

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 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

15           25th March 2020

16           **Abstract**

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

25        *Vgsc* gene in mosquito populations from 13 African countries. In addition to the three  
26        known *kdr* resistance alleles, we describe 20 non-synonymous nucleotide substitutions  
27        (at appreciable frequency in one or more populations) that are previously unknown  
28        in *Anopheles* mosquitoes and we mapped these variants on a molecular model of the  
29        protein. Thirteen of these novel alleles were found to occur almost exclusively on hap-  
30        lotypes carrying the known L995F resistance allele (L1014F in *Musca domesticus* codon  
31        numbering), and may enhance or compensate for the L995F resistance phenotype. A  
32        novel mutation I1527T, which is adjacent to a predicted pyrethroid binding site, was  
33        found in tight linkage with either of two alleles causing a V402L substitution, similar  
34        to a combination of substitutions found to cause pyrethroid resistance in several other  
35        insect species. We analyse the genetic backgrounds on which non-synonymous alleles  
36        are found, to determine which alleles have experienced recent positive selection, and to  
37        refine our understanding of the spread of resistance between species and geographical  
38        locations. We describe ten distinct *kdr* carrying haplotype groups with evidence of  
39        recent positive selection, five of which carry the known L995F resistance allele, five  
40        of which carry the known L995S resistance allele. Five of these groups are localised  
41        to a single geographical location, and five comprise haplotypes from different coun-  
42        tries, in one case separated by over 3000 km, providing new information about the  
43        geographical distribution and spread of resistance. Two "non-*kdr*" haplotype groups  
44        with evidence of recent selection were also detected, one of which carries the novel  
45        I1527T allele, and one of which carries a novel M490I allele. We also find evidence  
46        for multiple introgression events transmitting resistance alleles between *An. gambiae*  
47        and *An. coluzzii*. We identify markers that could be used to design high-throughput,  
48        low-cost genetic assays for improved surveillance of pyrethroid resistance in the field.  
49        Our results demonstrate that the molecular basis of target-site pyrethroid resistance  
50        in malaria vectors is more complex than previously appreciated, and provide a founda-  
51        tion for the development of new genetic tools to track the spread insecticide resistance  
52        and improve the design of strategies for insecticide resistance management.

## 53        **Introduction**

54        Pyrethroid insecticides have been the cornerstone of malaria prevention in Africa for almost  
55        two decades [1]. Pyrethroids are currently used in all insecticide-treated bed-nets (ITNs),  
56        and are widely used in indoor residual spraying (IRS) campaigns as well as in agriculture.

57 Resistance to these insecticides is now widespread in malaria vector populations across  
58 Africa [2]. The World Health Organization (WHO) has published plans for insecticide  
59 resistance management (IRM), which emphasise the need for improvements in both our  
60 knowledge of the molecular mechanisms of resistance and our ability to monitor them in  
61 natural populations [3, 4].

62 The voltage-gated sodium channel (VGSC) is the physiological target of pyrethroid in-  
63 secticides, and is integral to the insect nervous system. The sodium channel protein con-  
64 sists of four homologous domains (DI-IV) each of which comprises six transmembrane seg-  
65 ments (S1-S6) connected by intracellular and extracellular loops [5]. Pyrethroid molecules  
66 bind to this protein, stabilise the ion-conducting active state and thus disrupt normal  
67 nervous system function to produce paralysis (“knock-down”) and death. However, amino  
68 acid substitutions at key positions within the protein alter the interaction with insecticide  
69 molecules (target-site resistance), increasing the dose of insecticide required for knock-  
70 down, and leading to this type of resistance to also be known as knock-down resistance or  
71 *kdr* [6, 5].

72 In the African malaria vectors *Anopheles gambiae* and *An. coluzzii*, three substitutions  
73 have been found to cause pyrethroid resistance. Two of these substitutions occur in codon  
74 995<sup>1</sup>, with L995F prevalent in West and Central Africa [7, 8], and L995S found in Central  
75 and East Africa [9, 8]. A third substitution, N1570Y, has been found in West and Central  
76 Africa and shown to increase resistance in association with L995F [11]. However, studies in  
77 other insect species have found a variety of other *Vgsc* substitutions inducing a resistance  
78 phenotype [12, 13, 5]. To our knowledge, no studies in malaria vectors have analysed  
79 genetic variation across the full *Vgsc* coding sequence, thus the molecular basis of target-  
80 site resistance to pyrethroids has not been fully explored.

81 Basic information is also lacking about the spread of pyrethroid resistance in malaria  
82 vectors [3]. For example, it is not clear when, where or how many times pyrethroid  
83 target-site resistance has emerged. Geographical paths of transmission, carrying resistance  
84 alleles between mosquito populations, are also not known. Previous studies have found  
85 evidence that L995F occurs on several different genetic backgrounds, suggesting multiple

---

<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.

86 independent outbreaks of resistance driven by this allele [14, 15, 16, 17]. However, these  
87 studies analysed only small gene regions in a limited number of mosquito populations, and  
88 therefore had limited resolution to make inferences about relationships between haplotypes  
89 carrying this allele. It has also been shown that the L995F allele spread from *An. gambiae*  
90 to *An. coluzzii* in West Africa [18, 19, 20, 21]. However, both L995F and L995S now have  
91 wide geographical distributions [8], and to our knowledge no attempts have been made to  
92 infer or track the geographical spread of either allele across Africa.

93 Here we report an in-depth analysis of genetic variation in the *Vgsc* gene, using whole-  
94 genome Illumina sequence data from phase 2 of the *Anopheles gambiae* 1000 Genomes  
95 Project (Ag1000G) [22]. The Ag1000G phase 2 resource includes data on nucleotide vari-  
96 ation in 1,142 wild-caught mosquitoes sampled from 13 countries, with representation of  
97 West, Central, Southern and East Africa, and of both *An. gambiae* and *An. coluzzii*.  
98 We investigate variation across the complete gene coding sequence, and report popula-  
99 tion genetic data for both known and novel non-synonymous nucleotide substitutions. We  
100 then use haplotype data from the chromosomal region spanning the *Vgsc* gene to study  
101 the genetic backgrounds carrying resistance alleles, infer the geographical spread of resis-  
102 tance between mosquito populations, and provide evidence for recent positive selection.  
103 Finally, we explore ways in which variation data from Ag1000G can be used to design  
104 high-throughput, low-cost genetic assays for surveillance of pyrethroid resistance, with  
105 the capability to differentiate and track resistance outbreaks.

## 106 Results

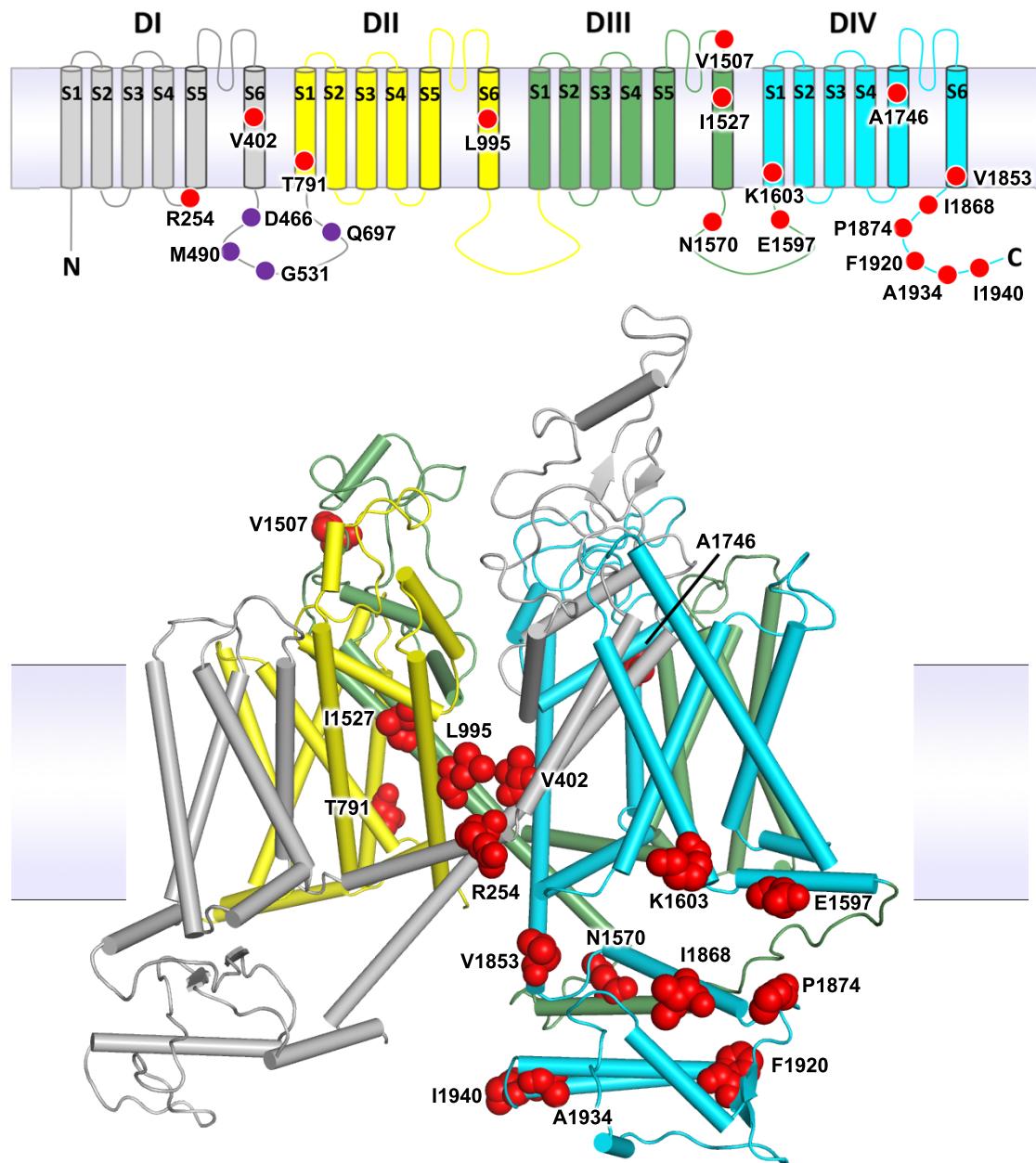
### 107 ***Vgsc* non-synonymous nucleotide variation**

