Copy number variants underlie the major selective sweeps in insecticide resistance genes in *Anopheles arabiensis* from Tanzania.

Eric R. Lucas, Sanjay C. Nagi, Bilali Kabula, Arjen E. Van't Hof, Emily J. Rippon, Dimitra Pipini, Nicholas J. Harding, Naomi A. Dyer, Chris S. Clarkson, Alistair Miles, Martin J. Donnelly, David Weetman

Electronic Supplementary Material

Supplementary figures and tables

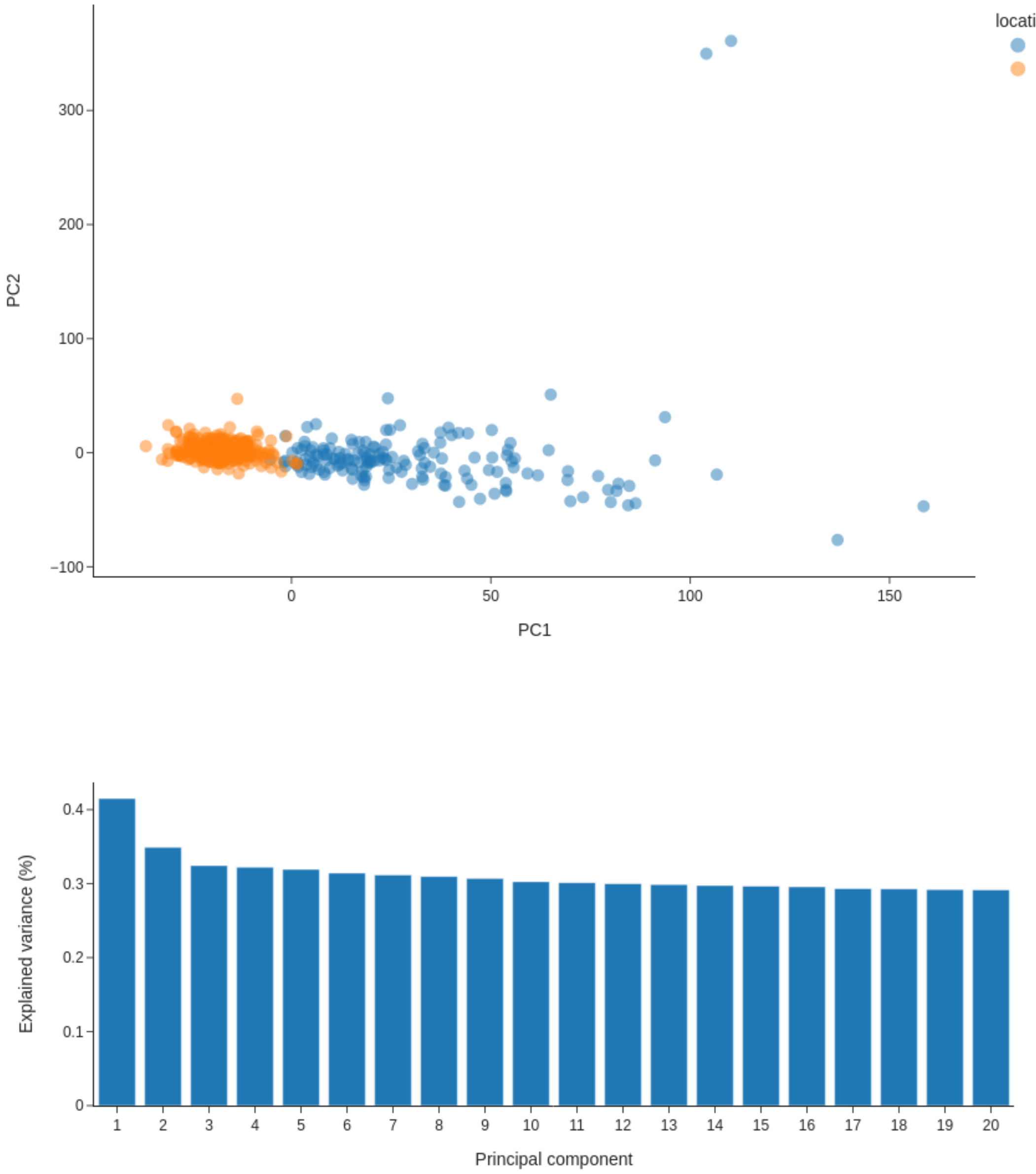


