Targeted genomic surveillance of insecticide resistance in African malaria vectors

Sanjay C Nagi[1*], Eric R Lucas[1], Faisal Ashraf[1], Trevor Mugoya[2], Edward Lukyamuzi[2], Shannan Summers[1], Calvin Yee[1], Christopher G Jacob[5], Harun Njoroge[1], Thomas Pemberton[1], John Essandoh[3], Martin Lukindu[2], Honorine Kaboré[1,4], Grégoire Sawadogo[1,4], Jessica Williams[1], Arjen E. Van't Hof[1], Anastasia Hernandez-Koutoucheva[7], Christina Hubbart[5], Kate Rowlands[5], Anna Jeffreys[5], Scott Goodwin[7], Naomi Park[6], Cristina Ariani[7], Alexander Egyir-Yawson[3], Sonia Goncalves[6], Shavanthi Rajatileka[7], Kirk Rockett[6], Victoria J. Simpson[7], Alistair Miles[6], David Weetman[1], Jonathan Kayondo[2], Tony Nolan[1], Martin J Donnelly[1,6]

[1]Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, UK.

[2]Department of Entomology, Uganda Virus Research Institute, Entebbe, Uganda

[3]Department of Biomedical Sciences, University of Cape Coast, Cape Coast, Ghana

[4]Institut de Recherche en Sciences de la Santé (IRSS), 01 BP 545, Bobo-Dioulasso 01, Burkina Faso

[5]Centre for Human Genetics, Nuffield Department of Medicine, Oxford, OX3 7BN, UK

[6]Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SD, UK

[7]Genomic Surveillance Unit, Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SD, UK

[*]Corresponding author: Email: Sanjay.Nagi@lstmed.ac.uk

Keywords: Genomic Surveillance, Amplicon sequencing, Anopheles, Insecticide resistance

Abstract

The emergence of insecticide resistance is threatening the efforts of malaria control programmes, which rely heavily on a limited arsenal of insecticidal tools, such as insecticide-treated bed nets. Importantly, genomic surveillance of malaria vectors can provide critical, policy-relevant insights into the presence and evolution of insecticide resistance, allowing us to maintain and extend the shelf life of these interventions. Yet the complex genetic architecture of resistance, combined with resource constraints in malaria-endemic settings, have thus far precluded the widespread use of genomics in routine surveillance. Meanwhile, stakeholders in sub-Saharan Africa are moving towards locally driven, decentralised generation of genomic data, underscoring the need for standardised and robust genomics workflows. To address this need, we demonstrate an approach to targeted genomic surveillance in *Anopheles gambiae s.l* with Illumina sequencing. We target 90 genomic loci in the *Anopheles gambiae s.l* genome, including 55 resistance-associated mutations and 35 ancestry informative markers. This protocol is coupled with advanced, automated software for accurate and reproducible variant analysis. We are able to elucidate population structure and ancestry in our cohorts and accurately identify most species in the *An. gambiae* species complex. We report frequencies of variants at insecticide-resistance loci and explore the continued evolution of the pyrethroid target site, the *Voltage-gated sodium channel.* Applying the platform to a recently established colony of field-caught resistant mosquitoes (Siaya, Kenya), we

identified seven independent resistance-associated variants contributing to reduced efficacy of insecticide-treated nets in East Africa. Additionally, we leverage a machine learning algorithm (XGBoost) to demonstrate the possibility of predicting bioassay mortality using genotypes alone. This achieved high accuracy (75%), demonstrating the potential of targeted genomics to predictively monitor insecticide resistance. Together these tools provide a practical, scalable solution for resistance monitoring while advancing the goal of building local genomic surveillance capacity in sub-Saharan Africa.

Introduction

Insecticidal interventions remain the cornerstone of malaria vector control, however, sustained control is challenged by the emergence and spread of insecticide resistance among the *Anopheline* vectors[1]. Addressing the problem of escalating resistance is vital to maintaining progress towards malaria elimination, which has stalled in recent years[2]. New active ingredients have recently become available, including long-lasting insecticidal nets (LLINs) with dual active-ingredients, such as pyrethroid-pyrrole (chlorfenapyr) and pyrethroid-synergist (PBO) combinations. After promising epidemiological results in randomised-controlled trials[3–6], dual chlorfenapyr and PBO nets already make up the vast majority of nets delivered to sub-Saharan Africa[7]. Their widespread deployment raises concerns about replicating the earlier over-reliance on pyrethroid-only nets over the past two decades; monitoring wild vector populations for existing and newly emerging resistance to these novel active ingredients will be crucial to extend their longevity.

Genomic surveillance has emerged as a pivotal approach to understanding the spread and evolution of pathogens in many infectious disease systems (for example, *SARS-CoV-2*[8], *mpox*[9], *mtb*[10]) and offers great potential to help mitigate insecticide resistance in malaria vectors. By monitoring known and identifying new resistance-conferring mutations before they become widespread, control programs can implement proactive targeted interventions, such as rotating insecticides, deploying alternative active ingredients, or prioritising resistant vector populations for intensified monitoring. These measures help limit the spread of resistance and sustain the efficacy of existing vector control strategies. Therefore, a thorough understanding of the link between resistance genotype and phenotype is essential to devise mitigation strategies, however, current phenotype-based methods for the surveillance of insecticide resistance are both labour-intensive and imprecise[11,12]. Whole genome sequencing has revealed insights into mosquito biology[13] and many genes and mutations in the *An. gambiae* genome involved in insecticide resistance[14–19]. Although whole genome sequencing is powerful, the relatively large

size of mosquito genomes (*An. gambiae* ≈ 278Mb) makes targeting specific genomic loci a cheaper and more tractable option for routine surveillance activities.

In this study, we develop a cost-effective approach to targeted genomic surveillance in *Anopheles gambiae sl*. We introduce Ag-vampIR: the *An. gambiae* Vector Amplicon Marker Panel for Insecticide Resistance, which contains 80 amplicons targeting 90 SNP markers, designed for Illumina sequencing platforms. The markers are selected to target known insecticide resistance loci and ancestry informative markers for the resolution of species within the *An. gambiae* complex. We also present open-source software, AmpSeeker, a computational workflow designed for the comprehensive analysis of any Illumina amplicon sequencing data. Leveraging the power of Snakemake[20], AmpSeeker streamlines the entire analytical pipeline, from raw data processing to variant calling and downstream analyses, and generates a local webpage for users to conveniently explore the results. Our study presents the development and validation of this workflow alongside the *An. gambiae s.l.* Ag-vampIR amplicon sequencing panel. Together, these tools aim to enhance malaria vector surveillance through scalable, efficient, and reproducible monitoring of insecticide resistance.

Results

Platform and protocol development

To perform targeted genomic surveillance of *Anopheles gambiae sl,* we developed an Illumina amplicon sequencing panel, Ag-vampIR, and a library preparation protocol. The Ag-vampIR panel is based on the *An. gambiae* AgamP4 PEST reference genome and targets 90 SNP markers in total, split over 82 amplicons (Figure 1A). It incorporates major known insecticide resistance loci (55 SNPs in total), including target-site mutations, such as *Kdr, Ace1* or *Rdl*, markers for metabolic resistance, or SNPs that tag selective sweeps prevalent in the genomes of wild specimens from the *Anopheles* 1000 genomes project (Ag1000G) (Supplementary Data 1, Supplementary Text 2). Thirty-five genomic targets are ancestry informative markers to allow the accurate resolution of species, and we include two amplicons at the intron 4–exon 5 boundary of *doublesex*, the target of a recently developed population suppression gene drive [21]. Each amplicon is approximately 200 bp long, and our method makes full use of both known and novel SNPs present on Ag-vampIR amplicons. The laboratory protocol utilises an indexing PCR reaction and a combination of 96 custom i7 adaptors and 16 i5 adaptors[22], to give a theoretical maximum of 1536 samples which can be multiplexed on a single flow cell. The entire amplicon sequencing protocol is summarised in Figure 1B.

