Agam-vgsc-report outline

**The genetic architecture of target-site resistance to pyrethroids and DDT in the malaria vectors *Anopheles gambiae* and *Anopheles coluzzii***

Chris S. Clarkson1\* , Alistair Miles2,1\* , Nicholas J. Harding2 , @@TODO?, Dominic Kwiatkowski1,2, Martin Donnelly3,1, and The Anopheles gambiae 1000 Genomes Consortium4

1Malaria Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA

2The Big Data Institute, Old Road, Oxford OX3 7FZ

3Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK

4Ag1000g Consortium, MalariaGEN, The Big Data Institute, Old Road, Oxford OX3 7FZ

**Abstract**

Increasing resistance to insecticides threatens malaria vector control campaign successes. As the only functional class of insecticide approved for use on treated bednets, resistance to pyrethroid insecticides is of particular concern. Pyrethroids target the voltage-gated sodium channel (VGSC), an essential component of the mosquito nervous system. Unfortunately exposure to insecticides generates a strong evolutionary pressure to evolve resistance. Three mutations have previously been identified within the VGSC gene (*Vgsc*), these disrupt toxin/active-site binding and produce a ‘*kdr’* resistant phenotype. Here, we use whole genome sequencing and haplotype estimation of 765 individual *A. gambiae*/*coluzzii* genomes by the Ag1000G Consortium, to investigate the whole gene genetic variation, evolution, recombination and gene flow involved in *Vgsc* mediated insecticide resistance. Multiple ‘origins’ of *kdr* mutations were evident even within populations, yet long distance gene flow of *kdr* carrying haplotypes was also found to be an important factor in the evolution of insecticide resistance across Africa. Haplotype networks revealed an abundance of high frequency non-synonymous mutations, unexpected due to the functional constraint suggested by conservation across dipterans. Even more striking was the non-random distribution of these protein altering mutations, many of which occurred on haplotypes carrying known resistance alleles and suggest additive or compensatory insecticide resistance phenotypic changes. Selective sweeps also appeared to have occurred on haplotypes with no known *kdr* mutations. These analyses and results reveal the *Vgsc* mediated resistance landscape to be much more complex than previously imagined while providing tools to identify and track mutations or haplotypes of medical importance. @@TODO improve

**Introduction**

**Malaria and insecticide resistance.** Recent research has quantified the huge gains made in the fight against malaria between 2000 and 2015, revealing that interventions reduced the incidence of clinical disease by 40%, averting 663 million cases with 68% of these prevented through use of insecticide treated bednets (ITNs) (Bhatt *et al*., 2015). However, the use of insecticides asserts strong selective pressures on the target vectors (Lynd *et al*., 2012). It is unsurprising, therefore, that in concert with intervention successes the incidence and strength of insecticide resistance has increased across Africa (Hemingway et al., 2016; Silva *et al*., 2014). This resistance risk to vector control campaigns is further compounded because, despite four chemical classes of insecticides being available for insect control (organophosphates and carbamates which target acetylcholinesterase; pyrethroids and organochlorines which target the voltage-gated sodium channel (VGSC)), only pyrethroids are approved by the World Health Organisation (WHO) for use in ITNs. With pyrethroids also commonly used for indoor residual spraying (IRS) campaigns, they are the most medically important class of insecticides (van den Berg *et al*., 2012) and increasing resistance to pyrethroids could cause total failure of vector control campaigns (Hemingway *et al*., 2016; World Health Organization, 2012). With pyrethroid resistance becoming increasingly common (Hemingway *et al*., 2016; Silva *et al*., 2014; Ranson *et al*., 2011), resistance mutations being shown to evolve *de novo* in the absence of gene flow (Reimer *et al*., 2005) and resistance already rendering ITNs ineffective in some regions (Toé *et al*., 2014; N’Guessan *et al*., 2007), there are increasing fears of major intervention failures. It is paramount that the evolution of pyrethroid insecticide resistance is understood for vector control campaigns to be targeted effectively to minimise further loss of efficacy (Jones *et al*., 2013).

**Voltage-gated sodium channel.** The physiological target of DDT and pyrethroids, the insect VGSC protein, is integral to the insect nervous system, allowing transmission of nerve impulses from neuron to neuron. Four domains make up the channel, each composed of six helical trans-membrane units including a voltage sensing unit and a unit which lines the interior of the ion channel as the four domains fit together on the cell. The channel is controlled by two gates; during an action potential (nerve impulse) the change in voltage across the cell membrane (depolarisation ) causes the protein channel to change shape, allowing Na+ (sodium ions) to pass from outside the cell down a concentration gradient to the lower Na+ concentrations found within , this is known as the ‘m-gate’. The channel is then closed by the inactivation particle, the ‘h-gate’, which blocks the channel before the m-gate conformation is restored as the resting voltage difference across the membrane returns (Davies *et al*., 2007a). The *Vgsc* gene was discovered within the paralysis locus in *Drosophila melanogaster*, consequently it is also referred to as ‘*para*’ in the literature (Loughney, Kreber and Ganetzky, 1989). In the sibling species *A. gambiae and A. coluzzii*, 31 exons code for the large (~60kb) gene (Giraldo-Calderón *et al*., 2014), which, situated centromere proximate on the 2L chromosome arm, lies within a large region of genetic divergence detected between the mosquito species pair (Turner et al., 2005).

**Evolution of insecticide resistance.** Both pyrethroids and dichlorodiphenyltrichloroethane (DDT - an organochlorine and the first widely used malaria vector control insecticide) have a similar mode of action. These insecticides are neurotoxic, targeting the VGSC protein present in the membranes of central and peripheral nervous system cells, responsible for controlling the rising phase of action potentials (nerve impulses). The insecticides bind to the sodium channel causing it to remain open, stimulating hyper excitability and causing paralysis or "knockdown" of the insect before ultimately death (Davies *et al*., 2007a). Through their effectiveness, these insecticides produce a strong evolutionary pressure to avoid the binding of insecticide to active-site, one path for the evolution of insect resistance to these xenobiotics is through conformational changes to the protein structure of insecticide binding sites, via non-synonymous point mutations.

These two classes of insecticides are used widely against a variety of vector and pest insects and insecticide resistance associated variants in the *Vgsc*, known as knockdown resistance (*kdr*) mutations, were first detected in the housefly, *Musca domestica* (Williamson *et al*., 1996). *Kdr* variants have since been discovered in the gene across Insecta (Davies *et al*., 2007a), with two such mutations widespread in African *Anopheles* malaria vector mosquitoes (Ranson *et al*., 2011; Silva *et al*., 2014), *Vgsc-995F* (Martinez‐Torres *et al*., 1998) and *Vgsc-995S* (Ranson *et al*., 2000). It should be noted that codon numbers for *Vgsc* mutations used throughout this publication relate directly to the AgamP4.4 reference gene-set to expedite future research in these insects (available through VectorBase: Giraldo-Calderón *et al*., 2014), rather than the codon numbering from other organisms which may have been used in initial variant discovery (@@reference my malariaGEN blog post/discuss more later? - <https://www.malariagen.net/> blog/untangling-mosquito-mutation-nomenclature). Another, more recently discovered *kdr* variant, *Vgsc-1570Y* is currently restricted to more westerly *Anopheles gambiae* and *A. coluzzii* populations (Jones *et al*., 2012; Silva *et al*., 2014).

The target of DDT and pyrethroid insecticides is the binding pocket, composed of IIS4-S5 linker and IIS5-III6 helices. Two of the target site mutations found in these malaria mosquitoes (*Vgsc-995F/S*) are situated on III6 helices, where they are thought to interact with the pocket reducing insecticide binding (O’Reilly *et al*., 2006). Interestingly, the only other *kdr* mutation that has been detected in *A. gambiae* is found in the III-IV linker (*Vgsc-1570Y*), near the region coding for the inactivation particle, where its role conferring resistance is less clear (Jones *et al*., 2012). Understanding the evolution of insecticide resistance in the VGSC is of great importance both medically (vector control) and for food security (crop pest control), while also providing a valuable evolutionary model; the great anthropogenic selection pressure due to insecticides is strong and recent (DDT use began in the 1940s, pyrethroids 1970s), allowing study of the emergence of adaptive variants in contemporary/tractable timescales (Davies *et al*., 2007a).

