Influence of genetic polymorphism on transcriptional enhancer activity in the malaria vector *Anopheles coluzzii*

Luisa Nardini 1,2*, Inge Holm 1,2*, Adrien Pain 1,2,3, Emmanuel Bischoff 1,2, Daryl M Gohl 4,5, Soumanaba Zongo 6, Wamdaogo M. Guelbeogo 6, N'Fale Sagnon 6, Kenneth D Vernick 1,2**, Michelle M Riehle 7**

1 Unit of Insect Vector Genetics and Genomics, Department of Parasites and Insect Vectors, Institut Pasteur, Paris, France
2 CNRS Unit of Evolutionary Genomics, Modeling, and Health (UMR2000), Institut Pasteur, Paris, France
3 Institut Pasteur Bioinformatics and Biostatistics Hub (C3BI), CNRS USR 3756, Institut Pasteur, Paris, France
4 University of Minnesota Genomics Center, Minneapolis, MN, USA
5 Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN, USA
6 Centre National de Recherche et de Formation sur le Paludisme (CNRFP), Ouagadougou, Burkina Faso
7 Department of Microbiology and Immunology, Medical College of Wisconsin, Milwaukee WI, USA

* These authors contributed equally to this work

** Equivalent corresponding authors
Email: MMR, mriehle@mcw.edu; KDV, kvernick@pasteur.fr

Author emails:
LN nardiniluisa1@gmail.com
IH inge.holm@pasteur.fr
AP adrien.pain@pasteur.fr
EB bischoff@pasteur.fr
DMG dmgohl@umn.edu
SZ zongosoumanaba@gmail.com
WMG guelbeogo.cnrfp@fasonet.bf
NS n.fale.cnlp@fasonet.bf
KDV kvernick@pasteur.fr
MMR mriehle@mcw.edu

ABSTRACT

Enhancers are cis-regulatory elements that control most of the developmental and spatial gene expression in eukaryotes. Genetic variation of enhancer sequences is known to influence phenotypes, but the effect of enhancer variation upon enhancer functional activity and downstream phenotypes has barely been examined in any species. In the African malaria vector, *Anopheles coluzzii*, we identified a pilot set of candidate enhancers in the proximity of genes relevant for immunity, insecticide resistance, and development. The candidate enhancers were functionally validated using luciferase reporter assays, and their activity was found to be essentially independent of their physical orientation, a typical property of enhancers. All of the enhancer intervals segregated genetically polymorphic alleles, which displayed significantly different levels of functional activity, and inactive null alleles were also observed. Deletion mutagenesis and functional testing revealed a modular structure of positive and negative regulatory elements within the tested enhancers. The enhancer alleles carry genetic polymorphisms that also segregate in wild *A. coluzzii* populations in West Africa, indicating that enhancer variants with likely phenotypic consequences are frequent in nature. These results demonstrate the feasibility of screening for naturally polymorphic *A. coluzzii* enhancers that underlie important aspects of malaria transmission and vector biology.

57 **INTRODUCTION**

58 Enhancers are short cis-acting regulatory elements in noncoding DNA that amplify
59 transcriptional levels of target genes by tens to hundreds fold over the basal level of core
60 promoter elements at the transcription start site. Core promoters are located proximal to
61 the transcription start site (TSS) and facilitate the binding of RNA polymerase and the
62 initiation of transcription[1-3]. Enhancers control transcriptional activity of target genes
63 through their interaction with activators and the promoter and are, in turn, responsible for
64 most regulated gene expression in the transcriptome. The precise mechanisms of enhancer
65 action is unknown and is an area of active study [4-6]. Enhancers can function at a distance
66 from target genes and independent of their physical orientation in the chromosome [7]. The
67 identities of enhancers and some of their interacting protein factors that lead to their
68 regulatory function have been described in well-studied model organisms, but enhancers
69 cannot be reliably predicted by sequence-based algorithms, and thus must be detected
70 directly by functional activity using reporter assays, or indirectly inferred using methods to
71 detect open or modified chromatin properties.

72

73 Sequence polymorphism within enhancers has been associated with phenotypic differences,
74 including predisposition to disease, as observed in diverse organisms [5,8-11]. Most of the
75 significant variants mapped in human genome-wide association studies (GWAS) are
76 noncoding [12], and at least 60-70% of significantly associated human GWAS single-nucleotide
77 point mutations (SNPs) lie within functional enhancers [8,13]. In cancer studies, the majority of
78 tumor-driving mutational changes are also thought to be in noncoding regulatory elements,
79 especially enhancers [14].

80

81 Genetic variation in enhancers can occur as SNPs, insertions and deletions (indels), and as
82 copy number variants [15-17]. Enhancer variation among individuals can underlie both
83 Mendelian and complex traits [7,18,19]. At the population level, positively-selected variation in
84 enhancers controlling key pathways likely plays an important role in differentiation and
85 evolution [20,21]. Indeed, some of the fastest-evolving parts of the human genome as
86 compared to other primates are functional embryonic enhancers related to central nervous
87 system development [22]. In another example, the vertebrate ZRS enhancer influencing limb
88 development displays strong conservation across a range of vertebrates, although in
89 advanced snakes where skeletal limb structures are absent, mutations render the enhancer
90 inactive [23]. Finally, stickleback fish display development of lateral pelvic spines in some
91 populations, which may be protective under certain predation regimes but may be an
92 adaptive liability in others, and spines have been independently lost in different populations.
93 Sequence differences in the Pitx1 enhancer among spined and spineless populations
94 correlate with the phenotype, and enhancer swaps restore or abolish spine development [24].
95 Work in Drosophila has shown that enhancer variation effects many aspects of morphology
96 including for example pigmentation[25], melanism [26] and larval tricome pattern and number[27].
97 Numerous Gal4 based enhancer trapping lines exist for Anopheles stephensi largely as a tool
98 for the study mosquito functional genomics and optimizing methods to drive transgene
99 expression[28]. Recent work in the African malaria vector Anopheles funestus, highlights the
100 profound effects of indel variation at a cis regulatory element in the enhancer/promoter
101 region of a cytochrome P450 and its association with insecticide resistance[29]. Thus,
102 relatively simple evolved sequence variation in enhancers can produce large phenotypic
103 shifts in the organism [30,31]. Despite these examples, the effect of genetic variation on
104 enhancers has barely been examined in any species, and to our knowledge, the functional

105 effect of variation on enhancer activity has not been systematically surveyed in any
106 organism.
107 Enhancer positional effects on gene expression are known to influence phenotype and can
108 lead to disease states. Viruses and mobile elements carrying an enhancer can place
109 surrounding genes under a different regulatory regime, as well as chromosome
110 translocations such as the human 'Philadelphia chromosome' and others that promote
111 chronic myeloid leukemia. In genetic vector control strategies, it could be important to know
112 the locations of enhancers that could cause unpredicted effects on transgene expression.
113
114 Enhancer discovery in mosquitoes is limited to a few previous studies using indirect methods
115 based on detection of chromatin properties to infer enhancer locations [32-34]. However,
116 almost nothing is known about mosquito enhancer genetic variation and influence on vector
117 phenotypes, and thus it is necessary to first develop the baseline criteria to distinguish and
118 study enhancers in *A. coluzzii*. To this end, here we analyze a pilot set of candidate
119 enhancers in close proximity to a small number of genes of interest in order to benchmark
120 the parameters needed for reliable enhancer discovery, validation, and determination of
121 polymorphism effects in *Anopheles*. We validate the pilot candidates as functional enhancers
122 using luciferase reporter assays, and measure the effects of genetic polymorphism on
123 enhancer activity. The results of the current report are a first step towards developing a
124 comprehensive genome-wide catalog of *Anopheles* enhancers, and biological studies to
125 characterize enhancer function in vector biology.

126 **RESULTS**

127 **Candidate enhancer selection**

128 The standard approach for enhancer detection is by functional testing using luciferase
129 reporter assays that directly measure enhancer activity, or by indirect methods such as ChIP-
130 seq, which can infer the presence of a subset of enhancers by correlation with chromatin
131 features. Here, we implemented for the first time in *Anopheles* a screen (Self-Transcribing
132 Active Regulatory Region sequencing, STARR-seq) that detects enhancers directly by a
133 functional reporter assay analogous to the luciferase reporter assay, but with the readout of
134 enhancer-dependent RNA transcript output measured as sequenced cDNA, rather than by
135 luciferase light output [35].

136

137 In order to identify candidate *A. coluzzii* enhancers, we examined our generated sequence
138 data from the functional screen in the vicinity of six selected genes using the Integrative
139 Genomics Viewer (IGV) [36]. Candidate enhancers were identified by using IGV to manually
140 search near the selected genes to detect intervals where coverage of the cDNA sequence
141 track, indicative of enhancer activity in *A. coluzzii* 4a3A cells (Figure 1, solid lines) was visibly
142 greater than the coverage of the plasmid sequence track, which is the internal baseline
143 control indicating background levels of the plasmid in 4a3A cells (Figure 1, dotted lines). The
144 target genes selected are involved in vector immunity: Krueppel-Like Factor 6/7 (KLF,
145 AGAP007038), Leucine-Rich Immune protein (LRIM1, AGAP006348); insecticide resistance:
146 Acetylcholinesterase (ACE1, AGAP001356), GABA-gated chloride channel subunit (Rdl,
147 AGAP006028); and developmental biology: LIM homeobox protein 2/9, ortholog of
148 Drosophila apterous FBgn0267978 (AP, AGAP008980), and Ovo, AGAP000114 (Table1). The
149 regulatory regions and enhancers of the latter two genes, apterous and Ovo, have been well

150 characterized for the Drosophila orthologs [37-39]. In addition, a negative control interval was
151 chosen as a size-matched interval located within intron 1-2 of the gene, homeobox protein
152 distal-less (DLX, AGAP007058) that has no visible divergence of cDNA and plasmid sequence
153 tracks by IGV examination, and thus no predicted enhancer function. The candidate
154 enhancer intervals are named according to the most proximal coding sequence.
155
156 **Functional validation of enhancer activity**
157 The predicted candidate enhancers predicted in Figure 1 were functionally tested to validate
158 the IGV-based predictions. The standard test for enhancer activity is by cloning the
159 candidate in a plasmid carrying a basal promoter and a luciferase reporter gene in an
160 episomal assay. An active enhancer will augment the rate of transcription from the basal
161 promoter, thus elevating the expression level of the luciferase gene. Luciferase expression is
162 measured by adding luciferase substrate to cell extract and detecting light output as relative
163 light units (RLU). Enhancer activity, if any, is measured as increased luciferase activation
164 above background.
165
166 Candidate amplicons from *A. coluzzii* (Table 1) were cloned into the firefly luciferase reporter
167 vector pGL-Gateway-DSCP, and co-transfected into 4a3A cells with the renilla luciferase
168 control vector pRL-ubi-63E. Firefly luciferase RLU measurements were corrected using the
169 renilla luciferase internal control values in the same well, and firefly/renilla RLU for the
170 experimental insert were statistically compared to the firefly/renilla mean value for the DLX
171 negative control, defined as the background level. At least one clone of each candidate
172 enhancer displayed luciferase activity levels above background ($p<0.005$), with activity
173 across candidates that varied from 2-fold to more than 20-fold over background (Figure 2).