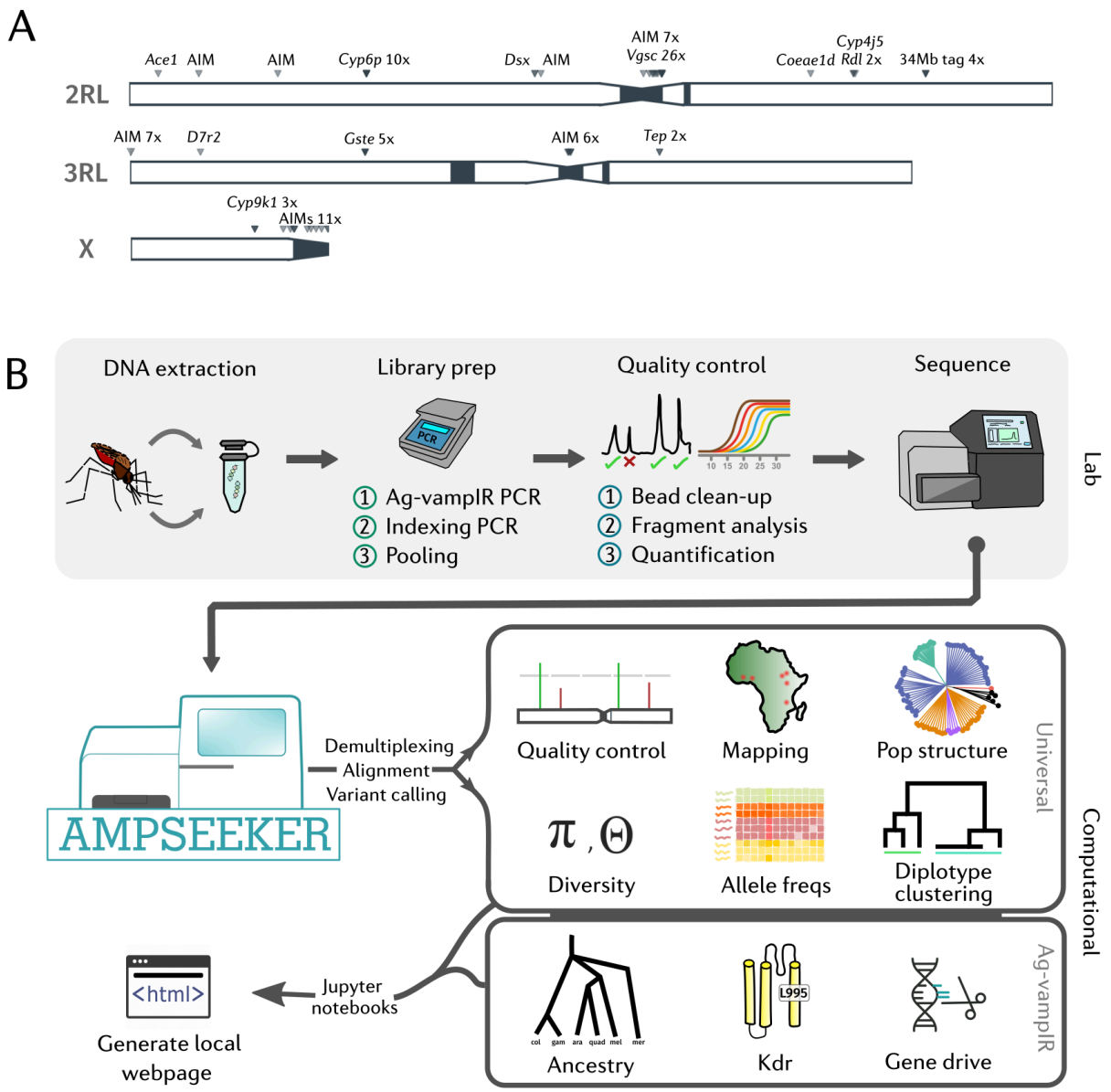

Figure 1 Ag-vampIR and AmpSeeker. A) A diagram of the genomic regions and SNP markers targeted by the Ag-vampIR panel (AgamP4 PEST reference genome). Heterochromatic regions are coloured black (for locus details see supplementary data 1). **B)** Schematic overview of the laboratory protocol and bioinformatic pipeline for targeted genomic surveillance. Upper: the laboratory and library preparation protocol. Lower: the AmpSeeker modules and computational workflow.

Pipeline development

To enable reproducible, decentralised analysis of the amplicon data, we developed an open-source computational pipeline, AmpSeeker. The pipeline is written in Snakemake [20], and facilitates automated, end-to-end analysis. AmpSeeker can be applied to any Illumina amplicon sequencing data, with additional optional analyses for the Ag-vampIR panel. It is modular in

architecture, allowing the analysis of further panels to be easily incorporated. Analyses are written in Python, the most popular programming language in the life sciences. The workflow is convenient to use and utilises Jupyter notebooks combined with Jupyter Book to generate a local web page of results for easy data exploration. Throughout this manuscript, we display outputs of the AmpSeeker pipeline.

Sample collections and sequencing

We collected *An. gambiae s.l* mosquitoes from four cohorts across sub-Saharan Africa; two field populations; Obuasi, Ghana (2017), and the Upper River Region, The Gambia (2011), and from two laboratory strains, Siaya (Western Kenya, established in 2023) and VK7 (Burkina Faso, 2014) (Supplementary Table 1). These cohorts were selected to provide a diverse mixture of ancestries and resistance profiles with which to evaluate the targeted surveillance platform. We sequenced 542 samples in total on an Illumina MiSeq with a reagent kit V2.

Quality control

Following sequencing, the MiSeq output was imported into AmpSeeker, which performs demultiplexing followed by alignment to the reference genome with bwa mem [23], and variant calling with bcftools mpileup [24]. Following variant calling, we perform sample quality control; removing samples with low depth, and samples with extreme levels of heterozygosity or outliers in a principal components analysis (PCA). This resulted in a reduction from 542 samples to 467.

We then evaluated the accuracy of AmpSeeker's variant calling, using whole-genome sequence data as a gold standard. 40 individuals (Obuasi, Ghana) were sequenced to 30x as part of the Ag1000g project, and previously analysed as part of a genome-wide association study [14]. We examined the concordance between genotypes for these 40 samples at 11 variant read depth thresholds (Supplementary Figure 1). At 20X and above, we observe highly concordant variant calls (> 99.3%), demonstrating that our method is reliable and accurate.

Taxon analysis and population structure

The *An. gambiae* species complex contains seven recognised species and at least four additional cryptic taxa, with varying contributions to malaria transmission[25–27]. Species within this complex are morphologically identical, underlying the importance of molecular methods to delineate the ancestry of an individual. Laboratory-based methods PCR assays[28,29] are used to resolve species, however, these methods rely on single loci and cannot resolve cryptic taxa. Methods that rely on single loci also have limited power to detect hybridisation between species, a phenomenon which can occur in areas of sympatry. In order to accurately identify species, we developed an

ensemble of three complementary methods: (1) an agnostic method based on ancestry informative markers (AIMs), (2) a PCA + SVM classifier and (3) an XGBoost classifier - both utilising reference data from 7790 published whole-genome sequences from the Ag1000g (see Methods). Taxons were assigned where at least two methods agreed, otherwise samples were designated as 'unassigned'. Our AIM approach targets thirty-five AIMs from those published in Miles *et al.*[13], allowing species delineation between *An. gambiae* and *An. coluzzii* at a greater resolution than traditional methods. AIMs are SNPs which show marked frequency or fixed differences between species and therefore give insights into the ancestry of a specimen. The PCA and XGBoost approaches allow us to expand our taxonomic analysis to other species in the *An. gambiae* complex whilst accounting for potential missing genotypes. To evaluate population structure alongside our taxonomic analysis, we built neighbour-joining trees and performed PCA.

Our analysis of population structure and ancestry revealed distinct patterns across sampling cohorts (Figure 2). The neighbour-joining tree and principal component analysis (Figure 2B, Supplementary Figure 2) separated our samples into five distinct clusters, which aligned with both geographic origin and assigned taxa. These clusters included: a large group of *An. arabiensis* from The Gambia, the Siaya colony samples (potentially the Pwani molecular form[27]), a cluster of *An. gambiae* primarily from Obuasi with four samples from The Gambia, and two clusters of *An. coluzzii* comprising the VK7 lab colony and a subset of the Gambian samples. Taxon assignments can be found in Supplementary Table 5 and Supplementary Data 3.

The AIM analysis provided further insights into the ancestry of these samples (Figure 2A). The cohort from Obuasi, Ghana, contains *An. gambiae* s.s ancestry alone, in correspondence with taxon calls from the Ag1000g project. All Obuasi samples were correctly assigned as *An. gambiae* by our ensemble taxon caller. The VK7 strain of *An. coluzzii* appears to have AIMs consistent with *An. coluzzii* ancestry, apart from eight AIMs on the 2L chromosomal arm. The AIMs on this contig are clustered near the beginning of the arm, close to, or within the *Voltage-gated sodium channel (Vgsc)*, and the *An. gambiae*-like AIM patterns seen here are due to VK7 harbouring the 995F F1 *Kdr* haplotype (see *Vgsc* analysis below). This haplotype confers resistance to pyrethroid insecticides, and has introgressed from *An. gambiae* into *An. coluzzii* in West Africa[30]. The ensemble caller called all VK7 individuals as *An. coluzzii*.

