Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Karyotypes of the Anopheles 2La chromosomal inversion carry functionally distinct transcriptional enhancers

Kathryn S Taquet
Medical College of Wisconsin
Adrien Pain
Institut Pasteur, Universite Paris Cite
Cameron E Anderson
Medical College of Wisconsin
Christian Mitri
Institut Pasteur, Université Paris Cité
Kenneth D Vernick
Institut Pasteur, Université Paris Cité
Michelle M Riehle
mriehle@mcw.edu

Medical College of Wisconsin

Research Article

Keywords: Chromosomal inversion, enhancers, genetic and structural variation

Posted Date: September 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10870879/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Chromosome inversions are an adaptive mechanism to capture an allelic block in a structural feature that inhibits recombinational separation, thus forming a stable haplotype of co-evolved alleles that are selected as a unit. In the Anopheles gambiae species complex of mosquitoes, the 2La inversion comprises ~ 10% of the genome, and the inverted and wild-type forms are strongly associated with phenotypes related to malaria transmission efficiency. Transcriptional enhancers control the expression of suites of genes in response to cellular signals. Enhancer variation private to inversion alleles, if it leads to inversion-specific enhancer regulatory function, could be an important mechanism underlying inversion phenotypic differences.

Results

Here we employ two independent and complementary methods, specifically, genome assembly based and population resequencing based, to identify candidate chromosomal inversion-specific enhancers in a large paracentric inversion segregating in the *Anopheles coluzzii* malaria vector. Using comparison of whole genome assemblies differing in their 2La inversion karyotype (PEST (2L + ^{a}), MOPTI (2La)), we identified twenty-eight candidate karyotype-specific enhancers. Using Ag1000G population resequencing data to screen for inversion allele-fixed SNPs within enhancers, we identified eight candidate karyotype-specific enhancers. Criteria filters yielded nine candidate inversion-specific enhancers that were functionally examined by mutagenesis and enhancer activity assays. Two candidates displayed clear inversion-specific functional difference. For one of these, inversion-specific variation was both necessary and sufficient to explain differences in enhancer activity, while for the other, inversion-specific variation, while necessary, was not sufficient to fully explain differences in enhancer activity between the inversion-specific enhancer alleles.

Conclusions

Inversion-specific enhancers were found for which the two allelic enhancer forms display functionally distinct activities. Functional differences of these, and other unknown, inversion-specific enhancers, are likely to lead to karyotype differences in gene regulation profiles, and, in the aggregate, probably at least partly explain the known large phenotypic differences among carriers of the two 2La karyotypes. This knowledge could inform vector control methods specifically directed at a particular segment of the vector population, particularly those that differ in phenotypes affecting malaria transmission such as 2La-associated indoor or outdoor-resting behavior.

Background

Analysis of genome-wide biological features, for example transcript abundance, genetic regulatory elements, and chromatin accessibility, requires mapping of sequence reads to a defined reference genome. Many experimental systems define a single reference genome for ease of analysis and the ability to compare across studies. However, use of a single reference genome will, by definition, introduce an ascertainment bias as it is blind to variable regions segregating in the population that are absent or significantly diverged from the reference genome[1–3].

Along with single nucleotide polymorphisms (SNPs) and insertion/deletion (indel) events, chromosomal inversions are a significant source of genetic variation and serve diverse roles in the evolution of phenotypic traits across kingdoms[4–6]. For example, chromosomal inversions affect body size in fruit flies [7], the adaptation to freshwater in sticklebacks [8], hemophilia A pathogenesis in humans[9], and the relationship between stick insects and their host plants[10]. Limited recombination in inversion heterozygotes results in genetic variation private to a single inversion karyotype and the consequent establishment of co-adapted gene complexes of closely linked loci that are inherited together.[4, 11]. In addition, the physical rearrangement of the inversion can cause rewiring of genes and their regulatory elements at the inversion boundaries due to the physical switching of their genomic neighborhoods [12–14].

Given the lack of gene flow across chromosomal inversion allelic forms (called karyotypes) due to inhibition of recombination, chromosomal segments of opposite orientation evolve largely independently of one another which results in the accumulation of mutations private to a single inversion karyotype leading to changes in gene expression patterns, and ultimately phenotypes often driven by selection pressures favoring sets of alleles and their potentially diverged regulatory regimes [15].

Chromosomal inversions are prevalent within dipterans [16] and are associated with specific traits within the major vectors of malaria, *Anopheles coluzzii* and *Anopheles gambiae*[6]. Within *Anopheles*, alternate inversion forms have been shown to influence traits important in pathogen transmission including resting behavior [17], insecticide resistance [18, 19], adaptation to aridity and thermotolerance [20–23] as well as demonstrating an intrinsic difference in susceptibility to malaria infection[24, 25]. Thus, *Anopheles* inversions widen the phenotypic and ecological spectrum exhibited among individuals of the same nominal species, thereby expanding malaria transmission both seasonally and geographically.

Given that most work performed on a genomic scale or -omics level utilizes a reference genome for the mapping of sequence reads and the calling of single nucleotide variants, the ascertainment of regulatory elements private to or significantly diverged between genomes may be biased. Reference genomes are by definition homozygous for a particular inversion allele and as a result, sequences absent or significantly diverged from the reference genome cannot be detected in analyses based on the reference.

Taken together, the accumulation of genetic variation private to a particular inversion karyotype may lead to the evolution of novel regulatory elements or the elimination of ancestral ones. However, without suitable reference genomes representing alternate inversion karyotypes, read mapping of genomics data aimed to identify the regulatory genome (e.g., STARR-seq, ATAC-seq, ChIP-seq), may be invisible to significant genetic variation between inversion forms.

Here we employ two discrete, yet complementary, computational methods using existing data to identify novel and divergent regulatory sequences between *Anopheles coluzzii* mosquitoes homozygous for the ancestral (2La) or derived (2L + ^{a}) form of the large paracentric 2La inversion. In the first approach, we map previously published STARR-seq data from *Anopheles coluzzii* [26] to the MOPTI reference genome (2La/2La) to identify candidate regulatory elements unique to the 2La inversion orientation, by comparison to the same STARR-seq reads mapped to the PEST (2L + ^{a} /2L + ^{a}) reference. In the second approach, we use publicly available *A. gambiae* and *A. coluzzii* sequence data from the Ag1000G project to identify differentially fixed nucleotide sites across mosquitoes homozygous for alternate 2La inversion karyotypes and overlay STARR-seq enhancer coordinates to identify STARR-seq enhancer sequences with inversion-specific SNPs. Each method has its own set of unique ascertainment biases, and the coupling of approaches limits the impact of these acknowledged limitations. For nine inversion-specific enhancers within the 2La inversion, experimental validation was performed to confirm enhancer activity for inversion associated alleles and to identify nucleotide positions necessary and sufficient for observed differences in enhancer activity.

Methods

Mapping of STARR-seq enhancers to the MOPTI 2La/2La genome

STARR-seq reads generated and previously analyzed relative to the A. gambiae 2L + ^{a}/2L + ^{a} homokaryotic PEST genome (VectorBase-59_AgambiaePEST_Genome, [26]), were here mapped to the A. coluzzii MOPTI 2La/2La genome ((VectorBase-59_AcoluzziiMOPTI_Genome) using BWA-MEM [27] with default parameters to identify candidate enhancers specific to the MOPTI genome. As the input material for the STARR-seq was a pool of 30 2La/2La females, we reasoned that mapping to a 2La/2La genome could highlight regulatory regions novel to or highly diverged in the 2La karyotype. During the conversion of SAM files to BAM files, two filters were applied to remove unmapped reads and supplementary alignments (-F 4 -F 2048). The mapping pipeline to call STARR-seq peaks is identical to that described previously [26] .

In contrast to previous work which defined candidate enhancers as those detected at the intersection of in all three biological replicates, here, to avoid missing potential candidates, we initially considered the union of peaks across all replicates. The small number of candidates chosen for functional validation were supported by at least two biological replicates, ensuring higher confidence.

To identify candidate regulatory regions specific to one inversion arrangement, we compared the STARR-seq peak sets identified in the PEST and MOPTI reference genomes. Enhancer sequences from both genomes were combined and compared in an all-against-all BLASTN search using BLAST+ v2.16.0 with the following parameters: e-value < 1e-10, percent identity ≥ 80, max_target_seqs = 100. Self-hits and within-genome hits were removed, and groups were curated to avoid counting duplicated enhancer regions multiple times. Enhancers with a significant reciprocal hit were classified as shared; enhancers lacking a significant hit in the other peak set were classified as reference specific. Overall mapping rates of STARR-seq reads were highly similar between the PEST 2L + a/2L + a and MOPTI 2La/2La genomes (∼1% difference overall; ∼1–2% difference for properly paired reads), with marginally higher rates against MOPTI, as expected given that the input material was derived from 2La/2La females. The absence of a large global difference in mapping rates indicates that the detected differences in peak sets are not driven by a genome-wide mapping bias. Rather, the approach specifically leverages local ascertainment effects: reads originating from regions that are divergent or rearranged between the 2La and 2L + ^{a} arrangements are expected to align better to the reference genome matching the sample karyotype. Such locally improved alignments can reveal regulatory regions that are absent or poorly represented in a single reference genome, and these regions are the focus of our downstream analysis.

Pilot assessment of ascertainment bias from limited reference genomes

The 3 candidate inversion-specific enhancers identified through mapping to the MOPTI genome, comprised of unique sequence and detected in at least 2 of 3 STARR-seq replicates were chosen for further follow-up (Table 1). For these three candidates, primers were designed from the PEST and MOPTI genomes to amplify the inversion-specific enhancers from genomic DNA isolated from at least 2 wild mosquitoes homozygous for the 2La or 2L + ^{a} karyotypes of the 2La inversion. Wild mosquito DNA was derived from samples collected along a Mali to Guinea transect [28]. We extended observations beyond the two high-quality reference genomes to collect additional evidence from wild mosquito samples to determine whether variation among the PEST and MOPTI genomes was associated with the 2La inversion or instead reflected ascertainment bias arising from the two reference genomes.

Table 1

Candidate inversion-specific enhancers identified by mapping STARR-seq sequencing reads against the MOPTI genome. Dfam in the sequence descriptio similar to a predicted transposable/mobile element catalogued in the Dfam database. Enhancers that were explored functionally are noted in bold (E

STARR-seq enhancer	Mopti Chromosome	Start position	End position	STARR-seq reps	Sequence description	% identity	length (bp)	mismatch	g
1	CM029348	86701046	86701581	1 and 2 and 3	Acol_Pao_Bel_Ele49 Dfam				
2	CM029348	90912755	90913291	1 and 2 and 3	Acol_Pao_Bel_Ele49 Dfam				
3	**CM029348**	**94819982**	**94820673**	**1 and 2 and 3**	**Unique sequence**	**92.784**	**485**	**12**	**9**
4	CM029348	100838386	100838916	1 and 2 and 3	Acol_Pao_Bel_Ele74 Dfam				
5	**CM029348**	**102479076**	**102480092**	**1 and 2 and 3**	**Unique sequence**	**91.016**	**1024**	**42**	**1**
6	CM029348	100247096	100247677	1 and 2	Acol_CR1_Ele23 Dfam				
7	**CM029348**	**102537823**	**102538392**	**1 and 2**	**Unique sequence**	**94.751**	**362**	**12**	**5**
8	CM029348	102657110	102657665	1 and 2	Acol_Tc1-mar_Ele3 Dfam				
9	CM029348	103838862	103839465	1 and 2	Acol_gypsy_Ele65 Dfam				
10	CM029348	95346863	95347366	1 and 3	Acol_Pao_Bel_Ele49 Dfam				
11	CM029348	86147395	86147897	2 and 3	Repetitive sequence of TATCTCTC				
12	CM029348	100785459	100786005	2 and 3	Repetitive sequence of AGAGAGAT				
13	CM029348	97051436	97051941	2 and 3	Repetitive sequence TTTTTCAAACTCAATAAACTCAACAAACTCAC				

After sequencing all 3 candidate inversion-specific enhancers, sequences were trimmed and aligned using MEGA [29]. For each of the 3 candidate inversion-specific enhancers, five genetically variable regions (defined as regions of > 5bp in length with at least 3 SNPS or an indel) between the PEST and MOPTI reference genomes were selected and screened for 2La inversion-specificity using sequence generated from 2La and 2L + ^{a} homozygous wild mosquito. For candidate inversion-specific enhancers ISEN2L_33.9MB and ISEN2L_25.9MB, zero of the 5 regions displayed genetic variation in wild 2La and 2L + ^{a} mosquito samples consistent with the karyotypes of the PEST and MOPTI reference genomes indicating that these are unlikely to be inversion-specific enhancers. However, for enhancer ISEN2L_25.8MB, 3 of the 5 regions contained fixed variants segregating with the 2La inversion karyotype in wild mosquitoes, including a large (32–40 bp) indel. ISEN2L_25.8MB was chosen for functional characterization to determine the role of the inversion-specific variation in underlying enhancer activity.