174 These results indicate that the IGV-based predictions were accurate for all six predicted
175 candidates, and thus validate these genomic intervals as functional *A. coluzzii* enhancers in
176 the cell culture model. Further work will be required to determine enhancer activity in the
177 native chromosomal context. The current information provides the first benchmark criteria
178 that can be used to develop the definitions and methods for subsequent algorithmic
179 genome-wide detection of *A. coluzzii* enhancers.
180
181 **Screening for polymorphic alleles of validated enhancer intervals**
182 Having confirmed that all six predicted candidates are functional enhancers, we next wished
183 to identify genetically variable alleles for each enhancer interval and measure their
184 luciferase activity. For this, alleles of the enhancer intervals were amplified and sequenced
185 from *A. coluzzii* colonies initiated from the populations in Cameroon, Mali or Burkina Faso.
186 For each of the six enhancer intervals, at least two distinct genetic variants were chosen for
187 tests of enhancer activity. The goal was to identify and test a range of genetic variants, not
188 the mosquito colonies. Therefore, a given colony may or may not be represented for a given
189 enhancer, depending on the variation it segregates. The enhancer interval alleles were
190 cloned and sequenced, and neighbor joining (N-J) trees depict the evolutionary relatedness
191 and degree of sequence difference of the alleles (Figure 3). Complete sequences for all
192 tested enhancer interval alleles are presented in Supplementary File S1.
193
194 **Genetic alleles of validated enhancer intervals display distinct enhancer activity**
195 Luciferase activity was measured for all alleles to determine the effect of genetic variation
196 on differences in functional enhancer activity. For five of the six enhancers, alleles displayed
197 significantly different levels of enhancer activity (Figure 4). For a given enhancer, alleles with

198 the greatest difference in activity tended to be the most genetically different from each
199 other (see also Figure 3). For example, the alleles of the KLF interval cloned from colonies
200 Fd05 and Fd03 are the most closely related genetically, and these also do not display a
201 difference in luciferase activity as compared to the allele from colony Fd09. For two
202 enhancer intervals (LRIM1 and ACE1), at least one genetic variant displayed activity levels
203 that were not significantly different from background, which effectively represents a
204 naturally occurring functionally inactive null enhancer allele. Genetic variation segregating at
205 the enhancer of Ovo did not display a significant influence on luciferase activity, and the Ovo
206 enhancer appears to display the consistently highest luciferase activity over all alleles tested
207 for any of the six enhancer intervals. These results indicate that genetic alleles of validated
208 enhancer intervals can display significantly different levels of functional activity. It is
209 reasonable to expect that large observed differences in enhancer activity will be translated
210 into phenotypic differences in the organism, related to the functions of the target genes that
211 are regulated by the polymorphic enhancer alleles. This prediction will need to be tested in
212 manipulative experiments.
213
214 **Enhancer activity is essentially independent of physical orientation**
215 Enhancers tend to function independently of their physical orientation in the genome, which
216 is testable when the candidate is cloned in a luciferase reporter plasmid. For three of the
217 above validated enhancers, we recloned two alleles in both orientations in the reporter and
218 measured luciferase activity. For the KLF and AP enhancers, there was no detectable effect
219 of orientation (Figure 5). The LRIM1 enhancer displayed a weakly significant effect of
220 orientation for allele Fd05_#1 ($p=0.042$), although for both orientations of the LRIM1
221 enhancer the absolute activity values were lower than the other enhancers tested (indicated

222 by y-axis values in Figure 5), and thus the weak orientation difference for this one weak
223 allele is not robustly supported. Thus, most of the enhancer alleles tested displayed function
224 independent of their physical orientation with respect to the basal promoter.
225
226 **Enhancer** deletion mutagenesis **reveals a modular structure of positive and negative**
227 **regulatory elements**
228 To resolve the minimal portion of the enhancer interval that carries the enhancer function,
229 we carried out deletion mutagenesis for two different genetic alleles of the LRIM1 enhancer
230 interval, one allele with high enhancer activity and the other low. The LRIM1 enhancer was
231 used for proof of principle, and was chosen because it had alleles with distinct activity levels,
232 and the genetic variation was spread across the enhancer interval. The deletion derivatives
233 carried 50% or 25% of the length of the initial enhancer interval, reduced equally from both
234 ends. We tested the deletion clones for luciferase activity, along with the original undeleted
235 enhancer (Figure 6A). Surprisingly, for LRIM1 allele Fd03_#3, the 50% construct displayed
236 the highest luciferase activity, greater than either 100% or 25% constructs. This indicates
237 that the integral 100% Fd03_#3 allele carries negative regulators of enhancer function,
238 which were deleted in the 50% derivative to yield a derivative with elevated enhancer
239 activity. The 25% derivative of allele Fd03_#3 displays significantly lower activity than the
240 50% derivative, suggesting that positive regulators of enhancer function are located outside
241 the 25% derivative, but within the 50% derivative.
242
243 Deletion derivatives of LRIM1 allele Fd05_#1 display a pattern distinct from the Fd03_#3
244 allele. For Fd05_#1, each incrementally smaller derivative was more active. This result was
245 also surprising, because it indicates that a highly active core enhancer element within the

246 smallest 25% derivative is attenuated by negative regulators that are progressively removed
247 from 100% to 50% in length, and again from a 50% to 25% length interval. The deletion
248 results indicate that enhancer activity is not directly correlated with sequence length, that
249 there is a complex structure of functional elements and modifiers within the enhancer
250 interval, and that different alleles of the same enhancer are comprised of distinct
251 combinations of modular regulators that differentially influence transcription.
252
253 The density of variable sites between Fd03_#3 and Fd05_#1 varies across the interval, such
254 that there were 60 variable nucleotide sites in the integral 100% length alleles, 37 variable
255 sites in the 50% derivatives and 22 sites in the 25% derivatives (Figure 6B). Finally, it is
256 notable that the 25% derivative for allele Fd05_#1 displays activity levels indistinguishable
257 from the Fd03_#3 50% derivative (p=0.99), even though they are no more genetically similar
258 than the integral 100% enhancer sequences for both alleles (Figure 6C). This result highlights
259 the relative independence of enhancer functional level from primary sequence patterns,
260 unlike the fundamental dependence of protein coding gene function on the amino acid
261 primary sequence code, and the consequent requirement for identification of enhancers by
262 detecting functional activity. Testing of a large panel and manipulative experiments would
263 be required to identify consistent patterns of enhancer modular organization.
264
265 **Enhancer alleles segregate in the natural *Anopheles coluzzii* population**
266 To confirm that the genetic variation observed in the enhancer alleles was natural and not
267 an artifact of laboratory colonies, we compared sequence data for the six enhancer intervals
268 to genetic variation observed in wild *A. coluzzii* from whole genome sequence of the
269 *Anopheles* gambiae 1000 (Ag1000) Genomes Consortium[40]. The comparison indicates that

270 genetic variation is shared between the cloned *A. coluzzii* colony haplotypes used in
271 luciferase assays and the natural population (Figure 7). Representative short sequence
272 alignments are shown, and full-length alignments with larger numbers of wild mosquitoes
273 are presented in Supplementary File S2. Alignments are presented rather than phylogenetic
274 trees because the wild Ag1000 sequences were called for SNPs but not indels. Therefore all
275 Ag1000 sequences by default share the same indels as the PEST reference, and a tree would
276 artifactually make all wild sequences appear more similar to one another. This analysis
277 demonstrates that genetic variants within confirmed functional enhancer intervals,
278 associated with differential enhancer activity, segregate in nature and do not represent
279 variants unique to lab colonies. Natural segregation of variants associated with differential
280 enhancer function supports the interpretation that the differential function of enhancer
281 alleles (Figure 4) based on a modular structure of regulatory elements (Figure 6) likely result
282 from natural selection for distinct phenotypic outcomes of allelic enhancer function.

283 **DISCUSSION**

284 The development of a methodology for screening and evaluation of *Anopheles* enhancers is
285 an initial step towards a more comprehensive study of enhancers and their polymorphism
286 effects. The currently examined enhancers were chosen because they were located near
287 known functional genes, but further work will be required to determine the actual influence,
288 if any, of these enhancers upon the nearby genes, which is not known. Moreover, enhancer
289 function is also controlled on spatial and temporal scales, and understanding *Anopheles*
290 enhancers and their effects on phenotypes in detail will ultimately require incorporating this
291 information. Lastly, the ability to translate results in the cell culture system to the whole
292 organism will involve downstream work, but screening of enhancer candidates in a cell
293 culture system is an important first step in determining the significance of non-coding
294 variation to vector biology.

295

296 Here we identify and validate candidate enhancer noncoding regulatory elements in the
297 malaria vector, *A. coluzzii*. We show that naturally segregating genetic variation significantly
298 influences enhancer activity levels, which if replicated in vivo could result in differences in
299 biological function and ultimately mosquito phenotype. Some enhancer alleles display high
300 activity while others display little or no activity above background and are thus naturally
301 occurring enhancer null alleles. The enhancers also tend to display activity that is
302 independent of their physical orientation, a common property of enhancers [7]. A structure-
303 function study of two enhancer alleles by deletion mutagenesis revealed a complex modular
304 organization of positive and negative modifiers that modulate enhancer activity. The current
305 study provides proof of principle for the influence of enhancer genetic polymorphism for
306 enhancer functional activity levels. These results provide benchmark parameters that can

307 now be implemented to develop a comprehensive genome-wide *Anopheles* enhancer
308 catalog. The current work is thus a step toward the long-term goal to identify functionally
309 important transcription factor binding motifs and correlate enhancer output with
310 phenotypes related to *Anopheles* biology, immunity and pathogen transmission.
311
312 By modifying the level of enhancer activity, genetic variation in enhancers can cause
313 quantitative changes in expression of the target genes regulated by the enhancers. Altered
314 expression profiles of enhancer target genes probably in turn trigger distinct phenotypic
315 outcomes. Different from mutations in protein coding genes, enhancers are typically located
316 in noncoding DNA, and there is no sequence pattern to aid interpretation of noncoding
317 variants. Here, we identified enhancer intervals proximal to genes underlying the important
318 vector phenotypes of insecticide resistance, immunity, and development. Most mosquito
319 studies to date have focused solely on genes and proteins associated with these traits,
320 rather than regulatory elements controlling the genes, in part due to the limited information
321 available for the noncoding mosquito genome.
322
323 Enhancers were identified and individual genetic variants were tested for their activity by
324 means of standard luciferase reporter assays. Interestingly, we detected variant alleles with
325 significant difference in their functional enhancer activity, including functional null alleles
326 that lack enhancer activity above background. For example, the LRIM1 enhancer Fd05_#1
327 allele or the ACE1 Fd03_#1 allele likely represent the ablation of an important transcription
328 factor binding site, resulting in the absence of enhancer activity above background with
329 likely downstream functional consequences. The range in functional enhancer activity that
330 we observed is likely to affect phenotypes produced by the genes they transcriptionally

331 regulate. Moreover, we demonstrate that variant alleles tested by luciferase activity in
332 laboratory colonies also segregate in the wild population, and are therefore subject to
333 natural selection. Thus, it is intriguing that selection has apparently generated a wide range
334 of natural allelic forms of enhancers, including alleles that lack functional activity. This is
335 consistent with the observation that genetic variation for enhancer function offers powerful
336 raw material for adaptation and evolutionary change [11,20,23,24]. The members of the Gambiae
337 species complex, including *A. coluzzii*, are highly adaptable to a range of ecological
338 conditions, and durable to vector control measures. This is thought to be associated with
339 their high genetic diversity [40]. The current study addresses standing genetic variation and
340 enhancer function something that needs to be considered as a likely new contributing factor
341 influencing the success of this mosquito and its relatives.
342
343 We functionally dissected two alleles of the LRIM1 enhancer, high and low activity variants,
344 respectively, by deletion mutagenesis (Figure 6). By measuring functional activity of integral,
345 50% and 25% length derivatives of the intervals, we detected a modular structure of positive
346 and negative regulators comprising the enhancer. Interestingly, deletion derivatives of the
347 two alleles behaved differently, indicating that the deletions was not a simple consequence
348 of sequence length. The high-activity allele Fd03_#3 appears to carries a negative regulator
349 in the terminal one-quarter of its length on one or both ends, because removal of these
350 sequences led to significantly elevated activity in the remaining 50% derivative as compared
351 to the integral enhancer. However, removal of an additional one-quarter again of the
352 sequence from both ends of the 50% derivative then diminished activity to a level below
353 that of the integral enhancer, suggesting that the positive regulator(s) revealed in the 50%
354 length derivative was no longer present in the 25% length derivative.

