Mechanisms of transcriptional regulation in *Anopheles gambiae* revealed by allele specific expression

Authors

Naomi A. Dyer[1], Eric R. Lucas[1], Sanjay C. Nagi[1], Daniel P. McDermott[1], Jon H. Brenas[2], Alistair Miles[2], Chris S. Clarkson[2], Henry D. Mawejje[3], Craig S. Wilding[4], Marc S. Halfon[5], Hasiba Asma[5], Eva Heinz[1,6], Martin J. Donnelly[1]

Short title

Allele specific expression in *Anopheles gambiae*

Affiliations

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

2 Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

3 Infectious Diseases Research Collaboration (IDRC), Plot 2C Nakasero Hill Road, P.O.Box 7475, Kampala, Uganda

4 School of Biological and Environmental Sciences, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK

5 Department of Biochemistry, Jacobs School of Medicine & Biomedical Sciences, University at Buffalo-State University of New York, 955 Main Street, Buffalo, New York 14203, USA

6 Department of Clinical Sciences, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, UK

Keywords

Allele, cis-regulation, transcript, insecticide, resistance

Correspondence

Naomi.Dyer@lstmed.ac.uk

Abstract

Malaria control relies on insecticides targeting the mosquito vector, but this is increasingly compromised by insecticide resistance, which can be achieved by elevated expression of detoxifying enzymes that metabolize the insecticide. In diploid organisms, gene expression is regulated both in *cis*, by regulatory sequences on the same chromosome, and by *trans* acting factors, affecting both alleles equally. Differing levels of transcription can be caused by mutations in *cis*-regulatory modules (CRM), but few of these have been identified in mosquitoes. We crossed bendiocarb resistant and susceptible *Anopheles gambiae* strains to identify *cis*-regulated genes that might be responsible for the resistant phenotype using RNAseq, and *cis*-regulatory module sequences controlling gene expression in insecticide resistance relevant tissues were predicted using machine learning. We found 115 genes showing allele specific expression in hybrids of insecticide susceptible and resistant strains, suggesting *cis* regulation is an important mechanism of gene expression regulation in

Anopheles gambiae. The genes showing allele specific expression included a higher proportion of *Anopheles* specific genes on average younger than genes those with balanced allelic expression.

Author Summary

The evolution of insecticide resistance, including resistance that is due to changes in the expression levels of certain resistance associated genes is threatening progress in malaria control. We investigated how the expression of genes in the malaria vector *Anopheles gambiae* is controlled, by implementing a method for the first time in this species. Each mosquito inherits a set of chromosomes from both parents, so has a maternal and paternal copy of most genes. When a gene is expressed, the DNA encoding that gene is transcribed into messenger RNA. This process is controlled by the cellular environment and by other DNA sequences on the same chromosome as each gene. We crossed mosquitoes from insecticide resistant and susceptible strains to equalize the cellular environment and then measured the levels of messenger RNA from both gene copies. 115 genes showed consistently different messenger RNA levels between gene copies in most crosses, suggesting these genes are regulated by factors on the same chromosome. There were relatively more Anopheles specific genes with imbalanced expression. Using machine learning we identified DNA sequences that may be responsible for controlling gene expression in mosquito tissues; several of these sequences were close to genes with imbalanced expression.

65 # Introduction

66 Malaria prevalence in Sub-Saharan Africa has reduced by 50% since 2000, primarily due to
67 insecticide-based control of mosquito vectors [1]. Recently, progress has stagnated [2],
68 partly due to increasing levels of resistance against insecticides in mosquito populations [3].
69 *Anopheles gambiae* is one of the dominant malaria vectors in Sub-Saharan Africa, the
70 primary vector across most of Uganda [4] and the vector for which the largest resource of
71 genome data is available, with 7275 genomes sequenced [5-8]. A common cause of
72 insecticide resistance is increased degradation of insecticides (termed metabolic resistance)
73 [9] with overexpression of insecticide-metabolizing P450s repeatedly implicated [10-12].
74 This can be caused by mutations in *cis*-regulatory regions regulating expression of metabolic
75 resistance genes. In diploid organisms, such mutations only affect expression of the allele of
76 the gene located on the same chromosome. Although some *trans* factors involved in
77 metabolic resistance gene regulation in *Anopheles* are known [13, 14], few studies have
78 identified genetic variation causing metabolic resistance [15-18]. The multiallelic nature of
79 metabolic insecticide resistance which can involve different mutations affecting the same
80 gene in different populations, as well as the involvement of multiple genes, makes marker
81 identification challenging as it limits the power of association studies unless very large
82 sample sizes are used [15].
83 Despite the primary role of gene overexpression in metabolic resistance, only one *cis*-
84 regulatory variant for resistance-linked differential expression has been identified in
85 *Anopheles gambiae* [19], and markers for such variants are therefore absent in the current
86 genetic marker panel for resistance [20]. Copy number variants (CNV) have been observed
87 in *Anopheles gambiae* metabolic resistance gene clusters [21]. For example, in *Anopheles*

88 *coluzzi* copy number of *Cyp6AA1* is associated with deltamethrin resistance [22]. The
89 relative contribution of CNV and *cis*-regulation on *Anopheles* gene expression has not yet
90 been determined.
91 Uganda sees a high burden of malaria; comprising 7.8% of all global cases in 2021 [23]. To
92 address this public health burden, the insecticide bendiocarb has been used for indoor
93 residual spraying (IRS) to complement the distribution of long-lasting insecticidal nets (LLIN).
94 Some resistance to bendiocarb in *An*. *gambiae* was observed in Nagongera (southeast
95 Uganda) and Kihihi (southwest Uganda), with 83% and 70% mortality, respectively, to WHO
96 bioassays with a diagnostic dose of 0.1% bendiocarb in 2014 prior to the IRS [24]. The IRS
97 campaign starting in December 2014 succeeded in reducing human biting rate, cases of
98 malaria and test positivity rate in Nagongera [24], but the potential for further increases in
99 resistance puts the long-term usefulness of bendiocarb into question.
100 Mosquitoes collected from Nagongera in 2014, which have moderate resistance to
101 bendiocarb, showed significant differential expression of many genes compared to the
102 susceptible Kisumu strain, including salivary gland protein encoding *D7r2* and *D7r4* genes as
103 well as the detoxification associated genes *Gstd3* and *Cyp6m2* [25]. Expression of *D7r4* was
104 associated with a single-nucleotide polymorphism (SNP) in a non-coding transcript
105 downstream of the D7 cassette [25].
106 In diploid organisms, allele specific expression (ASE) provides strong evidence that genes
107 may be under differential *cis*-regulatory control [26, 27]. Using a method that has been
108 applied to a variety of taxa [26, 28-32] but not mosquitoes, we describe the identification of
109 genes showing ASE in *An. gambiae* which potentially confer metabolic resistance. The
110 challenges of applying this method that arise from mosquito biology and genome structure
111 are discussed. In complementary work we predicted the sequences of some of the *cis*

112 regulatory modules that may underlie the expression of genes involved in insecticide
113 resistance using machine learning. Predictions included potential CRMs proximal to the
114 genes showing ASE and genes that show consistent differential expression patterns in
115 multiple resistant *Anopheles* strains, providing a starting point for future investigations into
116 CRM variants during the evolution of insecticide resistance.
117

118 Materials and Methods

119 Strains and Crosses

120 Resting female mosquitoes were collected in Siwa Village, Nagongera, Tororo District in
121 Eastern Uganda (0°46'12.0"N, 34°01'34.0"E) in March 2013 and allowed to lay eggs. 65
122 *Anopheles gambiae* (as ascertained by species ID PCR [33]) egg batches laid by the collected
123 females were reared to larvae and combined to establish the Nagongera colony. All the
124 colony founding mothers were screened for a known bendiocarb target site resistance
125 mutation G119S in the *Ace1* gene using the TaqMan assay described in [34]. No resistance
126 associated variants were observed and full sequencing of the *Ace1* gene from six colony
127 mosquitoes that survived Bendiocarb exposure did not reveal any other potential target site
128 resistance mutations. The colony was assayed using WHO tube assays [35] at F2 and was
129 found to be highly resistant to DDT (at diagnostic dose 4%) and deltamethrin (at diagnostic
130 dose 0.05%), intermediate resistance to the carbamate insecticide bendiocarb (at diagnostic
131 dose 0.1%). Bioassays of Nagongera colony adults using the WHO tube assay [35, 36] with
132 preexposure to 4% piperonyl butoxide (which inhibits P450 mediated metabolism) for one
133 hour prior to bendiocarb exposure increased mortality compared to bendiocarb exposure
134 alone, implying that bendiocarb resistance is metabolic rather than target site mediated.

135 A schematic of the experimental design is shown in figure 1. In brief, reciprocal crosses
136 between Nagongera mosquitoes and an inbred insecticide susceptible colony (origin
137 Kisumu, Kenya, 1975) were performed in cages as described previously [6]. Reciprocal
138 crosses each involved a total of 13 males and multiple females; since *Anopheles gambiae*
139 are swarm maters, multiple males are required to induce mating. Following mating, females
140 were transferred to individual cups for egg laying. F1 progeny of three females from each
141 reciprocal cross were raised to adults under standard insectary conditions of 12 hours light
142 12 hours dark cycle, 26°C +/-2°C and 70% relative humidity, and fed on 10% sucrose
143 solution (Table 1, Figure 1).

144 Sequencing

145 RNA was extracted from pools of 10 female F1 progeny from each of the six crosses 3-5 days
146 after eclosion using RNAqueous4PCR total RNA isolation kit (Invitrogen). RNA quality and
147 quantity was checked to be adequate for library preparation using the Agilent Bioanalyser
148 profile and Qubit® 2.0 Fluorometer. Total RNA libraries were prepared for Illumina paired-
149 end indexed sequencing according to the Illumina TruSeq RNA sample preparation v2 guide
150 [37]. cDNA libraries were bar coded, pooled and sequenced using the Illumina HiSeq1500
151 platform, with 100bp paired end reads.
152 DNA was extracted using the Qiagen DNeasy Kit from the 6 individual mothers that laid
153 viable eggs, all individual potential fathers that were alive following the cross, and individual
154 male F1 siblings (Figure 2) and sequenced using the Illumina HiSeq 2000 as described
155 previously [6].

156 ## Genome analysis

157 Quality control of genomic sequences for mothers, potential fathers and F1 progeny, and
158 matching of fathers to the progeny using Mendelian Error analysis was performed as
159 described previously [6]. Briefly, for the autosomes for each cross where the mother's
160 genome sequence was of sufficient quality (K2, K4, K6 and B5), the per progeny sample
161 Mendelian error was calculated for each possible parental pair of the known mother with all
162 potential fathers using scikit-allel [38]. The father in the parental pair that had the smallest
163 median progeny sample Mendelian error on at least 3 autosomal arms was deemed to be
164 the true father.
165 Scikit-allel [38] was used to analyze the genome of potential fathers including generating a
166 consensus sequence of homozygous sites for Kisumu and Nagongera colonies, and
167 calculating the number of single nucleotide polymorphisms (SNPs) differing between the
168 colonies.

169 ## RNAseq analysis

170 Samples were compared using the RNA-Seq-Pop snakemake workflow [39] which includes
171 quality control of RNAseq reads using with FastQC [40] and MultiQC [41], read alignment to
172 the Anopheles gambiae PEST reference genome (AgamP4, INSDC Assembly
173 GCA_000005575.1, Feb 2006) using HISAT2 [42], principal components analysis to group
174 samples by gene expression, differential expression analysis using Kallisto [43] with DESeq2
175 [44], variant calling using Freebayes version 1.3.2 [45] with ploidy 20 (as 10 diploid
176 individuals were pooled in each sample), calculation of population diversity statistics with
177 scikit-allel [38], and gene set enrichment analysis using GSEA [46]. RNAseq data for the

178 Kisumu parental colony, Busia G28 deltamethrin selected colony [39] and Tiefora pyrethroid
179 resistant colony [47] was compared with the cross progeny RNAseq data in this study.