**Towards understanding the evolution, history and movement of insecticide resistance mutations.** Investigations into the presence/absence of malaria vector *kdr* resistance genotypes within populations of the major malaria vectors *A. gambiae*/*coluzzii* (often as a proxy for resistance phenotype), currently use molecular assays which detect only the three known individual nucleotide polymorphisms (SNPs) linked with resistance, *Vgsc-L995F*, *Vgsc-L995S* or *Vgsc-N1570Y* (*e.g.* *kdr* TaqMan assays of Jones *et al*., 2012; Bass *et al*., 2007). Though these assays are effective at establishing which known mutations are segregating within populations, they provide little or no information about the history, movement, evolution or indeed if additional mutations are involved in this medically important phenotype. Focal studies using Sanger capillary sequencing of a single intron from the *Vgsc* (described as intron 1 but intron 18 according to the Agam P4.4 geneset - Giraldo-Calderón *et al*., 2014) have suggested multiple origins for individual resistance mutations across Africa (Etang *et a*l., 2009 - 119 samples; Pinto *et al*., 2007 - 288 samples) and found high levels of nucleotide diversity (Santolamazza *et al*., 2015). These results hinted at the potential for a high levels of complexity in the *Vgsc* and in *kdr* resistance, something that the current high-throughput molecular assays cannot corroborate.

The limited understanding of resistance evolution at the *Vgsc* in higher was recently highlighted by the initial findings of the *Anopheles gambiae* 1000 Genomes Project (Ag1000g). 765 wild caught mosquitoes from across Africa underwent whole genome sequencing (WGS), the diploid data was phased into haplotypes allowing accurate *Vgsc* nucleotide sequence to be generated for both chromosomes from each individual mosquito (Miles *et al*., 2017). This first ‘genomic’ comparison of whole *Vgsc* gene haplotypes within and between vector populations allowed, for the first time, a pan-African snapshot of evolution at this locus and revealed that the high nucleotide diversity is not just seen in introns as might be expected in an important and conserved gene (Santolamazza *et al*., 2015; @@TODO Davies ref for conservation across Insecta), but as non-synonymous (protein altering) variation throughout coding regions of the *Vgsc* (Miles *et al*., 2017). These initial genomic findings underlined the need for higher resolution, focal analyses of the locus such as where and how *kdr* insecticide resistance originates, if it is augmented by additional mutations and how resistant haplotypes move, is now essential if this resource is to be translated into information that can inform and improve vector control campaigns by reducing or avoiding resistance issues.

Here we use the WGS sequencing from the Ag1000g project to examine over 1500 *A. gambiae/coluzzii* *Vgsc* haplotypes. By clustering them into haplogroups putatively under selection using the measure which accounts for haplotype length, mutations and recombination, rather than a simpler measure such as genetic distance (Mathieson and McVean, 2014), we can infer relative haplogroup age and direction of gene flow of haplogroups across Africa. Furthermore, exploring haplogroups using a network approach and analyses of recombination, will allow the first high resolution genomic analysis of this important gene.

**Methods**

**Data collection and processing.** For detailed information on Ag1000g WGS sample collection, sequencing, variant calling, quality control and phasing see Miles *et al.* (2017). In brief, malaria vector mosquitoes were collected from eight countries across Sub-Saharan Africa: Angola, Burkina Faso, Cameroon, Gabon, Guinea, Guinea Bissau, Kenya and Uganda. Guinea Bissau was an *A. gambiae/coluzzii* admixed population, Angola *A. coluzzii*, Burkina Faso had samples of both *A. gambiae* and *A. coluzzii* and all other populations were consisted of *A. gambiae*. Mosquitoes were individually whole genome sequenced on the Illumina HiSeq 2000 platform, generating 100bp paired-end reads. Sequenced reads were aligned to the *A. gambiae* AgamP3 reference genome assembly (Holt *et al*., 2002). 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 (2010). Variant filtering rules removed accessible variants if: FS < 60.0, HRun > 4, DP <18000 or >32000, MQ < 40, alleles > 2, QD < 5, ReadPosRankSum < -8, overlapped positions repeat masked by DUST (<http://blast.wustl.edu/pub/dust>). This process produced a call set containing WGS sequences for 765 wild caught individuals which was publicly released (Miles *et al*., 2017; [www.malariagen.net](http://www.malariagen.net/)).

The Ag1000g variant data was functionally annotated using the SnpEff software which allowed investigation of potential phenotype altering variants within *Vgsc* (v4.1b - Cingolani *et al*., 2012). Per-population non-synonymousSNPs were identified as all gene variants, AGAP004707 2L:2358158-2431617, with a SnpEff annotation of “missense” and an ALT allele frequency of >5% in at least one of the nine mosquito populations, with the exception of the multi-allelic SNP 2,400,071 G>A which is shown despite only being found in *A. gambiae* from Cameroon at 0.4% frequency, as the G>T variant at the same position which causes the same codon change (*Vgsc-M490I*), is found above 5% frequency in Kenya. These ‘population level’ results are shown in Table 1, in all other analyses, where populations are initially combined, an alternative allele (ALT) allele frequency cut off of >1% across all 765 samples was employed to determine which variants to include. Minimum ALT allele frequencies were used to discriminate towards variants that may be undergoing selective sweeps and against less informative low frequency alleles.

For ease of comparison with previous work on *Vgsc*, pan Insecta, *A. gambiae* codon numbering is also reported in the equivalent *Musca domestica* codon numbering, the species in which the gene was first discovered. The *M. domestica Vgsc* sequence (EMBL accession X96668 - Williamson *et al*., 1996)was aligned with the *A. gambiae* AGAP004707-RA sequence (AgamP4.4 gene-set), using the Mega software package (v7 - Kumar, Stecher and Tamura, 2016). A map of equivalent codon numbers between the two species 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 - Delaneau *et al*., 2013), see Miles *et al*. (2017) supplementary text for more details. The SHAPEIT2 algorithm is not able to phase multi-allelic positions, therefore the two multi-allelic non-synonymous SNPs within the *Vgsc* gene (>1% ALT frequency), altering codons *Vgsc-V402* and *Vgsc-M420*, were phased onto the haplotypes using MVNcall (v1.0 - Menelaou and Marchini, 2013). Conservative filtering had removed one of the three known insecticide resistance conferring *kdr* variants, *Vgsc-N1570Y* (Jones *et al*., 2012). 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 (Miles *et al*., 2017). To evaluate the linkage disequilibrium (LD) of non-synonymous *Vgsc* mutations with the two most widespread *kdr*  resistance mutations (*Vgsc-L995S/F*), the D1 statistic was calculated using haplotypes and the python library Scikit-allel (Miles and Harding, 2016).

**Haplotype/-group age.** @@TODO - AM

-Length of shared haplotype and number of ‘mutations’ between them are informative of age…

-Pairwise values were hierarchically clustered and visualised as a dendrogram using the Python library Scipy and its cluster.hierarchy functions (method = linkage\_method).

-Cutting the dendrogram at @@generations clustered haplotypes together into haplogroups…

- Naming of haplogroups with reference to Ag1000g...

-dendro figure/distro figures/map - Python libraries...

**Haplogroup networks.** Discerning the relationships between similar haplotypes that are clustered within haplogroups can be difficult when using bifurcating trees as, inherently, the distance between the leaves at the tips (haplotypes) will be small. As these relationships may be informative of the history of selection, we utilised a network approach to elucidate them. Due to large genetic differences between haplogroups, the *Vgsc-L995S* and *F* haplotypes were analysed separately; haplotypes (SNP variants from 2L:2358158-2431617) were networked using the median-joining algorithm, designed to be used specifically with interspecific data with small genetic distances (Bandelt *et al*., 1999), implemented in a custom Python script. Networks were visualised with the Python library graphviz with non-synonymous edges highlighted using the SnpEff “missense” annotations (v4.1b - Cingolani *et al*., 2012) and composite figures constructed using Inkscape. @@TODO decide on max\_distance when re-making figure and add here.

