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.

INTRODUCTION

Enhancers are short cis-acting regulatory elements in noncoding DNA that amplify transcriptional levels of target genes by tens to hundreds fold over the basal level of core promoter elements at the transcription start site. Enhancers control transcriptional activity of target genes and are responsible for most regulated gene expression in the transcriptome. The precise mechanisms of enhancer action is unknown and is an area of active study [1-3]. Enhancers can function at a distance from target genes and independent of their physical orientation in the chromosome [4]. The identities of enhancers and some of their interacting protein factors that lead to their regulatory function have been described in well-studied model organisms, but enhancers cannot be reliably predicted by sequence-based algorithms, and thus must be detected directly by functional activity using reporter assays, or indirectly inferred using methods to detect open or modified chromatin properties.

Sequence polymorphism within enhancers has been associated with phenotypic differences, including predisposition to disease, as observed in diverse organisms [2,5-8]. Most of the significant variants mapped in human genome-wide association studies (GWAS) are noncoding [9], and at least 60-70% of significantly associated human GWAS single-nucleotide point mutations (SNPs) lie within functional enhancers [5,10]. In cancer studies, the majority of tumor-driving mutational changes are also thought to be in noncoding regulatory elements, especially enhancers [11].

Genetic variation in enhancers can occur as SNPs, insertions and deletions (indels), and as copy number variants [12-14]. Enhancer variation among individuals can underlie both

Mendelian and complex traits [4,15,16]. At the population level, positively-selected variation in enhancers controlling key pathways likely plays an important role in differentiation and evolution [17,18]. Indeed, some of the fastest-evolving parts of the human genome as compared to other primates are functional embryonic enhancers related to central nervous system development [19]. In another example, the vertebrate ZRS enhancer influencing limb development displays strong conservation across a range of vertebrates, although in advanced snakes where skeletal limb structures are absent, mutations render the enhancer inactive [20]. Finally, stickleback fish display development of lateral pelvic spines in some populations, which may be protective under certain predation regimes but may be an adaptive liability in others, and spines have been independently lost in different populations. Sequence differences in the Pitx1 enhancer among spined and spineless populations correlate with the phenotype, and enhancer swaps restore or abolish spine development [21]. Thus, relatively simple evolved sequence variation in enhancers can produce large phenotypic shifts in the organism [22,23]. Despite these few examples, the effect of genetic variation on enhancers has barely been examined in any species, and to our knowledge, the functional effect of variation on enhancer activity has not been systematically surveyed in any organism.

Enhancer discovery in mosquitoes is limited to a few previous studies using indirect methods based on detection of chromatin properties to infer enhancer locations [24-26]. However, nothing is known about mosquito enhancer genetic variation and influence on vector phenotypes, and thus it is necessary to first develop the baseline criteria to distinguish and study enhancers in *A. coluzzii*. To this end, here we select a pilot set of candidate enhancers in order to benchmark the parameters needed for reliable enhancer discovery, validation,

and determination of polymorphism effects in *Anopheles*. We validate the pilot candidates as functional enhancers using luciferase reporter assays, and measure the effects of genetic polymorphism on enhancer activity. The results of the current report are a first step towards developing a comprehensive genome-wide catalog of *Anopheles* enhancers, and biological studies to characterize enhancer function in vector biology.

RESULTS

Candidate enhancer selection

The standard approach for enhancer detection is by functional testing using luciferase reporter assays that directly measure enhancer activity, or by indirect methods such as ChIP-seq, which can infer the presence of a subset of enhancers by correlation with chromatin features. Here, we implemented for the first time in *Anopheles* a screen (Self-Transcribing Active Regulatory Region sequencing, STARR-seq) that detects enhancers directly by a functional reporter assay analogous to the luciferase reporter assay, but with the readout of enhancer-dependent RNA transcript output measured as sequenced cDNA, rather than by luciferase light output [27].

In order to identify candidate *A. coluzzii* enhancers, we examined our generated sequence data from the functional screen in the vicinity of six selected genes using the Integrative Genomics Viewer (IGV) [28]. Candidate enhancers were identified by using IGV to manually search near the selected genes to detect intervals where coverage of the cDNA sequence track, indicative of enhancer activity in *A. coluzzii* 4a3A cells (Figure 1, solid lines) was visibly greater than the coverage of the plasmid sequence track, which is the internal baseline control indicating background levels of the plasmid in 4a3A cells (Figure 1, dotted lines). The target genes selected are involved in vector immunity: Krueppel-Like Factor 6/7 (KLF, AGAP007038), Leucine-Rich Immune protein (LRIM1, AGAP006348); insecticide resistance: Acetylcholinesterase (ACE1, AGAP001356), GABA-gated chloride channel subunit (Rdl, AGAP006028); and developmental biology: LIM homeobox protein 2/9, ortholog of Drosophila apterous FBgn0267978 (AP, AGAP008980), and Ovo, AGAP000114 (Table1). The regulatory regions and enhancers of the latter two genes, apterous and Ovo, have been well

characterized for the Drosophila orthologs [29-31]. In addition, a negative control interval was chosen as a size-matched interval located within intron 1-2 of the gene, homeobox protein distal-less (DLX, AGAP007058) that has no visible divergence of cDNA and plasmid sequence tracks by IGV examination, and thus no predicted enhancer function. The candidate enhancer intervals are named according to the most proximal coding sequence.

Functional validation of enhancer activity

The predicted candidate enhancers predicted in Figure 1 were functionally tested to validate the IGV-based predictions. The standard test for enhancer activity is by cloning the candidate in a plasmid carrying a basal promoter and a luciferase reporter gene in an episomal assay. An active enhancer will augment the rate of transcription from the basal promoter, thus elevating the expression level of the luciferase gene. Luciferase expression is measured by adding luciferase substrate to cell extract and detecting light output as relative light units (RLU). Enhancer activity, if any, is measured as increased luciferase activation above background.

Candidate amplicons from *A. coluzzii* (Table 1) were cloned into the firefly luciferase reporter vector pGL-Gateway-DSCP, and co-transfected into 4a3A cells with the renilla luciferase control vector pRL-ubi-63E. Firefly luciferase RLU measurements were corrected using the renilla luciferase internal control values in the same well, and firefly/renilla RLU for the experimental insert were statistically compared to the firefly/renilla mean value for the DLX negative control, defined as the background level. At least one clone of each candidate enhancer displayed luciferase activity levels above background ($p<0.005$), with activity across candidates that varied from 2-fold to more than 20-fold over background (Figure 2).