108 To identify variants with a potentially functional role in pyrethroid resistance, we ex-  
109 tracted single nucleotide polymorphisms (SNPs) that alter the amino acid sequence of the  
110 VGSC protein from the Ag1000G phase 2 data resource [22]. We then computed their  
111 allele frequencies among 16 mosquito populations defined by species and country of ori-  
112 gin. Alleles that confer resistance are expected to increase in frequency under selective  
113 pressure, therefore we filtered the list of potentially functional variant alleles to retain  
114 only those at or above 5% frequency in one or more populations (Table 1). The resulting

list comprises 23 variant alleles, including the known L995F, L995S and N1570Y resistance alleles, and a further 20 alleles which prior to Ag1000G had not previously been described in anopheline mosquitoes. We reported 12 of these novel alleles in our overall analysis of the 765 samples in the Ag1000G phase 1 data resource [23], and we extend the analyses here to incorporate SNPs which alter codon 531, 697, 1507, 1603 and two tri-allelic SNPs affecting codons 402 and 490 in the 1,142 phase 2 samples.

The 23 non-synonymous variants were located on a transmembrane topology map and on a 3-dimensional homology model of the *Vgsc* protein. (Figure 1). The substitutions were found to be distributed throughout the channel, in all of the four internally homologous domains (DI-DIV), in S1, S5 and S6 membrane-spanning segments, in two of the intracellular loops connecting domains and in the C-terminal tail. The S5 and S6 segments that form the central ion-conducting pore of the channel carry five of seven segment substitutions, including V402 and L995 which have been shown to produce insecticide resistance phenotypes [6, 5, 7, 8, 9]. Seven 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<sub>v</sub>PaS 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 1 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, hinting there may be a



**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.

**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 [23] @@REF-phase2. 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].

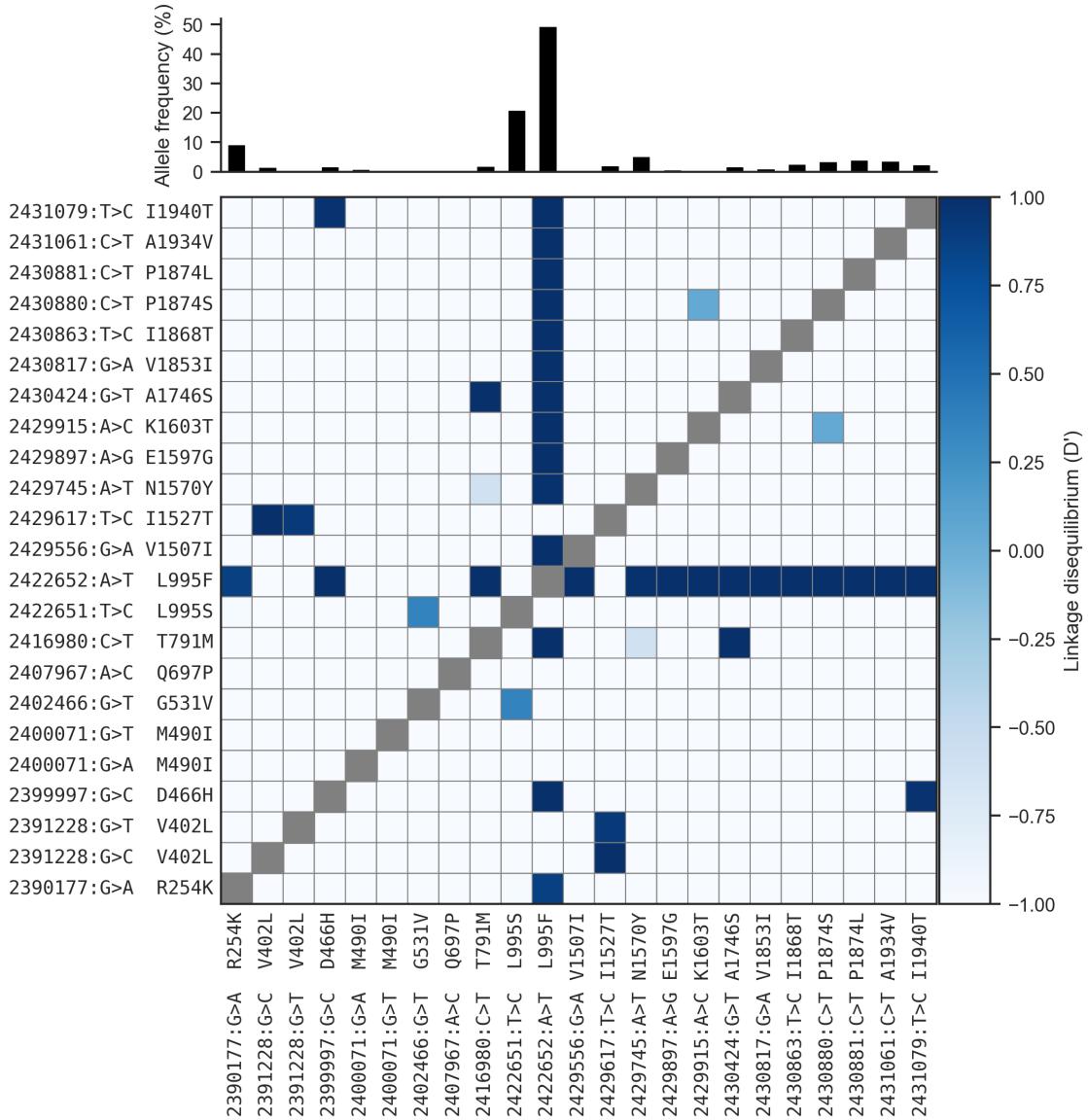
<sup>4</sup> Location of the variant within the protein structure. Transmembrane segments are named according to domain number (in Roman numerals) followed by 'S' then the number of the segment; e.g., 'IIS6' means domain two, transmembrane segment six. Internal linkers between segments within the same 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.

146 fitness advantage for mosquitoes carrying both alleles in some circumstances.

147 The N1570Y allele was present in Guinea, Burkina Faso (both species) and Cameroon.  
148 This allele has been shown to substantially increase pyrethroid resistance when it occurs  
149 in combination with L995F, both in association tests of phenotyped field samples [11]  
150 and functional tests using *Xenopus* oocytes [26]. To study the patterns of association  
151 among non-synonymous variants, we used haplotypes from the Ag1000G phase 2 resource  
152 to compute the normalised coefficient of linkage disequilibrium ( $D'$ ) between all pairs of  
153 variant alleles (Figure 2). As expected, we found N1570Y in almost perfect linkage with  
154 L995F. Of the 20 novel non-synonymous alleles, 13 also occurred almost exclusively in  
155 combination with L995F (Figure 2). These included two variants in codon 1874 (P1874S,  
156 P1874L), one of which (P1874S) has previously been associated with pyrethroid resistance  
157 in the crop pest moth *Plutella xylostella* [24].

158 The abundance of high-frequency non-synonymous variants occurring in combination  
159 with L995F is striking for two reasons. First, *Vgsc* is a highly conserved gene, expected  
160 to be under strong functional constraint and therefore purifying selection, so any non-  
161 synonymous variants are expected to be rare [12]. Second, in contrast with L995F, we did  
162 not observe any high-frequency non-synonymous variants occurring in combination with  
163 L995S. This contrast was clear when data on all variants within the gene were considered:  
164 for haplotypes carrying the L995 allele, the ratio of non-synonymous to synonymous nu-  
165 cleotide diversity ( $\pi_N/\pi_S$ ) was 20.04 times higher than haplotypes carrying the wild-type  
166 allele, but for those carrying L995S ( $\pi_N/\pi_S$ ) was 0.5 times lower than haplotypes carrying  
167 the wild-type allele. These results may indicate that L995F has substantially altered the  
168 selective regime for other amino acid positions within the protein, perhaps through relax-  
169 ation of purifying selection. Secondary substitutions have occurred and risen in frequency,  
170 suggesting that they are providing some selective advantage in the presence of insecticide  
171 pressure.