Fig. S1: PCA (using quality-filtered biallelic SNPs from genomic region 3L:15,000,000-41,000,000, euchromatic and free of chromosomal inverstions). Top panel shows clustering of samples by region. Bottom panel show variance explained by the first 10 PCs, indicating that PCs 3 onwards explain similar levels of variance and are thus likely only capturing noise.

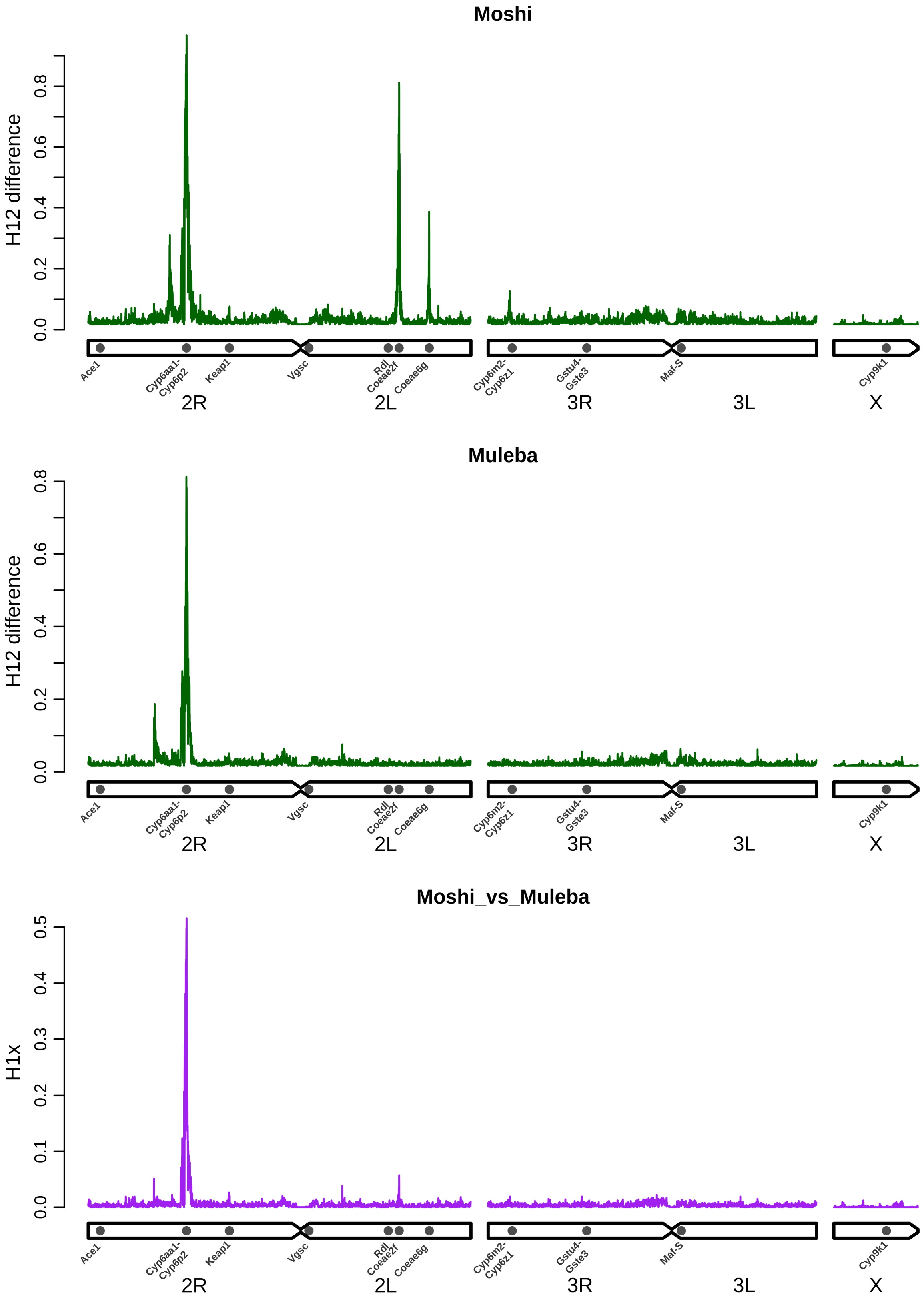
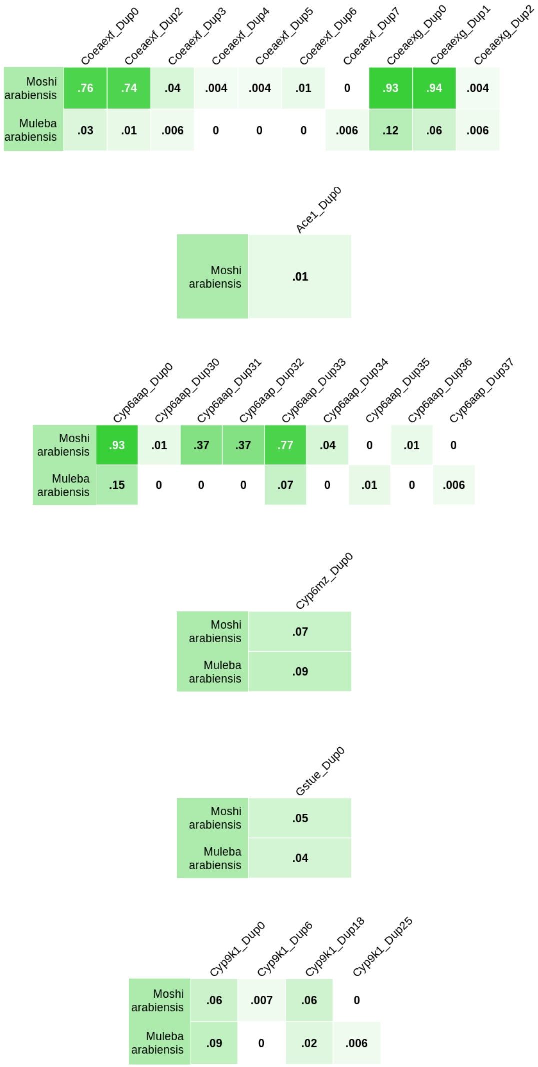


