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

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

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA

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

[3]Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA

[4]https://www.malariagen.net/projects/ag1000g#people

[*]These authors contributed equally

26th January 2018

Abstract

Resistance to pyrethroid insecticides is a major concern for malaria vector control, because these are the only compounds approved for use in insecticide-treated bed-nets (ITNs), and are also widely used for indoor residual spraying (IRS). Pyrethroids target the voltage-gated sodium channel (VGSC), an essential component of the mosquito nervous system, but substitutions in the amino acid sequence can disrupt the activity of these insecticides, inducing a resistance phenotype. Here we use Illumina whole-genome sequence data from phase 1 of the *Anopheles gambiae* 1000 Genomes Project (Ag1000G) to provide a comprehensive account of genetic variation in the *Vgsc* gene in mosquito populations from 8 African countries. In addition to the three known

resistance alleles, we describe 20 non-synonymous nucleotide substitutions at appreciable frequency in one or more populations that are previously unknown in *Anopheles* mosquitoes. We analyse the genetic backgrounds on which known and putative resistance alleles are found, to determine which alleles have experienced recent positive selection, and to refine our understanding of the spread of resistance between species and geographical locations. We describe twelve distinct haplotype clusters with evidence of recent positive selection, five of which carry the `L995F` resistance allele, five of which carry `L995S`, one of which carries `I1527T`, and one of which carries `M490I`. Seven of these clusters are localised to a single geographical location, and five comprise haplotypes from two or more countries, indicating the geographical spread of resistance. We also find evidence for multiple introgression events transmitting resistance alleles between *An. gambiae* and *An. coluzzii*. We discuss potential resistance phenotypes for these novel variants based on genetic evidence for positive selection, patterns of genetic linkage between variants, location of the variant within the protein domain architecture, and functional evidence from other species. Thirteen novel non-synonymous alleles were found to occur almost exclusively on haplotypes carrying the known `L995F` resistance allele, and may be secondary mutations which could enhance or compensate for the `L995F` resistance phenotype. The `I1527T` substitution, which is adjacent to a predicted pyrethroid binding site in the channel molecule, occurs in tight linkage with either of two alleles causing a `V402L` substitution, orthologous to a combination of substitutions found to cause pyrethroid resistance in several other insect species. We also discuss how high-throughput, low-cost genetic assays for monitoring resistance can be designed using these data. Our results demonstrate that the molecular basis of pyrethroid resistance in African malaria vectors is more complex than previously appreciated, and provide a foundation for the design of new genetic tools to track the spread insecticide resistance and to inform vector control.

Introduction

Pyrethroid insecticides have been the cornerstone of malaria prevention in Africa for almost two decades [1]. Pyrethroids are still the only class of insecticide approved for use in insecticide-treated bed-nets (ITNs), and are widely used in indoor residual spraying (IRS) campaigns as well as in agriculture. Pyrethroid resistance is, however, now widespread in malaria vector populations across Africa [2]. The World Health Organisation (WHO)

has published plans for insecticide resistance management (IRM), which emphasise the need for improvements in our ability to monitor resistance, and for improvements in our understanding of the molecular mechanisms of resistance [3].

The voltage-gated sodium channel (VGSC) is the physiological target of pyrethroid insecticides, and is integral to the insect nervous system. Pyrethroid molecules bind to sites within the protein channel and prevent normal nervous system function, causing paralysis ("knock-down") and then death. However, amino acid substitutions at key positions within the protein alter the interaction with insecticide molecules, increasing the dose of insecticide required for knock-down (target-site resistance) [4, 5]. In the African malaria vectors *Anopheles gambiae* and *An. coluzzii*, three substitutions have been found to cause pyrethroid resistance. Two of these substitutions occur in codon 995[1], with `L995F` prevalent in West and Central Africa [6, 7], and `L995S` found in Central and East Africa [8, 7]. A third substitution, `N1570Y`, has been found in Central Africa and shown to increase resistance in association with `L995F` [10]. However, studies in other insect species have found a variety of other *Vgsc* substitutions inducing a resistance phenotype [11, 12, 5]. To our knowledge, no studies in malaria vectors have analysed the full *Vgsc* coding sequence, thus the molecular basis of target-site resistance to pyrethroids has not been fully explored.

Basic information is also lacking about the spread of pyrethroid resistance in malaria vectors. For example, it is not known when, where or how many times pyrethroid target-site resistance has emerged. The paths of transmission, carrying resistance alleles between mosquito populations, are also not known. Previous studies have found evidence that `L995F` occurs on several different genetic backgrounds, suggesting multiple independent outbreaks of resistance driven by this allele [13, 14, 15]. However, these studies analysed only a small gene region in a limited number of mosquito populations, and therefore had limited resolution to make inferences about genetic relationships between gene sequences (haplotypes) carrying this allele. It has also been shown that the `L995F` allele spread from *An. gambiae* to *An. coluzzii* in West Africa [16, 17]. However, both `L995F` and `L995S` now have wide geographical distributions [7], and no attempts have been made to infer or

[1] Codon numbering is given here relative to transcript `AGAP004707-RA` as defined in the AgamP4.4 gene annotations. A mapping of codon numbers from `AGAP004707-RA` to *Musca domestica*, the system in which knock-down resistance mutations were first described [9], is given in Table 1.

track the geographical spread of either allele.

Here we report an in-depth analysis of the *Vgsc* gene, using whole-genome Illumina sequence data from phase 1 of the *Anopheles gambiae* 1000 Genomes Project (Ag1000G) [18]. The Ag1000G phase 1 resource includes data on nucleotide variation in 765 wild-caught mosquitoes sampled from 8 countries, with representation of West, Central and East Africa, and of both *An. gambiae* and *An. coluzzii*. We investigate variation across the complete gene coding sequence, and report population genetic data for both known and novel non-synonymous nucleotide substitutions. We then use haplotype data from the chromosomal region spanning the *Vgsc* gene to study the genetic backgrounds carrying resistance alleles, infer the geographical spread of resistance between mosquito populations and show evidence for recent positive selection. Finally, we explore ways in which variation data from Ag1000G could be used to design high-throughput, low-cost genetic assays for surveillance of pyrethroid resistance, with the capability to differentiate and track separate resistance outbreaks.

Results

Vgsc non-synonymous nucleotide variation

To identify variants with a potentially functional role in pyrethroid resistance, we extracted single nucleotide polymorphisms (SNPs) that alter the amino acid sequence of the VGSC protein from the Ag1000G phase 1 data resource. We then computed their allele frequencies among 9 mosquito populations defined by species and country of origin. Alleles that confer resistance are expected to increase in frequency under selective pressure, and we filtered the list of potentially functional variant alleles to retain 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 not previously described in these species. We reported 15 of these novel alleles in our global analysis of the Ag1000G phase 1 data resource [18], and we extend the analyses here to incorporate a SNP which alters codon 1603 and two tri-allelic SNPs affecting codons 402 and 490.

The two known resistance alleles affecting codon 995 had the highest overall allele fre-

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

Variant				Population allele frequency (%)								
Position[1]	*Ag*[2]	*Md*[3]	Domain[4]	AO*Ac*	BF*Ac*	GN*Ag*	BF*Ag*	CM*Ag*	GA*Ag*	UG*Ag*	KE	GW
`2,390,177 G>A`	`R254K`	`R261`	`IL45`	0	0	0	0	32	21	0	0	0
`2,391,228 G>C`	`V402L`	`V410`	`IS6`	0	7	0	0	0	0	0	0	0
`2,391,228 G>T`	`V402L`	`V410`	`IS6`	0	7	0	0	0	0	0	0	0
`2,399,997 G>C`	`D466H`	`-`	`LI/II`	0	0	0	0	7	0	0	0	0
`2,400,071 G>A`	`M490I`	`M508`	`LI/II`	0	0	0	0	0	0	0	18	0
`2,400,071 G>T`	`M490I`	`M508`	`LI/II`	0	0	0	0	0	0	0	0	0
`2,416,980 C>T`	`T791M`	`T810`	`IIS1`	0	1	13	14	0	0	0	0	0
`2,422,651 T>C`	`L995S`	`L1014`	`IIS6`	0	0	0	0	15	64	100	76	0
`2,422,652 A>T`	`L995F`	`L1014`	`IIS6`	86	85	100	100	53	36	0	0	0
`2,424,384 C>T`	`A1125V`	`K1133`	`LII/III`	9	0	0	0	0	0	0	0	0
`2,425,077 G>A`	`V1254I`	`I1262`	`LII/III`	0	0	0	0	0	0	0	0	5
`2,429,617 T>C`	`I1527T`	`I1532`	`IIIS6`	0	14	0	0	0	0	0	0	0
`2,429,745 A>T`*	`N1570Y`	`N1575`	`LIII/IV`	0	26	10	22	6	0	0	0	0
`2,429,897 A>G`	`E1597G`	`E1602`	`LIII/IV`	0	0	6	4	0	0	0	0	0
`2,429,915 A>C`	`K1603T`	`K1608`	`IVS1`	0	5	0	0	0	0	0	0	0
`2,430,424 G>T`	`A1746S`	`A1751`	`IVS5`	0	0	11	13	0	0	0	0	0
`2,430,817 G>A`	`V1853I`	`V1858`	`COOH`	0	0	8	5	0	0	0	0	0
`2,430,863 T>C`	`I1868T`	`I1873`	`COOH`	0	0	18	25	0	0	0	0	0
`2,430,880 C>T`	`P1874S`	`P1879`	`COOH`	0	21	0	0	0	0	0	0	0
`2,430,881 C>T`	`P1874L`	`P1879`	`COOH`	0	7	45	26	0	0	0	0	0
`2,431,019 T>C`	`F1920S`	`Y1925`	`COOH`	0	0	0	0	1	4	0	0	0
`2,431,061 C>T`	`A1934V`	`A1939`	`COOH`	0	12	0	0	0	0	0	0	0
`2,431,079 T>C`	`I1940T`	`I1945`	`COOH`	0	4	0	0	7	0	0	0	0

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

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

[3] Codon numbering according to *Musca domestica* EMBL accession X96668 [9].

