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

Luisa Nardini, Inge Holm, Adrien Pain, Emmanuel Bischoff, Daryl Gohl, Soumanaba Zongo, Wamdaogo Guelbéogo, N'fale Sagnon, Kenneth D Vernick, Michelle Riehle

SCIENTIFIC REPORTS
nature research

OPEN

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

Luisa Nardini[1,2,8], Inge Holm[1,2,8], 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*]

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 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 enhancers segregated genetically polymorphic alleles, which displayed significantly different levels of functional activity. Deletion mutagenesis and functional testing revealed a fine structure of positive and negative regulatory elements that modulate activity of the enhancer core. Enhancer polymorphisms segregate in wild *A. coluzzii* populations in West Africa. Thus, enhancer variants that modify target gene expression leading to likely phenotypic consequences are frequent in nature. These results demonstrate the existence of naturally polymorphic *A. coluzzii* enhancers, which may help explain important differences between individuals or populations for malaria transmission efficiency and vector adaptation to the environment.

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. Core promoters are located proximal to the transcription start site and facilitate the binding of RNA polymerase and the initiation of transcription[1–3]. Enhancers control transcriptional activity of target genes through their interaction with activators and the promoter and are, in turn, responsible for most regulated gene expression in the transcriptome. The precise mechanisms of enhancer action is unknown and is an area of active study[4–6]. Enhancers can function at a distance from target genes and independent of their physical orientation in the chromosome[7]. 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[5,8–11]. Most of the significant variants mapped in human genome-wide association studies (GWAS) are noncoding[12], and at least 60–70% of significantly associated human

[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. [8]These authors contributed equally: Luisa Nardini and Inge Holm. *email: kvernick@pasteur.fr; mriehle@mcw.edu

GWAS single-nucleotide point mutations (SNPs) lie within functional enhancers[8,13]. In cancer studies, the majority of tumor-driving mutational changes are also thought to be in noncoding regulatory elements, especially enhancers[14].

Genetic variation in enhancers can occur as SNPs, insertions and deletions (indels), and as copy number variants[15–17]. Enhancer variation among individuals can underlie both Mendelian and complex traits[7,18,19]. At the population level, positively-selected variation in enhancers controlling key pathways likely plays an important role in differentiation and evolution[20,21]. 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[22].

Enhancer discovery in mosquitoes is limited to a few previous studies using indirect methods based on detection of chromatin properties to infer enhancer locations[23–25]. In genetic vector control strategies, it could be important to know the locations of enhancers that could cause unpredicted effects on transgene expression. Gal4 based enhancer trapping lines have been used in *Anopheles stephensi* to optimize methods to drive transgene expression[26]. Recent work in the African malaria vector *Anopheles funestus* highlights the effects of indel variation at a cis-regulatory element of a cytochrome P450 and its association with insecticide resistance[27]. Thus, relatively simple evolved sequence variation in enhancers can produce large phenotypic shifts in the organism[28,29]. Despite these 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.

Here, we analyze a 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 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[30].

We identified candidate *A. coluzzii* enhancers by manual examination of our generated sequence data from the functional screen near six selected genes mentioned below. We detected intervals where coverage of the cDNA sequence track, indicative of enhancer activity (Fig. 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 transfected plasmid (Fig. 1, dotted lines). The genes were selected because of predicted functions important for the biology of vectors, and because they were near potential enhancer signals. The genes are involved in vector immunity: Krueppel-Like Factor 6/7 (KLF, AGAP007038), Leucine-Rich Immune protein 1 (LRIM1, AGAP006348); insecticide resistance: Acetylcholinesterase 1 (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 (Table 1). The regulatory regions and enhancers of the latter two genes, apterous and Ovo, have been well characterized for the Drosophila orthologs[31–33]. 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, and thus no predicted enhancer function. The candidate enhancers are named according to the most proximal coding sequence, but it is not known whether the enhancers influence the expression of the proximal gene.

Functional validation of enhancer activity. The predicted candidate enhancers predicted in Fig. 1 were functionally tested to validate the enhancer 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 (Fig. 2). These results indicate that the enhancer predictions were accurate for all six predicted candidates, and thus validate these genomic intervals as functional *A. coluzzii* enhancers. The current information provides the first benchmark criteria that can be used for algorithmic genome-wide detection of *A. coluzzii* enhancers.

Screening for polymorphic alleles of validated enhancers. Having confirmed that all six predicted candidates are functional enhancers, we next wished to identify genetically variable alleles for each enhancer and measure their luciferase activity. For this, alleles of the enhancers were amplified and sequenced from *A. coluzzii*

