**­Parallel evolution and adaptive introgression in mosquito vectors – a story of insecticide resistance**

**Sanjay C Nagi1\*, Eric Lucas1, Alistair Miles2, David We****etman1, Martin J Donnelly1**

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

\*Corresponding author:

Liverpool Tropical School of Medicine, Pembroke Place, Liverpool, L3 5QA, UK. Telephone +441517053225, Email: [Sanjay.Nagi@lstmed.ac.uk](mailto:Sanjay.Nagi@lstmed.ac.uk)

**Keywords**: Adaptation, Insecticide resistance, Anopheles, Parallel evolution

**Introduction**

Parallel evolution refers to the repeated evolution of a phenotype or genotype in distinct populations of organisms.

The spread of organophosphate resistance in the common house mosquito, *Culex pipiens*, is a textbook example of contemporary evolution in response to anthropogenic pressures. In this species, independent gene duplications and transposable element insertions at the Est2 and Est3 locus, confer resistance to organophosphate insecticides, with at least 16 distinct haplotypes competing across the mosquitoes’ worldwide range (ref raymond 1996).

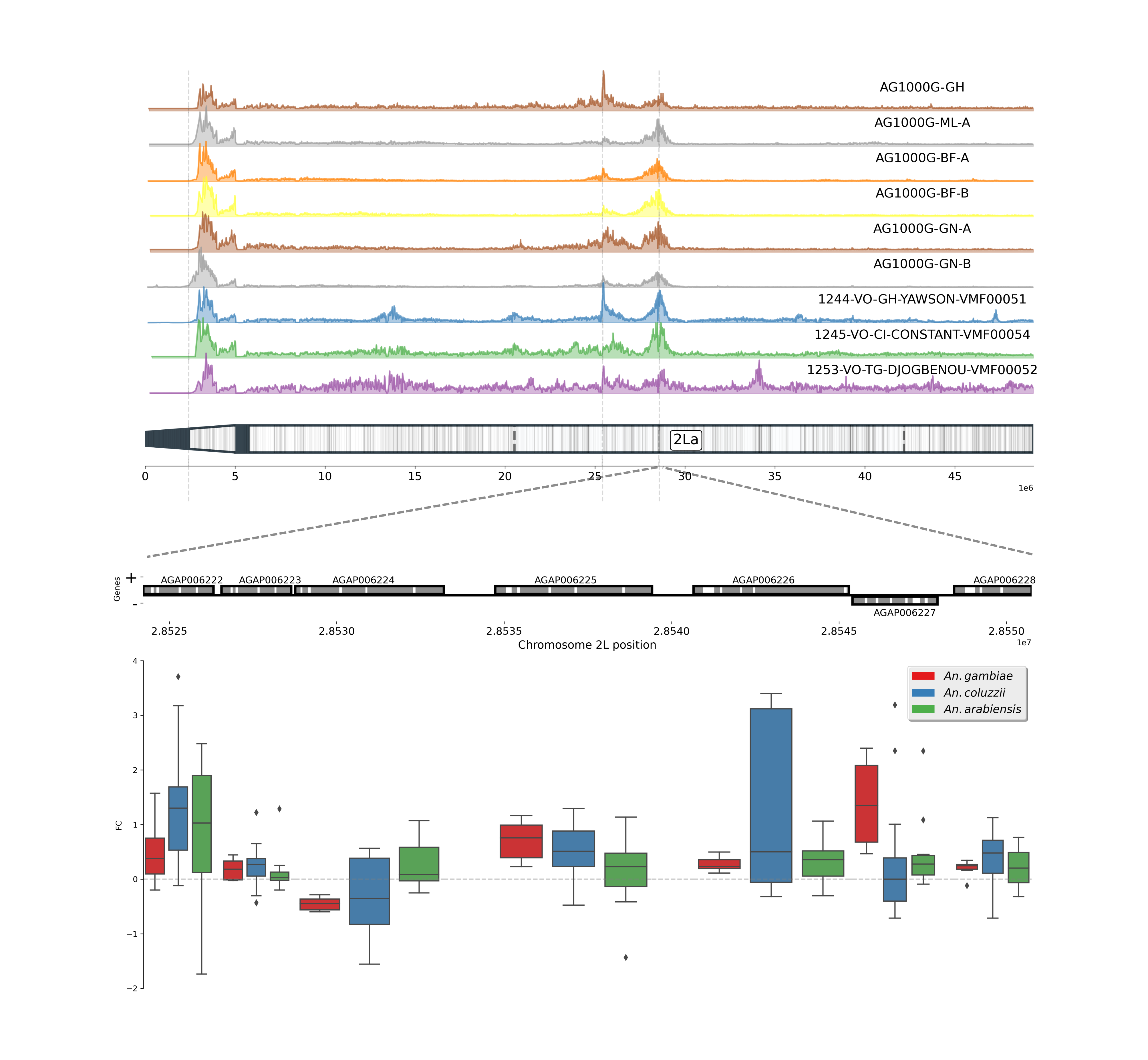
In the major malaria vector, *Anopheles gambiae*, resistance to organophosphates is associated primarily with the *Ace1* locus, the target of OP and carbamate insecticides. At this locus, a complex combination of heterozygous gene duplications and the *Ace1*-G119S non-synonymous mutation, confer varying levels of resistance whilst still