Categorizing SNPs from Ag1000G phase 1 AR3 variant data

Genotype frequencies and allele frequencies from Ag1000G phase 1 AR3 variant data [30] for chromosome 2L were used to understand SNP composition. Ag1000G phase 1 AR3 variant data includes 246 2La/2La mosquitoes (492 alleles), 195 2La/2L + ^{a} mosquitoes (390 alleles), and 324 2L + ^{a}/2L + ^{a} mosquitoes. Of the 246 2La/2La mosquitoes, there are 175 that are *A. gambiae* (from Guinea (n = 6), Burkina Faso (n = 66), Cameroon (n = 48), Uganda (n = 40), and Kenya (n = 15)), 68 that are *A. coluzzii*, all from Burkina Faso, and 3 that are admixed from Guinea-Bissau. Of the 195 2La/2L + ^{a} mosquitoes, there are a 172 that are *A. gambiae* (from Guinea (n = 12), Burkina Faso (n = 15), Cameroon (n = 71), Uganda (n = 48), Kenya (n = 22), and Gabon (n = 4)), 3 that are *A. coluzzii* (from Burkina Faso (n = 1) and Angola (n = 2)), and 20 that are admixed from Guinea-Bissau. Of the 324 2L + ^{a}/2L + ^{a} mosquitoes, there are a 243 that are *A. gambiae* (from Guinea (n = 13), Cameroon (n = 156), Uganda (n = 15), Kenya (n = 7), and Gabon (n = 52)), 58 that are *A. coluzzii* from Angola, and 23 that are admixed from Guinea-Bissau. *A. gambiae* and A. *coluzzii* are recently diverged species that are the only *Anopheles* species polymorphic for the 2La inversion, and they are responsible for most of the malaria transmission in Sub-Saharan Africa. [31, 32]. There are a total of 19,453,287 positions in the variant data for chromosome arm 2L. Because only a single mosquito segregating for an alternate nucleotide is required for a given position to be called variable, to filter out nearly fixed positions we determined the maximum allele frequency using vcftools (v0.1.16) [33]–freq command to identify positions that are not fixed, which we defined as having a maximum allele frequency < 0.95. 3,153,365 positions (16.2%) met this criterion. To only include reliable genotype calls, positions in which the missing genotype (./.) frequency was < 0.05 were determined. 17,551,796 positions (90.2%) met this criterion. The overlap of (1) positions that are not fixed and (2) positions having reliable genotype calls yields a total of 2,544,892 positions, which are the positions we refer to as “All 2L SNPs” (Additional File 1, Figure S1). To further characterize the differential SNPs between 2La/2La and 2L + ^{a} /2L + ^{a} mosquitoes, we looked at positions where individuals homozygous for 2La/2La and 2L + ^{a} /2L + ^{a} have a different maximum allele (503,169 positions) and a different mode genotype (667,033 positions). Applying both the maximum allele and distinct mode genotype filter yields 477,028 positions across the 2L chromosome arm, with 97.5% of those positions occurring within the 2La inversion (proximal breakpoint: PEST position 20528073bp; distal breakpoint: PEST position 42165625bp) (Additional File 1 Figure S2A). Next, we determined which positions were variable in 2La homozygous and/or 2L + ^{a} homozygous individuals. Variable was defined as having a maximum allele frequency < 0.95, and fixed was defined as having a maximum allele frequency ≥ 0.95. In 2La homozygous

and/or 2L + ^{a} homozygous individuals 393,603 positions were variable, with 121,713 being variable in 2L + ^{a}/2L + ^{a} individuals but fixed in 2La/2La individuals, 132,595 being variable in 2La/2La individuals but fixed in 2L + ^{a}/2L + ^{a} individuals, and 139,295 being variable in both 2La/2La individuals and 2L + ^{a}/2L + ^{a} individuals (Additional File 1, Figure S2B). The remaining 83,425 positions were differentially fixed where the maximum allele frequency was ≥ 0.95 when calculating allele frequency for each homozygous population for 2La inversion karyotype. 18,072 positions were 2La inversion-specific SNPs, using a 95% cut-off as described in methods below. The remaining 65,353 positions were inversion-specific-like where the minimum mode genotype for 2La/2La, 2L + ^{a} /2L + ^{a}, or heterozygous individuals ranged from ~ 75.4% to ~ 94.8% (Additional File 1 Figure S2B). The distribution of each SNP category is displayed as a histogram across chromosome arm 2L (Additional File 1 Figure S2C). For statistical analysis, a chi-square goodness-of-fit test was performed on each SNP category, binning at 5MB and using all 2L SNPs (2,544,892 positions) as the expected distribution.

Identification of 2La Inversion-specific SNPs

Genotype frequencies from Ag1000G phase 1 AR3 variant data[30] for chromosome arm 2L were used to identify 2La inversion-specific SNPs. Inversion-specific SNPs were defined as positions where > 95% of 2La/2La mosquitoes were homozygous for one nucleotide, > 95% of 2L + ^{a}/2L + ^{a} mosquitoes were homozygous for a different nucleotide, and > 95% of 2La/2L + ^{a} mosquitoes were heterozygous for the two nucleotides. To achieve this, for each 2La inversion karyotype (i.e., 2La homozygotes, 2L + ^{a} homozygotes, and heterozygotes), the mode genotype frequency was determined for every position in the Ag1000G phase 1 AR3 variant file. This strategy enabled identification of genotypes that were fixed in each 2La inversion karyotype. Only positions where the mode genotype frequency was > 0.95 for all three 2La inversion karyotypes were retained. To keep only reliably called variants, positions with a missing genotype frequency ≥ 0.05 amongst all 765 mosquitoes combined were removed. Lastly, to identify positions that were differentially fixed between the two inversion karyotype homozygotes and highly heterogenous in heterozygotes, the mode genotypes at each position were assessed, only retaining positions in which the mode genotype of 2La/2La mosquitoes was x/x, the mode genotype of 2L + ^{a}/2L + ^{a} mosquitoes was y/y, and the mode genotype of 2La/2L + ^{a} mosquitoes was x/y (where x and y represent any nucleotide but x does not equal y). This strategy identified 18,072 2La inversion-specific SNPs.

Identification of STARR-seq enhancers containing 2La Inversion-specific SNPs

Bedtools map (v2.30.0) [34, 35] was used to count and identify overlap between 2La inversion-specific SNPs and the 788 STARR-seq identified enhancers on chromosome 2L [26]. To utilize the 2La inversion-specific SNPs with bedtools, it was converted to a bed file in which the end coordinate was the SNP position + 1. Bedtools map was used where -a was set to the STARR-seq enhancer bed file, -b was set to the 2La inversion-specific SNP bed file, -c was set to 2 (specifying the SNP position column to map onto intervals in STARR-seq enhancer bed file), and -o was set to count,collapse (to both count and specify the 2La inversion-specific SNP positions that overlap with each enhancer).

Distribution of 2La Inversion-specific SNPs over exonic, intronic, and intergenic regions

Bedfiles of exons, introns, and intergenic regions were extracted from PEST (VectorBase-68_AgambiaePEST.gff)[36, 37]. Intergenic regions were determined as the intervals in the genome not covered by at least one interval in the GFF and were identified using bedtools complement (v2.30.0) [34, 35]. Exonic coordinates were extracted using awk '{if($3=="exon"){print $1"\t"$4"\t"$5"\t"$9}}' on the gff file, then sort | uniq to remove duplicate exon entries. Intronic regions were determined as the intervals in the genome that are neither exonic nor intergenic and were identified using bedtools complement. The number of SNPs within exonic, intronic, and intergenic regions was determined. Because region type (exonic, intronic, or intergenic) is determined by either strand, a single SNP can be classified as being within more than one region type. To avoid double-counting, a SNP was first prioritized as exonic, then intronic, and lastly intergenic. A chi-square goodness-of-fit test was performed to determine if the distribution of inversion-specific SNPs over exonic, intronic, and intergenic regions matches that of all 2La SNPs (2,544,892 positions as defined above).

Using SnpEff to predict the effects of 2La inversion-specific SNPs

SnpEff (v5.2) [38] was used to predict the effects of the 18,072 2La inversion-specific SNPs. The Ag1000G phase 1 AR3 vcf file was trimmed to only include the 18,072 2La inversion-specific SNP positions using bcftools (v1.22) [39, 40]. 5,124 positions are tri- or tetra- allelic. Predictions were based on *Anopheles gambiae* PEST reference genome version AgamP4 (VectorBase release 68) [33, 37].

Cloning of Enhancer Alleles

Nine candidate inversion-specific enhancer alleles prioritized by one of the two complementary methods (one from the mapping to alternate reference genomes and the remaining 8 containing ≥ 6 inversion-specific SNPs) described above were PCR amplified from genomic DNA isolated from *A. coluzzii* hemocyte-like cell lines, Sua 4.0 (2La fixed) and Ag55 (2L + ^{a} fixed) and from wild captured individual *A. gambiae* and *A. coluzzii* mosquito samples collected along a Mali to Guinea transect [28]. PCR reactions contained 1X Phusion Ultramix (PCR Biosystems), 10ng of template DNA, and 0.5μM each primer. Cycle conditions were an initial denaturation at 98°C for 30 s, 30 cycles of denaturation at 98°C for 10 s, primer annealing at 55–60°C for 30 s and extension at 72°C for 45 s, followed by a final extension of 10 min at 72°C. Enhancer PCR fragments were cloned into a Firefly luciferase reporter vector pGL-Gateway-DSCP (AddGene, vector ID#71506), transformed and plasmid purified. All plasmids were sequence verified. All primers are in Additional File 2 Table S1. Enhancer sequences can be found in Additional File 2 Table S2.

Insertion/Deletion Typing

One indel of interest was identified in each of the two functionally dissected candidate inversion-specific enhancers and given that Ag1000G phase 1 AR3 variant data[30] does not contain information on indel variation, alternate molecular diagnostics were used to genotype the indel in a panel of wild captured mosquitoes. An RFLP assay using BanI (New England Biolabs) was used to type an 8bp indel in ISEN2L_21.9MB with product digestion occurring only in the absence of the insert and resulting products were resolved on a 2% agarose gel, a 560 bp band indicated presence of the 8bp indel while bands at 464 and 96bp indicated the absence of the indel. Sanger sequence was used to type a 32-40bp indel in ISEN2L_25.8MB. Indel genotypes for both the ISEN2L_21.9MB and the ISEN2L_25.8MB indels were determined for 30 wild *A. gambiae* and *A. coluzzii* mosquitoes collected along a Mali-Guinea transfect [28], 15 2La/2La mosquitoes and 15 2L + ^{a}/2L + ^{a} including larval captured samples and adult females captured by human baited or spray catch. Concordance between 2La karyotype as determined by a molecular assay [41] and indel genotype was determined for all 30 wild mosquito alleles, for the indel in ISEN2L_21.9MB concordance with 2La karyotype was 95% and for the ISEN2L_25.8MB indel concordance was 90%. These indels were considered alongside SNP variation in the dissection of enhancer function.