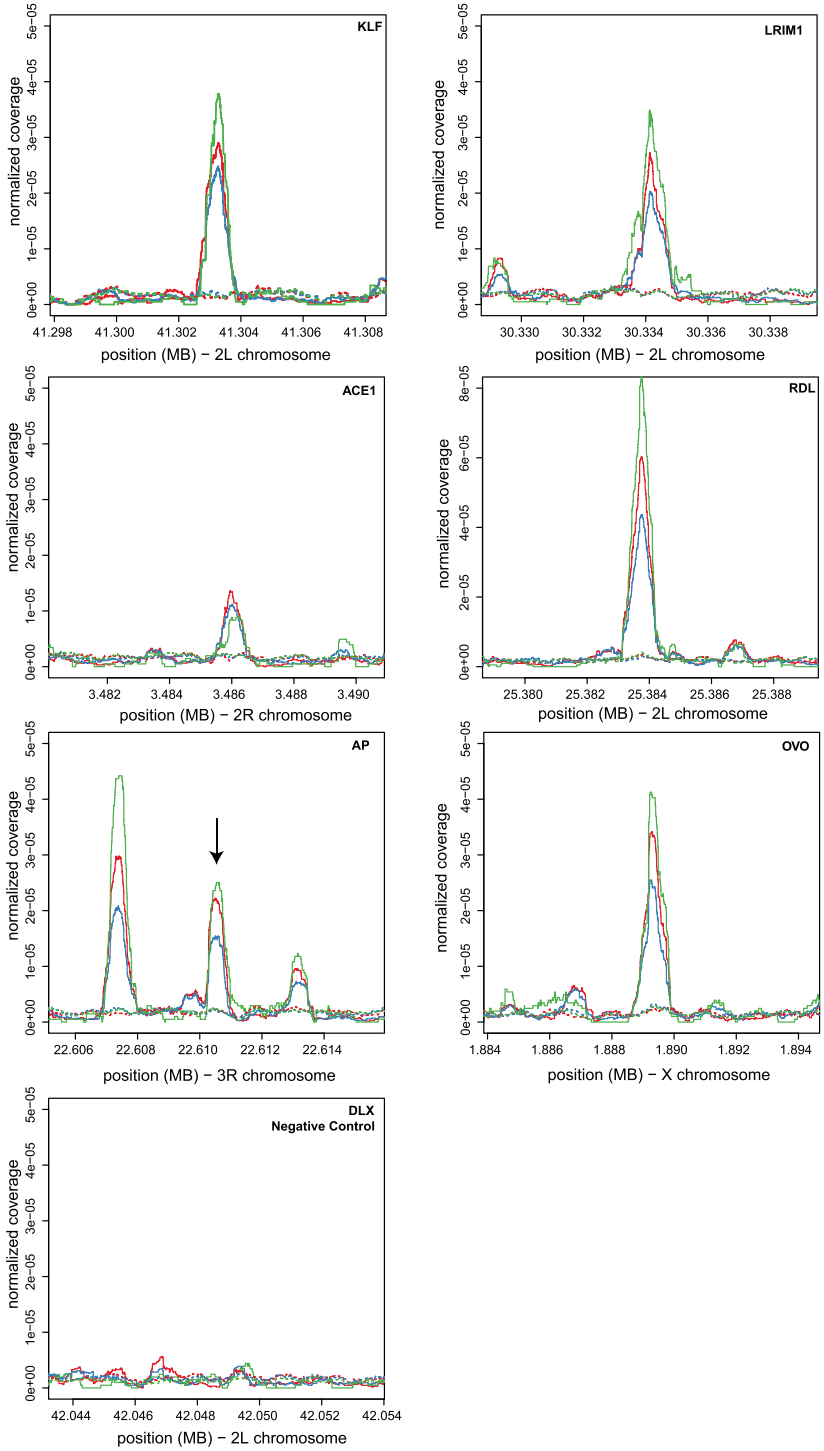

Figure 1. Detection of *Anopheles coluzzii* candidate enhancers. Sequence data near six *Anopheles coluzzii* genes were examined using the Integrative Genomics Viewer (IGV)[43] to screen for candidate enhancers. Enhancers were called where coverage of the cDNA sequence track (solid lines) was greater than the baseline coverage of the plasmid sequence track (dotted lines). The cDNA sequence track is analogous to light output from luciferase reporter assays. Line color (green, red, blue) represents three biological replicates. The enhancers are named by the most proximal genes: Krueppel-Like Factor 6/7 (KLF, AGAP007038), Leucine-Rich Immune protein 1 (LRIM1, AGAP006348), Acetylcholinesterase 1 (ACE1, AGAP001356), GABA-gated chloride channel subunit (Rdl, AGAP006028), LIM homeobox protein 2/9, ortholog of Drosophila apterous FBgn0267978 (AP, AGAP008980), and Ovo, AGAP000114 (Table 1). A negative control interval within intron 1–2 of distal-less (DLX, AGAP007058) was chosen because it lacks visible divergence of cDNA and plasmid sequence tracks. Graphs display cDNA and plasmid tracks in 10 kb windows centered on the candidate enhancers. Only one candidate enhancer is seen in all windows except AP, where the central peak (arrow) was used. X-axis indicates genomic coordinates in the PEST reference genome, y-axis indicates normalized sequence depth corrected for overall plasmid depth observed in the IGV display.

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

Table 1. Physical location of enhancers and proximal annotated gene. Enhancer coordinates are based on the locations of PCR primers given in Supplementary Table S2. Coordinates from the PEST AgamP4 genome assembly.

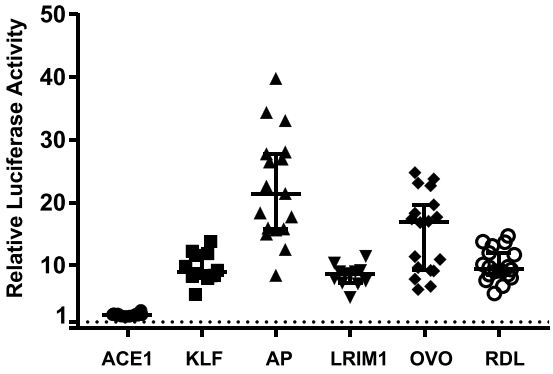

Figure 2. Candidate *Anopheles coluzzii* enhancers augment expression of a luciferase reporter. The six candidate enhancers from Fig. 1 were amplified from *Anopheles coluzzii* mosquitoes and cloned into the pGL-Gateway-DSCP plasmid carrying a basal core promoter and firefly luciferase reporter gene. 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).

colonies initiated from the populations in Cameroon, Mali or Burkina Faso. For each of the six enhancers, at least two distinct genetic variants were chosen for tests of enhancer activity. The goal was to identify and test a range of genetic variants, not the mosquito colonies. Therefore, a given colony may or may not be represented for a given enhancer, depending on the variation it segregates. The enhancer alleles were cloned and sequenced, and neighbor joining (N-J) trees depict the evolutionary relatedness and degree of sequence difference of the alleles (Fig. 3). Complete sequences for all tested enhancer alleles are presented in Supplementary File S1.