172 A novel allele, I1527T, was present in *An. coluzzii* from Burkina Faso at 13% fre-  
173 quency. Codon 1527 occurs within trans-membrane segment IIIS6, immediately adjacent  
174 to residues within a predicted binding site for pyrethroid molecules, thus it is plausible that  
175 I1527T could alter pyrethroid binding [27, 5]. We also found that the two variant alleles  
176 affecting codon 402, both of which induce a V402L substitution, were in strong linkage  
177 with I1527T ( $D' \geq 0.8$ ; Figure 2), and almost all haplotypes carrying I1527T also carried a  
178 V402L substitution. Substitutions in codon 402 have been found in a number of other insect



**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 1 cohort. See Table 1 for population allele frequencies.

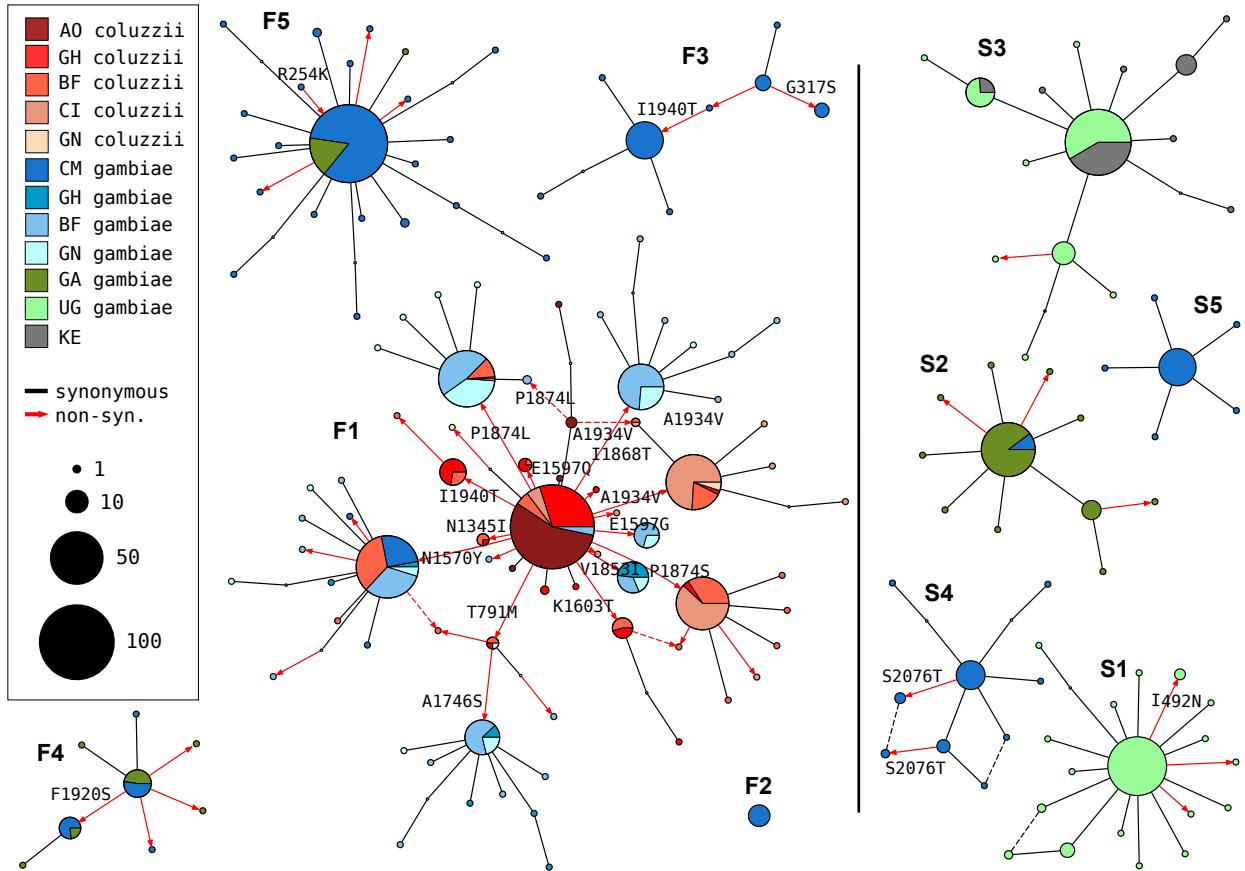
species and shown experimentally to confer pyrethroid resistance [5]. Because of the limited geographical distribution of these alleles, we hypothesize that the I1527T+V402L combination represents a pyrethroid resistance allele that arose in West African *An. coluzzii* populations. However, the L995F allele is at higher frequency (85%) in our Burkina Faso *An. coluzzii* population, and is known to be increasing in frequency [28], therefore L995F may provide a stronger resistance phenotype and is replacing I1527T+V402L.

185      The four remaining novel alleles, Q697P, G531V and two separate nucleotide substitutions  
186      causing M490I, did not occur in combination with any known resistance allele and were  
187      private to a single population (except a single haplotype carrying G531V from Bioko) (Table  
188      1), and to our knowledge, none have previously been found in other species [13, 5].

189      **Genetic backgrounds carrying resistance alleles**

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

203      To further investigate genetic backgrounds carrying resistance alleles, we used the  
204      Ag1000G phase 2 haplotype data from the *Vgsc* gene (2,284 haplotypes from 1,142 mosquitoes  
205      [22]), to construct median-joining networks [29] (Figure 3). The network analysis improves  
206      on hierarchical clustering by allowing for the reconstruction and placement of intermedi-  
207      ate haplotypes that may not be observed in the data. It also allows for non-hierarchical  
208      relationships between haplotypes, which may arise if recombination events have occurred  
209      between haplotypes. We constructed the network up to a maximum edge distance of 2 SNP  
210      differences, to ensure that each connected component captures a group of closely-related  
211      haplotypes. The resulting network contained 5 groups containing haplotypes carrying  
212      L995F, and a further 5 groups carrying L995S, in close correspondence with previous re-  
213      sults from hierarchical clustering (96.8% overall concordance in assignment of haplotypes  
214      to groups).



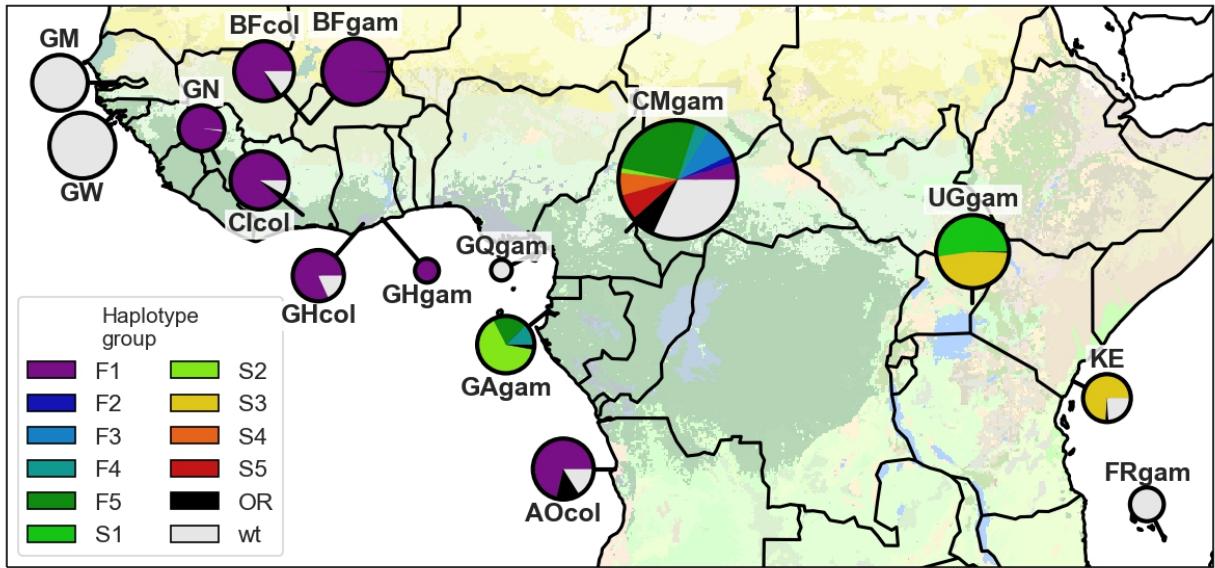
**Figure 3. Haplotype networks.** Median joining network for haplotypes carrying L995F (labelled F1-F5) or L995S variants (S1-S5) with a maximum edge distance of two SNPs. Labelling of network components is via concordance with hierarchical clusters discovered in [23]. Node size is relative to the number 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.

215 The haplotype network brings into sharp relief the explosive radiation of amino acid sub-  
 216 stitutions secondary to the L995F allele (Figure 3). Within the F1 group, nodes carrying  
 217 non-synonymous variants radiate out from a central node carrying only L995F, suggest-  
 218 ing that the central node represents the ancestral haplotype carrying L995F alone which  
 219 initially came under selection, and these secondary variants have arisen subsequently as  
 220 new mutations. Many of the nodes carrying secondary variants are large, consistent with  
 221 positive selection and a functional role for these secondary variants as modifiers of the  
 222 L995F resistance phenotype. The F1 network also allows us to infer multiple introgression  
 223 events between the two species. The central (putatively ancestral) node contains hap-

224 lotypes from individuals of both species, as do nodes carrying the N1570Y, P1874L and  
225 T791M variants. This structure is consistent with an initial introgression of the ancestral  
226 F1 haplotype, followed later by introgressions of haplotypes carrying secondary mutations.  
227 The haplotype network also illustrates the contrasting levels of non-synonymous varia-  
228 tion between L995F and L995S. Only eight non-synonymous variants are present within  
229 the L995S groups, and all are at low frequency, thus may be neutral or mildly deleterious  
230 variants that are hitch-hiking on selective sweeps for the L995S allele.