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

quencies within the Ag1000G phase 1 cohort. 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 alleles were present in *An. gambiae* populations sampled from Cameroon and Gabon, including some individuals with a hybrid L995F/S genotype (46/275 individuals in Cameroon, 36/56 in Gabon). In Cameroon these alleles were in Hardy Weinberg equilibrium (χ^2 = 0.02, P > 0.05), but there was an excess of heterozygotes in Gabon (χ^2 = 8.96, P < 0.005), suggesting a fitness advantage for mosquitoes carrying both alleles at least in some circumstances.

The N1570Y variant was present in Guinea, Burkina Faso (both species) and Cameroon. This variant has only ever been found in *An. gambiae* in association with L995F [10], and has been shown experimentally to substantially increase pyrethroid resistance when it occurs in combination with L995F [19]. To study the patterns of association among non-synonymous variants, we used haplotypes from the Ag1000G phase 1 resource to compute the normalised coefficient of linkage disequilibrium (D') between all pairs of variant alleles (Figure 1). As expected, we found N1570Y in almost perfect linkage with L995F, meaning that N1570Y was only ever found on haplotypes also carrying L995F. Of the 20 novel non-synonymous alleles, 13 also occurred almost exclusively in combination with L995F, exhibiting the same LD pattern as N1570Y (Figure 1). These included two variants in codon 1874 (P1874S, P1874L), one of which (P1874S) has previously been associated with pyrethroid resistance in the crop pest *Plutella xylostella* [20]. The abundance of high-frequency non-synonymous variants occurring in combination with L995F is striking for two reasons. First, *Vgsc* is a highly conserved gene, expected to be under strong functional constraint and therefore purifying selection, and so any non-synonymous variants should be rare [11]. Second, in contrast with L995F, we did not observe any high-frequency non-synonymous variants occuring in combination with L995S. This contrast was highly significant when data on all variants within the gene were considered: relative to haplotypes carrying the wild-type L995 allele, the ratio of non-synonymous to synonymous nucleotide diversity (π_N/π_S) was 1.5 (95% CI [0.8, 2.2]) times higher among haplotypes carrying L995S, but 28.1 (95% CI [25.2, 31.2]) times higher among haplotypes carrying L995F. These results indicate that L995F has substantially altered the selective regime for other amino acid positions within the protein. A number of secondary substitutions have occurred and risen in frequency, and therefore could be providing some selective advantage

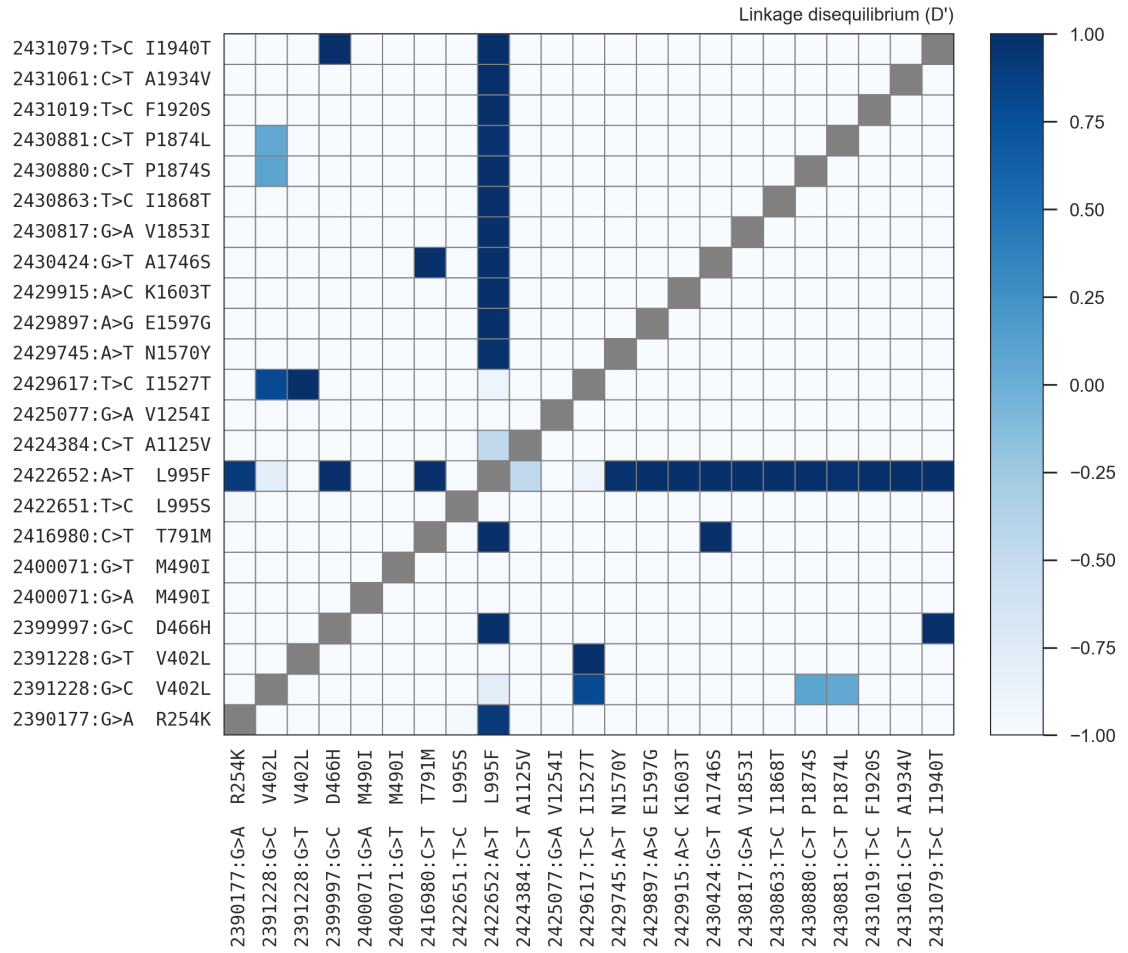

Figure 1. Linkage disequilibrium between non-synonymous variants. A value of 1 indicates that the two alleles are in perfect linkage, meaning that one of the two alleles is only ever found in combination with the other. Conversely, a value of -1 indicates that the two alleles are never found in combination with each other.

in the presence of insecticide pressure.

The `I1527T` allele was present in *An. coluzzii* from Burkina Faso at 14% frequency. Codon 1527 occurs within trans-membrane segment `IIIS6`, immediately adjacent to residues within a predicted binding site for pyrethroid molecules, thus it is plausible that `I1527T` could alter pyrethroid binding [21, 5]. We also found that the two variant alleles affecting codon 402, both of which induce a `V402L` substitution, were in strong linkage with `I1527T` ($D' \geq 0.8$; Figure 1), and almost all haplotypes carrying `I1527T` also carried a `V402L` substitution. The most parsimonious explanation for this pattern of linkage is that the `I1527T` mutation occurred first, and mutations in codon 402 subsequently arose on this genetic background or recombined with it. Substitutions in codon 402 have been found in a number of other insect species and shown experimentally to confer pyrethroid resis-

tance [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 [22], therefore `L995F` may provide a stronger resistance phenotype and is replacing `I1527T+V402L`.

The remaining 4 novel alleles (two separate nucleotide substitutions causing `M490I`; `A1125V`; `V1254I`) did not occur in combination with any known resistance allele. All are private to a single population, and (to our knowledge) none have previously been found in other species.

Genetic backgrounds carrying resistance alleles

Although it is known that pyrethroid resistance is increasing in prevalence in malaria vector populations across Africa, it has not been clear whether this is being driven by the spread of resistance alleles via gene flow, by resistance alleles emerging independently in multiple locations, or by some combination of both processes. The Ag1000G data resource provides a potentially rich source of information about the spread of insecticide resistance alleles in any given gene, because data are available not only for SNPs in gene coding regions, but also SNPs in introns and flanking intergenic regions, and in neighbouring genes. These additional variants can be used to analyse the genetic backgrounds (haplotypes) on which resistance alleles are found. If mosquitoes from different geographical locations or species carry the same resistance allele on identical or near-identical genetic backgrounds, this implies that the allele has been spread between mosquito populations by the movement and interbreeding of mosquitoes. Conversely, if the same resistance allele is found on different genetic backgrounds in different mosquito populations, this provides evidence that the allele has emerged independently in each population, either because of multiple mutational events since the introduction of pyrethroids, or because the allele was segregating at low frequency prior to the introduction of pyrethroids.