Genetic alleles of validated enhancers display distinct levels of enhancer activity. Luciferase activity was measured for all alleles to determine the effect of genetic variation on differences in functional enhancer activity. For five of the six enhancers, alleles displayed significantly different levels of enhancer activity (Fig. 4). For a given enhancer, alleles with the greatest difference in activity tended to be the most genetically different from each other (see also Fig. 3). For example, the alleles of the KLF enhancer cloned from colonies Fd05 and Fd03 are the most closely related genetically, and do not display a difference in luciferase activity as compared to the allele from colony Fd09. For two enhancers (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 enhancers. These results indicate that genetic alleles of validated enhancers can display significantly different levels of functional activity. The large observed differences in enhancer activity are expected to be translated into differential expression of the target genes that are regulated by the polymorphic enhancer alleles, likely leading to phenotypic differences in the organism related to the target gene functions. Enhancer allele outcomes for phenotype 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

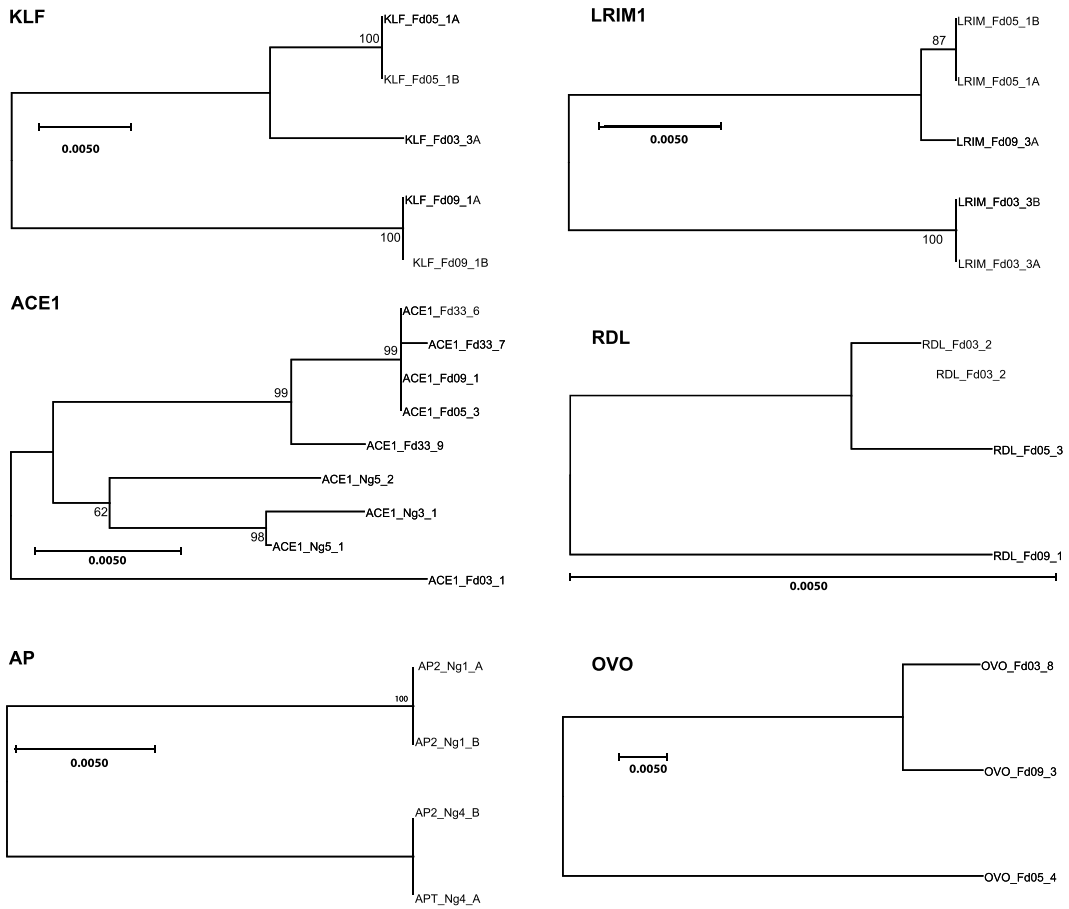

Figure 3. Phylogenetic comparison of enhancer allele genetic variation. Enhancer allelic variants were cloned and sequenced from *Anopheles coluzzii* colonies. Each sequenced clone represents a chromosomal haplotype. For each clone, individual sequences were aligned using MUSCLE and Mega was used to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

orientations in the reporter and measured luciferase activity. For the KLF and AP enhancers, there was no detectable effect of orientation (Fig. 5). The LRIM1 enhancer displayed a possible weak 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 Fig. 5), and thus the weak orientation difference for this one weak allele is not robustly supported. Thus, overall the enhancer alleles tested displayed function independent of their physical orientation with respect to the basal promoter.

Enhancer deletion mutagenesis reveals a modular structure of positive and negative regulatory elements. To resolve the minimal portion of the enhancer interval that carries the enhancer function, we carried out deletion mutagenesis for two different genetic alleles of the LRIM1 enhancer, one allele with high enhancer activity and the other low. The LRIM1 enhancer was used for proof of principle, and was chosen because it had alleles with distinct activity levels, and the genetic variation was spread across the enhancer. The deletion derivatives carried 50% or 25% of the length of the initial enhancer interval, reduced equally from both ends. We tested the deletion clones for luciferase activity, along with the original undeleted enhancer (Fig. 6A). Surprisingly, for LRIM1 allele Fd03_#3, the 50% construct displayed the highest luciferase activity, greater than either 100% or 25% constructs. This indicates that the integral 100% Fd03_#3 allele carries negative regulators of enhancer function, which were deleted in the 50% derivative to yield a derivative with elevated enhancer activity. The 25% derivative of allele Fd03_#3 displays significantly lower activity than the 50% derivative, suggesting that positive regulators of enhancer function are located outside the 25% derivative, but within the 50% derivative.

Deletion derivatives of LRIM1 allele Fd05_#1 display a pattern distinct from the Fd03_#3 allele. For Fd05_#1, each incrementally smaller derivative was more active. This result was also surprising, because it indicates that a highly active core enhancer element within the smallest 25% derivative is attenuated by negative regulators that