The Gambia cohort contains mostly (*n*=45) a subgroup with an apparent mix of *An. gambiae* vs *An. coluzzii* ancestry, but also 17 samples with predominantly *An. coluzzii* genotypes and 4 with

An. gambiae genotypes. While mixed patterns of *An. gambiae* vs *An. coluzzii* AIMs could indicate recent hybridisation or cryptic taxa (such as the recently reported Bissau molecular form[25]) both our XGBoost classifier and the principal components analysis predicted that the 'mixed' ancestry samples were in fact *An. arabiensis* (Supplementary Figure 3A). All three taxa are known to inhabit the Upper River Region of The Gambia [31].

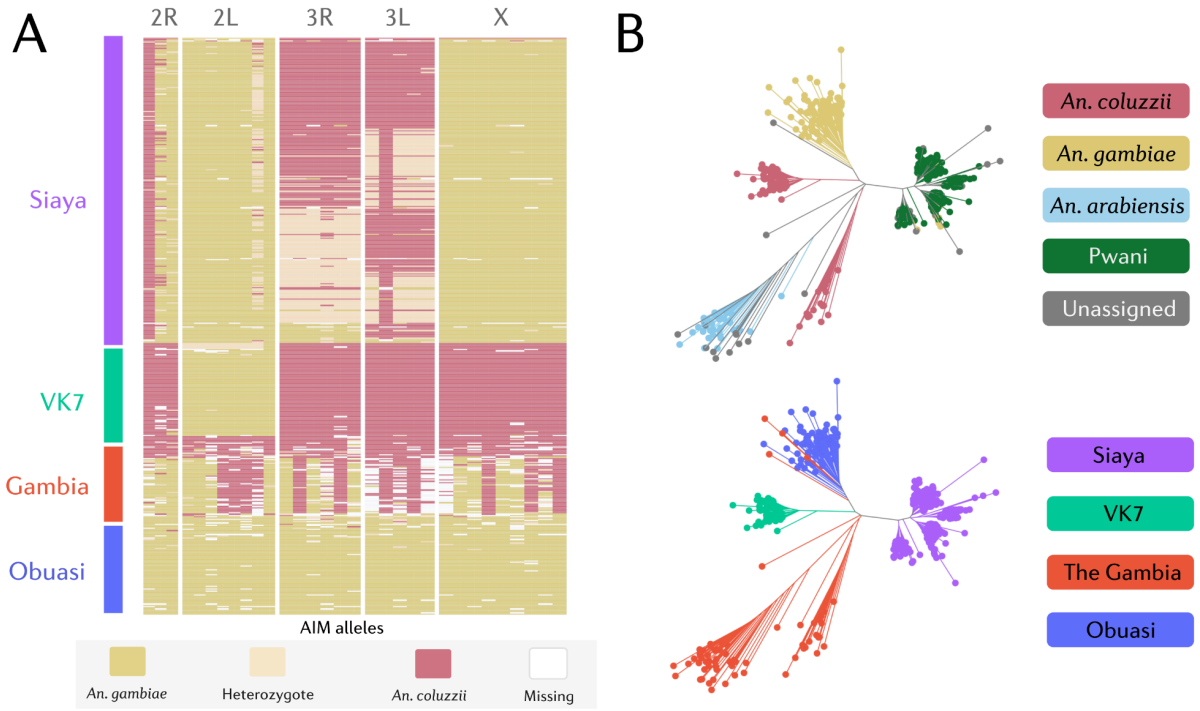

Figure 2. Ancestry and population structure. A) A heatmap of *An. gambiae* vs *An. coluzzii* ancestry informative marker (AIM) genotypes. Each row of the heatmap is a different sample, and each column a different AIM. AIMs are coloured according to the ancestral allele (gold: *An. gambiae*, maroon: *An. coluzzii*, and beige: heterozygote). B) A neighbour-joining tree of the 467 QC-passed amplicon sequencing samples coloured by sample taxon according to consensus taxon calls (upper) and sample location (lower).

The Siaya colony presented an intriguing pattern: ancestry consistent with *An. gambiae* on chromosome 2 and the X chromosome, but *An. coluzzii*-like ancestry on chromosome 3 (Figure 2A). This is surprising, as despite a recent finding of *An. coluzzii* in northern Kenya, this species has never been reported in Siaya county, western Kenya. Laboratory contamination is also unlikely, as the Siaya colony had never been maintained in the same insectary as a strain of *An. coluzzii*. Given the unusual AIM patterns in Siaya mosquitoes originating from Western Kenya, we hypothesised that this may relate to the origin of the AIMs themselves, which were ascertained

from an earlier study of West African mosquitoes. To investigate this further, we evaluated the performance of these AIMs using Ag1000g data from East Africa specifically (Kenya, Uganda, Tanzania, 2880 samples) and found that most AIMs remained highly discriminative (>97% concordant) for species identification in the region. Further analysis with PCA (Supplementary Figure 3B) showed Siaya clustering in close proximity to both the *An. gambiae* cluster and adjacent to the cluster of the recently described *Pwani* molecular form (labelled *Gcx3*)[27]. In Ag1000g data, the *Pwani* form exhibits mixed AIM patterns on autosomes, in comparison to samples assigned to *An. gambiae*, which consistently show *An. gambiae*-like AIMs on the autosomes (Supplementary Figure 4). *Pwani* was thought to be limited to the coast of East Africa, however, a single individual female mosquito assigned as *Pwani* in the Ag1000g was collected in Muleba, Tanzania, on the West coast of Lake Victoria in 2015 (Supplementary Figure 5), suggesting that this subspecies may also exist inland.

Surveillance of the Voltage-gated sodium channel

The Voltage-gated sodium channel (*Vgsc*) is the target of pyrethroids and DDT and across insect taxa has been the site of rapid evolution [32]. In *An. gambiae*, phenylalanine and serine mutations at the 995 codon, known as *kdr*, lower the binding potential of the channel, enabling mosquitoes to survive exposure [33,34]. Both 995F and 995S mutants have arisen independently on multiple occasions, with some origins remaining in more localised areas and others spreading over much of sub-Saharan Africa [18]. We calculated allele frequencies at known SNPs within the *Vgsc* across our sample populations (Supplementary Figure 6). We also used tagging SNPs targeted by the Ag-vampIR panel to estimate the origins of *Kdr* haplotypes in our data (Supplementary Table 5).

In the Siaya strain from Western Kenya, we observed a mixture of 995F (65.9%) and 995S (33.7%) alleles, consistent with recent evidence of the westward spread of 995F in *An. gambiae* across central and East Africa[35]. Siaya contained multiple origins of the 995S allele (S1=31% and S3=1.5%) and a single origin of 995F, the F2 haplotype. VK7 were fixed for 995F (100%), with no evidence of the recently described V402L/I1527T haplotype, in agreement with its low frequency in West Africa during the time period of the colony's establishment [18,36]. The VK7 995F *Kdr* alleles all belonged to the major F1 haplotype. In The Gambia, *An. arabiensis* mosquitoes contain the 995S mutation at 16% frequency. We were not able to identify the origin of this mutant, which could suggest a novel origin in this sub-species. In samples assigned as *An. gambiae* from The Gambia, mosquitoes were fixed for 995F, with 83% of those samples belonging to the F1 haplotype, and the other 17% undetermined. In *An. coluzzii* from The Gambia, most samples were assigned as wild-type, with a small proportion (7.6%) of 995F alleles.

To visualise haplotypes and mutations at the *Vgsc* in more detail, we performed diplotype clustering at this locus, plotting a dendrogram alongside sample heterozygosity and any *Vgsc* mutations present in our data (Figure 3A). In agreement with the *Kdr* origins analysis, we identified two distinct haplotype backgrounds associated with the 995F mutation. The VK7 and Obuasi cohorts shared a common 995F haplotype (F1) characterized by multiple secondary mutations across the *Vgsc* gene region, including T791M, N1570Y, A1746S, A1853I, and P1874S/L. In comparison, Siaya contains the F2 haplotype lacking these additional variants.

Allele frequencies at resistance-associated SNPs

We next surveyed other resistance-associated mutations targeted by Ag-vampIR (Figure 3B). The *Ace1*-G280S mutation, known to confer resistance to organophosphates and carbamates, was found close to fixation in *An. gambiae* from Obuasi (97% frequency), but was completely absent in VK7, with very few resistance alleles found in the Gambia (1) or Siaya (6). The two *Rdl* mutations A296G and A296S, which confer resistance to dieldrin, remain at moderate frequencies in sub-Saharan Africa, despite a ban on dieldrin since 1974. We found A296S at 24.2% in Siaya, but rare or absent in other cohorts, whereas A296G was at 50.7% frequency in the Obuasi cohort, but absent in other cohorts.