Measurement of Enhancer Activity by Luciferase Assay

A. coluzzii Sua4.0 [42] or Ag55 (Mos55) hemocyte-like cells [43] were seeded at 2.5 x 10^{4}cells/well in 65µL in an opaque 96 well plate. Cells were then agitated on a MixMate (Eppendorf) for 30 s at 300 rpm and incubated at 27°C for 24 h. Lipid-based transfections were performed with Lipofectamine 3000 (Invitrogen). Cells were dual transfected with two plasmids- GL-Gateway-DSCP (described above) containing a single candidate enhancer allele upstream of the firefly luciferase gene and Renilla vector pRL-ubi-63E (AddGene, #74280), a control plasmid. Plasmid transfection was performed at a ratio of 1:5 (renilla:firefly). Immediately following the addition of transfection reagents, the plate was again agitated on a MixMate (Eppendorf) for 30 s at 300 rpm and incubated at 27°C for 24 h. The Dual-Glo Luciferase Assay System (Promega) was then used for the luciferase assays, following supplier instructions. All measurements were recorded on the GloMax Discover (Promega) at 25°C. All plates contained cells transfected with control plasmids, a positive control enhancer peak near AGAP008980, and a negative control fragment located within intron one of AGAP007058, as previously described [44]. All samples were run with 6-fold technical replication within each biological replicate. Within each well, firefly luciferase measurements (Relative Light Units, RLUs) were normalized to the renilla measurements. These values were then expressed relative to the mean firefly/renilla of the negative control wells and this measure is referred to as relative luciferase activity.

Site Directed Mutagenesis

Wild type and mutant PEST and MOPTI alleles for ISEN2L_21.9MB and wild type PEST and MOPTI alleles for ISEN2L_25.8MB were synthesized by IDT (IDTDNA.com). ISEN2L_25.8MB candidate enhancer alleles were mutated at desired sites, introducing and eliminating the 32 bp indel on wildtype 2La/2La (MOPTI and the Sua4.0 cell line) and 2L + ^{a}/2L + ^{a} (PEST and Ag55 cell line) genetic backgrounds, respectively using the Q5 Site-Directed Mutagenesis Kit (New England Biolabs). Enhancers were PCR amplified using primers containing the mutation(s) of interest according to manufacturer instructions. Annealing temperatures ranged between 57–61°C. PCR products were then ligated, transformed and isolated plasmids were sequenced to confirm the introduction of desired mutations and the absence of additional mutations. Primer sequences are in Additional File 2 Table S1. Sequences for both wild type and mutated alleles can be found in File Additional File S2 Table S2.

Micro-C

Micro-C methods and analysis are described in detail elsewhere [45]. Briefly, three biological replicate samples were submitted to Dovetail Genomics for Micro-C library preparation and sequencing; samples were 3 biological replicates of 2La/2La hemocyte-like cell line Sua4.0 and 3 biological replicates of 2L + ^{a}/2L + ^{a} hemocyte cell line Ag55 collected under basal growth conditions. Cross-linked chromatin was digested with MNase, cells were lysed and chromatin fragments were bound to chromatin capture beads. Chromatin ends were proximity ligated, DNA was purified and converted to an Illumina compatible sequencing library. The resulting library was sequenced on an Illumina NovaSeq yielding 100 million unique mappable reads per replicate. Read processing, mapping and the generation of contact matrices used standard approaches and is described in detail elsewhere [45]. Micro-C raw data are available at PRJNA1439337.

ATAC-seq

Chromatin accessibility was assayed and compared for basally growing Sua4.0 and Ag55 cells in collaboration with ActiveMotif. Methods are described in detail elsewhere [45]. Briefly, ActiveMotif performed ATAC-seq and data analysis, using the nf-core/atacseq pipeline[46]. In addition we use published measures of chromatin accessibility including measurements made in hemocyte-like cells (our own [47] and that of others [48]) and in mosquito tissues, namely midguts and salivary glands for comparison [49, 50]. ATAC-seq raw data are available at PRJNA1439337.

Results

Use of two distinct and complementary approaches (Fig. 1) highlight candidate inversion-specific enhancers with differentially active 2La and 2L + ^{a} alleles. The first method used for the identification of candidate inversion-specific enhancers involved the published mapping of STARR-seq sequencing reads which had been originally mapped to the 2L + ^{a}/2L + ^{a} PEST reference genome [26] to the MOPTI 2La/2La genome assembly [51]. Using metrics for comparison described above, this analysis identified twenty-eight candidate enhancer sequences unique to the MOPTI genome with 13 of these detected across at least 2 STARR-seq biological replicates (Table 1) and the remaining 15 only in a single STARR-seq replicate. A closer examination of these 28 candidate inversion-specific enhancer sequences indicated that 8 were transposable or mobile elements identified through a search of the Dfam database [52], 11 were highly repetitive, non-complex, microsatellite-like sequence and may represent real differences in sequence, particularly at microsatellite loci, or sequencing or assembly errors, and the remaining 9 were unique sequence. Of these nine with unique sequence, two were detected across all three STARR-seq biological replicates and one across two biological replicates. While it is well known that transposable elements (TEs) have regulatory capacity [53, 54],, TE containing

candidate enhancers are not further explored here because it is not possible to know their position in a given individual mosquito's genome given the mobile nature of TEs.

Given that the 3 candidate inversion enhancers were ascertained via mapping to the two genomes of opposite karyotype status, we used genomic DNA isolated from mosquito hemocyte-like cell lines and from wild mosquitoes to determine whether the variants segregating between the PEST and MOPTI reference genomes were associated with the 2La inversion or instead differed between the two genomes for a reason unrelated to the 2La inversion. In sequencing these 3 candidate inversion-specific enhancers (ISEN2L_25.8MB, ISEN2L_25.9MB ISEN2L_33.9MB) across multiple gDNA sources, only one, ISEN2L_25.8MB, had genetic variation that was associated with the 2La inversion and was functionally characterized to determine whether the inversion-specific variation was necessary and sufficient to explain differences in enhancer activity alongside candidate inversion-specific enhancers identified using the second approach. All sequence is available in Additional File 2, Table S2.

2La inversion-specific SNPs create a U-shaped curve over the 2La inversion, are enriched in intergenic regions, and are largely non-coding variants.

The second method for the identification of inversion-specific SNPs analyzed Ag1000G phase 1 AR3 variant data for chromosome 2L for SNPs that segregate with the 2La inversion, using a > 95% genotype frequency threshold 18,072 inversion-specific SNPs were identified. Plotting the inversion-specific SNPs across the 2L chromosome reveals inversion-specific SNPs occur almost entirely within the 2La inversion with enrichment near inversion breakpoints with fewer SNPs in the middle of the inversion. This decrease in inversion-specific variation near the center of the inversion is consistent with double-recombination in the middle of the inversion allowing for homogenization between the 2La and 2L+ chromosomes (Fig. 2A). There are no inversion-specific SNPs that occur outside the proximal breakpoint, and only seven occur outside the distal breakpoint. These seven inversion-specific SNPs that occur outside the distal breakpoint occur within 832bp of the breakpoint. The distribution of SNPs over exonic, intronic, and intergenic regions was determined for the 18,072 inversion-specific SNPs and for comparison, all SNPs on the 2L chromosome arm (Fig. 2B). A chi-square goodness-of-fit test comparing the two distributions indicated enrichment of inversion-specific SNPs in intergenic regions and depletion in intronic regions ($p < 0.0001$). SnpEff was used to predict the effects of the 18,072 inversion-specific SNPs on genes and proteins based on *Anopheles gambiae* PEST reference genome version AgamP4 (VectorBase release 68). The inversion-specific SNPs, with 5,124 positions being multi-allelic, had 35,618 predicted effects. Over 95% of those effects (n = 34,032) were classified as modifier (typically defined as non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact), 3% (1,058 effects) were classified as low (defined as assumed to be mostly harmless or unlikely to change protein behavior), and 1.5% (528 effects) were classified as moderate (defined as a non-disruptive variant that might change protein effectiveness) (Fig. 2C). There were no effects classified as high (defined as the variant is assumed to have distributive impact in the protein, probably causing protein truncation, loss of function, or triggering nonsense mediated decay). Because the inversion-specific SNPs had predicted minimal effect on proteins and were largely non-coding variants in intergenic regions, we next explored their effect in enhancers.

Eight enhancers have a high burden of inversion-specific SNPs.

Using our previously published map of STARR-seq identified *Anopheles* enhancers, we identified enhancers on chromosome 2L that contained at least one inversion-specific SNP, then assessed the burden of inversion-specific SNPs. In total, 709 enhancers (89.6%) contained zero inversion-specific SNPs and were evenly distributed across chromosome 2L (Fig. 3A). Of the seventy-nine enhancers containing at least one inversion-specific SNP (10.4%), eight enhancers had a high burden of inversion-specific SNPs, which we defined as having six or more inversion-specific SNPs (Fig. 3B). The seventy-nine enhancers containing at least one inversion-specific SNP are fully inside the 2La inversion and create a similar U-shaped histogram as with inversion-specific SNPs. The 8 enhancers, 4 near the proximal breakpoint and 4 near the distal breakpoint, with at least 6 inversion-specific SNPs (ISEN2L_21.9MB, ISEN2L_22.3MB, ISEN2L_22.5MB, ISEN2L_22.6MB, ISEN2L_38.1MB, ISEN2L_41.4MB, ISEN2L_41.6MB, and ISEN2L_41.7MB) were further investigated to determine whether 2La and 2L + ^{a} alleles had significantly different enhancer activity and if so, whether the inversion-specific SNPs contributed to the difference in activity. All SNP variation, both inversion-specific SNPs and other segregating variation is shown for all eight enhancers and for the ISNE2L_25.8MB enhancer identified through the complementary analysis method in Additional File 1 Figure S3.

Enhancer activity and chromatin accessibility over inversion-specific enhancer regions

Activity of 2La and 2L + ^{a} enhancer alleles amplified and cloned from two hemocyte like cell lines of alternate inversion karyotypes (2La- Sua4.0 [42] and 2L + ^{a}- Ag55) and from two wild mosquitoes of alternate inversion karyotypes was assayed using a luciferase assay. Activity was assayed for the one candidate inversion-specific enhancer identified through mapping to the MOPTI 2La/2La genome and the eight candidate inversion-specific enhancers with a high burden (≥ 6) of inversion-specific SNPs. These nine enhancers were primarily located near the inversion breakpoints (Fig. 4A). All candidate inversion-specific enhancers displayed activity significantly above background, though relative luciferase activity varied widely from ~ 2 for ISEN2L_41.4MB to ≥ 30 for ISEN2L_22.3MB (Fig. 4B). Enhancers ISEN2L_21.9MB, ISEN2L_22.6MB, and ISEN2L_ 25.8MB displayed a pattern of significant differences in enhancer activity by inversion karyotype robust to both allele source (cell line or wild mosquito derived) and the assay environment (homokaryotic 2La or 2L + ^{a} cells). For ISE2L_22.5MB and ISEN2L_38.1MB the activity was significantly influenced by the cell environment, in both cases, activity assayed in cells homozygous for the 2La karyotype of the inversion was significantly higher. For the other 4 assayed enhancers, ISEN2L_22.3MB, ISEN2L_41.4MB, ISEN2L_41.6MB and ISEN 41.7MB there are not obvious patterns driven by allele source or cell environment and for the latter 3 of these, enhancer activity is quite low, though significantly above background, defined as a relative luciferase activity of 1. Considering the most biologically relevant comparisons, alleles assayed in the cell line matching their karyotype (comparing the left most and right most bars on the graph; 2La alleles in 2La cells and 2L + ^{a} alleles in 2L + ^{a} cells), there are significant differences for all candidate inversion-specific enhancers except ISEN2L_41.4MB. For both ISEN2L_21.9MB and ISEN2L_25.8MB additional functional dissection was done to determine if the inversion-specific variation was necessary and sufficient to account for the observed differences in activity. ISEN2L_21.9MB was further investigated as it carried the highest burden of inversion-specific variation (n = 13) and ISEN2L_25.8MB was further investigated as it segregated for a large inversion-specific indel even though it lacked inversion-specific variation as defined above.