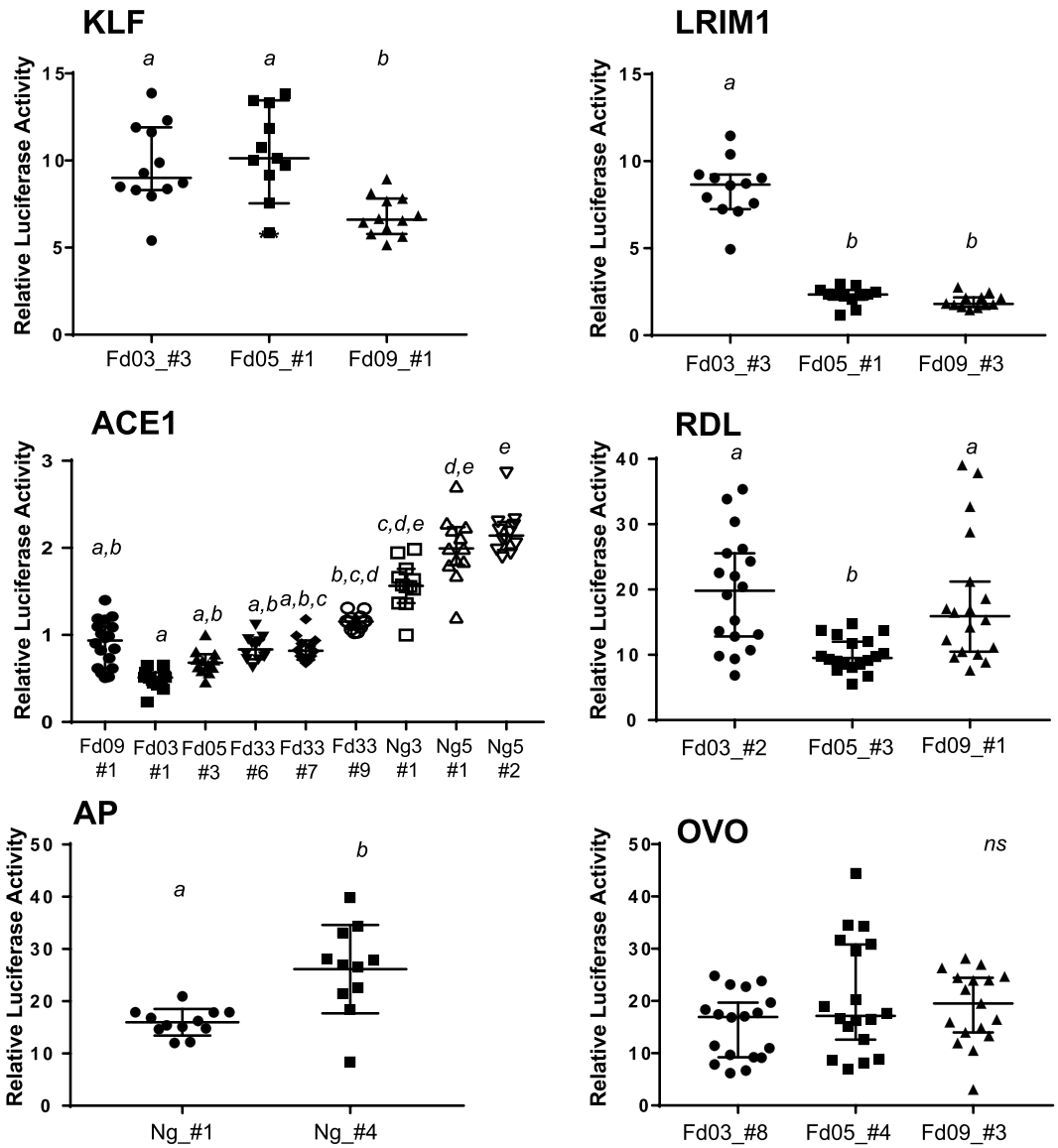

Figure 4. Genetic variation influences enhancer functional activity. To test the functional effect of genetic variation within enhancer sequences, the enhancer alleles shown in Fig. 3 were cloned into luciferase reporter plasmid pGL-Gateway-DSCP, and luciferase activity was measured. Statistically significant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

are progressively removed from 100% to 50% in length, and again from a 50% to 25% length interval. The deletion results indicate that enhancer activity is not directly correlated with sequence length, that there is a complex structure of functional elements and modifiers within the enhancer interval, and that different alleles of the same enhancer are comprised of distinct combinations of modular regulators that differentially influence transcription.

The density of variable sites between Fd03_#3 and Fd05_#1 varies across the interval, such that there were 60 variable nucleotide sites in the integral 100% length alleles, 37 variable sites in the 50% derivatives and 22 sites in the 25% derivatives (Fig. 6B). Finally, it is notable that the 25% derivative for allele Fd05_#1 displays activity levels indistinguishable from the Fd03_#3 50% derivative ($p = 0.99$), even though they are no more genetically similar than the integral 100% enhancer sequences for both alleles (Fig. 6C). This result highlights the relative independence of enhancer functional level from primary sequence patterns, unlike the fundamental dependence of protein coding gene function on the amino acid primary sequence code, and the consequent requirement for identification of enhancers by detecting functional activity. Testing of a large panel and manipulative experiments would be required to identify consistent patterns of enhancer modular organization.