We also monitor metabolic resistance variants, many of which were selected to track selective sweeps revealed by phase 1 of the Ag1000g (Supplementary Figure 9). Eight amplicons target the Cyp6p locus, and we find the *Cyp6p3*-I88T mutation at high (83.6%) frequency in VK7. The *Cyp6p4*-I236M mutation, which is associated with a gene duplication in East Africa [37], is at 61.7% frequency in Siaya. The *Cyp6p1*-L374M mutation at high frequency in Obuasi, tags an ongoing selective sweep found in *An. gambiae* from West Africa, originally identified in samples from Cameroon and Burkina Faso. Two mutations at *Cyp9k1* that tag a selective sweep are fixed in VK7 (100%), whilst again present at low frequencies in The Gambia (5.2%, 7.7%). Consistent with this, these SNPs tag a selective sweep at *Cyp9k1* originally found in *An. coluzzii* from Burkina Faso [13]. At the *Gste* gene cluster, members of which have been associated with resistance to DDT, organophosphates, and pyrethroids, the *Gste2*-I114T mutation is at 83.1% frequency in VK7, but at low frequencies in the Obuasi (6.7%) and Gambia (6.8%) cohorts and absent in Siaya. The *Gste2*-L119V mutation is at 61.9% frequency in Obuasi, and 5.4% in The Gambia. The low frequency of resistance mutations in The Gambia is likely due to the dominance of *An. arabiensis* in the cohort.

The *Cyp4j5*-L43F mutation, which has repeatedly been associated with resistance to pyrethroids in East Africa [38,39], is found in all cohorts at moderate frequencies, 19.8% in Obuasi, 65.5% in the

Gambia, 97.3% in VK7, and 46.2% in Siaya. In East Africa (and therefore Siaya), this mutation is at very high frequency on the 2La inversion karyotype, and is therefore acting as a tagging SNP for the inversion. The same may not be true in West Africa, where it can reach higher frequencies and has not been associated with pyrethroid resistance. Three tagging SNPs tag a selective sweep at approximately 34Mb on the 2L chromosomal arm, again found on the 2La inversion karyotype. This sweep is ongoing in *An. gambiae* from East Africa, and appears to be at 46% frequency in Siaya, whilst absent from other cohorts.

We also find the TEP1 *Plasmodium*-refractoriness allele fixed in VK7 (100%), and at low frequencies in The Gambia (8.3%). AmpSeeker also reports allele frequencies of mutations not specifically targeted by Ag-vampIR, which we provide as an exercise for the reader (https://sanjaynagi.github.io/agvampir002-results/notebooks/allele-frequencies.html) .

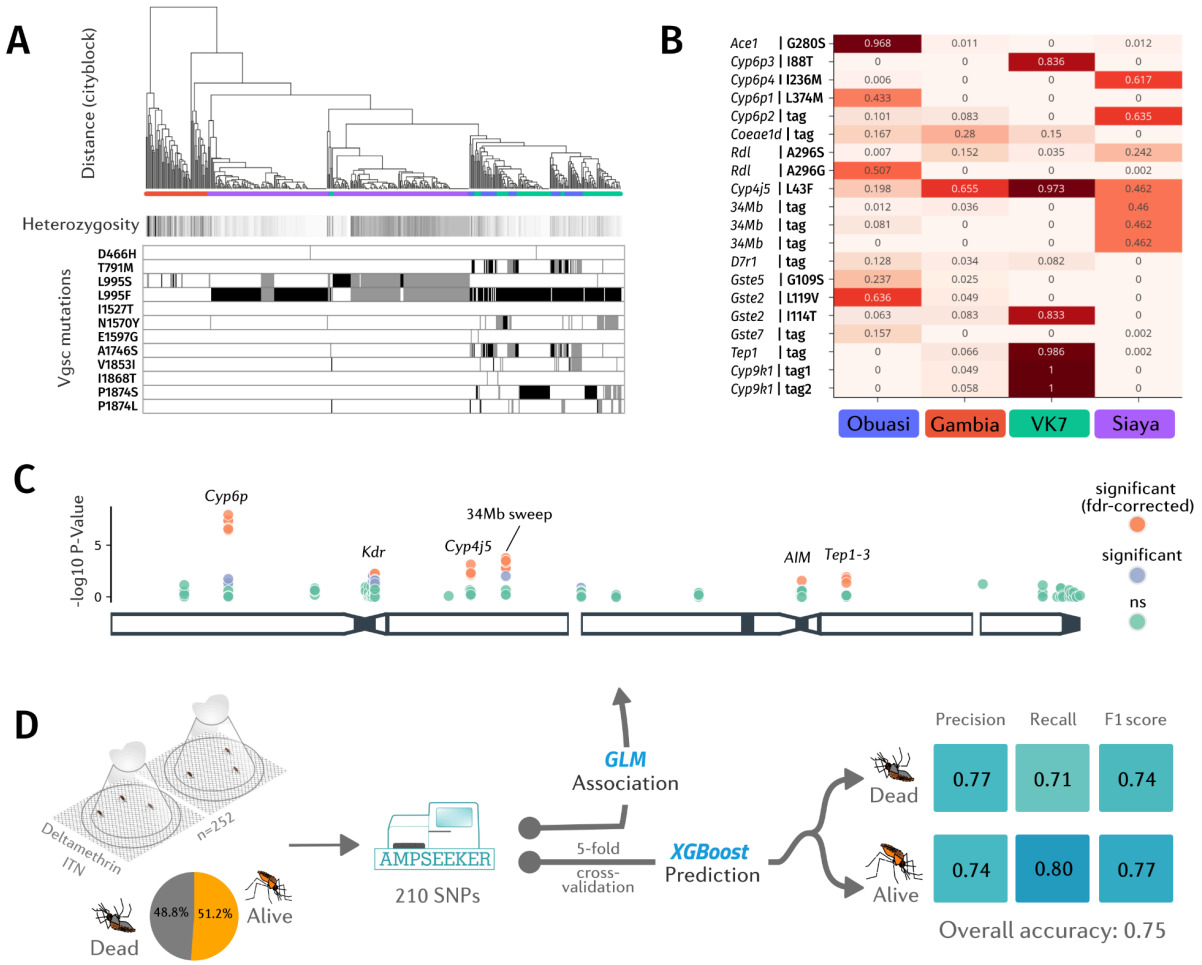

Figure 3. Genomic surveillance of insecticide resistance with Ag-vampIR and AmpSeeker. A) Diplotype clustering at the Voltage-gated sodium channel. We calculate pairwise distance between diplotypes at all *Vgsc* sites. Each column in the figure is a diplotype ordered by the dendrogram by hierarchical

clustering, using genetic distance based on city-block (Manhattan) distance and complete linkage. The leaves of the dendrogram are colored by the species of the individual to which they belong. Underneath the dendrogram, the heterozygosity (calculated as an average over all SNPs in the *Vgsc* locus) is displayed as horizontal bars. Clusters with low sample heterozygosity or inter-sample genetic distances of 0 are indicative of a selective sweep. Amino acid variation is displayed below the dendrogram - black indicates homozygosity and grey heterozygosity. B) Allele frequencies at other known insecticide-resistance SNPs targeted by Ag-vampIR. C) A Manhattan plot of the amplicon-wide association study using a generalised linear model. Siaya mosquitoes were exposed to deltamethrin-treated nets (PermaNet 2.0) in extended 40 minute WHO cone assays. D) A schematic of the approach to predicting the insecticide resistance phenotype from WHO cone assays, highlighting the Precision, Recall, F1-score and overall accuracy of the XGBoost algorithm in both prediction classes (Dead & Alive).

Genotype:phenotype association study

The much lower cost of amplicon sequencing compared to whole-genome sequencing enables association studies with larger sample sizes. The samples from the Siaya cohort were phenotyped against new PermaNet 2.0 deltamethrin-impregnated nets and we tested whether any variants genotyped by Ag-vampIR were associated with mosquito survival. For each SNP, we built a binomial GLM with the SNP as the predictor and phenotype as the response variable, followed by false discovery rate control to correct for multiple testing. For this analysis, we utilised any SNPs present on Ag-vampIR amplicons, after filtering, this resulted in 210 SNPs. The full results of the association study are in Supplementary Data 2 and are visualised as a manhattan plot in Figure 3C.

After false-discovery rate control, thirty-seven SNPs were significantly associated with survival following exposure to PermaNet 2.0 nets, corresponding to seven independent mutations. Thirteen SNPs in the Cyp6p gene cluster, including *Cyp6p4*-I236M, were positively and strongly associated with mosquito survival. All thirteen showed highly similar odds ratios ranging between 3.16 and 3.65. The presence of *Cyp6p4*-I236M suggests that Siaya harbors the triple-mutant haplotype described in Njoroge *et al.*, [37], and the twelve remaining SNPs are therefore all SNPs found on this mutant haplotype and are tagging *Cyp6aa1*-Dup1. At the *Vgsc*, the 995F mutation was significantly positively associated with survival (OR=1.78, 95% CIs=1.18-2.68), whereas 995S was negatively associated (OR=0.56, 95% CIs=0.37-0.84). These data are consistent with the literature, suggesting that 995F confers greater resistance to deltamethrin, and that 995F is currently replacing 995S in East Africa [17,35].

