

1           **The genetic architecture of target-site**  
2           **resistance to pyrethroid insecticides in the**  
3           **African malaria vectors *Anopheles gambiae***  
4           **and *Anopheles coluzzii***

5           Chris S. Clarkson<sup>1,\*</sup>, Alistair Miles<sup>2,1,\*</sup>, Nicholas J. Harding<sup>2</sup>, Andrias O. O'Reilly<sup>3</sup>, David  
6           Weetman<sup>4</sup>, Dominic Kwiatkowski<sup>1,2</sup>, Martin J. Donnelly<sup>4,1</sup>, and The *Anopheles gambiae* 1000  
7           Genomes Consortium<sup>5</sup>

8           <sup>1</sup>Wellcome Sanger Institute, Hinxton, Cambridge CB10 1SA

9           <sup>2</sup>Big Data Institute, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Old  
10           Road Campus, Oxford OX3 7LF

11           <sup>3</sup>Liverpool John Moores University, Brownlow Hill, Liverpool L3 5UG

12           <sup>4</sup>Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA

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

14           \*These authors contributed equally. Correspondence: chris.clarkson@sanger.ac.uk

15           11th February 2021

16           **Abstract**

17           Resistance to pyrethroid insecticides is a major concern for malaria vector control.  
18           Pyrethroids target the voltage-gated sodium channel (VGSC), an essential compo-  
19           nent of the mosquito nervous system. Substitutions in the amino acid sequence can  
20           inducing a resistance phenotype. We use whole-genome sequence data from phase 2 of  
21           the *Anopheles gambiae* 1000 Genomes Project (Ag1000G) to provide a comprehensive  
22           account of genetic variation in the *Vgsc* gene across 13 African countries. In addition  
23           to known resistance alleles, we describe 20 other non-synonymous nucleotide substi-  
24           tutions at appreciable population frequency, and map these variants onto a protein

model to investigate the likelihood of a pyrethroid resistance phenotypes. Thirteen of these novel alleles were found to occur almost exclusively on haplotypes carrying the known L995F *kdr* (knock-down resistance allele) and may enhance or compensate for the L995F resistance genotype. A novel mutation I1527T, adjacent to a predicted pyrethroid binding site, was found in tight linkage with V402L substitutions, similar to combinations associated with resistance in other insect species. We also analysed genetic backgrounds carrying resistance alleles, to determine which alleles have experienced recent positive selection, and describe ten distinct haplotype groups carrying known *kdr* resistance alleles. Five of these groups are observed in more than one country, in one case separated by over 3000 km, providing new information about the potential for the geographical spread of resistance. 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.

## 39 **Introduction**

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

47 The voltage-gated sodium channel (VGSC) is the physiological target of pyrethroid in-  
48 secticides, and is integral to the insect nervous system. The sodium channel protein con-  
49 sists of four homologous domains (DI-IV) each of which comprises six transmembrane seg-  
50 ments (S1-S6) connected by intracellular and extracellular loops [5]. Pyrethroid molecules  
51 bind to this protein, stabilise the ion-conducting active state, and thus disrupt normal  
52 nervous system function, producing paralysis (“knock-down”) and death. However, amino  
53 acid substitutions at key positions within the protein alter the interaction with insecticide  
54 molecules, increasing the dose of insecticide required for knock-down, known as knock-  
55 down resistance or *kdr* [6, 5].

56 In the African malaria vectors *Anopheles gambiae* and *An. coluzzii*, three substitutions  
57 have been found to cause pyrethroid resistance. Two of these substitutions occur in codon  
58 995<sup>1</sup>, with L995F prevalent in West and Central Africa [7, 8], and L995S found in Cen-  
59 tral and East Africa [9, 8]. A third substitution, N1570Y, has been found in West and  
60 Central Africa and shown to increase resistance in association with L995F [11]. However,  
61 studies in other insect species have found a variety of other *Vgsc* substitutions inducing  
62 a resistance phenotype [12, 13, 5]. To our knowledge, no studies in malaria vectors have  
63 analysed genetic variation across the full *Vgsc* coding sequence, thus the molecular basis  
64 of pyrethroid target-site resistance has not been fully explored.

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

77 Here we report an in-depth analysis of genetic variation in the *Vgsc* gene, using whole-  
78 genome Illumina sequence data from phase 2 of the *Anopheles gambiae* 1000 Genomes  
79 Project (Ag1000G) [22]. The Ag1000G phase 2 resource includes data on nucleotide vari-  
80 ation in 1,142 wild-caught mosquitoes sampled from 13 countries, with representation of  
81 West, Central, Southern and East Africa, and of both *An. gambiae* and *An. coluzzii*.  
82 We investigate variation across the complete gene coding sequence, and report popula-  
83 tion genetic data for both known and novel non-synonymous nucleotide substitutions. We  
84 then use haplotype data from the chromosomal region spanning the *Vgsc* gene to study

---

<sup>1</sup>Codon 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.

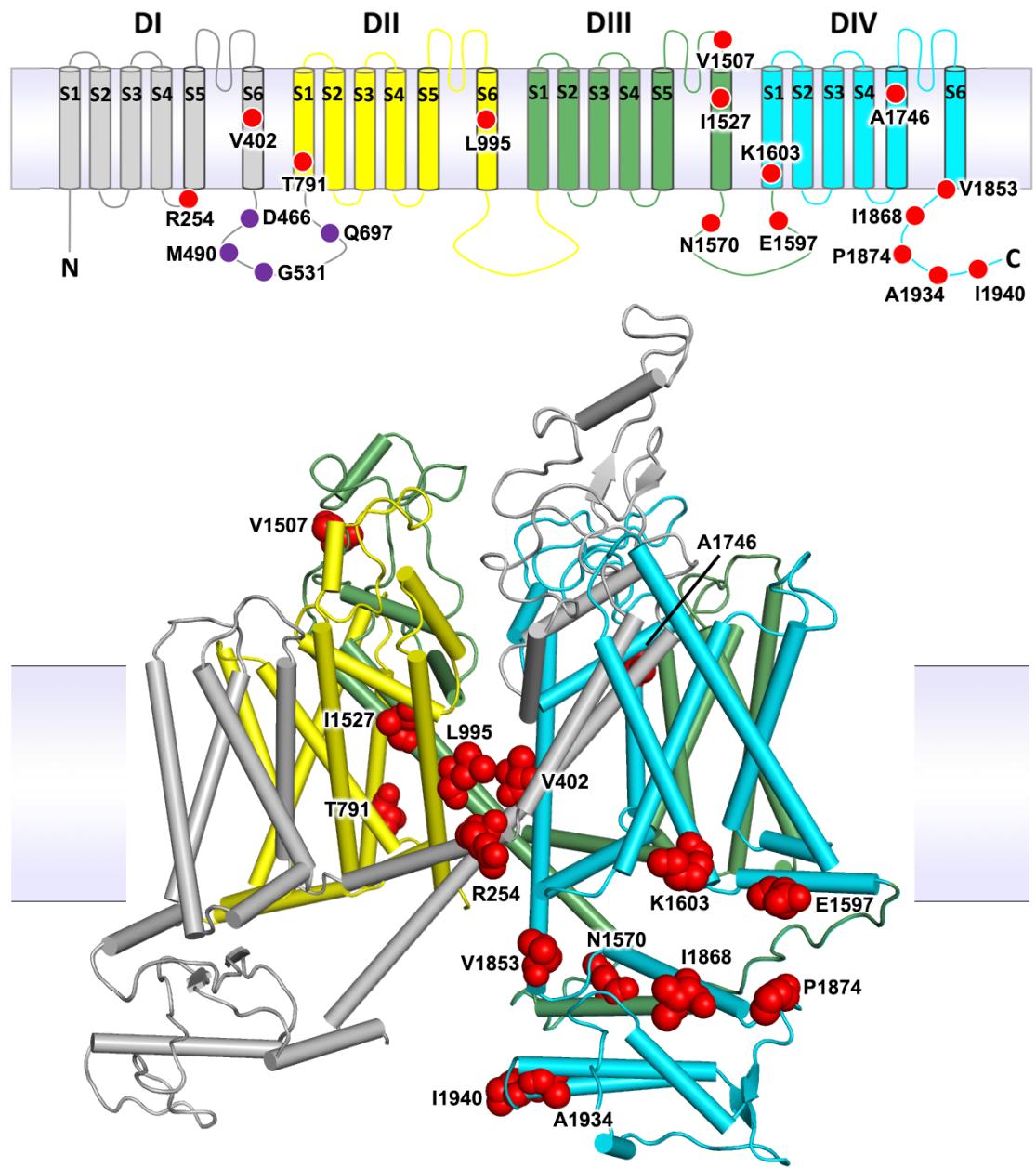
85 the genetic backgrounds carrying resistance alleles, investigate the geographical spread  
86 of resistance between mosquito populations, and provide evidence for recent positive se-  
87 lection. Finally, we explore ways in which variation data from Ag1000G can be used to  
88 design high-throughput, low-cost genetic assays for surveillance of pyrethroid resistance,  
89 with the capability to differentiate and track resistance outbreaks.