231 The F1 group contains haplotypes from mosquitoes of both species, and from mosquitoes  
232 sampled in six different countries (Angola, Burkina Faso, Cameroon, Côte d'Ivoire, Ghana,  
233 Guinea) (Figure 4). The F4, F5 and S2 groups each contain haplotypes from both  
234 Cameroon and Gabon. The S3 group contains haplotypes from both Uganda and Kenya.  
235 The haplotypes within each of these five groups (F1, F4, F5, S2, S3) were nearly identi-  
236 cal across the entire span of the *Vgsc* gene ( $\pi < 5.1 \times 10^{-5} \text{ bp}^{-1}$ ). In contrast, diversity  
237 among wild-type haplotypes was two orders of magnitude greater (Cameroon *An. gambiae*  
238  $\pi = 1.4 \times 10^{-3} \text{ bp}^{-1}$ ; Guinea-Bissau  $\pi = 5.7 \times 10^{-3} \text{ bp}^{-1}$ ). Thus it is reasonable to assume  
239 that each of these five groups contains descendants of an ancestral haplotype that carried  
240 a resistance allele and has risen in frequency due to selection for insecticide resistance.  
241 Given this assumption, these groups each provide evidence for adaptive gene flow between  
242 mosquito populations separated by considerable geographical distances.

243 A limitation of both the hierarchical clustering and network analyses is that they rely  
244 on genetic distances within a fixed genomic window from the start to the end of the  
245 *Vgsc* gene. *Anopheles* mosquitoes undergo homologous recombination during meiosis in  
246 both males and females, and any recombination events that occurred within this genomic  
247 window could affect the way that haplotypes are grouped together in clusters or network  
248 components. In particular, recombination events could occur during the geographical  
249 spread of a resistance allele, altering the genetic background upstream and/or downstream  
250 of the allele itself. An analysis based on a fixed genomic window might then fail to infer  
251 gene flow between two mosquito populations, because haplotypes with and without a  
252 recombination event could be grouped separately, despite the fact that they share a recent  
253 common ancestor. To investigate the possibility that recombination events may have  
254 affected our grouping of haplotypes carrying resistance alleles, we performed a moving



**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 or putative resistance alleles.

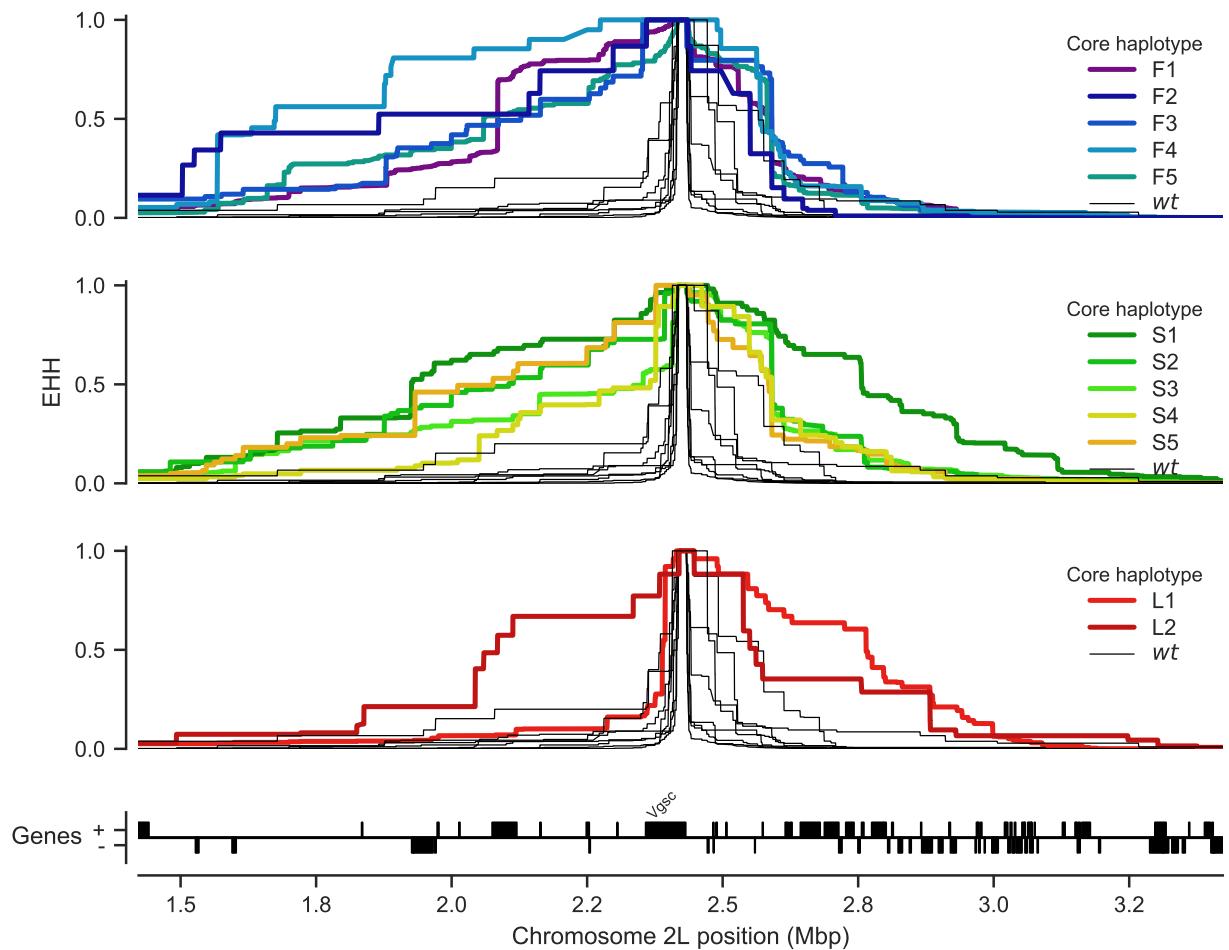
255 window analysis of haplotype homozygosity, spanning *Vgsc* and up to a megabase upstream  
 256 and downstream of the gene (Supplementary Figures S1, S2). This analysis supported a  
 257 refinement of our initial grouping of haplotypes carrying resistance alleles. All haplotypes  
 258 within groups S4 and S5 were effectively identical on both the upstream and downstream  
 259 flanks of the gene, but there was a region of divergence within the *Vgsc* gene itself that  
 260 separated them in the fixed window analyses (Supplementary Figure S2). The 13.8 kbp  
 261 region of divergence occurred upstream of codon 995 and contained 6 SNPs that were fixed  
 262 differences between S4 and S5. A possible explanation for this short region of divergence  
 263 is that a gene conversion event has occurred within the gene, bringing a segment from  
 264 a different genetic background onto the original genetic background on which the L995S  
 265 resistance mutation occurred.

#### 266 Positive selection for resistance alleles

267 To investigate evidence for positive selection on non-synonymous alleles, we performed  
 268 an analysis of extended haplotype homozygosity (EHH) [30]. Haplotypes under recent

positive selection will have increased rapidly in frequency, thus have had less time to be broken down by recombination, and should on average have longer regions of haplotype homozygosity relative to wild-type haplotypes. @@CHECK - We defined a core region spanning *Vgsc* codon 995 and an additional 6 kbp of flanking sequence, which was the minimum required to differentiate the haplotype groups identified via clustering and network analyses. Within this core region, we found 18 distinct haplotypes at a frequency above 1% within the cohort. These included core haplotypes corresponding to each of the 10 haplotype groups carrying L995F or L995S alleles identified above, as well as a core haplotype carrying I1527T which we labelled L1 (due to it carrying 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 dramatically 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 comparing F1-5 (carrying L995F) against S1-5 (carrying L995S) (Supplementary Figure S3). There were, however, some differences between core haplotypes carrying the same allele. For example, shared haplotypes were significantly longer for S1 (median 1.006 cM, 95% CI [0.986 - 1.040]) versus other core haplotypes carrying L995S (e.g., S2 median 0.593 cM, 95% CI [0.589 - 0.623]; Supplementary Figure S3). Longer shared haplotypes indicate a more recent common ancestor, and thus some of these core haplotypes may have experienced more recent and/or more intense selection than others. The L1 haplotype carrying I1527T+V402L exhibited a slow decay of EHH on the downstream flank of the gene, similar to haplotypes carrying L995F and L995S, indicating that this combination of alleles has experienced positive selection. EHH decay on the upstream gene flank was faster, being similar to wild-type haplotypes, however there were two separate nucleotide substitutions encoding V402L within this group of haplotypes, and a faster EHH decay on this flank is consistent with recombination events bringing V402L alleles from differ-



**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).

300 ent genetic backgrounds together with an ancestral haplotype carrying I1527T. The L2  
 301 haplotype carrying M490I exhibited EHH decay on both flanks comparable to haplotypes  
 302 carrying known resistance alleles. This could indicate evidence for selection on the M490I  
 303 allele, however these haplotypes are derived from a Kenyan mosquito population where  
 304 there is evidence for a severe recent bottleneck [23], and there were not enough wild-type  
 305 haplotypes from Kenya with which to compare, thus this signal may also be due to the  
 306 extreme demographic history of this population.

