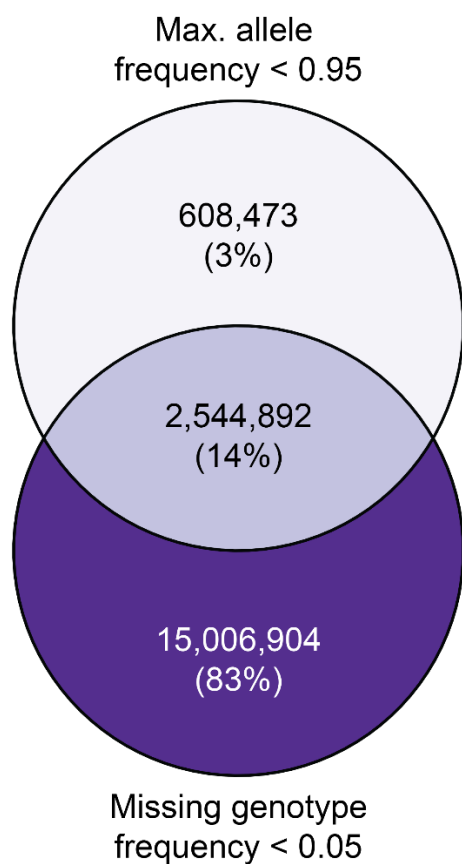
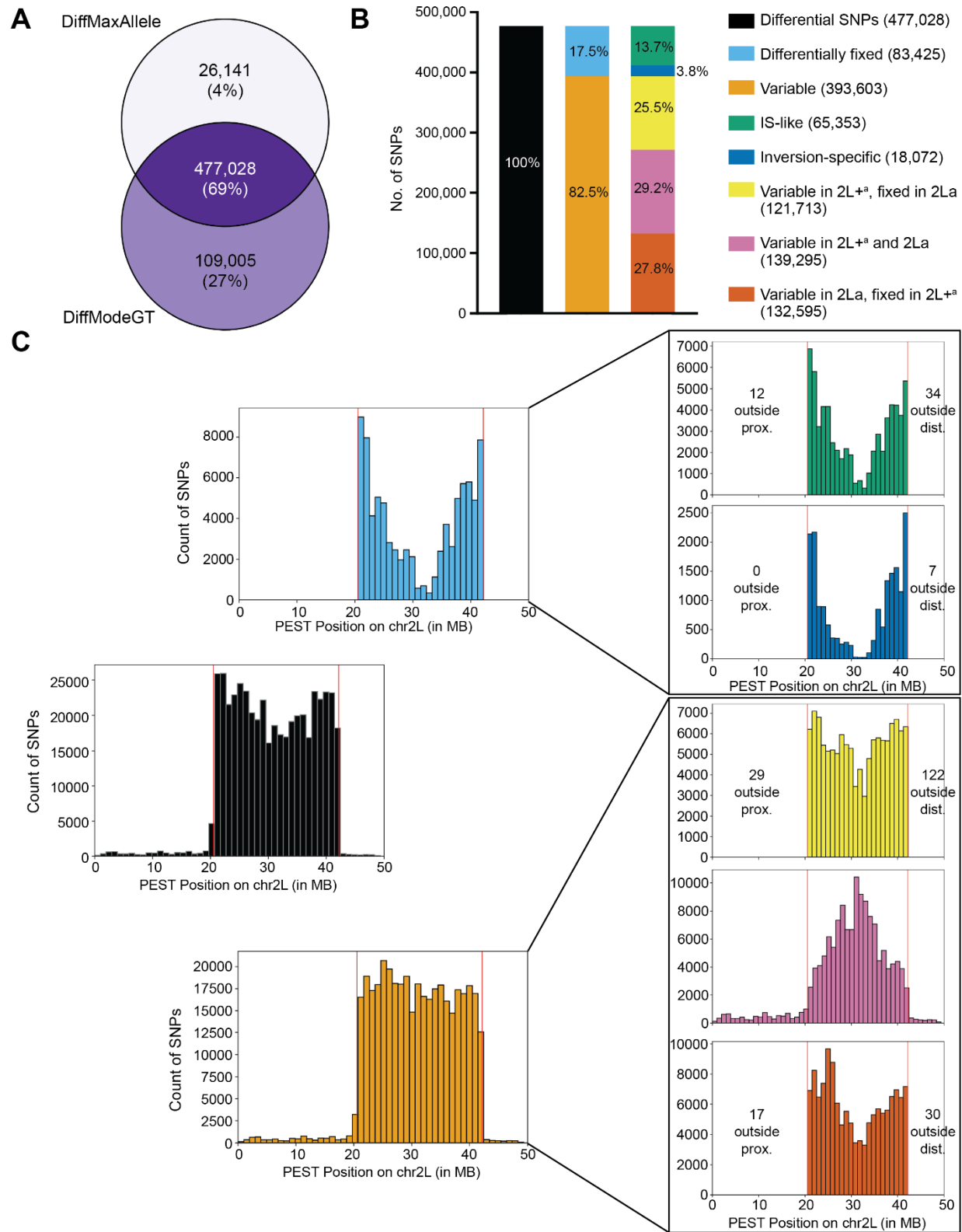




Additional File 1. Figure S1.

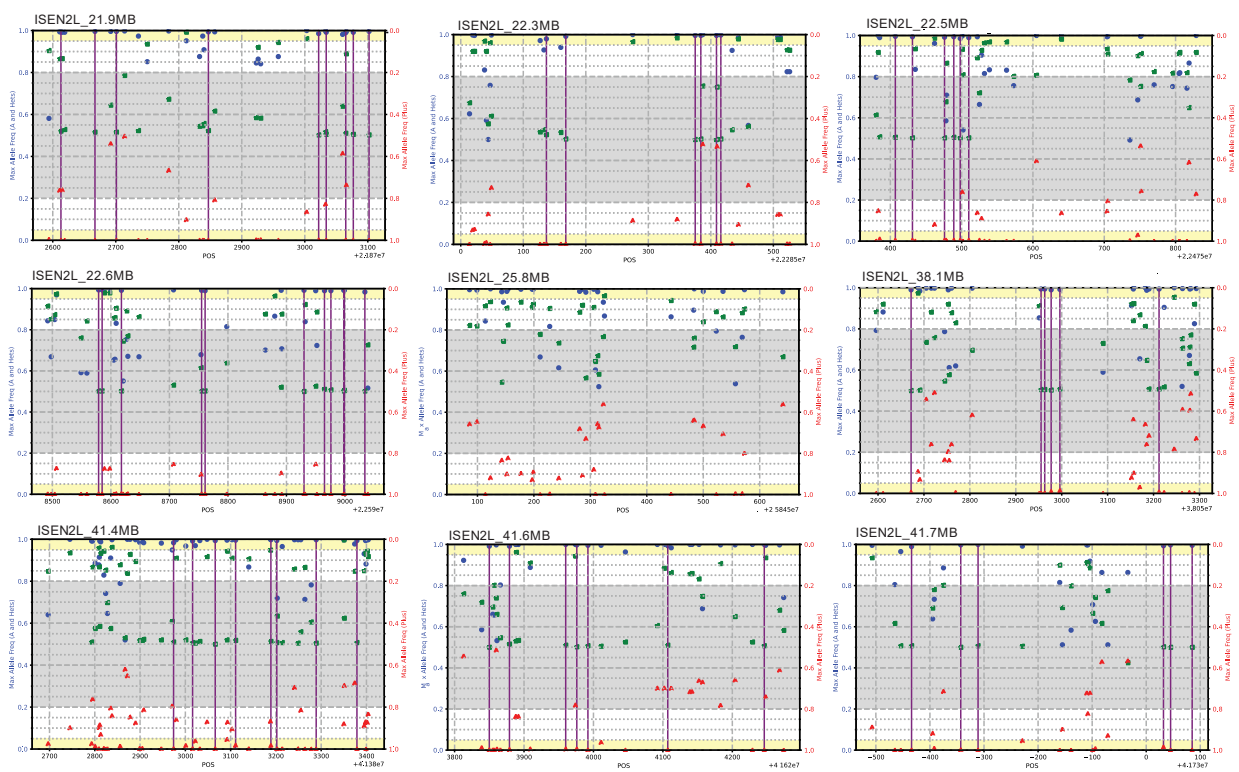
Additional File 1 Fig S1. Ag1000G phase 1 AR3 variant data was filtered down to 2,544,892 SNPs. To identify the number of reliably genotyped, variable positions in the Ag1000G phase1 AR3 variant dataset, SNPs with reliable genotypes (having a missing genotype frequency <0.05 ; 17,551,796 positions) and SNPs that were not fixed (having a maximum allele frequency <0.95 ; 3,153,365 positions) were determined, and the overlap (2,544,892 SNPs) was used for downstream data analysis.