90 **Results**

91 **Vgsc non-synonymous nucleotide variation**

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

105 The 23 non-synonymous variants were located on a transmembrane topology map and on  
106 a 3-dimensional homology model of the *Vgsc* protein. (Figure 1). The substitutions were  
107 found to be distributed throughout the channel, in all of the four internally homologous  
108 domains (DI-DIV), in S1, S5 and S6 membrane-spanning segments, in two of the intracel-  
109 lular loops connecting domains, and in the C-terminal tail. The S5 and S6 segments that  
110 form the central ion-conducting pore of the channel carry six of the eight segment substi-  
111 tutions, including V402 and L995 which have been shown to produce insecticide resistance  
112 phenotypes [6, 5, 7, 8, 9]. Two substitutions are located on the DIII-DIV linker including  
113 the resistance conferring N1570 [11]. A further six substitutions are found concentrated



**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.

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

118 intracellular linker, but this region is missing from the model as it was not resolved in the  
119 cockroach Na<sub>v</sub>PaS structure used as the model template [25].

120 The two known resistance alleles affecting codon 995 had the highest overall allele fre-  
121 quencies within the Ag1000G phase 2 cohort (Table 1). The L995F allele was at high  
122 frequency in populations of both species from West, Central and Southern Africa. The  
123 L995S allele was at high frequency among *An. gambiae* populations from Central and  
124 East Africa. Both of these alleles were present in *An. gambiae* populations sampled from  
125 Cameroon and Gabon. This included individuals with a heterozygous L995F/S genotype  
126 (50/297 individuals in Cameroon, 41/69 in Gabon). We calculated empirical p-values for  
127 these heterozygous genotype counts using the Dirichlet distribution and 1,000,000 Monte  
128 Carlo simulations. In Cameroon p=0.410 of simulations found higher proportions of het-  
129 erozygous genotypes, however in Gabon this dropped to p=0.005, suggesting there may  
130 be a fitness advantage for mosquitoes carrying both alleles in some circumstances.

131 The N1570Y allele was present in Guinea *An. gambiae*, Ghana *An. gambiae*, Burkina  
132 Faso (both species) and Cameroon *An. gambiae*. This allele has been shown to sub-  
133 stantially increase pyrethroid resistance when it occurs in combination with L995F, both  
134 in association tests of phenotyped field samples [11] and functional tests using *Xenopus*  
135 oocytes [26]. To study the patterns of association among non-synonymous variants, we  
136 used haplotypes from the Ag1000G phase 2 resource to compute the normalised coefficient  
137 of linkage disequilibrium ( $D'$ ) between all pairs of variant alleles (Figure 2). As expected,  
138 we found N1570Y in almost perfect linkage with L995F. Of the 20 novel non-synonymous  
139 alleles, 13 also occurred almost exclusively in combination with L995F (Figure 2). These  
140 included two variants in codon 1874 (P1874S, P1874L), one of which (P1874S) has previ-  
141 ously been associated with pyrethroid resistance in the crop pest moth *Plutella xylostella*  
142 [24].

143 The abundance of high-frequency non-synonymous variants occurring in combination  
144 with L995F is notable for two reasons. First, *Vgsc* is a highly conserved gene, expected  
145 to be under strong functional constraint and therefore purifying selection, so any non-  
146 synonymous variants are expected to be rare [12]. Second, in contrast with L995F, we did  
147 not observe any high-frequency non-synonymous variants occurring in combination with  
148 L995S. This contrast was clear when data on all variants within the gene were considered:

**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).

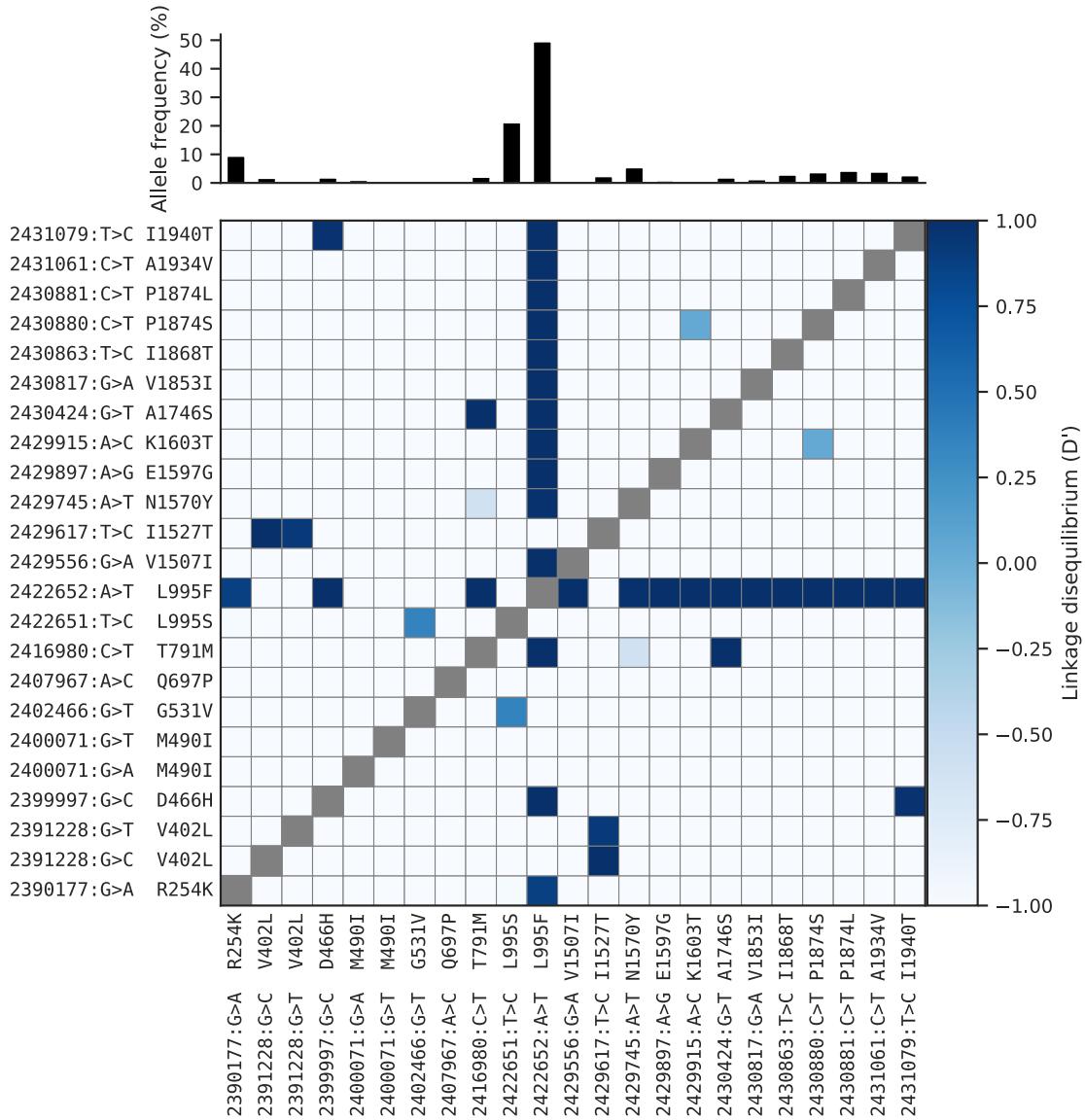
Position <sup>1</sup>	Variant			Population allele frequency (%)															
	Ag <sup>2</sup>	Md <sup>3</sup>	Domain <sup>4</sup>	AOAc	GHAc	BFAc	CIAc	GNAc	GW	GM	CMAg	GHAg	BFAg	GNAg	GAAg	UGAg	GQAg	FRAg	KE
2,390,177 G>A	R254K	R261	IL45	0.0	0.009	0.0	0.0	0.0	0.0	0.0	0.313	0.0	0.0	0.0	0.203	0.0	0.0	0.0	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	IIIS6	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	LIIV/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	LIIV/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	0.0	0.234	0.475	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

<sup>1</sup> Position relative to the AgamP3 reference sequence, chromosome arm 2L.