307 **Discussion**

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

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

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

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

342 The non-synonymous variants are distributed throughout the channel protein but can  
343 be considered in terms of three clusters: (i) the transmembrane domain, (ii) the DI-II  
344 intracellular linker and (iii) the DIII-DIV/C-terminal subdomain. The pyrethroid binding  
345 site is located in the transmembrane domain between the IIS4-S5 linker and the IIS5,  
346 IIS6 and IIIS6 helices [34]. The I1527T substitution that we discovered in *An. coluzzii*  
347 mosquitoes from Burkina Faso occurs in segment IIIS6 and is immediately adjacent to  
348 two pyrethroid-sensing residues in this binding site [5]. It is thus plausible that pyrethroid  
349 binding could be altered by this substitution. The I1527T substitution (*M. domestica*  
350 codon 1532) has been found in *Aedes albopictus* [35], and substitutions in the nearby  
351 codon 1529 (*M. domestica* codon 1534) have been reported in *Aedes albopictus* and in  
352 *Aedes aegypti* where it was found to be associated with pyrethroid resistance [5, 36, 37].  
353 We found the I1527T allele in tight linkage with two alleles causing a V402L substitution  
354 (*M. domestica* codon 410). Substitutions in codon 402 have been found in multiple insect  
355 species and are by themselves sufficient to confer pyrethroid resistance [5]. The fact that  
356 we find I1527T and V402L in such tight mutual association is intriguing because haplotypes  
357 carrying V402L alone should also have been positively selected and thus be present in one  
358 or more populations.

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

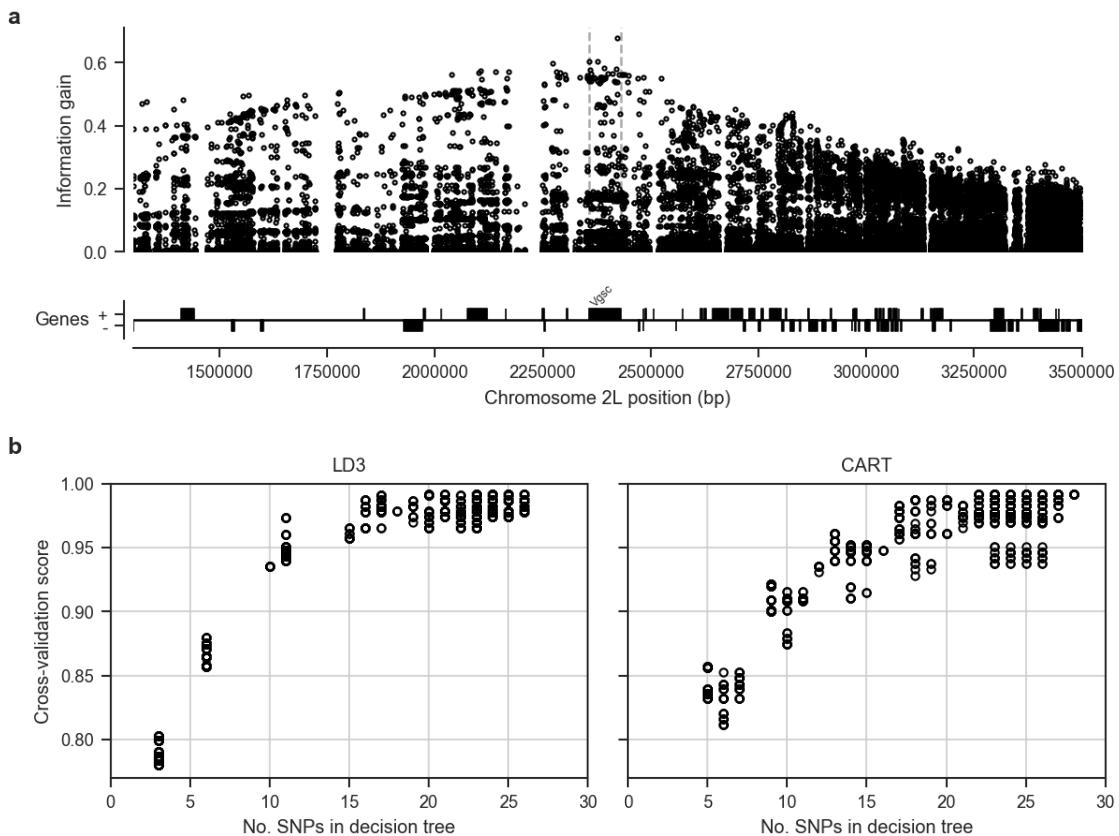
367 involves movement of the DIV domain to form a receptor for the DIII-DIV linker fast in-  
368 activation particle containing the 'MFM' sequence motif (equivalent to the 'IFM' motif in  
369 mammals) [40, 5]. Recent eukaryotic sodium channel structures reveal that the DIII-DIV  
370 linker is in complex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

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

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

396 Entomological surveillance teams in Africa regularly genotype mosquitoes for resistance al-  
397 leles in *Vgsc* codon 995, and use those results as an indicator for the presence of pyrethroid  
398 resistance alongside results from insecticide resistance bioassays. They typically do not,  
399 however, sequence the gene or genotype any other polymorphisms within the gene. Thus  
400 if there are other polymorphisms within the gene that cause or significantly enhance  
401 pyrethroid resistance, these will not be detected. Also, if a codon 995 resistance allele  
402 is observed, there is no way to know whether the allele is on a genetic background that  
403 has also been observed in other mosquito populations, and thus no way to investigate  
404 whether resistance alleles are emerging locally or being imported from elsewhere. Whole-  
405 genome sequencing of individual mosquitoes clearly provides data of sufficient resolution to  
406 answer these questions, and could be used to provide ongoing resistance surveillance. The  
407 cost of whole-genome sequencing continues to fall, with the present cost being approxi-  
408 mately 50 GBP to obtain ~30× coverage of an individual *Anopheles* mosquito genome with  
409 150 bp paired-end reads. However, to achieve substantial spatial and temporal coverage  
410 of mosquito populations, it is currently cheaper and more practical to develop targeted  
411 genetic assays for resistance outbreak surveillance. Technologies such as amplicon se-  
412 quencing [44] are already being trialled on mosquitoes [45], these could scale to tens of  
413 thousands of samples at low cost and could be implemented using existing platforms in  
414 national molecular biology facilities.

415 To facilitate the development of targeted genetic assays for surveillance of *Vgsc*-mediated  
416 pyrethroid resistance, we have produced several supplementary data tables. In Supple-  
417 mentary Table 1 we list all 64 non-synonymous variants found within the *Vgsc* gene in this  
418 study, with population allele frequencies. In Supplementary Table 2 we list 771 biallelic  
419 SNPs, within the *Vgsc* gene and up to 10 kbp upstream or downstream, that are poten-  
420 tially informative regarding which haplotype group a resistance haplotype belongs to, and  
421 thus could be used for tracking the spread of resistance. This table includes the allele  
422 frequency within each of the 12 haplotype groups defined here, to aid in identifying SNPs  
423 that are highly differentiated between two or more haplotype groups. We also provide  
424 Supplementary Table 3 which lists all 8,297 SNPs found within the *Vgsc* gene and up to



**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.

425 10 kbp upstream or downstream, which might need to be taken into account as flanking  
 426 variation when searching for PCR primers to amplify a SNP of interest. To provide some  
 427 indication for how many SNPs would need to be assayed in order to track the spread of  
 428 resistance, we used haplotype data from this study to construct decision trees that could  
 429 classify which of the 12 groups a given haplotype belongs to (Figure 6). This analysis  
 430 suggested that it should be possible to construct a decision tree able to classify haplo-  
 431 types with >95% accuracy by using 20 SNPs or less. In practice, more SNPs would be  
 432 needed, to provide some redundancy, and also to type non-synonymous polymorphisms in  
 433 addition to identifying the genetic background. However, it is still likely to be well within  
 434 the number of SNPs that could be assayed in a single multiplex via amplicon sequencing.  
 435 Thus it should be feasible to produce low-cost, high-throughput genetic assays for tracking

436 the spread of pyrethroid resistance. If combined with a limited amount of whole-genome  
437 sequencing at sentinel sites, this should also allow the identification of newly emerging  
438 resistance outbreaks.

