes by age. @@TODO bigger font. @@TODO change "kdr haplogroups" to something else.

past (Figure @@REF). This is expected at a locus experiencing recent positive selection and multiple selective sweeps. Within each sweep, all haplotypes share a very recent common ancestor, but between sweeps and among haplotypes without any resistance allele, genealogies reflect a more even distribution of ancestry. This overall bimodal distribution is also not simply a reflection of geographical population structure, because the same bimodality is observed within several populations (Figure @@REF). We take the midpoint between these two modes as an estimate for the maximum age of selective sweeps at this locus. When we then cut the haplotype tree at this age, and take the 11 largest clades, we find that the resulting clades are highly concordant with the clustering and network analyses based on genetic distance within the gene (Figure 3). Clusters F1-F5 are recapitulated, as are clusters S1-S4, with S4 and S5 merged into a single cluster reflecting their shared ancestry as discovered from previous analyses. We also label a new cluster "L@@" representing the lineage of haplotypes carrying the `I1527T` allele in combination with one or the other `V402L` allele. @@TODO what to say about the other "L@@" cluster?

Using these estimates for haplotype age, we draw some tentative conclusions regarding the history of these resistance outbreaks. @@TODO describe evidence for direction of spread within spreading haplogroups.

@@TODO talk about reconstructing ancestral haplotypes and studying recombination. Does this need a new sub-section? If so, what's the title?

@@TODO other part of recombination figure – both go to supplementary?

Discussion

@@TODO

Methods

@@TODO

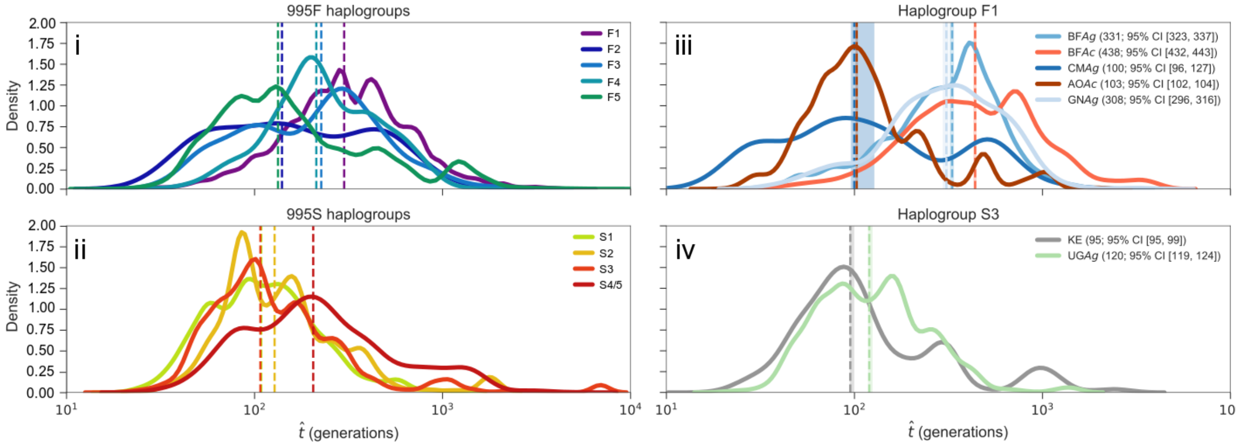

Figure 4. Haplotype age distributions. @@TODO rethink what goes in here, also if this needs to be here or can go to supplementary.

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.