Functional characterization of candidate inversion-specific enhancers.

Functional characterization of the observed differences in activity between 2La and 2L + ^{a} enhancer alleles was done for ISEN2L_21.9MB and ISEN2L_25.8MB. For ISEN2L_21.9MB, the wild type 2La/2La MOPTI allele had significantly higher activity compared to the wild type 2L + ^{a}/2L + ^{a} PEST allele. Mutating the 13 inversion-specific nucleotide positions on the 2La/2La MOPTI genetic background to the PEST (2L + ^{a}) alleles led to a significant decrease in activity; though still statistically distinguishable from the PEST wild type allele (Fig. 5A). Mutating the thirteen inversion-specific positions on the PEST genetic background to the MOPTI alleles did not significantly alter enhancer activity. This led us to conclude that the 13 inversion-specific sites are necessary, but not sufficient to explain the difference in activity between 2La and 2L + ^{a} enhancer alleles for ISEN2L_21.9MB. As the Ag1000 G data does not include insertion/deletion variation data, we examined the sequence of ISEN2L_21.9MB more closely and identified a short, 8 bp inversion-specific indel present on the 2La/2La karyotype. Utilizing a small panel of wild mosquitoes, consisting of 15 2La/2La homokaryotes and 15 2L + ^{a}/2L + ^{a}, we determined that the inversion genotypes and the insert were 95% concordant. To test the effect of this inversion-specific indel with the inversion-specific SNPs and this inversion-specific indel, we introduced the MOPTI indel sequence into the PEST genetic background and the PEST allele into the MOPTI genetic background individually and alongside the 13 previously mutated inversion-specific SNPs. Insertion or deletion of the 8bp indel did not alter enhancer activity significantly (Fig. 5A). Using the JASPAR transcription factor database and the FIMO function within MEME Suite seven of the thirteen inversion-specific SNPs along with the insertion/deletion lie over predicted transcription factor binding sites (Additional File 1 Supplementary Fig. 4A) and often the variable site is highly conserved within the predicted transcription factor binding site. All 33 SNPs between MOPTI and PEST alleles have also been annotated according to the categories defined above (see Fig. 2 and Additional File 1, Figure S4B)

For ISEN2L_25.8MB, we tested the functional significance of a 2La inversion associated indel (90% concordance between 2La inversion karyotype and the indel) which is present in the PEST 2L + ^{a}/2L + ^{a} genome and absent in the 2La/2La MOPTI genome. Using site directed mutagenesis we introduced the insertion into the MOPTI genetic background and deleted it from the PEST genetic background. The indel was necessary and sufficient to explain the differences in activity between 2La and 2L + ^{a} derived alleles (Fig. 5B). Using the JASPAR transcription factor database and the FIMO function within MEME Suite, we searched this 32-40bp insertion for predicted transcription factor binding sites. Multiple binding sites for the cg (combgap) transcription factor were identified along with single sites for Ets98B and ci were found in both the PEST and Ag55 alleles, and one site for BtbVII also found only on the Ag55 cell line allele (Additional file1 Figure S5)

Chromatin accessibility at candidate inversion-specific enhancers

ATAC-seq data [45] was used to assess chromatin accessibility in the candidate inversion-specific enhancer regions. For all 9 candidate inversion-specific enhancers there were 1 or 2 overlapping ATAC-seq peaks (+/- 1500bp), indicating that chromatin was accessible in both 2La homokaryotic cells and 2L + ^{a} homokaryotic cells (Additional File 2, Table S3). However, only 2 candidate inversion-specific enhancers overlapped ATAC-seq peaks displaying significant differences in accessibility between cells of opposite 2La inversion karyotype, with 2L + ^{a} cells displaying greater chromatin accessibility in both of these cases (ISEN2L_22.5MB overlapping with AgamP4_2L:22476460–22476766 and ISEN2L_41.6MB overlapping with AgamP4_2L:41623366_41623910) Seven of the nine inversion-specific enhancers are also detected in regions of open chromatin from other published work [47–50], while the remaining two inversion-specific enhancers had not been identified as regions of open chromatin previously, ISEN2L_22.6MB, and ISEN2L_41.6MB.

Regulatory network for inversion-specific enhancers.

Using previously published Micro-C data [45] and the distance decay normalization method for identifying enriched chromatin interactions, candidate target genes for the two inversion-specific enhancers were identified. As gene expression is often regulated by elements at distance from the promoter, the proximity ligation inherent in Micro-C allows the assessment of 3D chromatin interactions. Applying the distance decay normalization method to ISEN2L_21.9MB identifies AGAP028981, a ncRNA gene, as a top candidate target gene in 2La homozygotes hemocyte like cells and AGAP005822, encoding a protein of unknown function, as a top candidate target gene in 2L + ^{a} homozygous hemocyte like cells. Both AGAP028981 and AGAP005822 are expressed below the RNA-seq level of detection across cell lines of both karyotypes (Fig. 6A). For ISEN2L_25,8MB, the target gene in both 2La and 2L + ^{a} cells is AGAP006058, a heparan sulfate 2-0-sulfotransfrease which trends towards higher expression in the 2La^{+} cell line (1.5 fold difference in expression p = 0.07) (Fig. 6B). ISEN2L_25.8MB also has higher enhancer activity in 2L + ^{a} cells (Figs. 4 and 5), due to the presence of a 32-40bp indel that contains multiple transcription factor binding sites (Additional file1 Figure S5)

Non-coding variation in inversion-specific enhancers is conserved in the Gambiae species complex.

Sequence alignments from three 2La genomes (*A. coluzzii* MOPTI, *A. arabiensis* DONGOA 2021, and *A. merus* MAF2021) and 3 2L + ^{a} genomes (*A. gambiae* PEST, *A. melas* CM1001059, and *A. quadriannulatus* SANGWE) were examined for sequence conservation in both ISEN2L_21.9MB and ISEN2L_25.8MB. For the ISEN2L_21.9MB alignment, two species outside of the species complex, *A. epiroticus* Epiroticus2 and *A. christyi* ACHKN1017, were also included in the alignment. For ISEN2L_21.9MB, all three 2La genomes contained the 8bp insertion and in all 2L + ^{a} genomes the 8bp indel is deleted. For the thirteen inversion-specific SNPs, the *A. coluzzii* MOPTI and *A. arabiensis* DONGOLA genomes share the same nucleotide variant and for two of these thirteen the variant is also shared with *A. merus* a more distantly related species that is fixed for the 2La inversion karyotype (Additional File 1, Figure S6). For ISEN2L_25.8MB, the 32-40bp indel is present in all three 2L + ^{a} genomes and absent in all three 2La genomes (Additional File 1, Figure S7).

Discussion

The 2La paracentric inversion is a major genetic marker with significant influence on Plasmodium susceptibility in the Anopheles malaria vector [24, 25], and to date remains the only genetic marker for susceptibility confirmed in the natural population. Limited recombination between 2La heterokaryotic mosquito

individuals near the inversion breakpoints leads to genetic variation private to each of the two karyotypes, that is, inversion-specific SNPs as we detect here. The capture of divergent sequences by the 2La karyotypes is also consistent with previous observations [55]. Indeed, sequence and haplotype divergence of inversion karyotypes is a general feature of inversions in any system, and likely explains their adaptive advantage.

In the current system, genetically divergent and/or unique sequences exclusive to a given karyotype probably underlie the phenotypic variation between the 2La homozygous and the 2L + ^{a} homozygous mosquitoes including differences in intrinsic susceptibility to infection by human malaria parasites [24, 25] and differences in extrinsic factors affecting vectorial capacity including feeding and resting behaviors [17], insecticide resistance [18, 19], and thermo- and aridity tolerance [20–22]. In principle, the phenotypic differences associated with the 2La karyotypes can be assumed to result from differences in regulated gene expression profiles in carriers of the different karyotypes. Because enhancers typically regulate multiple genes, we reasoned that inversion-linked differences in enhancer sequences, if they led to a functional difference in enhancer activity, could generate larger phenotypic effects than divergence of single coding genes. Therefore, we used our recently reported genome-wide catalog of Anopheles enhancers [26] to identify the enhancer intervals within the 2La boundaries in both inversion karyotypes, filtered for sequence contrast between the karyotypes, and validated candidates by sequencing them in wild mosquito DNA of both karyotypes to confirm their sequence and 2La divergence in the population. Candidates were tested to determine inversion-dependent difference in enhancer activity, and prioritized candidates were also mutagenized to identify the variant positions that explain the enhancer activity difference of the karyotype alleles.

We implemented two complementary and independent approaches for identification of karyotype-divergent enhancers, which identified nine high quality candidates. The approach of mapping to alternate reference genomes fixed for alternate 2La karyotypes highlighted enhancer ISEN2L_25.8MB, which varies between the PEST and MOPTI reference genomes by a 32bp insertion in the PEST genome and displays significantly different activity associated with the 2La inversion. The alternate approach, using wild individual whole genome sequences to identify enhancers carrying a large burden of inversion-specific SNPs (≥ 6), highlighted 8 enhancers, with three of these (ISEN2L_21.9MB, ISEN2L_22.6MB and ISEN2L_41.6MB) displaying significantly different activity associated with the 2La inversion.

These distinct methods for candidate discovery each have their own inherent strengths and limitations. In the reference assembly approach, where we remapped the STARR-seq sequence reads to chromosome-assembled genomes that differ in inversion karyotype, there is the potential advantage to identify enhancers not present or significantly diverged from a single reference genome used for the initial STARR-seq read mapping for identification of enhancers [26]. However, this strength is balanced by an ascertainment bias caused by the small number of high-quality genomes of known inversion status, essentially only the PEST and MOPTI assemblies that we used here, one for each karyotype. In the second approach, aligning the enhancer catalog sequence intervals to whole genome sequence data from many wild individuals to identify fixed variants within the intervals, the main advantage is that the large sample size strengthens the power to determine sites that segregate in an inversion-specific manner, although balanced with the potential ascertainment bias that all reads from the wild individual mosquitoes were mapped to a single reference genome, the PEST 2L + ^{a}/2L + ^{a} reference genome, and thus enhancer variation or even entire enhancers not present in the 2L + ^{a} karyotype would be invisible. Thus, by combining these approaches, each with their potential ascertainment biases, we generated the most thorough analysis possible at the current level of genomic technology. The next technology step that would facilitate such an analysis would probably be the availability of massive numbers of wild individual sequences, generated by long-read sequencing, and de novo assembled without the bias of a reference genome, a level that clearly does not yet exist. Here, by integrating the results from the best currently available approaches, we identified divergent enhancers as well as queried their activity, chromatin accessibility and used proximity ligation to identify candidate target genes. This multi-omics approach is therefore a model for linking genetic variation with regulatory function using current technologies.