180 Selection of SNPs to detect ASE

181 Detection of ASE relies on sufficient SNP differing between the parents [48]. The number of
182 SNPs differing between maternal and paternal genomes varied (Table 2). For crosses B1 and
183 B3, with parent genomes unavailable, we instead attempted to use parental colony
184 consensus genome sequences for mapping. However, the number of SNPs differing in
185 Nagongera and Kisumu consensus sequences (Table 3) was insufficient for ASE analysis. The
186 low number of differing SNPs may be due to barriers to recombination imposed by the
187 genome structure of *Anopheles gambiae*, which limit the extent to which colonies become
188 inbred, with heterozygosity persisting even after many generations of inbreeding [49]. This
189 contrasts with *Drosophila* where colony genotypes have been used effectively for inference
190 of ASE [50, 51]. Since the mosquitoes were pooled prior to RNA extraction, inferring SNPs
191 from the RNAseq data to detect ASE is problematic, since most SNPs are heterozygous in
192 one or both parents, resulting in a mixture of homozygous and heterozygous SNPs in the
193 progeny, and we would not know at each SNP whether to test for deviation from 1:1
194 expression or 3:1 expression. To increase the number of SNPs to detect ASE in these pooled
195 samples we therefore used the siblings of the RNA sequenced F1 to infer which SNPs were
196 homozygous but different between the parents (opposite homozygous SNPs). For a
197 Mendelian inherited SNP, opposite homozygous SNP in parents should result in all progeny
198 having heterozygous biallelic SNP. If one or both parents were heterozygous (biallelic) then
199 50% of the progeny are expected to be heterozygous, and the probability of all the progeny
200 being heterozygous in this situation is given by the binomial probability density function
201 where p=0.5, the number of trials (per SNP) is the number of progeny and the number of

202 “successes” is also the number of progeny. The number of progeny and SNPs per
203 chromosome arm where all progeny were heterozygous is given in Supplementary Table 2.
204 Almost all sequenced siblings were male, the heterogametic sex in *An. gambiae*, so it was
205 not possible to infer SNPs on the X chromosome using this method.
206 The suitability of these SNPs to detect ASE was first tested by comparing the SNPs inferred
207 from the siblings to those inferred from the parents, where parental genotypes were known
208 (Supplementary table 3). Most SNPs were shared between the parents and siblings, with
209 only a small fraction (<7.5%) unique to the siblings. The total number of usable SNPs from
210 the siblings was lower than those in the parents: a reduction of between 6% and 16%. Per
211 SNP ASE was calculated as counts at reference SNP/counts at reference SNP plus counts at
212 alternative SNP for the sibling SNPs, and as counts at maternal SNP/ counts at maternal plus
213 paternal SNP for parent SNPs. The mean per SNP ASE for SNPs with at least 10 reads
214 counted remained between 0.49 and 0.51 confirming that the SNPs are likely to be
215 heterozygous in the F1 and that mapping bias is not a problem for this dataset. The standard
216 deviation of per SNP ASE increased very slightly (from 0.12-0.14 to 0.13 to 0.15). For crosses
217 B1 and B3, the per SNP ASE was in line with the other crosses with mean 0.5 and standard
218 deviation 0.13 and 0.15 respectively. However, when the per SNP ASE was plotted against
219 the total counts at each SNP, a cluster of SNPs a with high ASE and expression was seen for
220 the sibling but not parent inferred SNPs (Figure 3, Supplementary Figure 3). SNPs in crosses
221 K2, K4 and B5 that were unique to siblings and with high ASE and high expression were
222 therefore removed from the analysis for crosses B1 and B3 before inferring per gene ASE.
223

Analysis of allele specific expression (ASE)

For parent-based mapping, SNPs that were homozygous for different alleles in the parents were used to distinguish whether reads originated from a maternal or paternal allele. Due to the pooling of 10 females for RNA sequencing it was not possible to use parental heterozygous SNPs, since tests of ASE are based on deviation from an expected 1:1 expression ratio and not the 1:3 or 3:1 ratio that would result from one heterozygous and one homozygous parent. For sibling-based mapping, SNPs that distinguished maternal and paternal alleles on the autosomes were inferred from the genome sequences of the male siblings, using the following criteria: SNPs must be heterozygous and biallelic in all siblings where data was available. To eliminate potential low-quality SNPs, if five or more siblings had missing data at a SNP it was not used. Since F1 genome data was unphased, it was not possible to assign SNPs to either the Kisumu or Nagongera parent using this method.

Analysis of F1 RNAseq data for ASE was performed using ASEReadCounter* [52], which is based on ASEReadCounter [53], using individual parent genomes where available for each cross to assign reads to parental alleles (options –vcf_mat, -vcf_pat). Where only sibling inferred SNPs were known, option –vcf_joind was used on the unphased inferred F1 VCF file.

The properties of SNPs inferred from parents and siblings were compared by inferring parental pseudogenomes and running ASEReadCounter* with both sets of SNPs. SNP level ASE was calculated for all SNPs with >=10 total counts. SNPs which were inferred from siblings and not from parents, which had >100 counts and ASE < .25 or > .75 were removed from the sibling inferred SNPs for subsequent analysis.

Due to overdispersion, RNAseq data is not binomially distributed but can be better described by a beta binomial distribution. The distribution of counts data and

248 overdispersion parameter rho for the beta binomial distribution was calculated using the
249 counts data for each cross, including the counts for the SNP with the maximum counts in
250 each gene (with minimum 10 counts), using the R package glmmTMB [54], which uses
251 maximum likelihood estimation to fit binomial and beta binomial model to the data for each
252 cross. Models were compared using ANOVA in base R [55].
253 To obtain a gene level measurement of ASE, the R package MBASED [56] was used to
254 aggregate SNP level counts to gene level ASE by pseudo phasing SNPs in each gene such that
255 the SNPs with the higher read count at each variable site are combined into the major allele,
256 with frequency between 0.5 and 1. Only sites with at least 10 counts were used. For SNPs
257 inferred from the siblings of the RNAseq pool, MBASED was run assuming SNPs were not
258 phased. Where phase was known from the parents, MBASED was run in both phased and
259 non-phased mode to measure the impact on power to detect loci with ASE. In all cases the
260 same seed (988482) was set prior to running 10^6 simulations in which the total read count at
261 each SNP position in each gene are kept constant, but reference allele counts are drawn
262 from a null distribution beta binomial (mean 0.5 x total count, rho 0.038). The P value is the
263 proportion of simulations in which the major allele frequency is greater than or equal to the
264 observed major allele frequency. For genes with multiple SNPs available to infer ASE, the P
265 value for heterogeneity (Phet) between the level of ASE at each SNP within a gene was
266 additionally calculated. Low P het indicates possible isoform specific ASE. False discovery
267 correction was applied to both P values for ASE and Phet with a nominal rate of 5% [57].
268 For SNPs inferred from parents, allelic count tables were additionally analyzed using Qleelic
269 [52] with a binomial test of the null hypothesis of equal numbers of transcripts arising from
270 maternal and paternal alleles. ASE was considered significant at p=0.05 following Bonferroni
271 correction. Genes with fewer than 10 read counts were excluded. As technical replicates

272 were unavailable, an exploratory solution for correcting overdispersion was splitting reads
273 into groups (pseudoreplicates K4 14 groups, K2 three groups, K6 nine groups, B5 nine
274 groups). This produced a consistent quality correction constant (QCC) value of 1.56 for QCC
275 correction. R packages UpSetR [58], stats [55] and ggplot2 [59] were used for regressions
276 and plots.

277 Enrichment analysis

278 Hypergeometric tests for enrichment of gene ontology terms, Pfam protein domain
279 enrichment, and KEGG pathways were implemented using python [60]. The test set for each
280 cross was the set of genes showing significant ASE following FDR correction, and the
281 "universe set" was the set of genes for each cross that had at least 10 reads containing SNPs
282 distinguishing maternal and paternal alleles (for these genes ASE would have been
283 detectable if present).
284 The evolutionary age of genes was taken from the Princeton Protein Orthology Database
285 PPODv4_PTHR7-OrthoMCL [61], with age estimates based on Wagner parsimony. A gene
286 age enrichment test was implemented using the age_enrichment.py script in
287 ProteinHistorian [62], using option '-a ignore' to ignore the proteins not present in the
288 database. P values were obtained using a Mann Whitney U test for difference in overall age
289 distribution between sets and Fisher's exact test for difference in fraction of proteins in each
290 age in the set.
291 The null hypothesis, the proportion of genes showing significant ASE does not differ
292 between autosome arm, was tested using a two-sided Fisher's exact test (for those crosses
293 where the counts were too high for efficient computation P value was simulated using 2000
294 replicates).

295 To assess presence of CNV we used the CNV calls for Ag1000g phase 3, which applied a
296 Gaussian hidden Markov Model (HMM) to the normalized windowed coverage data for each
297 genotyped individual using the methods described in Lucas et al 2019 [21]. Copy number
298 variation was deemed possible if any of the parents and F1 siblings from crosses B5, K2, K4
299 or K6 (total 63 individuals) had a copy number call other than 2 on the autosomes. As ASE
300 could not be inferred on the X chromosome for crosses B1 and B3 we excluded it from the
301 analysis. The father of cross B5, AC0416-C had high variance in copy number calls and was
302 therefore removed from CNV analysis.
303 To assess selective sweeps, lists of genes either showing ASE or having SNPs to detect ASE
304 but not showing ASE were compared to a database of regions under selection in the
305 *Anopheles* selection atlas [63] as defined by H_{12} signal [64] both across the range of
306 *Anopheles gambiae* and restricted to Uganda using a custom Python script
307 (https://github.com/azurillandfriend/sweeps_ASE.git).

308 Prediction of *cis*-regulatory modules

309 *Anopheles cis*-regulatory modules (CRMs) were predicted using SCRMshaw [65-68]. Briefly,
310 *Drosophila* CRMs which drive gene expression in the tissue of interest were downloaded
311 from the Redfly database (v9.6.0, database updated 02/01/2023) [69, 70],with max size
312 2000. Only non-overlapping sequences (>100bp) were included. Syntenous regions from
313 related *Drosophila* species (putatively containing the equivalent CRM) were added to the
314 training data set using liftOver at https://genome.ucsc.edu/cgi-bin/hgLiftOver [71] as
315 described in [66]. SCRMshaw was trained using this augmented training set and a 10x bigger
316 set of non-CRM non-exonic regions. Repeats in CRMs, non-CRMs and the target *Anopheles*
317 *gambiae* PEST genome were masked using repeat masker [72]. Existing training sets for
318 adult peripheral nervous system, embryonic and larval excretory, and embryonic/ larval

319 Malpighian tubules were downloaded from GitHub

320 (https://github.com/HalfonLab/dmel_training_sets). These were compared with the 2272

321 Tn5 transposase hypersensitive sites identified by Ruiz et al [73].

322 Results and Discussion

323 Crosses

324 Three females from each reciprocal cross between the Kisumu and Nagongera strains laid

325 viable eggs (Figure 1, Table 1). For four crosses the father was identified by minimizing

326 median Mendelian error (Supplementary Table 1, but for the other two crosses the mother

327 failed sequencing quality control and none of the sequenced putative fathers were a good

328 match. We presume the true fathers for these crosses died during the experiment,

329 precluding extraction of good quality DNA.

330 RNAseq statistics for the pooled F1 females from each of these six crosses are shown in

331 Table 1. The number of reads recommended for 60% power to detect ASE at 1.5 fold is 500

332 per gene [48], which for the 13,796 annotated genes in *Anopheles gambiae* would require

333 around 6.9×10^6 reads. Five crosses exceeded this value with one cross, K2, having slightly

334 fewer (6.4×10^6reads). The proportion of reads mapping to the *Anopheles gambiae* genome

335 was similar for all crosses (mean 87.0%, standard deviation 1.1%).

336 Between pools gene expression comparison

337 Comparison between reciprocal crosses revealed similar gene expression in the F1 progeny,

338 with just 38 genes significantly downregulated and 38 genes significantly upregulated in the

339 F1 between reciprocal crosses at $P_{adj} \leq 0.05$ (Figure 2a). When comparing only F1 crosses in

340 PCA, B1 seemed to be an outlier in terms of expression (Supplementary figure 1a). However,

341 when compared with other colonies of *Anopheles gambiae* s.l., the F1 progeny formed a
342 distinct group (Supplementary figure 1b).
343 F1 gene expression differed from the Kisumu (parental) colony, with a total of 1362 genes
344 upregulated and 1265 genes downregulated in the F1 compared to the Kisumu colony (Padj
345 ≤ 0.001) (Figure 2b). RNAseq data was unavailable for the Nagongera colony, which no
346 longer exists. Gene set enrichment analysis of the differentially expressed genes indicated
347 that genes with Gene Ontology (GO) terms associated with odorant binding, olfactory
348 receptor activity, sensory perception of smell, response to stimulus, detection of chemical
349 stimulus involved in sensory perception of smell, structural constituent of cuticle and signal
350 transduction were downregulated in the F1 compared to Kisumu, whereas those with GO
351 terms associated with translation, ribosome, mitochondrion, structural component of
352 ribosome and serine type endopeptidase activity were upregulated in the F1 compared to
353 Kisumu (all at P adjusted 0.05). KEGG pathway analysis showed significant upregulation of
354 ribosome, citrate cycle (TCA cycle) and oxidative phosphorylation in the F1 compared to
355 Kisumu, and no KEGG pathway significantly downregulated.