In our global analysis of the Ag1000G phase 1 resource [18], we used 1710 biallelic SNPs from within the 73.5 kbp *Vgsc* gene (1607 exonic, 103 intronic) to compute the number of SNP differences between all pairs of 1530 haplotypes derived from the 765

wild-caught mosquitoes in the phase 1 cohort. We then used pairwise genetic distances to perform hierarchical clustering, and found that haplotypes carrying resistance alleles in codon 995 were grouped into 10 distinct clusters of near-identical haplotypes. Five of these clusters carried the L995F allele (labelled F1-F5), and a further five clusters carried L995S (labelled S1-S5). To confirm these initial findings, we used the same haplotype data to construct median-joining networks (Figure 2). The network analysis is similar to hierarchical clustering, but allows 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 occured between haplotypes. Furthermore, the visualisation of these networks allows relationships among closely-related haplotypes to be discerned. We constructed these networks up to a maximum edge distance of 2 SNP differences, to ensure that each connected component in the resulting networks captures a collection of closely-related haplotypes. The resulting networks confirmed the presence of 5 distinct groupings of haplotypes carrying L995F, and a further 5 haplotype groups carrying L995S, in close correspondence with the previous results from hierarchical clustering (97.1% overall concordance in assignment of haplotypes to clusters).

The haplotype networks bring into sharp relief the explosive radiation of amino acid substitutions secondary to the L995F allele. Within the F1 network, nodes carrying non-synonymous variants radiate out from a central node carrying only L995F, suggesting that the central node represents the ancestral haplotype carrying L995F alone which initially came under selection, and these secondary variants have arisen subsequently as new mutations. Many of the nodes carrying secondary variants are large, suggesting positive selection and a putatively 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 comprises haplotypes from 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 networks also illustrate the constrasting levels of non-synonymous variation between L995F and L995S. Two non-synonymous variants are present within the L995S networks,

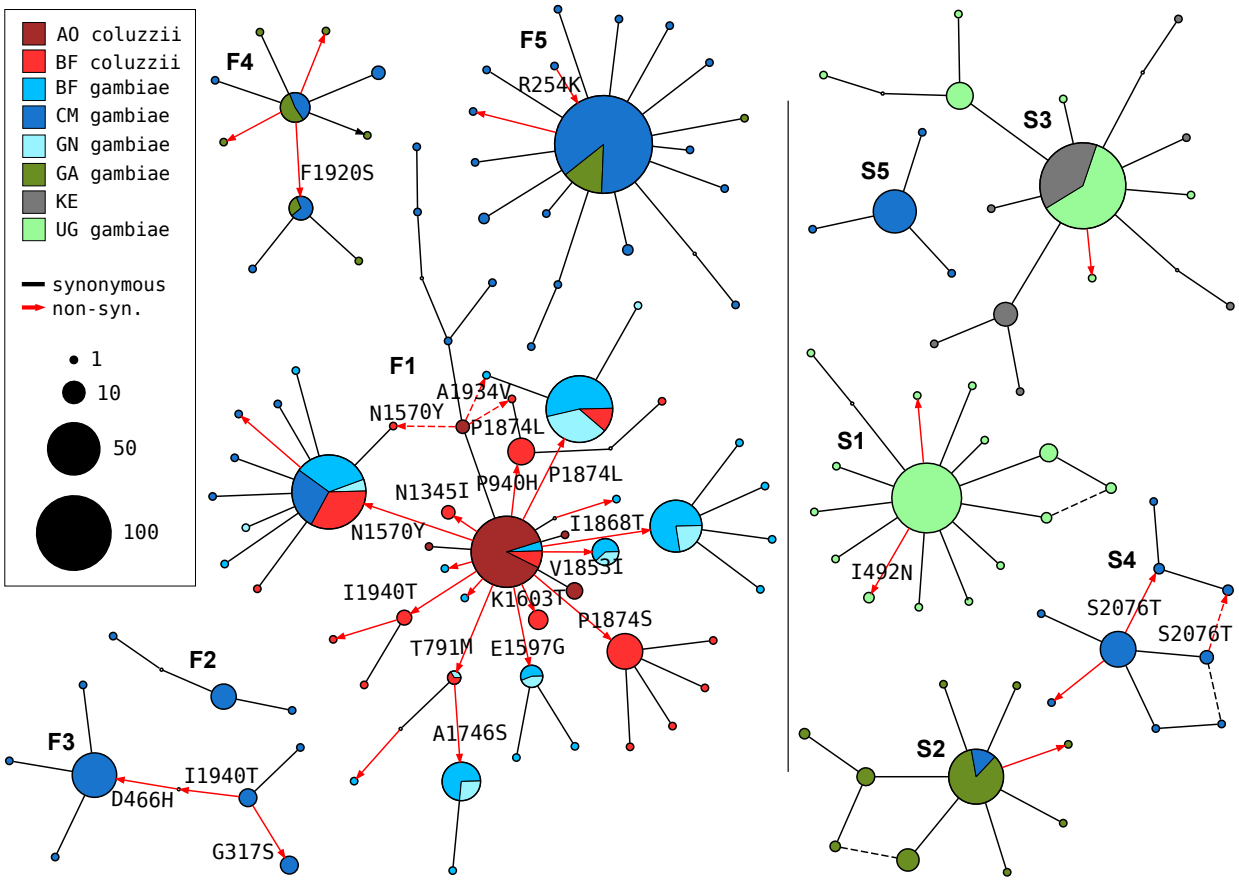

Figure 2. Haplotype networks. Median joining networks for haplotypes carrying `L995F` (labelled F1-F5) or `L995S` variants (S1-S5) with a maximum edge distance of two SNPs. Network labelling is via concordance with hierarchical clusters discovered in [18]. Node size is relative to the number of haplotypes contained and node colour represents the proportion of node haplotypes from mosquito populations/species. 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. Networks consisting of three or more haplotypes are shown.

but both are at low frequency, and thus may be neutral or mildly deleterious variants that are hitch-hiking on selective sweeps for the `L995S` allele.

As well as being found in mosquitoes of both species, F1 haplotypes were present in mosquitoes sampled from 4 different countries (Guinea, Burkina Faso, Cameroon, Angola) (Fig. 3). The F4, F5 and S2 haplotypes were each found in both Cameroon and Gabon. S3 haplotypes were present in both Uganda and Kenya. The haplotypes within each of these networks were nearly identical across the entire span of the *Vgsc* gene, and thus it is reasonable to assume that each network captures the descendants of an ancestral haplotype that has risen in frequency due to selection for insecticide resistance and subsequently accumulated other mutations. Given this assumption, these five networks each

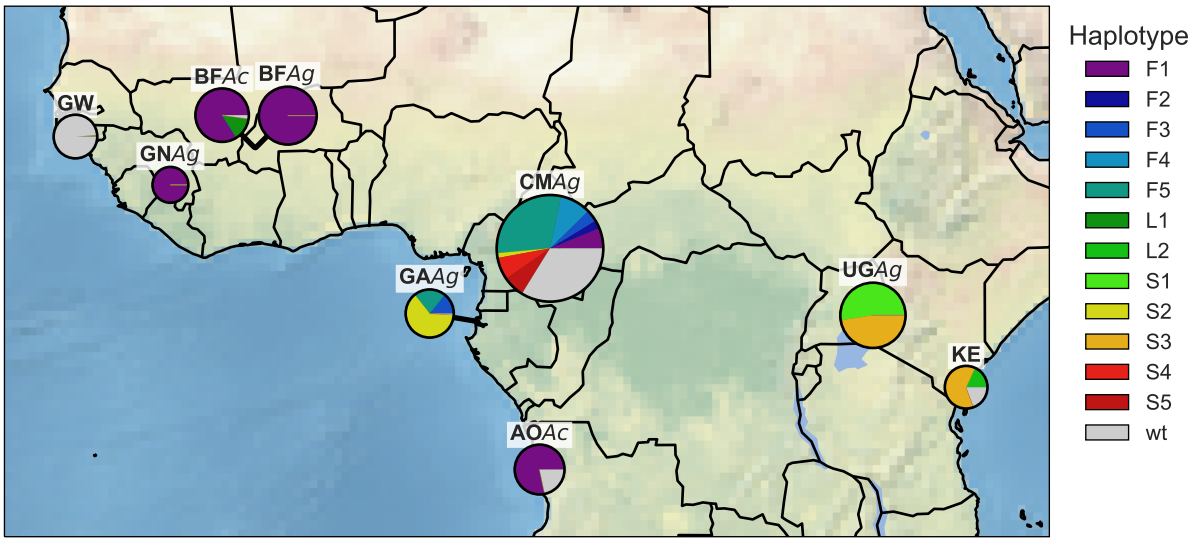

Figure 3. Map of haplotype frequencies. Each pie shows the frequency of different haplotypes within one of the populations sampled. The size of the pie is proportional to the number of haplotypes sampled. The size of each wedge within the pie is proportional to the frequency of a haplotype within the population. 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.

provide evidence for adaptive gene flow between mosquito populations separated by considerable geographical distances. However, the presence of haplotypes from two different countries within the same network does not imply direct gene flow, as haplotypes could be transmitted from or via a third location, which may be unsampled.