In the major malaria vector, *Anopheles gambiae*, we found evidence of large, repeated signals of selection at a locus orthologous to the *Culex pipiens Est2* and *Est3* genes. We integrate expression data from studies across sub-Saharan Africa, and perform haplotype analysis of this region, summarising copy number variation, and highlighting the *Coeae1f* and *Coeae2f* as major candidates for insecticide resistance in malaria vectors. Furthermore, we find that distinct, novel CNVs have arisen at this locus in both *An. gambiae* and *An. arabiensis*, as well as multiple selective sweeps in *An. coluzzii,* demonstrating the role of parallel evolution in insecticide resistance.

- How do we choose cohorts? When do we bring arabiensis in? Before CNV section, so then should It be plotted with other samples. But then there is the question of why no coluzzii?

- Need to re-look at coluzzii again in case there is introgression, but need to remember that UGT is probably driving any coluzzii sweep

**Results**



**A novel insecticide resistance locus?**

In the first phase of the ag1000g (ref), one of the largest novel signals of selection exists at a locus at approximately 28.5 megabases into the 2L chromosomal arm. The signals of selection are very large, with haplotype homozygosity extending beyond one megabase. Closer examination of the locus reveals a cluster of 7 putative detoxification genes, including two duplicated alpha-esterases, C*oeae1f*, and C*oeae2f*. These alpha-esterases sit in reverse orientation, 495 bases apart, and despite their recent shared ancestry, contain a varying number of exons, 7 in *Coeae1f*, and 4 in *Coeae2f*. Orthology searches revealed that these carboxylesterases were one-to-one orthologs with the *Est3* and *Est2* carboxylesterases of *Culex pipiens. Coeae1f* shares XX % amino acid similarity with *Est3*, and *Coeae2f* shares XX % amino acid similarity with Est4. Despite the earlier evidence in *Culex pipiens* for the alpha-esterases,UGTs are known to be involved in the phase 2 detoxification of pyrethroids (ref hanafy 2013), and aldehyde oxidases have been associated with resistance to neonicotinoids (hemongway 2000, shi 2009), so if a gene is causal for the sweep, which specific gene is not clear.

In order to gather further evidence for the locus’ role in insecticide resistance in *Anopheles gambiae,* we examined signals of selection from whole genome sequencing data in a larger cohort of *An. gambiae,* from throughout West Africa, collected between 2012 (CHECK!) and 2017 (see methods for further details of sample collections and sequencing). Figure 1A shows the H123 statistic calculated in XX windows across the genome for the 2L chromosomal arm. We preferred H123 to the other Garuds H statistics, as preliminary data showed the existence of multiple segregating selective sweeps in these populations, which H123 will have more power detect (ref). We find large H123 signals at this locus across all of these populations (Figure 1a), as well as at the VGSC and GABA receptor. The majority of the selection signals seem to peak over the two alpha-esterases, however, resolution at this genomic range is low, and genes in close proximity could instead be causal.

