

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. Harding², Andrias O. O'Reilly³, David Weetman⁴, Dominic Kwiatkowski^{1,2}, Martin Donnelly^{4,1}, and The *Anopheles gambiae* 1000 Genomes Consortium⁵

¹Wellcome Sanger Institute, Hinxton, Cambridge CB10 1SA

²Big Data Institute, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, Oxford OX3 7LF

³Liverpool John Moores University, Brownlow Hill, Liverpool L3 5UG

⁴Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA

⁵<https://www.malariagen.net/projects/ag1000g#people>

*These authors contributed equally

7th December 2020

Abstract

Resistance to pyrethroid insecticides is a major concern for malaria vector control because these compounds are used in almost all insecticide-treated bed-nets (ITNs), and are also used for indoor residual spraying (IRS). Pyrethroids target the voltage-gated sodium channel (VGSC), an essential component of the mosquito nervous system, but substitutions in the amino acid sequence can disrupt the activity of these insecticides, inducing a resistance phenotype. Here we use Illumina whole-genome sequence data from phase 2 of the *Anopheles gambiae* 1000 Genomes Project (Ag1000G) to provide a comprehensive account of genetic variation in the *Vgsc* gene in mosquito populations

from 13 African countries. In addition to three known resistance alleles, we describe 20 other non-synonymous nucleotide substitutions at appreciable population frequency, and map these variants onto a molecular model of the protein to investigate the likelihood of a pyrethroid resistance phenotype. Thirteen of these novel alleles were found to occur almost exclusively on haplotypes carrying the known L995F *kdr* resistance allele (L1014F in *Musca domestica* codon numbering) and may enhance or compensate for the L995F resistance phenotype. A novel mutation I1527T, adjacent to a predicted pyrethroid binding site, was found in tight linkage with V402L substitutions, similar to a combination of alleles found to cause pyrethroid resistance in several other insect species. We also analysed genetic backgrounds carrying resistance alleles, to determine which alleles have experienced recent positive selection, and to investigate the spread of resistance between species and geographical locations. We describe ten distinct haplotype groups carrying known *kdr* resistance alleles. Five of these groups are localised to a single geographical location, and five include haplotypes from different countries, in one case separated by over 3000 km, providing new information about the potential for the geographical spread of resistance. Markers are identified that could be used to design high-throughput, low-cost genetic assays for tracking the spread of pyrethroid resistance in the field. Our results demonstrate that the molecular basis of target-site pyrethroid resistance in malaria vectors is more complex than previously appreciated, and provide a foundation for the development of new genetic tools for insecticide resistance management.

Introduction

Pyrethroid insecticides have been the cornerstone of malaria prevention in Africa for almost two decades [1]. Pyrethroids are currently used in all insecticide-treated bed-nets (ITNs), and are used in indoor residual spraying (IRS) as well as in agriculture. Resistance to these insecticides is now widespread in malaria vector populations across Africa [2]. The World Health Organization (WHO) has published plans for insecticide resistance management (IRM) that emphasise the need for improvements in both our knowledge of the molecular mechanisms of resistance and our ability to monitor them in natural populations [3, 4].

The voltage-gated sodium channel (VGSC) is the physiological target of pyrethroid insecticides, and is integral to the insect nervous system. The sodium channel protein con-

sists of four homologous domains (DI-IV) each of which comprises six transmembrane segments (S1-S6) connected by intracellular and extracellular loops [5]. Pyrethroid molecules bind to this protein, stabilise the ion-conducting active state, and thus disrupt normal nervous system function, producing paralysis (“knock-down”) and death. However, amino acid substitutions at key positions within the protein alter the interaction with insecticide molecules, increasing the dose of insecticide required for knock-down, known as knock-down resistance or *kdr* [6, 5].

In the African malaria vectors *Anopheles gambiae* and *An. coluzzii*, three substitutions have been found to cause pyrethroid resistance. Two of these substitutions occur in codon 995¹, with L995F prevalent in West and Central Africa [7, 8], and L995S found in Central and East Africa [9, 8]. A third substitution, N1570Y, has been found in West and Central Africa and shown to increase resistance in association with L995F [11]. However, studies in other insect species have found a variety of other *Vgsc* substitutions inducing a resistance phenotype [12, 13, 5]. To our knowledge, no studies in malaria vectors have analysed genetic variation across the full *Vgsc* coding sequence, thus the molecular basis of pyrethroid target-site resistance has not been fully explored.

Basic information is also lacking about the spread of pyrethroid resistance in malaria vectors [3]. For example, it is not clear when, where or how many times pyrethroid target-site resistance has emerged. Geographical paths of transmission, carrying resistance alleles between mosquito populations, are also not known. Previous studies have found evidence that L995F occurs on several different genetic backgrounds, suggesting multiple independent outbreaks of resistance driven by this allele [14, 15, 16, 17]. However, these studies analysed only small gene regions in a limited number of mosquito populations, and therefore had limited resolution to make inferences about relationships between haplotypes carrying this allele. It has also been shown that the L995F allele spread from *An. gambiae* to *An. coluzzii* in West Africa [18, 19, 20, 21]. However, both L995F and L995S now have wide geographical distributions [8], and to our knowledge no attempts have been made to infer or track the geographical spread of either allele across Africa.

Here we report an in-depth analysis of genetic variation in the *Vgsc* gene, using whole-

¹Codon numbering is given here relative to transcript AGAP004707-RD as defined in the AgamP4.12 gene-set annotations. A mapping of codon numbers from AGAP004707-RD to *Musca domestica*, the system in which *kdr* mutations were first described [10], is given in Table 1.

85 genome Illumina sequence data from phase 2 of the *Anopheles gambiae* 1000 Genomes
 86 Project (Ag1000G) [22]. The Ag1000G phase 2 resource includes data on nucleotide vari-
 87 ation in 1,142 wild-caught mosquitoes sampled from 13 countries, with representation of
 88 West, Central, Southern and East Africa, and of both *An. gambiae* and *An. coluzzii*.
 89 We investigate variation across the complete gene coding sequence, and report popula-
 90 tion genetic data for both known and novel non-synonymous nucleotide substitutions. We
 91 then use haplotype data from the chromosomal region spanning the *Vgsc* gene to study
 92 the genetic backgrounds carrying resistance alleles, investigate the geographical spread
 93 of resistance between mosquito populations, and provide evidence for recent positive se-
 94 lection. Finally, we explore ways in which variation data from Ag1000G can be used to
 95 design high-throughput, low-cost genetic assays for surveillance of pyrethroid resistance,
 96 with the capability to differentiate and track resistance outbreaks.

97 **Results**

98 ***Vgsc* non-synonymous nucleotide variation**

99 To identify variants with a potentially functional role in pyrethroid resistance, we ex-
 100 tracted single nucleotide polymorphisms (SNPs) that alter the amino acid sequence of the
 101 VGSC protein from the Ag1000G phase 2 data resource [22]. We then computed their
 102 allele frequencies among 16 mosquito populations defined by species and country of ori-
 103 gin. Alleles that confer resistance are expected to increase in frequency under selective
 104 pressure, therefore we filtered the list of potentially functional variant alleles to retain
 105 only those at or above 5% frequency in one or more populations (Table 1). The resulting
 106 list comprises 23 variant alleles, including the known L995F, L995S and N1570Y resistance
 107 alleles, and a further 20 alleles which prior to Ag1000G had not previously been described
 108 in anopheline mosquitoes. We reported 12 of these novel alleles in our overall analysis of
 109 the 765 samples in the Ag1000G phase 1 data resource [23], and we extend the analyses
 110 here to incorporate SNPs which alter codon 531, 697, 1507, 1603 and two tri-allelic SNPs
 111 affecting codons 402 and 490.

112 The 23 non-synonymous variants were located on a transmembrane topology map and on
 113 a 3-dimensional homology model of the *Vgsc* protein. (Figure 1). The substitutions were

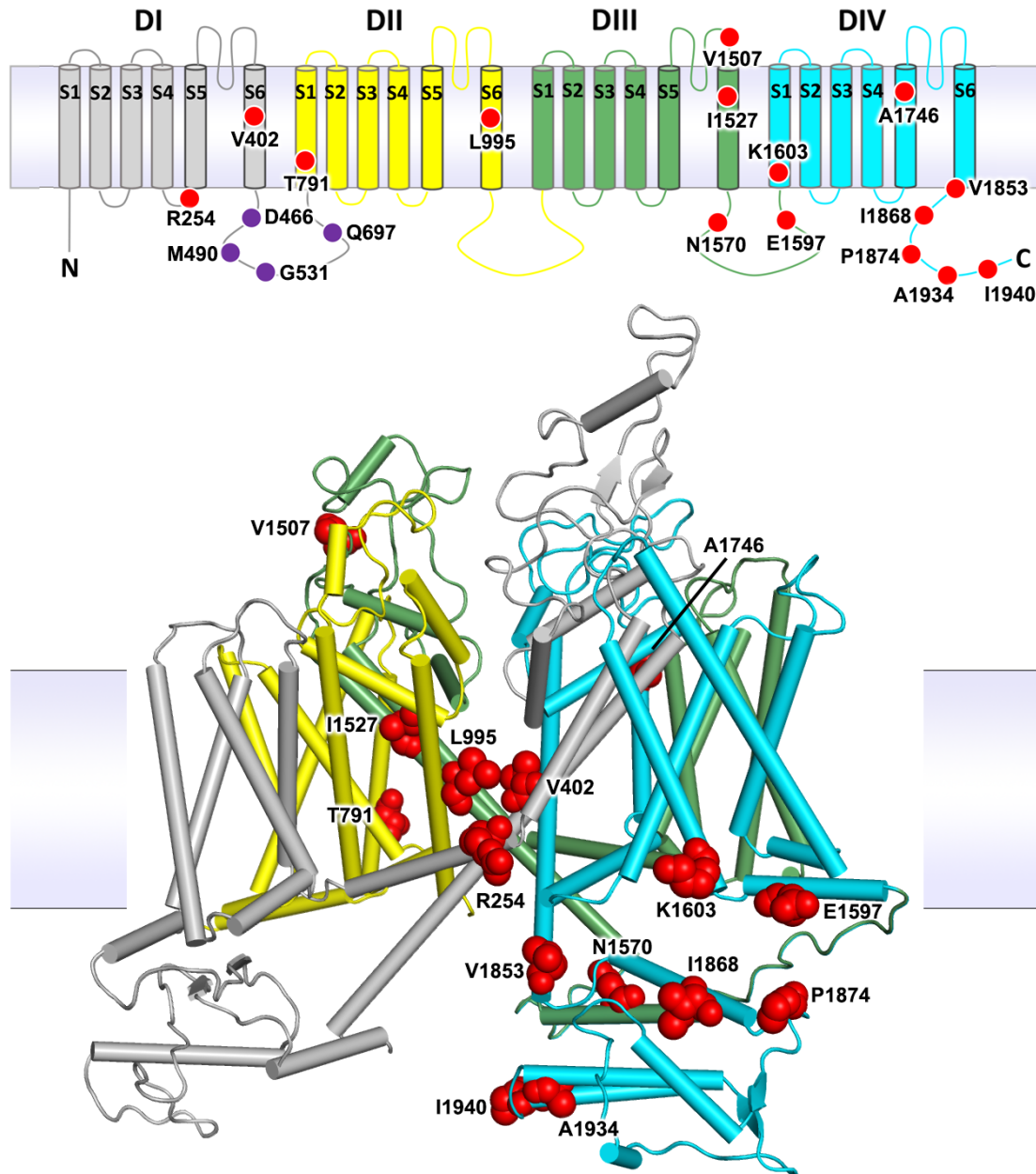


Figure 1. Voltage-gated sodium channel protein structure and non-synonymous variation. The *An. gambiae* voltage-gated sodium channel (AGAP004707-RD AgamP4.12) is shown as a transmembrane topology map (**top**) and as a homology model (**bottom**) in cartoon format coloured by domain. Variant positions are shown as red circles in the topology map and as red space-fill in the 3D model. Purple circles in the map show amino acids absent from the model due to the lack of modelled structure in this region.