## 439 **Methods**

### 440 **Code**

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

### 443 **Data**

444 We used variant calls and phased haplotype data from the Ag1000G Phase 2 AR1 data re-  
445 lease (<https://www.malariagen.net/data/ag1000g-phase-2-ar1>). Variant calls from  
446 Ag1000G Phase 2 are also available from the European Nucleotide Archive (ENA; <http://www.ebi.ac.uk/ena>) under study PRJEB36277.  
447

### 448 **Data collection and processing**

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

463     The Ag1000G variant data was functionally annotated using the SnpEff v4.1b software  
464     [47]. Non-synonymous *Vgsc* variants were identified as all variants in AgamP4.12 transcript  
465     AGAP004707-RD with a SnpEff annotation of “missense”. The *Vgsc* gene is known to  
466     exhibit alternative splicing [6], however at the time of writing the *An. gambiae* gene  
467     annotations did not include the alternative transcripts reported by Davies et al. We wrote  
468     a Python script to check for the presence of variants that are synonymous according to  
469     transcript AGAP004707-RD but non-synonymous according to one of the other transcripts  
470     present in the gene annotations or in the set reported by Davies et al. Supplementary Table  
471     1 includes the predicted effect for all SNPs that are non-synonymous in one or more of  
472     these transcripts. None of the variants that are non-synonymous in a transcript other  
473     than AGAP004707-RD were found to be above 5% frequency in any population.

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

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

489 **Haplotype networks**

490     Haplotype networks were constructed using the median-joining algorithm [29] as imple-  
491     mented in a Python module available from <https://github.com/malariagen/ag1000g-phase2-vgsc-repo>  
492     Haplotypes carrying either L995F or L995S mutations were analysed with a maximum edge

493 distance of two SNPs. Networks were rendered with the Graphviz library and a compos-  
494 ite figure constructed using Inkscape. Non-synonymous edges were highlighted using the  
495 SnpEff annotations [47].

496 **Positive selection**

497 Core haplotypes were defined on a 6,078 bp region spanning *Vgsc* codon 995, from chro-  
498 mosome arm 2L position 2,420,443 and ending at position 2,426,521. This region was  
499 chosen as it was the smallest region sufficient to differentiate between the ten genetic  
500 backgrounds carrying either of the known resistance alleles L995F or L995S. Extended  
501 haplotype homozygosity (EHH) was computed for all core haplotypes as described in [30]  
502 using scikit-allel version 1.1.9 [52], excluding non-synonymous and singleton SNPs. Analy-  
503 ses of haplotype homozygosity in moving windows (Supplementary Figs. S1, S2) and pair-  
504 wise haplotype sharing (Supplementary Figure S3) were performed using custom Python  
505 code available from <https://github.com/malariagen/ag1000g-phase2-vgsc-report>.

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

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

517 **Homology modelling**

518 A homology model of the *An. gambiae* voltage-gated sodium channel (AGAP004707-RD  
519 AgamP4.12) was generated using the 3.8 Å resolution structure of the *Periplaneta amer-*  
520 *icana* sodium channel Na<sub>v</sub>PaS structure (PDB code 5X0M) [25]. Sequences were aligned

521 using Clustal Omega [54]. 50 starting models were generated using MODELLER [55].  
522 The internal scoring function of MODELLER was used to select 10 models, which were  
523 visually inspected and submitted to the VADAR webserver [56] to assess stereochemistry  
524 in order to select the best final model. Figures were produced using PyMOL (DeLano  
525 Scientific, San Carlos, CA, USA).

526 **References**

- 527 [1] S. Bhatt et al. ‘The effect of malaria control on Plasmodium falciparum in Africa  
528 between 2000 and 2015’. In: *Nature* 526.7572 (2015), pp. 207–211. ISSN: 0028-0836.  
529 arXiv: [arXiv:1011.1669v3](https://arxiv.org/abs/1011.1669v3).
- 530 [2] Janet Hemingway et al. ‘Averting a malaria disaster: Will insecticide resistance derail  
531 malaria control?’ In: *The Lancet* 387.10029 (2016), pp. 1785–1788. ISSN: 1474547X.
- 532 [3] World Health Organization. *Global Plan for Insecticide Resistance Management  
(GPIRM)*. Tech. rep. Geneva: World Health Organization, 2012.
- 533 [4] World Health Organization et al. ‘Global vector control response 2017-2030.’ In:  
534 *Global vector control response 2017-2030.* (2017).
- 535 [5] Ke Dong et al. ‘Molecular biology of insect sodium channels and pyrethroid resis-  
536 tance’. In: *Insect Biochemistry and Molecular Biology* 50.1 (2014), pp. 1–17. ISSN:  
537 09651748.
- 538 [6] T. G.E. Davies et al. ‘A comparative study of voltage-gated sodium channels in the  
539 Insecta: Implications for pyrethroid resistance in Anopheline and other Neopteran  
540 species’. In: *Insect Molecular Biology* 16.3 (2007), pp. 361–375. ISSN: 09621075.
- 541 [7] D. Martinez-Torres et al. ‘Molecular characterization of pyrethroid knockdown resis-  
542 tance (kdr) in the major malaria vector *Anopheles gambiae* s.s.’ In: *Insect Molecular  
543 Biology* 7.2 (1998), pp. 179–184. ISSN: 09621075.
- 544 [8] Ana Paula B Silva et al. ‘Mutations in the voltage-gated sodium channel gene of  
545 anophelines and their association with resistance to pyrethroids: a review’. In: *Par-  
546 asites & Vectors* 7.1 (2014), p. 450. ISSN: 1756-3305.

- 548 [9] H. Ranson et al. ‘Identification of a point mutation in the voltage-gated sodium  
549 channel gene of Kenyan Anopheles gambiae associated with resistance to DDT and  
550 pyrethroids’. In: *Insect Molecular Biology* 9.5 (2000), pp. 491–497. ISSN: 09621075.
- 551 [10] Martin S. Williamson et al. ‘Identification of mutations in the housefly para-type  
552 sodium channel gene associated with knockdown resistance (kdr) to pyrethroid in-  
553 secticides’. In: *Molecular and General Genetics* 252.1-2 (1996), pp. 51–60. ISSN:  
554 00268925.
- 555 [11] Christopher M Jones et al. ‘Footprints of positive selection associated with a mu-  
556 tation (N1575Y) in the voltage-gated sodium channel of Anopheles gambiae.’ In:  
557 *Proceedings of the National Academy of Sciences of the United States of America*  
558 109.17 (2012), pp. 6614–9. ISSN: 1091-6490.
- 559 [12] T. G. E. Davies et al. ‘DDT, pyrethrins, pyrethroids and insect sodium channels’.  
560 In: *IUBMB Life* 59.3 (2007), pp. 151–162. ISSN: 1521-6543.
- 561 [13] Frank D. Rinkevich, Yuzhe Du and Ke Dong. ‘Diversity and convergence of sodium  
562 channel mutations involved in resistance to pyrethroids’. In: *Pesticide Biochemistry  
563 and Physiology* 106.3 (2013), pp. 93–100. ISSN: 00483575. arXiv: NIHMS150003.
- 564 [14] J Pinto et al. ‘Multiple origins of knockdown resistance mutations in the Afrotropical  
565 mosquito vector Anopheles gambiae’. In: *PLoS One* 2 (2007), e1243. ISSN: 19326203.
- 566 [15] Josiane Etang et al. ‘Polymorphism of intron-1 in the voltage-gated sodium channel  
567 gene of Anopheles gambiae s.s. populations from cameroon with emphasis on insecti-  
568 cide knockdown resistance mutations’. In: *Molecular Ecology* 18.14 (2009), pp. 3076–  
569 3086. ISSN: 09621083.
- 570 [16] Amy Lynd et al. ‘Field, genetic, and modeling approaches show strong positive  
571 selection acting upon an insecticide resistance mutation in anopheles gambiae s.s.’  
572 In: *Molecular Biology and Evolution* 27.5 (2010), pp. 1117–1125. ISSN: 07374038.
- 573 [17] Federica Santolamazza et al. ‘Remarkable diversity of intron-1 of the para voltage-  
574 gated sodium channel gene in an Anopheles gambiae/Anopheles coluzzii hybrid  
575 zone.’ In: *Malaria journal* 14.1 (2015), p. 9. ISSN: 1475-2875.