Apart from allelic variation, insecticide-detoxifying genes may need to be over-expressed in to produce a resistant phenotype. We therefore integrated gene expression data from previous microarray and RNA-Sequencing studies (ref), to provide additional evidence into the role of genes at the 28.5mb selection locus. Fig 1C shows boxplots of log2 fold changes, from transcriptomic experiments in which an insecticide resistant strain of *An. gambiae s.l* was compared to a susceptible strain.

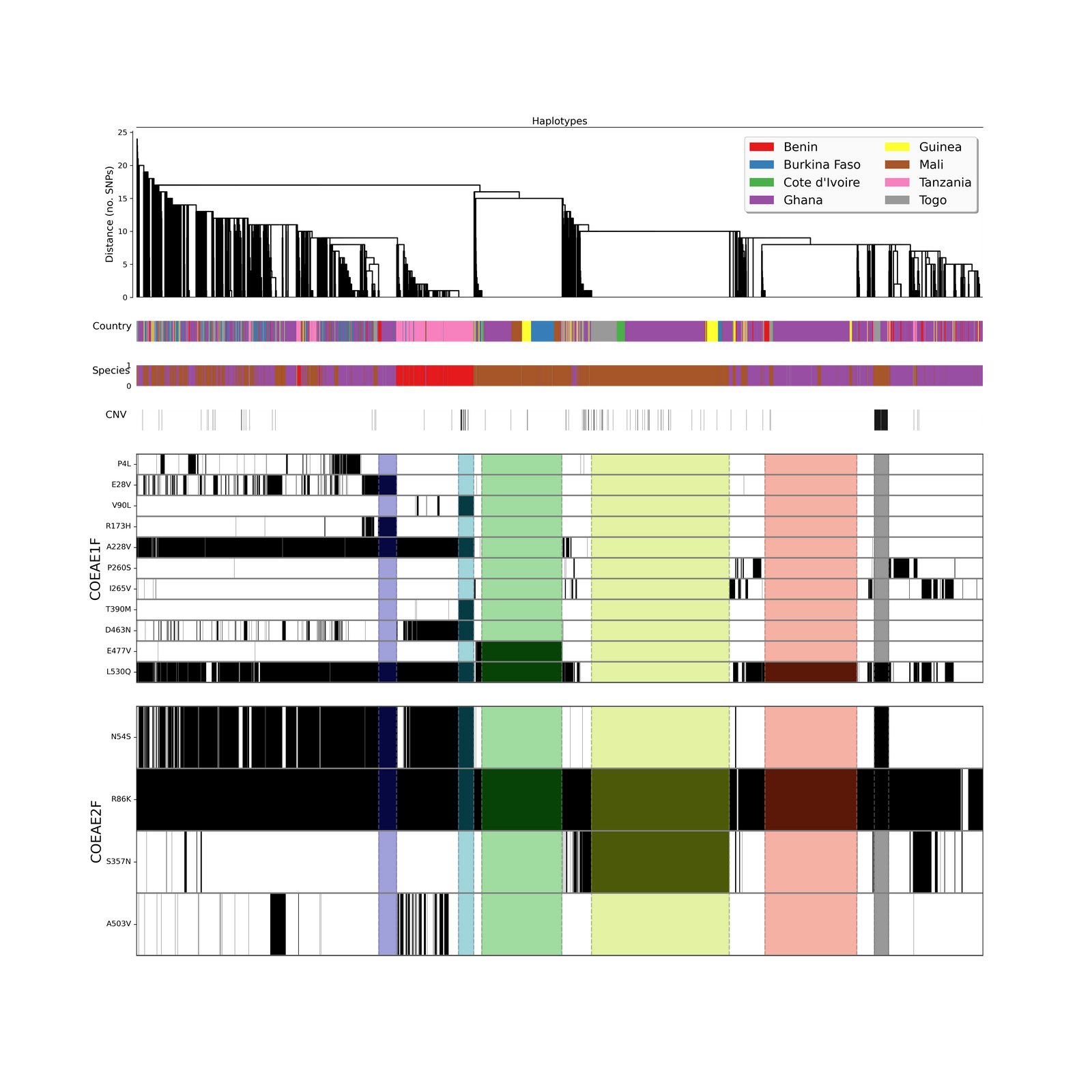
*Coeae1f* shows the highest median expression of all the genes in the detoxification cluster, with a fold change of XXX. In microarrays Bakaridjan in 2013 and 2014, the strain overexpressed *Coeae1f* 4.5 and 3.7 fold, respectively (ref). *Coeae1f* has been reported to be overexpressed in multiple populations across Africa, the two aforementioned An. *gambiae* from Burkina Faso, *coluzzii* from Cote D’Ivoire (3x), and Ugandan and Tanzanian An. *arabiensis* (Table 1). As it has high basal expression in RNA-Sequencing data (Supplementary appendix 1, ~7000 mean reads in this dataset), this five fold change may have larger consequences on absolute protein production. In recent RNA-Seq data from Cote D’ivoire (ref, DW), it is overexpressed 5.3 fold, in a resistant *An. gambiae* strain v Kisumu. The expression data is much less convincing for *Coeae2f.*