158 These results indicate that the IGV-based predictions were accurate for all six predicted
159 candidates, and thus validate these genomic intervals as functional *A. coluzzii* enhancers.
160 This information provides the first benchmark criteria that can be used to develop the
161 definitions and methods for subsequent algorithmic genome-wide detection of *A. coluzzii*
162 enhancers.
163
164 **Screening for polymorphic alleles of validated enhancer intervals**
165 Having confirmed that all six predicted candidates are functional enhancers, we next wished
166 to identify genetically variable alleles for each enhancer interval and measure their
167 luciferase activity. For this, alleles of the enhancer intervals were amplified and sequenced
168 from *A. coluzzii* colonies initiated from the populations in Cameroon, Mali or Burkina Faso.
169 For each of the six enhancer intervals, at least two distinct genetic variants were chosen for
170 tests of enhancer activity. The enhancer interval alleles were cloned and sequenced, and
171 neighbor joining (N-J) trees depict the evolutionary relatedness and degree of sequence
172 difference of the alleles (Figure 3). Complete sequences for all tested enhancer interval
173 alleles are presented in Supplementary File S1.
174
175 **Genetic alleles of validated enhancer intervals display distinct enhancer activity**
176 Luciferase activity was measured for all alleles to determine the effect of genetic variation
177 on differences in functional enhancer activity. For five of the six enhancers, alleles displayed
178 significantly different levels of enhancer activity (Figure 4). For a given enhancer, alleles with
179 the greatest difference in activity tended to be the most genetically different from each
180 other (see also Figure 3). For example, the alleles of the KLF interval cloned from colonies
181 Fd05 and Fd03 are the most closely related genetically, and these also do not display a

difference in luciferase activity as compared to the allele from colony Fd09. For two enhancer intervals (LRIM1 and ACE1), at least one genetic variant displayed activity levels that were not significantly different from background, which effectively represents a naturally occurring functionally inactive null enhancer allele. Genetic variation segregating at the enhancer of Ovo did not display a significant influence on luciferase activity, and the Ovo enhancer appears to display the consistently highest luciferase activity over all alleles tested for any of the six enhancer intervals. These results indicate that genetic alleles of validated enhancer intervals can display significantly different levels of functional activity. It is reasonable to expect that large observed differences in enhancer activity will be translated into phenotypic differences in the organism, related to the functions of the target genes that are regulated by the polymorphic enhancer alleles. This prediction will need to be tested in manipulative experiments.

Enhancer activity is essentially independent of physical orientation

Enhancers tend to function independently of their physical orientation in the genome, which is testable when the candidate is cloned in a luciferase reporter plasmid. For three of the above validated enhancers, we recloned two alleles in both orientations in the reporter and measured luciferase activity. For the KLF and AP enhancers, there was no detectable effect of orientation (Figure 5). The LRIM1 enhancer displayed a weakly significant effect of orientation for allele Fd05_#1 ($p=0.042$), although for both orientations of the LRIM1 enhancer the absolute activity values were lower than the other enhancers tested (indicated by y-axis values in Figure 5), and thus the weak orientation difference for this one weak allele is not robustly supported. Thus, most of the enhancer alleles tested displayed function independent of their physical orientation with respect to the basal promoter.

206

207 **Enhancer dissection reveals a modular structure of positive and negative regulatory**
208 **elements**

209 To resolve the minimal portion of the enhancer interval that carries the enhancer function,
210 we carried out deletion mutagenesis for two different genetic alleles of the LRIM1 enhancer
211 interval, one allele with high enhancer activity and the other low. The deletion derivatives
212 carried 50% or 25% of the length of the initial enhancer interval, reduced equally from both
213 ends. We tested the deletion clones for luciferase activity, along with the original undeleted
214 enhancer (Figure 6A). Surprisingly, for LRIM1 allele Fd03_#3, the 50% construct displayed
215 the highest luciferase activity, greater than either 100% or 25% constructs. This indicates
216 that the integral 100% Fd03_#3 allele carries negative regulators of enhancer function,
217 which were deleted in the 50% derivative to yield a derivative with elevated enhancer
218 activity. The 25% derivative of allele Fd03_#3 displays significantly lower activity than the
219 50% derivative, suggesting that positive regulators of enhancer function are located outside
220 the 25% derivative, but within the 50% derivative.

221

222 Deletion derivatives of LRIM1 allele Fd05_#1 display a pattern distinct from the Fd03_#3
223 allele. For Fd05_#1, each incrementally smaller derivative was more active. This result was
224 also surprising, because it indicates that a highly active core enhancer element within the
225 smallest 25% derivative is attenuated by negative regulators that are progressively removed
226 from 100% to 50% in length, and again from a 50% to 25% length interval. The deletion
227 results indicate that enhancer activity is not directly correlated with sequence length, that
228 there is a complex structure of functional elements and modifiers within the enhancer

229 interval, and that different alleles of the same enhancer are comprised of distinct
230 combinations of modular regulators that differentially influence transcription.
231
232 The density of variable sites between Fd03_#3 and Fd05_#1 varies across the interval, such
233 that there were 60 variable nucleotide sites in the integral 100% length alleles, 37 variable
234 sites in the 50% derivatives and 22 sites in the 25% derivatives (Figure 6B). Finally, it is
235 notable that the 25% derivative for allele Fd05_#1 displays activity levels indistinguishable
236 from the Fd03_#3 50% derivative (p=0.99), even though they are no more genetically similar
237 than the integral 100% enhancer sequences for both alleles (Figure 6C). This result highlights
238 the relative independence of enhancer functional level from primary sequence patterns,
239 unlike the fundamental dependence of protein coding gene function on the amino acid
240 primary sequence code, and the consequent requirement for identification of enhancers by
241 detecting functional activity.
242
243 **Enhancer alleles segregate in the natural *Anopheles coluzzii* population**
244 To confirm that the genetic variation observed in the enhancer alleles was natural and not
245 an artifact of laboratory colonies, we compared sequence data for the six enhancer intervals
246 to genetic variation observed in wild *A. coluzzii* from whole genome sequence of the
247 *Anopheles* gambiae 1000 (Ag1000) Genomes Consortium[32]. The comparison indicates that
248 genetic variation is shared between the cloned *A. coluzzii* colony haplotypes used in
249 luciferase assays and the natural population (Figure 7). Representative short sequence
250 alignments are shown, and full-length alignments with larger numbers of wild mosquitoes
251 are presented in Supplementary File S2. This analysis demonstrates that genetic variants
252 within confirmed functional enhancer intervals, associated with differential enhancer

253 activity, segregate in nature and do not represent variants unique to lab colonies. Natural
254 segregation of variants associated with differential enhancer function supports the
255 interpretation that the differential function of enhancer alleles (Figure 4) based on a
256 modular structure of regulatory elements (Figure 6) likely result from natural selection for
257 distinct phenotypic outcomes of allelic enhancer function.

DISCUSSION

Here we identify and validate candidate enhancer noncoding regulatory elements in the malaria vector, *A. coluzzii*. We show that naturally segregating genetic variation significantly influences enhancer activity levels, which likely leads to differences in biological function and ultimately mosquito phenotype. Some enhancer alleles display high activity while others display little or no activity above background and are thus naturally occurring enhancer null alleles. The enhancers also tend to display activity that is independent of their physical orientation, a common property of enhancers [4]. A structure-function dissection study of two enhancer alleles by deletion mutagenesis revealed a complex modular organization of positive and negative modifiers that modulate enhancer activity. The current study provides proof of principle for the influence of enhancer genetic polymorphism for enhancer functional activity levels. These results provide benchmark parameters that can now be implemented to develop a comprehensive genome-wide *Anopheles* enhancer catalog. The current work is thus a step toward the long-term goal to identify functionally important transcription factor binding motifs and correlate enhancer output with phenotypes related to *Anopheles* biology, immunity and pathogen transmission.