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

<sup>3</sup> Codon numbering according to *Musca domestica* EMBL accession X96668 [10].

<sup>4</sup> 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.



**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.

for haplotypes carrying the L995F allele, the ratio of non-synonymous to synonymous nucleotide diversity  $\pi_N/\pi_S$  was 20.04 times higher than haplotypes carrying the wild-type allele, but for those carrying L995S  $\pi_N/\pi_S$  was 0.5 times lower than haplotypes carrying the wild-type allele. These results indicate that L995F has substantially altered the selective regime for other amino acid positions within the protein. Secondary substitutions have occurred and risen in frequency, suggesting that they are providing some further selective advantage in the presence of insecticide pressure.

A novel allele, I1527T, was present in *An. coluzzii* from Ghana, Burkina Faso, Côte d'Ivoire and Guinea. Codon 1527 occurs within trans-membrane segment IIIIS6, imme-

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

## 172 **Genetic backgrounds carrying resistance alleles**

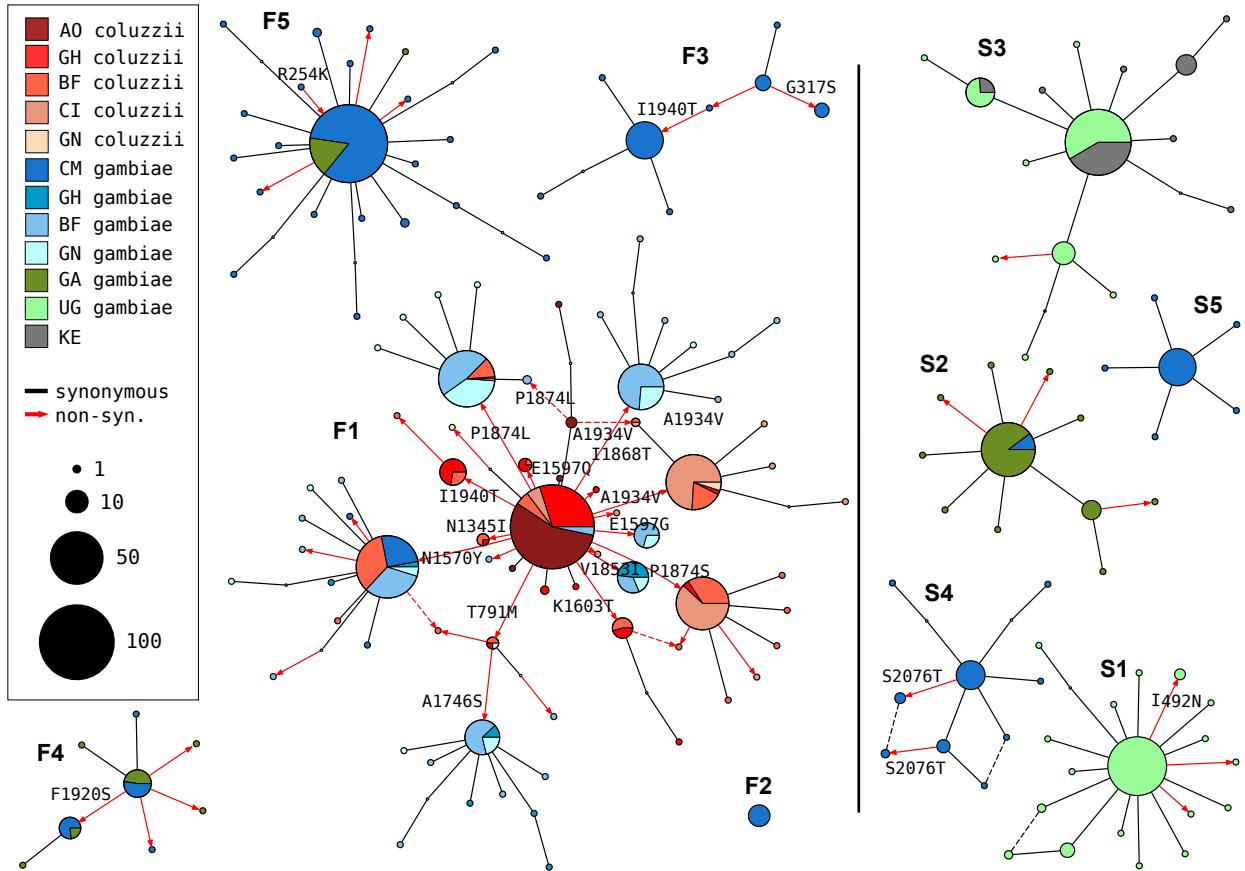
173 The Ag1000G data resource provides a rich source of information about the spread of  
174 insecticide resistance alleles in any given gene, because data are not only available for  
175 SNPs in protein coding regions, but also SNPs in introns, flanking intergenic regions,  
176 and in neighbouring genes. These additional variants can be used to analyse the genetic  
177 backgrounds (haplotypes) on which resistance alleles are found. In our initial report of  
178 the Ag1000G phase 1 resource [23], we used 1710 biallelic SNPs from within the 73.5 kbp  
179 *Vgsc* gene (1607 intronic, 103 exonic) to compute the number of SNP differences between  
180 all pairs of 1530 haplotypes derived from 765 wild-caught mosquitoes. We then used  
181 pairwise genetic distances to perform hierarchical clustering, and found that haplotypes  
182 carrying resistance alleles in codon 995 were grouped into 10 distinct clusters, each with  
183 near-identical haplotypes. Five of these clusters contained haplotypes carrying the L995F  
184 allele (labelled F1-F5), and a further five clusters contained haplotypes carrying L995S  
185 (labelled S1-S5).

186 To further investigate genetic backgrounds carrying resistance alleles, we used the  
187 Ag1000G phase 2 haplotype data from the *Vgsc* gene (2,284 haplotypes from 1,142 mosquitoes

[22]), to construct median-joining networks [29] (Figure 3). The network analysis improves on hierarchical clustering by allowing for the reconstruction and placement of intermediate haplotypes that may not be observed in the data. It also allows for non-hierarchical relationships between haplotypes, which may arise if recombination events have occurred between haplotypes. We constructed the network up to a maximum edge distance of 2 SNP differences, to ensure that each connected component captures a group of closely-related haplotypes. The resulting network contained 5 groups containing haplotypes carrying L995F, and a further 5 groups carrying L995S, in close correspondence with previous results from hierarchical clustering (96.8% overall concordance in assignment of haplotypes to groups).