A limitation of both the hierarchical clustering and network analyses is that they rely on genetic distances within a fixed genomic window from the start to the end of the *Vgsc* gene. *Anopheles* mosquitoes undergo homologous recombination during meiosis in both males and females, and any recombination events that occurred within this genomic window could affect the way that haplotypes are grouped together in clusters or networks. In particular, recombination events could occur during the geographical spread of a resistance allele, altering the genetic background upstream and/or downstream of the allele itself. An analysis based on a fixed genomic window might then fail to infer gene flow between two mosquito populations, because the calculation of genetic distances does not account for recombination events, and thus haplotypes with and without the recombination event could be grouped separately. To investigate the possibility that recombination events may

have affected our findings regarding the genetic backgrounds carrying resistance alleles, we performed a windowed analysis of haplotype homozygosity, spanning *Vgsc* and up to a megabase upstream and downstream of the gene (Supplementary Figs. S1, S2). This analysis supported a refinement of our initial classification of genetic backgrounds carrying resistance alleles. All haplotypes within clusters S4 and S5 were effectively identical on both the upstream and downstream flanks of the gene, but there was a region of divergence within the *Vgsc* gene itself that separated them in the fixed window analyses (Supplementary Fig. S2). The 13.8 kbp region of divergence occurred upstream of codon 995 and contained 8 SNPs that were fixed differences between S4 and S5. A possible explanation for this short region of divergence is that a gene conversion event has occurred within the gene, bringing a short segment from a different genetic background onto the original genetic background on which the `L995S` resistance mutation occurred. All haplotypes in clusters S4 and S5 were sampled from Cameroon, and thus considering this as a single genetic background does not imply any new gene flow events.

Positive selection for resistance alleles

To confirm that known resistance alleles are under positive selection, and investigate evidence for positive selection on non-synonymous alleles discovered in this study, we performed an analysis of extended haplotype homozygosity (EHH) [23]. Haplotypes under recent positive selection are expected to 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 spanning the selected allele, relative to wild-type haplotypes. We defined a core region spanning *Vgsc* codon 995 and an additional 6 kbp of flanking sequence. 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 genetic backgrounds carrying `L995F` and `L995S` alleles identified above, as well as a core haplotype carrying `I1527T` which we labelled L1. We also found a core haplotype corresponding to a collection 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 4).

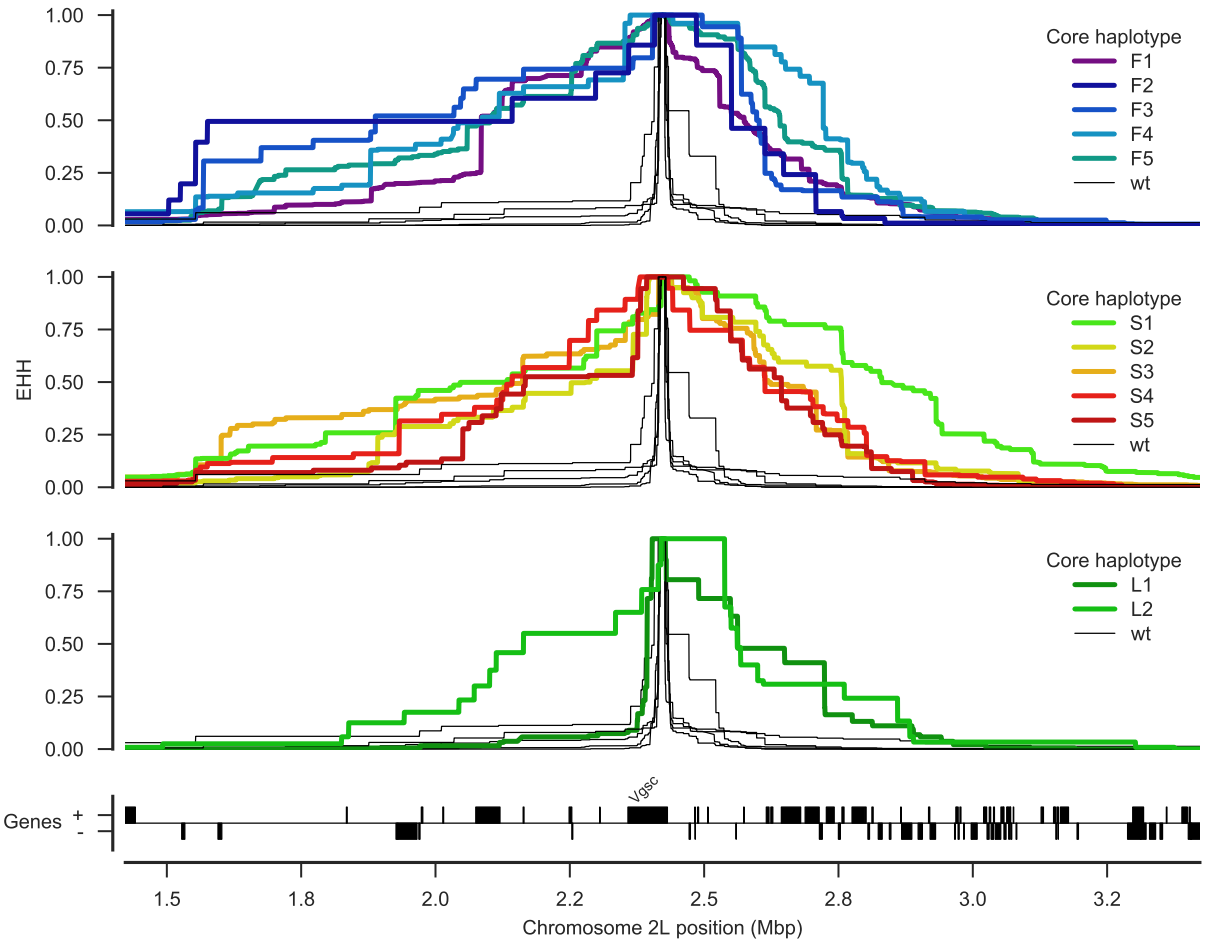

Figure 4. Evidence for positive selection on haplotypes carrying known or putative resistance alleles. Each panel plots the decay of extended haplotype homozygosity (EHH) for a set of core haplotypes centred on *Vgsc* codon 995. Core haplotypes F1-F5 carry the `L995F` allele; S1-S5 carry the `L995S` 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).

As expected, haplotypes carrying the `L995F` and `L995S` resistance alleles all experience a dramatically slower decay of EHH relative to wild-type haplotypes, confirming 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 [24]. However, we found no systematic difference in the length of shared haplotypes when comparing F1-5 (carrying `L995F`) against S1-5 (carrying `L995S`) (Supplementary Fig. S3). There were, however, some differences between core haplotypes carrying the same allele. For example, shared haplotypes were significantly longer for S1 (median 1.091 cM, 95% CI [1.076 - 1.091]) versus other core haplotypes carrying `L995S` (e.g., S2 median 0.699

cM, 95% CI [0.696 - 0.705]; Supplementary Fig. 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 different genetic backgrounds together with an ancestral haplotype carrying `I1527T`. The L2 haplotype carrying `M490I` exhibited EHH decay on both flanks comparable to haplotypes carrying known resistance alleles. This could indicate evidence for selection on the `M490I` allele, however these haplotypes are derived from a Kenyan mosquito population which is known to have experienced a severe recent bottleneck [18], and there were not enough wild-type haplotypes from Kenya with which to compare, thus this signal may also be due to the extreme demographic history of this population.

Discussion

Cross-resistance between pyrethroids and DDT

The VGSC protein is the physiological target of both pyrethroid insecticides and DDT [4]. The `L995F` and `L995S` alleles are known to increase resistance to both of these insecticide classes [6, 8]. Except in a few locations, DDT has not been used for IRS within the last two decades, and is not suitable for use in bed-nets due to potential carcinogenic effects [25]. DDT was, however, used in Africa for several pilot IRS projects carried out during the first global campaign to eradicate malaria, during the 1950s and 1960s [11]. DDT was also used in agriculture from the 1940s, and although agricultural usage has greatly diminished since the 1970s, some usage may remain [26, 25]. In this study we reported evidence of positive selection on the `L995F` and `L995S` alleles, as well as the `I1527T+V402L` combination and possibly also `M490I`. We also found 14 other non-synonymous substitutions within *Vgsc* that have arisen in association with `L995F` and appear to be positively selected.

