Sanjay C Nagi

**West African 2L sweep 28.5mb**

**Background**

H12 and iHS scans revealed a [large selective sweep](https://malariagen.github.io/agam-selection-atlas/0.1-alpha3/signal/H12/BFS/2/3/) within the 2LA inversion at ~28.5 mbp, found in *Anopheles* *gambiae* populations of Burkina Faso and Guinea, with hints of a sweep in the Ghanaian population of phase 2.

A number of detox genes lie underneath the swept region. Directly under the focal point of the sweep lie two paralogous alpha esterases, COEAE1F (AGAP006227), and COEAE2F (AGAP006228). Just before COEAE1F, there are five aldehyde oxidases (AGAP0006220/ 221/224/225/226) which are broken up by two glucosyl transferases (UGTs, AGAP006222, 223).

Based on the selection scans alone, the primary candidates would be the two esterase genes. However, UGTs are known to be involved in Phase 2 detoxification of pyrethroids (refs), and aldehyde oxidases have been associated with resistance to neonicotinoids (refs).

**Literature review**

Esterases are typically thought to be involved in organophosphate resistance…blah blah

The majority of genes with esterase activity can be divided into eight subfamilies: alpha-esterases, beta-esterases, juvenile hormone esterases, acetylcholinesterases, gliotactins, neurotactins, neuroligins and glucactin type.

All pyrethroids contain ester bonds, and are therefore subject to ester hydrolysis. Both es

Both esterases of interest here are alpha-esterases, which are generally believed to be more likely to metabolise pyrethroids than beta-carboxylesterases (ref?).

Type 2 pyrethroids are generally more potent, in part because their greater bulk makes them less susceptible by hydrolytic detox by carboxylesterases. The same is also true for cis and trans isomers – cis isomers are more potent, partly due to being less likely to be metabolised by esterases.

COEAE1F is controlled by maf-s, downregulated in vickys array. I should look for Maf-S ARE elements.

keap1 is swept in burkina faso gambiae

AOs

UGTs

**Haplotype clustering**

The sweep centres approximately on top of the two alpha esterases – approximately 28,548,000 bases into chromosome 2L. By performing haplotype clustering on the region, I identified a large cluster of 46 haplotypes which differ by less than 10 SNPs (there are 5118 segregating sites in this 50kb region, in these populations).

Nicks words in SAP2– “ this is a firm proxy for identity by descent in an organism with pi=0.01”. And this is a much much larger sweep!

The sweep is only found in individuals that are homozygous for the 2LA inversion polymorphism (karyotypes predicted by Grau-Bove *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 Grau-Bove *et al.*, 2019 (although limited recombination can occur and as evidenced in Anopheles over longer time periods (Grau-Bove *et al.*, 2019). If of large enough effect size, it is possible that the sweep could drive the 2LA inversion to higher frequency.

The shared haplotypes are found in Burkina Faso and Guinea, indicating that the sweep had spread over considerable geographic distance by 2012. That region in West Africa is relatively genetically uniform (there are no significant barriers to gene flow at least) (ref donnelly convo). Therefore, I hypothesise that the swept haplotype is likely to be found in Ghana and Cote D’ivoire, and certainly sub-Saharan Mali, which lies inbetween the Guinean and Burkinabe populations.

A CNV exists within the swept region (before detox genes), at low-moderate allele frequencies. This CNV covers GPRMTH2 (Methuselah-like GPCR), but (I have checked) is not associated with the swept haplotype – in fact it is in negative linkage.

**Selection scans**

We used SNP variation data from the Anopheles 1000 genomes project (Phase 2-AR1), retrieving phased genotype calls, and associated variant metadata.

We used 184 haplotypes calculated Garud’s H statistics across the genome of mosquitoes from

**Relate & Clues**

We estimated genome-wide geneaologies of *Anopheles gambiae* mosquitoes from Burkina Faso from phase 1 of the Ag1000g, using the program Relate (Speidal et al., 2019). We extracted the subtrees covering the E477V mutation in COEAE1F (2L:28545767), sampling branch lengths with mcmc and extracting coalescent times. We then used the likelihood-based approach implemented in CLUES (Stern et al., 2019) to estimate selection coefficients of the derived allele and allele frequency trajectories. CLUES samples from the ancestral recombination graph (ARG)

**Expression data**

Expression data for the aforementioned genes from Ingham et al (2018) as well as recent RNA-Seq experiments is reported in the Appendix.

COEAE1F has been reported to be overexpressed in multiple populations across Africa, in An. *gambiae* from Burkina Faso (2x), *coluzzii* from Cote D’Ivoire (3x), and Ugandan and Tanzanian An. *arabiensis* (Table 1). In recent RNA-Seq data from Cote D’ivoire (DW), it is overexpressed five-fold Resistant V susceptible strain. As it has high basal expression (~7000 mean reads in this dataset), this five fold change is particularly large.

The expression data is much less convincing for COEAE2F, although this gene is expressed slightly higher in susceptible populations.

Four out of five aldehyde oxidases show little differential expression between resistant and susceptible colonies, apart from AGAP006226. In *An. coluzzii*, this gene is heavily upregulated in microarrays from Bioko, Garre and Messa, which causes its average expression to be high, however, it is not consistently significantly upregulated across experiments.

AGAP006222, a glucoronyl transferase, is consistently upregulated across in populations across Africa, and is likely to be involved to some degree in phase 2 detoxification of xenobiotics, including insecticides. As this gene is part of a suite of phase 2 enzymes, we may not focus on it, although it is possible that co-amplification of a metabolising esterase and a UGT could form an insecticide detox pathway.

**Phylogenetic analysis of haplotypes**

**Tracking** **the** **sweep**

* Screen temporal tengrela samples (Burkina Faso, gambiae)
* Screen Ghana and Cote D’ivoire samples, including some which are phenotyped

**Bakaridjan**

* Screen bakaridjan for sweep
* qPCR for swept genes
* use an esterase inhibitor and phenotype?
* Phenotype against PBO nets for extra impact

**Sf9 expression & metabolism**

* express and do metabolism assays
* what classes do it/they bind
* Are the esterases inhibited by PBO?

**In silico modelling**

* modelling (itasser)
* docking with insecticides , PBO

**Methods**

**Selection scans**

H12, Nsl

**Haplotype clustering**

The sweep centres approximately on top of the two alpha esterases – approximately 28,548,000 bases into chromosome 2L. I first extracted haplotypes from the region 25kbp down and upstream from this locus, and constructed a distance matrix between haplotypes using the ‘hamming’ distance metric, and multiplying by the number of variants to conver hamming distance into number of SNPs, for interpretability.

We then performed hierarchical clustering to cluster the haplotypes, using single linkage, and cut the tree at 10 SNPs, identifying a large swept cluster.

**Tag SNP panel**

**References**

**Appendix**

**swept gene exp.**