Four out of five aldehyde oxidases show little differential expression between resistant and susceptible colonies, apart from AGAP006226, and RNA-Sequencing data shows little to no expression in AGAP006223 and AGAP006224. In *An. coluzzii*, this AGAP006226 is heavily upregulated inmicroarrays from Bioko, Garre and Messa, which causes its mean expression to be high, however, its upregulation is not consistent across experiments. AGAP006222, a glucoronyl transferase, is consistently upregulated in populations of *An. coluzzii* across Africa, and is likely to be involved to some degree in phase 2 detoxification of insecticides.

**Haplotype** **clustering** **of** **the *Coeae1f* locus**

The integrated selection and expression data strongly implicate the locus in having an involvement in insecticide resistance. To further identify specific haplotype clusters that may be driven by one or more selective sweeps, we performed hierarchical clustering on haplotypes from the West African *An. gambiae* cohorts above. Figure 2 shows the results of this clustering, aligned with metadata from the individual of each haplotype, such as sample country, the 2La karyotype of that haplotype, and whether an individual harbors at least one CNV at this locus.

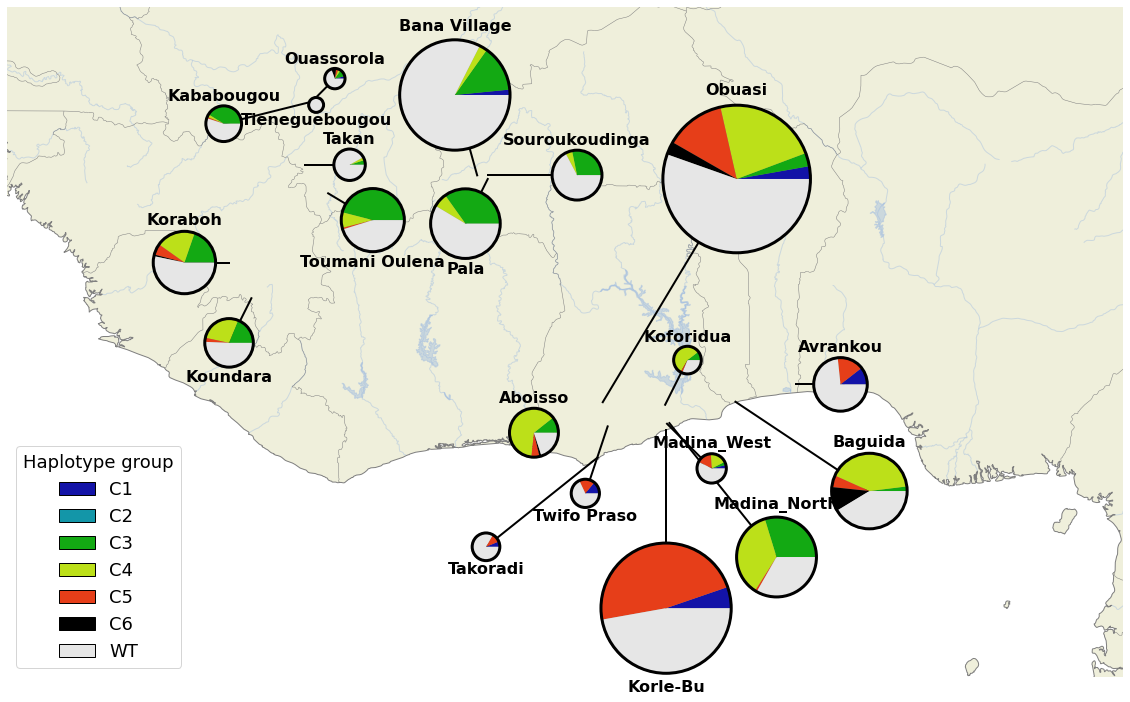
Loci under selection are likely to contain groups of haplotypes which are very similar to each other. In other words, a selective sweep has occurred, where one or more beneficial mutations is causing the haplotype to spread throughout the population. We extracted XXX haplotypes, and calculated the number of SNP differences in this XXX size window, using a cut-off of XXX SNPs to separate clusters. We then used a minimum threshold of 40 haplotypes in order to designate a cluster as swept, and other smaller clusters were designated as wild-type haplotypes.