**Recombination.** @@TODO - AM - Absolute divergence dxy...

**Code.** All custom Python code used to run analyses and generate figures is available from <https://github.com/malariagen/agam-vgsc-report>.

**Data.** All data, genomic, haplotypic and meta, is publically downloadable via ftp from [https://www.malariagen.net](https://www.malariagen.net/).

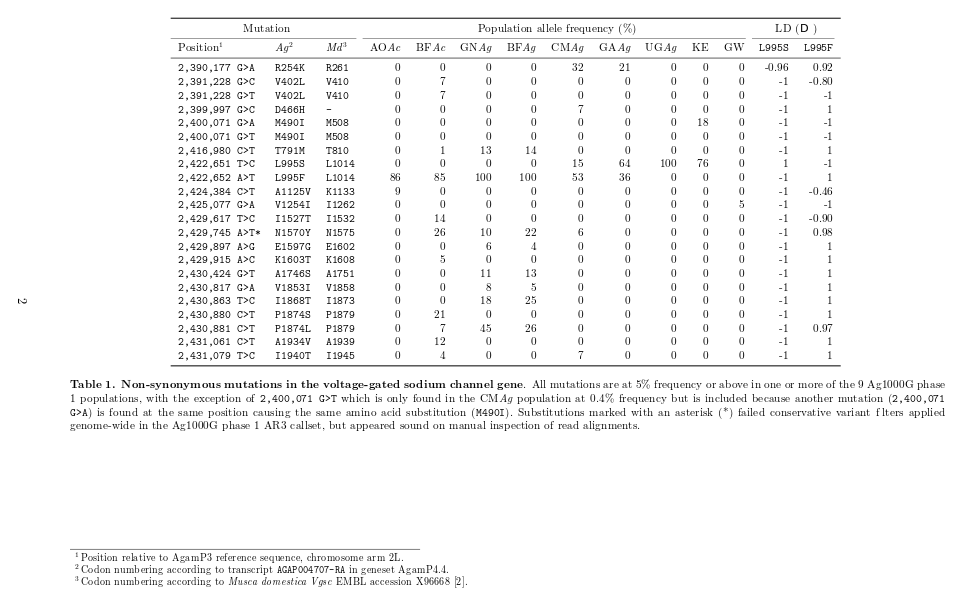
**Results**

Historically a simplistic models of genotype-phenotype association had been proposed for *Vgsc* mediated insecticide resistance in malaria vectors (and indeed across Insecta). Mutations in just two codons *Vgsc-L995S/F* and *Vgsc-N1570Y* in the gene have been associated with resistance and many African *A.gambiae/coluzzii* populations evaluated with molecular assays for one or two of these mutations for almost 20 years (*e.g.* the studies reviewed in Silva *et al*., 2014). However, multiple origins, gene flow and the propensity for *de novo* mutation of resistance loci revealed in genetic studies of the *Vgsc* (Santolamazza *et al*., 2015; Etang *et al*., 2009; Pinto *et al*., 2007; Reimer *et al*., 2005), suggested that a complexity of resistance that was not being brought to light by molecular assays. The number of *Vgsc* codon altering mutations discovered at high frequency in the Ag1000g phase 1 data set, strengthen these suggestions (Table 1).

*Kdr*  insecticide resistance mutations are widespread in Sub-Saharan Africa, all populations except for Guinea Bissau (GW) carry at least one *kdr* mutation at high frequency, Cameroon (CM*Ag*) and Gabon (GA*Ag*) carry both *Vgsc-995F* and *S*. Despite two *kdr* mutations segregating in Cameroon, the population presents the highest proportion of non-*kdr* (insecticide “susceptible”) carrying haplotypes (32%) after Guinea Bissau, while Gabon is fixed for *Vgsc-995* resistance linked mutations with 36% of haplotypes carrying *995F* and 64% carrying *995S*; these two *kdr* mutations are not found on the same haplotype (Table 1). The third previously reported *kdr* mutation, *Vgsc-N1570Y*, is also detected in samples from West Africa, but at much lower frequencies than the *Vgsc-995* mutations (Jones *et al*., 2012). The *Vgsc-N1570Y* mutation was previously only detected on haplotypes carrying *Vgsc-L995F* but the extensive Ag1000g sampling reveals that it is not in perfect linkage disequilibrium (LD) with this mutation, L995F D1 = 0.98 (Table 1).

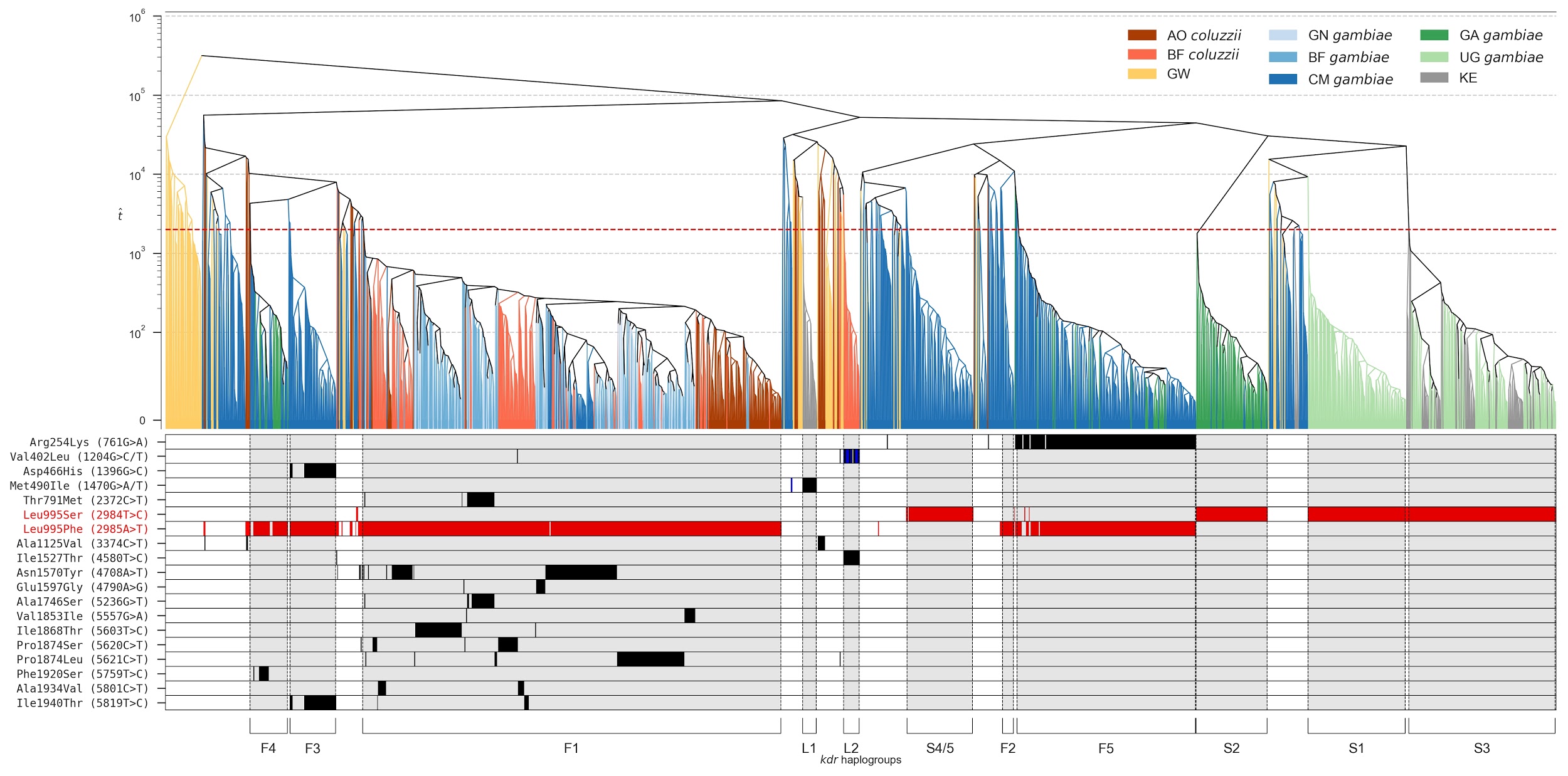
Three mutations have previously been associated with *Vgsc* mediated resistance in *Anopheles gambiae*/*coluzzii*, however, 18 further non-synonymous mutations are found at appreciable frequency in at least one population within the *Vgsc* coding sequence (Table 1). Mutations altering codon *Vgsc-1874* have been associated with insecticide resistance in the crop pest, *Plutella xylostella* (Sonoda *et al*., 2008) but to our knowledge none of the other mutations have been linked with insecticide resistant phenotypes (Rinkevich *et al*., 2013).