356 **ASE inference**

357 For the four crosses where the parent was known, per gene ASE was calculated as reads
358 mapping to maternal genome/total reads mapped at that gene. In addition the major allele
359 frequency at each gene was calculated for all crosses. Read count data were fitted to
360 binomial and beta binomial models. ANOVA indicated that the beta binomial model was a
361 better fit to the data (P < 2.2e-16). The dispersion parameter phi(disp) for the beta binomial
362 distribution of counts data for crosses B1, B3, B5, K2, and K4 was estimated at 25.8, giving a
363 rho 1/(1+phi(disp)) of 0.038. Per gene ASE statistics for all crosses are shown in Table 4. Due
364 to variations in parental genome and sequencing depth the power to detect ASE varies

365 between crosses, so it is not possible to infer whether the proportion of genes with ASE is
366 different between the crosses. Three crosses showed similar numbers of genes with
367 maternal or paternal ASE except cross K6 which showed extreme paternal ASE. This could
368 either be due to K6 showing a very different pattern of ASE to the other crosses, or a
369 sampling error occurred so that the RNA sequenced samples were not from the same cross
370 as the parents and siblings. We therefore checked the effect of using SNPs from the wrong
371 parents to infer ASE. Using SNPs from non-matching parents resulted in inflated ASE
372 estimates (Supplementary Table 4). Using SNPs from the parents of cross K4 to infer allelic
373 counts for cross K6 shifted mean imbalance back to 0.5 (figure 3, Supplementary Table 4).
374 Furthermore, when SNPs called from the RNAseq data were compared, the progeny of cross
375 K4 and K6 were extremely genetically similar and clustered closely in PCA for all
376 chromosomal arms (Supplementary Figure 4), suggesting they may have hatched from two
377 egg batches from the same cross. Gene level ASE for cross K6 was therefore inferred using
378 SNPs from K4 parents.
379 Genes showing ASE showed some overlap between the different crosses (Figure 4). At the
380 most conservative estimate, 13 genes showed ASE in all crosses. This exceeds the overlap
381 expected by random genes showing ASE, as in 100,000 simulations of randomly drawing the
382 observed number of significant genes from the set of genes where SNPs were available to
383 detect ASE in all crosses, the maximum overlap was 1. Most genes showing ASE were unique
384 to individual crosses (Figure 4), but there were many genes showing ASE in combinations of
385 multiple crosses that may also be under consistent differential *cis*-regulation between
386 Tororo and Nagongera colonies, Table 5 shows the 115 genes with significant ASE in at least
387 4 out of the 6 crosses. Directionality of ASE (maternal or paternal bias) was inferred for
388 crosses B5, K2, K4 and K6 (Supplementary Table 7).

389

390 Analysis of gene ages that showed or did not show ASE revealed an enrichment of younger,
391 *Anopheles* specific genes showing significant ASE (Figure 5 and Supplementary Table 5). The
392 same trend was seen in all crosses.
393 Genes showing ASE showed an unequal distribution along chromosomal arms compared to
394 genes not showing ASE (Supplementary table 6, Figure 6, Supplementary figure 2). This
395 trend was significant for all crosses except for K2. The trend was for slight enrichment on 3L.
396 For the 4 crosses where parents were available, this was also compared with the X
397 chromosome, which had a lower proportion of ASE/total detectable genes (Figure 6,
398 Supplementary figure 2). ANOVA indicated both chromosome and cross explained the
399 variance in the proportion of ASE/total detectable genes per chromosome arm ($P < 0.05$).

400 **Copy number variation**

401 True ASE is caused by *cis*-regulation, but copy number variation of the expressed gene could
402 lead to apparent ASE if there are different numbers of gene copies containing the SNPs used
403 to count reads; e.g. if two copies of a duplicated gene with total of three copies bear one
404 SNP and the third the other SNP, apparent ASE would be inferred without differential *cis*-
405 regulation. We therefore checked for CNV in the parental and F1 sibling genomes at the
406 genes that showed ASE in at least four out of six crosses, and in the genes which contained
407 sufficient SNPs to infer ASE but expression appeared to be in balance. A total of 60/114
408 (53%) autosomal genes showing ASE contained possible CNV. For genes that did not show
409 ASE and had available SNPs to infer ASE in all crosses, 481/ 1333 (36%) had possible CNV. A
410 two-sided Fisher's exact test rejected the hypothesis that the odds ratio is 1 (odds ratio 2.0,
411 P=0.0006). This suggests that for some genes CNV may be contributing to the observed ASE.

412 The number of individuals with a possible CNV in each gene varied between the genes
413 (Table 5).
414

415 Overlaps between ASE genes and other datasets

416 The list of 115 genes showing ASE in at least four crosses was compared with published
417 gene expression data. All other available *Anopheles* expression data is based on between
418 sample comparison and represents a mixture of *cis* and *trans* regulation, whereas ASE data
419 is specific to *cis* regulation and excludes any *trans* regulation. Therefore even for extremely
420 similar ASE and between sample comparison datasets only *cis*-regulated genes are expected
421 to overlap.
422 Genes showing ASE in at least four crosses were compared with genes with consistently high
423 median fold change in a meta-analysis of 35 experiments comparing RNAseq data between
424 *Anopheles gambiae* s.l. and *Anopheles funestus* strains [60], and with genes showing
425 significant fold change with consistent directionality in microarray data comparing
426 susceptible and resistant populations of *Anopheles coluzzi* [74]. For the RNAseq metadata
427 set, the genes with the top 5% of median fold changes between susceptible and resistant
428 populations (429/8599 total genes in the dataset), six genes were also present in the ASE
429 gene set. These were AGAP001251 (*Eupolytin*), AGAP008218 (*Cyp6Z2*), AGAP012296
430 (*Cyp9J5*), AGAP011068 (A*ldose reductase*), AGAP003583 (*L-iditol 2-dehydrogenase*) and
431 AGAP008331 (*WD repeat-containing protein 59*). There was no overlap between ASE and
432 microarray data.
433

434 Gene set enrichment analysis did not reveal any consistent patterns across crosses for
435 enrichment of genes showing significant ASE in any GO term or Kegg pathway. Trypsin and
436 Ubiquitin pfam domains were significantly enriched in 4/6 crosses.
437
438 Of the gene families previously implicated in metabolic resistance, we observed four P450s
439 with ASE in four of the six crosses: *Cyp12F3* (AGAP008019), *Cyp12F2* (AGAP008020), *Cyp6Z2*
440 (AGAP008212) and *Cyp9J5* (AGAP012296).
441
442 The D7 protein family has previously been implicated in bendiocarb resistance in Uganda
443 [25]. AGAP008282 (*D7r2*) showed significant ASE in two crosses (detectable in 4) and *D7r4* in
444 none (detectable in 3). Other *D7r* genes, *D7r1* (AGAP008284), *D7r5* (AGAP008280) also
445 showed significant ASE in two crosses, *D7L1*, *D7L2* and *D7r3* in one cross.
446
447 We finally asked whether genes showing ASE were more likely to be in a genomic region
448 that has undergone a recent selective sweep. Out of 115 genes showing ASE in at least 4/6
449 crosses, 13 were in a swept region, whereas 103 of the 1333 genes showing no evidence of
450 ASE in any of the 6 crosses were in swept regions. A two sided Fisher's exact test did not
451 reject the hypothesis that the odds ratio is 1 (odds ratio 1.5, P=0.2), suggesting that genes
452 showing ASE are no more likely to be in a swept region than genes that do not.
453

454 **CRM prediction**

455 *Drosophila* CRMs for adult midgut, adult Malpighian tubules, larval midgut and legs were
456 used as training data from SCRMshawHD to predict *Anopheles* CRMs operating in the same
457 tissues, together with previously developed training sets for the adult peripheral nervous

458 system, embryonic and larval excretory system and embryonic/ larval Malpighian tubules
459 (https://github.com/HalfonLab/dmel_training_sets). These tissues were selected based on
460 previous studies examining the tissue specific expression of genes involved in insecticide
461 resistance with roles including detoxification and cuticular resistance [75-80]. Training CRM
462 sets used for the first time in this study are shown in table 6, with full sequences at
463 https://github.com/azurillandfriend/traning_sets_IR.git. Oenocyte (FBbt:00004995), cuticle
464 (FBbt:00004970) and adult epidermis (FBbt:00005401) CRMs could not be used as training
465 data due to insufficient experimentally validated CRMs, highlighting the need for more
466 research to identify CRMs in these tissues. The top scoring 250 predictions for each training
467 set and method were combined, producing a total of 4122 unique CRM predictions. 62 of
468 these predicted CRMs were flanked by a gene showing ASE in at least 4/6 crosses (Table 7).
469 In total, CRMs were predicted for 33 of the 115 genes showing ASE in at least 4/6 crosses.
470 211 predicted CRMs were flanked by a gene showing consistently high median fold change
471 between resistant and susceptible strains (Supplementary table 8). CRMs were predicted for
472 a total of 141 of the 429 genes in this set. The predicted CRMs were also compared with a
473 published dataset of Tn5 transposase sensitive sites in the adult midgut [73, 81]. 50 of the
474 predicted CRMs overlapped with Tn5 transposase sensitive sites identified as *cis*-regulatory
475 elements by Ruiz et al [73], flanking a total of 48 genes. The vast majority of these CRMs
476 (47) were previously predicted by Kazemian [82] despite the different training sets. The
477 CRMs identified in these predictions provide a starting point for future studies to examine
478 genetic variation in Anopheles populations with different insecticide resistance phenotypes,
479 and for validation using reporter assays.
480

481 # Conclusions

482 This study of allele specific expression in *Anopheles* provides the first evidence that *cis*-
483 regulation of gene expression occurs across the genome and differs between strains in this
484 species.
485 Sample pooling, while a cost-effective solution for bulk RNAseq of small samples, does limit
486 the power of experiments targeting the study of ASE; in future, to maximize the power to
487 detect ASE, individual samples should be used. RNAseq on whole mosquitoes may have
488 masked tissue specific ASE. The true number of genes showing ASE in individual tissues is
489 likely higher than we observed on the whole mosquitoes analyzed here, since pleiotropic
490 effects may limit the potential for genes to up or down regulated in all tissues
491 simultaneously but permit the evolution of tissue specific regulation [83]. Future studies
492 should target specific tissues of interest. Despite these limitations, we were able to detect
493 genes showing ASE in pooled RNA from whole mosquitoes in crosses between *Anopheles*
494 strains with different carbamate resistance phenotypes, indicating different *cis*-regulation
495 patterns between the strains. Whilst we detected some genes previously implicated in
496 insecticide resistance, there was no consistent enrichment of these or any particular gene
497 ontology term amongst genes showing ASE. This probably indicates that the strains have
498 undergone extensive *cis*-regulatory divergence, affecting both genes involved in insecticide
499 resistance but also genes involved in many other functions. Comparing ASE in progeny of
500 crosses between multiple insecticide resistance and susceptible strains, together with
501 examining gene expression in the parental strains, would enable a more comprehensive
502 survey of the *cis* versus *trans* regulation of insecticide resistance genes. The bias towards

503 younger, *Anopheles* specific genes showing ASE in the F1 suggests there may be a higher
504 degree of *cis*-regulatory divergence between the parental strains for younger genes.
505 It was possible to computationally predict some CRM involved in tissue specific expression,
506 including potential CRM for genes showing ASE and genes previously implicated in
507 insecticide resistance. This was hampered by the lack of good quality training data for the
508 tissues thought to be most relevant to insecticide resistance in adult mosquitoes,
509 highlighting the need for future experimental CRM discovery.

510 **Data availability statement**

511 1. RNAseq data for all crosses and read counts per SNP are available at Gene Expression
512 Omnibus accession GSE241768.
513 2. Potential fathers sequencing data are available at European nucleotide archive
514 (accession numbers in Supplementary Table 9).
515 3. Sequencing data from siblings from crosses B1 and B3 are at European nucleotide
516 (accession numbers in supplementary Table 10)
517 4. All other nucleotide sequences used are already publicly available in the Phase 3
518 release of the *Anopheles gambiae* 1000 genomes project, at accession PRJEB42254.
519 For convenience, individual accessions for the parents and F1 male sibs in crosses B5,
520 K2, K4 and K6 are in Supplementary Table 11.