Additional File 1. Figure S2

Additional File 1 Fig2. Categorization of differential SNPs between 2La/2La and 2L⁺/_a/2L⁺/_a mosquitoes. (A) Venn diagram of positions where 2La/2La and 2L⁺/_a/2L⁺/_a mosquitoes have a different maximum allele (503,169 positions) and positions where 2La/2La and 2L⁺/_a/2L⁺/_a mosquitoes have a different mode genotype (667,033 positions). The overlap, which is differential SNPs between 2La/2La and 2L⁺/_a/2L⁺/_a mosquitoes, consists of 477,028 positions whose allele frequencies were further characterized. **(B)** Stacked bar graph characterizing the 477,028 differential SNPs between 2La/2La and 2L⁺/_a/2L⁺/_a mosquitoes. Variable SNPs were defined as positions where the maximum allele frequency <0.95, and fixed SNPs were defined as positions where the maximum allele frequency ≥0.95. Of the 477,028 differential SNPs between 2La/2La and 2L⁺/_a/2L⁺/_a mosquitoes, 17.5% are differentially fixed while 82.5% are variable. More specifically, 3.8% are inversion-specific SNPs, using a genotype frequency threshold of >95% to determine fixation for all inversion karyotype genotypes (2La homozygotes, 2L⁺/_a homozygotes, and 2La/2L⁺/_a heterozygotes) while 13.7% are inversion-specific-like (IS-like), where the mode genotype frequency for 2La/2La mosquitoes was at least 78% and the mode genotype frequency for 2L⁺/_a/2L⁺/_a mosquitoes was at least 83%. For the variable SNPs, 27.8% are variable in 2La/2La but fixed in 2L⁺/_a/2L⁺/_a, 29.2% are variable in both, and 25.5% are variable in 2L⁺/_a/2L⁺/_a but fixed in 2La/2La. **(C)** Histograms of the distribution on chromosome 