**Table 1. Non-synonymous mutations in the voltage-gated sodium channel gene.** The position and SNP of each non-synonymous mutation is shown relative to the AgamP3 reference genome sequence, chromosome arm 2L. Codon numbering is reported according to the *Anopheles gambiae Vgsc* transcript AGAP004707-RA from the AgamP4.4 geneset (***Ag*** column) and as the equivalent codon according to the *Musca domestica Vgsc* sequence where present (**Md** column). Population are represented by the country two letter code followed by species (either *Ag* - *A. gambiae*, *Ac* - *A. coluzzii*): AO - Angola, BF - Burkina Faso, GN - Guinea, CM - Cameroon, GA - Gabon, UG - Uganda, KE - Kenya, GW - Guinea Bissau. Linkage disequilibrium (LD) of each variant with each of the two most widespread *kdr* mutations, *Vgsc-L995S/F*, is calculated using the D1 statistic. All mutations shown are at 5% frequency or above in one or more of the nine 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% but is included because another mutation, 2,400,071 G>A, is found at the same position causing the the same substitution and is above 5% frequency or above in KE (M490I). Substitutions marked with an asterix (\*) failed conservative variant filters applied genome-wide in the the Ag1000g phase 1 AR3 callset, but appeared sound on manual inspection of read alignments.



The relationship between haplotypes, was explored within an age-informative framework by using a methodology derived from the statistic (Mathieson and McVean, 2014). By considering, in a pairwise fashion, the lengths of shared haplotypes and the number of sites which differed within the shared region, the most recent common ancestor (TMRCA), in effect the relative haplotype age in generations, can be estimated. The pairwise matrix of values was visualised as a cladogram of haplotypes clustered by age. Cutting this tree 2000 generations prior to sample collection demarcated 14 haplogroups (clusters of similar haplotypes), however, three of these haplogroups contained only wild-type haplotypes the remaining 11 non-synonymous mutations carrying clusters are highlighted (Figure 1). All but one of the haplogroups was composed of haplotypes from just one or two populations in relatively close geographic proximity. Haplogroups F4, F5 and S2 shared haplotypes from Cameroon and Gabon while S3 shared haplotypes from Uganda and Kenya. In contrast, the single ‘cosmopolitan’ haplogroup, F1, contained haplotypes from five populations spread across Western, Central and Southern Africa (Figures 1 and 2).

SNP ‘tracks’ below the cladogram showing which haplotypes carry non-synonymous mutations made it possible to visualise that much of the haplogroup clustering appeared driven by certain variants undergoing selective sweeps, increasing the frequency of the haplotypes carrying them. As expected the majority (9/11) of the haplogroups carried either the *Vgsc-L995S* or *F* insecticide resistance *kdr* mutations at or almost at fixation, these haplogroups were thus named either S*n* or F*n* (the *n* being taken from the similarity with genetic distance derived haplogroups defined by Miles *et al*. (2017)). However, the the placement of hitherto undescribed non-synonymous variation onto the cladogram, revealed these variants may also drive secondary selective sweeps on the haplotypes already carrying mutations known to be linked to resistant phenotypes, and also may have generated the selective sweeps on ‘susceptible’ (non-*kdr* carrying) haplogroups, creating non-*kdr* haplogroups L1 and L2 (Figure 1). Within the Kenyan haplogroup L1 and on fan otherwise wild-type background, the Met490Ile codon change, not found elsewhere, is fixed; the second non-*kdr* haplogroup L2, private to Burkina Faso *A. coluzzii*, is more complex with Ile1527Thr at fixation, both alleles of the multi-allelic (but same codon change) Val402Leu and a small number of haplotypes carrying Asn1570Tyr (*Vgsc-N1570Y*), the ‘third’ *kdr* mutation that has been shown to increase insecticide but has only ever been found in concert with Leu995Phe (*Vgsc-L995F*) previously (Jones *et al*., 2012). The F1 geographically cosmopolitan haplogroup was the most genetically diverse, with haplotypes carrying 13 of the 19 high frequency non-synonymous mutations, the highest protein altering diversity found in other haplogroups was only 3/19 (haplogroups F3 and F5) (Figure 1).



**Figure 1**. **Haplotype ‘age’ cladogram.** Pairwise values between haplotypes represented as a bifurcating tree. Red dashed horizontal line in upper panel denotes value (generations) the dendrogram was cut at to generate the haplogroup clustering (shown by black dashed vertical lines and named according to which *Vgsc-995* codon they carry - F/S/L and with respect to Miles *et al*. (2017)). Colour of dendrogram leaves shows origin population of each haplotype. Lower panel shows which non-synonymous mutations are carried by the haplotypes above, the two most widespread *kdr* insecticide mutations are shown by red bars (*Vgsc-L995F/S*). All other mutations are represented by black bars, except for multi-allelics (with two alternative alleles generating the same codon change - Val402Leu and Met490Ile) where the first alternative allele is shown in black and the the second in blue.

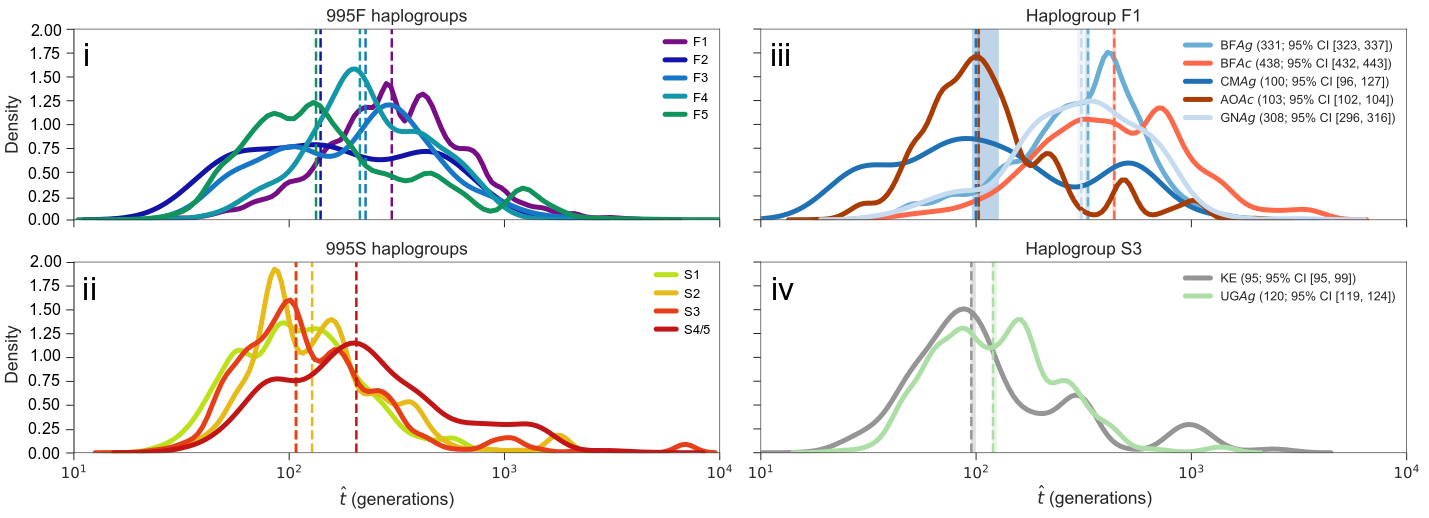
An advantage of using over genetic distance (*e.g.* Miles *et al*., 2017) was that, although the actual values of are only an estimation of the actual age or distance to TMRCA, the relative differences enable a prediction of gene-flow based on the assumption that populations carrying older haplogroup haplotypes are more likely to be the source of gene-flow for populations carrying younger haplotypes from the same haplogroup. Analyses of distributions revealed that median TMRCA of ‘F’ haplogroups tend to be older than ‘S’ haplogroups with the oldest ‘F’ haplogroup F1, coalescing ~250 generations ago and the oldest ‘S’, S4, ~230 generations ago; the ‘L’ haplogroups TMRCA appear intermediate to ‘F’ and ‘S’ (Figure 2ai-ii; Supplementary Table 1). @@TODO more statistics?