521 # Acknowledgements

522 This project was funded by Daphne Jackson Fellowship to N.A.D. sponsored by the
523 Biotechnology and Biosciences Research Council, with RNA sequencing funded by Infravec
524 (www.infravec.eu). Additional support to this work was provided by the National Institute of
525 Allergy and Infectious Diseases (NIAID R01-AI116811) and the Medical Research Council
526 (MR/T001070/1). M.J.D. is supported by a Royal Society Wolfson Fellowship (RSWF\FT
527 \180003).
528 This study was supported by the MalariaGEN Vector Observatory which is an international
529 collaboration working to build capacity for malaria vector genomic research and
530 surveillance, and involves contributions by the following institutions and teams. Wellcome
531 Sanger Institute: Lee Hart, Kelly Bennett, Anastasia Hernandez-Koutoucheva, Menelaos
532 Ioannidis, Julia Jeans, Paballo Chauke, Victoria Simpson, Eleanor Drury, Osama Mayet, Sónia

533 Gonçalves, Katherine Figueroa, Tom Madison, Kevin Howe, Mara Lawniczak; Liverpool
534 School of Tropical Medicine; Broad Institute of Harvard and MIT: Jessica Way, George Grant;
535 Pan-African Mosquito Control Association: Jane Mwangi, Edward Lukyamuzi, Sonia Barasa,
536 Ibra Lujumba, Elijah Juma. The authors would like to thank the staff of the Wellcome Sanger
537 Genomic Surveillance unit and the Wellcome Sanger Institute Sample Logistics, Sequencing
538 and Informatics facilities for their contributions.
539 The MalariaGEN Vector Observatory is supported by funding awarded to Dominic
540 Kwiatkowski and Mara Lawniczak from Wellcome (220540/Z/20/A, 'Wellcome Sanger
541 Institute Quinquennial Review 2021-2026') and funding awarded to Dominic Kwiatkowski
542 from the Bill and Melinda Gates Foundation (INV-001927). The Liverpool School of Tropical
543 Medicine's participation was supported by the National Institute of Allergy and Infectious
544 Diseases ([NIAID] R01-AI116811), with additional support from the Medical Research
545 Council (MR/P02520X/1). The latter grant is a UK-funded award and is part of the EDCTP2
546 programme supported by the European Union. The Pan-African Mosquito Control
547 Association’s participation was funded by the Bill and Melinda Gates Foundation (INV-
548 031595).
549 The authors are grateful for discussions with Louise Cerdeira, Laura Brettell and Shannon
550 Quek, and thank Luciene Salas Jennings and Andrew Carey for providing administrative
551 support to the project.
552
553
554

References

1. Bhatt, S., et al., *The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015.* Nature, 2015. **526**(7572): p. 207-+.
2. World Health Organisation, *Malaria Report 2019.* 1019.
3. Ranson, H. and N. Lissenden, *Insecticide Resistance in African Anopheles Mosquitoes: A Worsening Situation that Needs Urgent Action to Maintain Malaria Control.* Trends Parasitol, 2016. **32**(3): p. 187-196.
4. Lynd, A., et al., *LLIN Evaluation in Uganda Project (LLINEUP): a cross-sectional survey of species diversity and insecticide resistance in 48 districts of Uganda.* Parasit Vectors, 2019. **12**(1): p. 94.
5. Holt, R.A., et al., *The genome sequence of the malaria mosquito Anopheles gambiae.* Science, 2002. **298**(5591): p. 129-49.
6. The Anopheles gambiae 1000 Genomes Consortium, *Genetic diversity of the African malaria vector Anopheles gambiae.* Nature, 2017. **552**(7683): p. 96-100.
7. The Anopheles gambiae 1000 Genomes Consortium, *Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii.* Genome Res, 2020. **30**(10): p. 1533-1546.
8. MalariaGen. *Anopheles gambiae genomic surveillance project*. 2023 17/11/2023; Available from: https://www.malariagen.net/anopheles-gambiae-genomic-surveillance-project.
9. Liu, N., *Insecticide resistance in mosquitoes: impact, mechanisms, and research directions.* Annu Rev Entom, 2015. **60**: p. 537-59.
10. Mitchell, S.N., et al., *Identification and validation of a gene causing cross-resistance between insecticide classes in Anopheles gambiae from Ghana.* Proc Natl Acad Sci U S A, 2012. **109**(16): p. 6147-52.
11. Edi, C.V., et al., *CYP6 P450 enzymes and ACE-1 duplication produce extreme and multiple insecticide resistance in the malaria mosquito Anopheles gambiae.* PLoS Genet, 2014. **10**(3): p. e1004236.
12. Muller, P., et al., *Field-caught permethrin-resistant Anopheles gambiae overexpress CYP6P3, a P450 that metabolises pyrethroids.* PLoS Genet, 2008. **4**(11): p. e1000286.
13. Ingham, V.A., et al., *The transcription factor Maf-S regulates metabolic resistance to insecticides in the malaria vector Anopheles gambiae.* BMC Genomics, 2017. **18**(1): p. 669.
14. Wilding, C.S., *Regulating resistance: CncC:Maf, antioxidant response elements and the overexpression of detoxification genes in insecticide resistance.* Curr Opin Insect Sci, 2018. **27**: p. 89-96.
15. Lucas, E.R., et al., *Genome-wide association studies reveal novel loci associated with pyrethroid and organophosphate resistance in Anopheles gambiae s.l.* bioRxiv, 2023: p. 2023.01.13.523889.
16. 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.* Sci Rep, 2018. **8**(1): p. 2920.

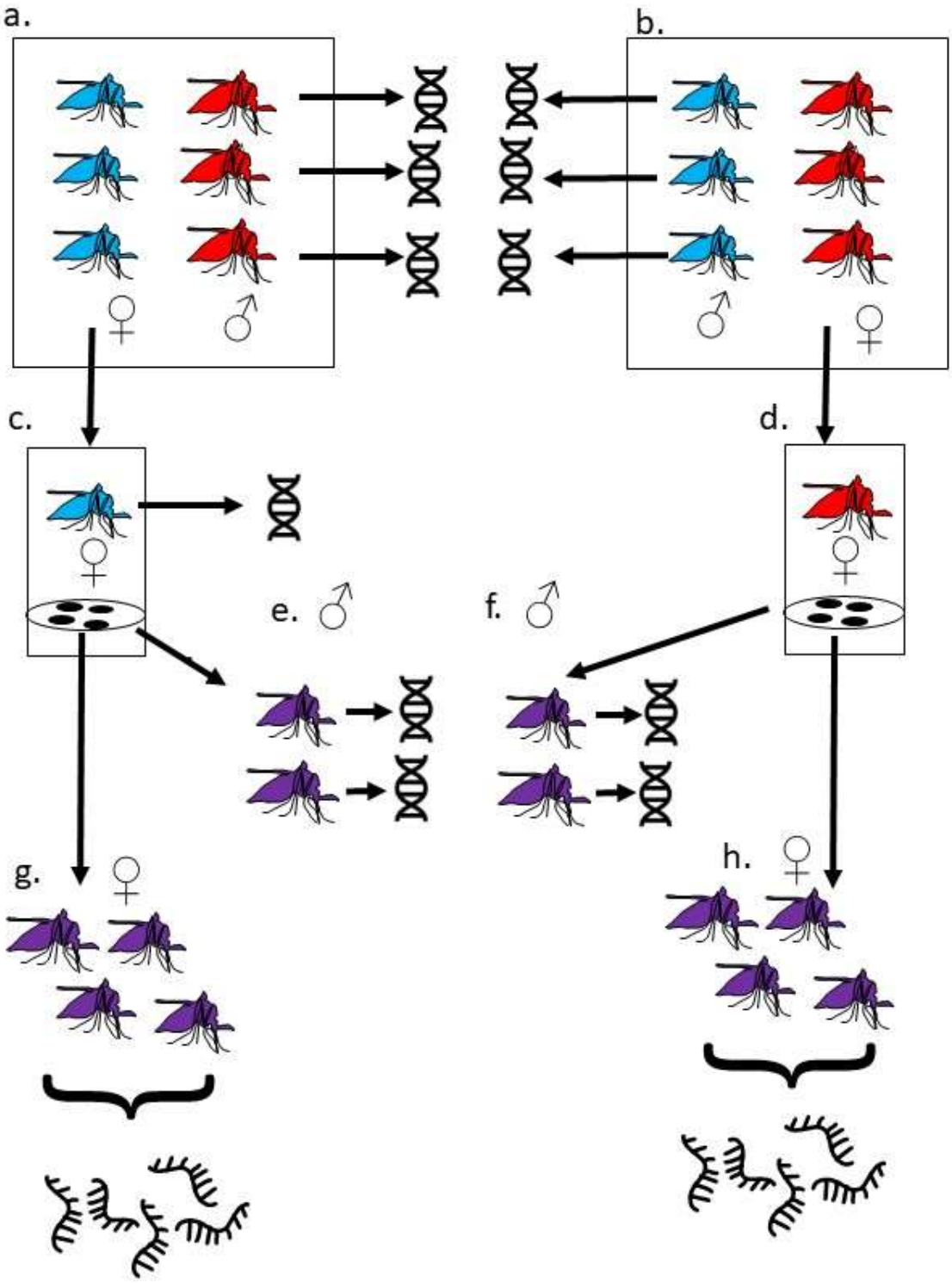
17. Ibrahim, S.S., et al., *Allelic Variation of Cytochrome P450s Drives Resistance to Bednet Insecticides in a Major Malaria Vector.* PLoS Genet, 2015. **11**(10): p. e1005618.
18. Main, B.J., et al., *Genetic variation associated with increased insecticide resistance in the malaria mosquito, Anopheles coluzzii.* Parasit Vectors, 2018. **11**(1): p. 225.
19. Ibrahim, S.S., et al., *Molecular drivers of insecticide resistance in the Sahelo-Sudanian populations of a major malaria vector.* bioRxiv, 2022: p. 2022.03.21.485146.
20. Lucas, E.R., et al., *A high throughput multi-locus insecticide resistance marker panel for tracking resistance emergence and spread in Anopheles gambiae.* Sci Rep, 2019. **9**(1): p. 13335.
21. Lucas, E.R., et al., *Whole-genome sequencing reveals high complexity of copy number variation at insecticide resistance loci in malaria mosquitoes.* Genome Res, 2019. **29**(8): p. 1250-1261.
22. Lucas, E.R., et al., *Genome-wide association studies reveal novel loci associated with pyrethroid and organophosphate resistance in Anopheles gambiae and Anopheles coluzzii.* Nat Commun, 2023. **14**(1): p. 4946.
23. World Health Organisation, *Malaria Report 2022*. 2022: Geneva.
24. Katureebe, A., et al., *Measures of Malaria Burden after Long-Lasting Insecticidal Net Distribution and Indoor Residual Spraying at Three Sites in Uganda: A Prospective Observational Study.* PLoS Med, 2016. **13**(11): p. e1002167.
25. Isaacs, A.T., et al., *Genome-wide transcriptional analyses in Anopheles mosquitoes reveal an unexpected association between salivary gland gene expression and insecticide resistance.* BMC Genomics, 2018. **19**(1): p. 225.
26. Tirosh, I., et al., *A yeast hybrid provides insight into the evolution of gene expression regulation.* Science, 2009. **324**(5927): p. 659-62.
27. Wittkopp, P.J., B.K. Haerum, and A.G. Clark, *Evolutionary changes in cis and trans gene regulation.* Nature, 2004. **430**(6995): p. 85-8.
28. Varkoohi, S., M.H. Banabazi, and M. Ghsemi-Siab, *Allele Specific Expression (ASE) analysis between Bos Taurus and Bos Indicus cows using RNA-Seq data at SNP level and gene level.* An Acad Bras Cienc, 2021. **93**(3): p. e20191453.
29. Graze, R.M., et al., *Allelic imbalance in Drosophila hybrid heads: exons, isoforms, and evolution.* Molecular Biology and Evolution, 2012. **29**(6): p. 1521-32.
30. Wang, M., S. Uebbing, and H. Ellegren, *Bayesian Inference of Allele-Specific Gene Expression Indicates Abundant Cis-Regulatory Variation in Natural Flycatcher Populations.* Genome Biol Evol, 2017. **9**(5): p. 1266-1279.
31. Reynes, C., et al., *ISoLDE: a data-driven statistical method for the inference of allelic imbalance in datasets with reciprocal crosses.* Bioinformatics, 2020. **36**(2): p. 504-513.
32. Pastinen, T. and T.J. Hudson, *Cis-acting regulatory variation in the human genome.* Science, 2004. **306**(5696): p. 647-50.
33. Scott, J.A., W.G. Brogdon, and F.H. Collins, *Identification of single specimens of the Anopheles gambiae complex by the polymerase chain reaction.* Am J Trop Med Hyg, 1993. **49**(4): p. 520-9.
34. Bass, C., et al., *Development of high-throughput real-time PCR assays for the identification of insensitive acetylcholinesterase (ace-1R) in Anopheles gambiae.* Pesticide Biochemistry and Physiology, 2010. **96**(2): p. 80-85.
35. World Health Organisation *Manual for monitoring insecticide resistance in mosquito vectors and selecting appropriate interventions*. 2022.