By modifying the level of enhancer activity, genetic variation in enhancers can cause quantitative changes in expression of the target genes regulated by the enhancers. Altered expression profiles of enhancer target genes probably in turn trigger distinct phenotypic outcomes. Different from mutations in protein coding genes, enhancers are typically located in noncoding DNA, and there is no sequence pattern to aid interpretation of noncoding variants. Here, we identified enhancer intervals proximal to genes underlying the important vector phenotypes of insecticide resistance, immunity, and development. Most mosquito

282 studies to date have focused solely on genes and proteins associated with these traits,
283 rather than regulatory elements controlling the genes, in part due to the limited information
284 available for the noncoding mosquito genome. The development of a methodology for
285 screening and evaluation of *Anopheles* enhancers is an initial step towards a more
286 comprehensive study of enhancers and their polymorphism effects. The current enhancers
287 were located near known functional genes, but further work will be required to determine
288 the actual influence, if any, of these enhancers upon the nearby genes, which is not known.
289 Moreover, enhancer function is also controlled on spatial and temporal scales, and
290 understanding *Anopheles* enhancers and their effects on phenotypes in detail will ultimately
291 require incorporating this information.
292
293 Enhancers were identified and individual genetic variants were tested for their activity by
294 means of standard luciferase reporter assays. Interestingly, we detected variant alleles with
295 significant difference in their functional enhancer activity, including functional null alleles
296 that lack enhancer activity above background. For example, the LRIM1 enhancer Fd05_#1
297 allele or the ACE1 Fd03_#1 allele likely represent the ablation of an important transcription
298 factor binding site, resulting in the absence of enhancer activity above background with
299 likely downstream functional consequences. The range in functional enhancer activity that
300 we observed is likely to affect phenotypes produced by the genes they transcriptionally
301 regulate. Moreover, we demonstrate that variant alleles tested by luciferase activity in
302 laboratory colonies also segregate in the wild population, and are therefore subject to
303 natural selection. Thus, it is intriguing that selection has apparently generated a wide range
304 of natural allelic forms of enhancers, including alleles that lack functional activity. This is
305 consistent with the observation that genetic variation for enhancer function offers powerful

raw material for adaptation and evolutionary change [8,17,20,21]. The members of the Gambiae species complex, including *A. coluzzii*, are highly adaptable to a range of ecological conditions, and durable to vector control measures. This is thought to be associated with their high genetic diversity [32]. The current study now reveals standing genetic variation for enhancer function as a likely new contributing factor for the success of this mosquito and its relatives.

We functionally dissected two alleles of the LRIM1 enhancer, high and low activity variants, respectively, by deletion mutagenesis (Figure 6). By measuring functional activity of integral, 50% and 25% length derivatives of the intervals, we detected a modular structure of positive and negative regulators comprising the enhancer. Interestingly, deletion derivatives of the two alleles behaved differently, indicating that the deletions was not a simple consequence of sequence length. The high-activity allele Fd03_#3 appears to carries a negative regulator in the terminal one-quarter of its length on one or both ends, because removal of these sequences led to significantly elevated activity in the remaining 50% derivative as compared to the integral enhancer. However, removal of an additional one-quarter again of the sequence from both ends of the 50% derivative then diminished activity to a level below that of the integral enhancer, suggesting that the positive regulator(s) revealed in the 50% length derivative was no longer present in the 25% length derivative.

That the low activity of the smallest derivative of Fd05_#1 was not a simple consequence of sequence length is made clear by a similar examination of the low-activity allele Fd05_#1. In this case, each incremental length decrease of the tested sequence led to increased enhancer activity. The Fd05_#1 allele result suggests that the integral enhancer displayed

low activity because it carried multiple negative regulators, which were resected by each successive deletion, revealing a highly active core enhancer element within the smallest interval tested. This latter minimal derivative of the low-activity Fd05_#1 allele carries an enhancer with, in fact, higher enhancer activity than the integral 100% sequence of the high-activity Fd03_#3 allele.

The LRIM1 deletion dissection results suggest that that large functional allelic diversity can be generated for a given enhancer interval by the combinatorial effect of positive and negative modifiers. Sequence changes in enhancers can generate or remove binding motifs for transcription factors and other regulatory proteins, which can modify transcription levels directly [33,34], or indirectly through loss of chromatin accessibility [14]. From the current results, we do not know whether the positive and negative modifiers within the LRIM1 alleles are comprised of reusable modular cassettes that are combined to fine-tune the activity of different enhancer intervals, or whether segregating SNPs in an enhancer can explain significant difference in functional activity. In the first model, such modifier modules should be recognizable with enough data, while under the second model, different modifiers, for example positive modifiers in different enhancers, may have little or no recognizable common pattern. Fine resolution nucleotide-level deletion series of a number of alleles would be required to determine the kind and extent of sequence difference necessary to alter enhancer functional levels in order to distinguish between the above models. Finally, the phenotypic implications of differentially active enhancer alleles will require determination of target gene networks regulated by an enhancer, as well as the protein-DNA interactions underlying differential enhancer allele activity.

354 Vector control has been central to the malaria control effort by use of indoor residual
355 spraying and long-lasting insecticide impregnated bednets. However, over-reliance on these
356 methods has led to widespread insecticide resistance in wild populations, and novel
357 methods of control are now required. The noncoding regulatory genome in *Anopheles* has
358 the potential to provide novel new targets for vector control, but until now has not been
359 interpretable or exploitable. The current work presents a necessary first step towards
360 establishing an efficient, effective method for associating noncoding variation with
361 important mosquito phenotypes.

METHODS

Wild mosquito samples and DNA library

Mosquito larvae were collected in Goundry village, Burkina Faso (latitude 12.5166876, longitude -1.3921092) using described methods [35], reared to adults, and were typed for species by the SINE200 X6.1 assay [36]. DNA from 60 *A. coluzzii* were pooled at equal volume and sheared using an S220 ultrasonicator (Covaris) to produce DNA fragments 800-1000 bp in length. Subsequently, DNA was processed as described for the STARR-seq assay [27], cloned into the plasmid pSTARR-seq_fly (AddGene 71499), transformed into MegaX DH10B T1R Electrocomp Cells (Invitrogen), cultured in LB + ampicillin (1ug/ml), and plasmid DNA was purified using the Plasmid Plus Mega Kit (Qiagen). The *Anopheles* gambiae PEST AgamP4 genome assembly available at Vectorbase was used as the reference genome (https://www.vectorbase.org/organisms/anopheles-gambiae/pest/agamp4).

Culture of plasmid library in *Anopheles* 4a3A cells

Hemocyte-like 4a3A cells [37] were maintained on Insect X-Press media (Lonza) supplemented with 10% Fetal Calf Serum, at 27°C. We confirmed that cells were derived from *A. coluzzii* by species typing using the Fanello assay [38]. The plasmid DNA library was transfected and cultured in 4a3A cells as described [27] using Lipofectamine 3000 Reagent (Invitrogen) and cultured for 24 h, in three biological replicates. RNA was extracted from cells using the RNeasy Midi Kit (Qiagen) followed by mRNA purification using Dynabeads mRNA Purification Kit (ThermoFisher). Plasmid DNA was isolated using the Plasmid Plus Midi or Mini Kit (Qiagen).