2L of each SNP category. Vertical red lines indicate 2La inversion breakpoints (proximal: PEST position 20528073bp; distal: PEST position 42165625bp). The 477,028 positions where 2La/2La and 2L⁺/_a/2L⁺/_a mosquitoes have a different maximum allele and a different mode genotype (black histogram) are distributed mostly within the 2La inversion, with some positions occurring elsewhere on chromosome 2L. The variable SNPs (orange histogram) have the same pattern as the 477,028 differential SNP positions while the differentially fixed SNPs (light blue histogram) show a U-shaped curved within the 2La inversion. Inversion-specific SNPs (dark blue histogram), IS-like SNPs (green histogram), SNPs variable only in 2L⁺/_a/2L⁺/_a mosquitoes (yellow histogram), and SNPs variable only in 2La/2La mosquitoes (red histogram) are largely confined within the 2La inversion breakpoints

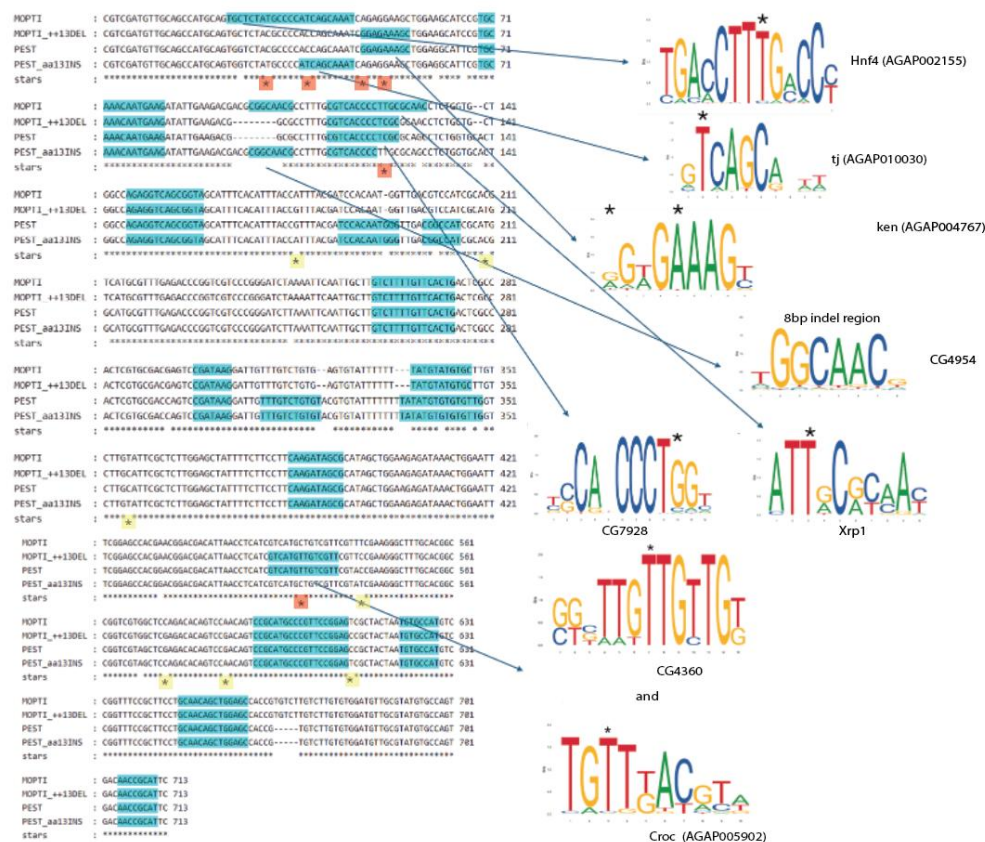
with enrichment near the breakpoints and reduction in the middle of the inversion (chi-square goodness-of-fit, $p < 0.0001$). The number of SNPs that occur outside of the 2La breakpoints are indicated to the left and right of the histogram. For the inversion-specific SNPs and IS-like SNPs, the SNPs outside of the 2La inversion fall within 4,430bp of each breakpoint. For the SNPs variable in only one of the homokaryotes, the SNPs outside of the 2La inversion fall within 44,593bp of each breakpoint. SNPs variable in both 2L⁺_a/2L⁺_a and 2La/2La mosquitoes (pink histogram) show the inverse pattern where there is enrichment in the middle of the inversion compared to the breakpoints.