36. World Health Organisation *Standard operating procedure for determining the ability of PBO to restore susceptibility of adult mosquitoes to pyrethroid insecticides in WHO tube tests*. 2022.
37. Illumina *TruSeq® RNA Sample Preparation v2 Guide*. 2012.
38. Miles, A., et al., *scikit-allel*. 2021: https://doi.org/10.5281/zenodo.4759368.
39. Nagi, S.C., et al., *RNA-Seq-Pop: Exploiting the sequence in RNA sequencing-A Snakemake workflow reveals patterns of insecticide resistance in the malaria vector Anopheles gambiae.* Mol Ecol Resour, 2023. **23**(4): p. 946-961.
40. Andrews, S., *FastQC: A Quality Control Tool for High Throughput Sequence Data* 2010, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
41. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-8.
42. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nat Biotechnol, 2019. **37**(8): p. 907-915.
43. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nat Biotechnol, 2016. **34**(5): p. 525-7.
44. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
45. Garrison E and M. G., *Haplotype-based variant detection from short-read sequencing.* arXiv:1207.3907 [q-bio.GN], 2012.
46. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
47. Williams, J., et al., *Sympatric Populations of the Anopheles gambiae Complex in Southwest Burkina Faso Evolve Multiple Diverse Resistance Mechanisms in Response to Intense Selection Pressure with Pyrethroids.* Insects, 2022. **13**(3).
48. Fontanillas, P., et al., *Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing.* Mol Ecol, 2010. **19 Suppl 1**: p. 212-27.
49. Turissini, D.A., S. Gamez, and B.J. White, *Genome-wide patterns of polymorphism in an inbred line of the African malaria mosquito Anopheles gambiae.* Genome Biol Evol, 2014. **6**(11): p. 3094-104.
50. Frochaux, M.V., et al., *cis-regulatory variation modulates susceptibility to enteric infection in the Drosophila genetic reference panel.* Genome Biol, 2020. **21**(1): p. 6.
51. Floc'hlay, S., et al., *Cis-acting variation is common across regulatory layers but is often buffered during embryonic development.* Genome Res, 2020.
52. Mendelevich, A., et al., *Replicate sequencing libraries are important for quantification of allelic imbalance.* Nat Commun, 2021. **12**(1): p. 3370.
53. Castel, S.E., et al., *Tools and best practices for data processing in allelic expression analysis.* Genome Biol, 2015. **16**: p. 195.
54. Brooks, M.E., et al., *glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling.* The R Journal, 2017. **9(2)**: p. 378-400.
55. R Core Team, *R: A language and environment for statistical computing*. 2018, R Foundation for Statistical Computing https://www.R-project.org/: Vienna, Austria.
56. Mayba, O., et al., *MBASED: allele-specific expression detection in cancer tissues and cell lines.* Genome Biol, 2014. **15**(8): p. 405.

57. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society: Series B (Methodological), 1995. **57**(1): p. 289-300.
58. Conway, J.R., A. Lex, and N. Gehlenborg, *UpSetR: an R package for the visualization of intersecting sets and their properties.* Bioinformatics, 2017. **33**(18): p. 2938-2940.
59. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016, Springer-Verlag New York.
60. Nagi, S.C., *AnoExpress*. 2023, Zenodo: https://doi.org/10.5281/zenodo.7965646.
61. Heinicke, S., et al., *The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists.* PLoS One, 2007. **2**(8): p. e766.
62. Capra, J.A., A.G. Williams, and K.S. Pollard, *ProteinHistorian: tools for the comparative analysis of eukaryote protein origin.* PLoS Comput Biol, 2012. **8**(6): p. e1002567.
63. Nagi, S., *anopheles-genomic-surveillance/selection-atlas: v0.0.1-beta*. 2023, Zenodo: https://doi.org/10.5281/zenodo.10364083.
64. Garud, N.R., et al., *Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps.* PLoS Genet, 2015. **11**(2): p. e1005004.
65. Asma, H. and M.S. Halfon, *Computational enhancer prediction: evaluation and improvements.* BMC Bioinformatics, 2019. **20**(1): p. 174.
66. Kazemian, M. and M.S. Halfon, *CRM Discovery Beyond Model Insects.* Methods Mol Biol, 2019. **1858**: p. 117-139.
67. Kazemian, M., et al., *Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison.* Nucleic Acids Res, 2011. **39**(22): p. 9463-72.
68. Kantorovitz, M.R., et al., *Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse.* Dev Cell, 2009. **17**(4): p. 568-79.
69. Rivera, J., et al., *REDfly: the transcriptional regulatory element database for Drosophila.* Nucleic Acids Res, 2019. **47**(D1): p. D828-D834.
70. Keranen, S.V.E., et al., *REDfly: An Integrated Knowledgebase for Insect Regulatory Genomics.* Insects, 2022. **13**(7).
71. WJ, K., et al., *The human genome browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.
72. Smit, A., R. Hubley, and P. Green, *RepeatMasker Open-4.0.* 2013-2015.
73. Ruiz, J.L., L.C. Ranford-Cartwright, and E. Gomez-Diaz, *The regulatory genome of the malaria vector Anopheles gambiae: integrating chromatin accessibility and gene expression.* NAR Genom Bioinform, 2021. **3**(1): p. lqaa113.
74. Ingham, V.A., S. Wagstaff, and H. Ranson, *Transcriptomic meta-signatures identified in Anopheles gambiae populations reveal previously undetected insecticide resistance mechanisms.* Nat Commun, 2018. **9**(1): p. 5282.
75. Ingham, V.A., et al., *Dissecting the organ specificity of insecticide resistance candidate genes in Anopheles gambiae: known and novel candidate genes.* BMC Genomics, 2014. **15**: p. 1018.
76. Kefi, M., et al., *Transcriptomic analysis of resistance and short-term induction response to pyrethroids, in Anopheles coluzzii legs.* BMC Genomics, 2021. **22**(1): p. 891.

738 77. Pignatelli, P., et al., *The Anopheles gambiae ATP-binding cassette transporter family:*
739 *phylogenetic analysis and tissue localization provide clues on function and role in*
740 *insecticide resistance.* Insect Mol Biol, 2018. **27**(1): p. 110-122.
741 78. Mohanty, A.K., et al., *Proteome data of female Anopheles stephensi antennae.* Data
742 Brief, 2019. **24**: p. 103911.
743 79. Balabanidou, V., et al., *Cytochrome P450 associated with insecticide resistance*
744 *catalyzes cuticular hydrocarbon production in Anopheles gambiae.* Proc Natl Acad Sci
745 U S A, 2016. **113**(33): p. 9268-73.
746 80. Grigoraki, L., et al., *Isolation and transcriptomic analysis of Anopheles gambiae*
747 *oenocytes enables the delineation of hydrocarbon biosynthesis.* Elife, 2020. **9**.
748 81. Ruiz, J.L., et al., *Chromatin changes in Anopheles gambiae induced by Plasmodium*
749 *falciparum infection.* Epigenetics Chromatin, 2019. **12**(1): p. 5.
750 82. Kazemian, M., et al., *Evidence for deep regulatory similarities in early developmental*
751 *programs across highly diverged insects.* Genome Biol and Evol, 2014. **6**(9): p. 2301-
752 20.
753 83. Quiver, M.H. and J. Lachance, *Adaptive eQTLs reveal the evolutionary impacts of*
754 *pleiotropy and tissue-specificity while contributing to health and disease.* HGG Adv,
755 2022. **3**(1): p. 100083.
756

757 **Figure 1: Crossing, DNA and RNA extraction schema**



758

759

760

761 Figure 1 legend:

762 a. 13 Kisumu females (blue) were crossed to 13 Nagongera males (red). Females mate
763 only once. Following mating, genomic DNA from Nagongera males was extracted and
764 sequenced.

765 b. In the reciprocal cross, 13 Nagongera females (red) were crossed to 13 Kisumu males
766 (blue). Following mating, genomic DNA was extracted and sequenced from Kisumu males.

767 c. Individual mated Kisumu females were transferred to laying cups. Following egg
768 laying, genomic DNA was extracted and sequenced from three of these

769 d. Individual mated Nagongera females were transferred to laying cups. Following egg
770 laying, genomic DNA was extracted and sequenced from three of these.

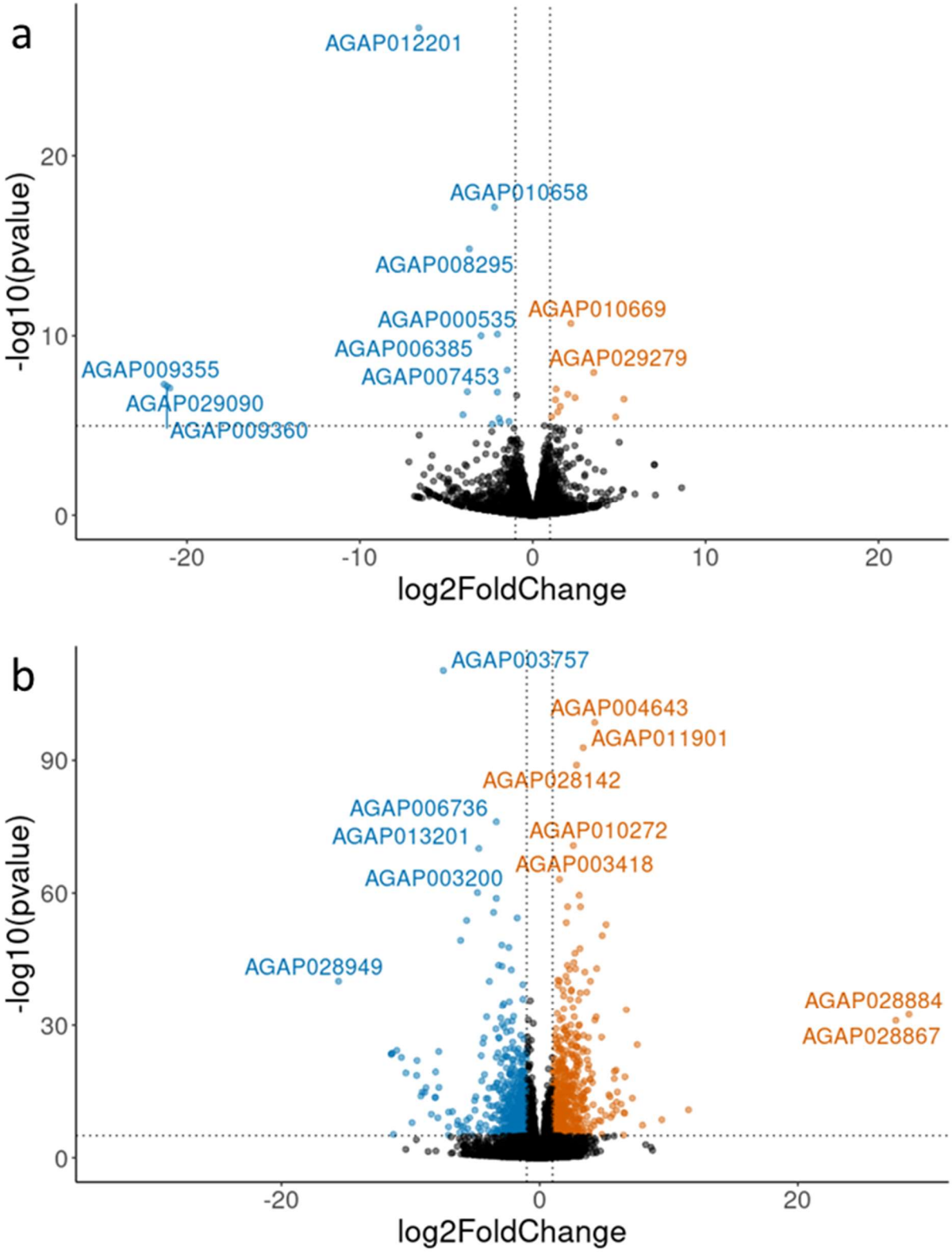
771 e. The F1 progeny (purple) from the three Kisumu mothers sequenced at step c. were
772 raised to adulthood. Genomic DNA was extracted and sequenced from individual males

773 f. The F1 progeny (purple) from the three Nagongera mothers sequenced at step d.
774 were raised to adulthood. Genomic DNA was extracted and sequenced from
775 individual males

776 g. and h. Female F1 from each of the six mothers were raised to adulthood. RNA was
777 extracted and sequenced from a pool of ten F1 females three to five days after eclosion.