114 found to be distributed throughout the channel, in all of the four internally homologous
 115 domains (DI-DIV), in S1, S5 and S6 membrane-spanning segments, in two of the intracel-
 116 lular loops connecting domains, and in the C-terminal tail. The S5 and S6 segments that
 117 form the central ion-conducting pore of the channel carry six of the eight segment substi-

tutions, including V402 and L995 which have been shown to produce insecticide resistance phenotypes [6, 5, 7, 8, 9]. Two substitutions are located on the DIII-DIV linker including the resistance conferring N1570 [11]. A further six substitutions are found concentrated in the protein’s carboxyl tail (C-terminus), including two alternative substitutions at the resistance associated P1874 residue [24]. The DIII-DIV linker and the C-terminus segment interact in the closed-state channel and substitutions are found throughout this intracellular subdomain. Finally, there are four novel substitutions located on the DI-DII intracellular linker, but this region is missing from the model as it was not resolved in the cockroach Na_vPaS structure used as the model template [25].

The two known resistance alleles affecting codon 995 had the highest overall allele frequencies within the Ag1000G phase 2 cohort (Table 1). The L995F allele was at high frequency in populations of both species from West, Central and Southern Africa. The L995S allele was at high frequency among *An. gambiae* populations from Central and East Africa. Both of these alleles were present in *An. gambiae* populations sampled from Cameroon and Gabon. This included individuals with a heterozygous L995F/S genotype (50/297 individuals in Cameroon, 41/69 in Gabon). We calculated empirical p-values for these heterozygous genotype counts using the Dirichlet distribution and 1,000,000 Monte Carlo simulations. In Cameroon p=0.410 of simulations found higher proportions of heterozygous genotypes, however in Gabon this dropped to p=0.005, suggesting there may be a fitness advantage for mosquitoes carrying both alleles in some circumstances.

The N1570Y allele was present in Guinea *An. gambiae*, Ghana *An. gambiae*, Burkina Faso (both species) and Cameroon *An. gambiae*. This allele has been shown to substantially increase pyrethroid resistance when it occurs in combination with L995F, both in association tests of phenotyped field samples [11] and functional tests using *Xenopus* oocytes [26]. To study the patterns of association among non-synonymous variants, we used haplotypes from the Ag1000G phase 2 resource to compute the normalised coefficient of linkage disequilibrium (D') between all pairs of variant alleles (Figure 2). As expected, we found N1570Y in almost perfect linkage with L995F. Of the 20 novel non-synonymous alleles, 13 also occurred almost exclusively in combination with L995F (Figure 2). These included two variants in codon 1874 (P1874S, P1874L), one of which (P1874S) has previously been associated with pyrethroid resistance in the crop pest moth *Plutella xylostella*

Table 1. Non-synonymous nucleotide variation in the voltage-gated sodium channel gene. AO=Angola; GH=Ghana; BF=Burkina Faso; CI=Côte d’Ivoire; GN=Guinea; GW=Guinea-Bissau; GM=Gambia; CM=Cameroon; GA=Gabon; UG=Uganda; GQ=Bioko; FR=Mayotte; KE=Kenya; *Ac=An. coluzzii*; *Ag=An. gambiae*. Species status of specimens from Guinea-Bissau, Gambia and Kenya is uncertain [22]. All variants are at 5% frequency or above in one or more of the 16 Ag1000G phase 2 populations, with the exception of 2,400,071 G>T which is only found in the CMAg population at 0.3% frequency but is included because another mutation is found at the same position (2,400,071 G>A) at >5% frequency and which causes the same amino acid substitution (M490I).

Variant				Population allele frequency (%)															
Position ¹	Ag ²	Md ³	Domain ⁴	AOAc	GHAc	BFAC	CIAC	GNAC	GW	GM	CMAg	GHAg	BFAG	GNAg	GAAG	UGAg	GQAg	FRAg	KE
2,390,177 G>A	R254K	R261	IL45	0.0	0.009	0.0	0.0	0.0	0.0	0.0	0.313	0.0	0.0	0.0	0.203	0.0	0.0	0.0	0.0
2,391,228 G>C	V402L	V410	IS6	0.0	0.127	0.073	0.085	0.125	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2,391,228 G>T	V402L	V410	IS6	0.0	0.045	0.06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2,399,997 G>C	D466H	-	LI/II	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.069	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2,400,071 G>A	M490I	M508	LI/II	0.0	0.0	0.0	0.0	0.0	0.0	0.031	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.188
2,400,071 G>T	M490I	M508	LI/II	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2,402,466 G>T	G531V	G549	LI/II	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.007	0.0	0.056	0.0	0.0
2,407,967 A>C	Q697P	Q724	LI/II	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.056	0.0	0.0
2,416,980 C>T	T791M	T810	IIS1	0.0	0.009	0.02	0.0	0.0	0.0	0.0	0.0	0.292	0.147	0.112	0.0	0.0	0.0	0.0	0.0
2,422,651 T>C	L995S	L1014	IIS6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.157	0.0	0.0	0.0	0.674	1.0	0.0	0.0	0.76
2,422,652 A>T	L995F	L1014	IIS6	0.84	0.818	0.853	0.915	0.875	0.0	0.0	0.525	1.0	1.0	1.0	0.326	0.0	0.0	0.0	0.0
2,429,556 G>A	V1507I	-	IIIL56	0.0	0.0	0.0	0.0	0.125	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2,429,617 T>C	I1527T	I1532	IIS6	0.0	0.173	0.133	0.085	0.125	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2,429,745 A>T	N1570Y	N1575	LIII/IV	0.0	0.0	0.267	0.0	0.0	0.0	0.0	0.057	0.167	0.207	0.088	0.0	0.0	0.0	0.0	0.0
2,429,897 A>G	E1597G	E1602	LIII/IV	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.065	0.062	0.0	0.0	0.0	0.0	0.0
2,429,915 A>C	K1603T	K1608	IVS1	0.0	0.055	0.047	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2,430,424 G>T	A1746S	A1751	IVS5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.292	0.141	0.1	0.0	0.0	0.0	0.0	0.0
2,430,817 G>A	V1853I	V1858	COOH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.542	0.049	0.062	0.0	0.0	0.0	0.0	0.0
2,430,863 T>C	I1868T	I1873	COOH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.261	0.2	0.0	0.0	0.0	0.0	0.0
2,430,880 C>T	P1874S	P1879	COOH	0.0	0.027	0.207	0.345	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2,430,881 C>T	P1874L	P1879	COOH	0.0	0.0	0.073	0.007	0.25	0.0	0.0	0.0	0.0	0.234	0.475	0.0	0.0	0.0	0.0	0.0
2,431,061 C>T	A1934V	A1939	COOH	0.0	0.018	0.107	0.465	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2,431,079 T>C	I1940T	I1945	COOH	0.0	0.118	0.04	0.0	0.0	0.0	0.0	0.067	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

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

² Codon numbering according to *Anopheles gambiae* transcript AGAP004707-RD in geneset AgamP4.12.

³ Codon numbering according to *Musca domestica* EMBL accession X96668 [10].

⁴ Location of the variant within the protein structure. Transmembrane segments are named according to domain number (in Roman numerals) followed by ‘S’ then the number of the segment; e.g., ‘IIS6’ means domain two, transmembrane segment six. Internal linkers between segments within the same domain are named according to domain (in Roman numerals) followed by ‘L’ then the numbers of the linked segments; e.g., ‘IL45’ means domain one, linker between transmembrane segments four and five. Internal linkers between domains are named ‘L’ followed by the linked domains; e.g., ‘LI/II’ means the linker between domains one and two. ‘COOH’ means the internal carboxyl tail.

149 [24].

150 The abundance of high-frequency non-synonymous variants occurring in combination
151 with L995F is notable for two reasons. First, *Vgsc* is a highly conserved gene, expected
152 to be under strong functional constraint and therefore purifying selection, so any non-
153 synonymous variants are expected to be rare [12]. Second, in contrast with L995F, we did
154 not observe any high-frequency non-synonymous variants occurring in combination with
155 L995S. This contrast was clear when data on all variants within the gene were considered:
156 for haplotypes carrying the L995 allele, the ratio of non-synonymous to synonymous nu-
157 cleotide diversity π_N/π_S was 20.04 times higher than haplotypes carrying the wild-type
158 allele, but for those carrying L995S π_N/π_S was 0.5 times lower than haplotypes carrying
159 the wild-type allele. These results indicate that L995F has substantially altered the se-
160 lective regime for other amino acid positions within the protein. Secondary substitutions
161 have occurred and risen in frequency, suggesting that they are providing some further
162 selective advantage in the presence of insecticide pressure.

163 A novel allele, I1527T, was present in *An. coluzzii* from Ghana, Burkina Faso, Cote
164 d'Ivoire and Guinea. Codon 1527 occurs within trans-membrane segment IIIS6, imme-
165 diately adjacent to residues within a predicted binding site for pyrethroid molecules, thus
166 it is plausible that I1527T could alter pyrethroid binding [27, 5]. We also found that the
167 two variant alleles affecting codon 402, both of which induce a V402L substitution, were
168 in strong linkage with I1527T ($D' \geq 0.8$; Figure 2), and almost all haplotypes carrying
169 I1527T also carried a V402L substitution. Substitutions in codon 402 have been found in
170 a number of other insect species and shown experimentally to confer pyrethroid resistance
171 [5]. The species and geographical distribution of the I1527T+V402L alleles suggest they
172 arose in West African *An. coluzzii* and had not spread to other regions or to *An. gambiae*
173 at the time of sampling. The I1527T allele was present at lower frequency than L995F
174 in all of the West African *An. coluzzii* populations. L995F is known to have increased in
175 frequency in West African *An. coluzzii* [28] and thus could be replacing I1527T+V402L
176 in these populations. The four remaining novel alleles, Q697P, G531V and two separate
177 nucleotide substitutions causing M490I, did not occur in combination with any known
178 resistance allele and were almost exclusively private to a single population (Table 1).