Fig. S2: Selection scans showing genome-wide H12 signal in Moshi (top) and Muleba (middle), as well as shared signals of selection (H1x) between the two sites (bottom).



	Coeaexf_Dup0	Coeaexf_Dup2	Coeaexf_Dup3	Coeaexf_Dup4	Coeaexf_Dup5	Coeaexf_Dup6	Coeaexf_Dup7	Coeaexg_Dup0	Coeaexg_Dup1	Coeaexg_Dup2
Moshi arabiensis	.76	.74	.04	.004	.004	.01	0	.93	.94	.004
Muleba arabiensis	.03	.01	.006	0	0	0	.006	.12	.06	.006

	Ace1_Dup0
Moshi arabiensis	.01

	Cyp6aap_Dup0	Cyp6aap_Dup30	Cyp6aap_Dup31	Cyp6aap_Dup32	Cyp6aap_Dup33	Cyp6aap_Dup34	Cyp6aap_Dup35	Cyp6aap_Dup36	Cyp6aap_Dup37
Moshi arabiensis	.93	.01	.37	.37	.77	.04	0	.01	0
Muleba arabiensis	.15	0	0	0	.07	0	.01	0	.006

	Cyp6mz_Dup0
Moshi arabiensis	.07
Muleba arabiensis	.09

	Gstue_Dup0
Moshi arabiensis	.05
Muleba arabiensis	.04

	Cyp9k1_Dup0	Cyp9k1_Dup6	Cyp9k1_Dup18	Cyp9k1_Dup25
Moshi arabiensis	.06	.007	.06	0
Muleba arabiensis	.09	0	.02	.006