778

779 **Figure 2: Reciprocal crosses show similar overall gene expression**



780

781 Figure 2 Legend

a. Volcano plot of log2 fold change against -log10 P value comparing gene expression in the F1 progeny of reciprocal crosses between Kisumu and Nagongera strains. Blue points indicate genes downregulated in progeny of Kisumu mothers compared to progeny of Nagongera mothers, orange points indicate genes upregulated in progeny of Kisumu mothers compared to progeny of Nagongera mothers.
b. Volcano plot of log2 fold change against -log10 P value comparing gene expression between F1 progeny of Nagongera and Kisumu with the Kisumu parental strain. Blue points indicate genes downregulated in the Kisumu compared to cross progeny, orange points indicate genes upregulated in Kisumu compared to cross progeny.

Figure 3: ASE in progeny of crosses between strains

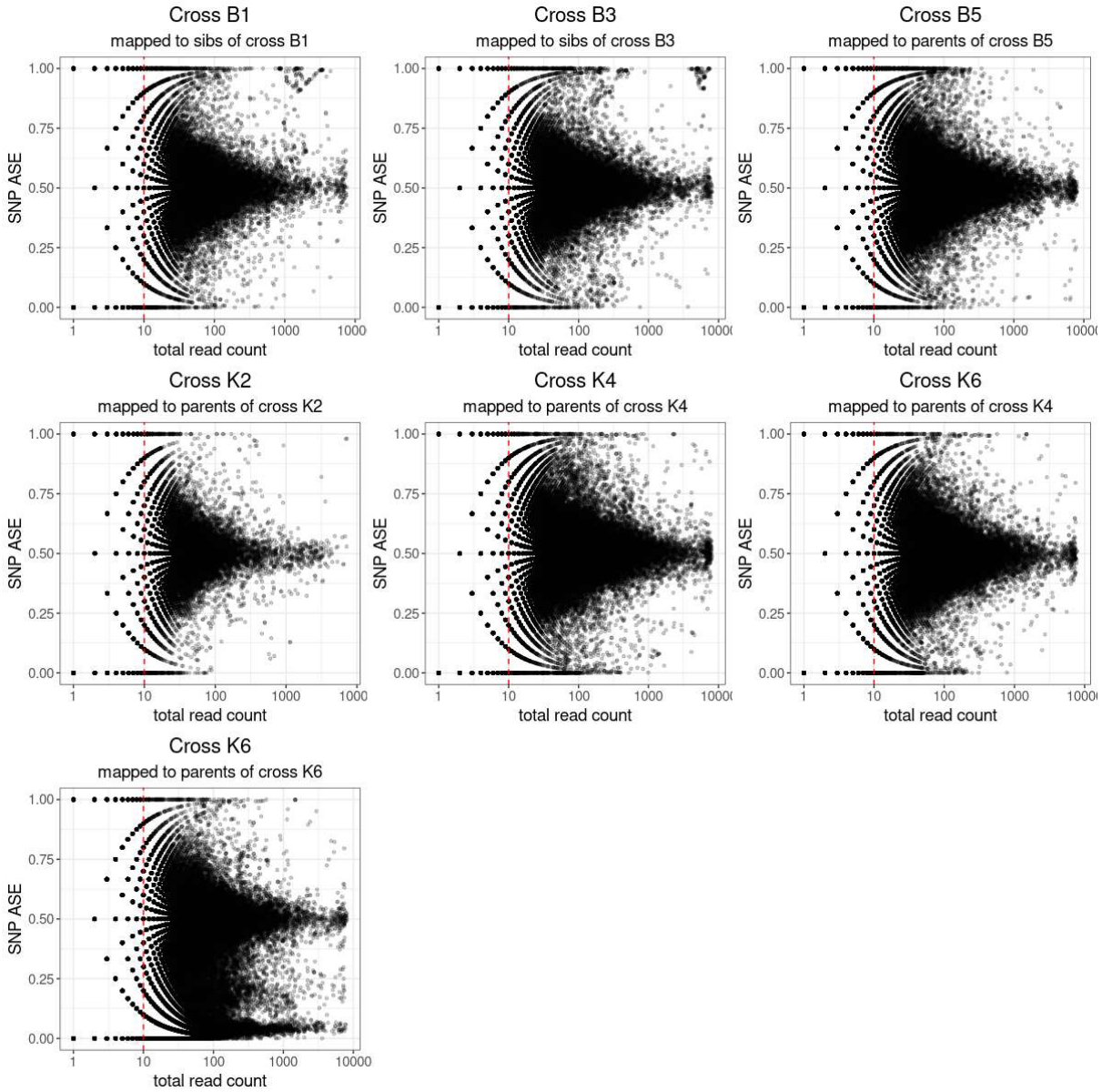
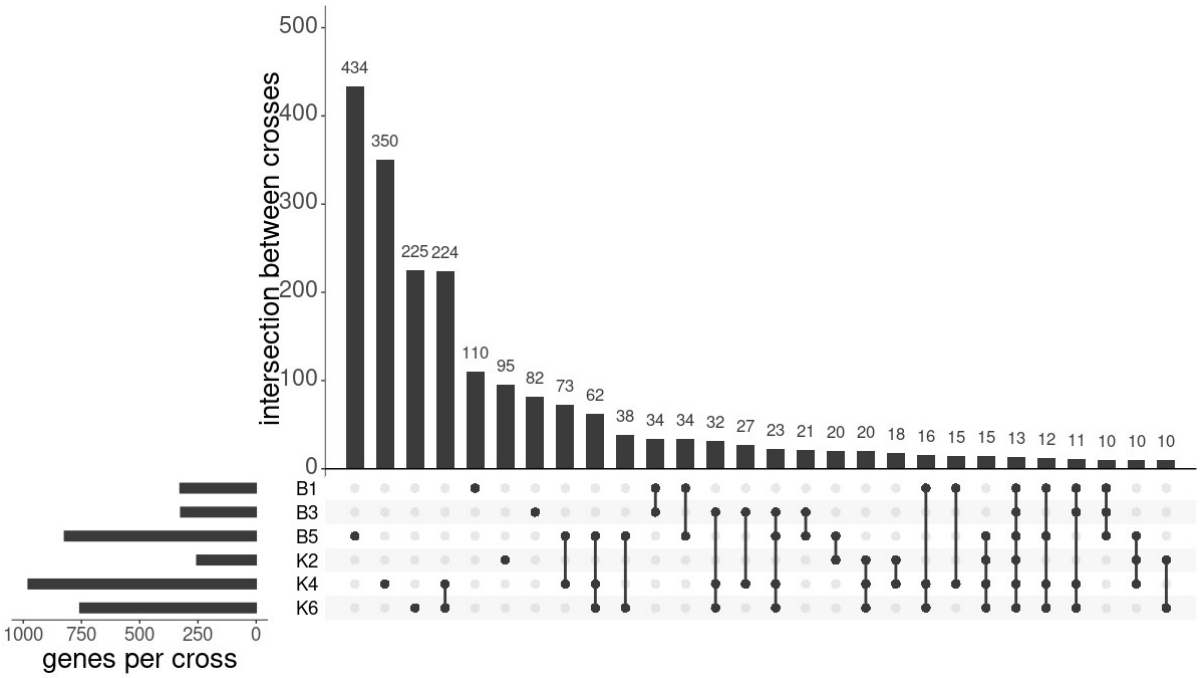


Figure 3 legend:

Plot of total read count against ASE at each SNP. Cross and source for the SNPs used to map and count reads are indicated at the top of each plot. Cross K6 is shown with ASE inferred using SNPs from both the parents of cross K4 and for the initially assumed parents of cross K6.

800 **Figure 4: Intersection of genes showing ASE between crosses**



801

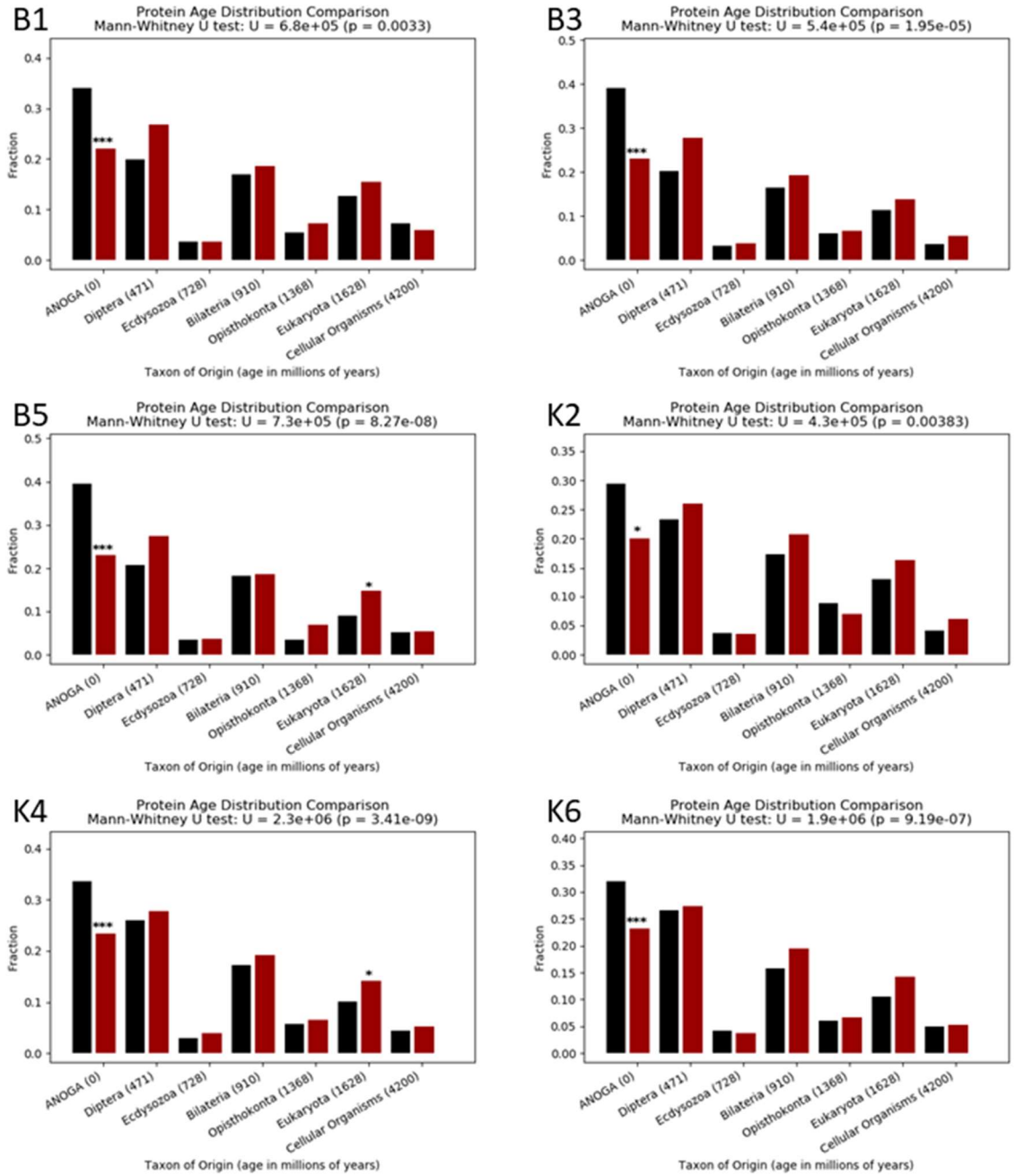
802 Figure 4 legend

803 UpSet plot with the number of genes showing ASE for each cross and the intersection of

804 these genes between crosses. Only the first 28 sets of overlaps are shown.