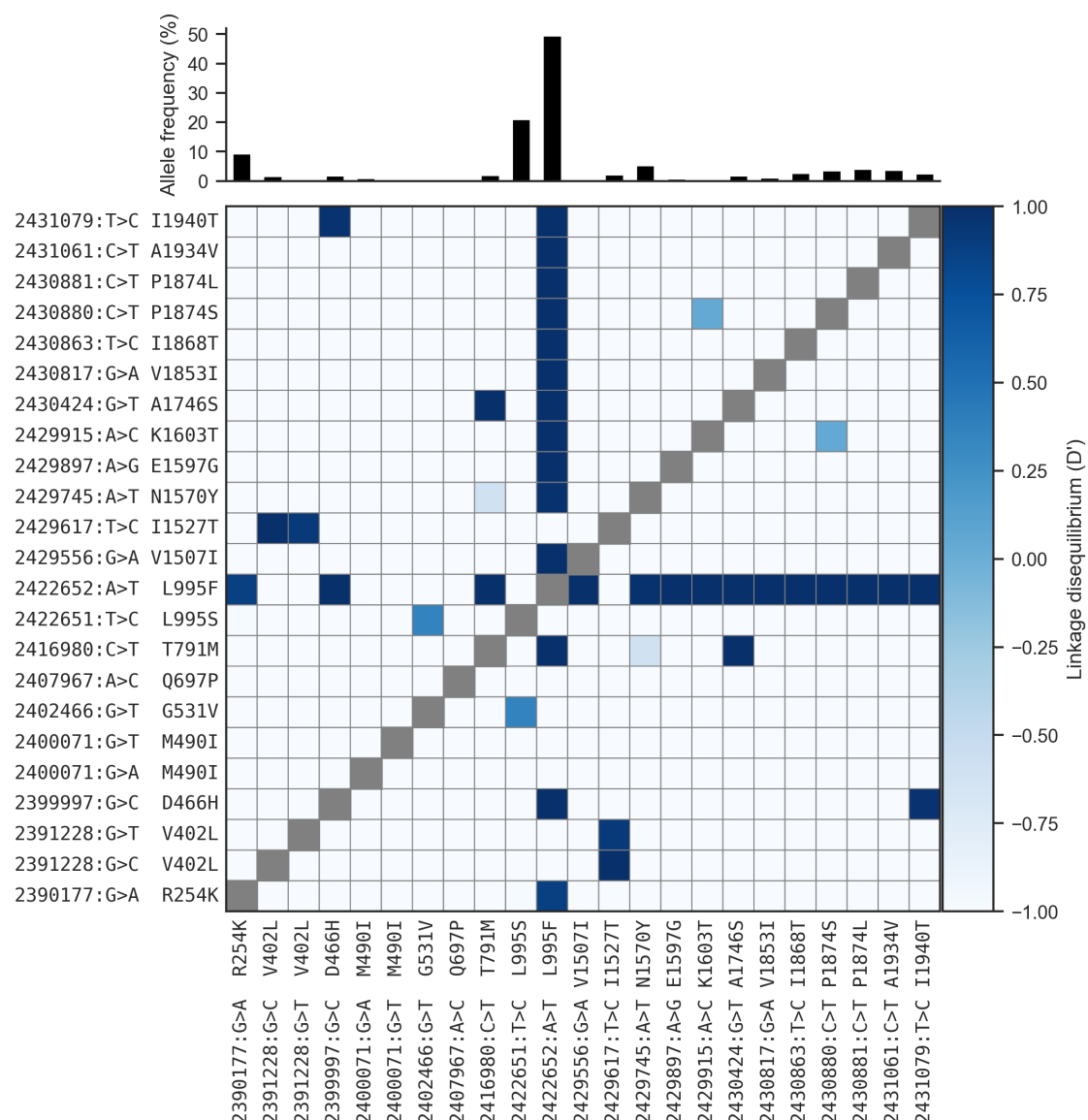


Figure 2. Linkage disequilibrium (D') between non-synonymous variants. A value of 1 indicates that two alleles are in perfect linkage, meaning that one of the alleles is only ever found in combination with the other. Conversely, a value of -1 indicates that two alleles are never found in combination with each other. The bar plot at the top shows the frequency of each allele within the Ag1000G phase 2 cohort. See Table 1 for population allele frequencies.

179 Genetic backgrounds carrying resistance alleles

180 The Ag1000G data resource provides a rich source of information about the spread of
 181 insecticide resistance alleles in any given gene, because data are not only available for
 182 SNPs in protein coding regions, but also SNPs in introns, flanking intergenic regions,
 183 and in neighbouring genes. These additional variants can be used to analyse the genetic
 184 backgrounds (haplotypes) on which resistance alleles are found. In our initial report of

185 the Ag1000G phase 1 resource [23], we used 1710 biallelic SNPs from within the 73.5 kbp
186 *Vgsc* gene (1607 intronic, 103 exonic) to compute the number of SNP differences between
187 all pairs of 1530 haplotypes derived from 765 wild-caught mosquitoes. We then used
188 pairwise genetic distances to perform hierarchical clustering, and found that haplotypes
189 carrying resistance alleles in codon 995 were grouped into 10 distinct clusters, each with
190 near-identical haplotypes. Five of these clusters contained haplotypes carrying the L995F
191 allele (labelled F1-F5), and a further five clusters contained haplotypes carrying L995S
192 (labelled S1-S5).

193 To further investigate genetic backgrounds carrying resistance alleles, we used the
194 Ag1000G phase 2 haplotype data from the *Vgsc* gene (2,284 haplotypes from 1,142 mosquitoes
195 [22]), to construct median-joining networks [29] (Figure 3). The network analysis improves
196 on hierarchical clustering by allowing for the reconstruction and placement of intermedi-
197 ate haplotypes that may not be observed in the data. It also allows for non-hierarchical
198 relationships between haplotypes, which may arise if recombination events have occurred
199 between haplotypes. We constructed the network up to a maximum edge distance of 2 SNP
200 differences, to ensure that each connected component captures a group of closely-related
201 haplotypes. The resulting network contained 5 groups containing haplotypes carrying
202 L995F, and a further 5 groups carrying L995S, in close correspondence with previous re-
203 sults from hierarchical clustering (96.8% overall concordance in assignment of haplotypes
204 to groups).

205 The haplotype network brings into sharp relief the explosive radiation of amino acid sub-
206 stitutions secondary to the L995F allele (Figure 3). Within the F1 group, nodes carrying
207 non-synonymous variants radiate out from a central node carrying only L995F, suggest-
208 ing that the central node represents the ancestral haplotype carrying just L995F which
209 initially came under selection, and these secondary variants have arisen subsequently as
210 new mutations. In F1 alone, 30 network edges (shown as red arrows - Figure 3) lead to
211 non-synonymous nodes. Many of the nodes carrying secondary variants are large, consis-
212 tent with positive selection and a functional role for these secondary variants as modifiers
213 of the L995F resistance phenotype. The F1 network also allows us to infer multiple intro-
214 gression events between the two species. The central (putatively ancestral) node contains
215 haplotypes from individuals of both species, as do nodes carrying the N1570Y, P1874L and

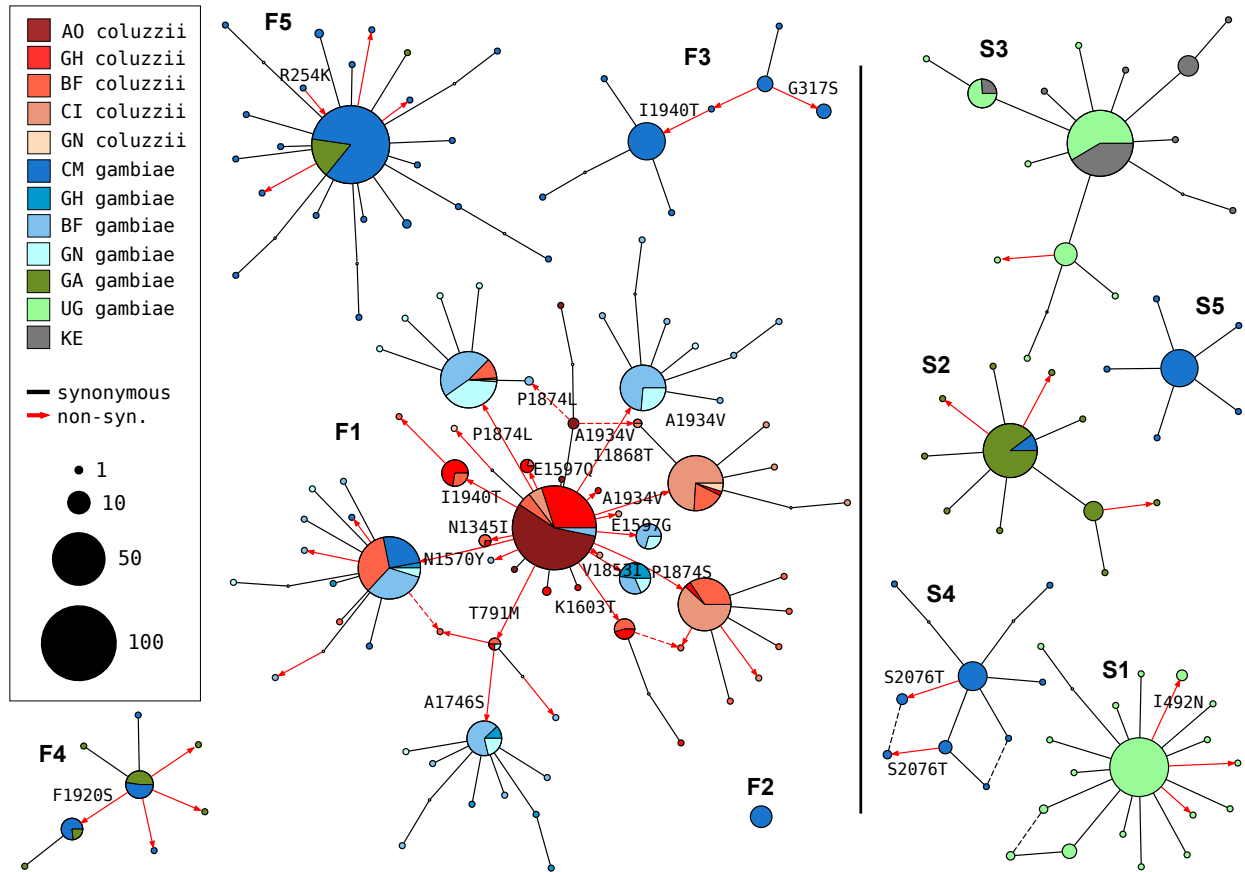
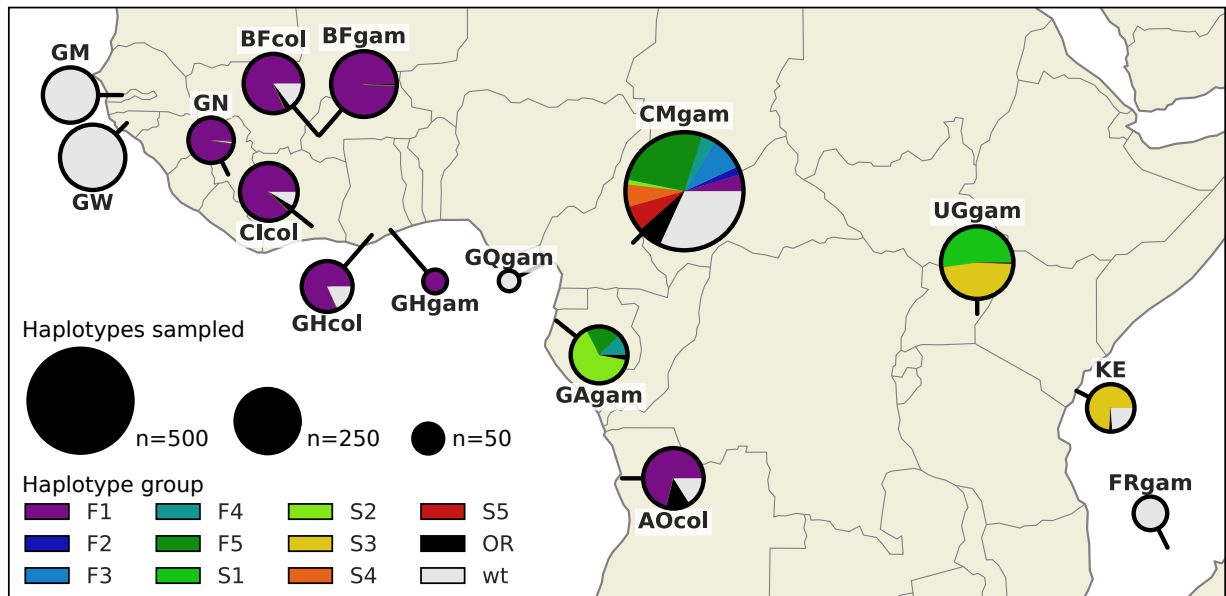


Figure 3. Haplotype networks. Median joining network for haplotypes carrying L995F (labelled F1-F5) or L995S variants (S1-S5) with a maximum edge distance of two SNPs. Labelling of network components is via concordance with hierarchical clusters discovered in [23]. Node size is relative to the number of haplotypes contained and node colour represents the proportion of haplotypes from mosquito populations/species - AO=Angola; GH=Ghana, BF=Burkina Faso; CI=Côte d'Ivoire; GN=Guinea; CM=Cameroon; GA=Gabon; UG=Uganda; KE=Kenya. Non-synonymous edges are highlighted in red and those leading to non-singleton nodes are labelled with the codon change, arrow head indicates direction of change away from the reference allele. Network components with fewer than three haplotypes are not shown.