Given that pyrethroids have dominated public health insecticide use for two decades, it is reasonable to assume that the selection pressure on these alleles is primarily due to pyrethroids. The `L995S` allele has a stronger DDT resistance phenotype than `L995F` [8], and it has been suggested that `L995S` may have been primarily selected by DDT usage [24]. However, we did not find any systematic difference in the extent of haplotype homozygosity between these two alleles, suggesting that both alleles have been under selection over a similar time frame. We did find some significant differences in haplotype homozygosity between different genetic backgrounds carrying resistance alleles, suggesting differences in the timing and/or strength of selection these may have experienced. However, there have been differences in the scale-up of pyrethroid-based interventions in different regions, and this could in turn generate heterogeneities in selection pressures. Nevertheless, it is possible that some if not all of the alleles we have reported provide some level of cross-resistance to DDT as well as pyrethroids, and that earlier DDT usage may have contributed at least in part to their selection. The differing of resistance profiles to the two types of pyrethroids (type I, e.g., permethrin; and type II, e.g., deltamethrin) [27], may also be affecting the selection landscape. Further sampling and analysis is required to investigate the timing of different selection events and relate these to historical patterns of insecticide use in different regions.

Resistance phenotypes for novel non-synonymous variants

The sodium channel protein consists of four homologous domains (I-IV) each of which comprises six transmembrane segments (S1-S6) connected by intracellular and extracellular loops [5]. Two analogous pyrethroid binding sites have been predicted within the pore-forming modules of the protein, the first (PyR1) involving residues from transmembrane segments IIS5 and IIIS6 and the internal linker between IIS4 and IIS5 (IIL45) [28], the second (PyR2) involving segments IS5, IS6, IIS6 and IL45 [21, 5]. Many of the amino acid substitutions known to cause pyrethroid resistance in insects affect residues within one of these two pyrethroid binding sites, and thus can directly alter pyrethroid binding [5]. For example, the `L995F` and `L995S` substitutions occur in segment IIS6 and belong to binding site PyR2 [21]. The `I1527T` substitution that we discovered in *An. coluzzii* mosquitoes from Burkina Faso occurs in segment IIIS6 and is immediately adjacent to two

pyrethroid-sensing residues in site PyR1 [5]. It is thus plausible that pyrethroid binding could be altered by this substitution. The I1527T substitution (*M. domestica* codon 1532) has been found in *Aedes albopictus* [29], and multiple substitutions in codon 1529 (*M. domestica* codon 1534) have been reported in *Aedes aegypti* and associated with pyrethroid resistance [5, 30]. We found the I1527T allele in tight linkage with two alleles causing a V402L substitution (*M. domestica* codon 410), with haplotype structure indicating that an initial I1527T mutation was subsequently brought together with V402L alleles from different genetic backgrounds via recombination. Substitutions in codon 402 have been found in multiple insect species and are by themselves sufficient to confer pyrethroid resistance [5]. Codon 402 is within segment IS6, immediately adjacent to a pyrethroid sensing residue in site PyR2. The fact that we find I1527T and V402L in such tight association, with V402L apparently secondary to I1527T, is intriguing because (a) these two residues appear to affect different pyrethroid binding sites, and (b) haplotypes carrying V402L alone should also have been positively selected and thus be present in one or more populations.

A number of substitutions in segments of the protein that are not involved either of the two pyrethroid binding sites have also been shown to confer pyrethroid resistance. For example, the N1570Y substitution causes substantially enhanced pyrethroid resistance when combined with L995F, although codon 1570 occurs in the internal linker between domains III and IV (LIII/IV) [21]. Computer modelling of the protein structure has suggested that substitutions in codon 1570 could allosterically alter site PyR2 and thus affect pyrethroid binding [21]. In addition to N1570Y, we found thirteen other substitutions at appreciable frequency occurring exclusively in association with L995F (Table 1). Of these, two (D466H, E1597G) occurred in the larger internal linkers between protein domains, one (R254K) occurred within a smaller internal linker between domain subunits, two (T791M, K1603T) occurred within an outer ("voltage-sensing") transmembrane segment, one (A1746S) occurred within an inner ("pore-forming") transmembrane segment, and the remaining seven occurred in the internal carboxyl-terminal tail. The novel non-synonymous mutation found on the Kenyan haplotypic background potentially under selection, M490I, also occurs in an internal linker between protein domains (L1/II). Substitutions within various locations in the protein have been shown to confer pyrethroid resistance either independently or in combination with other substitutions not by altering pyrethroid binding but by altering

the channel gating kinetics or the voltage-dependence of activation [5]. Thus there are a number of potential mechanisms by which a pyrethroid resistance phenotype can be obtained, and clearly much remains to be unravelled regarding the molecular biology of pyrethroid resistance in this gene.

Design of genetic assays for surveillance of pyrethroid resistance

Entomological surveillance teams in Africa do regularly genotype mosquitoes for resistance alleles in *Vgsc* codon 995, and use those results as an indicator for the presence of pyrethroid resistance alongside results from insecticide resistance bioassays. They typically do not, however, sequence the gene or genotype any other polymorphisms within the gene. Thus if there are other polymorphisms within the gene that cause or significantly enhance pyrethroid resistance, these will not be detected. Also, if a codon 995 resistance allele is observed, there is no way to know whether the allele is on a genetic background that has also been observed in other mosquito populations, and thus no way to investigate whether resistance alleles are emerging locally or being imported from elsewhere, and if so, what is the probable source population. Whole-genome sequencing of individual mosquitoes clearly provides data of sufficient resolution to identify resistance sweeps, and could also be used to provide ongoing resistance surveillance. The cost of whole-genome sequencing continues to fall, with the present cost being approximately 100 GBP to obtain ~30× coverage of an individual *Anopheles* mosquito genome with 150 bp paired-end reads. There is an interim period, however, during which it may be more practical to develop targeted genetic assays for resistance outbreak surveillance that could scale to tens of thousands of mosquitoes at low cost and that could be implemented using existing platforms in regional molecular biology facilities.

To facilitate the development of targeted genetic assays for surveillance of *Vgsc*-mediated pyrethroid resistance, we have produced two supplementary data tables and explore a potential process for assay design. In Supplementary Table 1 we provide a list of all @@N biallelic SNPs discovered with high confidence in the Ag1000G phase 1 cohort within the *Vgsc* gene and in the 100 kbp upstream and downstream flanking regions. To aid in PCR primer design, for each SNP we provide the flanking sequence for 250 bp upstream and downstream of the SNP position, including information about any polymorphisms within

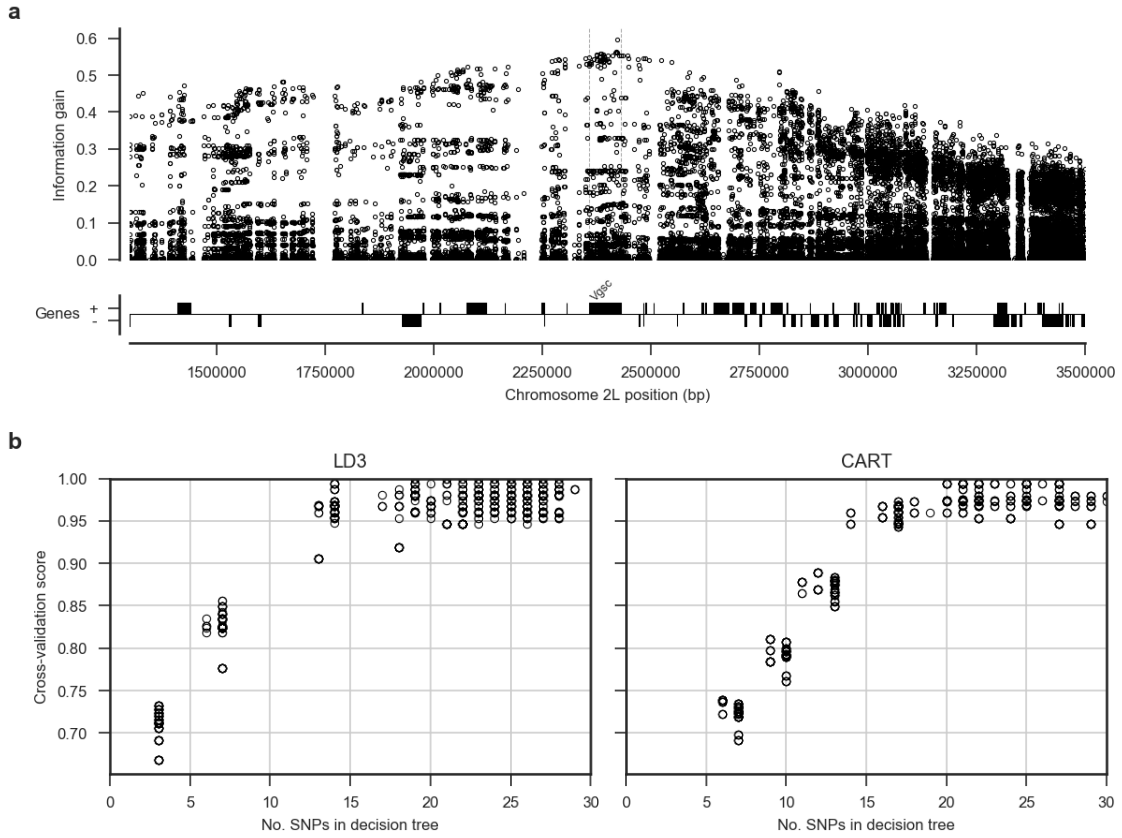

Figure 5. Informative SNPs for haplotype surveillance. **a**, Each data point represents a single SNP. The information gain value for each SNP provides an indication of how informative the SNP is likely to be if used as part of a genetic assay for testing whether a mosquito carries a resistance haplotype, and if so, which of the known resistance haplotype clusters it derives from. **b**, Number of SNPs required to accurately classify which cluster a haplotype derives from. Decision trees were constructed using either the LD3 (left) or CART (right) algorithm for comparison. Accuracy was evaluated using 10-fold stratified cross-validation.