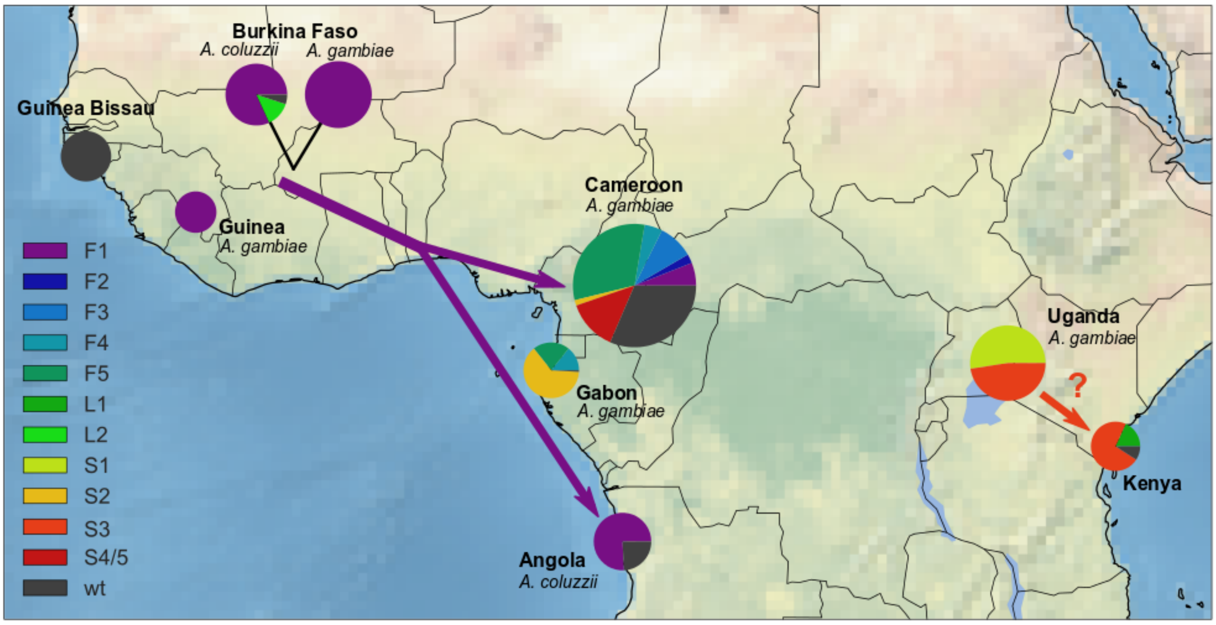

Figure 5. Geographical distribution of resistance haplogroups.

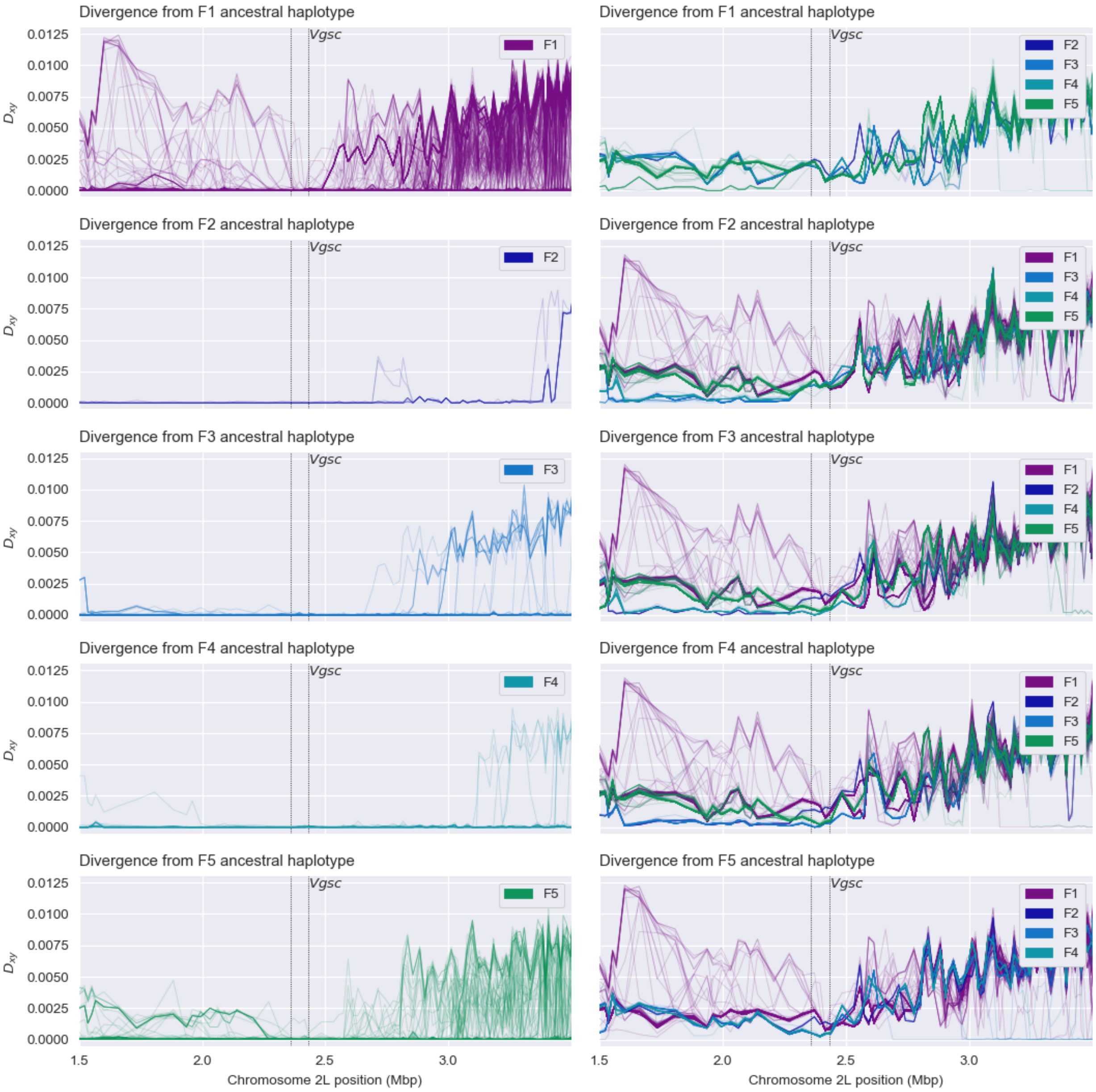

Figure 6. Recombination. @@TODO legend