Analysis of 4a3A library culture results

386 The mRNA purified from cells was reverse transcribed using SuperScript IV First-Strand cDNA
387 Synthesis System (Invitrogen) as described for the STARR-seq assay [27] using a plasmid-
388 specific primer (RT_Rev, Supplementary Table S1), the cDNA was then amplified using
389 primers Report_Fwd and Report_Rev (Supplementary Table S1), and the products were
390 sequenced on an Illumina HiSeq 2500 in 2x125 bp high output mode. Cell plasmid DNA was
391 amplified and sequenced in the same manner as the cDNA samples but using primers
392 Plasmid_Fwd and Plasmid_Rev (Supplementary Table S1).
393
394 **Selection of enhancer candidates**
395 The Integrative Genomics Viewer (IGV) [28] was used to select candidate enhancer intervals by
396 visual examination in the proximity of six annotated genes of interest. For determination of
397 enhancer activity, the RNA output transcribed from the STARR-seq reporter plasmid,
398 converted to cDNA and sequenced as described above, is compared to the levels of the
399 plasmid DNA, to control for differential plasmid replication. Thus, candidate enhancers were
400 predicted in intervals where coverage of the cDNA sequence track was visibly greater than
401 the baseline coverage of the plasmid sequence track. The target genes examined were
402 Krueppel-Like Factor 6/7 (KLF, AGAP007038), Leucine-Rich Immune protein (LRIM1,
403 AGAP006348), Acetylcholinesterase (ACE1, AGAP001356), GABA-gated chloride channel
404 subunit (Rdl, AGAP006028), LIM homeobox protein 2/9, ortholog of Drosophila apterous
405 FBgn0267978 (AP, AGAP008980), and Ovo, AGAP000114. In addition, a negative control
406 interval was cloned, which was a size-matched interval located within intron 1 of the gene,
407 homeobox protein distal-less (DLX, AGAP007058) that displayed no visible divergence of
408 cDNA and plasmid sequence tracks by IGV examination, and thus no predicted enhancer

409 function. The candidate enhancer intervals are named according to the most proximal
410 coding sequence (above and Table 1).
411
412 Candidate enhancers were amplified from DNA of mosquitoes from the following *A. coluzzii*
413 colonies: Ngousso, initiated in Cameroon in 2006 [39], Fd03, Mali, 2008, Fd05, Mali 2008, Fd09,
414 Burkina Faso, 2008, and Fd33, Burkina Faso, 2014. Fd colonies were previously described [40].
415 Primers are listed in Supplementary Table S2. Amplicons were cloned into the
416 pCR8/GW/TOPO vector (Invitrogen) and sequenced with GW1 and GW2 primers. At least
417 two genetically distinct sequences per candidate were then cloned into the firefly luciferase
418 reporter plasmid pGL-Gateway-DSCP (AddGene 71506) using Gateway LR Clonase II
419 (Invitrogen), transformed into OneShot OmniMax 2T1 Phage-Resistant Cells (Invitrogen), and
420 plasmid was purified from overnight culture.
421
422 To test the effect of enhancer orientation, the enhancer was cloned in the opposite
423 orientation in pGL-Gateway-DSCP. To test resolved enhancer intervals, the relevant
424 enhancer insert was amplified with primers that generated either 50% or 25% of the initial
425 insert size, equally reduced on both ends, and products were cloned in pGL-Gateway-DSCP.
426 In all cases, plasmids were resequenced to confirm insert identity using the primers LucNrev
427 and RVprimer3 (Supplementary Table S1).
428
429 **Quantitation and statistical analysis of enhancer activity by luciferase assay**
430 The Dual-Glo Luciferase Assay System (Promega) was used for luciferase assays. *A. coluzzii*
431 4a3A cells were seeded in 96 well plates at $1x10^5$cells/well, the difference in volume if any
432 was made up to 65 µl with medium, and cells were incubated for 24 h at 27°C. Two plasmids

were transfected into the 4a3A cells, the enhancer candidate in firefly luciferase vector pGL-Gateway-DSCP, and the renilla luciferase control vector pRL-ubi-63E (AddGene 74280), at a ratio of 1:5 (renilla:firefly), using transfection reagent Lipofectamine 3000 (Invitrogen), and were then incubated for 24 h at 27°C.

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

Analysis of enhancer allelic variants

The sequences of genetically polymorphic variants of a given enhancer interval, cloned from *A. coluzzii* colonies as described above, were analyzed for genetic relatedness. To generate neighbor joining (N-J) trees to depict the relationships between genetic variants for the

457 same enhancer, complete sequences were aligned using MUSCLE within the package
458 Molecular Evolutionary Genetics Analysis Mega version X [41], and N-J trees constructed using
459 Mega. When at least four variants were tested, bootstrapping was performed and bootstrap
460 values are included on N-J trees. Scale bars of trees represent 0.5, but each bar is a different
461 length. The longer the 0.5 scale bar, the more genetically similar the sequences.
462
463 **Analysis of wild *Anopheles* variation data**
464 Sequence information for 309 wild *A. coluzzii* from 6 West African countries; Angola (AR),
465 Burkina Faso (AB), Cote d'Ivoire (AY), Ghana (AA), Guinea (AV) and Guinea-Bissau (AJ),
466 generated as reported [32] were downloaded from MalariaGen
467 (https://www.malariagen.net/projects/ag1000g) as VCF files from the Ag1000G phase 2 AR1
468 data release and sequences of the six validated enhancer intervals were extracted. Next, we
469 aligned the diploid sequences from the wild sequences, corresponding to the cloned
470 sequences generated from the six tested enhancer intervals, which were validated for
471 enhancer function by luciferase activity. Sequence alignments were visually examined for
472 shared variation. Short representative sequence alignments are presented in Figure 7 (not
473 including indels), and complete alignments relative to the PEST AgamP4 genome assembly,
474 including indels, are presented in Supplementary File S2. Indel genotypes of the wild
475 sequences shown in Supplementary File S2 are relative to the PEST reference haplotype,
476 because the wild sequences were called for SNPs but not indels [32].

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

584 40 Redmond, S. N. *et al.* Association mapping by pooled sequencing identifies TOLL 11
585 as a protective factor against Plasmodium falciparum in Anopheles gambiae. *BMC*
586 *genomics* **16**, 779, doi:10.1186/s12864-015-2009-z (2015).

587 41 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary
588 Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549,
589 doi:10.1093/molbev/msy096 (2018).
590

Acknowledgments

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

Author contributions statement

Conceived and designed the experiments: MMR, DMG, KDV

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

Analyzed the data: AP, EB, MMR

Wrote the manuscript: LN, MMR, KDV

All authors read and approved the final manuscript.

Ethics statement

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

Data availability

All short read sequence files are available from the EBI European Nucleotide Archive database (http://www.ebi.ac.uk/ena/) under ENA study accession number [REQUESTED]. All other sequences are available in this article as Supplementary Files S1 and S2.

Competing interests statement

The authors declare no competing interests.