805

806 **Figure 5: Ages of genes showing ASE or not showing ASE**



807

808 Figure 5 legend, bar plots comparing the age of genes which showed ASE or did not in the

809 progeny of six crosses between Nagongera and Kisumu strains, using the Wagner parsimony

810 method. The cross name is indicated at the top left of each plot, Fisher's exact test P values

811 for the difference in fraction of genes in each age between genes showing ASE (red bars)

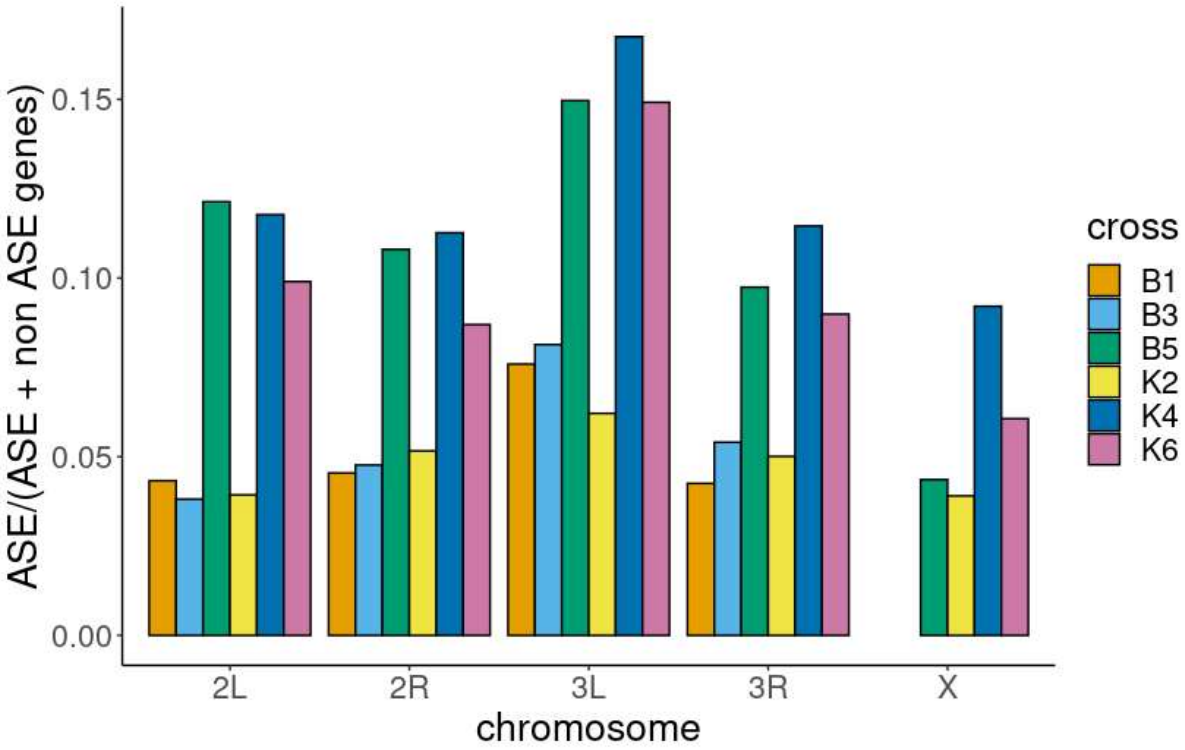
812 and genes not showing ASE (black bars) are displayed above the bars; *: 0.001<P<0.05,

813 ***:P≤0.001.

814

815 **Figure 6: proportion of genes showing ASE versus genes with**

816 **sufficient SNPs to detect ASE along each chromosome arm**



817

818

819 Figure 6 legend:

820 Bar plot of the proportion of genes showing allele specific expression/ total genes

821 containing suitable SNPs to detect ASE per chromosome arm. Colours indicate the different

822 crosses. X chromosomal SNPs were not available for crosses B1 and B3

823

824

825 **Table 1: Crosses between Kisumu and Nagongera and RNAseq summary**

826 **statistics**

RNAseq sample name	Cross name	Mother origin	Mother-Sanger code	Father origin	Father – Sanger code	Total reads counted (F1 RNAseq)	Percent of reads mapping to *Anopheles gambiae* PEST genome
Wilding_1	B1	Nagongera	failed QC	Kisumu	unknown	18,496,263	88.6%
Wilding_2	B3	Nagongera	failed QC	Kisumu	unknown	38,526,717	87.1%
Wilding_3	B5	Nagongera	AC0382-C	Kisumu	AC0416-C	28,969,694	87.7%
Wilding_4	K2	Kisumu	AC0300-C	Nagongera	AC0406-C	6,400,460	85.5%
Wilding_5	K4	Kisumu	AC0317-C	Nagongera	AC0398-C	46,229,505	86.3%
Wilding_6	K6	Kisumu	AC0334-C	Nagongera	AC0398-C	29,491,773	86.9%

827

828 **Table 2: Number of SNPs in exonic regions differing between the parents that**

829 **could be used to detect ASE**

Cross	2L	2R	3L	3R	X	Y / other	total
K2	13849	30134	19375	20877	15543	0	99778
K4	33139	35048	19074	22405	15250	0	124916
K6	30303	30811	14534	17801	15601	2	109052
B5	29627	36203	20865	25414	10539	1	122649

830

831 2L, 2R, 3L, 3R and X are the chromosome arms. The numbers are the number of SNPs in

832 exonic regions that are homozygous in the parents of each cross but differ between the

833 parents. These sites should be heterozygous in all cross progeny.

834

835 **Table 3: Number of SNPs differing between Kisumu and Nagongera colony**

836 **consensus sequences**

Chromosome	SNPs
2L	2575
2R	13507
3L	6216
3R	2949
X	10552
Total	35799

837

838 "SNPs" shows the number of SNPs differing across all sites in the consensus genome

839 sequences of Kisumu and Nagongera colonies

840

841

842 **Table 4: Number of genes showing ASE in each cross**

Cross	Genes showing ASE	Total genes with SNP(s) and sufficient counts to detect ASE
B1	327	6486
B3	324	6065
B5	823	7537
K2	255	5185
K4	979	8166
K6	757	7852
All (intersection)	13	2934

843

844 ASE detected with FDR of 5%

845 The totals for B1 and B3 are after the removal of SNPs which were inferred from siblings and

846 not from parents, which had >100 counts and ASE < .25 or > .75

847

848

849

850

851

852

853 **Table 5: Genes showing ASE in at least four crosses**

ID	chr	Gene Name	Description	overlap	het ASE	Freq CNV	Gene ASE in cross			
							B5	K2	K4	K6
AGAP001002	X		Toll protein	4	0	NA	0.00	1.00	0.96	0.98
AGAP001190	2R		female reproductive tract protease GLEANR_896	6	0	18	0.92	0.06	0.22	0.12
AGAP001243	2R			4	4	NA	0.38	0.03	0.28	0.28
AGAP001251	2R		eupolytin	5	0	3	0.24	NA	0.83	0.92
AGAP001356	2R	ACE1	Acetylcholinesterase	5	2	15	0.52	0.35	0.22	0.27
AGAP001365	2R			4	0	15	0.57	0.44	0.38	0.32
AGAP001594	2R			4	4	1	0.45	0.63	0.71	0.75
AGAP001726	2R			4	0	4	0.75	0.47	0.23	0.31
AGAP001971	2R		Polyubiquitin-C	4	5	11	0.55	0.59	0.64	0.69
AGAP002008	2R			4	0	NA	0.17	0.80	0.76	0.79
AGAP002058	2R		beta-galactosidase	4	0	3	0.14	0.83	0.82	0.93
AGAP002728	2R		Isocitrate dehydrogenase [NAD] subunit, mitochondrial	4	0	5	0.60	0.38	0.33	0.36
AGAP003364	2R			4	0	NA	0.61	0.44	0.33	0.36
AGAP003417	2R			4	0	NA	0.41	0.53	0.67	0.62
AGAP003444	2R			4	2	22	0.27	0.62	0.84	0.78
AGAP003474	2R			6	0	NA	0.49	0.63	0.61	0.63
AGAP003514	2R			4	3	NA	0.33	0.84	0.72	0.66
AGAP003525	2R		CCR4-NOT transcription complex, subunit 2	4	2	2	0.49	0.72	0.66	0.64
AGAP003583	2R		L-iditol 2-dehydrogenase	4	0	8	0.30	NA	0.32	0.06
AGAP003657	2R			4	0	NA	0.08	0.76	0.67	0.64
AGAP003663	2R		ATP-dependent RNA helicase DBP2	4	3	NA	0.43	0.62	0.60	0.58
AGAP003669	2R		homeobox protein MSX	4	2	NA	0.33	0.58	0.68	0.64
AGAP003686	2R			4	0	1	0.40	0.58	0.62	0.66
AGAP003701	2R		zinc finger protein 593 homolog	4	0	62	NA	NA	0.14	0.34
AGAP003722	2R	ANXB10C	annexin B10C	4	3	6	NA	NA	0.73	0.71
AGAP003728	2R		Uncharacterized methyltransferase WBSCR22	4	0	1	0.34	0.53	0.72	0.69
AGAP003731	2R		vacuolar protein sorting-associated protein 37B	4	4	1	0.49	0.57	0.74	0.77
AGAP003802	2R			4	3	NA	0.37	0.62	0.60	0.53
AGAP003861	2R		Peptide deformylase	4	0	1	0.34	0.61	0.73	0.73
AGAP004102	2R			4	0	1	0.36	0.49	0.65	0.66

AGAP004109	2R		Xaa-Pro aminopeptidase	6	6	NA	0.74	0.28	0.37	0.40
AGAP004110	2R			6	3	NA	0.79	0.23	0.23	0.25
AGAP004113	2R	mRpL55	39S ribosomal protein L55, mitochondrial	5	0	4	0.75	0.34	0.36	0.36
AGAP004133	2R			4	1	35	0.35	0.97	0.63	0.75
AGAP004233	2R		transcription factor IIIB 90 kDa subunit	6	1	NA	0.19	0.76	0.74	0.76
AGAP004534	2R		cathepsin B precursor	4	0	NA	0.87	NA	0.21	0.23
AGAP004562	2R		MFS transporter, PCFT/HCP family, solute carrier family 46 (folate transporter)	4	0	NA	0.86	0.49	0.14	0.27
AGAP004790	2L		Up-regulated during skeletal muscle growth 5 homolog	4	0	4	0.05	NA	0.79	0.79
AGAP004891	2L		metallophosphoesterase domain-containing protein 1	4	2	NA	NA	NA	0.70	0.65
AGAP005012	2L		galactokinase	4	0	30	0.92	NA	0.21	0.18
AGAP005243	2L			4	2	NA	0.37	0.67	0.47	0.44
AGAP005548	2L		alpha-crystallin B chain	4	1	1	0.67	0.96	0.75	0.82
AGAP006690	2L		BTB (POZ) domain containing 9	4	0	NA	0.68	NA	0.62	0.70
AGAP006756	2L			4	0	NA	0.57	NA	0.74	0.68
AGAP006775	2L		glucosyl/glucuronosyl transferases	4	0	NA	0.87	NA	0.80	0.86
AGAP007159	2L		alpha-crystallin B chain	4	3	4	0.27	NA	0.17	0.13
AGAP007402	2L			4	1	33	0.61	0.90	0.87	0.86
AGAP007407	2L	CTLMA4	C-type lectin (CTL) - mannose binding	4	0	2	0.47	0.07	0.27	0.26
AGAP007413	2L			4	1	NA	0.56	0.76	0.64	0.60
AGAP007484	2L		Sugar transporter ERD6-like 4	5	0	NA	0.86	0.58	0.17	0.30
AGAP008019	3R	CYP12F3	cytochrome P450	4	0	14	0.78	NA	0.12	0.09
AGAP008020	3R	CYP12F2	cytochrome P450	4	0	46	0.87	0.47	0.21	0.27
AGAP008122	3R		U6 snRNA-associated Sm-like protein LSm7	4	2	NA	0.61	NA	0.40	0.40
AGAP008218	3R	CYP6Z2	cytochrome P450	4	1	1	0.89	NA	0.66	0.47
AGAP008325	3R		U3 small nucleolar RNA-associated protein 23	4	0	NA	0.58	0.29	0.27	0.30
AGAP008331	3R		WD repeat-containing protein 59	4	0	56	0.70	0.77	0.03	0.03
AGAP008551	3R		mothers against decapentaplegic homolog 1	4	0	NA	0.55	0.53	0.90	0.80

AGAP008849	3R		D-3-phosphoglycerate dehydrogenase	4	5	NA	0.54	0.48	0.48	0.46
AGAP009251	3R			4	2	26	0.86	NA	0.20	0.13
AGAP009418	3R			4	2	63	0.17	0.91	0.48	0.77
AGAP009518	3R			4	1	NA	0.47	0.41	0.45	0.41
AGAP009527	3R			5	0	6	NA	0.38	0.37	0.34
AGAP009585	3R			5	0	41	0.23	NA	0.07	0.28
AGAP009782	3R		methylosome protein 50	4	3	10	0.28	0.55	1.00	1.00
AGAP009791	3R	Nep1	neprilysin, neutral endopeptidase 1	4	0	NA	0.32	0.66	0.69	0.63
AGAP010658	3L			5	0	1	0.30	NA	0.94	0.85
AGAP010814	3L	TEP6	thioester-containing protein 6	5	5	14	0.73	0.30	0.68	0.64
AGAP010873	3L		Cytosolic Fe-S cluster assembly factor NUBP2 homolog	5	3	NA	0.56	0.06	0.19	0.15
AGAP011065	3L			4	3	63	0.65	0.51	0.76	0.65
AGAP011091	3L			4	0	1	0.51	NA	0.70	0.72
AGAP011155	3L		dual specificity phosphatase	4	0	NA	0.77	0.27	0.32	0.32
AGAP011202	3L			4	0	14	0.48	0.58	0.60	0.60
AGAP011379	3L	GPRFZ1	frizzled receptor	4	4	NA	0.75	0.51	0.37	0.38
AGAP011499	3L			4	4	63	0.49	0.56	0.66	0.69
AGAP011616	3L			4	0	NA	0.40	0.55	0.69	0.78
AGAP011750	3L		alpha-N-acetylglucosaminidase	4	0	NA	0.70	0.36	0.28	0.35
AGAP011783	3L	CLIPA13	CLIP-domain serine protease	5	0	NA	0.38	0.91	0.73	0.74
AGAP011784	3L			4	0	NA	0.83	0.53	0.16	0.53
AGAP011792	3L	CLIPA7	CLIP-domain serine protease	4	0	NA	0.63	0.41	0.36	0.38
AGAP012008	3L			5	3	NA	0.83	0.23	0.78	0.86
AGAP012053	3L		alpha,alpha-trehalase	6	5	NA	0.68	0.33	0.42	0.38
AGAP012100	3L	RpS26	40S ribosomal protein S26	4	0	5	0.20	0.79	0.77	0.71
AGAP012118	3L		FGFR1 oncogene partner 2-like protein	5	4	NA	0.54	0.47	0.63	0.59
AGAP012126	3L		lysophospholipase II	5	4	NA	0.95	0.26	0.09	0.11
AGAP012130	3L	ATN	allatotropin	6	4	NA	0.89	0.03	0.05	0.07
AGAP012135	3L		26S proteasome regulatory subunit N8	5	3	8	0.35	0.55	0.65	0.64
AGAP012148	3L		Mitogen-activated protein kinase	6	6	NA	0.71	0.18	0.15	0.18
AGAP012152	3L			4	2	NA	0.72	0.29	0.24	0.27
AGAP012154	3L		solute carrier family 15 member 1	6	5	NA	0.75	0.18	0.29	0.25