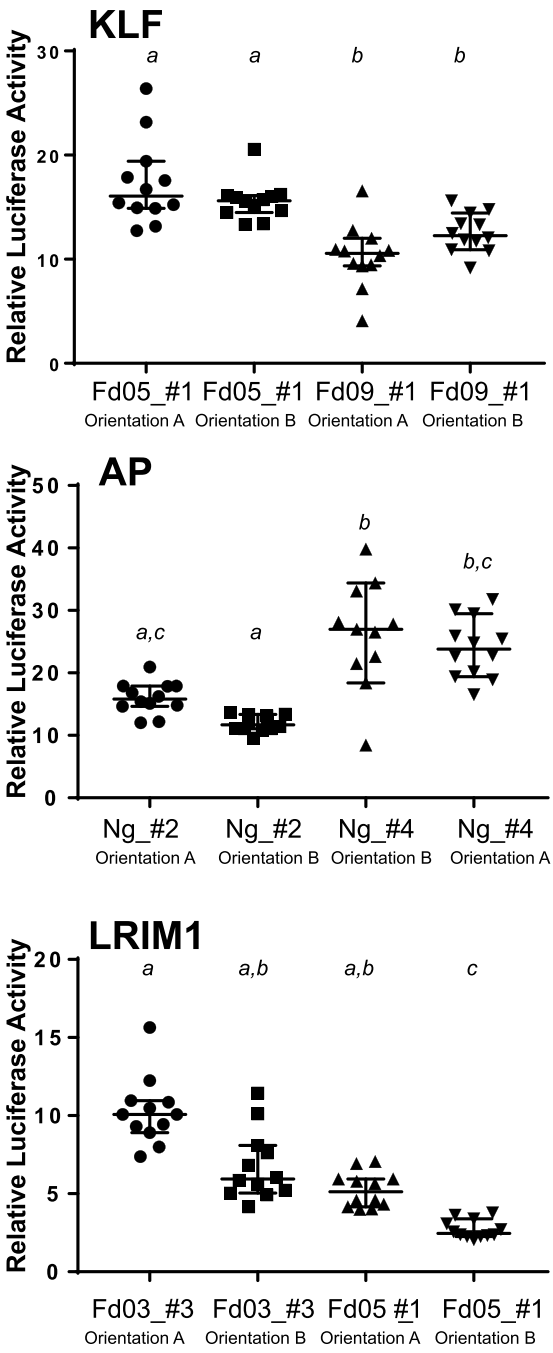

Figure 5. Enhancer activity is essentially independent of orientation. The influence of physical orientation of the enhancer within the luciferase reporter plasmid pGL-Gateway-DSCP was tested by cloning three enhancers, KLF, AP and LRIM1, in both orientations in the plasmid, and luciferase activity was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

Enhancer alleles segregate in the natural *Anopheles coluzzii* population. To confirm that the genetic variation observed in the enhancer alleles was natural and not an artifact of laboratory colonies, we compared sequence data for the six enhancers to genetic variation observed in wild *A. coluzzii* from whole genome sequence of the *Anopheles* gambiae 1000 (Ag1000) Genomes Consortium[34]. The comparison indicates that genetic variation is shared between the cloned *A. coluzzii* colony haplotypes used in luciferase assays and the natural population (Fig. 7). Representative short sequence alignments are shown, and full-length alignments with larger numbers of wild mosquitoes are presented in Supplementary File S2. Alignments are presented rather

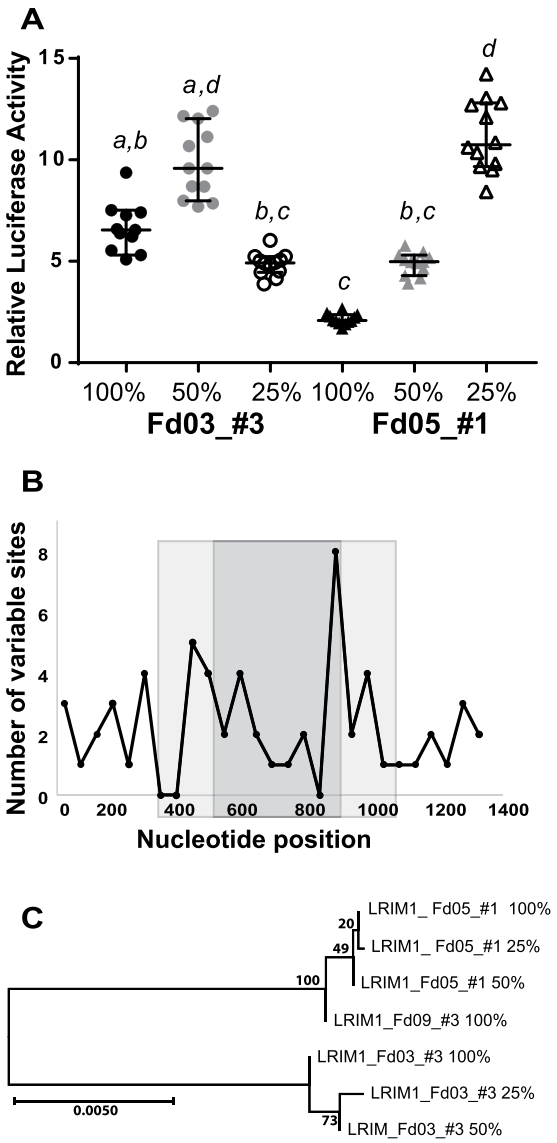

Figure 6. Enhancer deletion mutagenesis reveals positive and negative regulatory elements. Deletion mutagenesis was carried out for two alleles of the LRIM1 enhancer, the high-activity allele Fd03_#3 and low-activity allele Fd05_#1 (Figs 4 and 5). The integral enhancer alleles (100%) were each deleted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

than phylogenetic trees because the wild Ag1000 sequences were called for SNPs but not indels. Therefore, all Ag1000 sequences by default share the same indels as the PEST reference, and a tree would artifactually make all wild sequences appear more similar to one another. This analysis demonstrates that genetic variants within confirmed functional enhancers, associated with differential enhancer activity, segregate in nature and do not represent variants unique to lab colonies. Natural segregation of variants associated with differential enhancer