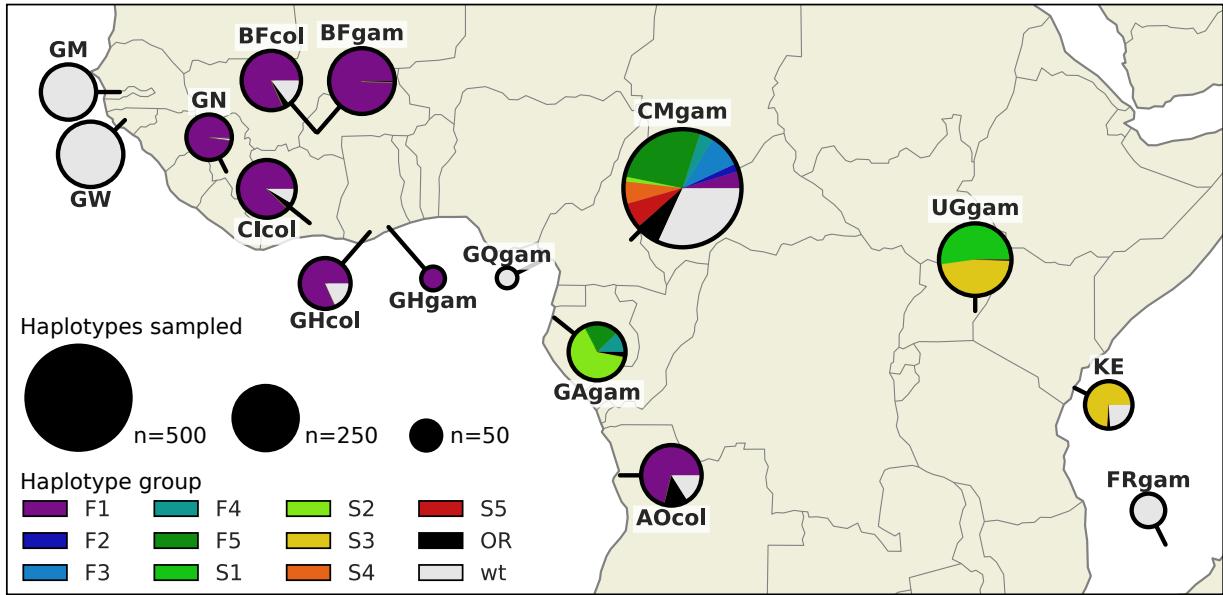
The haplotype network brings into sharp relief the explosive radiation of amino acid substitutions secondary to the L995F allele (Figure 3). Within the F1 group, nodes carrying non-synonymous variants radiate out from a central node carrying only L995F, suggesting that the central node represents the ancestral haplotype carrying just L995F which initially came under selection, and these secondary variants have arisen subsequently as new mutations. In F1 alone, 30 network edges (shown as red arrows - Figure 3) lead to non-synonymous nodes. Many of the nodes carrying secondary variants are large, consistent with positive selection and a functional role for these secondary variants as modifiers of the L995F resistance phenotype. The F1 network also allows us to infer multiple introgression events between the two species. The central (putatively ancestral) node contains haplotypes from individuals of both species, as do nodes carrying the N1570Y, P1874L and T791M variants. This structure is consistent with an initial introgression of the ancestral F1 haplotype, followed later by introgressions of haplotypes carrying secondary mutations. The haplotype network also illustrates the contrasting levels of non-synonymous variation between L995F and L995S. Within all of the L995S groups, only eight edges lead to non-synonymous nodes and all these nodes are small (low frequency variants), thus may be neutral or mildly deleterious variants that are hitch-hiking on selective sweeps for the L995S allele.

The F1 group contains haplotypes from mosquitoes of both species, and from mosquitoes 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



**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.

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



**Figure 4. Map of haplotype frequencies.** Each pie shows the frequency of different haplotype groups within one of the populations sampled. The size of the pie is proportional to the number of haplotypes sampled. The size of each wedge within the pie is proportional to the frequency of a haplotype group within the population. Haplotypes in groups F1-5 carry the L995F *kdr* allele. Haplotypes in groups S1-5 carry the L995S *kdr* allele. Haplotypes in group other resistant (OR) carry either L995F or L995S but did not cluster within any of the haplotype groups. Wild-type (*wt*) haplotypes do not carry any known resistance alleles.

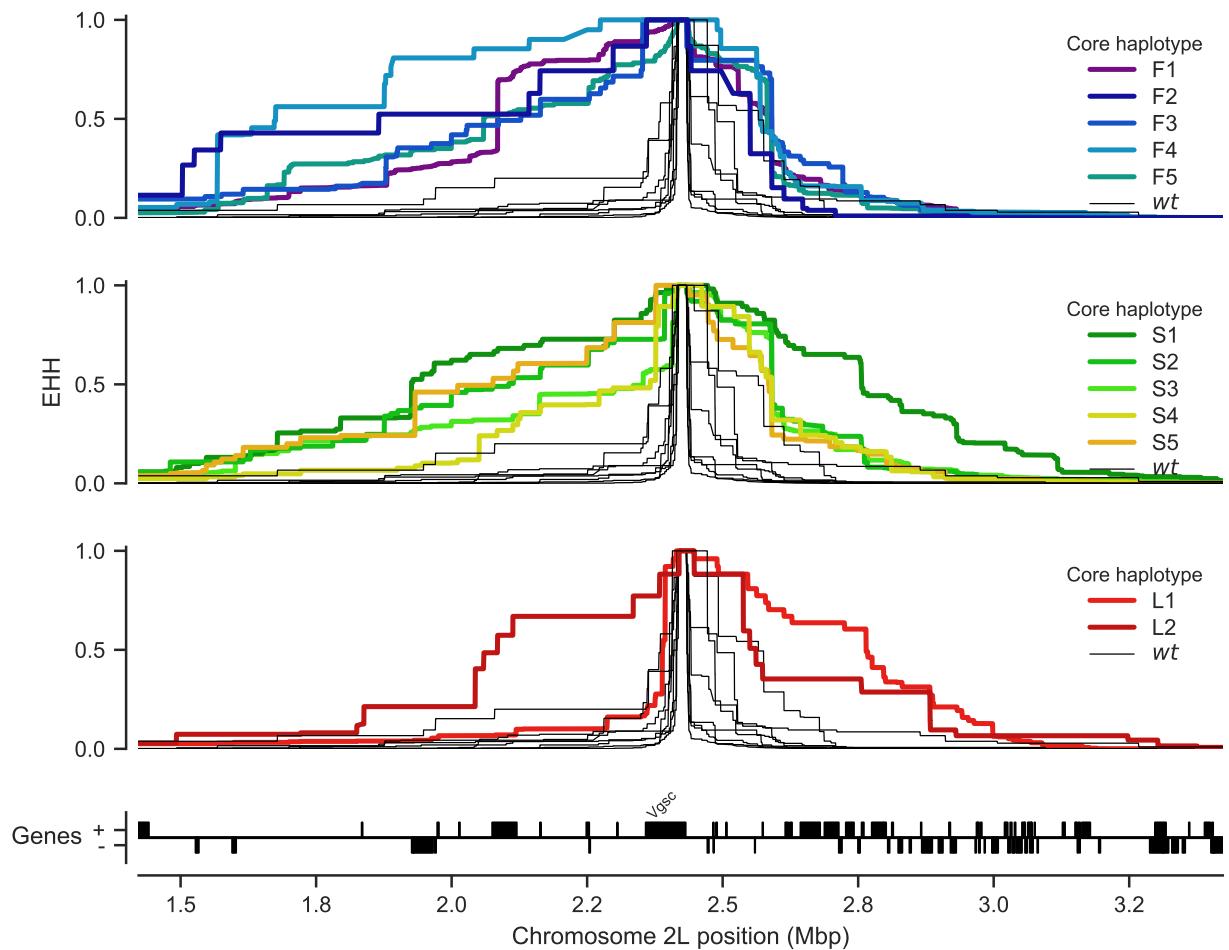
228 Populations carrying *kdr* alleles were collected between the years 2009 and 2012, with the  
 229 exception of Gabon, which was collected in 2000. This temporal spread allows, albeit with  
 230 low-resolution, tracking of haplotypes through time. The spatially widespread F1 group  
 231 contains haplotypes from samples collected between 2009-2012 (Figure 4, [22]) (Figure  
 232 3). We still do not know how fast insecticide resistance alleles can travel between these  
 233 countries, but the large geographic spread suggests the F1 haplotype group originated some  
 234 considerable time before the earliest collection in 2009. Haplotype groups F4, F5 and S2,  
 235 all carry haplotypes from samples collected in Cameroon (2009) and Gabon (2000). These  
 236 observations demonstrate that, even in mosquito populations with high levels of genetic  
 237 diversity and large effective population size [23], nucleotide sequences carrying alleles under  
 238 strong selection can persist unchanged for almost a decade.