Fig. S3: Frequency (proportion of samples carrying at least one copy) of known CNV alleles detected using diagnostic reads around the *Coeaexf* cluster, the *Coeaexg* cluster (combined into a single subfigure), *Ace1*, the *Cyp6aa / Cyp6p* cluster, the *Cyp6m / Cyp6z* cluster, the *Gste* cluster and *Cyp9k1*. Only CNV alleles with frequency $>$ 0% are shown. Cell darkness provided as a visual aid for the magnitude of the value in each cell. The genomic coordinates of each CNV allele can be found in Supplementary Data S@. In each cluster, the "Dup0" column indicates the presence of increased copy number in any of the genes in the cluster. Where this is larger than the sum of known alleles, it suggests the presence of uncharacterised CNV alleles. "Del" alleles in Ace1 represent secondary deletions within the Ace1-Dup1 CNV.

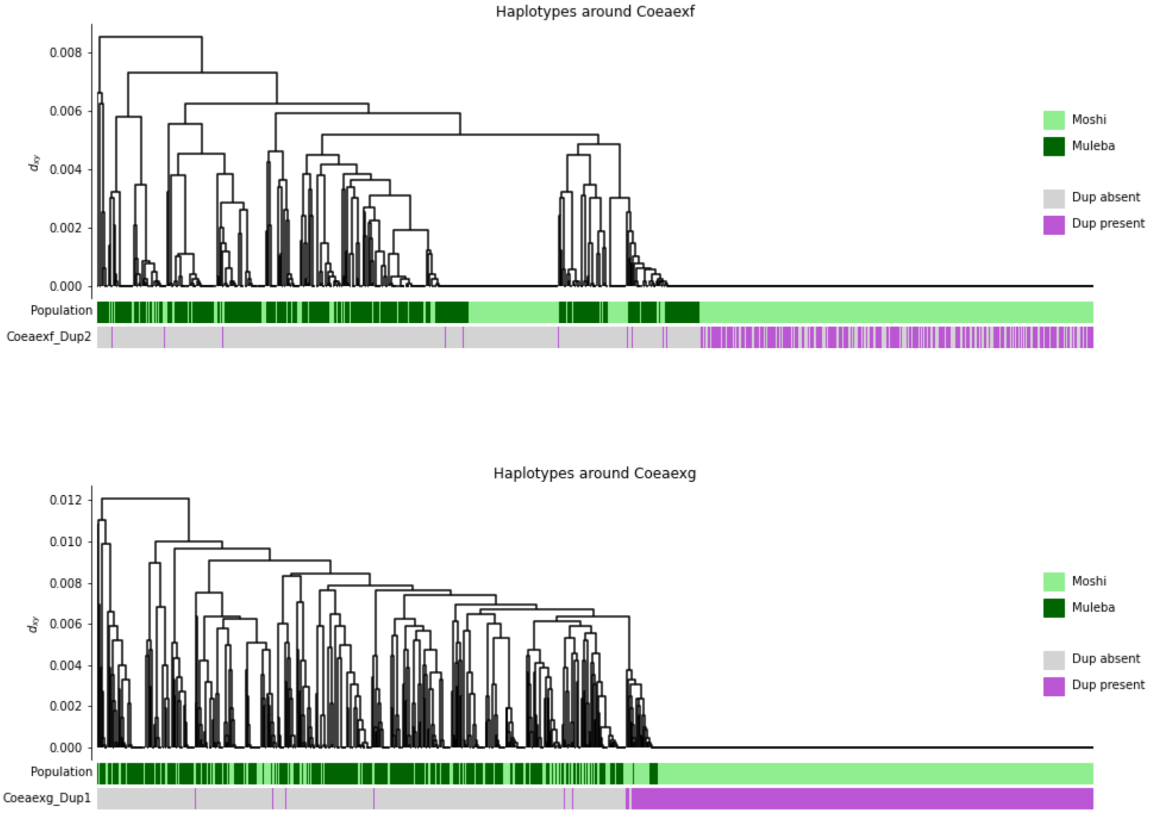
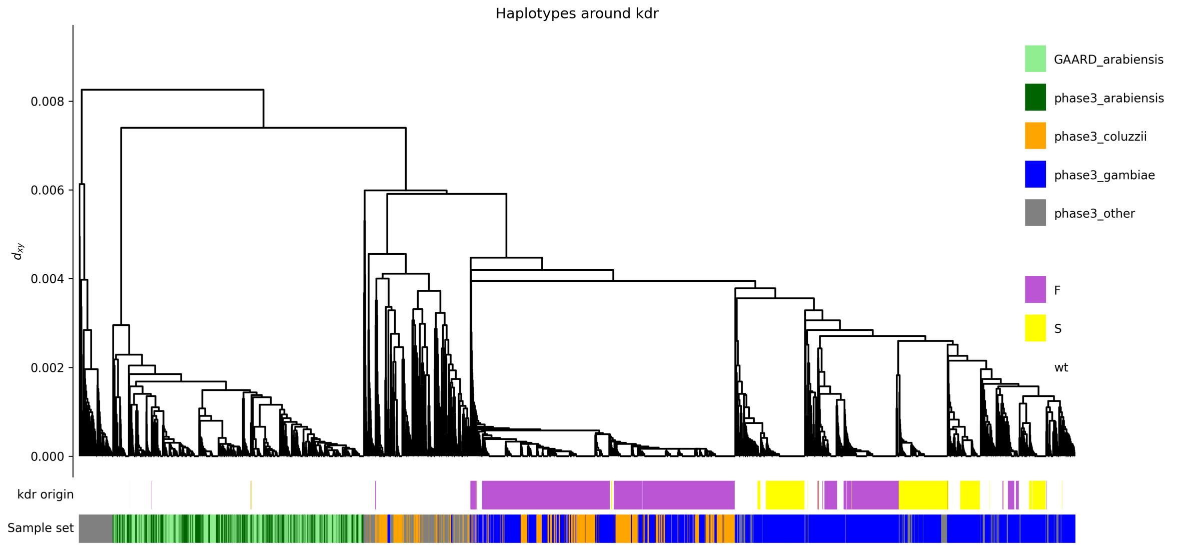
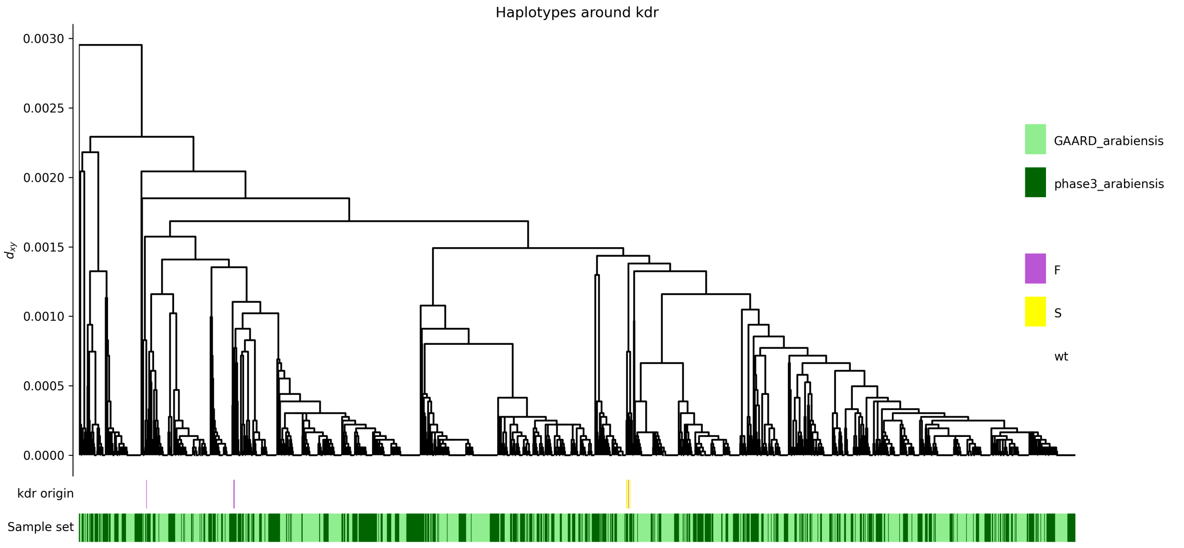
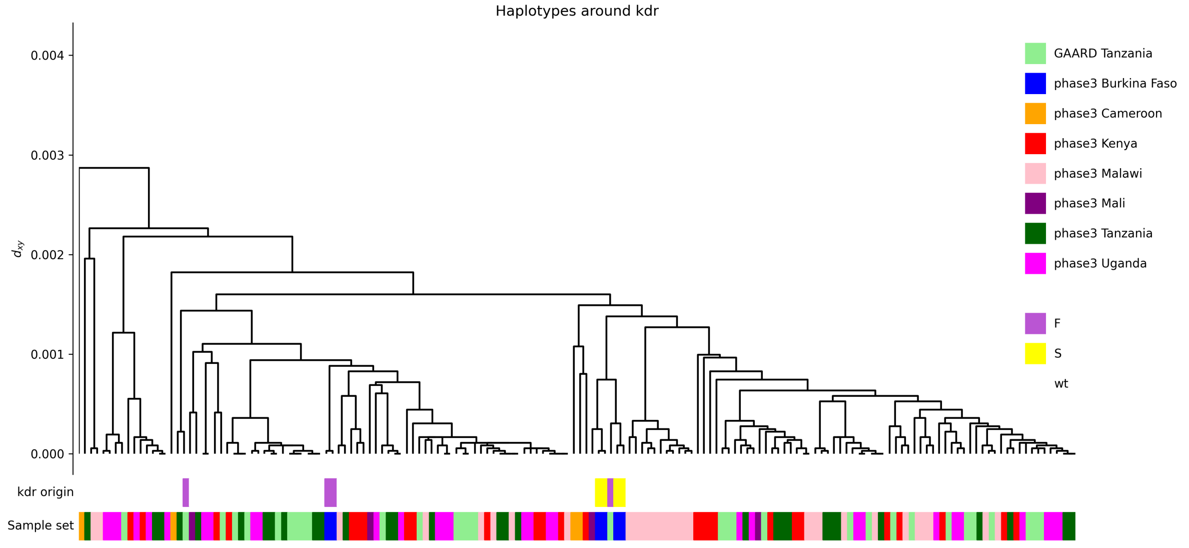