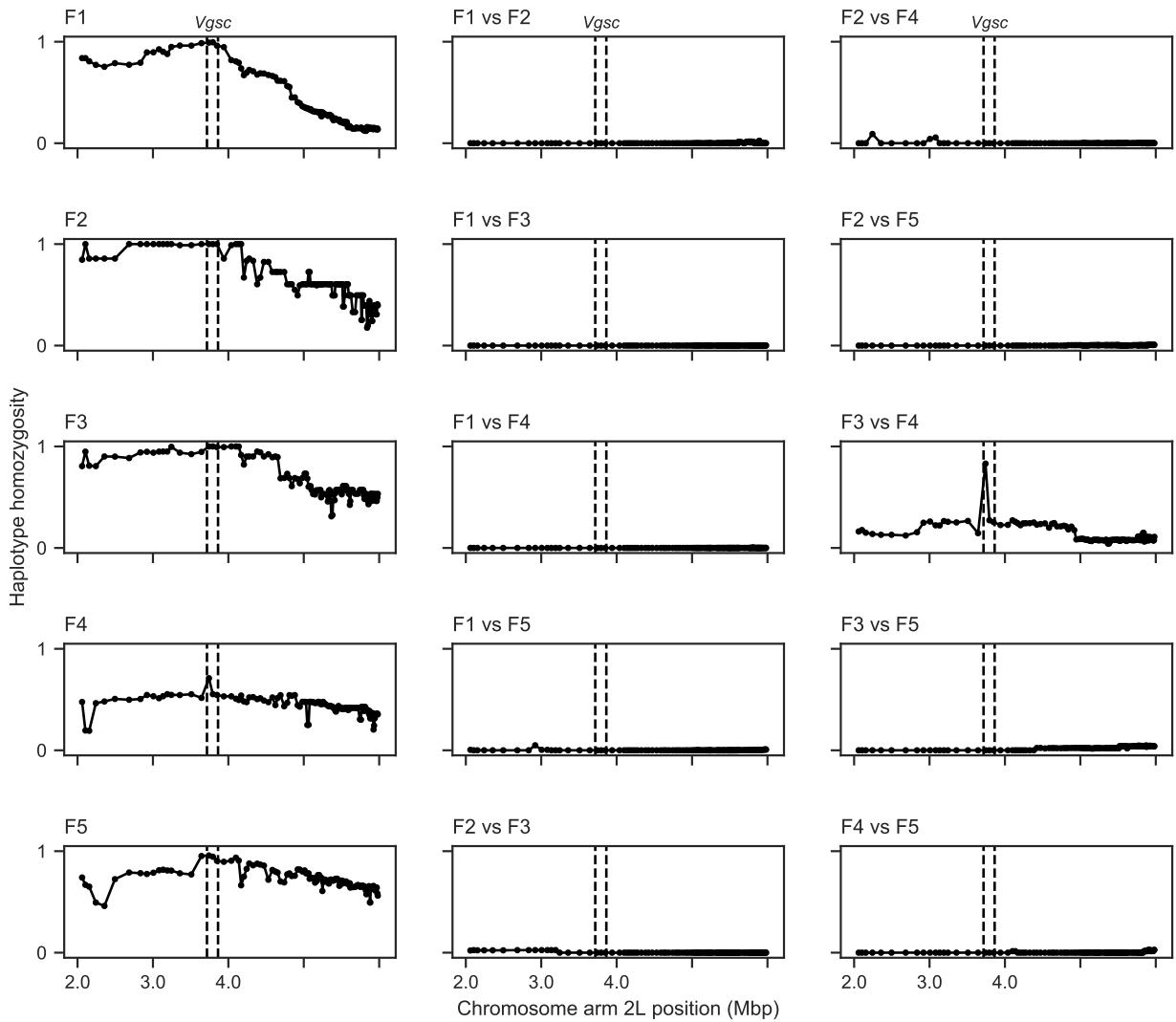
- 576 [18] Mylène Weill et al. ‘The kdr mutation occurs in the Mopti form of Anopheles gam-  
577 biae s. through introgression’. In: *Insect molecular biology* 9.5 (2000), pp. 451–455.
- 578 [19] Abdoulaye Diabaté et al. ‘The spread of the Leu-Phe kdr mutation through Anophe-  
579 les gambiae complex in Burkina Faso: genetic introgression and de novo phenomena’.  
580 In: *Tropical Medicine & International Health* 9.12 (2004), pp. 1267–1273.
- 581 [20] Chris S. Clarkson et al. ‘Adaptive introgression between Anopheles sibling species  
582 eliminates a major genomic island but not reproductive isolation’. In: *Nature Com-  
583 munications* 5 (2014). ISSN: 2041-1723.
- 584 [21] Laura C. Norris et al. ‘Adaptive introgression in an African malaria mosquito coin-  
585 cident with the increased usage of insecticide-treated bed nets’. In: *Proceedings of  
586 the National Academy of Sciences* (2015), p. 201418892. ISSN: 0027-8424.
- 587 [22] Chris S Clarkson et al. ‘Genome variation and population structure among 1,142  
588 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles  
589 coluzzii’. In: *bioRxiv* (2019), p. 864314.
- 590 [23] The Anopheles gambiae 1000 Genomes Consortium. ‘Natural diversity of the malaria  
591 vector Anopheles gambiae’. In: *Nature* 552 (2017), pp. 96–100.
- 592 [24] Shoji Sonoda et al. ‘Genomic organization of the para-sodium channel ??-subunit  
593 genes from the pyrethroid-resistant and -susceptible strains of the diamondback  
594 moth’. In: *Archives of Insect Biochemistry and Physiology* 69.1 (2008), pp. 1–12.  
595 ISSN: 07394462.
- 596 [25] Huaizong Shen et al. ‘Structure of a eukaryotic voltage-gated sodium channel at  
597 near-atomic resolution’. In: *Science* (2017), eaal4326.
- 598 [26] L Wang et al. ‘A mutation in the intracellular loop III/IV of mosquito sodium  
599 channel synergizes the effect of mutations in helix IIS6 on pyrethroid resistance’. In:  
600 *Molecular Pharmacology* 87.3 (2015), pp. 421–429.
- 601 [27] Yuzhe Du et al. ‘Molecular evidence for dual pyrethroid-receptor sites on a mosquito  
602 sodium channel’. In: *Proceedings of the National Academy of Sciences* 110.29 (2013),  
603 pp. 11785–11790.

- 604 [28] Kobié H. Toé et al. ‘Increased pyrethroid resistance in malaria vectors and decreased  
605 bed net effectiveness Burkina Faso’. In: *Emerging Infectious Diseases* 20.10 (2014),  
606 pp. 1691–1696. ISSN: 10806059.
- 607 [29] H. J. Bandelt, P. Forster and A. Rohl. ‘Median-joining networks for inferring in-  
608 traspecific phylogenies’. In: *Molecular Biology and Evolution* 16.1 (1999), pp. 37–48.  
609 ISSN: 0737-4038.
- 610 [30] Pardis C. Sabeti et al. ‘Detecting recent positive selection in the human genome from  
611 haplotype structure’. In: *Nature* 419.6909 (2002), pp. 832–837. ISSN: 0028-0836.
- 612 [31] Molly C Reid and F Ellis McKenzie. ‘The contribution of agricultural insecticide use  
613 to increasing insecticide resistance in African malaria vectors’. In: *Malaria journal*  
614 15.1 (2016), p. 107.
- 615 [32] Sara A Abuelmaali et al. ‘Impacts of agricultural practices on insecticide resistance  
616 in the malaria vector Anopheles arabiensis in Khartoum State, Sudan’. In: *PLoS*  
617 *One* 8.11 (2013), e80549.
- 618 [33] Zhaonong Hu et al. ‘A sodium channel mutation identified in Aedes aegypti se-  
619 lectively reduces cockroach sodium channel sensitivity to type I, but not type II  
620 pyrethroids’. In: *Insect biochemistry and molecular biology* 41.1 (2011), pp. 9–13.
- 621 [34] Andrias O. O’Reilly et al. ‘Modelling insecticide-binding sites in the voltage-gated  
622 sodium channel’. In: *Biochemical Journal* 396.2 (2006), pp. 255–263. ISSN: 0264-6021.
- 623 [35] Jiabao Xu et al. ‘Multi-country survey revealed prevalent and novel F1534S muta-  
624 tion in voltage-gated sodium channel (VGSC) gene in Aedes albopictus’. In: *PLoS*  
625 *neglected tropical diseases* 10.5 (2016), e0004696.
- 626 [36] Intan H Ishak et al. ‘Contrasting patterns of insecticide resistance and knockdown  
627 resistance (kdr) in the dengue vectors Aedes aegypti and Aedes albopictus from  
628 Malaysia’. In: *Parasites & vectors* 8.1 (2015), p. 181.
- 629 [37] Yiji Li et al. ‘Evidence for multiple-insecticide resistance in urban Aedes albopictus  
630 populations in southern China’. In: *Parasites & vectors* 11.1 (2018), p. 4.