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

624

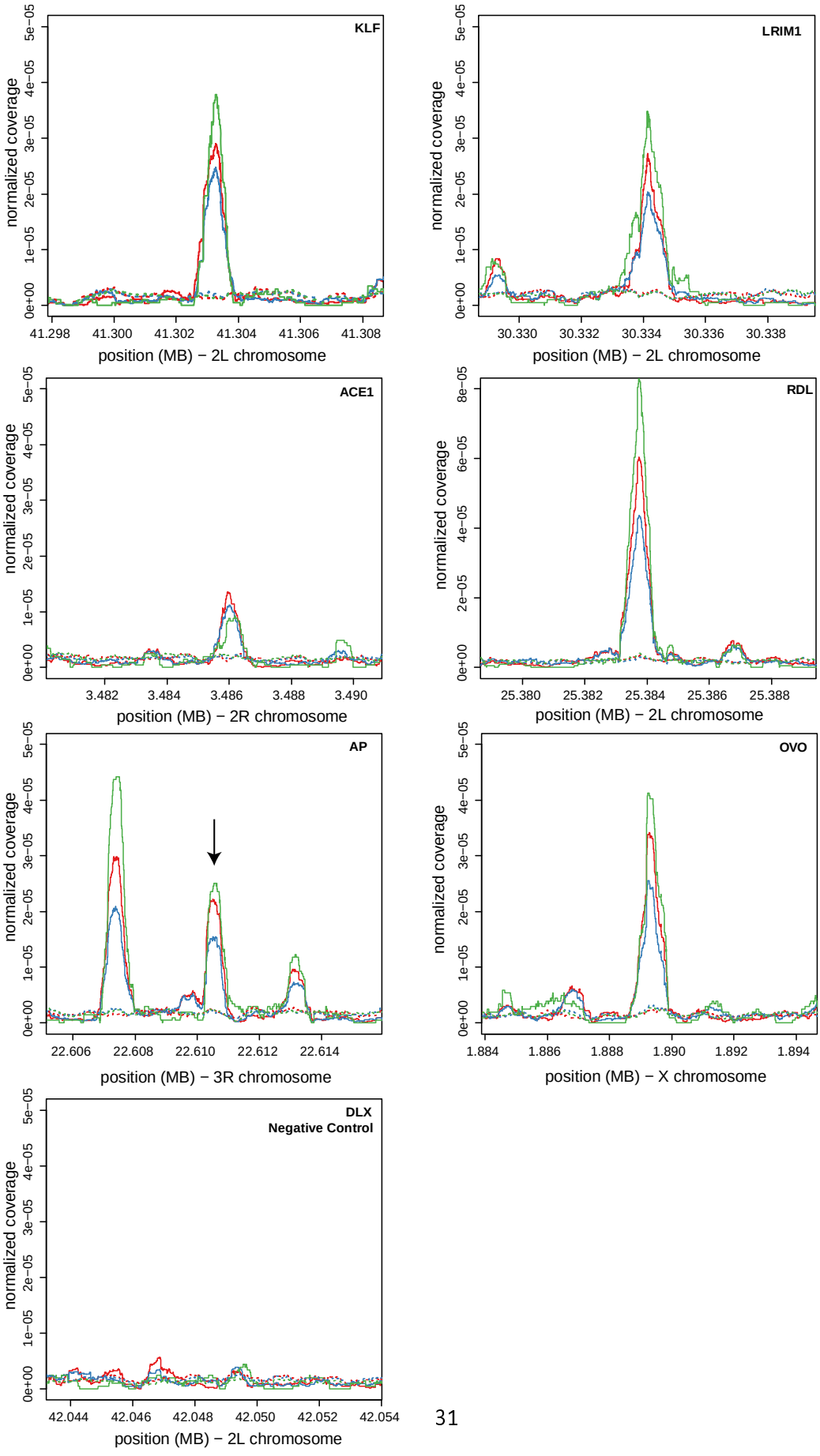
Proximal gene	Enhancer Interval	Proximal Gene Coordinates
AP (AGAP008980)	3R:22609939-22611138	3R:22543990-22609635
OVO (AGAP000114)	X:1888505-1890055	X:1852650-1884326
KLF (AGAP007038)	2L:41302647-41303886	2L:41287202-41308450
LRIM1 (AGAP006348)	2L:30333431-30334787	2L:30329656-30331296
ACE1 (AGAP001356)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

625

626 **FIGURE LEGENDS**

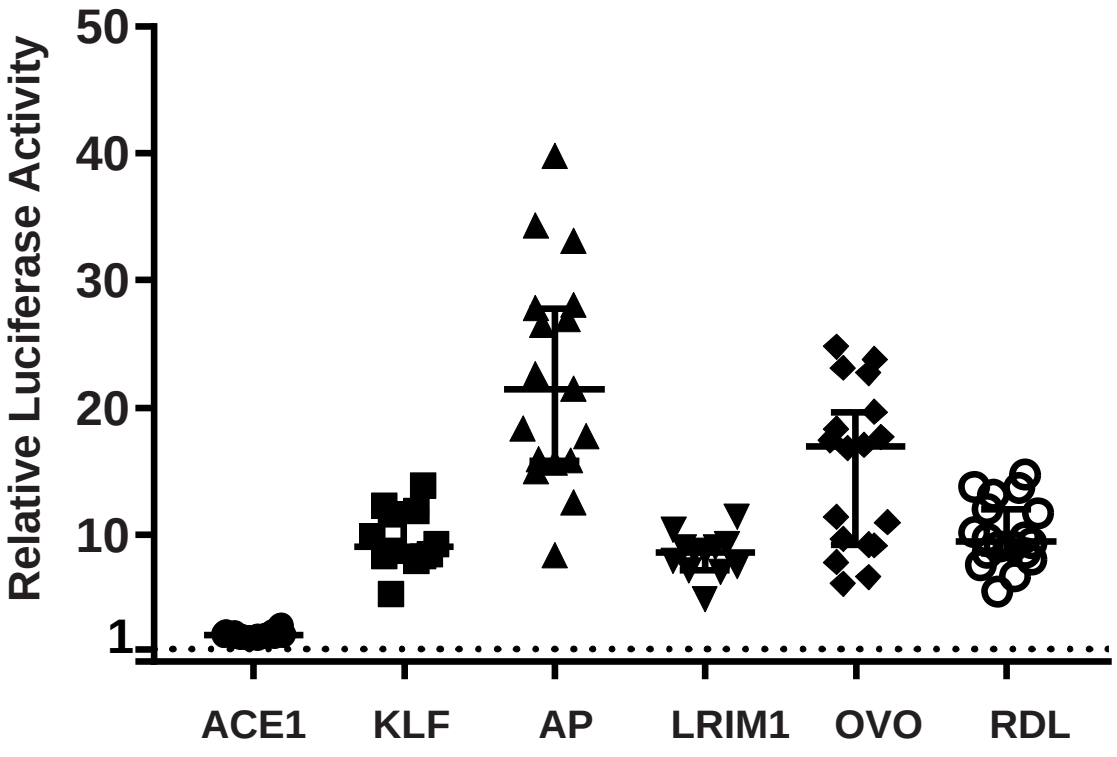
627

628 **Figure 1. Detection of *Anopheles coluzzii* candidate enhancers.** Sequence data near six
629 *Anopheles coluzzii* genes were examined using the Integrative Genomics Viewer (IGV) [28] to
630 screen for candidate enhancers, identified visually where coverage of the cDNA sequence
631 track (solid lines) was greater than the baseline coverage of the plasmid sequence track
632 (dotted lines). The cDNA sequence track is analogous to light output from luciferase reporter
633 assays, but where the readout is enhancer-dependent RNA transcript output in *A. coluzzii*
634 4a3A cells, measured as sequenced cDNA, rather than by luciferase light output. Line color
635 (green, red, blue) represents three biological replicates. The enhancers are named by the
636 most proximal genes: Krueppel-Like Factor 6/7 (KLF, AGAP007038), Leucine-Rich Immune
637 protein (LRIM1, AGAP006348), Acetylcholinesterase (ACE1, AGAP001356), GABA-gated
638 chloride channel subunit (Rdl, AGAP006028), LIM homeobox protein 2/9, ortholog of
639 Drosophila apterous FBgn0267978 (AP, AGAP008980), and Ovo, AGAP000114 (Table1). A
640 negative control interval within intron 1-2 of distal-less (DLX, AGAP007058) was chosen
641 because it lacks visible divergence of cDNA and plasmid sequence tracks. Graphs display
642 cDNA and plasmid tracks in 10 kb windows centered on the candidate enhancers. Only one
643 candidate enhancer is seen in all windows except AP, where the central peak (arrow) was
644 used. X-axis indicates genomic coordinates in the PEST reference genome, y-axis indicates
645 normalized sequence depth corrected for overall plasmid depth observed in the IGV display.