**Supplementary Table 1. Haplogroup medians @@TODO add number of haps in HG column**

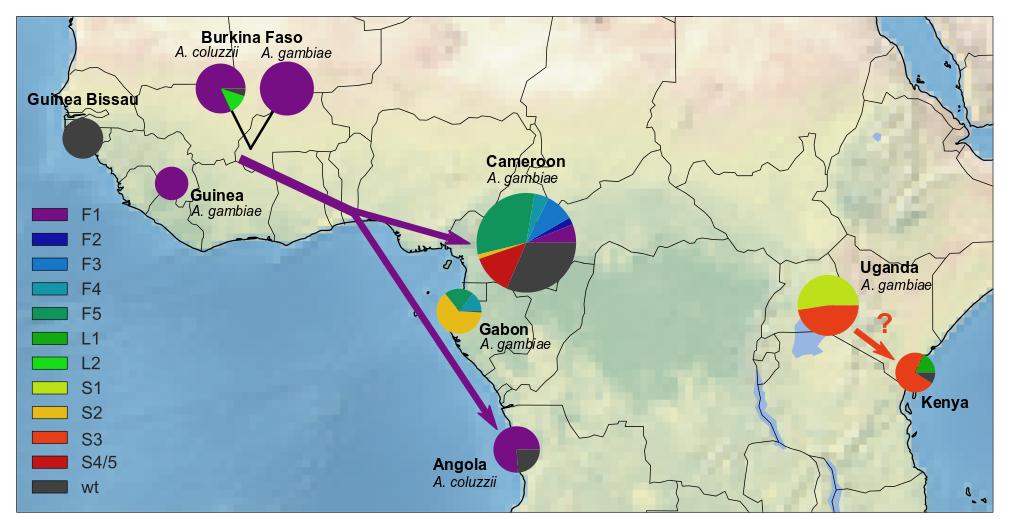
|  |  |  |
| --- | --- | --- |
| **Haplogroup** | ***n* haplotypes** | **Median** |
| F1 | 461 | 2.4766 |
| F2 | 13 | 2.1459 |
| F3 | 51 | 2.3548 |
| F4 | 42 | 2.3284 |
| F5 | 197 | 2.1247 |
| L1 | 16 | 2.1884 |
| L2 | 18 | 2.209 |
| S1 | 108 | 2.0357 |
| S2 | 79 | 2.1065 |
| S3 | 162 | 2.0309 |
| S4/5 | 73 | 2.3120 |

Examining the distributions of individual populations within haplogroups provides a higher-resolution inference of gene flow, potentially more relevant for informing vector control. Striking from earlier analyses into the cosmopolitan F1 haplogroup, both here and in Miles *et al*. (2017), was the apparent gene flow between the geographically distant populations it spans. Population level examination of the distributions within F1 show that the haplotypes in Central and Southern populations of Cameroon and Angola respectively coalesce more recently than the more northerly populations of Burkina Faso Guinea, a result that suggests gene-flow of this insecticide resistance mutation carrying haplogroup from north to south (Figure 2). To infer direction of gene-flow in the S3 haplogroup shared between Kenyan and Uganda appears more difficult. Though confidence intervals do not overlap, suggesting the Ugandan haplotypes are older, the relatively small age difference of just 35 generations and the bi-modal distributions reduce confidence (Figure 2). It is estimated that mosquitoes have ~12 generations per year in the wild (Lehmann *et al*., 1998), therefore inference of relative age is strongest for samples collected temporally as closely as possible. Fortunately all populations were sampled between 2009 and 2012, with the exception of Gabon which was sampled in 2000. We therefore do not show distribution analyses in haplogroups shared with Gabon, F4 and F5 (Figure 1).

**a)**

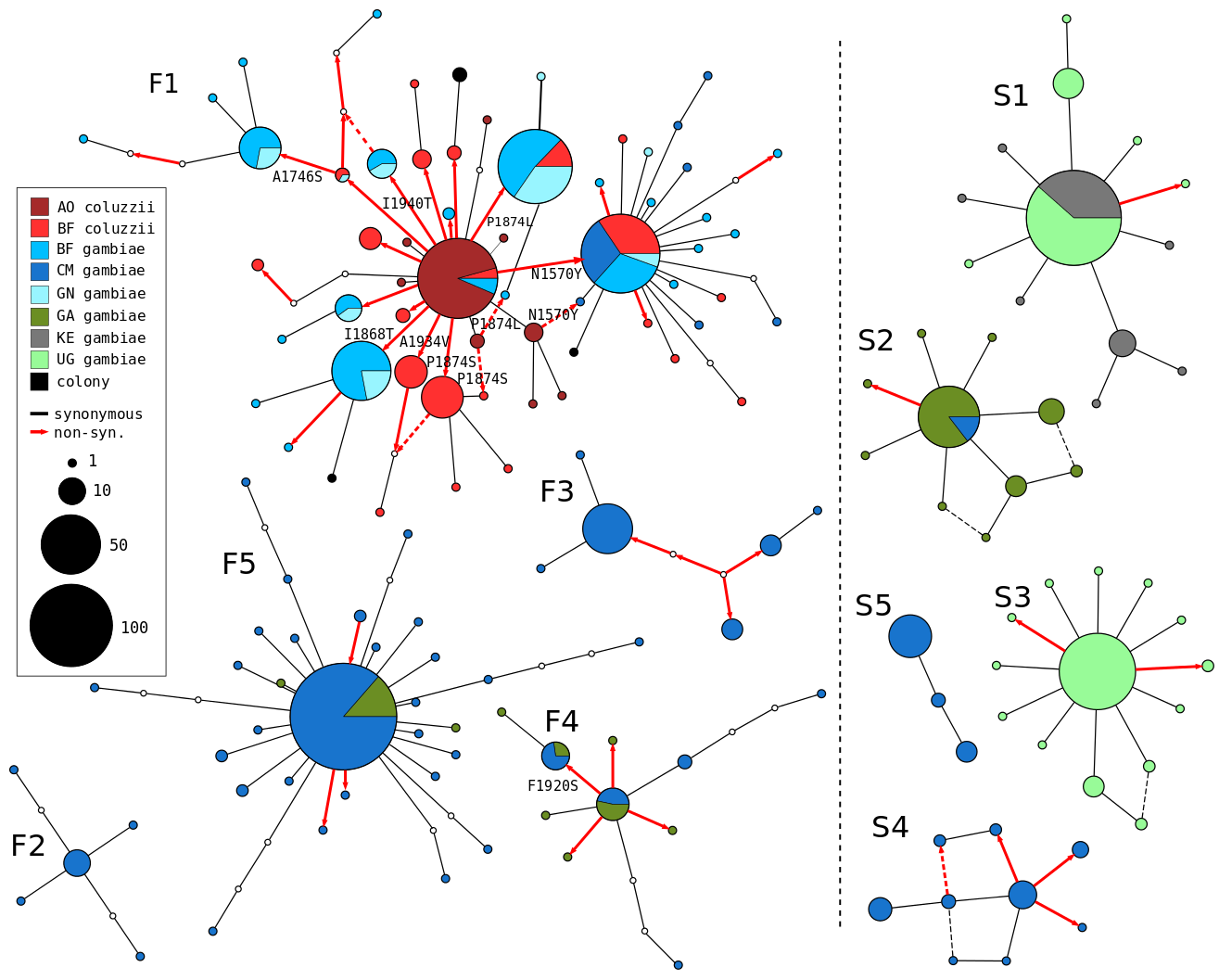


**b)**



**Figure 2**. **a)** **Haplogroup ‘age’ distributions.** **i-ii)** Comparing distributions of between haplogroups, medians shown by vertical dashed lines. **iii-iv)** Comparing distributions of between the populations within haplogroups, median and 95% confidence limits shown by dashed vertical lines and vertical shaded regions respectively. **b)** **Spatial distribution of haplogroups.** Pie charts represent the nine sample populations, country name and population species (as defined in Miles *et al.,* (2017)) shown. Pie chart size indicates relative number of haplotypes in population and pie slice indicates proportion of haplogroup in each population. Arrows denote direction of gene flow of haplogroups inferred by distribution and arrow colour refers to haplogroup, question mark indicates reduced confidence in flow direction.