355

356 That the low activity of the smallest derivative of Fd05_#1 was not a simple consequence of
357 sequence length is made clear by a similar examination of the low-activity allele Fd05_#1. In
358 this case, each incremental length decrease of the tested sequence led to increased
359 enhancer activity. The Fd05_#1 allele result suggests that the integral enhancer displayed
360 low activity because it carried multiple negative regulators, which were resected by each
361 successive deletion, revealing a highly active core enhancer element within the smallest
362 interval tested. This latter minimal derivative of the low-activity Fd05_#1 allele carries an
363 enhancer with, in fact, higher enhancer activity than the integral 100% sequence of the high-
364 activity Fd03_#3 allele.

365

366 The LRIM1 deletion mutagenesis results suggest that that large functional allelic diversity
367 can be generated for a given enhancer interval by the combinatorial effect of positive and
368 negative modifiers. Sequence changes in enhancers can generate or remove binding motifs
369 for transcription factors and other regulatory proteins, which can modify transcription levels
370 directly [41,42], or indirectly through loss of chromatin accessibility [17]. From the current results,
371 we do not know whether the positive and negative modifiers within the LRIM1 alleles are
372 comprised of reusable modular cassettes that are combined to fine-tune the activity of
373 different enhancer intervals, or whether segregating SNPs in an enhancer can explain
374 significant difference in functional activity. In the first model, such modifier modules should
375 be recognizable with enough data, while under the second model, different modifiers, for
376 example positive modifiers in different enhancers, may have little or no recognizable
377 common pattern. Fine resolution nucleotide-level deletion series of a number of alleles
378 would be required to determine the kind and extent of sequence difference necessary to

379 alter enhancer functional levels in order to distinguish between the above models. Finally,
380 the phenotypic implications of differentially active enhancer alleles will require
381 determination of target gene networks regulated by an enhancer, as well as the protein-DNA
382 interactions underlying differential enhancer allele activity.
383
384 Vector control has been central to the malaria control effort by use of indoor residual
385 spraying and long-lasting insecticide impregnated bednets. However, over-reliance on these
386 methods has led to widespread insecticide resistance in wild populations, and novel
387 methods of control are now required. The noncoding regulatory genome in *Anopheles* has
388 the potential to provide novel new targets for vector control, but until now has not been
389 interpretable or exploitable. For example, understanding the regulatory genome landscape
390 could improve the choice of insertion sites for exogenous transgenes for proper expression,
391 and enhancers themselves could be targeted in order to alter transcription of immune or
392 disease important genes. The current work presents a necessary first step towards
393 establishing an efficient, effective method for associating noncoding variation with
394 important mosquito phenotypes.

395 **METHODS**

396 **Wild mosquito samples and DNA library**

397 Mosquito larvae were collected in Goundry village, Burkina Faso (latitude 12.5166876,
398 longitude -1.3921092) using described methods [43], reared to adults, and were typed for
399 species by the SINE200 X6.1 assay [44]. DNA from 60 *A. coluzzii* were pooled at equal volume
400 and sheared using an S220 ultrasonicator (Covaris) to produce DNA fragments 800-1000 bp
401 in length. Subsequently, DNA was processed as described for the STARR-seq assay [35], cloned
402 into the plasmid pSTARR-seq_fly (AddGene 71499), transformed into MegaX DH10B T1R
403 Electrocomp Cells (Invitrogen), cultured in LB + ampicillin (1ug/ml), and plasmid DNA was
404 purified using the Plasmid Plus Mega Kit (Qiagen). The *Anopheles* gambiae PEST AgamP4
405 genome assembly available at Vectorbase was used as the reference genome
406 (https://www.vectorbase.org/organisms/anopheles-gambiae/pest/agamp4).

407

408 **Culture of plasmid library in *Anopheles* 4a3A cells**

409 Hemocyte-like 4a3A cells [45] were maintained on Insect X-Press media (Lonza) supplemented
410 with 10% Fetal Calf Serum (heat inactivated at 56°C for 30 minutes), at 27°C. No antibiotics
411 were used. We confirmed that cells were derived from *A. coluzzii* by species typing using the
412 Fanello assay [46]. The plasmid DNA library was transfected and cultured in 4a3A cells as
413 described [35] using Lipofectamine 3000 Reagent (Invitrogen) and cultured for 24 h, in three
414 biological replicates. RNA was extracted from cells using the RNeasy Midi Kit (Qiagen)
415 followed by mRNA purification using Dynabeads mRNA Purification Kit (ThermoFisher).
416 Plasmid DNA was isolated using the Plasmid Plus Midi or Mini Kit (Qiagen).

417

418 **Analysis of 4a3A library culture results**

419 The mRNA purified from cells was reverse transcribed using SuperScript IV First-Strand cDNA
420 Synthesis System (Invitrogen) as described for the STARR-seq assay [35] using a plasmid-
421 specific primer (RT_Rev, Supplementary Table S1), the cDNA was then amplified using
422 primers Report_Fwd and Report_Rev (Supplementary Table S1), and the products were
423 sequenced on an Illumina HiSeq 2500 in 2x125 bp high output mode. Cell plasmid DNA was
424 amplified and sequenced in the same manner as the cDNA samples but using primers
425 Plasmid_Fwd and Plasmid_Rev (Supplementary Table S1).
426
427 **Selection of enhancer candidates**
428 The Integrative Genomics Viewer (IGV) [36] was used to select candidate enhancer intervals by
429 visual examination in the proximity of six annotated genes of interest. For determination of
430 enhancer activity, the RNA output transcribed from the STARR-seq reporter plasmid,
431 converted to cDNA and sequenced as described above, is compared to the levels of the
432 plasmid DNA, to control for differential plasmid replication. Thus, candidate enhancers were
433 predicted in intervals where coverage of the cDNA sequence track was visibly greater than
434 the baseline coverage of the plasmid sequence track. The target genes examined were
435 Krueppel-Like Factor 6/7 (KLF, AGAP007038), Leucine-Rich Immune protein (LRIM1,
436 AGAP006348), Acetylcholinesterase (ACE1, AGAP001356), GABA-gated chloride channel
437 subunit (Rdl, AGAP006028), LIM homeobox protein 2/9, ortholog of Drosophila apterous
438 FBgn0267978 (AP, AGAP008980), and Ovo, AGAP000114. In addition, a negative control
439 interval was cloned, which was a size-matched interval located within intron 1 of the gene,
440 homeobox protein distal-less (DLX, AGAP007058) that displayed no visible divergence of
441 cDNA and plasmid sequence tracks by IGV examination, and thus no predicted enhancer
442 function. In all but one case, the candidate enhancer was the only one in the vicinity of the

443 target gene, for AP there were three peaks, the most gene proximal peak is likely a promoter
444 so the next most proximal peak was chosen as shown in Figure 1. The candidate enhancer
445 intervals are named according to the most proximal coding sequence (above and Table 1).
446
447 Candidate enhancers were amplified from DNA of mosquitoes from the following *A. coluzzii*
448 colonies: Ngousso, initiated in Cameroon in 2006 [47], Fd03, Mali, 2008, Fd05, Mali 2008, Fd09,
449 Burkina Faso, 2008, and Fd33, Burkina Faso, 2014. Fd colonies were previously described [48].
450 Primers are listed in Supplementary Table S2. Amplicons were cloned into the
451 pCR8/GW/TOPO vector (Invitrogen) and sequenced with GW1 and GW2 primers. A standard
452 plasmid was used for luciferase assays that incorporates the engineered Drosophila
453 synthetic core promoter (DSCP). The information about the fine structure of Anopheles core
454 promoters does not yet exist. The results indicate that the DSCP is functional in Anopheles
455 cells. At least two genetically distinct sequences per candidate were then cloned into the
456 firefly luciferase reporter plasmid pGL-Gateway-DSCP (AddGene 71506) using Gateway LR
457 Clonase II (Invitrogen), transformed into OneShot OmniMax 2T1 Phage-Resistant Cells
458 (Invitrogen), and plasmid was purified from overnight culture.
459
460 To test the effect of enhancer orientation, the enhancer was cloned in the opposite
461 orientation in pGL-Gateway-DSCP. To test resolved enhancer intervals, the relevant
462 enhancer insert was amplified with primers that generated either 50% or 25% of the initial
463 insert size, equally reduced on both ends, and products were cloned in pGL-Gateway-DSCP.
464 In all cases, plasmids were resequenced to confirm insert identity using the primers LucNrev
465 and RVprimer3 (Supplementary Table S1).
466

467 **Quantitation and statistical analysis of enhancer activity by luciferase assay**

468 The Dual-Glo Luciferase Assay System (Promega) was used for luciferase assays. *A. coluzzii*
469 4a3A cells were seeded in 96 well plates at $1x10^5$cells/well, the difference in volume if any
470 was made up to 65 µl with medium, and cells were incubated for 24 h at 27°C. Two plasmids
471 were transfected into the 4a3A cells, the enhancer candidate in firefly luciferase vector pGL-
472 Gateway-DSCP, and the renilla luciferase control vector pRL-ubi-63E (AddGene 74280), at a
473 ratio of 1:5 (renilla:firefly), using transfection reagent Lipofectamine 3000 (Invitrogen), and
474 were then incubated for 24 h at 27°C.