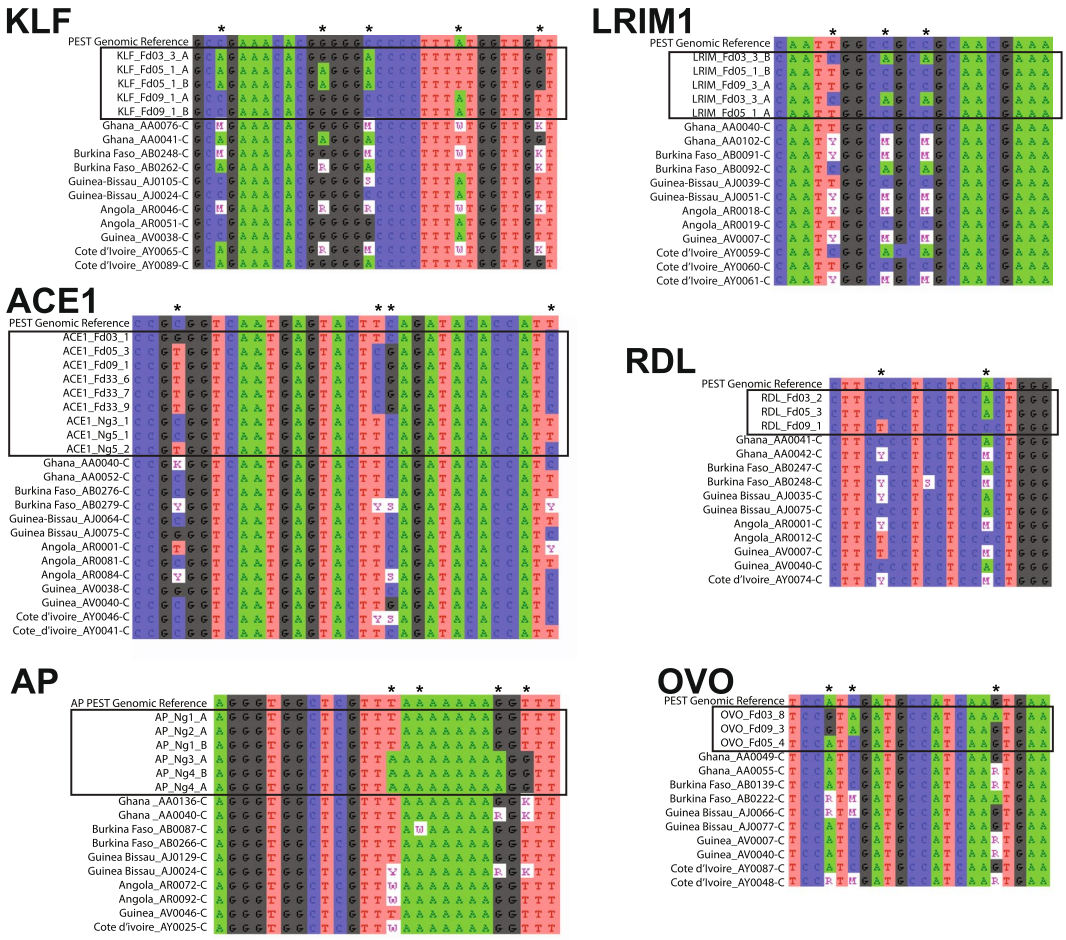

Figure 7. Genetic variants in differentially active enhancer alleles segregate in wild *Anopheles coluzzii*. Genetic variation observed in colonized and wild *A. coluzzii* was compared for the six studied enhancers. Representative short sequence alignments are shown (full-length alignments with additional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

function supports the interpretation that the differential function of enhancer alleles (Fig. 4) based on a modular structure of regulatory elements (Fig. 6) likely result from natural selection for distinct phenotypic outcomes of allelic enhancer function.

Discussion

The development of a methodology for screening and evaluation of *Anopheles* enhancers is an initial step towards a more comprehensive study of enhancers and their polymorphism effects. The currently examined enhancers were chosen because they were located near known functional genes, which were used to name the enhancers for convenience, but further work will be required to determine the actual influence, if any, of these enhancers upon the nearby genes, which is not known. Moreover, enhancer function is also controlled on spatial and temporal scales within the organism, and understanding *Anopheles* enhancers and their effects on phenotypes in detail will ultimately require incorporating this information.

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. 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, leading to differences in biological function and ultimately mosquito phenotype. 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. Most mosquito studies to date have focused solely on genes and proteins, rather than regulatory elements controlling the genes, in part due to the limited information available for the noncoding portion of the genome.

We detected variant alleles with significant difference in their functional enhancer activity, including functional null alleles that lack enhancer activity above background. For example, the LRIM1 enhancer Fd05_#1 allele or the ACE1 Fd03_#1 allele likely represent the ablation of an important transcription factor binding site, resulting in the absence of enhancer activity above background, which likely has downstream functional consequences. The range in functional enhancer activity that we observed is likely to affect phenotypes produced by the genes they transcriptionally regulate. Moreover, we demonstrate that variant alleles tested by luciferase activity in laboratory colonies also segregate in the wild population, and are therefore subject to natural selection. Thus, it is intriguing that selection has apparently generated a wide range of natural allelic forms of enhancers, including alleles that lack functional activity. This is consistent with the observation that genetic variation for enhancer function offers powerful raw material for adaptation and evolutionary change[11,20,35,36]. 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[34]. The current study is novel in that it addresses standing genetic variation for noncoding regulatory function, which is a poorly understood but likely important contributing factor influencing 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. By measuring functional activity of integral, 50% and 25% length derivatives of the intervals, we detected a modular structure of positive and negative regulators within the enhancer. Interestingly, deletion derivatives of the two alleles behaved differently, indicating that the outcome 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 were 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 removed 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 mutagenesis 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[37,38], or indirectly through loss of chromatin accessibility[17]. From the current results, we do not know whether the specific positive and negative modifiers found within the LRIM1 alleles are reused and combined to fine-tune the activity of different enhancers. If so, such regulator modules should be recognizable with enough genomic and functional data. Finally, the phenotypic implications of enhancer alleles will require determination of the target genes regulated by an enhancer, as well as the protein-DNA interactions underlying differential enhancer allele activity.

Vector control has been central to the malaria control effort by use of indoor residual spraying and long-lasting insecticide impregnated bednets. However, over-reliance on these methods has led to widespread insecticide resistance in wild populations, and novel methods of control are now required. The noncoding regulatory genome in *Anopheles* has the potential to provide novel new targets for vector control, but until now has not been interpretable or exploitable. For example, noncoding variants genetically associated with phenotypic traits could be interpreted by their functional enhancer activity. Consequently, enhancer variants could serve as markers for traits such as insecticide resistance, parasite susceptibility, behavior or adaptation to ecological conditions. Understanding the genomic enhancer landscape could improve the choice of insertion sites for exogenous transgenes for proper expression, and enhancers themselves could be genetically modified in order to alter transcription of immune or disease important genes with phenotypic consequences. The current work presents a necessary first step towards establishing an efficient, effective method for associating noncoding variation with 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[39], reared to adults, and were typed for species by the SINE200 X6.1 assay[40]. 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[30], cloned into the plasmid pSTARR-seq_fly (AddGene 71499), transformed into MegaX DH10B T1R Electrocomp Cells (Invitrogen), cultured in LB+ ampicillin (1 ug/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[41] were maintained on Insect X-Press media (Lonza) supplemented with 10% Fetal Calf Serum (heat inactivated at 56 °C for 30 minutes), at 27 °C. No antibiotics were used. We confirmed that cells were derived from *A. coluzzii* by species typing using the Fanello assay[42]. The plasmid DNA library was transfected and cultured in 4a3A cells as described[30] 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. The mRNA purified from cells was reverse transcribed using SuperScript III or IV First-Strand cDNA Synthesis System (Invitrogen) as described for the STARR-seq assay[30] using a plasmid-specific primer (RT_Rev, Supplementary Table S1), the cDNA was then amplified using primers Report_Fwd and Report_Rev (Supplementary Table S1), and the products were sequenced on an Illumina HiSeq. 2500 in 2×125 bp high output mode. Cell plasmid DNA was amplified and sequenced in the same manner as the cDNA samples but using primers Plasmid_Fwd and Plasmid_Rev (Supplementary Table S1).