216 T791M variants. This structure is consistent with an initial introgression of the ancestral
 217 F1 haplotype, followed later by introgressions of haplotypes carrying secondary mutations.
 218 The haplotype network also illustrates the contrasting levels of non-synonymous varia-
 219 tion between L995F and L995S. Within all of the L995S groups, only eight edges lead to
 220 non-synonymous nodes and all these nodes are small (low frequency variants), thus may
 221 be neutral or mildly deleterious variants that are hitch-hiking on selective sweeps for the
 222 L995S allele.

223 The F1 group contains haplotypes from mosquitoes of both species, and from mosquitoes
 224 sampled in six different countries (Angola, Burkina Faso, Cameroon, Côte d'Ivoire, Ghana,



Guinea) (Figure 4). The F4, F5 and S2 groups each contain haplotypes from both Cameroon and Gabon. The S3 group contains haplotypes from both Uganda and Kenya. The haplotypes within each of these five groups (F1, F4, F5, S2, S3) were nearly identical across the entire span of the *Vgsc* gene ($\pi < 4.5 \times 10^{-5} bp^{-1}$). In contrast, diversity among wild-type haplotypes was two orders of magnitude greater (Cameroon *An. gambiae* $\pi = 1.4 \times 10^{-3} bp^{-1}$; Guinea-Bissau $\pi = 5.7 \times 10^{-3} bp^{-1}$). Thus it is reasonable to assume that each of these five groups contains descendants of an ancestral haplotype that carried a resistance allele and has risen in frequency due to selection for insecticide resistance. Given this assumption, these groups each provide evidence for adaptive gene flow between mosquito populations separated by considerable geographical distances.

A limitation of both the hierarchical clustering and network analyses is that they rely on genetic distances within a fixed genomic window from the start to the end of the *Vgsc* gene. *Anopheles* mosquitoes undergo homologous recombination during meiosis both males and females, and any recombination events that occurred within this genomic window could affect the way that haplotypes are grouped together in clusters or network

240 components. In particular, recombination events could occur during the geographical
 241 spread of a resistance allele, altering the genetic background upstream and/or downstream
 242 of the allele itself. An analysis based on a fixed genomic window might then fail to infer
 243 gene flow between two mosquito populations, because haplotypes with and without a
 244 recombination event could be grouped separately, despite the fact that they share a recent
 245 common ancestor. To investigate the possibility that recombination events may have
 246 affected our grouping of haplotypes carrying resistance alleles, we performed a moving
 247 window analysis of haplotype homozygosity, spanning *Vgsc* and up to a megabase upstream
 248 and downstream of the gene (Supplementary Figures S1, S2). This analysis supported a
 249 refinement of our initial grouping of haplotypes carrying resistance alleles. All haplotypes
 250 within groups S4 and S5 were effectively identical on both the upstream and downstream
 251 flanks of the gene, but there was a region of divergence within the *Vgsc* gene itself that
 252 separated them in the fixed window analyses (Supplementary Figure S2). The 13.8 kbp
 253 region of divergence occurred upstream of codon 995 and contained 6 SNPs that were fixed
 254 differences between S4 and S5. A possible explanation for this short region of divergence
 255 is that a gene conversion event has occurred within the gene, bringing a segment from
 256 a different genetic background onto the original genetic background on which the L995S
 257 resistance mutation occurred.

258 **Positive selection for resistance alleles**

259 To investigate evidence for positive selection on non-synonymous alleles, we performed
 260 an analysis of extended haplotype homozygosity (EHH) [30]. Haplotypes under recent
 261 positive selection will have increased rapidly in frequency, thus have had less time to be
 262 broken down by recombination, and should on average have longer regions of haplotype
 263 homozygosity relative to wild-type haplotypes. We defined a core region spanning *Vgsc*
 264 codon 995 and an additional 6 kbp of flanking sequence, which was the minimum required
 265 to differentiate the haplotype groups identified via clustering and network analyses. Within
 266 this core region, we found 18 distinct haplotypes at a frequency above 1% within the cohort.
 267 These included core haplotypes corresponding to each of the 10 haplotype groups carrying
 268 L995F or L995S alleles identified above, as well as a core haplotype carrying I1527T which
 269 we labelled L1 (due to it carrying the the wild-type leucine codon at position 995). We also

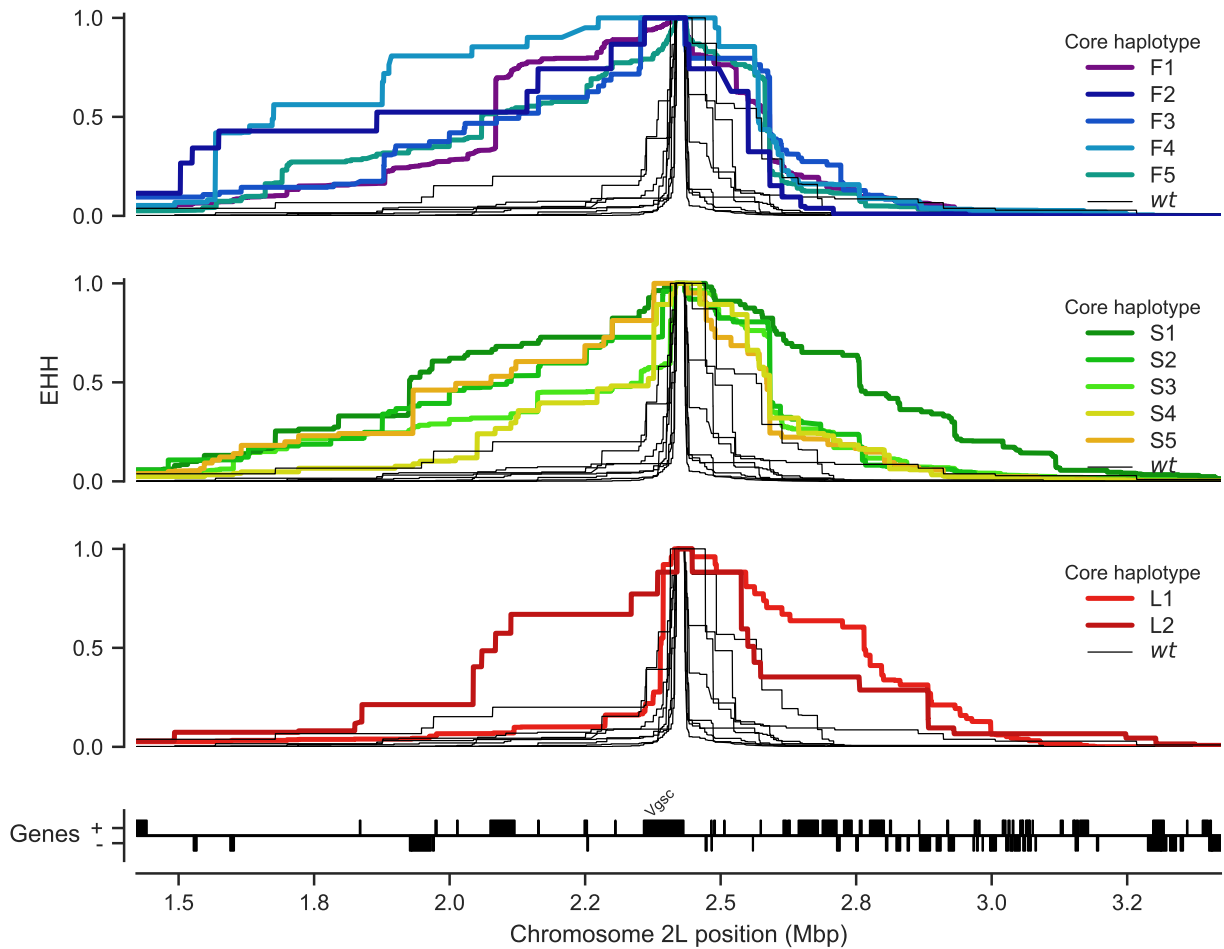


Figure 5. Evidence for positive selection on haplotypes carrying known or putative resistance alleles. Each panel plots the decay of extended haplotype homozygosity (EHH) for a set of core haplotypes centred on *Vgsc* codon 995. Core haplotypes F1-F5 carry the L995F allele; S1-S5 carry the L995S allele; L1 carries the I1527T allele; L2 carries the M490I allele. Wild-type (*wt*) haplotypes do not carry known or putative resistance alleles. A slower decay of EHH relative to wild-type haplotypes implies positive selection (each panel plots the same collection of wild-type haplotypes).

found a core haplotype corresponding to a group of haplotypes from Kenya carrying an M490I allele, which we labelled as L2. All other core haplotypes we labelled as wild-type (*wt*). We then computed EHH decay for each core haplotype up to a megabase upstream and downstream of the core locus (Figure 5).

As expected, haplotypes carrying the L995F and L995S resistance alleles all experience a slower decay of EHH relative to wild-type haplotypes, supporting positive selection. Previous studies have found evidence for different rates of EHH decay between L995F and L995S haplotypes, suggesting differences in the timing and/or strength of selection [16]. However, we found no systematic difference in the length of shared haplotypes when

279 comparing F1-5 (carrying L995F) against S1-5 (carrying L995S) (Supplementary Figure
280 S3). There were, however, some differences between core haplotypes carrying the same
281 allele. For example, shared haplotypes were significantly longer for S1 (median 1.006 cM,
282 95% CI [0.986 - 1.040]) versus other core haplotypes carrying L995S (e.g., S2 median
283 0.593 cM, 95% CI [0.589 - 0.623]; Supplementary Figure S3). Longer shared haplotypes
284 indicate a more recent common ancestor, and thus some of these core haplotypes may
285 have experienced more recent and/or more intense selection than others.