475

476 Luciferase activity was detected on a GloMax Discover instrument (Promega) at 25°C, with
477 two 20 min incubations, one after the addition of Dual-Glo Luciferase reagent (Promega) and
478 another after the addition of Stop & Glo reagent (Promega). All samples were run in 6-fold
479 replication within a single plate and across at least two independent plates, for at least two
480 biological replicates of each candidate, yielding at least 12 measurements. Firefly luciferase
481 measurements expressed in relative light units (RLU) were corrected using the
482 measurements of RLU for the renilla luciferase internal control in the same well. Values for
483 the DLX negative control on the same plate were defined as the background level. Values of
484 firefly/renilla RLU for the experimental insert were normalized to the firefly/renilla mean
485 value for DLX in order to combine results across replicates. Luciferase activity was declared
486 above background if the firefly/renilla RLU ratio for the experimental insert was significantly
487 higher than the firefly/renilla value for the DLX negative control. Luciferase activity was
488 statistically compared using a non-parametric ANOVA (Kruskal-Wallis) with post hoc pairwise
489 comparisons.

490

491 **Analysis of enhancer allelic variants**

492 The sequences of genetically polymorphic variants of a given enhancer interval, cloned from
493 *A. coluzzii* colonies as described above, were analyzed for genetic relatedness. To generate
494 neighbor joining (N-J) trees to depict the relationships between genetic variants for the
495 same enhancer, complete sequences were aligned using MUSCLE within the package
496 Molecular Evolutionary Genetics Analysis Mega version X [49], and N-J trees constructed using
497 Mega. When at least four variants were tested, bootstrapping was performed and bootstrap
498 values are included on N-J trees. Scale bars of trees represent 0.5, but each bar is a different
499 length. The longer the 0.5 scale bar, the more genetically similar the sequences.

500

501 **Analysis of wild *Anopheles* variation data**

502 Sequence information for 309 wild *A. coluzzii* from 6 West African countries; Angola (AR),
503 Burkina Faso (AB), Cote d'Ivoire (AY), Ghana (AA), Guinea (AV) and Guinea-Bissau (AJ),
504 generated as reported [40] were downloaded from MalariaGen
505 (https://www.malariagen.net/projects/ag1000g) as VCF files from the Ag1000G phase 2 AR1
506 data release and sequences of the six validated enhancer intervals were extracted from the
507 raw VCF files using GATK version 3.9 (SelectVariants mode) [50], the consensus sequences with
508 IUPAC ambiguity codes for the variants were extracted using BCFtools version 1.9
509 (consensus mode with –iupac-codes option) [51], the sequences of each interval were aligned
510 with Clustal W version 2.1 [52] and visualized with AliView aligner version 1.24 [53].

511

512 The diploid sequences from the wild sequences were aligned to the cloned sequences
513 generated from the six validated enhancer intervals. Sequence alignments were visually
514 examined for shared variation. Short representative sequence alignments are presented in

515 Figure 7 (not including indels), and complete alignments relative to the PEST
516 AgamP3genome assembly, including indels, are presented in Supplementary File S2. Indel
517 genotypes of the wild sequences shown in Supplementary File S2 are relative to the PEST
518 reference haplotype, because the wild sequences were called for SNPs but not indels [40].

519 **REFERENCES**

520 1 Ong, C. T. & Corces, V. G. Enhancer function: new insights into the regulation of
521 tissue-specific gene expression. *Nat Rev Genet* **12**, 283-293, doi:10.1038/nrg2957
522 (2011).

523 2 Moreau, P. *et al.* The SV40 72 base repair repeat has a striking effect on gene
524 expression both in SV40 and other chimeric recombinants. *Nucleic acids research* **9**,
525 6047-6068, doi:10.1093/nar/9.22.6047 (1981).

526 3 Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of
527 transcription initiation. *Nat Rev Mol Cell Biol* **19**, 621-637, doi:10.1038/s41580-018-
528 0028-8 (2018).

529 4 Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase
530 Separation Model for Transcriptional Control. *Cell* **169**, 13-23,
531 doi:10.1016/j.cell.2017.02.007 (2017).

532 5 Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease
533 associations with regulatory information in the human genome. *Genome research* **22**,
534 1748-1759, doi:10.1101/gr.136127.111 (2012).

535 6 Chen, H. *et al.* Dynamic interplay between enhancer-promoter topology and gene
536 activity. *Nat Genet* **50**, 1296-1303, doi:10.1038/s41588-018-0175-z (2018).

537 7 Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers:
538 five essential questions. *Nat Rev Genet* **14**, 288-295, doi:10.1038/nrg3458 (2013).

539 8 Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease
540 variants. *Nature*, doi:10.1038/nature13835 (2014).

541 9 Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in
542 Drosophila melanogaster. *Nature* **471**, 480-485 (2011).

543 10 Romanoski, C. E., Link, V. M., Heinz, S. & Glass, C. K. Exploiting genomics and natural
544 genetic variation to decode macrophage enhancers. *Trends Immunol* **36**, 507-518
545 (2015).

546 11 Sicard, A. *et al.* Standing genetic variation in a tissue-specific enhancer underlies
547 selfing-syndrome evolution in Capsella. *Proceedings of the National Academy of*
548 *Sciences of the United States of America* **113**, 13911-13916 (2016).

549 12 MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide
550 association studies (GWAS Catalog). *Nucleic acids research* **45**, D896-D901,
551 doi:10.1093/nar/gkw1133 (2017).

552 13 Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor
553 footprints. *Nature* **489**, 83-90, doi:10.1038/nature11212 (2012).

554 14 Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of
555 noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-1165,
556 doi:10.1038/ng.3101 (2014).

557 15 Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function.
558 *Nature* **503**, 487-492, doi:10.1038/nature12615 (2013).

559 16 Capellini, T. D. *et al.* Ancient selection for derived alleles at a GDF5 enhancer
560 influencing human growth and osteoarthritis risk. *Nat Genet* **49**, 1202-1210,
561 doi:10.1038/ng.3911 (2017).

562 17 Jacobs, J. *et al.* The transcription factor Grainy head primes epithelial enhancers for
563 spatiotemporal activation by displacing nucleosomes. *Nat Genet* **50**, 1011-1020,
564 doi:10.1038/s41588-018-0140-x (2018).

565 18 Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. & Shiroishi, T. Elimination of a long-
566 range cis-regulatory module causes complete loss of limb-specific Shh expression and
567 truncation of the mouse limb. *Development* **132**, 797-803, doi:10.1242/dev.01613
568 (2005).

569 19 Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital
570 heart disease. *Hum Mol Genet* **21**, 3255-3263, doi:10.1093/hmg/dds165 (2012).

571 20 Arnold, C. D. *et al.* Quantitative genome-wide enhancer activity maps for five
572 Drosophila species show functional enhancer conservation and turnover during cis-
573 regulatory evolution. *Nat Genet* **46**, 685-692, doi:10.1038/ng.3009 (2014).

574 21 Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-
575 regulatory evolution. *Science* **346**, 1007-1012, doi:10.1126/science.1246426 (2014).

576 22 Franchini, L. F. & Pollard, K. S. Can a few non-coding mutations make a human brain?
577 *Bioessays* **37**, 1054-1061, doi:10.1002/bies.201500049 (2015).

578 23 Kvon, E. Z. *et al.* Progressive Loss of Function in a Limb Enhancer during Snake
579 Evolution. *Cell* **167**, 633-642 e611, doi:10.1016/j.cell.2016.09.028 (2016).

580 24 Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent
581 deletion of a Pitx1 enhancer. *Science* **327**, 302-305, doi:10.1126/science.1182213
582 (2010).

583 25 Jeong, S. *et al.* The evolution of gene regulation underlies a morphological difference
584 between two Drosophila sister species. *Cell* **132**, 783-793,
585 doi:10.1016/j.cell.2008.01.014 (2008).

586 26 Rebeiz, M., Pool, J. E., Kassner, V. A., Aquadro, C. F. & Carroll, S. B. Stepwise
587 modification of a modular enhancer underlies adaptation in a Drosophila population.
588 *Science* **326**, 1663-1667, doi:10.1126/science.1178357 (2009).

589 27 Frankel, N. *et al.* Morphological evolution caused by many subtle-effect substitutions
590 in regulatory DNA. *Nature* **474**, 598-603, doi:10.1038/nature10200 (2011).

591 28 O'Brochta, D. A., Pilitt, K. L., Harrell, R. A., 2nd, Aluvihare, C. & Alford, R. T. Gal4-
592 based enhancer-trapping in the malaria mosquito Anopheles stephensi. *G3* **2**, 1305-
593 1315, doi:10.1534/g3.112.003582 (2012).

594 29 Weedall, G. D. *et al.* A cytochrome P450 allele confers pyrethroid resistance on a
595 major African malaria vector, reducing insecticide-treated bednet efficacy. *Sci Transl*
596 *Med* **11**, doi:10.1126/scitranslmed.aat7386 (2019).

597 30 McGregor, A. P. *et al.* Morphological evolution through multiple cis-regulatory
598 mutations at a single gene. *Nature* **448**, 587-590, doi:10.1038/nature05988 (2007).

599 31 Prud'homme, B., Gompel, N. & Carroll, S. B. Emerging principles of regulatory
600 evolution. *Proceedings of the National Academy of Sciences of the United States of*
601 *America* **104 Suppl 1**, 8605-8612, doi:10.1073/pnas.0700488104 (2007).

602 32 Behura, S. K. *et al.* High-throughput cis-regulatory element discovery in the vector
603 mosquito Aedes aegypti. *BMC genomics* **17**, 341, doi:10.1186/s12864-016-2468-x
604 (2016).

605 33 Mysore, K., Li, P. & Duman-Scheel, M. Identification of Aedes aegypti cis-regulatory
606 elements that promote gene expression in olfactory receptor neurons of distantly
607 related dipteran insects. *Parasit Vectors* **11**, 406, doi:10.1186/s13071-018-2982-6
608 (2018).

609 34 Ruiz, J. L. *et al.* Chromatin changes in Anopheles gambiae induced by Plasmodium
610 falciparum infection. *Epigenetics Chromatin* **12**, 5, doi:10.1186/s13072-018-0250-9
611 (2019).

612 35 Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by
613 STARR-seq. *Science* **339**, 1074-1077, doi:10.1126/science.1232542 (2013).

614 36 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer
615 (IGV): high-performance genomics data visualization and exploration. *Briefings in*
616 *bioinformatics* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

617 37 Bieli, D. *et al.* The Drosophila melanogaster Mutants apblot and apXasta Affect an
618 Essential apterous Wing Enhancer. *G3* **5**, 1129-1143, doi:10.1534/g3.115.017707
619 (2015).

620 38 Preger-Ben Noon, E. *et al.* Comprehensive Analysis of a cis-Regulatory Region Reveals
621 Pleiotropy in Enhancer Function. *Cell reports* **22**, 3021-3031,
622 doi:10.1016/j.celrep.2018.02.073 (2018).

623 39 Bieli, D. *et al.* Establishment of a Developmental Compartment Requires Interactions
624 between Three Synergistic Cis-regulatory Modules. *PLoS Genet* **11**, e1005376,
625 doi:10.1371/journal.pgen.1005376 (2015).

626 40 Anopheles gambiae Genomes, C. *et al.* Genetic diversity of the African malaria vector
627 Anopheles gambiae. *Nature* **552**, 96-100, doi:10.1038/nature24995 (2017).

628 41 Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of
629 transcription factor binding. *Science* **328**, 1036-1040, doi:10.1126/science.1186176
630 (2010).

631 42 Rada-Iglesias, A., Prescott, S. L. & Wysocka, J. Human genetic variation within neural
632 crest enhancers: molecular and phenotypic implications. *Philos Trans R Soc Lond B*
633 *Biol Sci* **368**, 20120360, doi:10.1098/rstb.2012.0360 (2013).