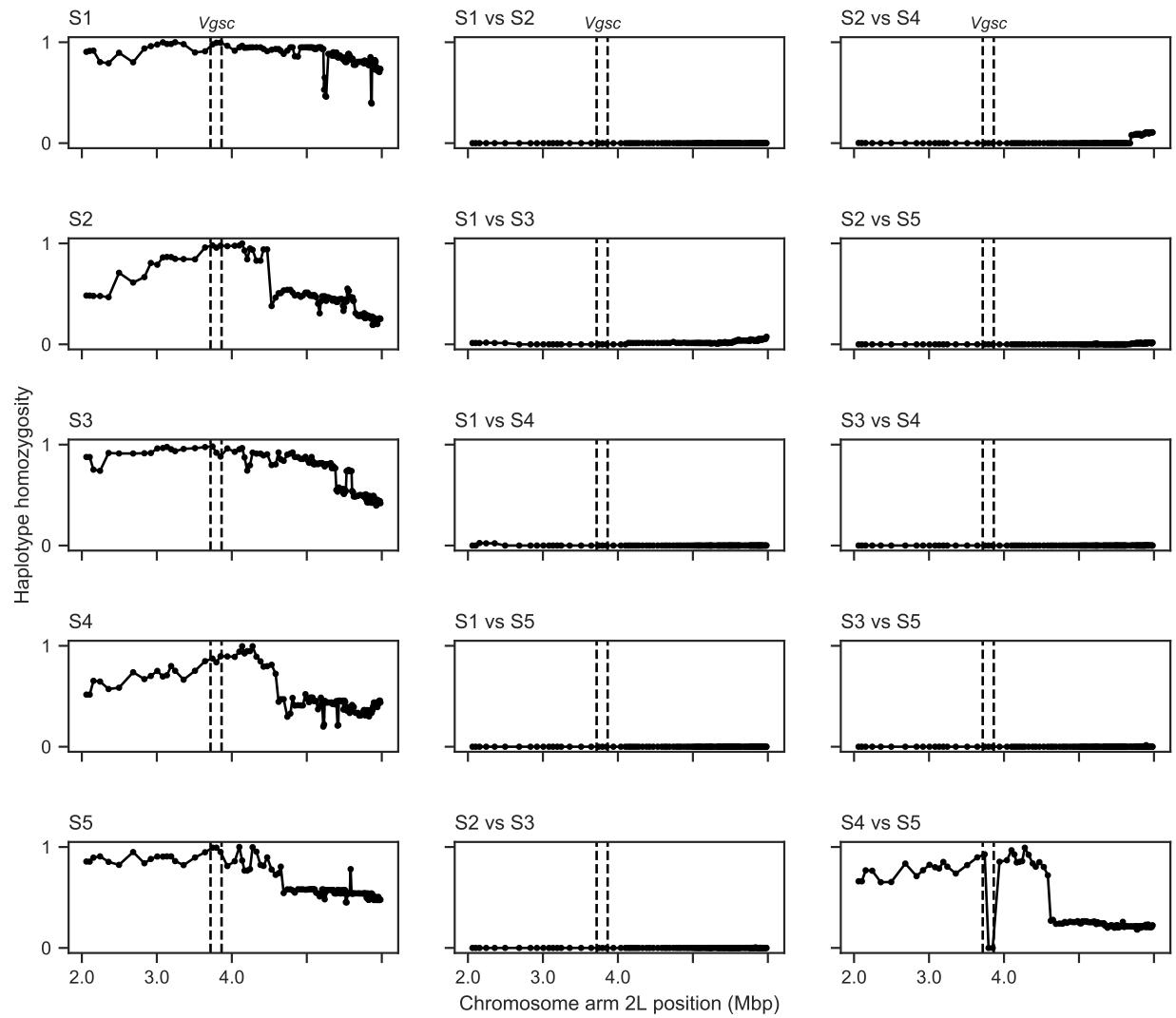
- 631 [38] Yuzhe Du et al. ‘Rotational Symmetry of Two Pyrethroid Receptor Sites in the  
632 Mosquito Sodium Channel’. In: *Molecular Pharmacology* 88.2 (Aug. 2015), pp. 273–  
633 280. ISSN: 1521-0111.
- 634 [39] H Vais et al. ‘Activation of Drosophila sodium channels promotes modification by  
635 deltamethrin. Reductions in affinity caused by knock-down resistance mutations’. In:  
636 *The Journal of general physiology* 115.3 (Mar. 2000), pp. 305–318. ISSN: 0022-1295.
- 637 [40] Deborah L. Capes et al. ‘Domain IV voltage-sensor movement is both sufficient and  
638 rate limiting for fast inactivation in sodium channels’. In: *The Journal of General  
639 Physiology* 142.2 (Aug. 2013), pp. 101–112. ISSN: 1540-7748.
- 640 [41] Zhen Yan et al. ‘Structure of the Nav1.4-B1 Complex from Electric Eel’. In: *Cell*  
641 170.3 (27th July 2017), 470–482.e11. ISSN: 0092-8674.
- 642 [42] Altin Sula et al. ‘The complete structure of an activated open sodium channel’. In:  
643 *Nature Communications* 8 (16th Feb. 2017), p. 14205. ISSN: 2041-1723.
- 644 [43] P N R Usherwood et al. ‘Mutations in DIIS5 and the DIIS4-S5 linker of Drosophila  
645 melanogaster sodium channel define binding domains for pyrethroids and DDT’. In:  
646 *FEBS letters* 581.28 (27th Nov. 2007), pp. 5485–5492. ISSN: 0014-5793.
- 647 [44] Andy Kilianski et al. ‘Bacterial and viral identification and differentiation by am-  
648 plicon sequencing on the MinION nanopore sequencer.’ In: *GigaScience* 4 (2015),  
649 p. 12. ISSN: 2047-217X.
- 650 [45] Eric R Lucas et al. ‘A high throughput multi-locus insecticide resistance marker  
651 panel for tracking resistance emergence and spread in *Anopheles gambiae*’. In: *BioRxiv*  
652 (2019), p. 592279.
- 653 [46] R A Holt et al. ‘The genome sequence of the malaria mosquito *Anopheles gambiae*’.  
654 In: *Science* 298.5591 (2002), pp. 129–149. ISSN: 0036-8075.
- 655 [47] Pablo Cingolani et al. ‘A program for annotating and predicting the effects of single  
656 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*  
657 strain w1118; iso-2; iso-3’. In: *Fly* 6.2 (2012), pp. 80–92. ISSN: 19336942.

- 658 [48] Sudhir Kumar, Glen Stecher and Koichiro Tamura. ‘MEGA7: Molecular Evolution-  
659       ary Genetics Analysis Version 7.0 for Bigger Datasets’. In: *Molecular biology and*  
660       *evolution* 33.7 (2016), pp. 1870–1874. ISSN: 15371719.
- 661 [49] Olivier Delaneau et al. ‘Haplotype estimation using sequencing reads’. In: *American*  
662       *Journal of Human Genetics* 93.4 (2013), pp. 687–696. ISSN: 00029297.
- 663 [50] Androniki Menelaou and Jonathan Marchini. ‘Genotype calling and phasing using  
664       next-generation sequencing reads and a haplotype scaffold’. In: *Bioinformatics* 29.1  
665       (2013), pp. 84–91. ISSN: 13674803.
- 666 [51] R. C. Lewontin. ‘The Interaction of Selection and Linkage. I. General Considerations;  
667       Heterotic Models’. In: *Genetics* 49.1 (1964), pp. 49–67. ISSN: 0016-6731.
- 668 [52] Alistair Miles and Nicholas Harding. *scikit-allel: A Python package for exploring and*  
669       *analysing genetic variation data*. 2016.
- 670 [53] F. Pedregosa et al. ‘Scikit-learn: Machine Learning in Python’. In: *Journal of Ma-*  
671       *chine Learning Research* 12 (2011), pp. 2825–2830.
- 672 [54] Fabian Sievers et al. ‘Fast, scalable generation of high-quality protein multiple se-  
673       quence alignments using Clustal Omega’. In: *Molecular systems biology* 7 (2011),  
674       p. 539. ISSN: 1744-4292.
- 675 [55] Narayanan Eswar et al. ‘Comparative protein structure modeling using MODELLER’.  
676       In: *Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et*  
677       *Al.] Chapter 2 (Nov. 2007), Unit 2.9. ISSN: 1934-3663.*
- 678 [56] Leigh Willard et al. ‘VADAR: a web server for quantitative evaluation of protein  
679       structure quality’. In: *Nucleic Acids Research* 31.13 (1st July 2003), pp. 3316–3319.

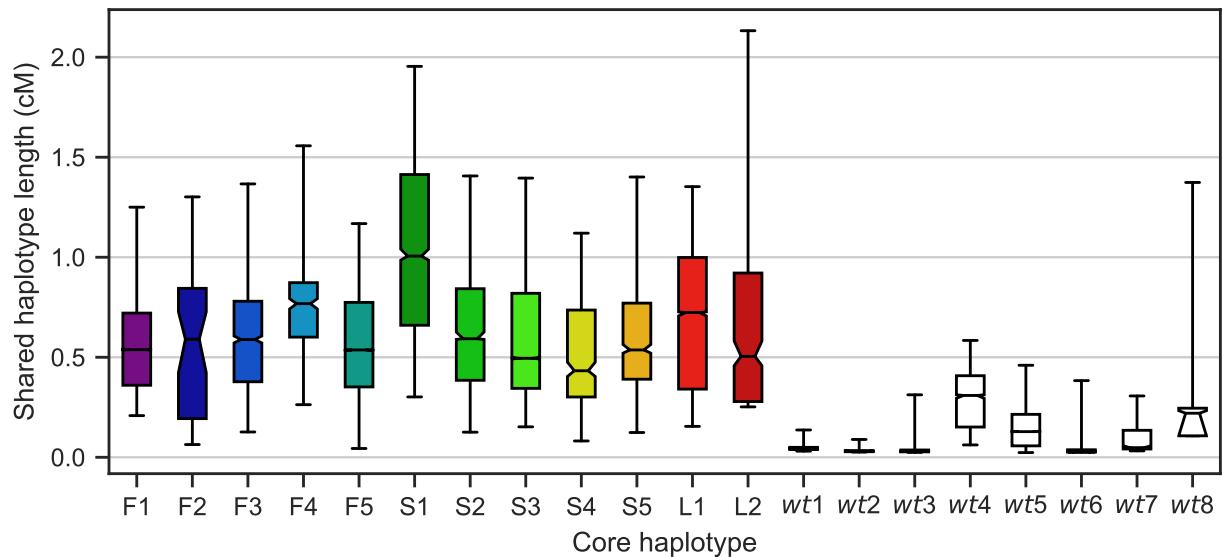
680 **Supplementary figures**



**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.