286 The L1 haplotype carrying I1527T+V402L exhibited a slow decay of EHH on the down-
287 stream flank of the gene, similar to haplotypes carrying L995F and L995S, indicating that
288 this combination of alleles has experienced positive selection. EHH decay on the upstream
289 gene flank was faster, being similar to wild-type haplotypes, however there were two sepa-
290 rate nucleotide substitutions encoding V402L within this group of haplotypes, and a faster
291 EHH decay on this flank is consistent with recombination events bringing V402L alleles
292 from different genetic backgrounds together with an ancestral haplotype carrying I1527T.
293 The L2 haplotype carrying M490I exhibited EHH decay on both flanks comparable to hap-
294 lotypes carrying known resistance alleles. This could indicate evidence for selection on the
295 M490I allele, but these haplotypes are derived from a Kenyan mosquito population where
296 there is evidence for a severe recent bottleneck [23], and there were not enough wild-type
297 haplotypes from Kenya with which to compare. Thus this signal may also be due to the
298 extreme demographic history of this population.

299 Discussion

300 Cross-resistance between pyrethroids and DDT

301 The VGSC protein is the physiological target of both pyrethroid insecticides and DDT [6].
302 The L995F and L995S alleles are known to increase resistance to both of these insecticide
303 classes [7, 9]. By 2012, over half of African households owned at least one pyrethroid
304 impregnated ITN and nearly two thirds of IRS programmes were using pyrethroids [2].
305 Pyrethroids were also introduced into agriculture in Africa prior to the scale-up of public
306 health vector control programmes, and continue to be used on a variety of crops such as
307 cotton [31]. DDT was used in Africa for several pilot IRS projects carried out during the

308 first global campaign to eradicate malaria, during the 1950s and 1960s [12]. DDT is still
 309 approved for IRS use by WHO and remains in use in some locations, however within the
 310 last two decades pyrethroid use has been far more common and widespread. DDT was also
 311 used in agriculture from the 1940s, and although agricultural usage has greatly diminished
 312 since the 1970s, some usage remains [32]. In this study we reported evidence of positive
 313 selection on the L995F and L995S alleles, as well as the I1527T+V402L combination and
 314 possibly M490I. We also found 14 other non-synonymous substitutions that have arisen in
 315 association with L995F and appear to be positively selected. Given that pyrethroids have
 316 dominated public health insecticide use for two decades, it is reasonable to assume that the
 317 selection pressure on these alleles is primarily due to pyrethroids rather than DDT. It has
 318 previously been suggested that L995S may have been initially selected by DDT usage [16].
 319 However, we did not find any systematic difference in the extent of haplotype homozygosity
 320 between these two alleles, suggesting that both alleles have been under selection over a
 321 similar time frame. We did find some significant differences in haplotype homozygosity
 322 between different genetic backgrounds carrying resistance alleles, suggesting differences
 323 in the timing and/or strength of selection these may have experienced. However, there
 324 have been differences in the scale-up of pyrethroid-based interventions in different regions,
 325 and this could in turn generate heterogeneities in selection pressures. Nevertheless, it is
 326 possible that some if not all of the alleles we have reported provide some level of cross-
 327 resistance to DDT as well as pyrethroids, and we cannot exclude the possibility that
 328 earlier DDT usage may have contributed at least in part to their selection. The differing
 329 of resistance profiles to the two types of pyrethroids (type I, e.g., permethrin; and type
 330 II, e.g., deltamethrin) [33], will also affect the selection landscape. Further sampling and
 331 analysis will be required to investigate the timing of different selection events and relate
 332 these to historical patterns of insecticide use in different regions.

333 **Resistance phenotypes for novel non-synonymous variants**

334 The non-synonymous variants are distributed throughout the channel protein but can
 335 be considered in terms of three clusters: (i) the transmembrane domain, (ii) the DI-II
 336 intracellular linker and (iii) the DIII-DIV/C-terminal subdomain. The pyrethroid binding
 337 site is located in the transmembrane domain between the IIS4-S5 linker and the IIS5,

338 IIS6 and IIIS6 helices [34]. The I1527T substitution that we discovered in *An. coluzzii*
 339 mosquitoes from Burkina Faso occurs in segment IIIS6 and is immediately adjacent to
 340 two pyrethroid-sensing residues in this binding site [5]. It is thus plausible that pyrethroid
 341 binding could be altered by this substitution. The I1527T substitution (*M. domestica*
 342 codon 1532) has been found in *Aedes albopictus* [35], and substitutions in the nearby
 343 codon 1529 (*M. domestica* I1534T) have been reported in *Aedes albopictus* and in *Aedes*
 344 *aegypti* where it was found to be associated with pyrethroid resistance [5, 36, 37]. We
 345 found the I1527T allele in tight linkage with two alleles causing a V402L substitution (*M.*
 346 *domestica* V410L). Substitutions in codon 402 have been found in multiple insect species
 347 and are by themselves sufficient to confer pyrethroid resistance [5]. The fact that we
 348 find I1527T and V402L in such tight mutual association is intriguing because haplotypes
 349 carrying V402L alone should also have been positively selected and thus be present in one
 350 or more populations.

351 The V402 residue is located towards the middle of the IS6 helix. The L995F and L995S
 352 substitutions occur at a similar position on the IIS6 helix. It was proposed these S6 sub-
 353 stitutions confer resistance by allosterically modifying formation of the pyrethroid binding
 354 site [34]. More recently the L995 *kdr* residue was speculated to form part of a second
 355 pyrethroid binding site in the insect channel termed 'PyR2' [27, 38]. A major functional
 356 effect of the L995F substitution is enhanced closed-state inactivation [39]. This contributes
 357 to *kdr* resistance by reducing the number of channels that undergo activation, which is
 358 the functional state that pyrethroids bind to with highest affinity [39]. Fast inactivation
 359 involves movement of the DIV domain to form a receptor for the DIII-DIV linker fast in-
 360 activation particle containing the 'MFM' sequence motif (equivalent to the 'IFM' motif in
 361 mammals) [40, 5]. Recent eukaryotic sodium channel structures reveal that the DIII-DIV
 362 linker is in complex with the C-terminal segment in the closed-state conformation but the
 363 DIII-DIV linker appears to dissociate and bind in close proximity in the DIV S6 helix upon
 364 transition to the inactivated state [25, 41]. It seems that binding of the DIII-DIV linker
 365 pushes the DIV S6 helix forward to occlude the pore and produce the inactivated state
 366 [41]. We suggest that substitutions located on the DIII-DIV linker and C-terminal tail may
 367 perturb the conformation of this subdomain when it assembles in the closed-state channel
 368 and may subsequently affect capture or release of the DIII-DIV linker from this complex.

369 The expected functional outcome would be altered channel inactivation, although whether
370 inactivation is enhanced or diminished and if this compensates for a deleterious effect of
371 L995F on channel function awaits elucidation. The N1570Y substitution on the DIII-DIV
372 linker has been functionally characterised but inactivation kinetics in the mutant channel
373 were found unaltered [26]. Pyrethroid sensitivity was also unaffected by N1570Y although
374 resistance was greatly enhanced in the N1570Y + L995F double mutant [26].

375 The final cluster of novel variants is located on the DI-DII intracellular linker. This
376 segment includes the novel M490I substitution that was found on the Kenyan L2 haplotypic
377 background potentially under selection. M490I did not occur in association with L995F or
378 any other non-synonymous substitutions. Although we were unable to model this region,
379 we speculate that the DI-DII linker passes under the DII S4-S5 linker and these regions
380 may interact, as was found in a bacterial sodium channel structure [42]. The structural
381 effects of DI-DII substitutions may be altered interactions with the DII S4-S5 linker, the
382 movement of which is critical for formation of the pyrethroid binding site [34, 43]. Overall,
383 there are a number of potential mechanisms by which a pyrethroid resistance phenotype
384 may arise and topology modelling reveals how many of the non-synonymous variants we
385 discover may be involved, though clearly much remains to be unravelled regarding the
386 molecular biology of pyrethroid resistance in this channel.

387 **Design of genetic assays for surveillance of pyrethroid resistance**

388 Entomological surveillance teams in Africa regularly genotype mosquitoes for resistance al-
389 leles in *Vgsc* codon 995, and use those results as an indicator for the presence of pyrethroid
390 resistance alongside results from insecticide resistance bioassays. They typically do not,
391 however, sequence the gene or genotype any other polymorphisms within the gene. Thus,
392 if there are other polymorphisms within the gene that cause or significantly enhance
393 pyrethroid resistance, these will not be detected. Also, if a codon 995 resistance allele is
394 observed, there is no way to know whether the allele is on a genetic background that has
395 also been observed in other mosquito populations, and thus no way to investigate whether
396 resistance alleles are emerging locally or being imported from elsewhere. Whole-genome
397 sequencing of individual mosquitoes clearly provides data of sufficient resolution to answer
398 these questions, and could be used to provide ongoing resistance surveillance. The cost of

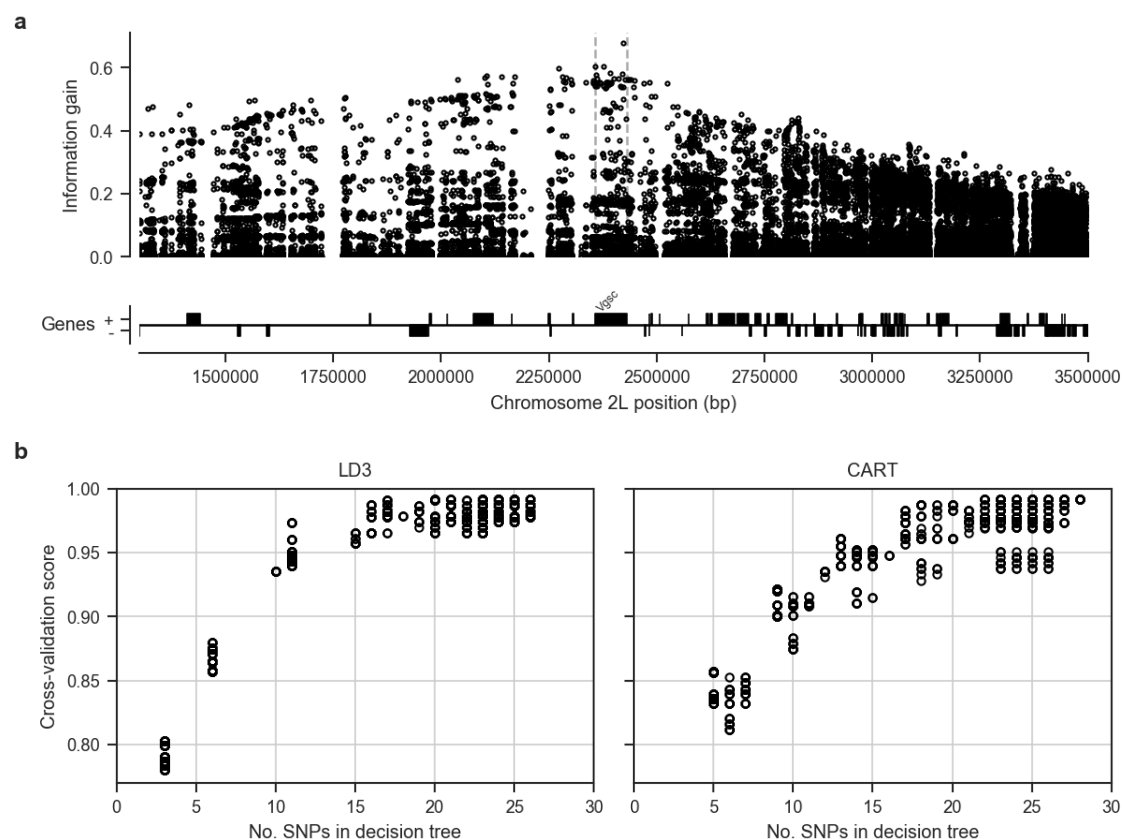


Figure 6. Informative SNPs for haplotype surveillance. **a**, Each data point represents a single SNP. The information gain value for each SNP provides an indication of how informative the SNP is likely to be if used as part of a genetic assay for testing whether a mosquito carries a resistance haplotype, and if so, which haplotype group it belongs to. **b**, Number of SNPs required to accurately predict which group a resistance haplotype belongs to. Each data point represents a single decision tree. Decision trees were constructed using either the LD3 (left) or CART (right) algorithm for comparison. Accuracy was evaluated using 10-fold stratified cross-validation.

whole-genome sequencing continues to fall, making it a practical tool for malaria vector surveillance. However, to achieve substantial spatial and temporal coverage of mosquito populations, it would also be necessary to develop targeted genetic assays for resistance outbreak surveillance. Technologies such as amplicon sequencing [44] are already being trialled on mosquitoes [45], these could scale to tens of thousands of samples at low cost and could be implemented using existing platforms in national molecular biology facilities.