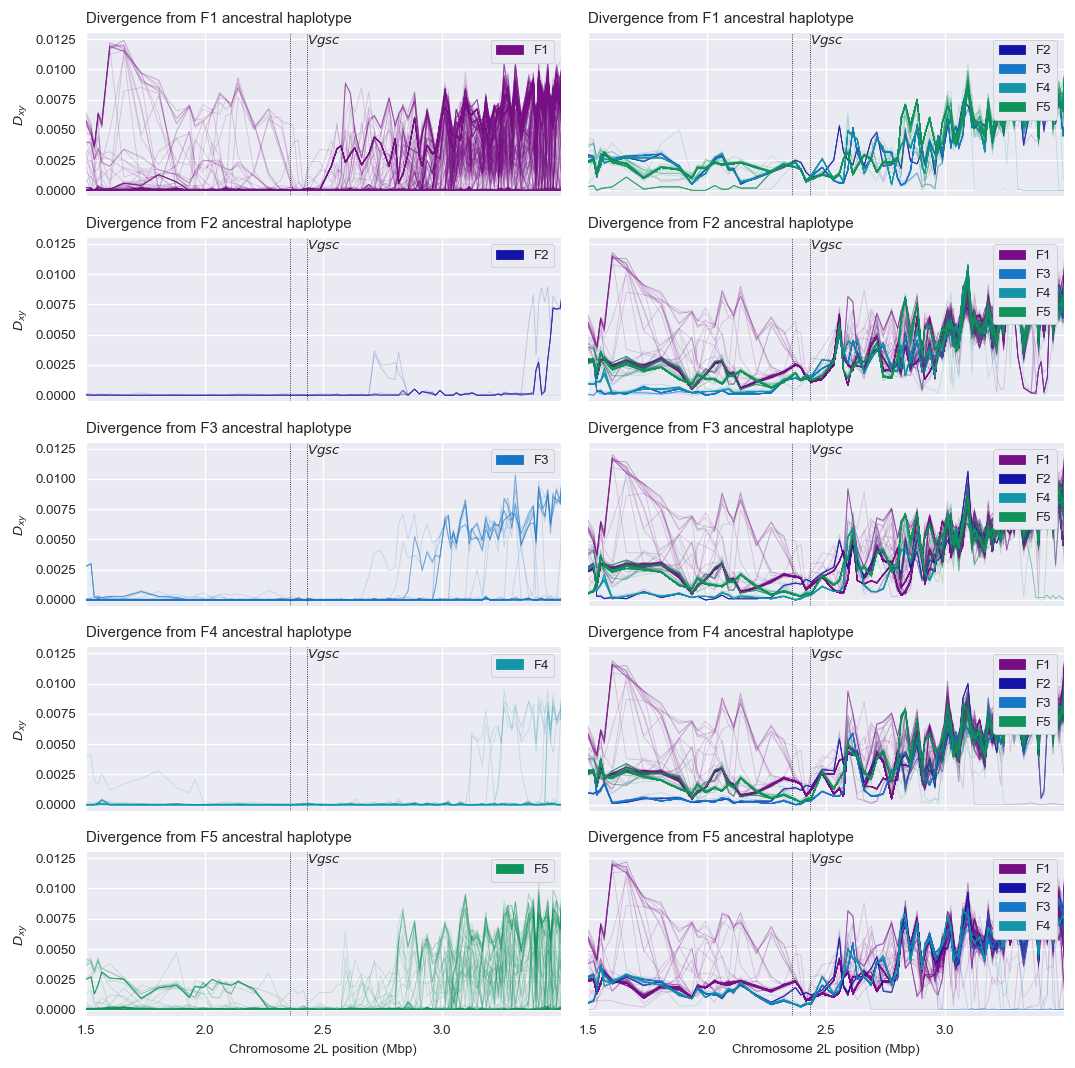
The *Vgsc* haplogroups differ greatly in the number of haplotypes they contain, from a maximum of 461 to a minimum a minimum of 16 (Supplementary Table 1) and harbour haplotypes from up to five populations (F1) down to just one population (F2, F3, S1, S4/5) (Figure 1); these haplogroups may, therefore, have been sampled at different stages of selective sweep and/or perhaps were undergoing different selective regimes. To elucidate the genetic relationships within haplogroups and explore signatures of evolution, a network approach was taken. By visualising the *Vgsc* SNP variation in each haplogroup as a network, where each edge represents a variant, the amount and effect of the variation could be explored… @@TO DO - need finished figure 3 so we can talk about L1/L2 etc - we found a bug in the software and are in the process of reproducing this figure. The figure shown is the ag1000g iteration haplogroups but should be qualitatively similar to the final figure just with L1/2 added. ...Most haplogroups star shape = recent expansion (REF?), therefore most look recent snapshot after the initial non-syn mut that created them (usually a kdr mutation - except L1/2) they don't show any more non-syn at high freq apart from F4 with just one node. However, the F1 haplogroup looks very different, @@nn non-synonymous mutations were present with ten leading to ‘high frequency’ nodes, suggesting selective sweeps occurring on the already swept/sweeping haplotype...



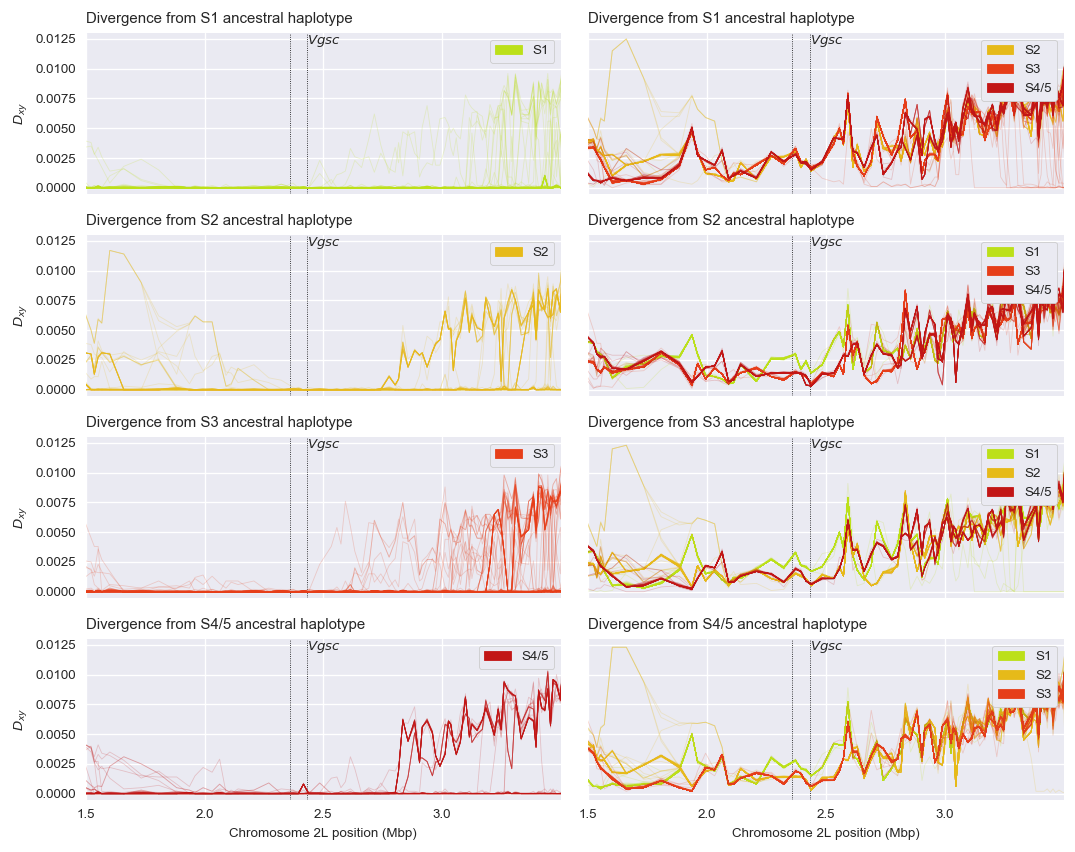
**Figure 3**. **Haplogroup networks** (this is the old figure - found bug when generating the haplogroup networks @@TODO fix code bugs and replace with new figure). Haplotype networks are shown for the defined haplogroups, each edge represents a SNP difference between nodes. Non-synonymous edges are coloured red with the codon change/position shown in text if the edge leads to a node with a haplotype frequency of ten or more, synonymous edges are black. Node size is relative to haplotype frequency and pie chart proportions and colour show node population (see key).