239 A limitation of both the hierarchical clustering and network analyses is that they rely  
 240 on genetic distances within a fixed genomic window from the start to the end of the  
 241 *Vgsc* gene. *Anopheles* mosquitoes undergo homologous recombination during meiosis in  
 242 both males and females, and any recombination events that occurred within this genomic

243 window could affect the way that haplotypes are grouped together in clusters or network  
244 components. In particular, recombination events could occur during the geographical  
245 spread of a resistance allele, altering the genetic background upstream and/or downstream  
246 of the allele itself. An analysis based on a fixed genomic window might then fail to infer  
247 gene flow between two mosquito populations, because haplotypes with and without a  
248 recombination event could be grouped separately, despite the fact that they share a recent  
249 common ancestor. To investigate the possibility that recombination events may have  
250 affected our grouping of haplotypes carrying resistance alleles, we performed a moving  
251 window analysis of haplotype homozygosity, spanning *Vgsc* and up to a megabase upstream  
252 and downstream of the gene (Supplementary Figures 1 and 2). This analysis supported a  
253 refinement of our initial grouping of haplotypes carrying resistance alleles. All haplotypes  
254 within groups S4 and S5 were effectively identical on both the upstream and downstream  
255 flanks of the gene, but there was a region of divergence within the *Vgsc* gene itself that  
256 separated them in the fixed window analyses (Supplementary Figure 3). The 13.8 kbp  
257 region of divergence occurred upstream of codon 995 and contained 6 SNPs that were fixed  
258 differences between S4 and S5. A possible explanation for this short region of divergence  
259 is that a gene conversion event has occurred within the gene, bringing a segment from  
260 a different genetic background onto the original genetic background on which the L995S  
261 resistance mutation occurred.

## 262 Positive selection for resistance alleles

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



**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).

we labelled L1 (due to it carrying the the wild-type leucine codon at position 995). We also 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

282 [16]. However, we found no systematic difference in the length of shared haplotypes when  
283 comparing F1-5 (carrying L995F) against S1-5 (carrying L995S) (Supplementary Figure  
284 3). There were, however, some differences between core haplotypes carrying the same  
285 allele. For example, shared haplotypes were significantly longer for S1 (median 1.006 cM,  
286 95% CI [0.986 - 1.040]) versus other core haplotypes carrying L995S (e.g., S2 median  
287 0.593 cM, 95% CI [0.589 - 0.623]; Supplementary Figure 3). Longer shared haplotypes  
288 indicate a more recent common ancestor, and thus some of these core haplotypes may  
289 have experienced more recent and/or more intense selection than others.

290 As sample collections took place over 12 years (2000-2012), it might be expected that  
291 core haplotypes appearing earlier in our sampling would have smaller shared haplotypes  
292 due to increased opportunity for recombination and mutation. However, no correlation  
293 was found between the year a core haplotype was first detected and the median length  
294 ( $r(8)=0.03$ ,  $p=0.93$ , Supplementary Figure 3).

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

308 **Discussion**

309 **Cross-resistance between pyrethroids and DDT**

310 The VGSC protein is the physiological target of both pyrethroid insecticides and DDT [6].  
311 The L995F and L995S alleles are known to increase resistance to both of these insecticide  
312 classes [7, 9]. By 2012, over half of African households owned at least one pyrethroid  
313 impregnated ITN and nearly two thirds of IRS programmes were using pyrethroids [2].  
314 Pyrethroids were also introduced into agriculture in Africa prior to the scale-up of public  
315 health vector control programmes, and continue to be used on a variety of crops such as  
316 cotton [31]. DDT was used in Africa for several pilot IRS projects carried out during the  
317 first global campaign to eradicate malaria, during the 1950s and 1960s [12]. DDT is still  
318 approved for IRS use by WHO and remains in use in some locations, however within the  
319 last two decades pyrethroid use has been far more common and widespread. DDT was also  
320 used in agriculture from the 1940s, and although agricultural usage has greatly diminished  
321 since the 1970s, some usage remains [32]. In this study we reported evidence of positive  
322 selection on the L995F and L995S alleles, as well as the I1527T+V402L combination and  
323 possibly M490I. We also found 14 other non-synonymous substitutions that have arisen in  
324 association with L995F and appear to be positively selected. Given that pyrethroids have  
325 dominated public health insecticide use for two decades, it is reasonable to assume that the  
326 selection pressure on these alleles is primarily due to pyrethroids rather than DDT. It has  
327 previously been suggested that L995S may have been initially selected by DDT usage [16].  
328 However, we did not find any systematic difference in the extent of haplotype homozygosity  
329 between these two alleles, suggesting that both alleles have been under selection over a  
330 similar time frame. We did find some significant differences in haplotype homozygosity  
331 between different genetic backgrounds carrying resistance alleles, suggesting differences  
332 in the timing and/or strength of selection these may have experienced. However, there  
333 have been differences in the scale-up of pyrethroid-based interventions in different regions,  
334 and this could in turn generate heterogeneities in selection pressures. Nevertheless, it is  
335 possible that some if not all of the alleles we have reported provide some level of cross-  
336 resistance to DDT as well as pyrethroids, and we cannot exclude the possibility that  
337 earlier DDT usage may have contributed at least in part to their selection. The differing

338 of resistance profiles to the two types of pyrethroids (type I, e.g., permethrin; and type  
339 II, e.g., deltamethrin) [33], will also affect the selection landscape. Further sampling and  
340 analysis will be required to investigate the timing of different selection events and relate  
341 these to historical patterns of insecticide use in different regions.

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

343 The non-synonymous variants are distributed throughout the channel protein but can  
344 be considered in terms of three clusters: (i) the transmembrane domain, (ii) the DI-II  
345 intracellular linker and (iii) the DIII-DIV/C-terminal subdomain. The pyrethroid binding  
346 site is located in the transmembrane domain between the IIS4-S5 linker and the IIS5,  
347 IIS6 and IIIS6 helices [34]. The I1527T substitution that we discovered in *An. coluzzii*  
348 mosquitoes from Burkina Faso occurs in segment IIIS6 and is immediately adjacent to  
349 two pyrethroid-sensing residues in this binding site [5]. It is thus plausible that pyrethroid  
350 binding could be altered by this substitution. The I1527T substitution (*M. domestica*  
351 codon 1532) has been found in *Aedes albopictus* [35], and substitutions in the nearby  
352 codon 1529 (*M. domestica* I1534T) have been reported in *Aedes albopictus* and in *Aedes*  
353 *aegypti* where it was found to be associated with pyrethroid resistance [5, 36, 37]. We  
354 found the I1527T allele in tight linkage with two alleles causing a V402L substitution (*M.*  
355 *domestica* V410L). Substitutions in codon 402 have been found in multiple insect species  
356 and are by themselves sufficient to confer pyrethroid resistance [5]. The fact that we  
357 find I1527T and V402L in such tight mutual association is intriguing because haplotypes  
358 carrying V402L alone should also have been positively selected and thus be present in one  
359 or more populations.

360 The V402 residue is located towards the middle of the IS6 helix. The L995F and L995S  
361 substitutions occur at a similar position on the IIS6 helix. It was proposed these S6 sub-  
362 stitutions confer resistance by allosterically modifying formation of the pyrethroid binding  
363 site [34]. More recently the L995 *kdr* residue was speculated to form part of a second  
364 pyrethroid binding site in the insect channel termed 'PyR2' [27, 38]. A major functional  
365 effect of the L995F substitution is enhanced closed-state inactivation [39]. This contributes  
366 to *kdr* resistance by reducing the number of channels that undergo activation, which is  
367 the functional state that pyrethroids bind to with highest affinity [39]. Fast inactivation