To facilitate the development of targeted genetic assays for surveillance of *Vgsc*-mediated pyrethroid resistance, we have produced several supplementary data tables. In Supplementary Table 1 we list all 82 non-synonymous variants found within the *Vgsc* gene in this study, with population allele frequencies. In Supplementary Table 2 we list 756 biallelic SNPs, within the *Vgsc* gene and up to 10 kbp upstream or downstream, that are poten-

tially informative regarding which haplotype group a resistance haplotype belongs to, and thus could be used for tracking the spread of resistance. This table includes the allele frequency within each of the 10 haplotype groups defined here, to aid in identifying SNPs that are highly differentiated between two or more haplotype groups. We also provide Supplementary Table 3 which lists all 10,244 SNPs found within the *Vgsc* gene and up to 10 kbp upstream or downstream, which might need to be taken into account as flanking variation when searching for PCR primers to amplify a SNP of interest. To provide some indication for how many SNPs would need to be assayed in order to track the spread of resistance, we used haplotype data from this study to construct decision trees that could classify which of the 12 groups a given haplotype belongs to (Figure 6). This analysis suggested that it should be possible to construct a decision tree able to classify haplotypes with >95% accuracy by using 20 SNPs or less. In practice, more SNPs would be needed, to provide some redundancy, and also to type non-synonymous polymorphisms in addition to identifying the genetic background. However, it is still likely to be well within the number of SNPs that could be assayed in a single multiplex via amplicon sequencing. Thus it should be feasible to produce low-cost, high-throughput genetic assays for tracking the spread of pyrethroid resistance. If combined with whole-genome sequencing of mosquitoes at sentinel sites, this should also allow the identification of newly emerging resistance outbreaks.

Methods

Code

All scripts and Jupyter Notebooks used to generate analyses, figures and tables are available from the GitHub repository <https://github.com/malariagen/ag1000g-phase2-vgsc-report>.

Data

We used variant calls and phased haplotype data from the Ag1000G Phase 2 AR1 data release (<https://www.malariagen.net/data/ag1000g-phase-2-ar1>). Variant calls from Ag1000G Phase 2 are also available from the European Nucleotide Archive (ENA; <http://www.ebi.ac.uk/ena>) under study PRJEB36277.

438 Data collection and processing

439 For detailed information on Ag1000G WGS sample collection, sequencing, variant call-
440 ing, quality control and phasing, see [23, 22]. In brief, *An. gambiae* and *An. coluzzii*
441 mosquitoes were collected from 33 sites in 13 countries across Sub-Saharan Africa: An-
442 gola, Bioko, Burkina Faso, Cameroon, Côte d’Ivoire, Gabon, The Gambia, Ghana, Guinea,
443 Guinea Bissau, Kenya, Mayotte and Uganda. From Angola and Côte d’Ivoire just *An.*
444 *coluzzii* were sampled, Burkina Faso, Ghana and Guinea had samples of both *An. gambiae*
445 and *An. coluzzii* and all other populations consisted of purely *An. gambiae*, except for
446 The Gambia, Guinea Bissau and Kenya where species status is uncertain [22]. Mosquitoes
447 were individually whole genome sequenced on the Illumina HiSeq 2000 platform, gener-
448 ating 100bp paired-end reads. Sequence reads were aligned to the *An. gambiae* AgamP3
449 reference genome assembly [46]. Aligned bam files underwent improvement, before variants
450 were called using GATK UnifiedGenotyper. Quality control included removal of samples
451 with mean coverage $\leq 14\times$ and filtering of variants with attributes that were correlated
452 with Mendelian error in genetic crosses.

453 The Ag1000G variant data was functionally annotated using the SnpEff v4.1b software
454 [47]. Non-synonymous *Vgsc* variants were identified as all variants in AgamP4.12 transcript
455 AGAP004707-RD with a SnpEff annotation of “missense”. The *Vgsc* gene is known to
456 exhibit alternative splicing [6], however at the time of writing the *An. gambiae* gene
457 annotations did not include the alternative transcripts reported by Davies et al. We wrote
458 a Python script to check for the presence of variants that are synonymous according to
459 transcript AGAP004707-RD but non-synonymous according to one of the other transcripts
460 present in the gene annotations or in the set reported by Davies et al. Supplementary Table
461 1 includes the predicted effect for all SNPs that are non-synonymous in one or more of
462 these transcripts. None of the variants that are non-synonymous in a transcript other
463 than AGAP004707-RD were found to be above 5% frequency in any population.

464 For ease of comparison with previous work on *Vgsc*, pan Insecta, in Table 1 and Supple-
465 mentary Table 1 we report codon numbering for both *An. gambiae* and *Musca domestica*
466 (the species in which the gene was first discovered). The *M. domestica* *Vgsc* sequence
467 (EMBL accession X96668 [10]) was aligned with the *An. gambiae* AGAP004707-RD se-

468 quence (AgamP4.12 gene-set) using the Mega v7 software package [48]. A map of equivalent codon numbers between the two species for the entire gene can be download from the
469 MalariaGEN website ([https://www.malariagen.net/sites/default/files/content/](https://www.malariagen.net/sites/default/files/content/blogs/domestica_gambiae_map.txt)
470 [blogs/domestica_gambiae_map.txt](https://www.malariagen.net/sites/default/files/content/blogs/domestica_gambiae_map.txt)).
471

472 Haplotypes for each chromosome of each sample were estimated (phased) using using
473 phase informative reads (PIRs) and SHAPEIT2 v2.r837 [49], see [23] supplementary text
474 for more details. The SHAPEIT2 algorithm is unable to phase multi-allelic positions,
475 therefore the two multi-allelic non-synonymous SNPs within the *Vgsc* gene, altering codons
476 V402 and M490, were phased onto the biallelic haplotype scaffold using MVNcall v1.0 [50].
477 Lewontin's D' [51] was used to compute the linkage disequilibrium (LD) between all pairs
478 of non-synonymous *Vgsc* mutations.

479 Haplotype networks

480 Haplotype networks were constructed using the median-joining algorithm [29] as implemented in a Python module available from <https://github.com/malariagen/ag1000g-phase2-vgsc-report>
481 Haplotypes carrying either L995F or L995S mutations were analysed with a maximum edge
482 distance of two SNPs. Networks were rendered with the Graphviz library and a composite figure constructed using Inkscape. Non-synonymous edges were highlighted using the
483 SnpEff annotations [47].
484
485

486 Positive selection

487 Core haplotypes were defined on a 6,078 bp region spanning *Vgsc* codon 995, from chromosome arm 2L position 2,420,443 and ending at position 2,426,521. This region was
488 chosen as it was the smallest region sufficient to differentiate between the ten genetic
489 backgrounds carrying either of the known resistance alleles L995F or L995S. Extended
490 haplotype homozygosity (EHH) was computed for all core haplotypes as described in [30]
491 using scikit-allel version 1.1.9 [52], excluding non-synonymous and singleton SNPs. Analyses of haplotype homozygosity in moving windows (Supplementary Figs. S1, S2) and pairwise
492 haplotype sharing (Supplementary Figure S3) were performed using custom Python
493 code available from <https://github.com/malariagen/ag1000g-phase2-vgsc-report>.
494
495

496 Design of genetic assays for surveillance of pyrethroid resistance

497 To explore the feasibility of indentifying a small subset of SNPs that would be sufficient
498 to identify each of the genetic backgrounds carrying known or putative resistance alleles,
499 we started with an input data set of all SNPs within the *Vgsc* gene or in the flanking
500 regions 20 kbp upstream and downstream of the gene. Each of the 2,284 haplotypes in
501 the Ag1000G Phase 2 cohort was labelled according to which core haplotype it carried,
502 combining all core haplotypes not carrying known or putative resistance alleles together as
503 a single "wild-type" group. Decision tree classifiers were then constructed using scikit-learn
504 version 0.19.0 [53] for a range of maximum depths, repeating the tree construction process
505 10 times for each maximum depth with a different initial random state. The classification
506 accuracy of each tree was evaluated using stratified 5-fold cross-validation.

507 Homology modelling

508 A homology model of the *An. gambiae* voltage-gated sodium channel (AGAP004707-RD
509 AgamP4.12) was generated using the 3.8 Å resolution structure of the *Periplaneta amer-*
510 *icana* sodium channel Na_vPaS structure (PDB code 5X0M) [25]. Sequences were aligned
511 using Clustal Omega [54]. 50 starting models were generated using MODELLER [55].
512 The internal scoring function of MODELLER was used to select 10 models, which were
513 visually inspected and submitted to the VADAR webserver [56] to assess stereochemistry
514 in order to select the best final model. Figures were produced using PyMOL (DeLano
515 Scientific, San Carlos, CA, USA).

516 References

- 517 [1] S. Bhatt et al. ‘The effect of malaria control on *Plasmodium falciparum* in Africa
518 between 2000 and 2015’. In: *Nature* 526.7572 (2015), pp. 207–211. ISSN: 0028-0836.
- 519 [2] Janet Hemingway et al. ‘Averting a malaria disaster: Will insecticide resistance derail
520 malaria control?’ In: *The Lancet* 387.10029 (2016), pp. 1785–1788. ISSN: 1474547X.
- 521 [3] World Health Organization. *Global Plan for Insecticide Resistance Management*
522 *(GPIRM)*. Tech. rep. Geneva: World Health Organization, 2012.