Selection of enhancer candidates. The Integrative Genomics Viewer (IGV)[43] was used to select candidate enhancers by visual examination in the proximity of six annotated genes of interest. For determination of enhancer activity, the RNA output transcribed from the STARR-seq reporter plasmid, converted to cDNA and sequenced as described above, is compared to the levels of the plasmid DNA, to control for differential plasmid replication. Thus, candidate enhancers were predicted in intervals where coverage of the cDNA sequence track was visibly greater than the baseline coverage of the plasmid sequence track. The target genes examined were Krueppel-Like Factor 6/7 (KLF, AGAP007038), Leucine-Rich Immune protein 1 (LRIM1, AGAP006348), Acetylcholinesterase 1 (ACE1, AGAP001356), GABA-gated chloride channel subunit (Rdl, AGAP006028), LIM homeobox protein 2/9, ortholog of Drosophila apterous FBgn0267978 (AP, AGAP008980), and Ovo, AGAP000114. In addition, a negative control interval was cloned, which was a size-matched interval located within intron 1 of the gene, homeobox protein distal-less (DLX, AGAP007058) that displayed no visible divergence of cDNA and plasmid sequence tracks by IGV examination, and thus no predicted enhancer function. In all but one case, the candidate enhancer was the only one in the vicinity of the target gene, for AP there were three peaks, the most gene proximal peak is likely a promoter so the next most proximal peak was chosen as shown in Fig. 1. The candidate enhancers are named according to the most proximal coding sequence (above and Table 1).

Candidate enhancers were amplified from DNA of mosquitoes from the following *A. coluzzii* colonies: Ngousso, initiated in Cameroon in 2006[44], Fd03, Mali, 2008, Fd05, Mali 2008, Fd09, Burkina Faso, 2008, and Fd33, Burkina Faso, 2014. Fd colonies were previously described[45]. Primers are listed in Supplementary Table S2. Amplicons were cloned into the pCR8/GW/TOPO vector (Invitrogen) and sequenced with GW1 and GW2 primers. A standard plasmid was used for luciferase assays that incorporates the engineered Drosophila synthetic core promoter (DSCP). The information about the fine structure of Anopheles core promoters does not yet exist. The results indicate that the DSCP is functional in Anopheles cells. At least two genetically distinct sequences per candidate were then cloned into the firefly luciferase reporter plasmid pGL-Gateway-DSCP (AddGene 71506) using Gateway LR Clonase II (Invitrogen), transformed into OneShot OmniMax 2T1 Phage-Resistant Cells (Invitrogen), and plasmid was purified from overnight culture.

To test the effect of enhancer orientation, the enhancer was cloned in the opposite orientation in pGL-Gateway-DSCP. To test resolved enhancers, the relevant enhancer insert was amplified with primers that generated either 50% or 25% of the initial insert size, equally reduced on both ends, and products were cloned in pGL-Gateway-DSCP. In all cases, plasmids were resequenced to confirm insert identity using the primers LucNrev and RVprimer3 (Supplementary Table S1).

Quantitation and statistical analysis of enhancer activity by luciferase assay. The Dual-Glo Luciferase Assay System (Promega) was used for luciferase assays. *A. coluzzii* 4a3A cells were seeded in 96 well plates at 1×10^5 cells/well, the difference in volume if any 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, 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 same enhancer, complete sequences were aligned using MUSCLE within the package Molecular Evolutionary Genetics Analysis Mega version X[46], and N-J trees constructed using Mega. When at least four variants were tested, bootstrapping was performed and bootstrap values are included on N-J trees. Scale bars of trees represent 0.5, but each bar is a different length. The longer the 0.5 scale bar, the more genetically similar the sequences. A workflow summary figure is shown in Supplementary Fig. S1.

Analysis of wild *Anopheles* variation data. Sequence information for 309 wild *A. coluzzii* from 6 West African countries; Angola (AR), Burkina Faso (AB), Cote d'Ivoire (AY), Ghana (AA), Guinea (AV) and Guinea-Bissau (AJ), generated as reported[34] were downloaded from MalariaGen (https://www.malariagen.net/projects/ag1000g) as VCF files from the Ag1000G phase 2 AR1 data release and sequences of the six validated enhancers were extracted from the raw VCF files using GATK version 3.9 (SelectVariants mode)[47], the consensus sequences with IUPAC ambiguity codes for the variants were extracted using BCFtools version 1.9 (consensus mode with –iupac-codes option)[48], the sequences of each interval were aligned with Clustal W version 2.1[49] and visualized with AliView aligner version 1.24[50].

The diploid sequences from the wild sequences were aligned to the cloned sequences generated from the six validated enhancers. Sequence alignments were visually examined for shared variation. Short representative sequence alignments are presented in Fig. 7 (not including indels), and complete alignments relative to the PEST AgamP3genome assembly, including indels, are presented in Supplementary File S2. Indel genotypes of the wild sequences shown in Supplementary File S2 are relative to the PEST reference haplotype, because the wild sequences were called for SNPs but not indels[34].

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 PRJEB34434. All other sequences are available in this article as Supplementary Files S1 and S2.