634 43 Riehle, M. M. *et al.* A cryptic subgroup of Anopheles gambiae is highly susceptible to
635 human malaria parasites. *Science* **331**, 596-598, doi:10.1126/science.1196759 (2011).

636 44 Santolamazza, F. *et al.* Insertion polymorphisms of SINE200 retrotransposons within
637 speciation islands of Anopheles gambiae molecular forms. *Malaria journal* **7**, 163,
638 doi:10.1186/1475-2875-7-163 (2008).

639 45 Muller, H. M., Dimopoulos, G., Blass, C. & Kafatos, F. C. A hemocyte-like cell line
640 established from the malaria vector Anopheles gambiae expresses six
641 prophenoloxidase genes. *The Journal of biological chemistry* **274**, 11727-11735
642 (1999).

643 46 Fanello, C., Santolamazza, F. & della Torre, A. Simultaneous identification of species
644 and molecular forms of the Anopheles gambiae complex by PCR-RFLP. *Medical and*
645 *Veterinary Entomology* **16** (2002).

646 47 Harris, C. *et al.* Polymorphisms in Anopheles gambiae immune genes associated with
647 natural resistance to Plasmodium falciparum. *PLoS Pathog* **6**, e1001112,
648 doi:10.1371/journal.ppat.1001112 (2010).

649 48 Redmond, S. N. *et al.* Association mapping by pooled sequencing identifies TOLL 11
650 as a protective factor against Plasmodium falciparum in Anopheles gambiae. *BMC*
651 *genomics* **16**, 779, doi:10.1186/s12864-015-2009-z (2015).

652 49 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary
653 Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549,
654 doi:10.1093/molbev/msy096 (2018).

655 50 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
656 analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303,
657 doi:10.1101/gr.107524.110 (2010).

658 51 Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting
659 autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749-1751,
660 doi:10.1093/bioinformatics/btw044 (2016).

661 52 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948,
662 doi:10.1093/bioinformatics/btm404 (2007).

663 53 Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large
664 datasets. *Bioinformatics* **30**, 3276-3278, doi:10.1093/bioinformatics/btu531 (2014).

665

666 **Acknowledgments**

667 We thank the Center for Production and Infection of *Anopheles* platform (CEPIA) at the
668 Institut Pasteur, and Corinne Genève and Emma Brito-Fravallo, GGIV Institut Pasteur, for
669 rearing mosquitoes. We thank Alexander Stark, Research Institute of Molecular Pathology,
670 Vienna for plasmids and helpful advice. This work received financial support to MMR from
671 National Institutes of Health, NIAID #AI121587; to KDV from the European Commission,
672 Horizon 2020 Infrastructures #731060 Infravec2; European Research Council, Support for
673 Frontier Research, Advanced Grant #323173 AnoPath; and French Laboratoire d'Excellence
674 "Integrative Biology of Emerging Infectious Diseases" #ANR-10-LABX-62-IBEID. The funders
675 had no role in study design, data collection and analysis, decision to publish, or preparation
676 of the manuscript.

677

678 **Author contributions statement**

679 Conceived and designed the experiments: MMR, DMG, KDV

680 Performed the experiments: LN, IH, DMG, SZ, WMG, NS, MMR

681 Analyzed the data: AP, EB, MMR

682 Wrote the manuscript: LN, MMR, KDV

683 All authors read and approved the final manuscript.

684

685 **Ethics statement**

686 No animals or human subjects were used. Mosquito colonies were maintained on
687 anonymous commercial human blood using an artificial membrane feeding device.

688

689 **Data availability**

690 All short read sequence files are available from the EBI European Nucleotide Archive
691 database (http://www.ebi.ac.uk/ena/) under ENA study accession number [REQUESTED]. All
692 other sequences are available in this article as Supplementary Files S1 and S2.
693
694 **Competing interests statement**
695 The authors declare no competing interests.

696 **Table 1: Physical location of enhancer intervals and proximal annotated gene**. Enhancer
697 interval coordinates are based on the locations of PCR primers given in Supplementary Table
698 S2. Coordinates from the PEST AgamP4 genome assembly.
699

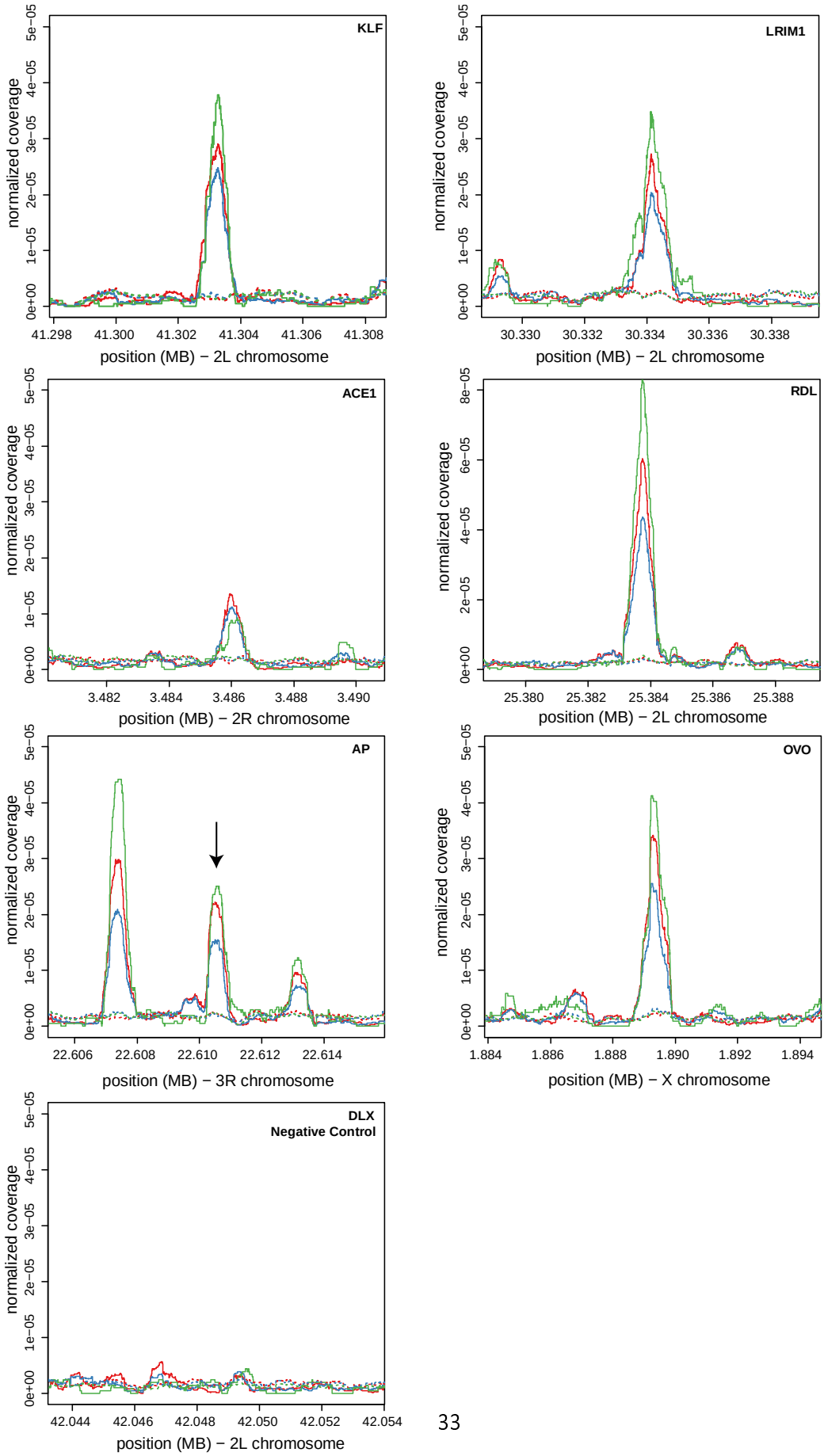
Proximal gene	Enhancer Interval	Proximal Gene Coordinates
AP (AGAP008980)	3R:22609939-22611138	3R:22543990-22609635
OVO (AGAP000114)	X:1888505-1890055	X:1852650-1884326
KLF (AGAP007038)	2L:41302647-41303886	2L:41287202-41308450
LRIM1 (AGAP006348)	2L:30333431-30334787	2L:30329656-30331296
ACE1 (AGAP001356)	2R:3485436-3486583	2R:3483099-3497400
RDL (AGAP006028)	2L:25382828-25384253	2L:25363652-25434556