- [4] World Health Organization et al. *Global vector control response 2017-2030*. Tech. rep. 2017.
- [5] Ke Dong et al. ‘Molecular biology of insect sodium channels and pyrethroid resistance’. In: *Insect Biochemistry and Molecular Biology* 50.1 (2014), pp. 1–17. ISSN: 09651748.
- [6] T. G.E. Davies et al. ‘A comparative study of voltage-gated sodium channels in the Insecta: Implications for pyrethroid resistance in Anopheline and other Neopteran species’. In: *Insect Molecular Biology* 16.3 (2007), pp. 361–375. ISSN: 09621075.
- [7] D. Martinez-Torres et al. ‘Molecular characterization of pyrethroid knockdown resistance (*kdr*) in the major malaria vector *Anopheles gambiae* s.s.’ In: *Insect Molecular Biology* 7.2 (1998), pp. 179–184. ISSN: 09621075.
- [8] Ana Paula B Silva, Joselita Maria M Santos and Ademir J Martins. ‘Mutations in the voltage-gated sodium channel gene of anophelines and their association with resistance to pyrethroids: a review’. In: *Parasites & Vectors* 7.1 (2014), p. 450. ISSN: 1756-3305.
- [9] H. Ranson et al. ‘Identification of a point mutation in the voltage-gated sodium channel gene of Kenyan *Anopheles gambiae* associated with resistance to DDT and pyrethroids’. In: *Insect Molecular Biology* 9.5 (2000), pp. 491–497. ISSN: 09621075.
- [10] Martin S. Williamson et al. ‘Identification of mutations in the housefly *para*-type sodium channel gene associated with knockdown resistance (*kdr*) to pyrethroid insecticides’. In: *Molecular and General Genetics* 252.1-2 (1996), pp. 51–60. ISSN: 00268925.
- [11] Christopher M Jones et al. ‘Footprints of positive selection associated with a mutation (N1575Y) in the voltage-gated sodium channel of *Anopheles gambiae*.’ In: *Proceedings of the National Academy of Sciences of the United States of America* 109.17 (2012), pp. 6614–9. ISSN: 1091-6490.
- [12] T. G. E. Davies et al. ‘DDT, pyrethrins, pyrethroids and insect sodium channels’. In: *IUBMB Life* 59.3 (2007), pp. 151–162. ISSN: 1521-6543.

- [13] Frank D. Rinkevich, Yuzhe Du and Ke Dong. ‘Diversity and convergence of sodium channel mutations involved in resistance to pyrethroids’. In: *Pesticide Biochemistry and Physiology* 106.3 (2013), pp. 93–100. ISSN: 00483575.
- [14] J Pinto et al. ‘Multiple origins of knockdown resistance mutations in the Afrotropical mosquito vector *Anopheles gambiae*’. In: *PLoS One* 2 (2007), e1243. ISSN: 19326203.
- [15] Josiane Etang et al. ‘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’. In: *Molecular Ecology* 18.14 (2009), pp. 3076–3086. ISSN: 09621083.
- [16] Amy Lynd et al. ‘Field, genetic, and modeling approaches show strong positive selection acting upon an insecticide resistance mutation in *Anopheles gambiae* s.s.’ In: *Molecular Biology and Evolution* 27.5 (2010), pp. 1117–1125. ISSN: 07374038.
- [17] Federica Santolamazza et al. ‘Remarkable diversity of intron-1 of the *para* voltage-gated sodium channel gene in an *Anopheles gambiae*/*Anopheles coluzzii* hybrid zone.’ In: *Malaria Journal* 14.1 (2015), p. 9. ISSN: 1475-2875.
- [18] Mylène Weill et al. ‘The *kdr* mutation occurs in the Mopti form of *Anopheles gambiae* s.s. through introgression’. In: *Insect Molecular Biology* 9.5 (2000), pp. 451–455.
- [19] Abdoulaye Diabaté et al. ‘The spread of the Leu-Phe *kdr* mutation through *Anopheles gambiae* complex in Burkina Faso: genetic introgression and de novo phenomena’. In: *Tropical Medicine & International Health* 9.12 (2004), pp. 1267–1273.
- [20] Chris S. Clarkson et al. ‘Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation’. In: *Nature Communications* 5 (2014). ISSN: 2041-1723.
- [21] Laura C. Norris et al. ‘Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets’. In: *Proceedings of the National Academy of Sciences* (2015), p. 201418892. ISSN: 0027-8424.
- [22] The *Anopheles gambiae* 1000 Genomes Consortium. ‘Genome variation and population structure among 1,142 mosquitoes of the African malaria vector species *Anopheles gambiae* and *Anopheles coluzzii*’. In: *bioRxiv* (2019), p. 864314.

- [23] The *Anopheles gambiae* 1000 Genomes Consortium. ‘Natural diversity of the malaria vector *Anopheles gambiae*’. In: *Nature* 552 (2017), pp. 96–100.
- [24] Shoji Sonoda et al. ‘Genomic organization of the para-sodium channel α -subunit genes from the pyrethroid-resistant and -susceptible strains of the diamondback moth’. In: *Archives of Insect Biochemistry and Physiology* 69.1 (2008), pp. 1–12. ISSN: 07394462.
- [25] Huaizong Shen et al. ‘Structure of a eukaryotic voltage-gated sodium channel at near-atomic resolution’. In: *Science* (2017), eaal4326.
- [26] L Wang et al. ‘A mutation in the intracellular loop III/IV of mosquito sodium channel synergizes the effect of mutations in helix IIS6 on pyrethroid resistance’. In: *Molecular Pharmacology* 87.3 (2015), pp. 421–429.
- [27] Yuzhe Du et al. ‘Molecular evidence for dual pyrethroid-receptor sites on a mosquito sodium channel’. In: *Proceedings of the National Academy of Sciences* 110.29 (2013), pp. 11785–11790.
- [28] Kobié H. Toé et al. ‘Increased pyrethroid resistance in malaria vectors and decreased bed net effectiveness Burkina Faso’. In: *Emerging Infectious Diseases* 20.10 (2014), pp. 1691–1696. ISSN: 10806059.
- [29] H. J. Bandelt, P. Forster and A. Rohl. ‘Median-joining networks for inferring intraspecific phylogenies’. In: *Molecular Biology and Evolution* 16.1 (1999), pp. 37–48. ISSN: 0737-4038.
- [30] Pardis C. Sabeti et al. ‘Detecting recent positive selection in the human genome from haplotype structure’. In: *Nature* 419.6909 (2002), pp. 832–837. ISSN: 0028-0836.
- [31] Molly C Reid and F Ellis McKenzie. ‘The contribution of agricultural insecticide use to increasing insecticide resistance in African malaria vectors’. In: *Malaria Journal* 15.1 (2016), p. 107.
- [32] Sara A Abuelmaali et al. ‘Impacts of agricultural practices on insecticide resistance in the malaria vector *Anopheles arabiensis* in Khartoum State, Sudan’. In: *PLoS One* 8.11 (2013), e80549.

- [33] Zhaonong Hu et al. ‘A sodium channel mutation identified in *Aedes aegypti* selectively reduces cockroach sodium channel sensitivity to type I, but not type II pyrethroids’. In: *Insect Biochemistry and Molecular Biology* 41.1 (2011), pp. 9–13.
- [34] Andrias O. O’Reilly et al. ‘Modelling insecticide-binding sites in the voltage-gated sodium channel’. In: *Biochemical Journal* 396.2 (2006), pp. 255–263. ISSN: 0264-6021.
- [35] Jiabao Xu et al. ‘Multi-country survey revealed prevalent and novel F1534S mutation in voltage-gated sodium channel (VGSC) gene in *Aedes albopictus*’. In: *PLoS Neglected Tropical Diseases* 10.5 (2016), e0004696.
- [36] Intan H Ishak et al. ‘Contrasting patterns of insecticide resistance and knockdown resistance (*kdr*) in the dengue vectors *Aedes aegypti* and *Aedes albopictus* from Malaysia’. In: *Parasites & Vectors* 8.1 (2015), p. 181.
- [37] Yiji Li et al. ‘Evidence for multiple-insecticide resistance in urban *Aedes albopictus* populations in southern China’. In: *Parasites & Vectors* 11.1 (2018), p. 4.
- [38] Yuzhe Du et al. ‘Rotational symmetry of two pyrethroid receptor sites in the mosquito sodium channel’. In: *Molecular Pharmacology* 88.2 (Aug. 2015), pp. 273–280. ISSN: 1521-0111.
- [39] H Vais et al. ‘Activation of *Drosophila* sodium channels promotes modification by deltamethrin. Reductions in affinity caused by knock-down resistance mutations’. In: *The Journal of General Physiology* 115.3 (Mar. 2000), pp. 305–318. ISSN: 0022-1295.
- [40] Deborah L. Capes et al. ‘Domain IV voltage-sensor movement is both sufficient and rate limiting for fast inactivation in sodium channels’. In: *The Journal of General Physiology* 142.2 (Aug. 2013), pp. 101–112. ISSN: 1540-7748.
- [41] Zhen Yan et al. ‘Structure of the Nav1.4-B1 Complex from Electric Eel’. In: *Cell* 170.3 (27th July 2017), 470–482.e11. ISSN: 0092-8674.
- [42] Altin Sula et al. ‘The complete structure of an activated open sodium channel’. In: *Nature Communications* 8 (16th Feb. 2017), p. 14205. ISSN: 2041-1723.
- [43] P N R Usherwood et al. ‘Mutations in DIIS5 and the DIIS4-S5 linker of *Drosophila melanogaster* sodium channel define binding domains for pyrethroids and DDT’. In: *FEBS Letters* 581.28 (27th Nov. 2007), pp. 5485–5492. ISSN: 0014-5793.

- [44] Andy Kilianski et al. ‘Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer.’ In: *GigaScience* 4 (2015), p. 12. ISSN: 2047-217X.
- [45] Eric R Lucas et al. ‘A high throughput multi-locus insecticide resistance marker panel for tracking resistance emergence and spread in *Anopheles gambiae*’. In: *Scientific reports* 9.1 (2019), pp. 1–10.
- [46] R A Holt et al. ‘The genome sequence of the malaria mosquito *Anopheles gambiae*’. In: *Science* 298.5591 (2002), pp. 129–149. ISSN: 0036-8075.
- [47] Pablo Cingolani et al. ‘A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3’. In: *Fly* 6.2 (2012), pp. 80–92. ISSN: 19336942.
- [48] Sudhir Kumar, Glen Stecher and Koichiro Tamura. ‘MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets’. In: *Molecular Biology and Evolution* 33.7 (2016), pp. 1870–1874. ISSN: 15371719.
- [49] Olivier Delaneau et al. ‘Haplotype estimation using sequencing reads’. In: *American Journal of Human Genetics* 93.4 (2013), pp. 687–696. ISSN: 00029297.
- [50] Androniki Menelaou and Jonathan Marchini. ‘Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold’. In: *Bioinformatics* 29.1 (2013), pp. 84–91. ISSN: 13674803.
- [51] R. C. Lewontin. ‘The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models’. In: *Genetics* 49.1 (1964), pp. 49–67. ISSN: 0016-6731.
- [52] Alistair Miles and Nicholas Harding. *scikit-allel: A Python package for exploring and analysing genetic variation data*. 2016.
- [53] F. Pedregosa et al. ‘Scikit-learn: Machine Learning in Python’. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [54] Fabian Sievers et al. ‘Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega’. In: *Molecular Systems Biology* 7 (2011), p. 539. ISSN: 1744-4292.

- 665 [55] Narayanan Eswar et al. ‘Comparative protein structure modeling using MODELLER’.
666 In: *Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et*
667 *Al.]* Chapter 2 (Nov. 2007), Unit 2.9. ISSN: 1934-3663.
- 668 [56] Leigh Willard et al. ‘VADAR: a web server for quantitative evaluation of protein
669 structure quality’. In: *Nucleic Acids Research* 31.13 (1st July 2003), pp. 3316–3319.