these flanking regions. Not all SNPs are informative for detecting whether an individual mosquito carries a resistance allele, or diagnosing which genetic background is present, and we provide some summary statistics for each SNP to aid in the identification of the most informative SNPs. This includes allele frequencies for each of the 12 haplotype clusters identified here as carrying known or putative resistance alleles, as well as for wild-type haplotypes from different locations. To help with designing classifiers than can accurately call resistance haplotypes with a minimal number of SNPs, we also provide the information gain [31] and the Gini impurity [32] for each SNP. Note that recombination events are more likely at increasing distances upstream and downstream of the resistance variants under selection, and thus the most informative SNPs are found closest to the resistance variants within the gene (Figure 5). However, SNPs with some information gain are available

throughout the gene and in flanking regions.

A possible strategy for the design of a genetic assay could proceed by (1) performing an initial round of filtering to remove SNPs which are not informative (e.g., low information gain); (2) performing a round of primer design to remove SNPs for which primers are unlikely to be successful; (3) performing a full analysis of the remaining SNPs to select a subset that is sufficient to classify all resistance haplotypes identified here, including some redundancy; (4) finalise primer designs for the chosen panel of SNPs. A possible methodology for step 3 would be to use an algorithm such as ID3 [31] or CART [32] to build a decision tree, although many other algorithms for building classifiers are also applicable. To aid in the development of a classifier, in Supplementary Table 2 we provide our classification for each of the 1530 haplotypes sampled here, along with the alleles carried by each haplotype for each of the SNPs included in Supplementary Table 1. To test the methodology, we constructed decision trees using either LD3 or CART algorithms, and using all available SNPs from within the *Vgsc* plus 20 kbp flanking regions as input features (i.e., assuming primers could be designed in all cases). Figure 5b shows the cross-validation scores obtained for trees constructed allowing increasing numbers of SNPs. This analysis suggests that it should be possible to construct a decision tree able to classify these resistance 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 specific non-synonymous polymorphisms in addition to identifying known genetic backgrounds carrying resistance alleles. However, it is still likely to be well within the number of SNPs that could be assayed via a technology such as amplicon sequencing [33]. Thus it should be feasible to produce low-cost, high-throughput genetic assays for tracking the spread of pyrethroid resistance. If combined with a limited amount of whole-genome sequencing at sentinel sites, this should also allow the identification of newly emerging resistance outbreaks.

Methods

Code

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

Data

We used variant calls from the Ag1000G Phase 1 AR3 data release (`https://www.malariagen.net/data/ag1000g-phase1-ar3`) and phased haplotype data from the Ag1000G Phase 1 AR3.1 data release (`https://www.malariagen.net/data/ag1000g-phase1-ar3.1`). Variant calls from Ag1000G Phase 1 are also available from the European Nucleotide Archive (ENA; `http://www.ebi.ac.uk/ena`) under study PRJEB18691.

Data collection and processing

For detailed information on Ag1000g WGS sample collection, sequencing, variant calling, quality control and phasing see [18]. In brief, *An. gambiae* and *An. coluzzii* mosquitoes were collected from eight countries across Sub-Saharan Africa: Angola, Burkina Faso, Cameroon, Gabon, Guinea, Guinea Bissau, Kenya and Uganda. From Angola just *An. coluzzii* were sampled, Burkina Faso had samples of both *An. gambiae* and *An. coluzzii* and all other populations consisted of purely *An. gambiae* except for Kenya and Guinea Bissau, where species status is uncertain [18]. Mosquitoes were individually whole genome sequenced on the Illumina HiSeq 2000 platform, generating 100bp paired-end reads. Sequence reads were aligned to the *An. gambiae* AgamP3 reference genome assembly [34]). Aligned bam files underwent improvement, before variants were called using GATK UnifiedGenotyper. Quality control included removal of samples with mean coverage <= 14x and an accessibility map was employed following a similar approach to that used for human data by The 1000 Genomes Project Consortium [35]). Various quality control filters were applied to remove samples and SNPs with poor quality data.

The Ag1000g variant data was functionally annotated using the SnpEff v4.1b software which allowed investigation of potential phenotype altering variants within *Vgsc* [36]. Non-synonymous *Vgsc* variants were identified as all variants in transcript AGAP004707-RA with a SnpEff annotation of "missense".

For ease of comparison with previous work on *Vgsc*, pan Insecta, in Table 1 we report codon numbering for both *An. gambiae* and *Musca domestica* (the species in which the gene was first discovered). The *M. domestica Vgsc* sequence (EMBL accession X96668 [9]) was aligned with the *An. gambiae* AGAP004707-RA sequence (AgamP4.4 gene-set), using

the Mega v7 software package [37]. A map of equivalent codon numbers between the two species for the entire gene can be download from the MalariaGEN website (`https://www.malariagen.net/sites/default/files/content/blogs/domestica_gambiae_map.txt`).

Haplotypes for each chromosome of each sample were estimated (phased) using using phase informative reads (PIRs) and SHAPEIT2 v2.r837 [38], see [18] supplementary text for more details. The SHAPEIT2 algorithm is unable to phase multi-allelic positions, therefore the two multi-allelic non-synonymous SNPs within the *Vgsc* gene, altering codons `V402` and `M490`, were phased onto the haplotypes using MVNcall v1.0 [39]. Conservative filtering had removed one of the three known insecticide resistance conferring kdr variants, `N1570Y` [10]. After manual inspection of the read alignment revealed that the SNP call could be confidently made, it was added back into the data set and then also phased onto the haplotypes using MVNcall. Lewontin's D' [40] was used to compute the linkage disequilibrium (LD) between all pairs of non-synonymous *Vgsc* mutations.

Haplotype networks

Haplotype networks were constructed using the median-joining algorithm [41] as implemented in a Python module available from `https://github.com/malariagen/agam-vgsc-report`. Haplotypes carrying either `L995F` or `L995S` mutations were analysed with a maximum edge distance of two SNPs, to ensure networks contained haplotypes with recent common ancestors. Networks were rendered with the Graphviz library and a composite figure constructed using Inkscape. Non-synonymous edges were highlighted using the SnpEff annotations [36].

Positive selection

Core haplotypes were defined on a 6,078 bp region spanning *Vgsc* codon 995, from chromosome arm 2L position 2,420,443 and ending at position 2,426,521. This region was chosen as it was the smallest region sufficient to differentiate between the ten genetic backgrounds carrying either of the known resistance alleles `L995F` or `L995S`. Extended haplotype homozygosity (EHH) was computed for all core haplotypes as described in [23] using scikit-allel version 1.1.9 [42], excluding non-synonymous and singleton SNPs. Analyses of haplotype homozygosity in moving windows (Supplementary Figs. S1, S2)

and pairwise haplotype sharing (Supplementary Fig. S3) were performed using custom Python code available from `https://github.com/malariagen/agam-vgsc-report`.

Design of genetic assays for surveillance of pyrethroid resistance

To explore the feasibility of indentifying a small subset of SNPs that would be sufficient to identify each of the genetic backgrounds carrying known or putative resistance alleles, we started with an input data set of all SNPs within the *Vgsc* gene or in the flanking regions 20 kbp upstream and downstream of the gene. Each of the 1530 haplotypes in the Ag1000G Phase 1 cohort was labelled according to which core haplotype it carried, combining all core haplotypes not carrying known or putative resistance alleles together as a single "wild-type" group. Decision tree classifiers were then constructed using scikit-learn version 0.19.0 [43] for a range of maximum depths, repeating the tree construction process 10 times for each maximum depth with a different initial random state. The classification accuracy of each tree was evaluated using stratified 5-fold cross-validation.

References

[1] S. Bhatt et al. 'The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015'. In: *Nature* 526.7572 (2015), pp. 207–211. ISSN: 0028-0836. arXiv: `arXiv:1011.1669v3`.

[2] Janet Hemingway et al. 'Averting a malaria disaster: Will insecticide resistance derail malaria control?' In: *The Lancet* 387.10029 (2016), pp. 1785–1788. ISSN: 1474547X.

[3] World Health Organization. *Global Plan for Insecticide Resistance Management (GPIRM)*. Tech. rep. Geneva: World Health Organization, 2012.

[4] T. G.E. Davies et al. 'A comparative study of voltage-gated sodium channels in the Insecta: Implications for pyrethroid resistance in Anopheline and other Neopteran species'. In: *Insect Molecular Biology* 16.3 (2007), pp. 361–375. ISSN: 09621075.