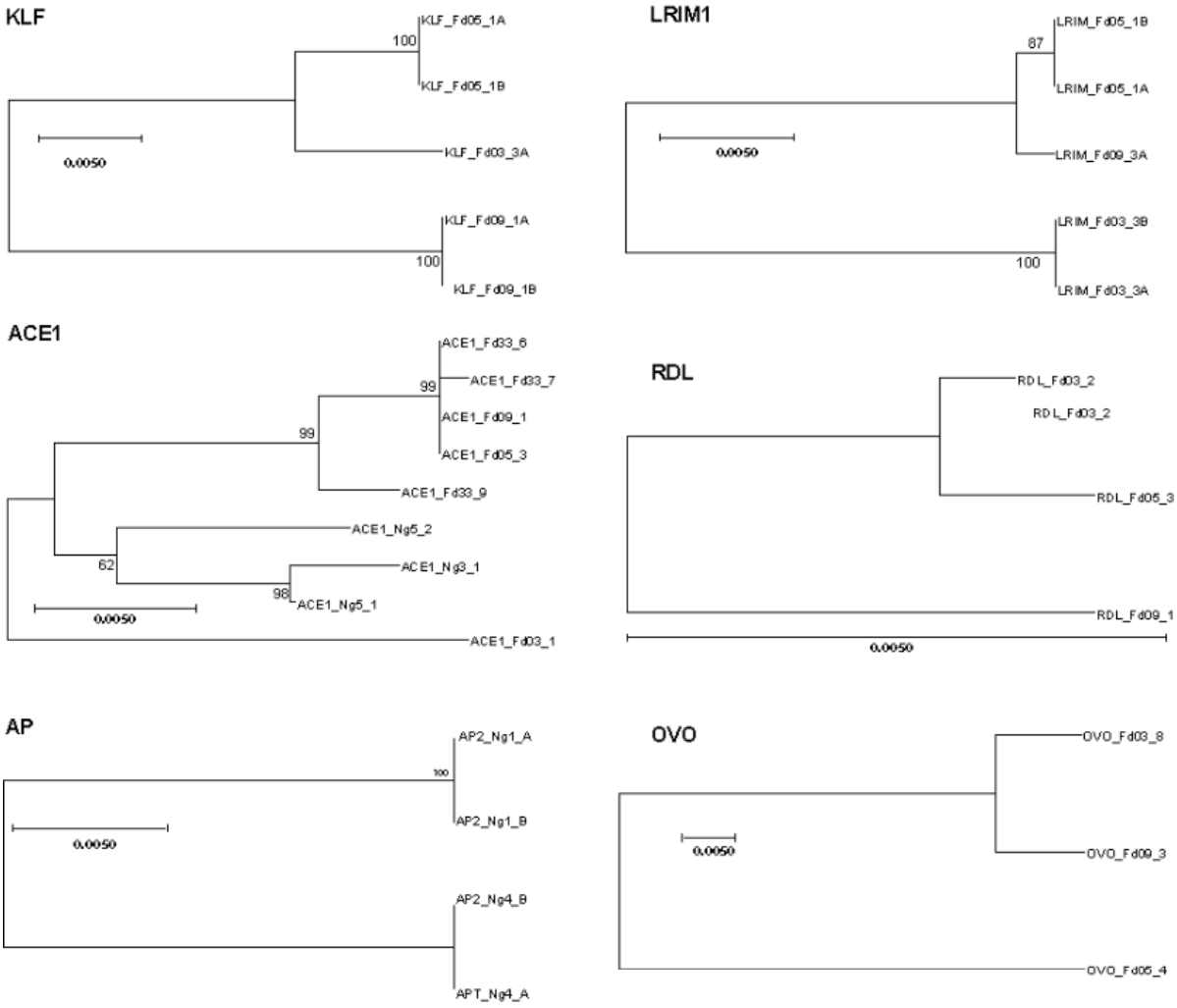
646

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



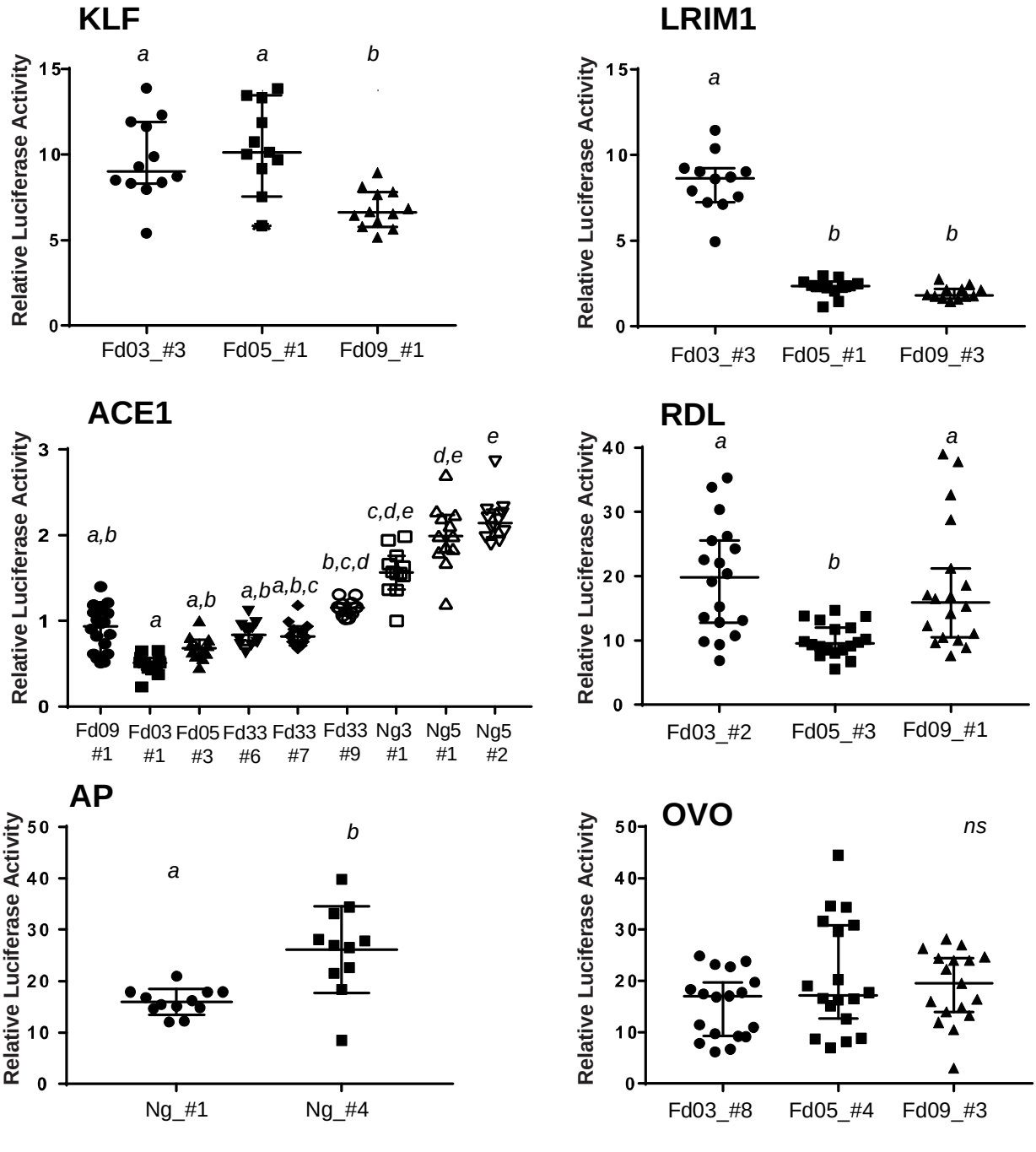
661

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



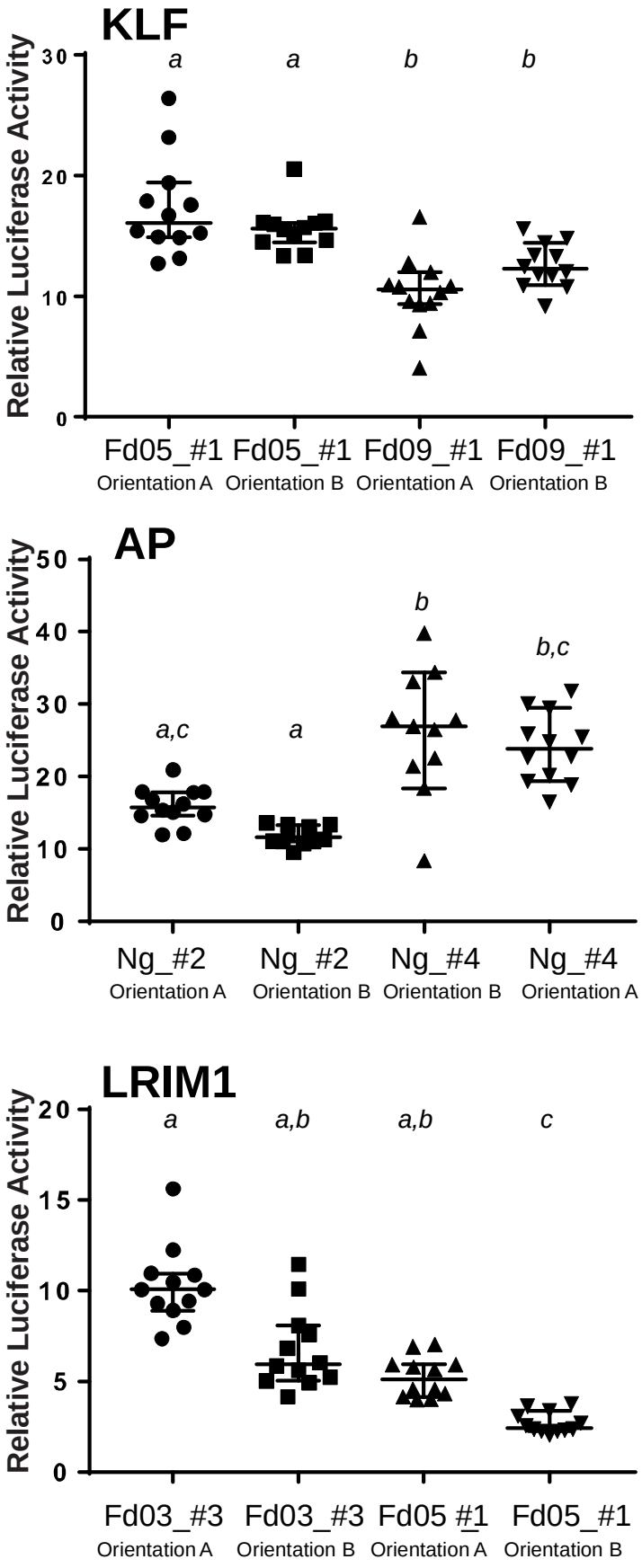
673
674

675 **Figure 4. Genetic variation influences enhancer functional activity**. To test the functional
676 effect of genetic variation within enhancer sequences, the enhancer alleles shown in Figure
677 3 were cloned into luciferase reporter plasmid pGL-Gateway-DSCP, and luciferase activity
678 was measured. Statistically significant differences in luciferase activity as determined using a
679 non-parametric ANOVA are indicated by letters, samples labelled with different letters are
680 significantly different from each other and samples with the same letter are not significantly
681 different (thus samples labeled a,b are not statistically different from samples labelled either
682 a or b). Bars indicate the median and 95% confidence intervals, n=12 for all tests. X-axis
683 labels indicate colony origin (Ng, Ngousso, other colony names as given) and allele name, y-
684 axis indicates the relative luciferase activity for each measurement determined as in Figure
685 2.