Four SNPs at *Cyp4j5* were associated with resistance, including the non-synonymous mutations *Cyp4j5*-A47P (OR=1.78, 1.16-2.72), *Cyp4j5*-L43F (OR=2.25, 1.40-3.60), and *Cyp4j5*-T30W (OR=1.94, 1.22-3.07). Fourteen SNPs at the 34Mb sweep locus are associated with survival; all with similar

odds ratios ranging from 2.10 to 2.48. In Ag1000g data, there is only a single selective sweep at this locus, and so these fourteen SNPs are likely to be tagging this selective sweep. This is the first evidence that this haplotype confers resistance to insecticides used in vector control. A SNP (3L:380924, C), which is found on the same amplicon of an ancestry informative marker (3L:380974, C), is negatively associated with survival (OR=0.56, 0.37-0.83). The SNP is located in an intron of AGAP029563, a gene without a description or any functional annotation. Three intergenic SNPs at the TEP locus, previously associated with *Plasmodium* refractoriness, were also associated with mosquito survival, in agreement with recent GWAS results suggesting that this locus may also be important for insecticide resistance[14]. One SNP (3L:11214067, G) was positively associated (OR=1.96, 1.20-3.21), whilst two others, likely to be found on a competing haplotype, were strongly negatively associated with survival with odds ratios of 0.29 (0.14-0.63) and 0.33 (0.16-0.68).

Predicting bioassay survival with machine learning

Given the number of loci associated with bioassay survival to insecticide-treated nets, we hypothesised that it may be possible to build models using machine learning that are predictive of bioassay mortality. We trained an XGBoost model on the 252 phenotyped Siaya samples at the 210 genotyped SNPs using five-fold cross-validation, to predict the mortality of individual mosquitoes. Figure 3D shows a schematic of the phenotypic prediction using XGBoost.

The model was able to predict with 75% accuracy (weighted F1-score of 0.75) the classification of individual mosquitoes as either dead or alive in the bioassay. Cross-validation demonstrated consistent performance across different data splits (CV score: 0.754 ± 0.116). Analysis of feature importance revealed several key genetic variants driving the predictions, with the most influential SNPs located at 2L:28502739 (*Cyp6p* cluster), 3L:11208428 (*Tep* cluster), and 2L:34100775 (34Mb sweep), in agreement with the earlier association study. The model showed balanced performance between predicting both dead (precision: 0.77, recall: 0.71) and alive phenotypes (precision: 0.74, recall: 0.80), suggesting robust predictive capability across both outcomes. This substantially outperformed random baseline models (0.37-0.49), demonstrating that the genetic variants captured meaningful signals predictive of insecticide resistance.

Cost-effectiveness of Ag-vampIR

In Supplementary Table 6, we estimate costs of reagents and kits used in the Ag-vampIR protocol. The estimated per sample cost, based on 500 samples run on a MiSeq reagent kit V2, is $9.23, which equates to ~$0.10 for each of the ninety loci targeted. For comparison, whole-genome sequencing is estimated to cost over $100 dollars per sample, whilst an LNA

probe-based qPCR assay would cost an estimated ~$2.475 per locus genotyped. These costs include performing DNA extractions with a column-based proprietary kit (Nexttec), the most expensive part of the protocol, which could feasibly be removed for a simpler, cheaper extraction [40].

Discussion

We have developed an integrated lab protocol and computational workflow to perform effective, targeted, end to end genomic surveillance of *An. gambiae s.l.* We show that Ag-vampIR and AmpSeeker can rapidly generate accurate and readily-interpretable data at an affordable cost, opening up novel areas of research and enabling widespread genomic surveillance.

Overall, the cost-effectiveness of the platform allows genomic surveillance on a much larger scale, in comparison to whole-genome sequencing [14,41]. We demonstrate how this can be particularly useful in performing genotype:phenotype association studies, boosting power to detect variants involved in the insecticide-resistant (or any other) phenotype. The association study with the Siaya colony yielded important insights into pyrethroid resistance mechanisms, identifying seven independent variants significantly associated with survival against PermaNet 2.0 nets. The study's findings validated several known resistance mechanisms while uncovering new associations, for example, at the previously uncharacterized 34Mb sweep locus.

The demonstration that Ag-vampIR genotypes alone could predict WHO cone bioassay mortality with good accuracy is particularly promising. The XGBoost model's performance (weighted F1-score of 0.75) suggests that genetic data could potentially serve as a proxy for phenotypic resistance testing, though further validation across different populations and bioassay protocols is necessary. Whilst strong predictive power for organophosphate and carbamate resistance has been demonstrated for the Ace-1 G280S mutation [16,42], high prediction capacity for pyrethroid resistance is impressive given the multivariate nature of resistance mechanisms in contemporary African populations of *An. gambiae* [14,43]. This capability could transform resistance monitoring by enabling rapid, cost-effective screening of vector populations, allowing control programmes to make data-driven decisions about insecticide deployment. Whilst phenotypic testing will still play a role, this predictive approach could enable more frequent and widespread resistance surveillance than is currently feasible with bioassays.

Our approach also proved highly effective for exploring population structure and ancestry in our samples. By targeting ancestry informative markers (AIMs), we were able to both distinguish

between *An. gambiae* and *An. coluzzii* and accurately identify other sibling species including *An. arabiensis* and *An. melas*. The high accuracy of our machine learning approaches in species classification (F1 scores >0.99) demonstrates the discriminatory power of these markers, a potentially surprising finding, given that they were intended to differentiate only *An. gambiae* and *An. coluzzii*.

Robust and standardised tools that facilitate the rapid and reproducible analysis of genomic data are required for effective genomic surveillance. Whilst bioinformatic capacity in sub-Saharan Africa is increasing, there remains a paucity of standardised analysis methods for genomic data, outside of the Vector Observatory[44]. In developing AmpSeeker, we provide an all-purpose tool for scalable and reproducible analysis of Illumina amplicon sequencing data. AmpSeeker's novelty is in its generation of a local webpage of results, allowing the user to explore the data interactively, and enabling the rapid, interpretable dissemination of results to partners and stakeholders alike. The interactive nature of the results, which includes descriptive, explanatory text, we hope facilitates an understanding of complex genomic data, potentially bridging the gap between bioinformaticians and field researchers or policymakers. AmpSeeker is open-source and written in Python, a mature and popular programming language for bioinformatics, increasing the likelihood of long-term sustainability and ease of maintenance. Its open-source nature aligns with the principles of open science, promoting transparency and collaboration. We are currently delivering workshops and collaborating with research groups across sub-Saharan Africa to facilitate its implementation. We are also expanding the utility of AmpSeeker to amplicon panels from other organisms. Given its applicability to any amplicon sequencing data, and the ease to which analysis of other panels can be incorporated, we expect AmpSeeker to prove useful to the research community beyond vector amplicon panels. We strongly encourage the vector and genomics communities to get involved in AmpSeeker's development, which we hope will foster community ownership and potentially drive wider adoption.

Centralised genome sequencing efforts such as the Ag1000g and Pf7, have advanced the field, generating large amounts of high-quality genomic data and revealing major evolutionary and public health insights [45,46]. Routine genomic surveillance, however, would be more effective if decentralised and conducted in-country, to enable the rapid turnaround from sample collection to results, to cultivate research ownership, and to facilitate effective communication with NMCPs and malaria stakeholders. The development of complete workflows for genomic surveillance, alongside ongoing efforts in building genomics capacity, will be pivotal to this decentralisation

process. However, major challenges remain - for example, the cost of obtaining reagents for sequencing (and indeed, other molecular biology applications), in sub-Saharan Africa (SSA) remains unacceptably high; Although we show that our protocol is a cost-effective route to high-throughput genotyping, the cost in SSA is likely to be two or three times higher. Existing genomic infrastructure must be supported and expanded further with sustained investment in both equipment and expertise [47]. Critical attention must be paid to establishing reliable supply chains for reagents and consumables, while developing regional centers of excellence that can provide training and technical support. Importantly, in the case of amplicon sequencing, analysis can be conducted on any standard bioinformatic laptop running Linux or Mac, eliminating the need for specialized computing infrastructure. Additionally, mechanisms for data sharing and standardization across countries need to be established to ensure that decentralized efforts contribute to a coherent understanding of vector populations across the continent.