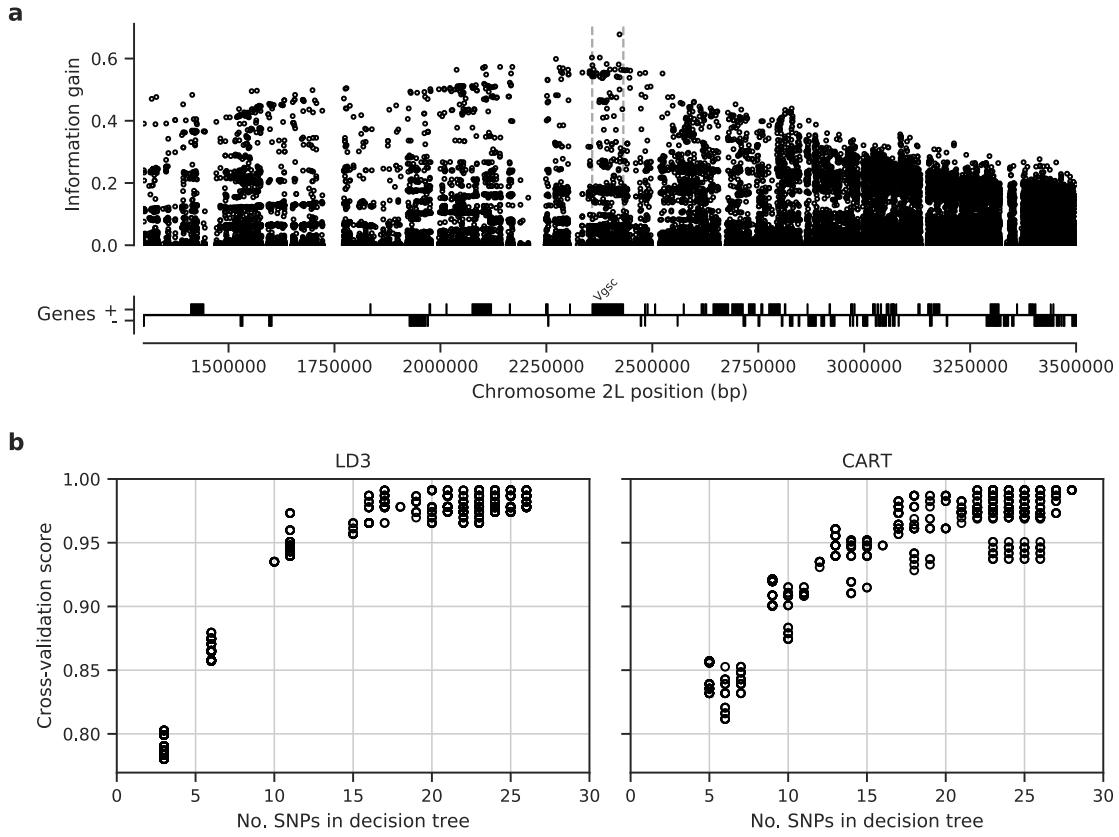
368 involves movement of the DIV domain to form a receptor for the DIII-DIV linker fast in-  
369 activation particle containing the 'MFM' sequence motif (equivalent to the 'IFM' motif in  
370 mammals) [40, 5]. Recent eukaryotic sodium channel structures reveal that the DIII-DIV  
371 linker is in complex with the C-terminal segment in the closed-state conformation but the  
372 DIII-DIV linker appears to dissociate and bind in close proximity in the DIV S6 helix upon  
373 transition to the inactivated state [25, 41]. It seems that binding of the DIII-DIV linker  
374 pushes the DIV S6 helix forward to occlude the pore and produce the inactivated state  
375 [41]. We suggest that substitutions located on the DIII-DIV linker and C-terminal tail may  
376 perturb the conformation of this subdomain when it assembles in the closed-state channel  
377 and may subsequently affect capture or release of the DIII-DIV linker from this complex.  
378 The expected functional outcome would be altered channel inactivation, although whether  
379 inactivation is enhanced or diminished and if this compensates for a deleterious effect of  
380 L995F on channel function awaits elucidation. The N1570Y substitution on the DIII-DIV  
381 linker has been functionally characterised but inactivation kinetics in the mutant channel  
382 were found unaltered [26]. Pyrethroid sensitivity was also unaffected by N1570Y although  
383 resistance was greatly enhanced in the N1570Y + L995F double mutant [26].

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

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

397 Entomological surveillance teams in Africa regularly genotype mosquitoes for resistance al-  
398 leles in *Vgsc* codon 995, and use those results as an indicator for the presence of pyrethroid  
399 resistance alongside results from insecticide resistance bioassays. They typically do not,  
400 however, sequence the gene or genotype any other polymorphisms within the gene. Thus,  
401 if there are other polymorphisms within the gene that cause or significantly enhance  
402 pyrethroid resistance, these will not be detected. Also, if a codon 995 resistance allele is  
403 observed, there is no way to know whether the allele is on a genetic background that has  
404 also been observed in other mosquito populations, and thus no way to investigate whether  
405 resistance alleles are emerging locally or being imported from elsewhere. Whole-genome  
406 sequencing of individual mosquitoes clearly provides data of sufficient resolution to answer  
407 these questions, and could be used to provide ongoing resistance surveillance. The cost of  
408 whole-genome sequencing continues to fall, making it a practical tool for malaria vector  
409 surveillance. However, to achieve substantial spatial and temporal coverage of mosquito  
410 populations, it would also be necessary to develop targeted genetic assays for resistance  
411 outbreak surveillance. Technologies such as amplicon sequencing [44] are already being  
412 trialled on mosquitoes [45], these could scale to tens of thousands of samples at low cost  
413 and could be implemented using existing platforms in national molecular biology facilities.

414 To facilitate the development of targeted genetic assays for surveillance of *Vgsc*-mediated  
415 pyrethroid resistance, we have produced several supplementary data tables. In Supple-  
416 mentary Table 1 we list all 82 non-synonymous variants found within the *Vgsc* gene in this  
417 study, with population allele frequencies. In Supplementary Table 2 we list 756 biallelic  
418 SNPs, within the *Vgsc* gene and up to 10 kbp upstream or downstream, that are poten-  
419 tially informative regarding which haplotype group a resistance haplotype belongs to, and  
420 thus could be used for tracking the spread of resistance. This table includes the allele  
421 frequency within each of the 10 haplotype groups defined here, to aid in identifying SNPs  
422 that are highly differentiated between two or more haplotype groups. We also provide  
423 Supplementary Table 3 which lists all 10,244 SNPs found within the *Vgsc* gene and up to  
424 10 kbp upstream or downstream, which might need to be taken into account as flanking  
425 variation when searching for PCR primers to amplify a SNP of interest. To provide some



**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.

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

437 resistance outbreaks.

## 438 Methods

### 439 Code

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

### 442 Data

443 We used variant calls and phased haplotype data from the Ag1000G Phase 2 AR1 data re-  
444 lease (<https://www.malariagen.net/data/ag1000g-phase-2-ar1>). Variant calls from  
445 Ag1000G Phase 2 are also available from the European Nucleotide Archive (ENA; <http://www.ebi.ac.uk/ena>) under study PRJEB36277.  
446

### 447 Data collection and processing

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

462 The Ag1000G variant data was functionally annotated using the SnpEff v4.1b software  
463 [47]. Non-synonymous *Vgsc* variants were identified as all variants in AgamP4.12 transcript

464 AGAP004707-RD with a SnpEff annotation of “missense”. The *Vgsc* gene is known to  
465 exhibit alternative splicing [6], however at the time of writing the *An. gambiae* gene  
466 annotations did not include the alternative transcripts reported by Davies et al. We wrote  
467 a Python script to check for the presence of variants that are synonymous according to  
468 transcript AGAP004707-RD but non-synonymous according to one of the other transcripts  
469 present in the gene annotations or in the set reported by Davies et al. Supplementary Table  
470 1 includes the predicted effect for all SNPs that are non-synonymous in one or more of  
471 these transcripts. None of the variants that are non-synonymous in a transcript other  
472 than AGAP004707-RD were found to be above 5% frequency in any population.

473 For ease of comparison with previous work on *Vgsc*, pan Insecta, in Table 1 and Supple-  
474 mentary Table 1 we report codon numbering for both *An. gambiae* and *Musca domestica*  
475 (the species in which the gene was first discovered). The *M. domestica* *Vgsc* sequence  
476 (EMBL accession X96668 [10]) was aligned with the *An. gambiae* AGAP004707-RD se-  
477 quence (AgamP4.12 gene-set) using the Mega v7 software package [48]. A map of equiva-  
478 lent codon numbers between the two species for the entire gene can be download from the  
479 MalariaGEN website ([https://www.malariagen.net/sites/default/files/content/blogs/domestica\\_gambiae\\_map.txt](https://www.malariagen.net/sites/default/files/content/blogs/domestica_gambiae_map.txt)).  
480

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

## 488 **Haplotype networks**

489 Haplotype networks were constructed using the median-joining algorithm [29] as im-  
490 plemented in a Python module available from <https://github.com/malariagen/ag1000g->  
491 phase2-vgsc-report. Haplotypes carrying either L995F or L995S mutations were analysed  
492 with a maximum edge distance of two SNPs. Networks were rendered with the Graphviz  
493 library and a composite figure constructed using Inkscape. Non-synonymous edges were

494 highlighted using the SnpEff annotations [47].