[5] Ke Dong et al. 'Molecular biology of insect sodium channels and pyrethroid resistance'. In: *Insect Biochemistry and Molecular Biology* 50.1 (2014), pp. 1–17. ISSN: 09651748.

[6] D. Martinez-Torres et al. 'Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector Anopheles gambiae s.s.' In: *Insect Molecular Biology* 7.2 (1998), pp. 179–184. ISSN: 09621075.

[7] Ana Paula B Silva et al. 'Mutations in the voltage-gated sodium channel gene of anophelines and their association with resistance to pyrethroids: a review'. In: *Parasites & Vectors* 7.1 (2014), p. 450. ISSN: 1756-3305.

[8] H. Ranson et al. 'Identification of a point mutation in the voltage-gated sodium channel gene of Kenyan Anopheles gambiae associated with resistance to DDT and pyrethroids'. In: *Insect Molecular Biology* 9.5 (2000), pp. 491–497. ISSN: 09621075.

[9] Martin S. Williamson et al. 'Identification of mutations in the housefly para-type sodium channel gene associated with knockdown resistance (kdr) to pyrethroid insecticides'. In: *Molecular and General Genetics* 252.1-2 (1996), pp. 51–60. ISSN: 00268925.

[10] Christopher M Jones et al. 'Footprints of positive selection associated with a mutation (N1575Y) in the voltage-gated sodium channel of Anopheles gambiae.' In: *Proceedings of the National Academy of Sciences of the United States of America* 109.17 (2012), pp. 6614–9. ISSN: 1091-6490.

[11] T. G. E. Davies et al. 'DDT, pyrethrins, pyrethroids and insect sodium channels'. In: *IUBMB Life* 59.3 (2007), pp. 151–162. ISSN: 1521-6543.

[12] Frank D. Rinkevich, Yuzhe Du and Ke Dong. 'Diversity and convergence of sodium channel mutations involved in resistance to pyrethroids'. In: *Pesticide Biochemistry and Physiology* 106.3 (2013), pp. 93–100. ISSN: 00483575. arXiv: NIHMS150003.

[13] J Pinto et al. 'Multiple origins of knockdown resistance mutations in the Afrotropical mosquito vector Anopheles gambiae'. In: *PLoS One* 2 (2007), e1243. ISSN: 19326203.

[14] Josiane Etang et al. 'Polymorphism of intron-1 in the voltage-gated sodium channel gene of Anopheles gambiae s.s. populations from cameroon with emphasis on insecticide knockdown resistance mutations'. In: *Molecular Ecology* 18.14 (2009), pp. 3076–3086. ISSN: 09621083.

[15] Federica Santolamazza et al. 'Remarkable diversity of intron-1 of the para voltage-gated sodium channel gene in an Anopheles gambiae/Anopheles coluzzii hybrid zone.' In: *Malaria journal* 14.1 (2015), p. 9. ISSN: 1475-2875.

[16] 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.

[17] 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.

[18] The Anopheles gambiae 1000 Genomes Consortium. 'Natural diversity of the malaria vector Anopheles gambiae'. In: *Nature* 552 (2017), pp. 96–100.

[19] 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.

[20] Shoji Sonoda et al. 'Genomic organization of the para-sodium channel ??-subunit genes from the pyrethroid-resistant and -susceptible strains of the diamondback moth'. In: *Archives of Insect Biochemistry and Physiology* 69.1 (2008), pp. 1–12. ISSN: 07394462.

[21] 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.

[22] 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.

[23] Pardis C. Sabeti et al. 'Detecting recent positive selection in the human genome from haplotype structure'. In: *Nature* 419.6909 (2002), pp. 832–837. ISSN: 0028-0836.

[24] Amy Lynd et al. 'Field, genetic, and modeling approaches show strong positive selection acting upon an insecticide resistance mutation in anopheles gambiae s.s.' In: *Molecular Biology and Evolution* 27.5 (2010), pp. 1117–1125. ISSN: 07374038.

[25] Vladimir Turusov, Valery Rakitsky and Lorenzo Tomatis. 'Dichlorodiphenyltrichloroethane (DDT): ubiquity, persistence, and risks.' In: *Environmental health perspectives* 110.2 (2002), p. 125.

[26] Thomas Dunlap. *DDT: scientists, citizens, and public policy*. Princeton University Press, 1981.

[27] Zhaonong Hu et al. 'A sodium channel mutation identified in Aedes aegypti selectively reduces cockroach sodium channel sensitivity to type I, but not type II pyrethroids'. In: *Insect biochemistry and molecular biology* 41.1 (2011), pp. 9–13.

[28] Andrias O. O'Reilly et al. 'Modelling insecticide-binding sites in the voltage-gated sodium channel'. In: *Biochemical Journal* 396.2 (2006), pp. 255–263. ISSN: 0264-6021.

[29] Jiabao Xu et al. 'Multi-country survey revealed prevalent and novel F1534S mutation in voltage-gated sodium channel (VGSC) gene in Aedes albopictus'. In: *PLoS neglected tropical diseases* 10.5 (2016), e0004696.

[30] Yiji Li et al. 'Evidence for multiple-insecticide resistance in urban Aedes albopictus populations in southern China'. In: *Parasites & vectors* 11.1 (2018), p. 4.

[31] J. R. Quinlan. 'Induction of decision trees'. In: *Machine Learning* 1.1 (1986), pp. 81–106. ISSN: 0885-6125.

[32] L Breiman et al. *Classification and Regression Trees*. Vol. 19. 1984, p. 368. ISBN: 0412048418.

[33] Andy Kilianski et al. 'Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer.' In: *GigaScience* 4 (2015), p. 12. ISSN: 2047-217X.

[34] R A Holt et al. 'The genome sequence of the malaria mosquito Anopheles gambiae'. In: *Science* 298.5591 (2002), pp. 129–149. ISSN: 0036-8075.

[35] The 1000 Genomes Project Consortium. 'A map of human genome variation from population-scale sequencing.' In: *Nature* 467.7319 (2010), pp. 1061–73. ISSN: 1476-4687. arXiv: 1302.2710v1.

[36] Pablo Cingolani et al. 'A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3'. In: *Fly* 6.2 (2012), pp. 80–92. ISSN: 19336942.

[37] Sudhir Kumar, Glen Stecher and Koichiro Tamura. 'MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets'. In: *Molecular biology and evolution* 33.7 (2016), pp. 1870–1874. ISSN: 15371719.

[38] Olivier Delaneau et al. 'Haplotype estimation using sequencing reads'. In: *American Journal of Human Genetics* 93.4 (2013), pp. 687–696. ISSN: 00029297.

[39] Androniki Menelaou and Jonathan Marchini. 'Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold'. In: *Bioinformatics* 29.1 (2013), pp. 84–91. ISSN: 13674803.

[40] R. C. Lewontin. 'The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models'. In: *Genetics* 49.1 (1964), pp. 49–67. ISSN: 0016-6731.

[41] H. J. Bandelt, P. Forster and A. Rohl. 'Median-joining networks for inferring intraspecific phylogenies'. In: *Molecular Biology and Evolution* 16.1 (1999), pp. 37–48. ISSN: 0737-4038.

[42] Alistair Miles and Nicholas Harding. *scikit-allel: A Python package for exploring and analysing genetic variation data*. 2016.