A fundamental limitation of any targeted sequencing approach is the inherent bias towards known variants and genomic regions of interest. While our panel captures many known resistance mechanisms in *An. gambiae s.l*, novel mutations arising outside our targeted amplicon regions will inevitably be missed. This is particularly relevant given the remarkable adaptability of *Anopheles* populations and their demonstrated capacity to evolve new resistance mechanisms under insecticide pressure [13]. Efforts are underway to generate a web resource of selection signals in the Vector Observatory (https://anopheles-genomic-surveillance.github.io/selection-atlas/home-page.html), with the aim of building a database of haplotypes under putative selection, which we will use to update the Ag-vampIR panel. A second limitation of the approach is its limited ability to detect large structural variants which could mean that CNVs that confer metabolic resistance are missed, although they may be tagged by tagging SNPs. The modular nature of AmpSeeker and our panel design allows for updating the panel. This will be vital for ensuring long-term success of the platform. Using AnoPrimer, we can design primers for additional loci with ease whilst avoiding regions where genetic variants are at appreciable frequency in wild populations *An. gambiae s.l* [48]. Future versions of the panel could incorporate recently discovered markers of resistance [19,49] and tagging SNPs for major chromosomal inversions.

Conclusion

The development of complete workflows for genomic surveillance, alongside ongoing efforts in building genomics capacity, will be central to effective malaria vector control. Through our

approach, we provide an integrated solution for targeted genomic surveillance of insecticide resistance and species composition in the *An. gambiae* species complex. The platform's ability to track multiple resistance mechanisms and generate reproducible and interpretable results makes it a valuable tool for evidence-based decision making. As insecticide resistance continues to threaten malaria control efforts, accessible and standardized approaches to surveillance will become increasingly critical. By providing open-source tools and protocols, we hope to contribute to the development of sustainable, decentralized genomic surveillance capabilities across disease-endemic regions.

Materials and Methods

Sample collections and bioassay testing

Samples from four cohorts were collected and sequenced with the Ag-vampIR protocol (Supplementary Table 1), that spanned different species and geographies across sub-Saharan Africa. This included two laboratory colonies, Siaya and VK7, and two field cohorts from Obuasi, Ghana, and the eastern Upper River Region, The Gambia. Siaya (*An. gambiae)* were established in 2023 from Siaya county, Western Kenya, and VK7 (*An. coluzzii)* in 2014 from Valle de Kou, Burkina Faso. Obuasi samples were collected in 2017 with indoor resting aspiration by Prokopak aspirators, whilst the Gambia samples were sampled as part of the SANTE study with CDC miniature light traps [50].

For the association study, we exposed 3-5 day old female mosquitoes from the Siaya colony to new, unopened PermaNet 2.0 (deltamethrin) LLINs in extended WHO cone test bioassays[51]. The WHO cone test exposure time was extended from 3 to 40 minutes to induce close to 50% mortality. The details of each specimen involved in the study are specified in Supplementary Data 3.

Ag-vampIR primer and panel design

Details of the design of Ag-vampIR are found in Supplementary Text 2. In brief, primer design for all amplicons except Agam-79 (*Vgsc*-V402L), Agam-80 (I1527T) and *Dsx* was performed using an in-house modified version of mprimer[52]. The program modifications speed up the process of multiplex PCR design without any effect on the resultant primer design output itself. The V402L and I1527T amplicons were added later, designed with AnoPrimer, a primer design tool that considers the presence of genetic variation from the Ag1000g in primer binding sites[48]. For each input file, mprimer was run using the following parameters: primer min_gc20, primer opt_gc 50,

primer max_gc 80, primer min_tm 50, primeropt_tm 58, primer max_tm 68, product_min_size 190 bp, product_max_size 250 bp, penalty allowance 10.

To prepare samples for Illumina sequencing with sample barcoding (detailed below), we modified each primer by adding standard Illumina adapter sequences. The forward adapter sequence 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT- was added to the locus-specific forward primer, while the reverse adapter sequence 5′-TCGGCATTCCTGCTGAACCGCTCTTCCGATCT- was added to the locus-specific reverse primer. Additionally, to reduce primer dimer formation, we replaced the second-to-last base with a 2'-O-Methyl base, following the approach described by McKerrell et al. [53]. Primer sequences are detailed in Supplementary Data 1.

Laboratory protocol and sequencing

DNA was extracted from individual mosquitoes using Qiagen DNeasy extraction kits. For convenience, we did not quantify or normalise DNA concentrations prior to the Ag-vampIR protocol, however, this would be expected to result in more even coverage between sequencing plates. Full protocols are published in Jacob et al.,[22], with modifications below. In brief, an initial locus-specific PCR (PCR1) is run on each individual DNA sample, using a primer pool of all primers at 10µm. PCR1 is immediately followed by PCR2, in which PCR1 products are added to plates containing lyophilised i5 and i7 sample-specific barcodes and Illumina adaptors. In PCR2, these barcodes are added to the ends of the locus-specific amplicons, and permit hybridisation to the Illumina flow cell. After PCR2, each plate is pooled together, followed by an AMPure XP bead clean up and fragment analysis with an Agilent Tapestation or BioAnalyzer. qPCR is then used to determine the concentration of the final libraries, which are then diluted and pooled and prepared for sequencing according to Illumina protocols. We sequenced the libraries on an Illumina MiSeq with a MiSeq reagent kit V2, which provides 150 Bp reads and a total output yield of ~5 Gb.

Bioinformatic analysis

Full details of the AmpSeeker pipeline are designated in Supplementary Text 3. AmpSeeker is written in snakemake[20], according to the best practices. In brief, AmpSeeker converts Illumina BCL files to fastq with BCL-Convert (v4.3.6). Reads are then trimmed and assessed for quality with fastp [54], and the quality of index reads is assessed separately with fastqc[55]. Reads are then aligned to the reference genome with bwa mem[23], with indexing performed by samtools[56]. Coverage is assessed with mosdepth[57], and alignments visualised with an IGV notebook[58]. Variants are then called with bcftools mpileup[24] and annotated with snpEff [59].

For analysis of the resulting variant calls, AmpSeeker utilises a series of custom-built Jupyter notebooks written in Python. Notebooks are executed by snakemake using the python tool papermill, which allows parameters to be passed to the notebook prior to its execution. For genomic analysis, we use scikit-allel v1.3.7[60], whilst all visualisation is performed with Plotly [61] for its interactive plotting capabilities. Executed jupyter notebooks, containing data and interactive plots, are then combined with Jupyter Book[62] to generate a HTML web page containing all data and results.

Ancestry analysis and taxon assignment

We employed three complementary approaches to assign taxonomic status to our samples: Ancestry Informative Markers (AIMs), and two machine learning methods applied to principal components (PCA), and genotypes directly (XGBoost). Ancestry informative markers to distinguish *An. gambiae* from *An. coluzzii* were ascertained from Miles *et al.,*[13] which used a subset of SNPs from Neafsey *et al.,*[63], with samples from Mali. We used the AIMs to paint genotypes in our samples and to assign a putative taxon to samples. To check the reliability of our AIMs against *An. gambiae* from East Africa, we checked concordance of each AIM with any public samples from Uganda, Kenya, or Tanzania which were assigned as *An. gambiae* by the vector observatory pipelines[44,46]. For taxon assignment, we only use AIMs found on chromosome 3 and the X chromosome, to avoid known regions of adaptive introgression on chromosome 2 [30]. We called samples with an average AIM genotype of less than 0.5 as *An. gambiae,* between 0.5 and 1.5 uncertain, and above 1.5 called as *An. coluzzii*.

We built an extreme gradient boosting (XGBoost) model, using all Ag-vampIR SNPs as input, by training a model on 7363 curated, QC-controlled, whole-genome sequenced individuals from the Ag1000g, and used SNP data from five taxa in the complex: *An. gambiae, An. coluzzii, An. arabiensis, An. melas,* and the Pwani and Bissau molecular forms (Supplementary Table 2). We removed species with less than 10 samples (*An. quadrianulatus*, *An. merus* and *An. fontenillei*). The model was implemented using the XGBoost framework in Python, using 5-fold cross-validation. Feature selection was performed using a mean threshold criterion to identify informative AIMs, and hyperparameter optimization was conducted through a randomised cross-validation search. For the XGBoost model, we optimized parameters including learning rate (0.01-0.3), maximum tree depth (3-9), number of estimators (100-300), minimum child weight (1-5), and subsample ratio (0.6-1.0). The XGBoost model was highly accurate on Ag1000g data

(Supplementary Table 3). Model performance was evaluated using precision, recall, and F1-scores, with confusion matrices generated to assess classification accuracy across all species.

We also performed Principal Component Analysis (PCA) on combined data from our amplicon samples and 7,363 whole-genome sequenced reference samples from the Vector Observatory. After removing invariant sites, we conducted PCA on the remaining variable sites and used the first four principal components for classification. To automatically assign taxon status, we implemented a Support Vector Machine (SVM) classifier with a probability threshold of 0.8 to assign taxonomic status. Samples with classification probability below this threshold were marked as "unassigned." This approach allowed us to visualize the genetic relationships between our samples and reference populations in reduced dimensional space while providing statistical confidence for taxonomic assignments. For each sample, we assigned the taxonomic status that was predicted by at least two of the three methods. In cases where no consensus was reached (i.e., all three methods predicted different taxa), samples were labeled as "unassigned."