495 **Positive selection**

496 Core haplotypes were defined on a 6,078 bp region spanning *Vgsc* codon 995, from chro-  
497 mosome arm 2L position 2,420,443 and ending at position 2,426,521. This region was  
498 chosen as it was the smallest region sufficient to differentiate between the ten genetic  
499 backgrounds carrying either of the known resistance alleles L995F or L995S. Extended  
500 haplotype homozygosity (EHH) was computed for all core haplotypes as described in [30]  
501 using scikit-allel version 1.1.9 [52], excluding non-synonymous and singleton SNPs. Anal-  
502 yses of haplotype homozygosity in moving windows (Supplementary Figures 1 and 2) and  
503 pairwise haplotype sharing (Supplementary Figure 3) were performed using custom Python  
504 code available from <https://github.com/malariagen/ag1000g-phase2-vgsc-report>.

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

506 To explore the feasibility of indentifying a small subset of SNPs that would be sufficient  
507 to identify each of the genetic backgrounds carrying known or putative resistance alleles,  
508 we started with an input data set of all SNPs within the *Vgsc* gene or in the flanking  
509 regions 20 kbp upstream and downstream of the gene. Each of the 2,284 haplotypes in  
510 the Ag1000G Phase 2 cohort was labelled according to which core haplotype it carried,  
511 combining all core haplotypes not carrying known or putative resistance alleles together as  
512 a single "wild-type" group. Decision tree classifiers were then constructed using scikit-learn  
513 version 0.19.0 [53] for a range of maximum depths, repeating the tree construction process  
514 10 times for each maximum depth with a different initial random state. The classification  
515 accuracy of each tree was evaluated using stratified 5-fold cross-validation.

516 **Homology modelling**

517 A homology model of the *An. gambiae* voltage-gated sodium channel (AGAP004707-RD  
518 AgamP4.12) was generated using the 3.8 Å resolution structure of the *Periplaneta amer-*  
519 *icana* sodium channel Na<sub>v</sub>PaS structure (PDB code 5X0M) [25]. Sequences were aligned  
520 using Clustal Omega [54]. 50 starting models were generated using MODELLER [55].  
521 The internal scoring function of MODELLER was used to select 10 models, which were

522 visually inspected and submitted to the VADAR webserver [56] to assess stereochemistry  
523 in order to select the best final model. Figures were produced using PyMOL (DeLano  
524 Scientific, San Carlos, CA, USA).

## 525 References

- 526 [1] S. Bhatt et al. ‘The effect of malaria control on *Plasmodium falciparum* in Africa  
527 between 2000 and 2015’. In: *Nature* 526.7572 (2015), pp. 207–211. ISSN: 0028-0836.
- 528 [2] Janet Hemingway et al. ‘Averting a malaria disaster: Will insecticide resistance derail  
529 malaria control?’ In: *The Lancet* 387.10029 (2016), pp. 1785–1788. ISSN: 1474547X.
- 530 [3] World Health Organization. *Global Plan for Insecticide Resistance Management  
531 (GPIRM)*. Tech. rep. Geneva: World Health Organization, 2012.
- 532 [4] World Health Organization et al. *Global vector control response 2017-2030*. Tech.  
533 rep. 2017.
- 534 [5] Ke Dong et al. ‘Molecular biology of insect sodium channels and pyrethroid resis-  
535 tance’. In: *Insect Biochemistry and Molecular Biology* 50.1 (2014), pp. 1–17. ISSN:  
536 09651748.
- 537 [6] T. G.E. Davies et al. ‘A comparative study of voltage-gated sodium channels in the  
538 Insecta: Implications for pyrethroid resistance in Anopheline and other Neopteran  
539 species’. In: *Insect Molecular Biology* 16.3 (2007), pp. 361–375. ISSN: 09621075.
- 540 [7] D. Martinez-Torres et al. ‘Molecular characterization of pyrethroid knockdown resis-  
541 tance (kdr) in the major malaria vector *Anopheles gambiae* s.s.’ In: *Insect Molecular  
542 Biology* 7.2 (1998), pp. 179–184. ISSN: 09621075.
- 543 [8] Ana Paula B Silva, Joselita Maria M Santos and Ademir J Martins. ‘Mutations in  
544 the voltage-gated sodium channel gene of anophelines and their association with  
545 resistance to pyrethroids: a review’. In: *Parasites & Vectors* 7.1 (2014), p. 450. ISSN:  
546 1756-3305.
- 547 [9] H. Ranson et al. ‘Identification of a point mutation in the voltage-gated sodium  
548 channel gene of Kenyan *Anopheles gambiae* associated with resistance to DDT and  
549 pyrethroids’. In: *Insect Molecular Biology* 9.5 (2000), pp. 491–497. ISSN: 09621075.

- 550 [10] Martin S. Williamson et al. ‘Identification of mutations in the housefly *para*-type  
551 sodium channel gene associated with knockdown resistance (*kdr*) to pyrethroid  
552 insecticides’. In: *Molecular and General Genetics* 252.1-2 (1996), pp. 51–60. ISSN:  
553 00268925.
- 554 [11] Christopher M Jones et al. ‘Footprints of positive selection associated with a mu-  
555 tation (N1575Y) in the voltage-gated sodium channel of *Anopheles gambiae*.’ In:  
556 *Proceedings of the National Academy of Sciences of the United States of America*  
557 109.17 (2012), pp. 6614–9. ISSN: 1091-6490.
- 558 [12] T. G. E. Davies et al. ‘DDT, pyrethrins, pyrethroids and insect sodium channels’.  
559 In: *IUBMB Life* 59.3 (2007), pp. 151–162. ISSN: 1521-6543.
- 560 [13] Frank D. Rinkevich, Yuzhe Du and Ke Dong. ‘Diversity and convergence of sodium  
561 channel mutations involved in resistance to pyrethroids’. In: *Pesticide Biochemistry  
562 and Physiology* 106.3 (2013), pp. 93–100. ISSN: 00483575.
- 563 [14] J Pinto et al. ‘Multiple origins of knockdown resistance mutations in the Afrotropical  
564 mosquito vector *Anopheles gambiae*’. In: *PLoS One* 2 (2007), e1243. ISSN: 19326203.
- 565 [15] Josiane Etang et al. ‘Polymorphism of intron-1 in the voltage-gated sodium channel  
566 gene of *Anopheles gambiae* s.s. populations from cameroon with emphasis on insecti-  
567 cide knockdown resistance mutations’. In: *Molecular Ecology* 18.14 (2009), pp. 3076–  
568 3086. ISSN: 09621083.
- 569 [16] Amy Lynd et al. ‘Field, genetic, and modeling approaches show strong positive  
570 selection acting upon an insecticide resistance mutation in *Anopheles gambiae* s.s.’  
571 In: *Molecular Biology and Evolution* 27.5 (2010), pp. 1117–1125. ISSN: 07374038.
- 572 [17] Federica Santolamazza et al. ‘Remarkable diversity of intron-1 of the *para* voltage-  
573 gated sodium channel gene in an *Anopheles gambiae*/*Anopheles coluzzii* hybrid zone.’  
574 In: *Malaria Journal* 14.1 (2015), p. 9. ISSN: 1475-2875.
- 575 [18] Mylène Weill et al. ‘The *kdr* mutation occurs in the Mopti form of *Anopheles gambiae*  
576 s.s. through introgression’. In: *Insect Molecular Biology* 9.5 (2000), pp. 451–455.