700

701 **FIGURE LEGENDS**

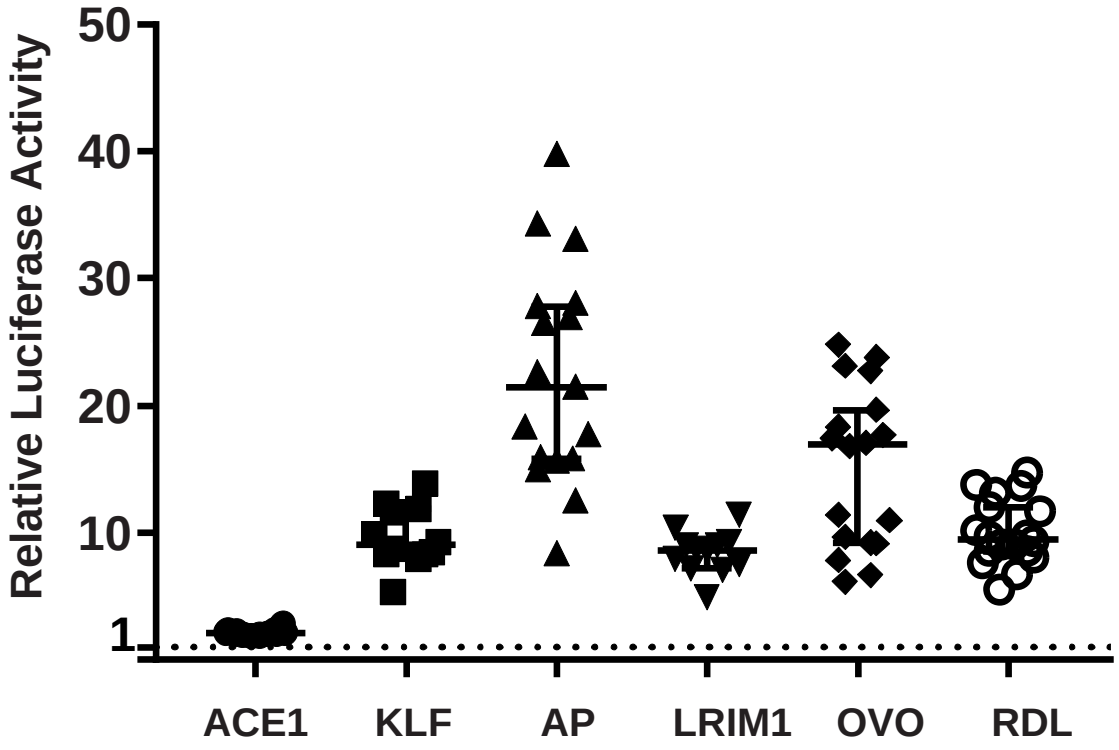
702

703 **Figure 1. Detection of *Anopheles coluzzii* candidate enhancers.** Sequence data near six
704 *Anopheles coluzzii* genes were examined using the Integrative Genomics Viewer (IGV) [36] to
705 screen for candidate enhancers, identified visually where coverage of the cDNA sequence
706 track (solid lines) was greater than the baseline coverage of the plasmid sequence track
707 (dotted lines). The cDNA sequence track is analogous to light output from luciferase reporter
708 assays, but where the readout is enhancer-dependent RNA transcript output in *A. coluzzii*
709 4a3A cells, measured as sequenced cDNA, rather than by luciferase light output. Line color
710 (green, red, blue) represents three biological replicates. The enhancers are named by the
711 most proximal genes: Krueppel-Like Factor 6/7 (KLF, AGAP007038), Leucine-Rich Immune
712 protein (LRIM1, AGAP006348), Acetylcholinesterase (ACE1, AGAP001356), GABA-gated
713 chloride channel subunit (Rdl, AGAP006028), LIM homeobox protein 2/9, ortholog of
714 Drosophila apterous FBgn0267978 (AP, AGAP008980), and Ovo, AGAP000114 (Table1). A
715 negative control interval within intron 1-2 of distal-less (DLX, AGAP007058) was chosen
716 because it lacks visible divergence of cDNA and plasmid sequence tracks. Graphs display
717 cDNA and plasmid tracks in 10 kb windows centered on the candidate enhancers. Only one
718 candidate enhancer is seen in all windows except AP, where the central peak (arrow) was
719 used. X-axis indicates genomic coordinates in the PEST reference genome, y-axis indicates
720 normalized sequence depth corrected for overall plasmid depth observed in the IGV display.


KLF
LRIM1
ACE1
RDL
AP
OVO
DLX
Negative Control
normalized coverage
position (MB) − 2L chromosome
position (MB) − 2R chromosome
position (MB) − 3R chromosome
position (MB) − X chromosome

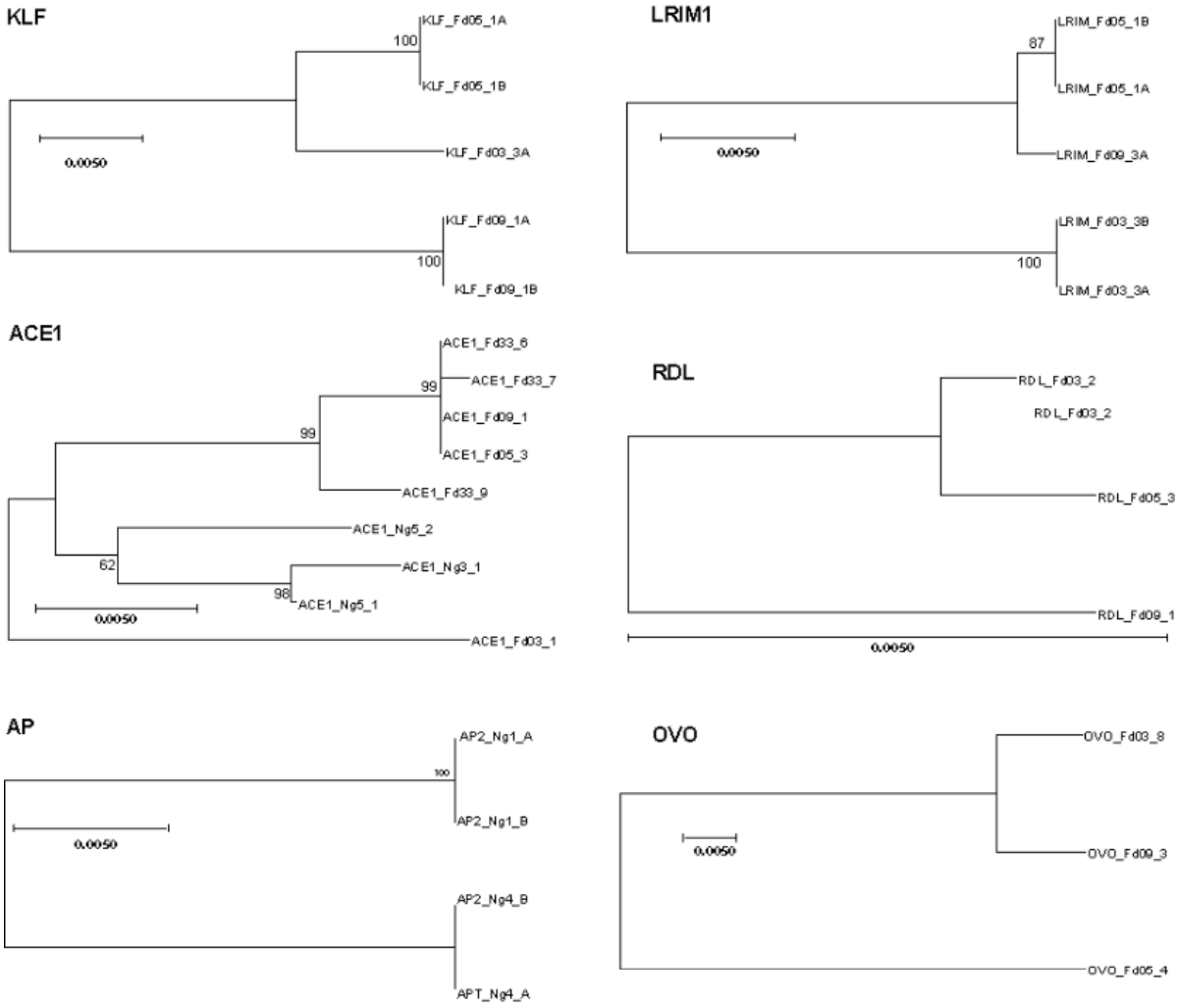
721

722 **Figure 2. Candidate *Anopheles coluzzii* enhancers augment expression of a luciferase**
723 **reporter.** The six candidate enhancer intervals from Figure 1 were amplified from *Anopheles*
724 *coluzzii* mosquitoes and cloned into the pGL-Gateway-DSCP plasmid carrying a basal core
725 promoter and firefly luciferase reporter gene. The cloned candidates were tested for
726 influence upon luciferase expression using a dual luciferase assay system to quantify
727 luciferase activity above background, defined by the DLX negative control (horizontal dotted
728 line). Each of the six tested candidates displayed normalized luciferase activity significantly
729 above background ($p<0.005$), thus validating the candidates as functional *A. coluzzii*
730 enhancers. Each point represents an individual replicate measure of luciferase activity for
731 the tested candidate. Bars indicate the median and 95% confidence intervals. X-axis
732 indicates the name of the candidate enhancer according to the nearest gene (Table 1), y-axis
733 indicates the relative luciferase activity for each measurement, expressed as firefly luciferase
734 corrected to the renilla luciferase internal control value, and normalized for the value of the
735 DLX negative control (see Methods).