670 **Supplementary figures**

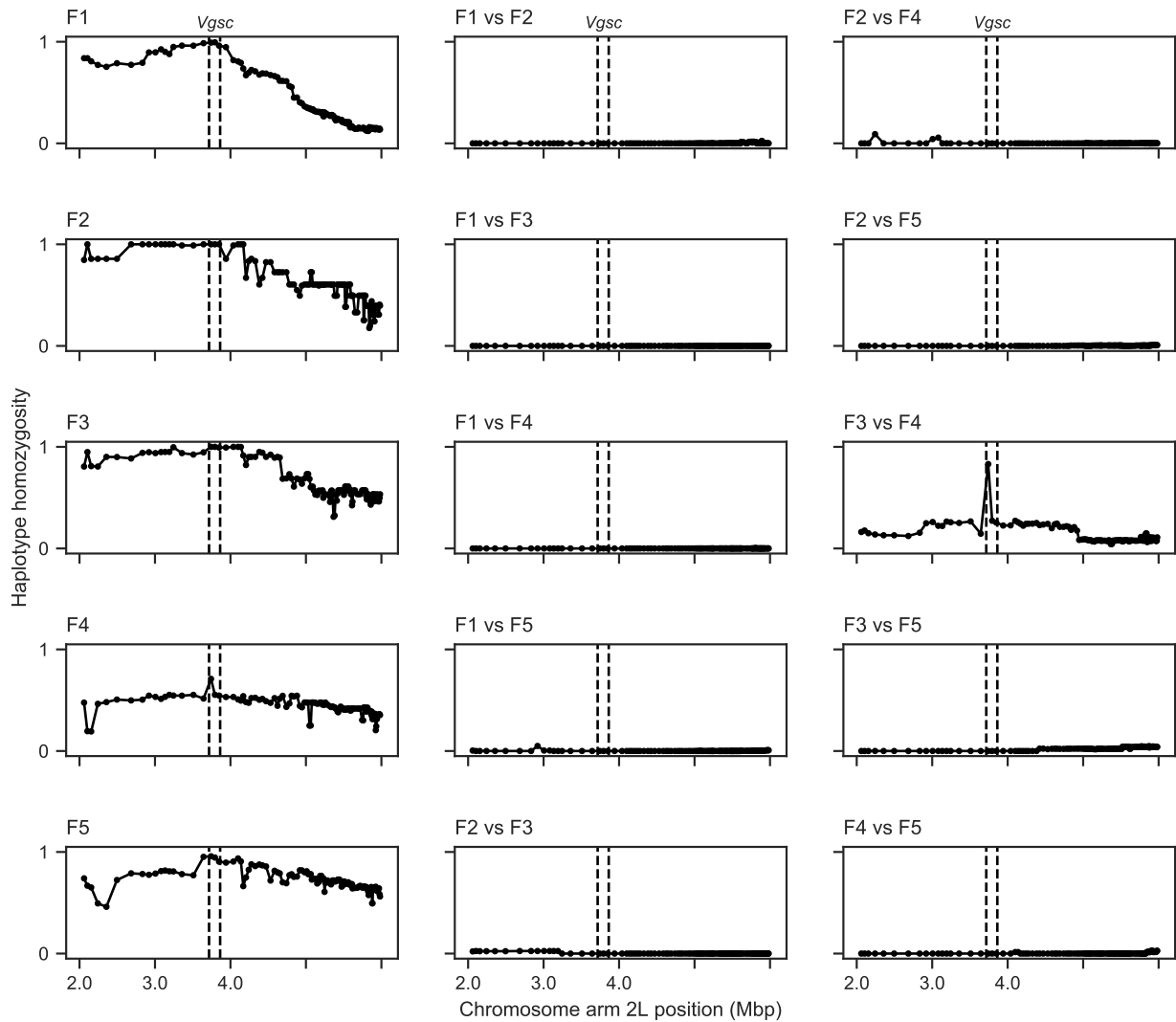


Figure S1. Windowed analysis of haplotype homozygosity for genetic backgrounds carrying the L995F allele. Each sub-plot shows the fraction of haplotype pairs that are identical within half-overlapping moving windows of 1000 SNPs. Each sub-plot in the left-hand column shows homozygosity for haplotype pairs within one of the haplotype groups identified by the network analysis. Sub-plots in the central and right-hand columns show homozygosity for haplotype pairs between two haplotype groups. If two haplotype groups are truly unrelated, haplotype homozygosity between them should be close to zero across the whole genome region. Dashed vertical lines show the location of the *Vgsc* gene.

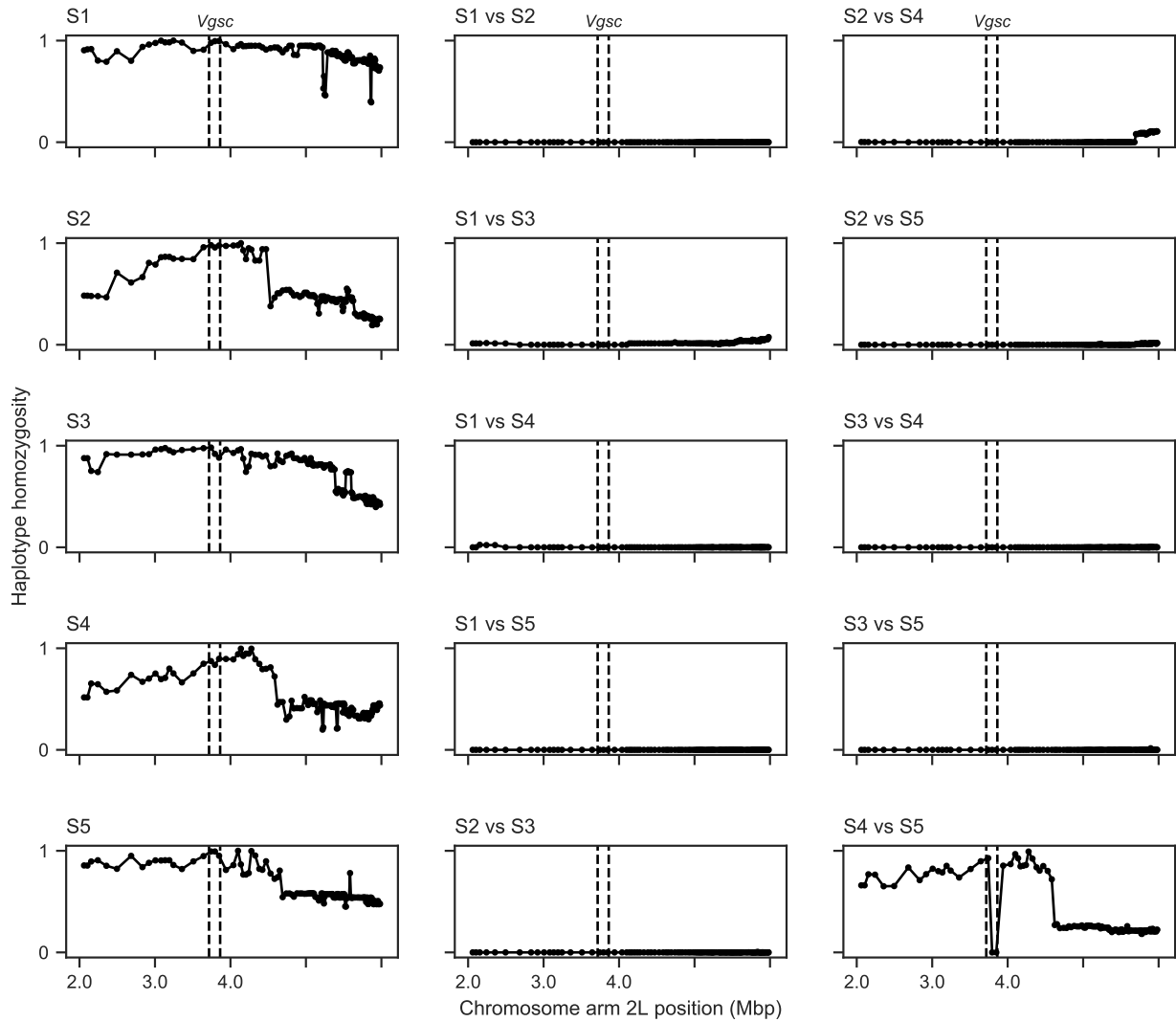


Figure S2. Windowed analysis of haplotype homozygosity for genetic backgrounds carrying the L995S allele. See Supplementary Figure S1 for explanation. Haplotype homozygosity is high between groups S4 and S5 on both flanks of the gene, indicating that haplotypes from both groups are in fact closely related.

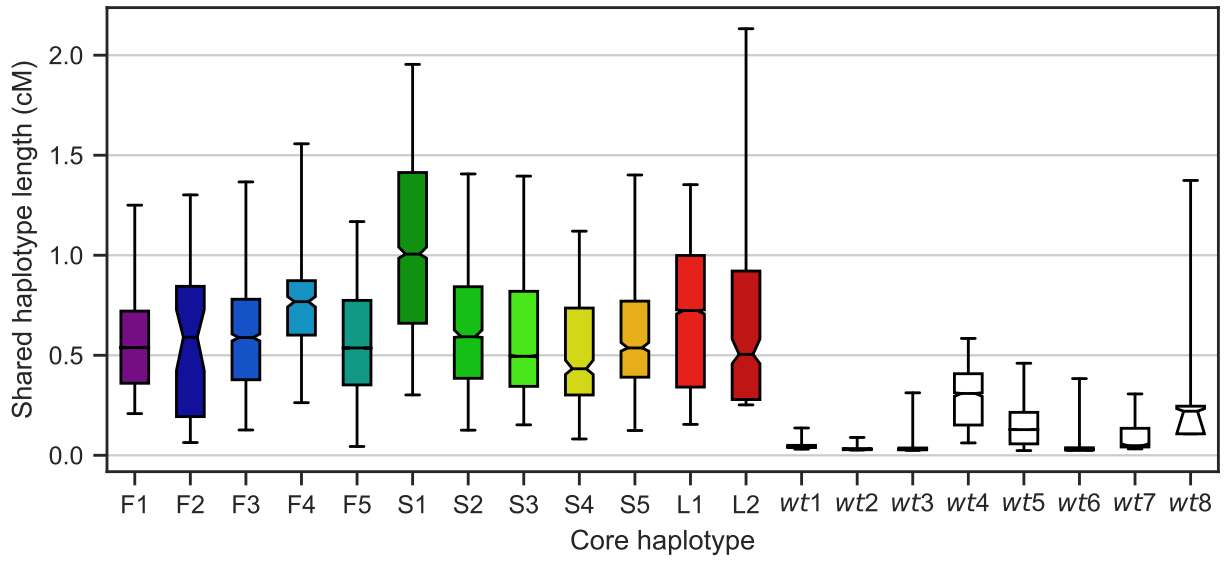


Figure S3. Shared haplotype length. Each bar shows the distribution of shared haplotype lengths between all pairs of haplotypes with the same core haplotype. For each pair of haplotypes, the shared haplotype length is computed as the region extending upstream and downstream from the core locus (*Vgsc* codon 995) over which haplotypes are identical at all non-singleton variants. The *Vgsc* gene sits on the border of pericentromeric heterochromatin and euchromatin, and we assume different recombination rates in upstream and downstream regions. The shared haplotype length is expressed in centiMorgans (cM) assuming a constant recombination rate of 2.0 cM/Mb on the downstream (euchromatin) flank and 0.6 cM/Mb on the upstream (heterochromatin) flank. Bars show the inter-quartile range, fliers show the 5-95th percentiles, horizontal black line shows the median, notch in bar shows the 95% bootstrap confidence interval for the median. Haplotypes F1-5 each carry the L995F resistance allele. Haplotypes S1-5 each carry the L995S resistance allele. Haplotype L1 carries the I1527T allele. Haplotype L2 carries the M490I allele. Wild-type (*wt*) haplotypes do not carry any known or putative resistance alleles.