- 577 [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.
- 578
- 579
- 580 [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.
- 581
- 582
- 583 [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.
- 584
- 585
- 586 [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: *Genome Research* (2020), pp. 1533–1546.
- 587
- 588
- 589 [23] The *Anopheles gambiae* 1000 Genomes Consortium. ‘Natural diversity of the malaria vector *Anopheles gambiae*’. In: *Nature* 552 (2017), pp. 96–100.
- 590
- 591 [24] Shoji Sonoda et al. ‘Genomic organization of the para-sodium channel *a*-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.
- 592
- 593
- 594 ISSN: 07394462.
- 595 [25] Huaizong Shen et al. ‘Structure of a eukaryotic voltage-gated sodium channel at near-atomic resolution’. In: *Science* (2017), eaal4326.
- 596
- 597 [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.
- 598
- 599
- 600 [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.
- 601
- 602
- 603 [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.
- 604
- 605

- 606 [29] H. J. Bandelt, P. Forster and A. Rohl. ‘Median-joining networks for inferring in-  
607       traspecific phylogenies’. In: *Molecular Biology and Evolution* 16.1 (1999), pp. 37–48.  
608       ISSN: 0737-4038.
- 609 [30] Pardis C. Sabeti et al. ‘Detecting recent positive selection in the human genome from  
610       haplotype structure’. In: *Nature* 419.6909 (2002), pp. 832–837. ISSN: 0028-0836.
- 611 [31] Molly C Reid and F Ellis McKenzie. ‘The contribution of agricultural insecticide use  
612       to increasing insecticide resistance in African malaria vectors’. In: *Malaria Journal*  
613       15.1 (2016), p. 107.
- 614 [32] Sara A Abuelmaali et al. ‘Impacts of agricultural practices on insecticide resistance  
615       in the malaria vector *Anopheles arabiensis* in Khartoum State, Sudan’. In: *PLoS*  
616       One 8.11 (2013), e80549.
- 617 [33] Zhaonong Hu et al. ‘A sodium channel mutation identified in *Aedes aegypti* se-  
618       lectively reduces cockroach sodium channel sensitivity to type I, but not type II  
619       pyrethroids’. In: *Insect Biochemistry and Molecular Biology* 41.1 (2011), pp. 9–13.
- 620 [34] Andrias O. O’Reilly et al. ‘Modelling insecticide-binding sites in the voltage-gated  
621       sodium channel’. In: *Biochemical Journal* 396.2 (2006), pp. 255–263. ISSN: 0264-6021.
- 622 [35] Jiabao Xu et al. ‘Multi-country survey revealed prevalent and novel F1534S muta-  
623       tion in voltage-gated sodium channel (VGSC) gene in *Aedes albopictus*’. In: *PLoS*  
624       Neglected Tropical Diseases 10.5 (2016), e0004696.
- 625 [36] Intan H Ishak et al. ‘Contrasting patterns of insecticide resistance and knockdown  
626       resistance (*kdr*) in the dengue vectors *Aedes aegypti* and *Aedes albopictus* from  
627       Malaysia’. In: *Parasites & Vectors* 8.1 (2015), p. 181.
- 628 [37] Yiji Li et al. ‘Evidence for multiple-insecticide resistance in urban *Aedes albopictus*  
629       populations in southern China’. In: *Parasites & Vectors* 11.1 (2018), p. 4.
- 630 [38] Yuzhe Du et al. ‘Rotational symmetry of two pyrethroid receptor sites in the mosquito  
631       sodium channel’. In: *Molecular Pharmacology* 88.2 (Aug. 2015), pp. 273–280. ISSN:  
632       1521-0111.

- 633 [39] H Vais et al. ‘Activation of *Drosophila* sodium channels promotes modification by  
634 deltamethrin. Reductions in affinity caused by knock-down resistance mutations’. In:  
635 *The Journal of General Physiology* 115.3 (Mar. 2000), pp. 305–318. ISSN: 0022-1295.
- 636 [40] Deborah L. Capes et al. ‘Domain IV voltage-sensor movement is both sufficient and  
637 rate limiting for fast inactivation in sodium channels’. In: *The Journal of General*  
638 *Physiology* 142.2 (Aug. 2013), pp. 101–112. ISSN: 1540-7748.
- 639 [41] Zhen Yan et al. ‘Structure of the Nav1.4-B1 Complex from Electric Eel’. In: *Cell*  
640 170.3 (27th July 2017), 470–482.e11. ISSN: 0092-8674.
- 641 [42] Altin Sula et al. ‘The complete structure of an activated open sodium channel’. In:  
642 *Nature Communications* 8 (16th Feb. 2017), p. 14205. ISSN: 2041-1723.
- 643 [43] P N R Usherwood et al. ‘Mutations in DIIS5 and the DIIS4-S5 linker of *Drosophila*  
644 *melanogaster* sodium channel define binding domains for pyrethroids and DDT’. In:  
645 *FEBS Letters* 581.28 (27th Nov. 2007), pp. 5485–5492. ISSN: 0014-5793.
- 646 [44] Andy Kilianski et al. ‘Bacterial and viral identification and differentiation by am-  
647 plicon sequencing on the MinION nanopore sequencer.’ In: *GigaScience* 4 (2015),  
648 p. 12. ISSN: 2047-217X.
- 649 [45] Eric R Lucas et al. ‘A high throughput multi-locus insecticide resistance marker  
650 panel for tracking resistance emergence and spread in *Anopheles gambiae*’. In: *Sci-  
651 entific reports* 9.1 (2019), pp. 1–10.
- 652 [46] R A Holt et al. ‘The genome sequence of the malaria mosquito *Anopheles gambiae*’.  
653 In: *Science* 298.5591 (2002), pp. 129–149. ISSN: 0036-8075.
- 654 [47] Pablo Cingolani et al. ‘A program for annotating and predicting the effects of single  
655 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*  
656 strain w1118; iso-2; iso-3’. In: *Fly* 6.2 (2012), pp. 80–92. ISSN: 19336942.
- 657 [48] Sudhir Kumar, Glen Stecher and Koichiro Tamura. ‘MEGA7: Molecular Evolution-  
658 ary Genetics Analysis Version 7.0 for Bigger Datasets’. In: *Molecular Biology and*  
659 *Evolution* 33.7 (2016), pp. 1870–1874. ISSN: 15371719.
- 660 [49] Olivier Delaneau et al. ‘Haplotype estimation using sequencing reads’. In: *American*  
661 *Journal of Human Genetics* 93.4 (2013), pp. 687–696. ISSN: 00029297.

- 662 [50] Androniki Menelaou and Jonathan Marchini. ‘Genotype calling and phasing using  
663 next-generation sequencing reads and a haplotype scaffold’. In: *Bioinformatics* 29.1  
664 (2013), pp. 84–91. ISSN: 13674803.
- 665 [51] R. C. Lewontin. ‘The Interaction of Selection and Linkage. I. General Considerations;  
666 Heterotic Models’. In: *Genetics* 49.1 (1964), pp. 49–67. ISSN: 0016-6731.
- 667 [52] Alistair Miles and Nicholas Harding. *scikit-allel: A Python package for exploring and  
668 analysing genetic variation data*. 2016.
- 669 [53] F. Pedregosa et al. ‘Scikit-learn: Machine Learning in Python’. In: *Journal of Ma-  
670 chine Learning Research* 12 (2011), pp. 2825–2830.
- 671 [54] Fabian Sievers et al. ‘Fast, scalable generation of high-quality protein multiple se-  
672 quence alignments using Clustal Omega’. In: *Molecular Systems Biology* 7 (2011),  
673 p. 539. ISSN: 1744-4292.
- 674 [55] Narayanan Eswar et al. ‘Comparative protein structure modeling using MODELLER’.  
675 In: *Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et  
676 Al.]* Chapter 2 (Nov. 2007), Unit 2.9. ISSN: 1934-3663.
- 677 [56] Leigh Willard et al. ‘VADAR: a web server for quantitative evaluation of protein  
678 structure quality’. In: *Nucleic Acids Research* 31.13 (1st July 2003), pp. 3316–3319.

## 679 Data Accessibility and Benefit-Sharing Statement

680 Sequence read alignments and variant calls from Ag1000G phase 2 are available from the  
681 European Nucleotide Archive under study accession PRJEB36277 (ENA - <http://www.ebi.ac.uk/ena>).  
682 Sequence read alignments for samples in Ag1000G phase 1 are available under study ac-  
683 cession PRJEB18691.