Transposable elements and the regulatory genome

A number of the candidate inversion-specific enhancers identified in mapping to the MOPTI reference genome share significant sequence similarly to annotated transposable elements (TEs) (Table 1). Work in vertebrate and invertebrate systems has demonstrated the central role that TEs have for the regulatory genome, with about a quarter of human candidate *cis* regulatory elements being TE-derived [56]. Across insect genomes, TEs occupy from 0.13% to 81.5% of the genome [57]; in Anopheline genomes from 0.13–19.55%, while some Culicidae genomes exceed 40%TE composition. Moreover, TEs have been shown to contain tissue-specific transcription factor binding motifs [58], contribute to the regulation of lncRNAs in Drosophila [59], and regulate the wing pigmentation in the peppered moth [60]. In *Anopheles funestus*, a transposon carrying structural variation, in the form of a 4.3kb insertion, is significantly associated with P450-based insecticide resistance and susceptibility to *P. falciparum* infection [61]. Understanding the functional role of TEs is challenging given their inherit mobility, coupled with their multicopy nature and repetitive sequence, a trifecta making cataloguing of TEs a challenge. Long read sequencing technology has aided the ability to determine TE positions, but doing this across multiple individuals remains a significant challenge. Consequently, the putative location of the TE-associated enhancers in the MOPTI reference genome may or may not indicate their true genomic location, which could vary across individuals in any case, and which would require many long-read individual genome sequences to resolve. Therefore, these potentially TE-associated enhancers are noted here to aid future work by us or others, but it is not currently practical to carry out further genetic or functional characterization of them.

Microsatellites and regulatory function

In addition to TEs, candidate inversion-specific enhancers identified in mapping to the MOPTI reference genome were often of simple, non-unique sequence including microsatellite-like sequences. We previously reported that *A. coluzzii* STARR-seq enhancers are enriched for certain dinucleotide repeats (CA, GA, GC) and depleted of others (TA) [26], and this pattern has also been shown across vertebrates and invertebrates [62, 63]. Microsatellites can also serve as binding sites for transcription factors [64]. However, the functional significance in the current system remains to be determined.

Inversion-specific variation

Previous work by others has identified SNPs that are correlated with 2La inversion karyotype. Love et al. [65, 66] used data from Ag1000G to identify 209 tag SNPs within the 2La inversion that had high concordance (> 80%) with inversion karyotype. All of the 209 are included in our 18,072 inversion-specific SNPs. Alvarez et al. [67] showed that biallelic SNPs can be used to reliably detect chromosomal inversions. They used a sliding window analysis of PCA components to identify regions containing SNPs with high linkage-disequilibrium, indicative of inverted chromosomal regions due to the lack of recombination over these areas. Their study also used Ag1000G data to validate their pipeline, which correctly identified the 2La and 2Rb inversion in *Anopheles gambiae.* While the list of biallelic SNPs used to detect the 2La inversion is not publicly available, the SNPs occur between positions 20,466,374 and 42,223,401, and 100% of our inversion-specific SNPs fall within this region.

Conclusions

A major Anopheles genetic marker of Plasmodium susceptibility, the 2La inversion, is associated with a wide range of phenotypes influencing vectorial capacity, but this association with phenotype has no genetic explanation to date. The current study represents a step towards understanding 2La control over malaria transmission phenotypes. Here, we identify functionally divergent transcriptional enhancers across 2La karyotypes. The multiomics approach to identify candidate inversion-specific SNPs in *Anopheles* has the potential to inform the development of novel control strategies directed at segments of the mosquito population that may disproportionately contribute to malaria transmission, specifically the 2L + ^{a} carrying group that serve as malaria superspreaders.

Abbreviations

ISEN
Inversion-specific enhancer
SNP
single nucleotide polymorphism
Indel
insertion/deletion
TE
transposable element

Declarations

Declarations

Ethics Approval and consent to participate

.

Consent for publication

All authors have seen and approved the manuscript, and it has not been accepted for publication elsewhere.

Additional File 1 – pdf-

Additional File 1 Figure S1. Ag1000G phase 1 AR3 variant data

was filtered down to 2,544,892 SNPs.

Additional File 1 Figure S2. Categorization of differential SNPs between 2La/2La and 2L + ^{a} /2L + ^{a} mosquitoes

Additional File 1 Figure S3.

Genetic variation at each of the 9 candidate inversion-specific enhancers assayed in Fig. 1.

Additional File 1. Figure S4. Transcription factor binding sites and genetic variation in ISEN2L_21.9MB.

Additional File 1. Figure S5.

Predicted transcription factor binding sites within the indel in ISEN2L_25.8MB.

Additional File 1 Figure S6.

Alignment of ISEN2L_21.9MB sequences from representative genomes for six members of the A. gambiae species complex, A. coluzzii, A. arabiensis, A. merus, and A. gambiae, A. melas and A. quadriannulatus.

Additional File 1 Figure S7.

Alignment of ISEN2L_25.8MB sequences from representative genomes for six members of the A. gambiae species complex, A. coluzzii, A. arabiensis, A. merus, and A. gambiae, A. melas and A. quadriannulatus.

Additional file 2 – xls

Table S1- Primer sequences.

For primers used in SDM, mutant positions are indicated by lowercase letters.

Funding

This study received financial support to KST from National Institutes of Health, # AI188600 and to MMR from National Institutes of Health, NIAID #AI145999 and #AI191531.

Author Contribution

Authors' contributionsKST, CM, KDV, MMR designed the research. KST, AP, and CEA and performed the research. KST, AP and CEA analyzed the data, and KST and MMR wrote the manuscript with input from all authors.

Acknowledgement

We thank Hans-Michael Müller (Sua4.0) and Michael Adang (Ag55) for sharing the cell lines.

Data Availability

Availability of data and materialsThe Micro-C and ATAC-seq data used to support findings of this project have been previously published [45] and can be found at PRJNA1439337. The STARR-seq data have been previously published [26] and are available at PRJEB34434.

References

1. Gunther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genet. 2019;15(7):e1008302.
2. Akopyan M, Genchev M, Armstrong EE, Mooney JA. Reference genome choice compromises population genetic analyses. Cell. 2025;188(24):6939–e695211.
3. Maurstad MF, Hoff SNK, Cerca J, Ravinet M, Bradbury I, Jakobsen KS, Praebel K, Jentoft S. Reference genome bias light species-specific chromosomal reorganization translocations Genome Biol. 2025;26(1):355.
4. Berdan EL, Barton NH, Butlin R, Charlesworth B, Faria R, Fragata I, Gilbert KJ, Jay P, Kapun M, Lotterhos KE, Merot C, Durmaz Mitchell E, Pascual M, Peichel CL, Rafajlovic M, Westram AM, Schaeffer SW, Johannesson K, Flatt T. How chromosomal inversions reorient the evolutionary process. J Evol Biol. 2023;36(12):1761–82.
5. Gompert Z, Feder JL, Parchman TL, Planidin NP, Whiting FJH, Nosil P. Adaptation repeatedly uses complex structural genomic variation. Science. 2025;388(6744):eadp3745.
6. Ayala D, Ullastres A, Gonzalez J. Adaptation through chromosomal inversions in Anopheles. Front Genet. 2014;5:129.
7. Weeks AR, McKechnie SW, Hoffmann AA. Dissecting adaptive clinal variation: markers, inversions and size/stress associations in Drosophila melanogaster from a central field population. Ecol Lett. 2002;5(6):756–63.
8. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E, Searle S, Schmutz J, Grimwood J, Dickson MC, Myers RM, Miller CT, Summers BR, Knecht AK, Brady SD, Zhang H, Pollen AA, Howes T, Amemiya C, P., Baldwin T, Bloom DB, Jaffe R, Nicol J, Wilkinson ES, Lander F, Di Palma K, Lindblad-Toh, Kingsley DM. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484(7392):55–61. Broad Institute Genome Sequencing, T. Whole Genome Assembly, J.
9. Lakich D, Kazazian HH Jr., Antonarakis SE, Gitschier J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. Nat Genet. 1993;5(3):236–41.
10. Nosil P, Soria-Carrasco V, Villoutreix R, De-la-Mora M, de Carvalho CF, Parchman T, Feder JL, Gompert Z. Complex evolutionary processes maintain an ancient chromosomal inversion. Proc Natl Acad Sci U S A. 2023;120(25):e2300673120.

11. Kapun M. The Influence of Chromosomal Inversions on Genetic Variation and Clinal Patterns in Genomic Data of Drosophila melanogaster. Methods Mol Biol. 2025;2935:141–77.
12. Ghavi-Helm Y, Jankowski A, Meiers S, Viales RR, Korbel JO, Furlong EEM. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nat Genet. 2019;51(8):1272–82.
13. Okabe A, Kaneda A. Transcriptional dysregulation by aberrant enhancer activation and rewiring in cancer. Cancer Sci. 2021;112(6):2081–8.
14. Wang M, He B, Hao Y, Srinivasan D, Shrinet J, Fraser P. Cellular reprogramming is driven by widespread rewiring of promoter-enhancer interactions. BMC Biol. 2023;21(1):264.
15. Stevison LS, Hoehn KB, Noor MA. Effects of inversions on within- and between-species recombination and divergence. Genome Biol Evol. 2011;3:830–41.
16. Singh BN. Hundred years of research on inversion polymorphism in Drosophila. Curr Sci. 2019;117(5):761–75.
17. Coluzzi M, Sabatini A, Petrarca V, Di Deco MA. Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. Trans R Soc Trop Med Hyg. 1979;73(5):483–97.
18. Brooke BD, Hunt RH, Chandre F, Carnevale P, Coetzee M. Stable chromosomal inversion polymorphisms and insecticide resistance in the malaria vector mosquito Anopheles gambiae (Diptera: Culicidae). J Med Entomol. 2002;39(4):568–73.
19. Brooke BD, Hunt RH, Coetzee M. Resistance to dieldrin + fipronil assorts with chromosome inversion 2La in the malaria vector Anopheles gambiae. Med Vet Entomol. 2000;14(2):190–4.
20. Rocca KA, Gray EM, Costantini C, Besansky NJ. *2La chromosomal inversion enhances thermal tolerance of Anopheles gambiae larvae.* Malar J, 2009. 8: p. 147.
21. Gray EM, Rocca KA, Costantini C, Besansky NJ. Inversion 2La is associated with enhanced desiccation resistance in Anopheles gambiae. Malar J. 2009;8:215.
22. Fouet C, Gray E, Besansky NJ, Costantini C. Adaptation to aridity in the malaria mosquito Anopheles gambiae: chromosomal inversion polymorphism and body size influence resistance to desiccation. PLoS ONE. 2012;7(4):e34841.
23. Ibrahim SS, Mukhtar MM, Muhammad A, Wondji CS. *2La Paracentric Chromosomal Inversion and Overexpressed Metabolic Genes Enhance Thermotolerance and Pyrethroid Resistance in the Major Malaria Vector Anopheles gambiae.* Biology (Basel), 2021. 10(6).
24. Petrarca V, Beier JC. Intraspecific chromosomal polymorphism in the Anopheles gambiae complex as a factor affecting malaria transmission in the Kisumu area of Kenya. Am J Trop Med Hyg. 1992;46(2):229–37.
25. Riehle MM, Bukhari T, Gneme A, Guelbeogo WM, Coulibaly B, Fofana A, Pain A, Bischoff E, Renaud F, Beavogui AH, Traore SF, Sagnon N, Vernick KD. The Anopheles gambiae 2La chromosome inversion is associated with susceptibility to Plasmodium falciparum in Africa. Elife, 2017. 6.
26. Holm I, Nardini L, Pain A, Bischoff E, Anderson CE, Zongo S, Guelbeogo WM, Sagnon N, Gohl DM, Nowling RJ, Vernick KD, Riehle MM. Comprehensive Genomic Discovery of Non-Coding Transcriptional Enhancers in the African Malaria Vector Anopheles coluzzii. Front Genet. 2021;12:785934.
27. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
28. Coulibaly B, Kone R, Barry MS, Emerson B, Coulibaly MB, Niare O, Beavogui AH, Traore SF, Vernick KD, Riehle MM. Malaria vector populations across ecological zones in Guinea Conakry and Mali, West Africa. Malar J. 2016;15:191.
29. Kumar S, Stecher G, Suleski M, Sanderford M, Sharma S, Tamura K. MEGA12: Molecular Evolutionary Genetic Analysis Version 12 for Adaptive and Green Computing. Mol Biol Evol, 2024. 41(12).
30. Anopheles gambiae Genomes C, g. Data analysis, g. Partner working, c.-A., Sample F, Burkina, Cameroon, Gabon, Guinea B, Guinea. Kenya, Uganda, Crosses, Sequencing, p. data, d. Web application and c. Project, *Genetic diversity of the African malaria vector Anopheles gambiae.* Nature, 2017. 552(7683): pp. 96–100.
31. Sharakhov IV, White BJ, Sharakhova MV, Kayondo J, Lobo NF, Santolamazza F, Della Torre A, Simard F, Collins FH, Besansky NJ. Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the Anopheles gambiae complex. Proc Natl Acad Sci U S A. 2006;103(16):6258–62.
32. Sangbakembi-Ngounou C, Ngoagouni C, Akone-Ella O, Kengne P, Costantini C, Nakouné E, Ayala D. Temporal and biting dynamics of the chromosomal inversion 2La in the malaria vectors and in Bangui, Central African Republic. Frontiers in Ecology and Evolution; 2022. p. 10.
33. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G. Durbin and G. Genomes Project Analysis, *The variant call format and VCFtools*. Bioinformatics. 2011;27(15):2156–8.
34. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
35. Quinlan AR. *BEDTools: The Swiss-Army Tool for Genome Feature Analysis.* Curr Protoc Bioinformatics, 2014. 47: p. 11 12 1–34.
36. Harb OS, McDowell MA, Roos DS. VEuPathDB Resources: A Platform for Free Online Data Exploration, Integration, and Analysis. Methods Mol Biol. 2024;2802:573–86.
37. Alvarez-Jarreta J, Amos B, Aurrecoechea C, Bah S, Barba M, Barreto A, Basenko EY, Belnap R, Blevins A, Bohme U, Brestelli J, Brown S, Callan D, Campbell LI, Christophides GK, Crouch K, Davison HR, DeBarry JD, Demko R, Doherty R, Duan Y, Dundore W, Dyer S, Falke D, Fischer S, Gajria B, Galdi D, Giraldo-Calderon GI, Harb OS, Harper E, Helb D, Howington C, Hu S, Humphrey J, Iodice J, Jones A, Judkins J, Kelly SA, Kissinger JC, Kittur N, Kwon DK, Lamoureux K, Li W, Lodha D, MacCallum RM, Maslen G, McDowell MA, Myers J, Nural MV, Roos DS, Rund SSC, Shanmugasundram A, Sitnik V, Spruill D, Starns D, Tomko SS, Wang H, Warrenfeltz S, Wieck R, Wilkinson PA, Zheng J. VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center in 2023. Nucleic Acids Res. 2024;52(D1):D808–16.

38. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.
39. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
40. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience, 2021. 10(2).
41. White BJ, Santolamazza F, Kamau L, Pombi M, Grushko O, Mouline K, Brengues C, Guelbeogo W, Coulibaly M, Kayondo JK, Sharakhov I, Simard F, Petrarca V. Della Torre and N.J. Besansky, *Molecular karyotyping of the 2La inversion in Anopheles gambiae*. Am J Trop Med Hyg. 2007;76(2):334–9.
42. Muller HM, Dimopoulos G, Blass C, Kafatos FC. A hemocyte-like cell line established from the malaria vector Anopheles gambiae expresses six prophenoloxidase genes. J Biol Chem. 1999;274(17):11727–35.
43. Pudney M, Varma M, Leake C. *Establishment of cell lines from larvae of culicine (Aedes species) and anopheline mosquitoes.* TCA manual / Tissue Culture Association, 1979. 5(1): pp. 997–1002.
44. Nardini L, Holm I, Pain A, Bischoff E, Gohl DM, Zongo S, Guelbeogo WM, Sagnon N, Vernick KD, Riehle MM. Influence of genetic polymorphism on transcriptional enhancer activity in the malaria vector Anopheles coluzzii. Sci Rep. 2019;9(1):15275.
45. Taquet KS, Anderson CE, Vernick KD, Riehle MM. The Anopheles gambiae 2La chromosomal inversion influences chromatin organization and 3D landscape of genes related to malaria transmission. Research Square; 2026.
46. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38(3):276–8.
47. Zmarlak-Feher NM, Taquet KS, Zakhia R, Pain A, Brito-Fravallo E, Anderson CE, Vernick KD, Mitri C, Riehle MM. Functional characterization of transcriptional enhancers in an Anopheles genetic locus controlling natural resistance to the malaria parasite, Plasmodium falciparum. Epigenetics Chromatin. 2025;18(1):37.
48. Perez-Zamorano B, Rosas-Madrigal S, Lozano OAM, Castillo Mendez M, Valverde-Garduno V. Identification of cis-regulatory sequences reveals potential participation of Iola and Deaf1 transcription factors in Anopheles gambiae innate immune response. PLoS ONE. 2017;12(10):e0186435.
49. Ruiz JL, Yerbanga RS, Lefevre T, Ouedraogo JB, Corces VG. Gomez-Diaz, *Chromatin changes in Anopheles gambiae induced by Plasmodium falciparum infection*. Epigenetics Chromatin. 2019;12(1):5.
50. Ruiz JL, Ranford-Cartwright LC, Gomez-Diaz E. The regulatory genome of the malaria vector Anopheles gambiae: integrating chromatin accessibility and gene expression. NAR Genom Bioinform. 2021;3(1):lqaa113.
51. Zamyatin A, Avdeyev P, Liang J, Sharma A, Chen C, Lukyanchikova V, Alexeev N, Tu Z, Alekseyev MA, Sharakhov IV. Chromosome-level genome assemblies of the malaria vectors Anopheles coluzzii and Anopheles arabiensis. Gigascience, 2021. 10(3).
52. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12(1):2.
53. Sundaram V, Wysocka J. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. Philos Trans R Soc Lond B Biol Sci. 2020;375(1795):20190347.
54. Gonzalez J, Petrov DA. The adaptive role of transposable elements in the Drosophila genome. Gene. 2009;448(2):124–33.
55. Cheng C, White BJ, Kamdem C, Mockaitis K, Costantini C, Hahn MW, Besansky NJ. Ecological genomics of Anopheles gambiae along a latitudinal cline: a population-resequencing approach. Genetics. 2012;190(4):1417–32.
56. Du AY, Chobirko JD, Zhuo X, Feschotte C, Wang T. Regulatory transposable elements in the encyclopedia of DNA elements. Nat Commun. 2024;15(1):7594.
57. Sproul JS, Hotaling S, Heckenhauer J, Powell A, Marshall D, Larracuente AM, Kelley JL, Pauls SU, Frandsen PB. Analyses of 600 + insect genomes reveal repetitive element dynamics and highlight biodiversity-scale repeat annotation challenges. Genome Res. 2023;33(10):1708–17.
58. Zhang Y, She J, Hu X, Jin Y, Zhao J, Hou S, Tao C, Du M, Yang E. Transposable elements drive species-specific and tissue-specific transcriptomes in human development. Genome Biol. 2025;26(1):379.
59. Gan Y, Wang L, Liu G, Guo X, Zhou Y, Chang K, Zhang Z, Yan F, Liu Q, Chen B. Transposable Elements Contribute to the Regulation of Long Noncoding RNAs in Drosophila melanogaster. Insects, 2024. 15(12).
60. Van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ. The industrial melanism mutation in British peppered moths is a transposable element. Nature. 2016;534(7605):102–5.
61. Mugenzi LMJ, Tekoh TA, Ntadoun ST, Chi AD, Gadji M, Menze BD, Tchouakui M, Irving H, Wondji MJ, Weedall GD, Hearn J, Wondji CS. Association of a rapidly selected 4.3kb transposon-containing structural variation with a P450-based resistance to pyrethroids in the African malaria vector Anopheles funestus. PLoS Genet. 2024;20(7):e1011344.
62. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, Turner JM, Bertelsen MF, Murchison EP, Flicek P, Odom DT. Enhancer evolution across 20 mammalian species. Cell. 2015;160(3):554–66.
63. Andolfatto P. Adaptive evolution of non-coding DNA in Drosophila. Nature. 2005;437(7062):1149–52.
64. Bagshaw ATM. Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes. Genome Biol Evol. 2017;9(9):2428–43.
65. Love RR, Redmond SN, Pombi M, Caputo B, Petrarca V, Della Torre A. C. Anopheles gambiae Genomes and N.J. Besansky, *In Silico Karyotyping of Chromosomally Polymorphic Malaria Mosquitoes in the Anopheles gambiae Complex.* G3 (Bethesda), 2019. 9(10): pp. 3249–3262.

66. Love RR, Pombi M, Guelbeogo MW, Campbell NR, Stephens MT, Dabire RK, Costantini C, Della Torre A, Besansky NJ. *Inversion Genotyping in the Anopheles gambiae Complex Using High-Throughput Array and Sequencing Platforms.* G3 (Bethesda), 2020. 10(9): pp. 3299–3307.

67. Alvarez MVN, Bozoni FT, Alonso DP, Ribolla PEM. *A Multi-Approach for In Silico Detection of Chromosome Inversions in Mosquito Vectors.* Microorganisms, 2025. 13(10).