The *Vgsc* gene falls in a putative low recombination region, centromere proximal (Pinto et al., 2007; Carneiro, Ferrand and Nachman, 2008), however more recent research has revealed evidence of recombination in the region (Lee *et al*., 2013). To investigate the potential for recombination to be driving the observed clustering, putative ancestral haplotypes were reconstructed for each haplogroup using majority rule (@@REF), then within and between haplogroup divergence comparisons were performed against each ancestral haplotype (Supplementary figure 1). These analyses reveal that although much recombination is seen with the @@kb region (2L-@@:@@), particularly centromere distal, very few recombination events are detected within the *Vgsc* itself (2L:2358158-2431617) with almost no divergence from ancestral haplotype in the genic region within haplogroups (*Dxy* ≈ 0.00 - Supplementary figure 1). This result is further compounded by the divergence found in the *Vgsc* between haplogroups, which all have a *Dxy* > 0.00 around the gene, however, more absolute divergence was found between *Vgsc-995S* haplogroups than *Vgsc-995F* (Supplementary figure 1). @@TODO these are new results and proabably need thinking about a bit more, some actual numbers from the notebook might be good here too. Do we need to include L1/2 divergence plots - not defined by carrying a SNP so perhaps uninformative?

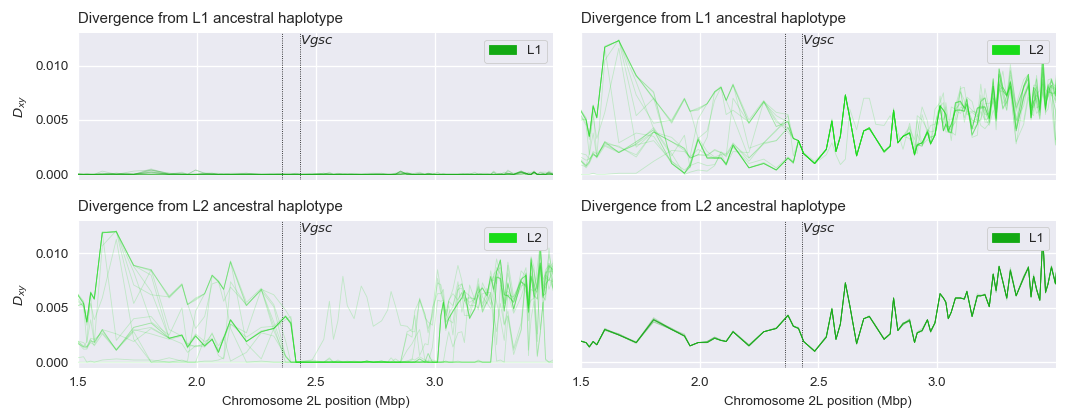
**a)**



**b)**



**c)**



**Supplementary Figure 1.** Recombination in the vgsc region. @@TODO

**Discussion**

In the major African malaria vectors, *Anopheles gambiae* and *A. coluzzii*, the VGSC and the nucleotide sequence of the gene that encodes it (*Vgsc*) are of particular medical and biological interest. The channel, an essential part of the insect nervous system, carries the target site for pyrethroid and DDT insecticides. With the rise of insecticide resistance and the reliance of ITN control campaigns on pyrethroids, there are fears of complete control failures and reversals of the huge decreases in malaria incidence (Hemingway *et al*., 2016). Three loci in the gene have been associated with resistance and many African *A.gambiae/coluzzii* populations evaluated with molecular assays for at least two of these mutations for almost 20 years (*e.g.* the studies reviewed in Silva *et al*., 2014). However, multiple origins, gene flow, recombination and the propensity for *de novo* mutation of resistance loci revealed in genetic studies of the *Vgsc* (Santolamazza *et al*., 2015; Lee *et al*., 2013; Etang *et al*., 2009; Pinto *et al*., 2007; Reimer *et al*., 2005), suggest that a complexity of resistance that was not being brought to light by these molecular assays.

The Ag1000g project revealed a much higher level of non-synonymous *Vgsc* complexity than expected while describing a snapshot of *Vgsc* genomics Africa-wide (Miles *et al*., 2017), and has presented an opportunity using the 765 WGS Ag1000g individuals to uncover complexity and to investigate the origins, evolution and gene flow of resistant haplotypes; if vector control campaigns are to predict resistance evolution and react effectively to it, rather than just identifying resistance strength or presence/absence, this complexity must be understood. @@TODO once results are finalised - finish discussion:

* Ag1000g genetic distance clusters seem robust, recapitulates most of haplogroups from Ag1000g paper AND allows elucidation of history with knowing demographic history of pops
* Results show complexity of selective sweeps revealing underestimation of resistance and urgency of need to improve genotype/phenotype research. Particularly interesting are the high frequency non-syn mutations found in “non-kdr” haplogroups, perhaps not just additive/compensatory mutations sweeping on kdr backgrounds but also completely new resistance mutations on non-kdr haplotypes.
* F1 networks - something different appears to be happening here, not a simple star shape, this haplogroup is older so we might expect more divergence than others but multiple overlayed sweeps suggest directional selection on increased resistance/compensatory phenotypes. Perhaps this haplogroup had lower fitness in the ancestral haplotype?
* Recombination isn’t driving clustering, though there recombination close to centromere which may not be expected.
* Kdr west/east doesn’t hold anymore if it ever did...Cam/Gabon
* Developing a framework for predicting the evolution of *Vgsc* mediated insecticide resistance that can be applying to other resistance loci in these mosquitoes...
* Not just malaria - existing and emerging threats from arbovirus - Aedes

**References**

The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." *Nature* 467(2010): 1061–1073.

Bagi, J. *et al*. "When a discriminating dose assay is not enough: measuring the intensity of

insecticide resistance in malaria vectors." *Malaria Journal* 14.1 (2015): 210.

Bandelt, H.-J., P. Forster and A. Röhl. "Median-joining networks for inferring intraspecific phylogenies." *Molecular Biology and Evolution* 16.1 (1999): 37-48.

Barrett, J.C. *et al*. "Haploview: analysis and visualization of LD and haplotype maps."

*Bioinformatics* 21.2 (2005): 263-265.

Bass, C. *et al*. "Detection of knockdown resistance (*kdr*) mutations in *Anopheles gambiae*: a comparison of two new high-throughput assays with existing methods." *Malaria Journal* 6.1 (2007): 1.

van den Berg, H. *et al*. "Global trends in the use of insecticides to control vector-borne diseases." *Environmental Health Perspectives* 120.4 (2012): 577-582.

Bhatt, S. *et al*. "The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015." *Nature* 526.7572 (2015): 207-211.

Brito, L. P. *et al*. "Assessing the effects of *Aedes aegypti* *kdr* mutations on pyrethroid resistance and its fitness cost." *PLoS One* 8.4 (2013): e60878.

Carneiro, M., Ferrand N. and M. W. Nachman. Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics* 181 (2008): 593-606.

Cingolani, P. 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." *Fly* 6.2 (2012): 80-92.

Clement M, D. Posada D and K. Crandall. "TCS: a computer program to estimate gene genealogies." *Molecular Ecology* 9(10) (2000): 1657-1660.

Crawford, J. E. and B. P. Lazzaro. "The demographic histories of the M and S molecular forms of *Anopheles gambiae* ss." *Molecular Biology and Evolution* 27.8 (2010): 1739-1744.