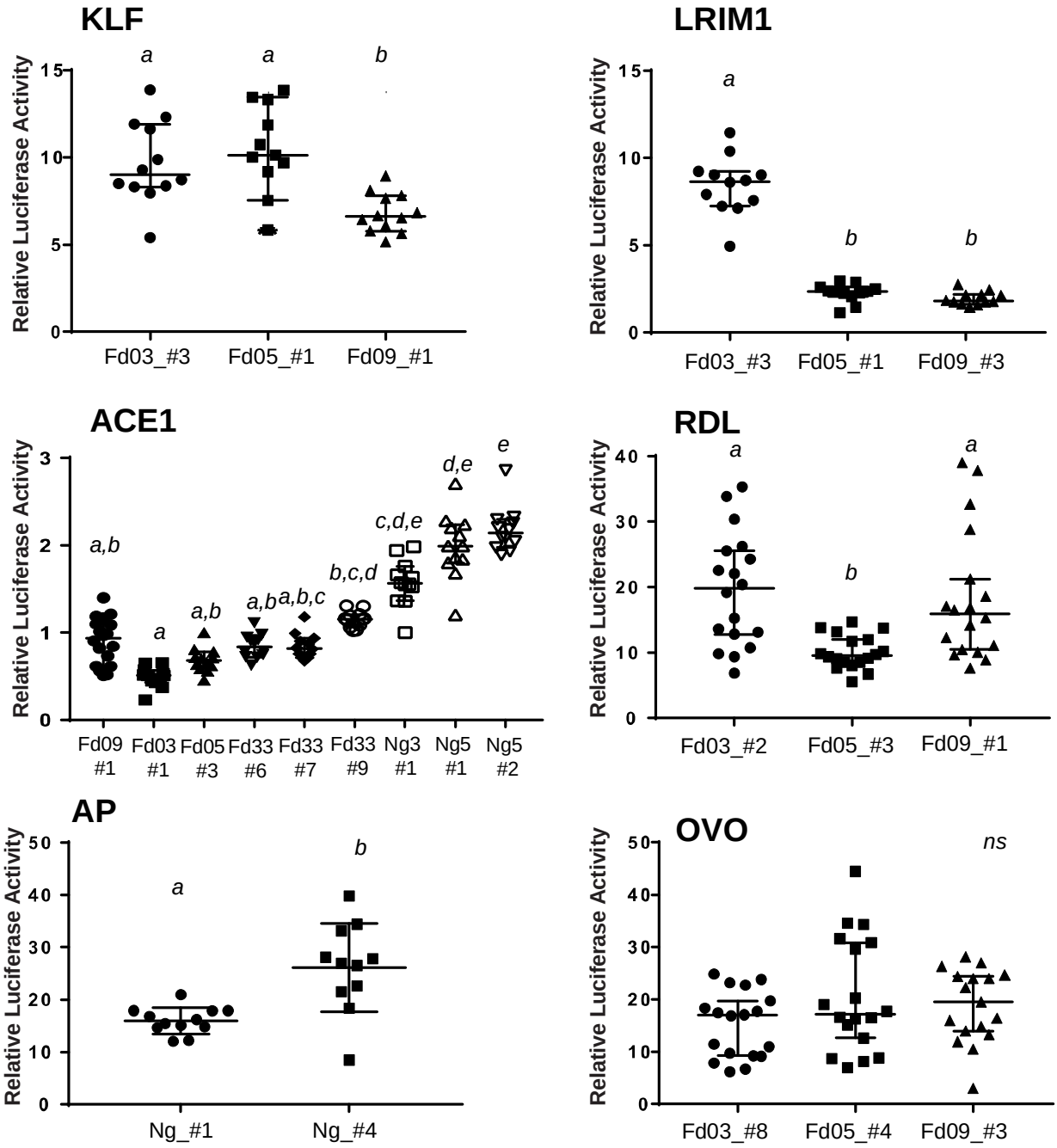

Relative Luciferase Activity
50
40
30
20
10
1
ACE1
KLF
AP
LRIM1
OVO
RDL

736

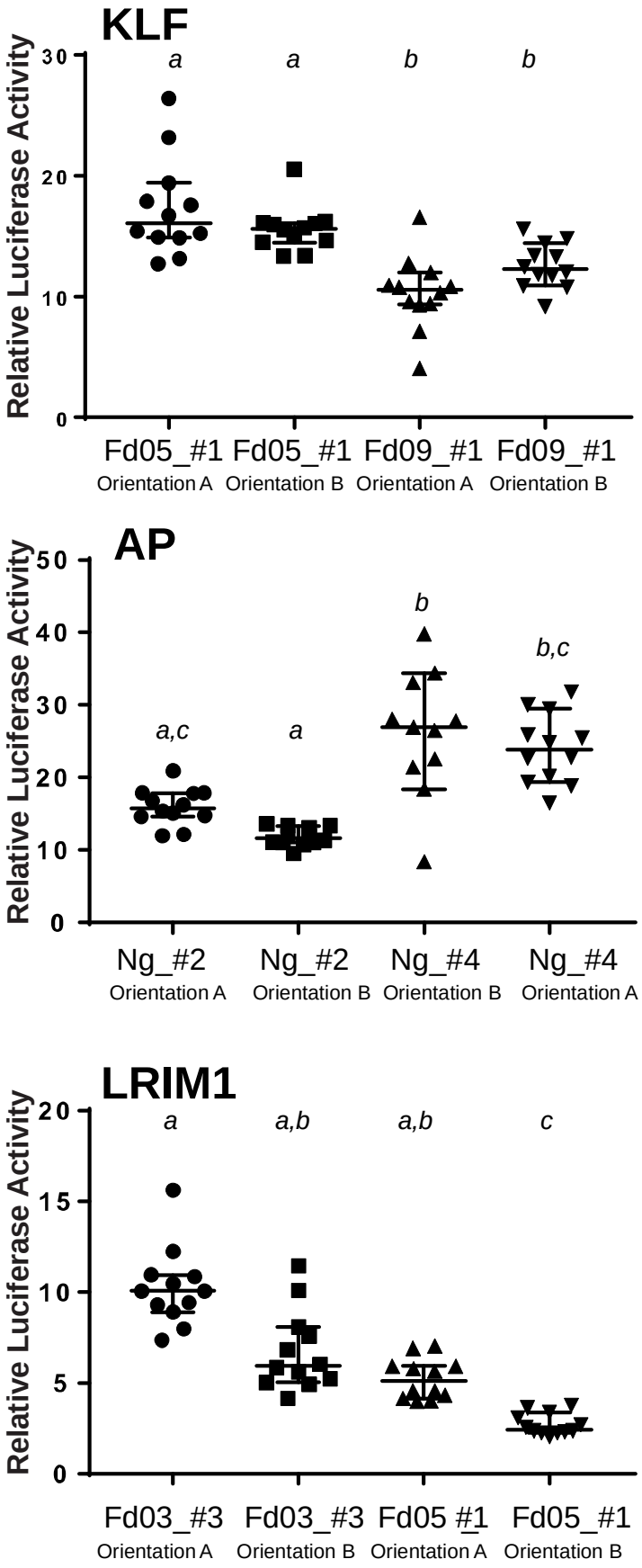
737 **Figure 3. Phylogenetic comparison of enhancer allele genetic variation**. Enhancer allelic
738 variants were cloned and sequenced from *Anopheles coluzzii* colonies. Each sequenced clone
739 represents a chromosomal haplotype. For each clone, individual sequences were aligned
740 using MUSCLE and Mega was used to construct neighbor joining (N-J) trees for complete
741 sequences for all haplotypes for each enhancer. Trees depict the degree of genetic similarity
742 among alleles, and phylogenetic scale bars represent 0.5 nucleotide substitutions per site.
743 The scale bar for the Rdl tree is long (pairwise distance 0.008 between alleles Fd03_#2 and
744 Fd09_#1), indicating that the Rdl alleles segregate relatively little variation, while the Ovo
745 tree scale bar is short (pairwise distance 0.0445 between alleles Fd03_#8 and Fd05_#4),
746 indicating more than 5-fold greater genetic diversity among Ovo alleles as compared to Rdl.
747 Alignments for complete sequences of alleles are presented in Supplementary File S1.


KLF
KLF_Fd05_1A
100
KLF_Fd05_1B
KLF_Fd03_3A
0.0050
KLF_Fd09_1A
100
KLF_Fd09_1B
LRIM1
LRIM_Fd05_1B
87
LRIM_Fd05_1A
LRIM_Fd09_3A
0.0050
LRIM_Fd03_3B
100
LRIM_Fd03_3A
ACE1
ACE1_Fd33_6
ACE1_Fd33_7
99
ACE1_Fd09_1
99
ACE1_Fd05_3
ACE1_Fd33_9
ACE1_Ng5_2
62
ACE1_Ng3_1
98
ACE1_Ng5_1
0.0050
ACE1_Fd03_1
RDL
RDL_Fd03_2
RDL_Fd03_2
RDL_Fd05_3
RDL_Fd09_1
0.0050
AP
AP2_Ng1_A
100
AP2_Ng1_B
0.0050
AP2_Ng4_B
APT_Ng4_A
OVO
OVO_Fd03_8
0.0050
OVO_Fd09_3
OVO_Fd05_4

750 **Figure 4. Genetic variation influences enhancer functional activity**. To test the functional
751 effect of genetic variation within enhancer sequences, the enhancer alleles shown in Figure
752 3 were cloned into luciferase reporter plasmid pGL-Gateway-DSCP, and luciferase activity
753 was measured. Statistically significant differences in luciferase activity as determined using a
754 non-parametric ANOVA are indicated by letters, samples labelled with different letters are
755 significantly different from each other and samples with the same letter are not significantly
756 different (thus samples labeled a,b are not statistically different from samples labelled either
757 a or b). Bars indicate the median and 95% confidence intervals, n=12 for all tests. X-axis
758 labels indicate colony origin (Ng, Ngousso, other colony names as given) and allele name, y-
759 axis indicates the relative luciferase activity for each measurement determined as in Figure
760 2.

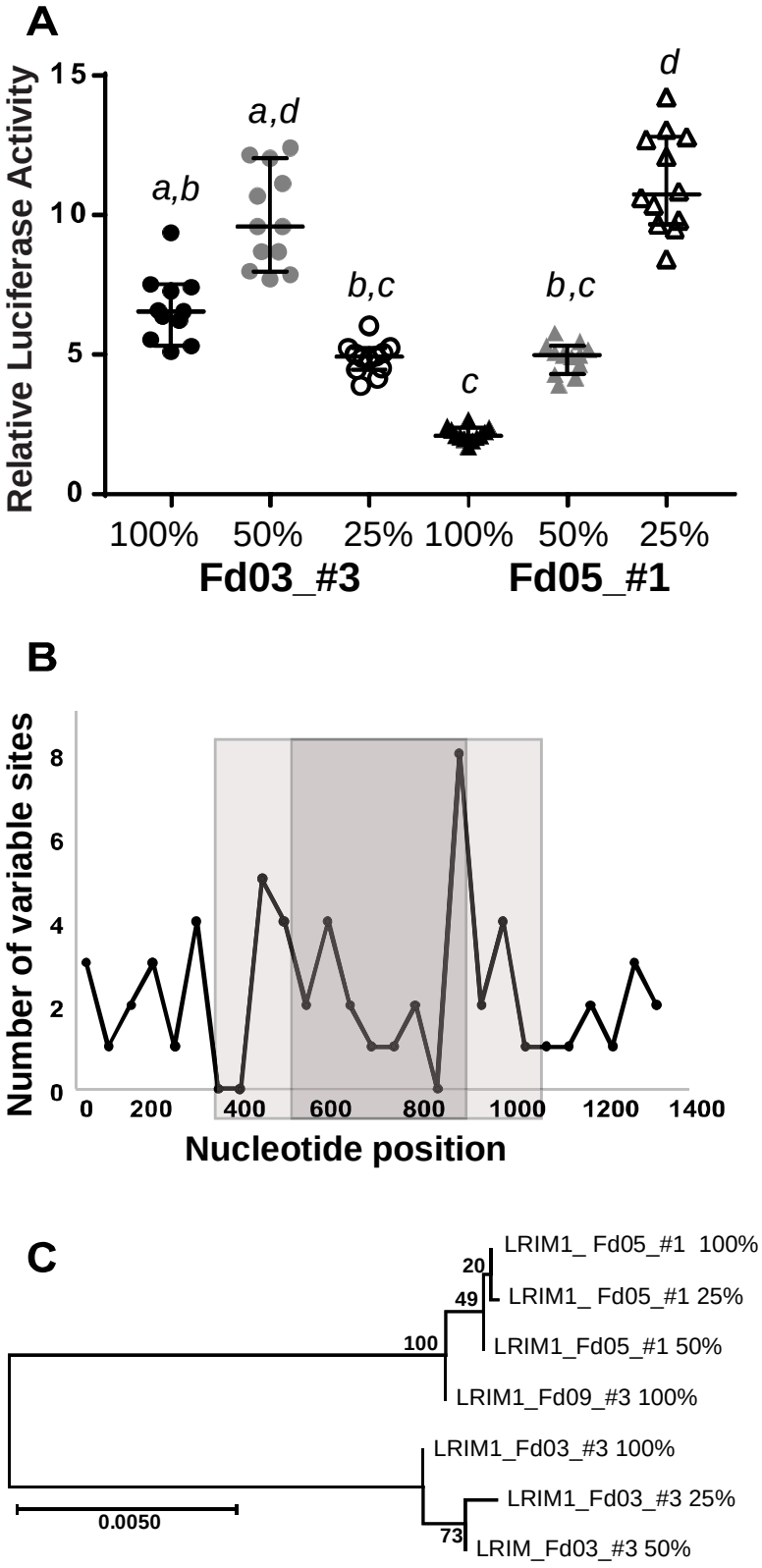

KLF
Relative Luciferase Activity
a
a
b
Fd03_#3
Fd05_#1
Fd09_#1
LRIM1
Relative Luciferase Activity
a
b
b
Fd03_#3
Fd05_#1
Fd09_#3
ACE1
Relative Luciferase Activity
a,b
a
a,b
a,b
a,b,c
b,c,d
c,d,e
d,e
e
Fd09 #1
Fd03 #1
Fd05 #3
Fd33 #6
Fd33 #7
Fd33 #9
Ng3 #1
Ng5 #1
Ng5 #2
RDL
Relative Luciferase Activity
a
b
a
Fd03_#2
Fd05_#3
Fd09_#1
AP
Relative Luciferase Activity
a
b
Ng_#1
Ng_#4
OVO
Relative Luciferase Activity
ns
Fd03_#8
Fd05_#4
Fd09_#3

763 **Figure 5. Enhancer activity is essentially independent of orientation.** The influence of
764 physical orientation of the enhancer within the luciferase reporter plasmid pGL-Gateway-
765 DSCP was tested by cloning three enhancers, KLF, AP and LRIM1, in both orientations in the
766 plasmid, and luciferase activity was measured. KLF and AP enhancers displayed no
767 detectable effect of orientation on luciferase activity, while LRIM1 displayed a slightly
768 significant difference (p=0.042) in luciferase activity for the allele Fd05_#1. Statistical
769 differences indicated by letters as in Figure 4, error bars as in Figure 4, n=12 for all tests. X-
770 axis indicates the name of the enhancer allele tested and the enhancer insert orientation
771 (arbitrarily defined as A and B), n indicates the number of wells measured, y-axis indicates
772 the relative luciferase activity for each measurement as in Figure 2.