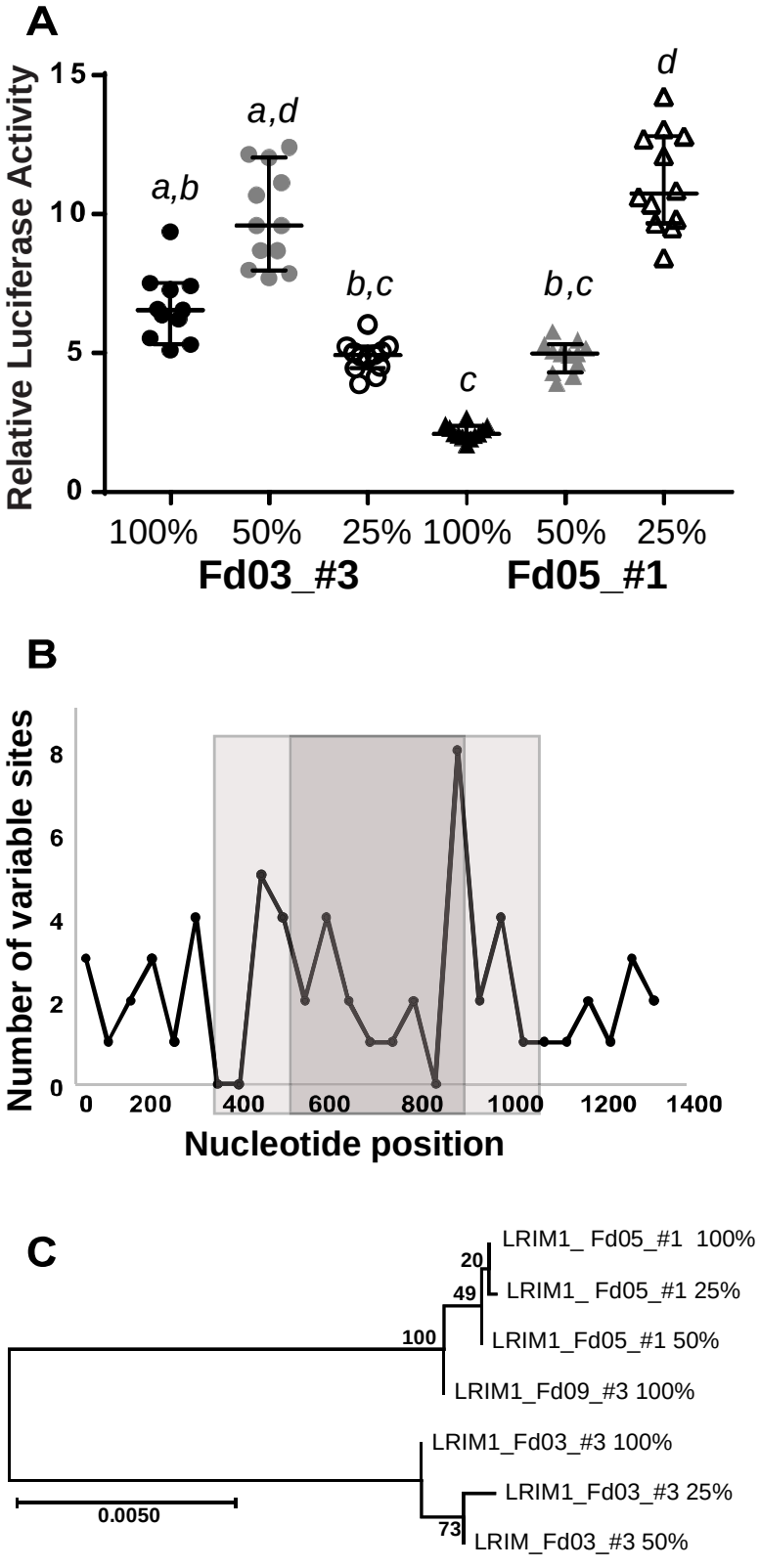
686
687

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



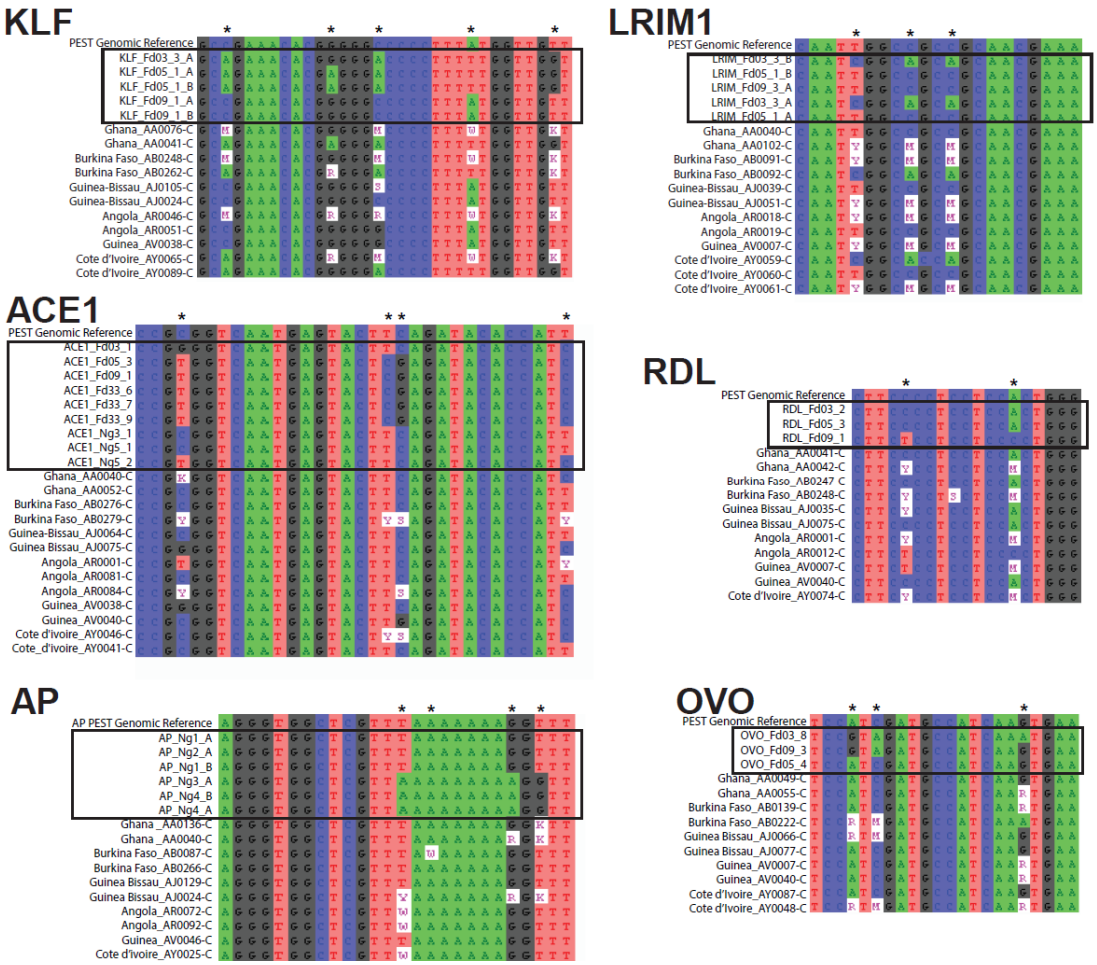
698

699 **Figure 6. Enhancer dissection reveals positive and negative regulatory elements.** Deletion
700 mutagenesis was carried out for two alleles of the LRIM1 enhancer, the high-activity allele
701 Fd03_#3 and low-activity allele Fd05_#1 (Figures 4 and 5). The integral enhancer alleles
702 (100%) were each deleted for one-quarter of their length from both termini (50%
703 derivative), and one-quarter length again (25% derivative). **A.** Deletion derivatives were
704 tested for luciferase activity, along with the original integral alleles. Statistical differences
705 indicated by letters as in Figure 4, error bars as in Figure 4, n=12 for all tests. X-axis indicates
706 allele name and deletion derivatives, y-axis indicates the relative luciferase activity for each
707 measurement as in Figure 2. Enhancer activity is not directly correlated with sequence
708 length, and enhancer alleles are structured from distinct combinations of positive and
709 negative regulators of transcription. **B.** Plot of the number of variant nucleotide positions
710 between the Fd03_#3 and Fd05_#1 alleles along the length of the enhancer sequence.
711 Variant sites are counted within a 50 bp non-overlapping window and plotted at the
712 midpoint of the window. The light gray shading indicates the extent of the 50% length
713 derivatives and the dark gray shading the 25% derivatives. X-axis indicates nucleotide
714 position in derivatives, y-axis indicates number of variable sites between the Fd03_#3 and
715 Fd05_#1 alleles in 50 bp windows. There were a total of 60 variable sites between Fd03_#3
716 and Fd05_#1 alleles in the 100% integral enhancer, 37 variable sites in the 50% derivatives
717 and 22 sites in the 25% derivatives. **C.** Neighbor-joining tree depicting sequence relatedness
718 between the integral 100% enhancer and the 50% and 25% derivatives for LRIM1 Fd03_#3
719 and Fd05_#1 alleles. The Fd09_#3 allele is included as an outgroup. Scale bar description as
720 in Figure 3.



721

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



733