The S3 sweep is only found in individuals that are homozygous for the 2La inversion polymorphism (karyotypes predicted by (Grau-bové *et al.*, 2019)⁠. This may restrict the spread of the sweep, as recombination is suppressed between 2La heterozygotes, and therefore the swept haplotype is likely to remain only on a 2La background, although limited recombination can occur as previously evidenced in *Anopheles* (Grau-bové *et al.*, 2019)⁠. The swept haplotype is towards the centre of the inversion (8.02mb from the beginning of the breakpoint, and 13.62mb from the end), which makes inter-karyotypic introgression more likely than if close to end points (Cheng *et al.*, 2012)⁠. Surprisingly, the hierarchical clustering did not seem to pull out two clades corresponding to 2l+a and 2la haplotypes. This would be expected given the approximate time since the inversion occurred (XXX years, ref), and suggests some inter-karyotypic introgression may have occurred in the intervening period.

We examined patterns of haplotype clustering down and upstream from the *Coeae1f* region, in order to ensure the patterns we see are indeed resulting from selection, and not demography and population structure.

**The spatial distribution of swept haplotype clusters**



**Amino acid variation**

As a first step to identify potential causal mutations in each selective sweep, we calculated allele frequencies at non-synonymous sites in the region under selection. We first phased multiallelic non-synonymous sites in the onto the existing biallelic haplotype scaffold of the above *An. gambiae* cohorts, using the program *mvncall v1.0.* We only phase non-synonymous multiallelic variants which are present at a frequency above 5%. In total, we discover XXX biallelic non-synonymous, and XXX multiallelic non-synonymous mutations.

A full table of non-synonymous variation in the region of interest can be found in the supplementary data XXX. Overall, the highest frequency variants were XXX and XXX. The most widespread geographically was XXX, found in etc etc.

* LD plots, overall and in each sweep in panel
* Median joining network????

**Haplotype association tests**

**Copy number variation**

* Map of CNVs and their frequencies?



Coeae1f is regulated by the transcription factor maf-s

COEAE1F is controlled by the MAF-S transcription factor, demonstrated by a large downregulation (how large?? ) of the gene when during a MAF-S knockout (Ingham et al., 2017). COEAE2F was also downregulated (by how much?????).

We extracted motifs from the JASPAR database, to look for CnC binding sites in the region under selection. We found XXXXXX

A distinct selective sweep at the KEAP1 locus is spreading in the same populations (ref/suppl). If a KEAP1 sweep served to increase expression of Maf-S controlled genes,

*In silico* protein models suggest shared inter-species function

We used alphafold (openfold?) to predict protein structures for the two *An. gambiae* carboxylesterases (*Coeae1f* and *Coeae2f*) and *Culex pipiens* (*Est2* and *Est3*), and examined mutations close to the Ser-His-Glu catalytic triad and active site.

We then performed molecular docking with XXX different insecticides from XX different classes, examining the resulting models for productive poses in which metabolism of the insecticide could occur.

**Discussion**

- There exists a CNV allele in Cx.pipiens which also contains half of the AO 006226

**Limitations**

- missing small indels, Svs, insertions