Figures

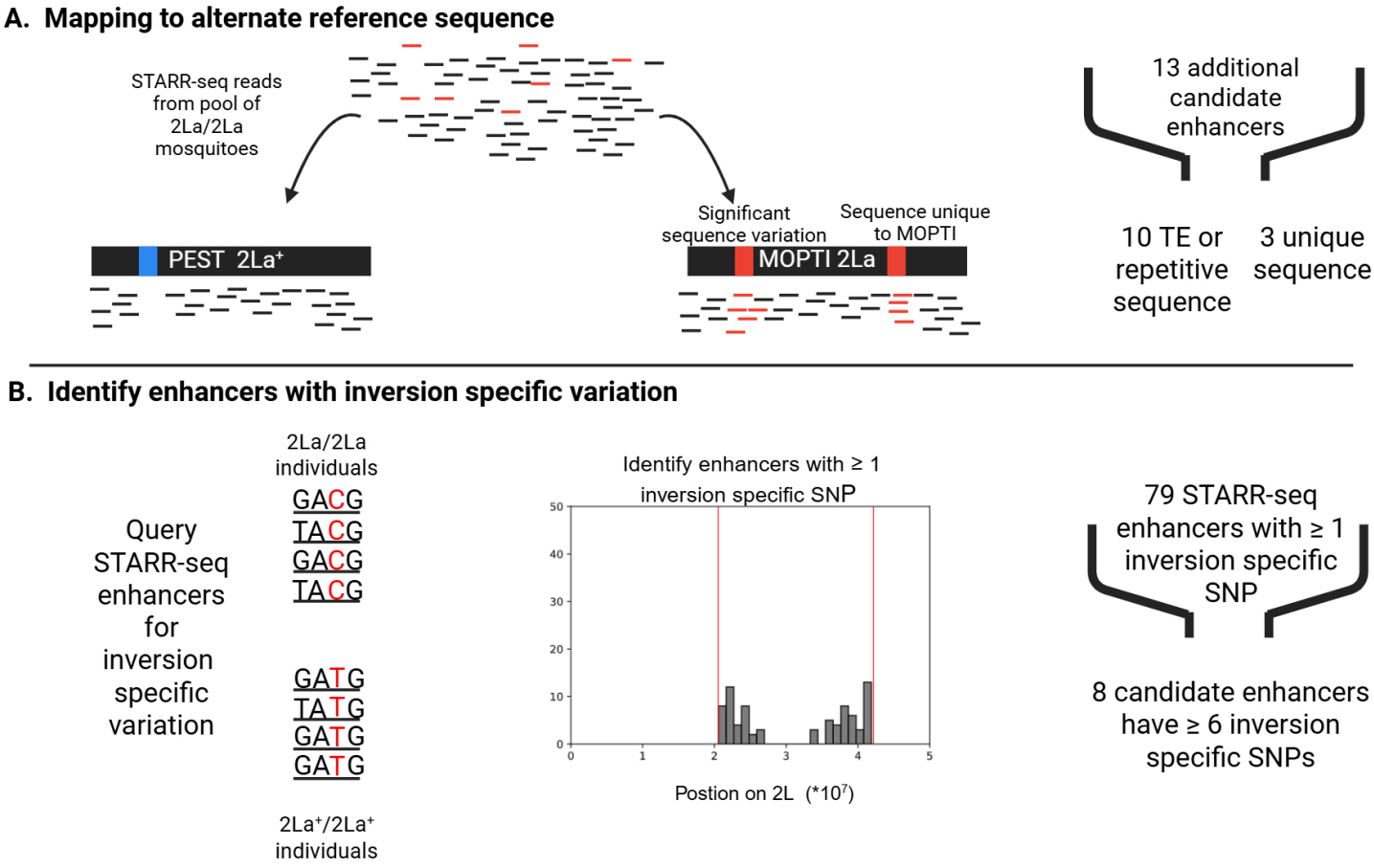

Figure 1

Figure 1

Schematic representation of the two complementary approaches used to identify candidate 2La inversion-specific enhancers A. Mapping of STARR-seq reads, originally mapped to the PEST genome [26], and now mapped to a reference genome colinear with the 2La samples. **B.** Identification of STARR-seq identified candidate enhancers carrying a high (≥ 6) burden of inversion-specific SNPs. Created with Biorender.

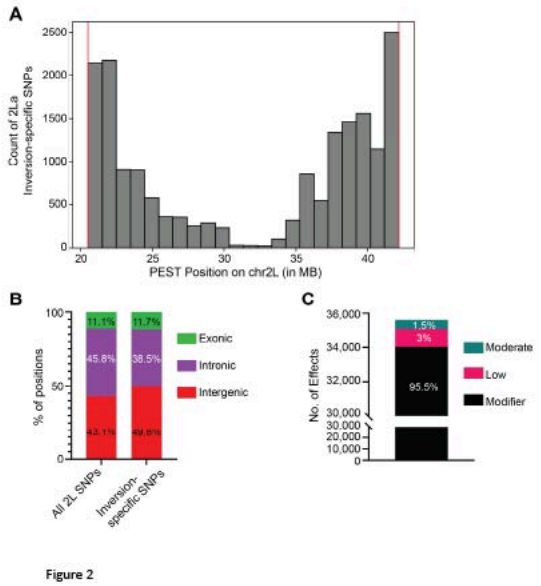

Figure 2

Figure 2

2La inversion-specific SNPs are enriched near inversion breakpoints and in intergenic regions and are mostly non-coding variants. A. Histogram of 2La inversion-specific SNPs on chr2L**.** Inversion-specific SNP identification was performed on Ag1000G data that contained whole genome sequences of 765 mosquitoes (246 2La/2La; 324 2L+^{a} /2L+^{a}; 195 2La/2L+^{a}) from 8 different countries across Africa. Inversion-specific SNPs were defined as positions where 2La/2La mosquitoes were fixed for one allele, 2L+^{a}/2L+^{a} mosquitoes were fixed for another allele, and 2La/2L+^{a} mosquitoes had high heterozygosity, using a threshold of >95%. Using these criteria, 18,072 inversion-specific SNPs were identified that segregate with the 2La inversion, with enrichment near the inversion breakpoints and few inversion-specific SNPs in the middle of the inversion. Greater than 99% of inversion-specific SNPs occur within the bounds of the 2La inversion breakpoints (vertical red lines). Zero inversion-specific SNPs occur outside the proximal breakpoint (PEST position 20528073bp), and seven occur outside the distal breakpoint (PEST position 42165625bp), but these occur within 832bp of the distal breakpoint. Each vertical gray bar represents the number of inversion-specific SNPs within a ~1MB window. **B.** Distribution of inversion-specific SNPs over exonic, intronic, and intergenic regions on chromosome 2L. SNP distribution over exonic, intronic, and intergenic regions was determined for all 2L SNPs (2,544,892 positions) as well as 2La inversion-specific SNPs. Exonic, intronic, and intergenic regions were extracted from PEST (VectorBase-68_AgambiaePEST.gff). A chi-square goodness-of-fit test comparing inversion-specific SNP counts to expected counts based on proportion of all 2L SNPs in exonic, intronic, and intergenic regions indicates that inversion-specific SNPs are enriched in intergenic regions and depleted in intronic regions, $p < 0.0001$. (C) Predicted effect classification of inversion-specific SNPs. SnpEff was used to predict the effects of the 18,072 inversion-specific SNPs on genes and proteins. Predictions were based on *Anopheles gambiae* PEST reference genome version AgamP4 (VectorBase release 68). 35,618 effects were predicted. 95.5% of predicted effects are modifiers, defined as usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact, 3% are low, defined as assumed to be mostly harmless or unlikely to change protein behavior, and 1.5% of effects are moderate, defined as a non-disruptive variant that might change protein effectiveness.

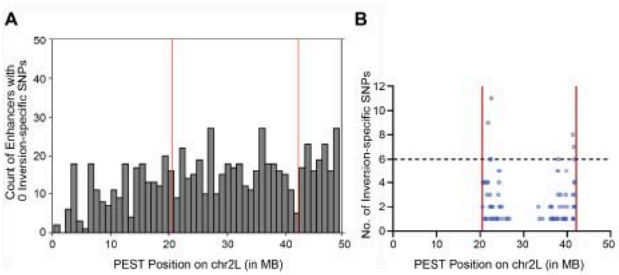

Figure 3

Figure 3

Anopheles enhancers that contain 2La inversion-specific SNPs. A, Histogram of the STARR-seq identified enhancers ([26]) lacking inversion-specific SNPs. Enhancers lacking inversion-specific SNPs are evenly distributed across chromosome 2L. Vertical red lines indicate 2La inversion breakpoints (proximal: PEST position 20528073bp; distal: PEST position 42165625bp). Each bar represents the number of enhancers without inversion-specific SNPs within an approximately 1MB window. **B.** Scatter plot of inversion-specific SNP burden of STARR-seq identified enhancers containing ≥1 inversion-specific SNP. Blue dots are plotted at the middle of each enhancer. Vertical red lines indicate 2La inversion breakpoints (proximal: PEST position 20528073bp; distal: PEST position 42165625bp). Eight enhancers containing at least six inversion-specific SNPs (at or above the horizontal dotted line) were selected for further investigation.

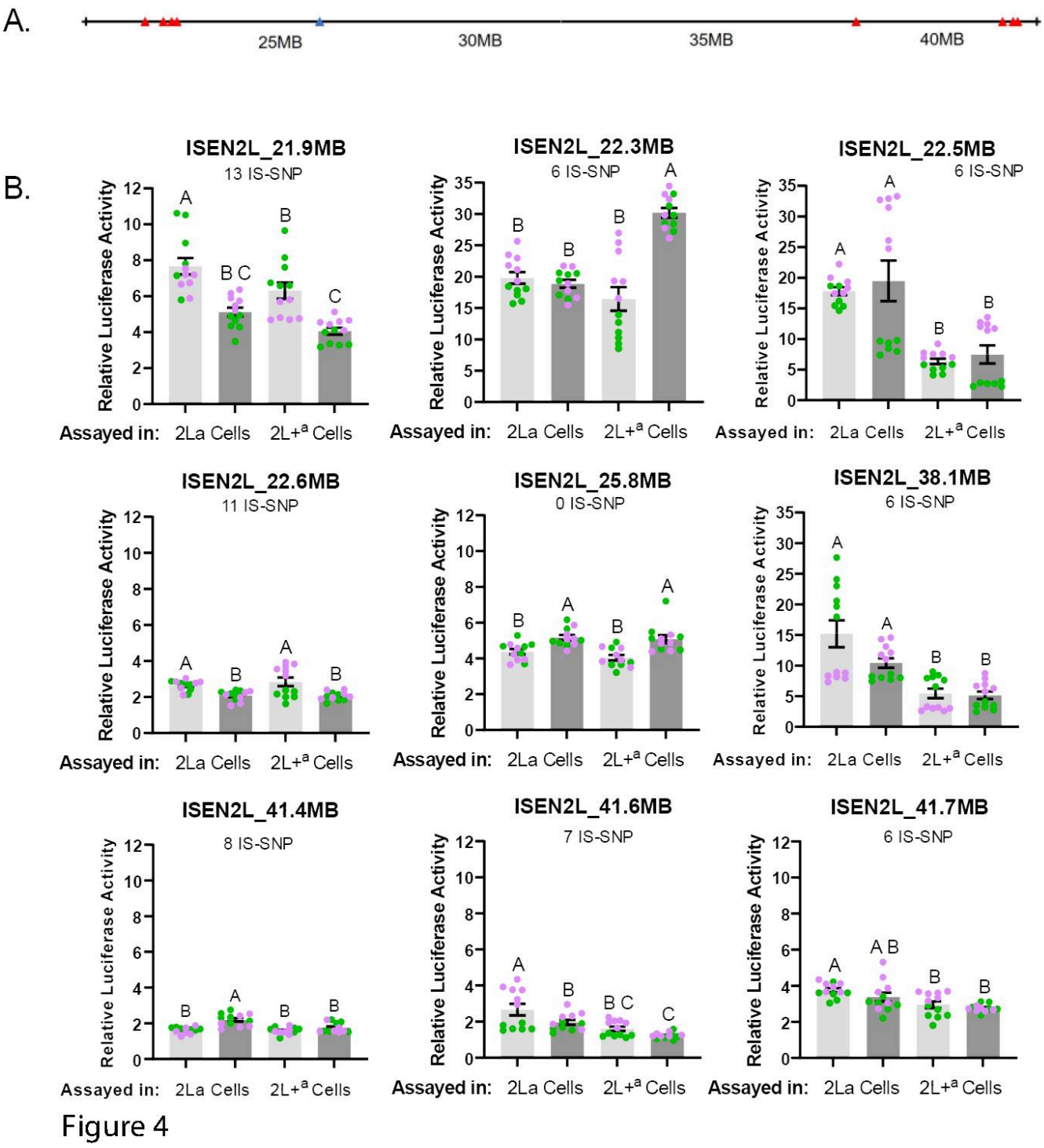

Figure 4

Figure 4

Candidate Inversion-specific Enhancer Activity Highlights Enhancer Activity Correlated with Inversion-specific Variation A. Graphic of the 2La inversion portion of the 2L chromosome arm with the location of candidate inversion-specific enhancers highlighted. The single candidate inversion-specific enhancer selected by mapping reads to the MOPTI genome is in blue and the eight candidate inversion-specific enhancers with a high burden of inversion-specific SNPs (≥ 6) are in red. **B.** All nine inversion-specific enhancers were assayed for activity. Activity was assayed for 2 alleles for each of the inversion karyotypes; one allele cloned from a wild mosquito homozygous for either the 2La of 2L+^{a} inversion form and the other alleles from a homozygous hemocyte like cell line (2La/2La- Sua4.0 or 2L+^{a}/2L+^{a} Ag55). The sequences of all tested alleles are in Additional File 2 Table S2. The x-axis indicates the cellular environment in which the activity was tested. The y-axis of each plot indicates the relative luciferase activity expressed as firefly luciferase corrected to the renilla luciferase internal control and normalized of the value for the negative control (see Methods). Beneath the name of each enhancer is the number of inversion-specific SNPs (IS-SNP) alternatively fixed between 2La and 2L+^{a} homokaryotes. Enhancer activity was assayed in 2 cellular backgrounds, the

Sua4.0 hemocyte like cell line fixed for 2La and the Ag55 (Mos55) hemocyte like cell line fixed for 2L+^{a}. Enhancer activity of alleles cloned from homozygous 2La cell lines or wild mosquitoes were averaged and are shown in the light gray bars and enhancer activity of alleles cloned from homozygous 2L+^{a} cell lines or wild mosquitoes are shown in the dark gray bars. Relative luciferase activity was measured in 6-fold technical replicates per allele/cell type combination. Data from the cell line derived allele is shown in green dots and data from the wild mosquito derived allele is shown in purple dots. A one-way ANOVA followed by Tukey's multiple comparison test was used to statistical analysis. Within a given plot, bars labelled with the same letter indicate a lack of statistically significant differences in activity enhancer activity while bars labelled with different letters are indicative of significantly different enhancer activity ($p<0.05$). The most biologically relevant cases are the outermost bars in which enhancer alleles are assayed in their native cell environment.