Population structure

For population structure analyses, we utilised variants present on any amplicon, not just those SNPs specifically targeted by the Ag-vampIR panel. Performing population structure analyses with resistance loci comes with the caveat that in many cases these loci will be non-neutral, given that they may be under selection, whilst ancestry informative markers are expected to segregate by species. After filtering out any SNPs that were invariant or missing in more than 5% of samples, 2389 SNPs remained. AmpSeeker both performs principal components analysis (PCA) and builds Neighbour-joining trees (NJT) generating interactive plots, coloured by user-selected metadata.

Kdr analysis and Diplotype clustering

Diplotype clustering of the Voltage-gated sodium channel was performed as in Nagi et al.,[19], and is as follows: We extracted diplotypes from the Vgsc gene region and performed complete-linkage hierarchical clustering analysis. Diplotypes, which can also be called multi-locus genotypes, represent sequences of diploid genotypes. We first converted diplotypes into allele count arrays, where each site is represented by an array of four elements corresponding to the possible alleles. For each site, we calculated the pairwise differences between these allele count arrays for all pairs of individuals. The clustering was performed using city-block (Manhattan) distance as the distance metric with complete linkage. We also

identified the non-synonymous variants present on each diplotype and plotted this information alongside sample heterozygosity across the region.

We included heterozygosity in the diplotype clustering plot, to permit easy identification of diplotype clusters homozygous for a selective sweep, as opposed to diplotype clusters which contain two separate haplotypes under selection. To calculate individual-level heterozygosity, we used the scikit-allel v1.3.7 implementation (allel.GenotypeArray.is_het). For a given diplotype region, we calculated the number of heterozygote calls and then divided by the number of called alleles to get a single value of heterozygosity for the region for each individual sample.

Genotype:phenotype associations

For the Siaya colony genotype:phenotype association tests, we loaded SNPs from all amplicons, removing invariant, highly missing (>20% of calls), or low minor allele frequency SNPs (<2%). We then split multi-allelic calls, and converted genotypes to the number of alternate calls. For each SNP, we ran a binomial generalised linear model (GLM) with logit link function using the number of alternate calls at the SNP of interest as a predictor variable and the phenotype (dead or alive in WHO cone test) the response variable. We then applied FDR correction to the resulting p-values with statsmodels[64].

Predicting bioassay mortality with an XGBoost machine learning model

We developed a machine learning approach using XGBoost (eXtreme Gradient Boosting). We used data from the Siaya colony mosquitoes that had been exposed to PermaNet 2.0 nets in extended WHO cone bioassays. For features, we utilized 210 quality-controlled SNPs that passed filtering for variance and missingness from the Ag-vampIR amplicons. We used 5-fold cross-validation to assess model performance, ensuring consistent evaluation across different data partitions.

We optimized hyperparameters through a cross-validation grid search (scikit-learn's GridSearchCV), exploring ranges for key parameters including learning rate (<u>0.01</u>, 0.1, 0.3), maximum tree depth (3, 7, <u>12</u>), number of estimators (<u>100</u>, 500), minimum child weight (<u>1</u>, 5), gamma (0, <u>0.3</u>), subsample ratio (0.7, <u>1</u>), reg_alpha (<u>0</u>, 1) and reg_lambda (<u>1</u>,4). The final best model parameters are underlined.

Data availability

The results book for this analysis is located at sanjaynagi.github.io/agvampir002-results/, which we encourage the reader to explore. Read data is deposited in the SRA (BioProject accession PRJNA1207724).

Author contributions

MJD, SCN, TN, JK, DW and ERL conceptualised and designed the study. SCN, ERL, CGJ, TP, AVH, CH, KR, AJ, ScG, NP, CA, SR and KR contributed to the development of the Ag-vampIR panel and laboratory protocol. SCN, TM, EL, ERL, AHK and AM contributed to the development of the AmpSeeker pipeline. FA, CY, ML and SS carried out DNA extractions and performed sequencing. MJD, VJS, SoG, TN, JK secured the funding to support the sequencing and analysis. HN, JE, AEY and JW provided samples. SCN led the data analysis with contributions from ERL, FA, SS and AM. SCN wrote the manuscript with contributions from all co-authors.

Acknowledgements

The authors would like to extend our deepest gratitude to the late Professor Dominic Kwiatkowski, whose pioneering contributions to the field of malaria genomic epidemiology paved the way for this research. We would also like to acknowledge the MalariaGEN laboratory teams at the University of Oxford and the Wellcome Sanger Institute for their contributions to the laboratory protocol, and to the SANTE trial team for mosquito collections. We thank Petra Korlevic for her illustrative logo for the Ag-vampIR panel.

This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID R01-AI116811 to M.J.D. and D.W.) and the Medical Research Council (MR/T001070/1 to M.J.D., D.W., and E.R.L., MR/P02520X/1 to M.J.D. and D.W., MR/W002159/1 to T.N.). The latter grant is a UK-funded award and is part of the EDCTP2 program supported by the European Union. M.J.D. was supported by a Royal Society Wolfson Fellowship (RSWF\FT \180003). SCN was funded by an MRC CASE studentship (MR/R015678/1). VJS is supported by the Bill and Melinda Gates Foundation (INV-068808). This study/project is funded by the NIHR [NIHR Global Health Research Group on Establishing Regional Hubs for Genomic Surveillance in West Africa (NIHR 134717)]. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.[14]

References

1. Hemingway J et al. Averting a malaria disaster: Will insecticide resistance derail malaria control? The Lancet. 2016. http://dx.doi.org/10.1016/S0140-6736(15)00417-1. https://doi.org/10.1016/S0140-6736(15)00417-1

2. WHO. World malaria report 2023. 2023 Nov 30 [accessed 2024 Dec 4]. https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023

3. Staedke SG et al. Effect of long-lasting insecticidal nets with and without piperonyl butoxide on malaria indicators in Uganda (LLINEUP): a pragmatic, cluster-randomised trial embedded in a national LLIN distribution campaign. The Lancet. 2020;395(10232):1292–1303.

4. Protopopoff N et al. Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions , separately and together , against malaria transmitted by pyrethroid-resistant mosquitoes : a cluster , randomised controlled , trial. The Lancet. 2018;391(10130):1577–1588.

5. Accrombessi M et al. Efficacy of pyriproxyfen-pyrethroid long-lasting insecticidal nets (LLINs) and chlorfenapyr-pyrethroid LLINs compared with pyrethroid-only LLINs for malaria control in Benin: a cluster-randomised, superiority trial. The Lancet. 2023 Jan 24.

6. Mosha JF et al. Effectiveness of long-lasting insecticidal nets with pyriproxyfen-pyrethroid, chlorfenapyr-pyrethroid, or piperonyl butoxide-pyrethroid versus pyrethroid only against malaria in Tanzania: final-year results of a four-arm, single-blind, cluster-randomised trial. The Lancet infectious diseases. 2024 [accessed 2024 Dec 17];24(1):87–97.

7. Oxborough RM, Chilito KCF, Tokponnon F, Messenger LA. Malaria vector control in sub-Saharan Africa: complex trade-offs to combat the growing threat of insecticide resistance. The lancet. Planetary health. 2024 [accessed 2024 Dec 4];8(10):e804–e812.

8. Wilkinson E et al. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Science (New York, N.Y.). 2021 [accessed 2025 Jan 1];374(6566):423–431.

9. Otieno JR et al. Global genomic surveillance of monkeypox virus. Nature medicine. 2024 Oct 23 [accessed 2025 Jan 1]:1–9.

10. Pruthi SS et al. Leveraging large-scale Mycobacterium tuberculosis whole genome sequence data to characterise drug-resistant mutations using machine learning and statistical approaches. Scientific reports. 2024 [accessed 2025 Jan 1];14(1):27091.

11. Bagi J et al. When a discriminating dose assay is not enough: measuring the intensity of insecticide resistance in malaria vectors. Malaria journal. 2011;14. http://dx.doi.org/10.1186/s12936-015-0721-4. https://doi.org/10.1186/s12936-015-0721-4

12. Praulins G, McDermott DP, Spiers A, Lees RS. Reviewing the WHO Tube Bioassay Methodology: Accurate Method Reporting and Numbers of Mosquitoes Are Key to Producing Robust Results. Insects. 2022;13(6). http://dx.doi.org/10.3390/insects13060544. https://doi.org/10.3390/insects13060544

13. Miles A et al. Genetic diversity of the African malaria vector Anopheles gambiae. Nature. 2017;552:96–100.

14. Lucas ER et al. Genome-wide association studies reveal novel loci associated with pyrethroid and organophosphate resistance in Anopheles gambiae and Anopheles coluzzii. Nature communications. 2023;14(1):4946.