Fig. S4: Haplotype clustering using 500 SNPs around the *Coeaexf* (top) and *Coeaexg* (bottom) genomic regions. Haplotypes bearing the CNV allele *Coeaexf*_Dup2 represented a subset of haplotypes from the large swept cluster in *Coeaexf*, whereas haplotypes with the CNV *Coeaexg*_Dup1 were closely associated with the large swept cluster in *Coeaexg*. The CNV alleles were phased by identifying SNPs that were highly (*Coeaexf*_Dup2) or perfectly (*Coeaexg*_Dup1) correlated with their presence / absence level. Full workings to reproduce this analysis can be found at https://github.com/vigg-lstm/GAARD_east/blob/main/CNV_analysis/sweeps.


Haplotypes around kdr
d_{xy}
0.008
0.006
0.004
0.002
0.000
kdr origin
Sample set
GAARD_arabiensis
phase3_arabiensis
phase3_coluzzii
phase3_gambiae
phase3_other
F
S
wt


Haplotypes around kdr
d_{xy}
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
kdr origin
Sample set
GAARD_arabiensis
phase3_arabiensis
F
S
wt


Haplotypes around kdr
d_{xy}
0.004
0.003
0.002
0.001
0.000
kdr origin
Sample set
GAARD Tanzania
phase3 Burkina Faso
phase3 Cameroon
phase3 Kenya
phase3 Malawi
phase3 Mali
phase3 Tanzania
phase3 Uganda
F
S
wt