[43] F. Pedregosa et al. 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

Supplementary figures

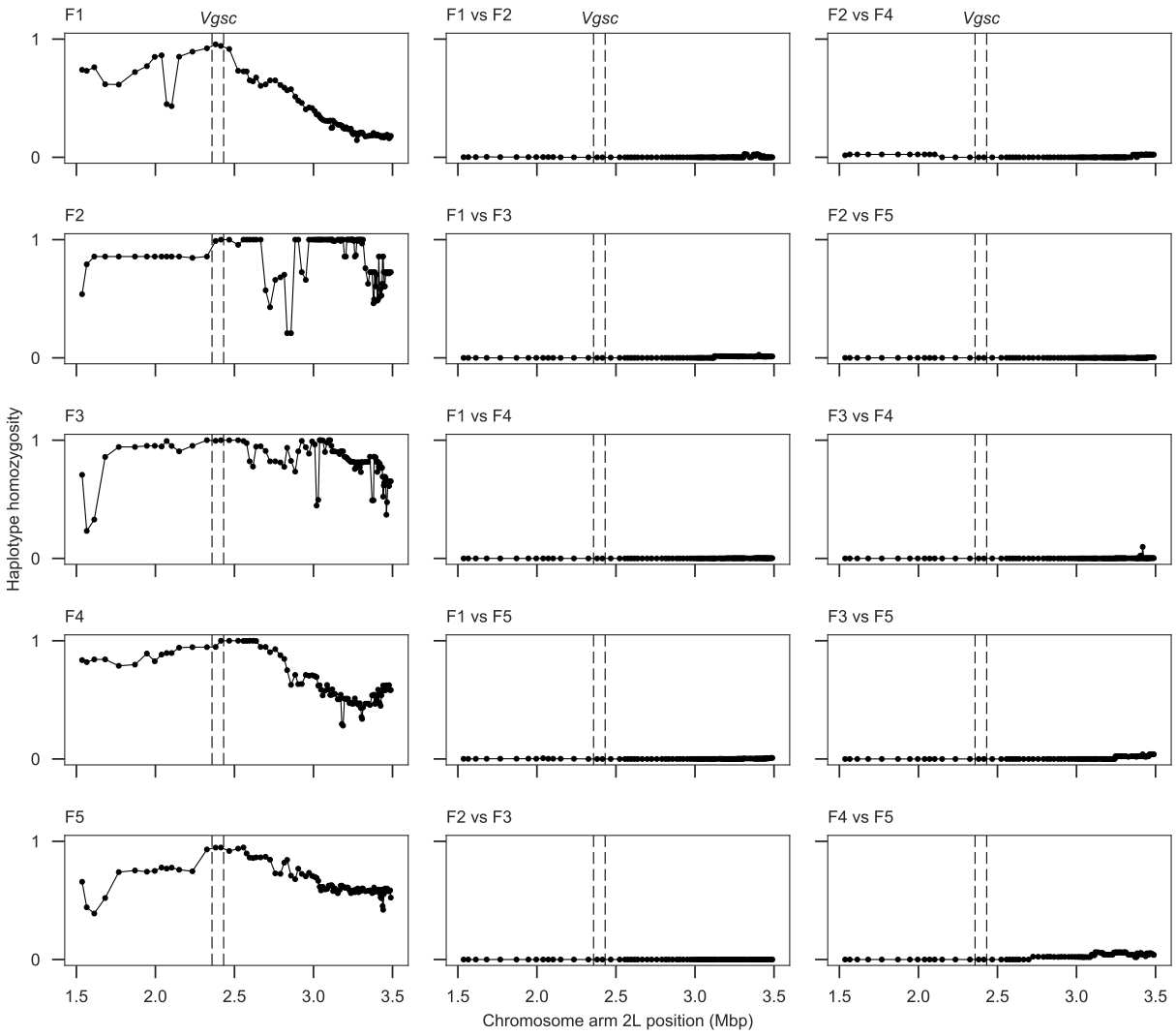

Figure S1. Windowed analysis of haplotype homozygosity for genetic backgrounds carrying the `L995F` allele. Each sub-plot shows the fraction of haplotype pairs that are identical within half-overlapping moving windows of 1000 SNPs. Each sub-plot in the left-hand column shows homozygosity for haplotype pairs within one of the haplotype clusters identified by the hierarchical clustering and network analyses. Sub-plots in the central and right-hand columns show homozygosity for haplotype pairs between two haplotype clusters. If two haplotype clusters 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.

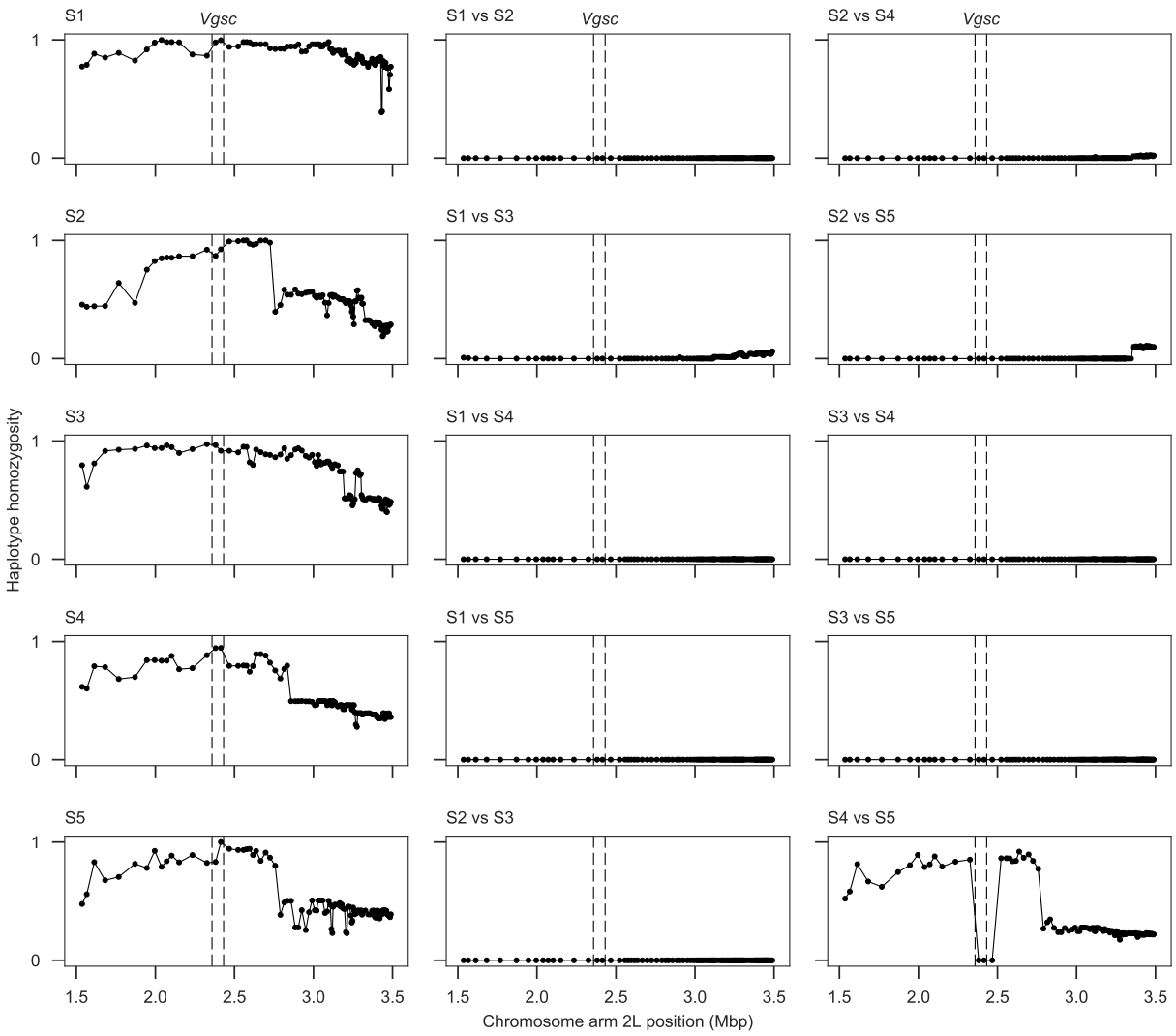

Figure S2. Windowed analysis of haplotype homozygosity for genetic backgrounds carrying the `L995S` allele. See Supplementary Fig. S1 for explanation. Haplotype homozygosity is high between clusters S4 and S5 on both flanks of the gene, indicating that haplotypes from both clusters are in fact closely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

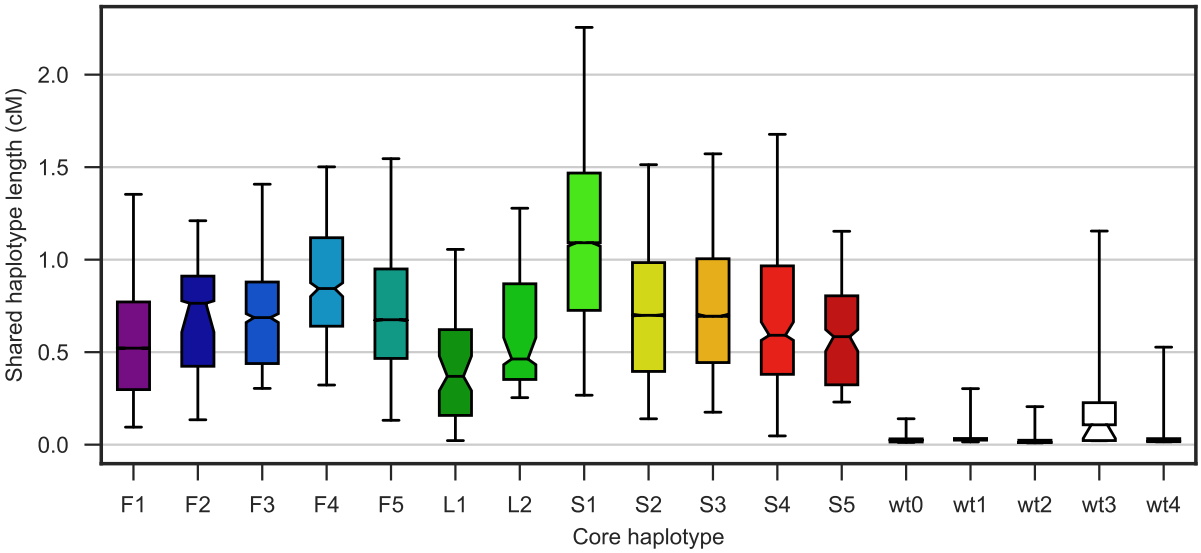

Figure S3. Shared haplotype length. Each bar shows the distribution of shared haplotype lengths between all pairs of haplotypes with the same core haplotype. For each pair of haplotypes, the shared haplotype length is computed as the region extending upstream and downstream from the core locus (*Vgsc* 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.