15. Ingham V, Wagstaff S, Ranson H. Transcriptomic meta-signatures identified in Anopheles gambiae populations reveal previously undetected insecticide resistance mechanisms. Nature communications. 2018.

16. Grau-Bové X et al. Resistance to pirimiphos-methyl in West African Anopheles is spreading via duplication and introgression of the Ace1 locus. PLoS genetics. 2021;17(1):e1009253.

17. Lucas ER et al. A high throughput multi-locus insecticide resistance marker panel for tracking resistance emergence and spread in Anopheles gambiae. bioRxiv. 2019;(March 2020). http://dx.doi.org/10.1101/592279. https://doi.org/10.1101/592279

18. Clarkson CS et al. The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors Anopheles gambiae and Anopheles coluzzii. Molecular ecology. 2021;30(21):5303–5317.

19. Nagi SC et al. Parallel evolution in mosquito vectors-A duplicated esterase locus is associated with resistance to pirimiphos-methyl in Anopheles gambiae. Molecular biology and evolution. 2024 [accessed 2024 Oct 9];41(7):msae140.

20. Mölder F et al. Sustainable data analysis with Snakemake. F1000Research. 2021;10(33):33.

21. Kyrou K et al. A CRISPR–Cas9 gene drive targeting doublesex causes complete population suppression in caged Anopheles gambiae mosquitoes. Nature biotechnology. 2018 [accessed 2023 Jan 23];36(11):1062–1066.

22. Jacob CG et al. Genetic surveillance in the Greater Mekong subregion and South Asia to support malaria control and elimination. eLife. 2021 [accessed 2024 Oct 16];10.

23. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009;25(14):1754–1760.

24. Danecek P et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2). http://dx.doi.org/10.1093/gigascience/giab008. https://doi.org/10.1093/gigascience/giab008

25. Caputo B et al. Population genomic evidence of a putative “far-west” African cryptic taxon in the Anopheles gambiae complex. Communications biology. 2024;7(1):1115.

26. Zhong D et al. Extensive new Anopheles cryptic species involved in human malaria transmission in western Kenya. Scientific reports. 2020 [accessed 2025 Feb 11];10(1):16139.

27. Mwinyi SH et al. Genomic analysis reveals a new cryptic taxon within the Anopheles gambiae complex with a distinct insecticide resistance profile in the coast of east Africa. Molecular ecology. 2025 Apr 16 [accessed 2025 Apr 28]:e17762.

28. Scott JA, Brogdon WG, Collins FH. Identification of single specimens of the Anopheles gambiae complex by the polymerase chain reaction. The American journal of tropical medicine and hygiene. 1993;49(4):520–529.

29. Santolamazza F et al. Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms. Malaria journal. 2008;7(1):163.

30. Clarkson CS et al. Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. Nature communications. 2014 [accessed 2023 Jan 10];5(1):1–10.

31. Caputo B et al. Anopheles gambiae complex along The Gambia river, with particular reference to the molecular forms of An. gambiae s.s. Malaria journal. 2008 [accessed 2025 Apr 28];7(1):182.

32. Dong K et al. Molecular biology of insect sodium channels and pyrethroid resistance. Insect biochemistry and molecular biology. 2014 [accessed 2024 Dec 17];50:1–17.

33. Ranson H 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. Insect molecular biology. 2000;9(5):491–497.

34. Martinez-Torres D et al. Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector Anopheles gambiae s.s. Insect molecular biology. 1998;7(2):179–184.

35. Lynd A et al. LLIN Evaluation in Uganda Project (LLINEUP)-effects of a vector control trial on Plasmodium infection prevalence and genotypic markers of insecticide resistance in Anopheles vectors from 48 districts of Uganda. Scientific reports. 2024 [accessed 2024 Dec 2];14(1):14488.

36. Williams J et al. Characterisation of Anopheles strains used for laboratory screening of new vector control products. Parasites and Vectors. 2019;12(1):1–14.

37. Njoroge H, Van't Hof A, Oruni A, Pipini D. Identification of a rapidly-spreading triple mutant for high-level metabolic insecticide resistance in Anopheles gambiae provides a real-time molecular diagnostic for .... Molecular Ecology. 2022. https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16591

38. Weetman D et al. Candidate-gene based GWAS identifies reproducible DNA markers for metabolic pyrethroid resistance from standing genetic variation in East African Anopheles gambiae. Scientific reports. 2018;8(1):2920.

39. Nagi SC, Oruni A, Weetman D, Donnelly MJ. RNA-Seq-Pop: Exploiting the sequence in RNA-Seq - a Snakemake workflow reveals patterns of insecticide resistance in the malaria vector Anopheles gambiae. Molecular ecology resources. 2023 Jan 25.

40. Makunin A et al. A targeted amplicon sequencing panel to simultaneously identify mosquito species and Plasmodium presence across the entire Anopheles genus. Molecular ecology resources. 2022;22(1):28–44.

41. Lucas ER et al. Copy number variants underlie the major selective sweeps in insecticide resistance genes in Anopheles arabiensis from Tanzania. bioRxiv. 2024 Mar 13. http://biorxiv.org/lookup/doi/10.1101/2024.03.11.583874. https://doi.org/10.1101/2024.03.11.583874

42. Essandoh J, Yawson AE, Weetman D. Acetylcholinesterase (Ace-1) target site mutation 119S is strongly diagnostic of carbamate and organophosphate resistance in Anopheles gambiae s.s. and Anopheles coluzzii across southern Ghana. Malaria journal. 2013;12(1):1–10.

43. Lucas ER et al. Copy number variants underlie major selective sweeps in insecticide resistance genes in Anopheles arabiensis. PLoS biology. 2024;22(12):e3002898.

44. MalariaGEN. MalariaGEN Vector Observatory. Ag3.5: *Anopheles gambiae* data resource. 2023. https://malariagen.github.io/vector-data/ag3/ag3.5.html

45. MalariaGEN. Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples. Wellcome Open Research. 2023 [accessed 2023 Jan 18];8(22). http://dx.doi.org/10.12688/wellcomeopenres.18681.1. https://doi.org/10.12688/wellcomeopenres.18681.1

46. Ag1000G. Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii. Genome Res. 2020:1–14.

47. de Oliveira T. The third era of genomics is at risk of being dismantled across the global south. Science (New York, N.Y.). 2024 [accessed 2024 Oct 11];386(6718). https://www.science.org/doi/10.1126/science.adt4843. https://doi.org/10.1126/science.adt4843

48. Nagi SC, Ashraf F, Miles A, Donnelly MJ. AnoPrimer: Primer Design in malaria vectors informed by range-wide genomic variation. Wellcome open research. 2024 [accessed 2024 Oct 9];9:255.

49. Kengne-Ouafo JA et al. A single E205D allele of a key P450 CYP6P3 is driving metabolic pyrethroid resistance in the major African malaria vector Anopheles gambiae. bioRxiv. 2024 Jan 3.

50. Pinder M et al. To assess whether indoor residual spraying can provide additional protection against clinical malaria over current best practice of long-lasting insecticidal mosquito nets in The Gambia: study protocol for a two-armed cluster-randomised trial. Trials. 2011 [accessed 2024 Dec 2];12(1):147.

51. WHO. Guidelines for laboratory and field testing of long-lasting insecticidal nets. 2013 Feb 9 [accessed 2024 Dec 6]. https://www.who.int/publications/i/item/9789241505277

52. Shen Z et al. MPprimer: a program for reliable multiplex PCR primer design. BMC bioinformatics. 2010;11(1):143.

53. McKerrell T et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. Cell reports. 2015 [accessed 2024 Dec 2];10(8):1239–1245.

54. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England). 2018;34(17):i884–i890.

55. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

56. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics . 2009;25(16):2078–2079.

57. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics (Oxford, England). 2018;34(5):867–868.

58. Robinson JT et al. Integrative genomics viewer. Nature biotechnology. 2011;29(1):24–26.

59. 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. 2012;6(2):80–92.

60. Miles A, Harding NJ. Scikit-allel. 2017. http://dx.doi.org/10.5281/zenodo.3935797. https://doi.org/10.5281/zenodo.3935797

61. Inc. PT. Collaborative data science. 2015. https://plot.ly

62. Executable Books Community. Jupyter Book. Zenodo; 2020. https://zenodo.org/records/4539666. https://doi.org/10.5281/ZENODO.4539666

63. Neafsey DE et al. SNP genotyping defines complex gene-flow boundaries among African

malaria vector mosquitoes. Science. 2010;330(6003):514–517.

64. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: Proceedings of the Python in Science Conference. SciPy; 2010. p 92–96.