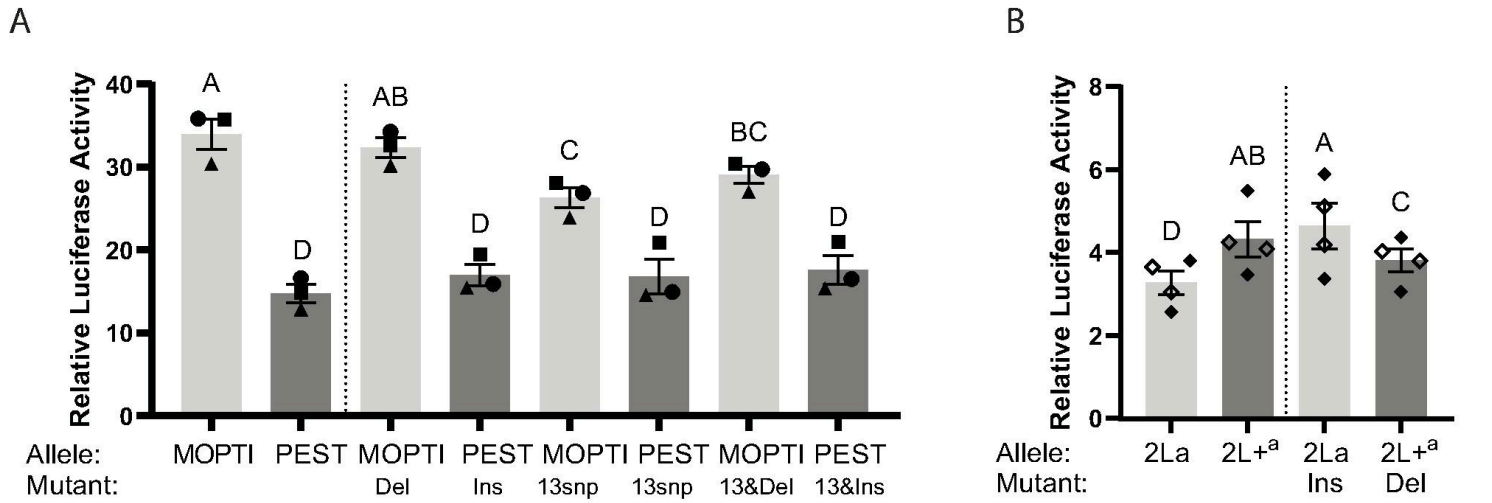

Figure 5

Figure 5

Inversion-specific genetic variation underlies differences in enhancer activity. Based on consistent findings across cell lines (see Figure 4), enhancer activity was functionally dissected by assaying activity in the 2La homozygous cell line (Sua 4.0). **A.** For ISEN2L_21.8MB allele sequence from the A. coluzzii MOPTI 2La reference genome was used as a representative 2La allele and allele sequence from the A. gambiae PEST reference genome was used as a representative 2L+^{a} allele. ISEN2L_21.9MB alleles were synthesized (IDTdna.com) in two forms (1) wild type sequence and (2) wild type mutated only at the 13 inversion-specific variant positions (all sequences are in Additional File 2 Table S2). Impacts of an eight bp inversion-specific insertion/deletion present in the MOPTI reference genome were also assayed, in isolation from and in combination with the thirteen inversion-specific SNPs. Neither SNP mutations nor insertion variation had a significant impact, individually or in combination, on the enhancer activity of the PEST allele of the ISEN2L_21.9Mb enhancer. However, mutation of all 13 inversion-specific SNPs, significantly ($p<0.0001$) decreased activity level of the MOPTI allele to a level intermediate between the PEST and MOPTI wild type alleles. The 8bp deletion did not significantly impact enhancer activity. All wild type alleles and mutants were tested across 3 biological replicates (depicted as square, circle and triangle symbols), with 6 technical replicates within each biological replicate. Data for wild type alleles is shown to the left of the vertical dashed line and data for mutant alleles to the right. Enhancer activity was analyzed using a repeated measure 2-way ANOVA with the Geisses-greenhouse correction, considering enhancer alleles and biological replicate as factors. Tukey's multiple comparisons test was used for pairwise comparisons. Letters above the bars indicate statistical significance ($p<0.05$); if bars are labeled with the same letter measured enhancer activity was not statistically different. **B**. Wild type PEST and MOPTI alleles were synthesized (IDTDNA.com) and alleles from the Sua4.0 and Ag55 cell lines were PCR amplified. The 32 bp indel associated with the 2La inversion was deleted from 2L+^{a} alleles (PEST and Ag55) and inserted into 2La alleles (MOPTI and Sua4.0) using site directed mutagenesis. Cell line alleles derived from Sua 4.0 (2La) and Ag55 (2L+^{a}) cells are represented by filled diamonds and 2La MOPTI and PEST 2L+^{a} reference genome alleles are represented by open diamonds. Wild type alleles are shown to the left of the vertical dashed line with mutant alleles to the right. The inversion-specific 32 bp indel within the newly identified inversion-specific enhancer, was necessary and sufficient to account for the significant difference in enhancer activity observed between 2La and 2L+^{a} wild type alleles. Alleles containing the insertion have significantly greater activity than those with this segment deleted (2La= $p<0.0001$, 2L+^{a} $p=0.04$). Statistical analysis and significance labelling as in A.

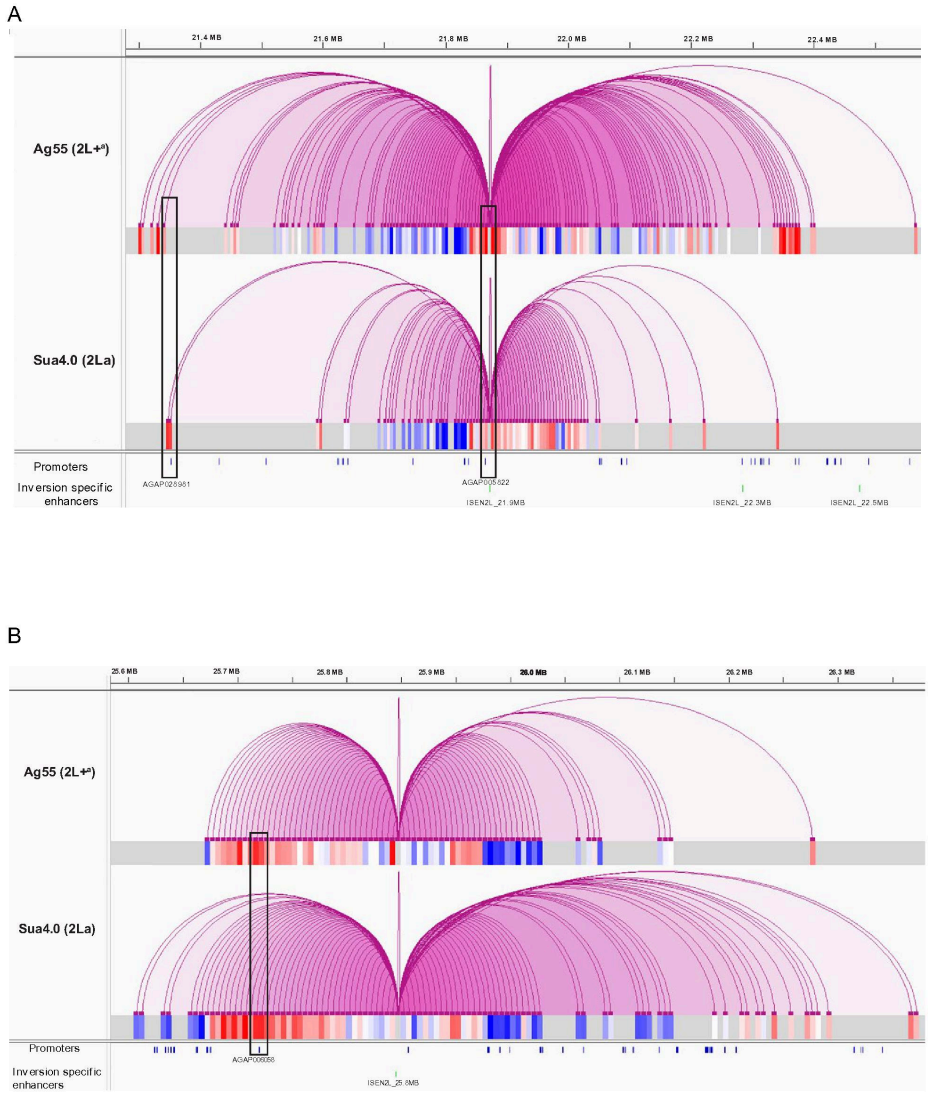

Figure 6

Figure 6

Micro-C enabled identification of candidate target genes for functionally characterized enhancers with 2La inversion-specific activity. Briefly, Micro-C was performed on basally growing *Anopheles coluzzii* hemocyte cells of alternate inversion forms (Sua4.0 for 2La and Ag55 for 2L+^{a}). The R package multiHiCcompare[3] was used to identify all interacting chromatin regions between 2La and 2L+^{a} cell lines (log_2FC>1, and p-adj <0.05), focusing on the regions containing ISEN2L_21.9MB in A and ISEN2L_25.8MB in B. A 500kb window is shown. Purple bars are 5kb regions. Purple arcs indicate an interaction between two 5kb regions. Chromatin interactions anchored on the inversion-specific enhancer of interest, in **A**, ISEN2L_21.9MB and in **B**, ISEN2L_25.8MB. Chromatin interactions are shown for both Ag55 (2L+^{a}) (top) and Sua4.0 (2La) (bottom) cells with data mapped to the PEST reference genome to enable comparison. Promoters (blue rectangles) and inversion-specific enhancers (green rectangles) are shown below the chromatin loops and above are the PEST coordinates on the 2L arm of chromosome 2 **A.** Used distance decay metrics to identify interacting regions enriched for their distance, overlayed with the

location of Anopheles gambiae gene promoters (defined at 1000bp upstream of the TSS) to identify AGAP028981 and AGAP005822 as the target genes (the anchors of greatest interaction enrichment) for ISEN2L_21.9MB in 2La and 2L+^{a} cells, respectively. Distance decay z scores are depicted in the heatmap, with red indicating interactions more frequent than expected given their distance, blue indicating interactions less frequent than expected given their distance and white indicating interaction frequencies as predicted for distance. **B.** As in A, the candidate target gene of ISEN2L_25.8MB in both 2La and 2L+^{a} is AGAP006058.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.docx
- AdditionalFile2July28.xlsx
- Additionalfile1SupplementaryFiguresAugust28.pdf