Received: 19 May 2019; Accepted: 7 October 2019;
Published online: 24 October 2019

References

1. Ong, C. T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**, 283–293, https://doi.org/10.1038/nrg2957 (2011).
2. Moreau, P. *et al*. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic acids research* **9**, 6047–6068, https://doi.org/10.1093/nar/9.22.6047 (1981).
3. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* **19**, 621–637, https://doi.org/10.1038/s41580-018-0028-8 (2018).
4. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23, https://doi.org/10.1016/j.cell.2017.02.007 (2017).
5. 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, https://doi.org/10.1101/gr.136127.111 (2012).
6. Chen, H. *et al*. Dynamic interplay between enhancer-promoter topology and gene activity. *Nat Genet* **50**, 1296–1303, https://doi.org/10.1038/s41588-018-0175-z (2018).
7. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat Rev Genet* **14**, 288–295, https://doi.org/10.1038/nrg3458 (2013).
8. Farh, K. K. *et al*. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, https://doi.org/10.1038/nature13835 (2014).
9. Kharchenko, P. V. *et al*. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**, 480–485 (2011).
10. 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).
11. 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).
12. MacArthur, J. *et al*. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research* **45**, D896–D901, https://doi.org/10.1093/nar/gkw1133 (2017).
13. Neph, S. *et al*. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90, https://doi.org/10.1038/nature11212 (2012).
14. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160–1165, https://doi.org/10.1038/ng.3101 (2014).
15. Heinz, S. *et al*. Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487–492, https://doi.org/10.1038/nature12615 (2013).
16. 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, https://doi.org/10.1038/ng.3911 (2017).
17. Jacobs, J. *et al*. The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. *Nat Genet* **50**, 1011–1020, https://doi.org/10.1038/s41588-018-0140-x (2018).
18. 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, https://doi.org/10.1242/dev.01613 (2005).
19. Smemo, S. *et al*. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum Mol Genet* **21**, 3255–3263, https://doi.org/10.1093/hmg/dds165 (2012).
20. 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, https://doi.org/10.1038/ng.3009 (2014).
21. Vierstra, J. *et al*. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007–1012, https://doi.org/10.1126/science.1246426 (2014).
22. Franchini, L. F. & Pollard, K. S. Can a few non-coding mutations make a human brain? *Bioessays* **37**, 1054–1061, https://doi.org/10.1002/bies.201500049 (2015).
23. Behura, S. K. *et al*. High-throughput cis-regulatory element discovery in the vector mosquito Aedes aegypti. *BMC genomics* **17**, 341, https://doi.org/10.1186/s12864-016-2468-x (2016).

24. 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, https://doi.org/10.1186/s13071-018-2982-6 (2018).
25. Ruiz, J. L. *et al.* Chromatin changes in Anopheles gambiae induced by Plasmodium falciparum infection. *Epigenetics Chromatin* **12**, 5, https://doi.org/10.1186/s13072-018-0250-9 (2019).
26. O'Brochta, D. A., Pilitt, K. L., Harrell, R. A. 2nd, Aluvihare, C. & Alford, R. T. Gal4-based enhancer-trapping in the malaria mosquito Anopheles stephensi. *G3* **2**, 1305–1315, https://doi.org/10.1534/g3.112.003582 (2012).
27. Weedall, G. D. *et al.* A cytochrome P450 allele confers pyrethroid resistance on a major African malaria vector, reducing insecticide-treated bednet efficacy. *Sci Transl Med* **11**, https://doi.org/10.1126/scitranslmed.aat7386 (2019).
28. McGregor, A. P. *et al.* Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature* **448**, 587–590, https://doi.org/10.1038/nature05988 (2007).
29. 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, https://doi.org/10.1073/pnas.0700488104 (2007).
30. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077, https://doi.org/10.1126/science.1232542 (2013).
31. Bieli, D. *et al.* The Drosophila melanogaster Mutants apblot and apXasta Affect an Essential apterous Wing Enhancer. *G3* **5**, 1129–1143, https://doi.org/10.1534/g3.115.017707 (2015).
32. Preger-Ben Noon, E. *et al.* Comprehensive Analysis of a cis-Regulatory Region Reveals Pleiotropy in Enhancer Function. *Cell reports* **22**, 3021–3031, https://doi.org/10.1016/j.celrep.2018.02.073 (2018).
33. Bieli, D. *et al.* Establishment of a Developmental Compartment Requires Interactions between Three Synergistic Cis-regulatory Modules. *PLoS Genet* **11**, e1005376, https://doi.org/10.1371/journal.pgen.1005376 (2015).
34. Anopheles gambiae Genomes, C. *et al.* Genetic diversity of the African malaria vector Anopheles gambiae. *Nature* **552**, 96–100, https://doi.org/10.1038/nature24995 (2017).
35. Kvon, E. Z. *et al.* Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167**, 633–642 e611, https://doi.org/10.1016/j.cell.2016.09.028 (2016).
36. Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**, 302–305, https://doi.org/10.1126/science.1182213 (2010).
37. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040, https://doi.org/10.1126/science.1186176 (2010).
38. 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, https://doi.org/10.1098/rstb.2012.0360 (2013).
39. Riehle, M. M. *et al.* A cryptic subgroup of Anopheles gambiae is highly susceptible to human malaria parasites. *Science* **331**, 596–598, https://doi.org/10.1126/science.1196759 (2011).
40. Santolamazza, F. *et al.* Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms. *Malaria journal* **7**, 163, https://doi.org/10.1186/1475-2875-7-163 (2008).
41. 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).
42. 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).
43. 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, https://doi.org/10.1093/bib/bbs017 (2013).
44. Harris, C. *et al.* Polymorphisms in Anopheles gambiae immune genes associated with natural resistance to Plasmodium falciparum. *PLoS pathogens* **6**, e1001112, https://doi.org/10.1371/journal.ppat.1001112 (2010).
45. Redmond, S. N. *et al.* Association mapping by pooled sequencing identifies TOLL 11 as a protective factor against Plasmodium falciparum in Anopheles gambiae. *BMC genomics* **16**, 779, https://doi.org/10.1186/s12864-015-2009-z (2015).
46. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547–1549, https://doi.org/10.1093/molbev/msy096 (2018).
47. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297–1303, https://doi.org/10.1101/gr.107524.110 (2010).
48. Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751, https://doi.org/10.1093/bioinformatics/btw044 (2016).
49. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948, https://doi.org/10.1093/bioinformatics/btm404 (2007).
50. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278, https://doi.org/10.1093/bioinformatics/btu531 (2014).

Acknowledgements

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

Conceived and designed the experiments: M.M.R., D.M.G. and K.D.V. Performed the experiments: L.N., I.H., D.M.G., S.Z., W.M.G., N.S. and M.M.R. Analyzed the data: A.P., E.B., M.M.R. Wrote the manuscript: L.N., M.M.R. and K.D.V. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-51730-8.

Correspondence and requests for materials should be addressed to K.D.V. or M.M.R.