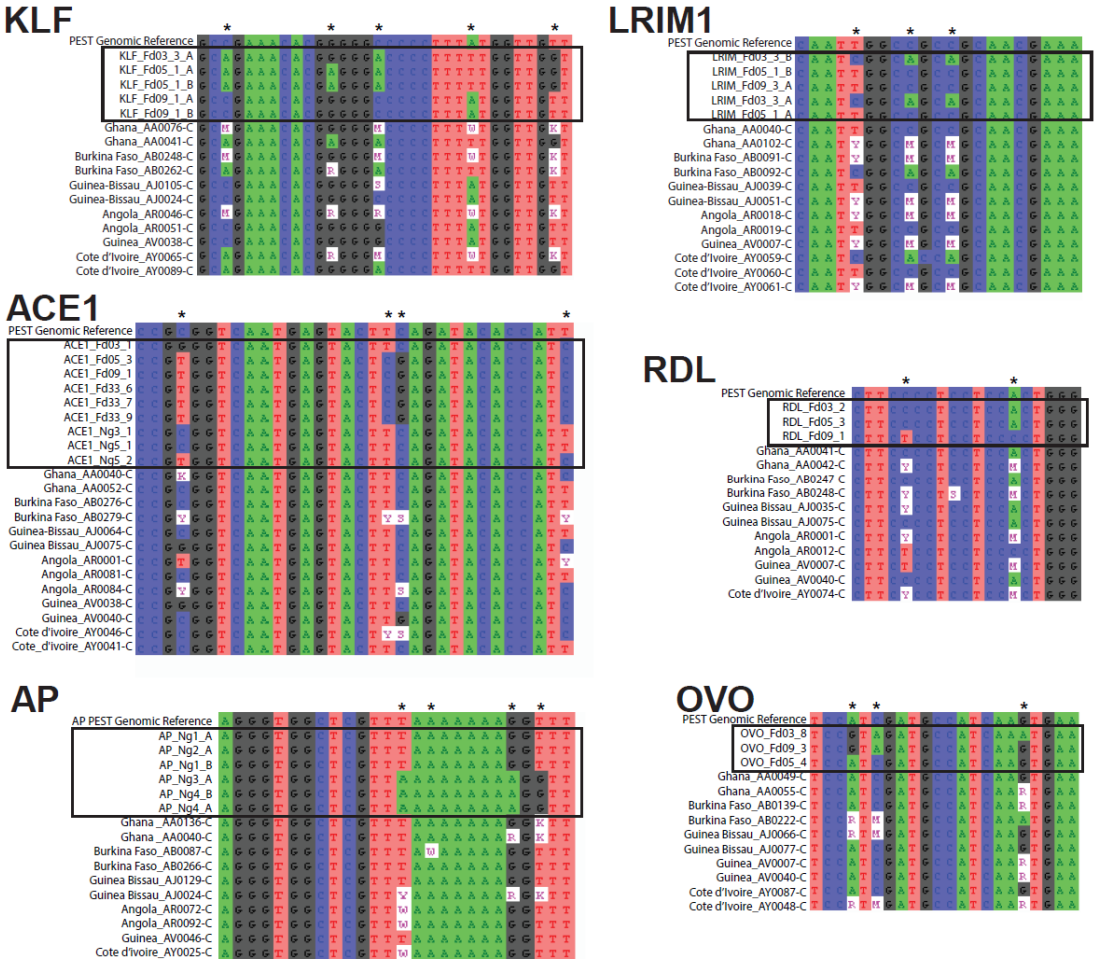

KLF
Relative Luciferase Activity
30
20
10
0
a
a
b
b
Fd05_#1 Orientation A
Fd05_#1 Orientation B
Fd09_#1 Orientation A
Fd09_#1 Orientation B
AP
Relative Luciferase Activity
50
40
30
20
10
0
a,c
a
b
b,c
Ng_#2 Orientation A
Ng_#2 Orientation B
Ng_#4 Orientation B
Ng_#4 Orientation A
LRIM1
Relative Luciferase Activity
20
15
10
5
0
a
a,b
a,b
c
Fd03_#3 Orientation A
Fd03_#3 Orientation B
Fd05 #1 Orientation A
Fd05_#1 Orientation B

773

774 **Figure 6. Enhancer** deletion mutagenesis **reveals positive and negative regulatory**
775 **elements.** Deletion mutagenesis was carried out for two alleles of the LRIM1 enhancer, the
776 high-activity allele Fd03_#3 and low-activity allele Fd05_#1 (Figures 4 and 5). The integral
777 enhancer alleles (100%) were each deleted for one-quarter of their length from both termini
778 (50% derivative), and one-quarter length again (25% derivative). **A.** Deletion derivatives
779 were tested for luciferase activity, along with the original integral alleles. Statistical
780 differences indicated by letters as in Figure 4, error bars as in Figure 4, n=12 for all tests. X-
781 axis indicates allele name and deletion derivatives, y-axis indicates the relative luciferase
782 activity for each measurement as in Figure 2. Enhancer activity is not directly correlated with
783 sequence length, and enhancer alleles are structured from distinct combinations of positive
784 and negative regulators of transcription. **B.** Plot of the number of variant nucleotide
785 positions between the Fd03_#3 and Fd05_#1 alleles along the length of the enhancer
786 sequence. Variant sites are counted within a 50 bp non-overlapping window and plotted at
787 the midpoint of the window. The light gray shading indicates the extent of the 50% length
788 derivatives and the dark gray shading the 25% derivatives. X-axis indicates nucleotide
789 position in derivatives, y-axis indicates number of variable sites between the Fd03_#3 and
790 Fd05_#1 alleles in 50 bp windows. There were a total of 60 variable sites between Fd03_#3
791 and Fd05_#1 alleles in the 100% integral enhancer, 37 variable sites in the 50% derivatives
792 and 22 sites in the 25% derivatives. **C.** Neighbor-joining tree depicting sequence relatedness
793 between the integral 100% enhancer and the 50% and 25% derivatives for LRIM1 Fd03_#3
794 and Fd05_#1 alleles. The Fd09_#3 allele is included as an outgroup. Scale bar description as
795 in Figure 3.


A
Relative Luciferase Activity
15
10
5
0
a,b
a,d
b,c
c
b,c
d
100%
50%
25%
100%
50%
25%
Fd03_#3
Fd05_#1
B
Number of variable sites
8
6
4
2
0
0
200
400
600
800
1000
1200
1400
Nucleotide position
C
LRIM1_ Fd05_#1 100%
20
49
LRIM1_ Fd05_#1 25%
100
LRIM1_Fd05_#1 50%
LRIM1_Fd09_#3 100%
LRIM1_Fd03_#3 100%
LRIM1_Fd03_#3 25%
0.0050
73
LRIM_Fd03_#3 50%

797 **Figure 7. Genetic variants in differentially active enhancer alleles segregate in wild**
798 ***Anopheles coluzzii*.** Genetic variation observed in colonized and wild *A. coluzzii* was
799 compared for the six studied enhancer intervals. Representative short sequence alignments
800 are shown (full-length alignments with additional samples in Supplementary File S2).
801 Asterisks above sequence alignments indicate variant positions shared between the cloned
802 *A. coluzzii* colony haplotypes used in luciferase assays and the natural population. The top
803 sequence row in each alignment is the PEST genome reference sequence, followed by
804 sequences of alleles tested by luciferase assays (boxed by rectangles), followed by
805 sequences of wild *A. coluzzii*. Ambiguous nucleic acid codes are used for heterozygous sites
806 only in wild samples because the cloned sequences from *A. coluzzii* colonies are haplotypes,
807 which are unambiguous.


KLF
PEST Genomic Reference
KLF_Fd03_3_A
KLF_Fd05_1_A
KLF_Fd05_1_B
KLF_Fd09_1_A
KLF_Fd09_1_B
Ghana_AA0076-C
Ghana_AA0041-C
Burkina Faso_AB0248-C
Burkina Faso_AB0262-C
Guinea-Bissau_AJ0105-C
Guinea-Bissau_AJ0024-C
Angola_AR0046-C
Angola_AR0051-C
Guinea_AV0038-C
Cote d'Ivoire_AY0065-C
Cote d'Ivoire_AY0089-C
LRIM1
PEST Genomic Reference
LRIM_Fd03_3_B
LRIM_Fd05_1_B
LRIM_Fd09_3_A
LRIM_Fd03_3_A
LRIM_Fd05_1_A
Ghana_AA0040-C
Ghana_AA0102-C
Burkina Faso_AB0091-C
Burkina Faso_AB0092-C
Guinea-Bissau_AJ0039-C
Guinea-Bissau_AJ0051-C
Angola_AR0018-C
Angola_AR0019-C
Guinea_AV0007-C
Cote d'Ivoire_AY0059-C
Cote d'Ivoire_AY0060-C
Cote d'Ivoire_AY0061-C
ACE1
PEST Genomic Reference
ACE1_Fd03_1
ACE1_Fd05_3
ACE1_Fd09_1
ACE1_Fd33_6
ACE1_Fd33_7
ACE1_Fd33_9
ACE1_Ng3_1
ACE1_Ng5_1
ACE1_Ng5_2
Ghana_AA0040-C
Ghana_AA0052-C
Burkina Faso_AB0276-C
Burkina Faso_AB0279-C
Guinea-Bissau_AJ0064-C
Guinea Bissau_AJ0075-C
Angola_AR0001-C
Angola_AR0081-C
Angola_AR0084-C
Guinea_AV0038-C
Guinea_AV0040-C
Cote d'Ivoire_AY0046-C
Cote_d'ivoire_AY0041-C
RDL
PEST Genomic Reference
RDL_Fd03_2
RDL_Fd05_3
RDL_Fd09_1
Ghana_AA0041-C
Ghana_AA0042-C
Burkina Faso_AB0247 C
Burkina Faso_AB0248-C
Guinea Bissau_AJ0035-C
Guinea Bissau_AJ0075-C
Angola_AR0001-C
Angola_AR0012-C
Guinea_AV0007-C
Guinea_AV0040-C
Cote d'Ivoire_AY0074-C
AP
AP PEST Genomic Reference
AP_Ng1_A
AP_Ng2_A
AP_Ng1_B
AP_Ng3_A
AP_Ng4_B
AP_Ng4_A
Ghana _AA0136-C
Ghana _AA0040-C
Burkina Faso_AB0087-C
Burkina Faso_AB0266-C
Guinea Bissau_AJ0129-C
Guinea Bissau_AJ0024-C
Angola_AR0072-C
Angola_AR0092-C
Guinea_AV0046-C
Cote d'Ivoire_AY0025-C
OVO
PEST Genomic Reference
OVO_Fd03_8
OVO_Fd09_3
OVO_Fd05_4
Ghana_AA0049-C
Ghana_AA0055-C
Burkina Faso_AB0139-C
Burkina Faso_AB0222-C
Guinea Bissau_AJ0066-C
Guinea Bissau_AJ0077-C
Guinea_AV0007-C
Guinea_AV0040-C
Cote d'Ivoire_AY0087-C
Cote d'Ivoire_AY0048-C

808