AGAP012188	3L		ubiquinol-cytochrome c reductase subunit 7	5	3	4	0.68	0.23	0.30	0.42
AGAP012189	3L		abhydrolase domain-containing protein 12	6	3	NA	0.88	0.19	0.09	0.11
AGAP012217	3L		sterol O-acyltransferase	6	4	21	0.80	0.41	0.78	0.71
AGAP012219	3L	Med1	mediator of RNA polymerase II transcription subunit 1	5	4	1	0.41	0.40	0.32	0.30
AGAP012296	3L	CYP9J5	cytochrome P450	4	2	36	0.28	0.64	NA	NA
AGAP012305	3L		Muscle LIM protein at 84B	4	4	NA	0.37	0.66	NA	NA
AGAP012314	3L			4	4	2	0.23	NA	0.86	0.86
AGAP012984	2R		Protease m1 zinc metalloprotease	4	0	2	0.69	0.33	0.30	0.31
AGAP012996	2R		juvenile hormone-inducible protein	5	0	NA	0.38	0.63	0.59	0.60
AGAP013039	2R			4	0	NA	0.84	0.17	0.26	0.38
AGAP013188	2R	APN4	aminopeptidase 4	5	4	61	0.31	0.86	0.62	0.58
AGAP013255	2R	APN3	aminopeptidase N1	4	3	62	0.34	0.93	0.37	0.36
AGAP013359	2R			4	0	1	0.95	0.06	0.04	0.14
AGAP013393	2R	APN2	aminopeptidase N1	4	3	62	0.50	0.17	0.26	0.35
AGAP028064	3R			4	1	NA	0.25	NA	0.88	0.85
AGAP028157	3L			4	0	36	0.32	0.75	0.84	0.78
AGAP028209	3R			4	1	1	0.38	0.41	0.35	0.39
AGAP028542	3L			6	0	6	1.00	0.21	0.00	0.00
AGAP029122	3L			4	5	NA	0.40	0.49	0.68	0.62
AGAP029263	3L			4	4	4	0.76	0.17	0.25	0.29
AGAP029287	3L			5	4	NA	0.89	0.00	0.06	0.07
AGAP029362	2L			4	0	NA	0.19	0.71	0.80	0.63
AGAP029417	3R			4	1	37	0.91	0.06	0.12	0.12
AGAP029453	3R			5	4	NA	0.54	0.63	0.75	0.69
AGAP029526	2L			4	2	2	0.61	1.00	0.60	0.57
AGAP029597	3L			6	2	NA	0.96	0.15	0.03	0.03

Chrom: chromosome

ID: Vectorbase gene ID

Overlap: the number of crosses that showed significant ASE of the gene, when using SNPs from the siblings to infer allelic expression.

859 Expression biased to the maternal allele is highlighted in red and expression biased to the
860 paternal allele highlighted in blue; for cross B5 maternal bias means biased to Nagongera
861 whereas for K2, K4 and K6 crosses maternal bias means biased to Kisumu.
862 Freq CNV: how many of the 63 parents and siblings with CNV calls (crosses K2, K4, K6 and
863 B5) had a possible CNV covering that gene. NA indicates no CNV found in that gene.
864 Het ASE: the number of crosses which had significant heterogeneity in the MAF at SNPs
865 within that gene. 0 can indicate either no heterogeneity or only one SNP available to score
866 ASE.
867
868

869 **Table 6: training CRM**

Tissue	Anatomy ontology term	Date of search	Total Number of CRM <2000bp in Redfly at time of search	Number of CRM meeting criteria (non overlapping)
Adult Midgut	FBbt:00003138	12/12/22	12	11
Larval Midgut	FBbt:00005825	12/12/22	16	15
Adult Malpighian Tubules	FBbt:00005725	12/12/22	10	9
Leg	FBbt:00004640	12/12/22	35	28
Antennae	FBbt:00004511	24/04/23	25	20
Oenocyte	FBbt:00004995	25/04/23	3	2

870

Table 7: CRM predictions for genes that showed ASE in at least 4/6 crosses

chrom	Flanking gene 1	Name flanking gene 1	flanking_gene_2	Name flanking gene 2	CRM length	models	score	CRM_start	CRM_end	training_sets
2L	AGAP005012		AGAP005013		500	imm	7.12	8246250	8246750	Adult Malpighian tubules
2L	AGAP005012		AGAP005013		500	imm	4.85	8288000	8288500	Adult midgut
2L	AGAP005244		AGAP005243		500	hexmcd	26.41	12797250	12797750	malpig.mapping1
2L	AGAP006690		AGAP006689		500	pac	0.39	36723500	36724000	Larval midgut
2L	AGAP006755	snRNP-U1-70K	AGAP006756		500	imm	17.54	37866500	37867000	emb-larv_excretory,malpig.mapping1
2R	AGAP001194		AGAP001190		500	imm	6.75	1190000	1190500	Adult Malpighian tubules
2R	AGAP001356	ACE1	AGAP001355		500	hexmcd	12.68	3467250	3467750	legs
2R	AGAP013039		AGAP001990		500	pac	0.35	13217500	13218000	Adult Malpighian tubules
2R	AGAP013039		AGAP001990		750	hexmcd, pac	11.86	13224000	13224750	Adult Malpighian tubules
2R	AGAP002007		AGAP002008		500	pac	0.36	13703500	13704000	Adult midgut
2R	AGAP002007		AGAP002008		500	pac	0.41	13710000	13710500	Adult pns
2R	AGAP002007		AGAP002008		500	pac	0.42	13711750	13712250	Adult pns
2R	AGAP002007		AGAP002008		500	hexmcd, imm, pac	16.10	13717000	13717500	Adult midgut, antenna
2R	AGAP002007		AGAP002008		500	pac	0.44	13721000	13721500	emb-larv excretory
2R	AGAP002007		AGAP002008		500	hexmcd, imm	14.42	13727000	13727500	legs
2R	AGAP002007		AGAP002008		500	hexmcd	13.10	13773750	13774250	Larval midgut
2R	AGAP002008		AGAP002007		500	pac	0.41	13804500	13805000	legs

2R	AGAP002008		AGAP002009		500	pac	0.40	13843500	13844000	legs
2R	AGAP003473		AGAP003474		500	hexmcd	10.65	38103750	38104250	Adult Malpighian tubules
2R	AGAP003663		AGAP013457		500	imm	5.17	41330500	41331000	Adult pns
2R	AGAP003669		AGAP003667		500	pac	0.39	41423000	41423500	Adult pns
2R	AGAP003669		AGAP003670		500	hexmcd	11.76	41493500	41494000	Adult pns
2R	AGAP003727		AGAP003728		500	hexmcd	12.09	42615500	42616000	Adult pns
2R	AGAP013081		AGAP004110		500	pac	0.33	50078500	50079000	Adult midgut
2R	AGAP004112		AGAP004113	mRpL55	500	pac	0.39	50141750	50142250	antenna
2R	AGAP004112		AGAP004113	mRpL55	500	pac	0.40	50151250	50151750	legs
2R	AGAP004112		AGAP004113	mRpL55	500	hexmcd	12.06	50160250	50160750	antenna
2R	AGAP004232		AGAP004233		500	pac	0.38	52891500	52892000	Larval midgut
2R	AGAP004232		AGAP004233		500	pac	0.34	52894750	52895250	Adult Malpighian tubules
2R	AGAP004232		AGAP004233		500	pac	0.34	52921250	52921750	Adult midgut
3L	AGAP010813	TEP18	AGAP010814	TEP6	500	hexmcd	13.38	11182250	11182750	Adult pns
3L	AGAP011064		AGAP011065		500	imm	5.20	16137250	16137750	Adult pns
3L	AGAP011379	GPRFZ1	AGAP011378		500	pac	0.36	22779500	22780000	Larval midgut
3L	AGAP011379	GPRFZ1	AGAP011378		500	pac	0.37	22796500	22797000	Larval midgut
3L	AGAP012008		AGAP012007		500	imm	10.24	36201750	36202250	Larval midgut
3L	AGAP012026		AGAP028157		500	imm	10.72	36626250	36626750	legs
3L	AGAP012026		AGAP028157		500	imm	10.90	36645500	36646000	antenna
3L	AGAP012026		AGAP028157		500	hexmcd	29.43	36652000	36652500	emb-larv excretory, legs, malpig.mapping1
3L	AGAP029616		AGAP012053		500	hexmcd	12.69	36968750	36969250	legs
3L	AGAP029616		AGAP012053		500	hexmcd, imm	29.45	37009250	37009750	emb-larv_excretory, legs, malpig.mapping1
3L	AGAP029122		AGAP012093		750	hexmcd, imm	35.98	37603750	37604500	emb-larv excretory, malpig.mapping1

3L	AGAP029122		AGAP012093		500	imm	11.19	37608750	37609250	antenna
3L	AGAP029612		AGAP012118		500	hexmcd	26.81	37824750	37825250	emb-larv_excretory ,malpig.mapping1
3L	AGAP029612		AGAP012118		500	imm	9.59	37830500	37831000	Larval midgut
3L	AGAP012313		AGAP012314		500	pac	0.45	40131000	40131500	malpig.mapping1
3L	AGAP012313		AGAP012314		500	hexmcd	11.57	40135250	40135750	antenna
3L	AGAP012313		AGAP012314		500	pac	0.44	40137750	40138250	malpig.mapping1
3L	AGAP012313		AGAP012314		500	pac	0.37	40140500	40141000	Larval midgut
3R	AGAP028064		AGAP028015		500	pac	0.38	333000	333500	Larval midgut
3R	AGAP008019	CYP12F3	AGAP008018	CYP12F4	500	imm	8.26	4322500	4323000	Adult midgut
3R	AGAP008551		AGAP008552	CYP4H27	500	imm	11.67	12539500	12540000	Adult midgut
3R	AGAP008553		AGAP029453		500	pac	0.40	12582000	12582500	Adult pns
3R	AGAP029453		AGAP008557		500	pac	0.37	12662750	12663250	Larval midgut
3R	AGAP029417		AGAP009416		500	pac	0.40	32426500	32427000	Adult pns
3R	AGAP009418		AGAP009417		500	pac	0.36	32603250	32603750	Adult midgut
3R	AGAP009518		AGAP029410		500	imm	17.77	35145750	35146250	emb-larv_excretory, malpig.mapping1
3R	AGAP009518		AGAP029410		500	pac	0.40	35148250	35148750	antenna
3R	AGAP009522		AGAP009527		500	hexmcd	20.04	35340750	35341250	Adult midgut
3R	AGAP009522		AGAP009527		500	hexmcd, imm	11.01	35348500	35349000	Larval midgut
3R	AGAP009527		AGAP009522		500	pac	0.37	35364000	35364500	Larval midgut
3R	AGAP009790	CPAP3-B	AGAP009791	Nep1	500	hexmcd	17.04	43310250	43310750	Adult midgut
3R	AGAP009790	CPAP3-B	AGAP009791	Nep1	500	imm	4.90	43317000	43317500	Adult midgut

The positions of the 62 predicted CRMs for genes that showed ASE in at least 4/6 crosses. The two closest genes flanking each CRM are shown, with the flanking gene that shoed ASE highlighted in green