Davies, T. G. E. *et al*. "DDT, pyrethrins, pyrethroids and insect sodium channels." *IUBMB* Life 59 (2007a): 151–162.

Davies, T. G. E. *et al*. "A comparative study of voltage‐gated sodium channels in the Insecta: implications for pyrethroid resistance in Anopheline and other Neopteran species." *Insect Molecular Biology* 16.3 (2007b): 361-375.

Delaneau O. *et al*. "Haplotype estimation using sequencing reads." *American Journal of Human Genetics* 93 (2013): 687–96.

Etang, J. *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." *Molecular Ecology* 18 (2009): 3076-3086.

Flaxman, A. D. *et al*. "Rapid scaling up of insecticide-treated bed net coverage in Africa and its relationship with development assistance for health: a systematic synthesis of supply, distribution, and household survey data." *PLoS Medicine* 7.8 (2010): 1011.

Foster, S. P., *et al*. "Analogous pleiotropic effects of insecticide resistance genotypes in peach–potato aphids and houseflies." *Heredity* 91.2 (2003): 98-106.

Giraldo-Calderón, G. I. et al. "VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases." Nucleic Acids Research (2014).

Griffin, J. T. *et al*. "Reducing Plasmodium falciparum malaria transmission in Africa: a model-based evaluation of intervention strategies." *PLoS Medicine* 7.8 (2010): 1028.

Hemingway, J. *et al*. "Averting a malaria disaster: will insecticide resistance derail malaria control?" *The Lancet* 387.10029 (2016): 1785-1788.

Jones, C. M. et al. "Footprints of positive selection associated with a mutation (*N1575Y*) in the voltage-gated sodium channel of *Anopheles gambiae*." *Proceedings of the National Academy of Sciences* 109 (2012): 6614-6619.

Jones, C. M. *et al*. "The dynamics of pyrethroid resistance in *Anopheles arabiensis* from Zanzibar and an assessment of the underlying genetic basis." *Parasite and Vectors* 6.1 (2013): 343.

Karasov, T., P. W. Messer and D. A. Petrov. "Evidence that adaptation in *Drosophila* is not limited by mutation at single sites." *PLoS Genetics* 6.6 (2010): e1000924.

Kumar, S., G. Stecher, and K. Tamura. "MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets." *Molecular Biology and Evolution* 33.7 (2016): 1870-1874.

Lee, Y. *et a*l. "Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, *Anopheles gambiae*." *Proceedings of the National Academy of Sciences* 110.49 (2013): 19854-19859.

Lehmann, T. *et al*. "The effective population size of Anopheles gambiae in Kenya: implications for population structure." *Molecular Biology and Evolution* 15.3 (1998): 264-276.

Loughney, K., R. Kreber, and B. Ganetzky. "Molecular analysis of the para locus, a sodium channel gene in *Drosophila*." *Cell* 58.6 (1989): 1143-1154.

Lynd, A. *et al*. "Field, genetic, and modeling approaches show strong positive selection acting upon an insecticide resistance mutation in *Anopheles gambia*e *ss*" *Molecular Biology and Evolution* 27.5 (2010): 1117-1125.

Martinez‐Torres, D. *et al*. "Molecular characterization of pyrethroid knockdown resistance (*kdr*) in the major malaria vector *Anopheles gambiae ss*." *Insect Molecular Biology* 7.2 (1998): 179-184.

Mathieson, I. and G. McVean. "Demography and the age of rare variants." *PLoS Genetics* 10.8 (2014): e1004528.

Menelaou, A. and J. Marchini. “Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold.” *Bioinformatics* 29.1 (2013):84-91.

Messer, P. W. and D. A. Petrov. "Population genomics of rapid adaptation by soft selective sweeps." *Trends in Ecology & Evolution* 28.11 (2013): 659-669.

Miles, A. and Harding, N. J. "cggh/scikit-allel: v0.21.2." (2016). Zenodo. <http://doi.org/10.5281/zenodo.157626>

Miles A. *et al*., Ag1000g…

N’Guessan, R. *et al*. "Reduced efficacy of insecticide-treated nets and indoor residual spraying for malaria control in pyrethroid resistance area, Benin." *Emerging Infectious Diseases* 13.2 (2007): 199.

O'Reilly, A., *et al*. "Modelling insecticide-binding sites in the voltage-gated sodium channel." *Biochemistry Journal* 396 (2006): 255-263.

Pinto, J., *et al*. "An island within an island: genetic differentiation of *Anopheles gambiae* in Sao Tome, West Africa, and its relevance to malaria vector control." *Heredity* 91.4 (2003): 407-414.

Pinto, J. *et al*. "Multiple origins of knockdown resistance mutations in the Afrotropical mosquito vector *Anopheles gambiae*." *PLoS One* 2.11 (2007): e1243.

Posada, D. and K. A. Crandall. "Intraspecific gene genealogies: trees grafting into networks." *Trends in Ecology & Evolution* 16.1 (2001): 37-45.

R Development Core Team. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna. (2014) Available at: [http://www.R-project.org](http://www.r-project.org/).

Ranson, H. *et a*l. "Identification of a point mutation in the voltage‐gated sodium channel gene of Kenyan *Anopheles gambiae* associated with resistance to DDT and pyrethroids." *Insect Molecular Biology* 9.5 (2000): 491-497.

Ranson, H. *et al*. "Pyrethroid resistance in African anopheline mosquitoes: what are the implications for malaria control?" *Trends in Parasitology* 27.2 (2011): 91-98.

Reimer, L. J. *et al*. "An unusual distribution of the *kdr* gene among populations of *Anopheles gambiae* on the island of Bioko, Equatorial Guinea." *Insect Molecular Biology* 14.6 (2005): 683-688.

Rinkevich, F. D., D. Yuzhe Du, and K. Dong. "Diversity and convergence of sodium channel

mutations involved in resistance to pyrethroids." *Pesticide Biochemistry and Physiology*

106.3 (2013): 93-100.

Sabeti, P. C. *et al*. "Detecting recent positive selection in the human genome from haplotype structure." *Nature* 419.6909 (2002): 832-837.

Sabeti, P. C. *et al*. "Positive natural selection in the human lineage." *Science* 312.5780 (2006): 1614-1620.

Santolamazza, F. *et al*. "Remarkable diversity of intron-1 of the para voltage-gated sodium channel gene in an *Anopheles gambiae/Anopheles coluzzii* hybrid zone." *Malaria Journal* 14.1 (2015): 1-10.

Silva, A. P. B. *et al*. "Mutations in the voltage-gated sodium channel gene of anophelines and their association with resistance to pyrethroids–a review." *Parasites and Vector*s 7 (2014): 450.

Sonoda, S. *et al*. "Genomic organization of the para‐sodium channel α‐subunit genes from

the pyrethroid‐resistant and‐susceptible strains of the diamondback moth." *Archives of Insect*

*Biochemistry and Physiology* 69.1 (2008): 1-12.

Stump, A. D. *et al*. Centromere-proximal differentiation and speciation in *Anopheles gambiae. Proceedings of the National Academy of Sciences* 102 (2005): 15930–15935.

Toé, K. H. et al. "Increased pyrethroid resistance in malaria vectors and decreased bed net effectiveness, Burkina Faso." *Emerging Infectious Disease*s 20.10 (2014): 1691.

Turner, T. L., M. W. Hahn, and S. V. Nuzhdin. "Genomic islands of speciation in *Anopheles gambiae*." *PLoS Biology* 3.9 (2005): 1572.

Weetman, D. *et al*. "Contemporary evolution of resistance at the major insecticide target site

gene *Ace‐1* by mutation and copy number variation in the malaria mosquito *Anopheles*

*gambiae*." *Molecular Ecology* 24.11 (2015): 2656-2672.

Williamson, M. S. *et al*. "Identification of mutations in the housefly para-type sodium channel gene associated with knockdown resistance (*kdr*) to pyrethroid insecticides. *Molecular Genetics and Genomics* 252 (1996): 51-60.

World Health Organization (WHO). "Global Plan for Insecticide Resistance Management (GPIRM)." Geneva (2012).

World Health Organisation. "Test procedures for insecticide resistance monitoring in malaria

vector mosquitoes." Geneva (2013).