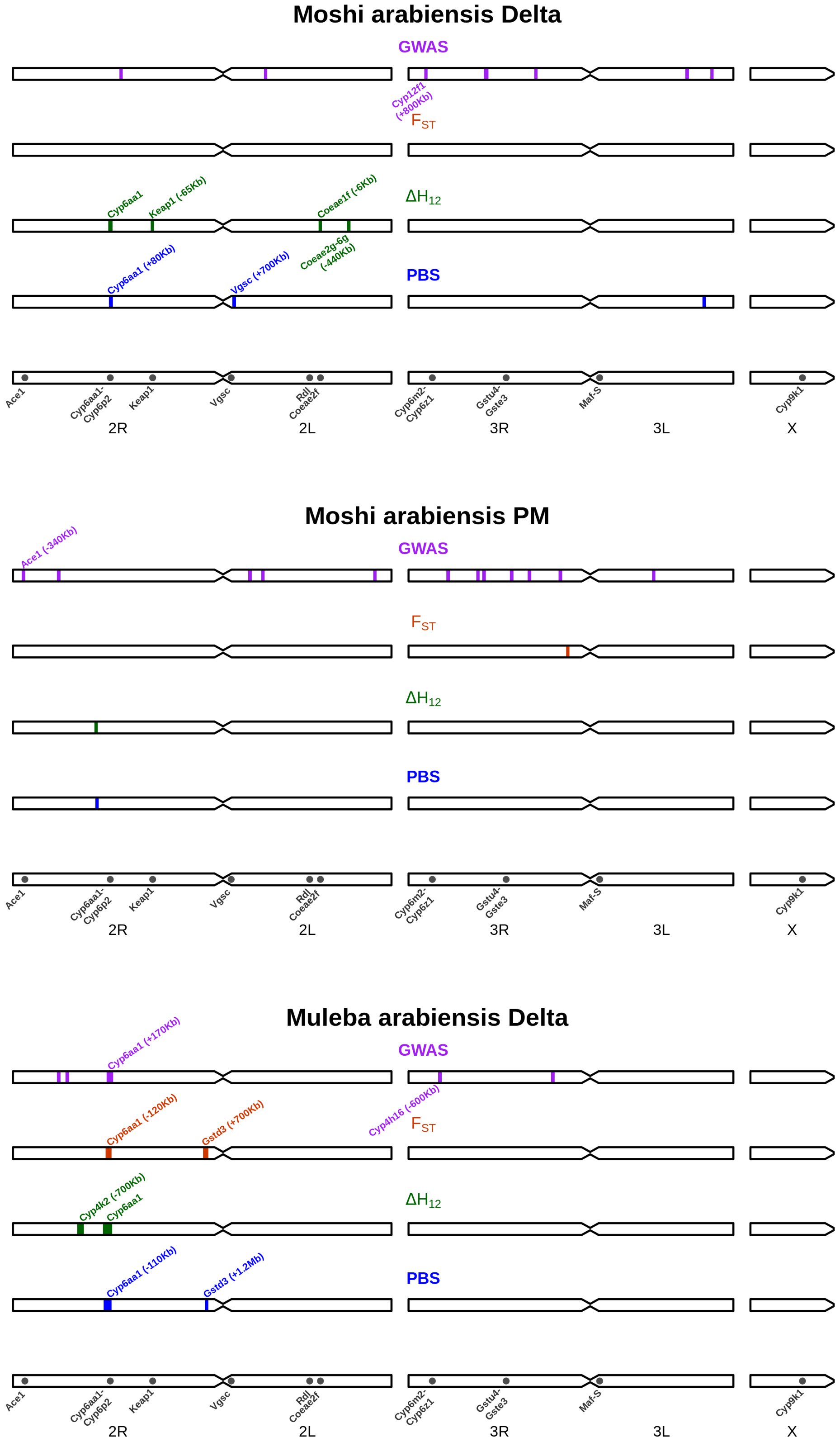


Fig. S6: Genomic regions implicated in insecticide resistance by each of our four approaches. For the global GWAS method, these are 100,000 bp windows which contained at least 10 of the top 1000 significant SNPs. For F_{ST}, these are significant peaks which contained at least one haplotype significantly positively associated with resistance (Supplementary Data S2). For ΔH_{12} and PBS, these are significant positive peaks (ie: indicating stronger signals of selection in resistant compared to susceptible samples). Regions are annotated with genes discussed in the manuscript as possibly causing the signal. Genomic distances in brackets indicate the distance of the peak either to the left (-) or right (+) of the gene in question.