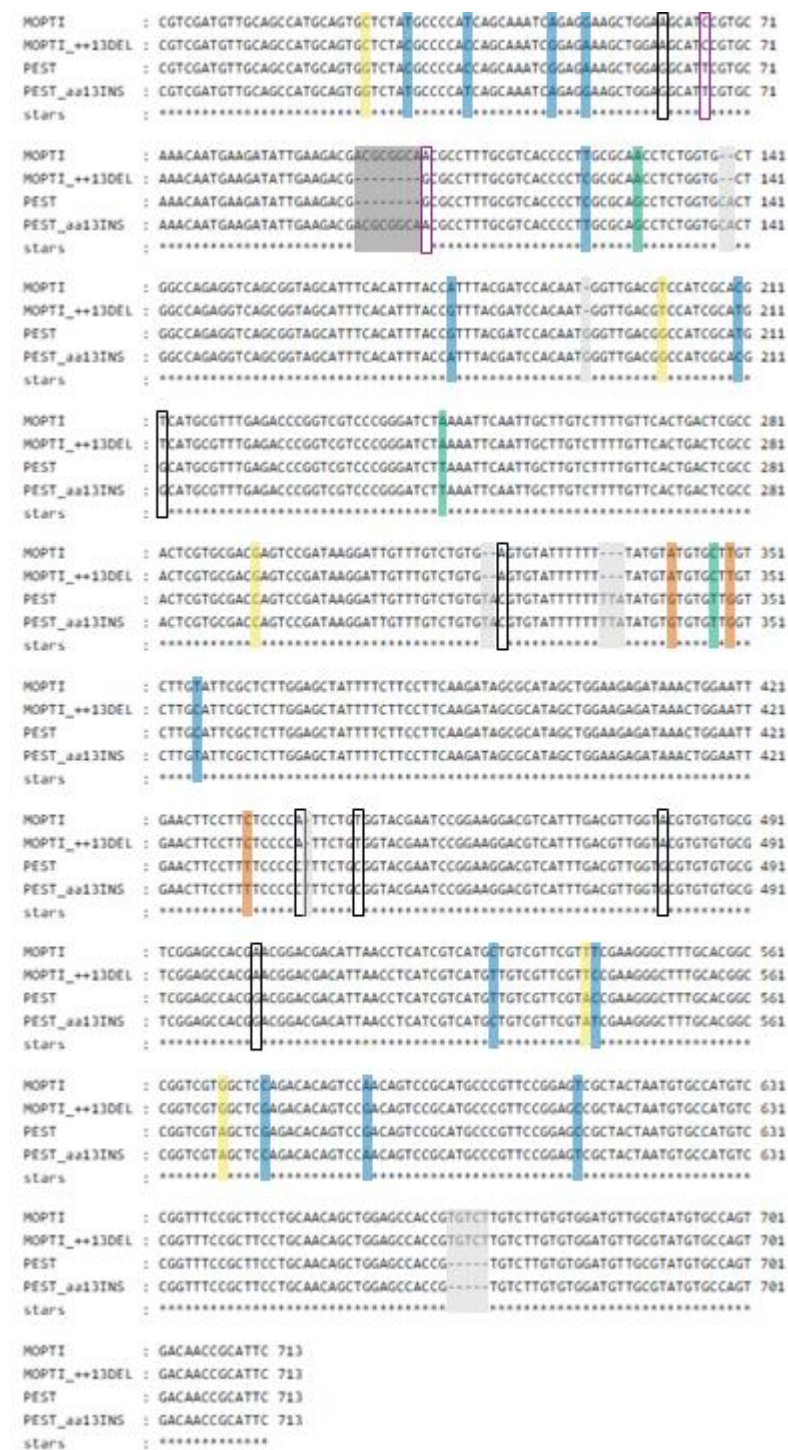
Additional File 1. Figure S3

Additional File 1 FigS3. Genetic variation at each of the 9 candidate inversion specific enhancers assayed in Figure 1. The left y-axis gives the maximum allele frequency for both 2La homozygotes (blue dots) and 2La/2L+a heterozygotes (green dots) with 1 at the top of the plot and 0 at the bottom. The right y-axis displays the maximum allele frequency for 2La+a homozygotes (red dots), with 0 at the top of the plot and 1 at the bottom. Yellow shading indicates allele frequencies >95% which indicates allele fixation (top for 2La and bottom for 2+a). Gray shading indicates allele frequencies between 20-80% which represents variable positions. Vertical purple bars highlight the locations of inversion specific SNPs. Note, at these vertical bars, the blue dot is in the upper yellow region, the red dot is in the lower yellow region and the green dot is at ~50%, meeting the criteria applied to identify an inversion specific SNP. These plots depict the STARR-seq enhancer regions defined in Holm et al. To ensure amplification of the entire enhancer for luciferase assays, primers extended left and right on these regions and in one case altered the number of inversion specific SNPs; here ISEN2L_21.9MB has nine inversion specific SNPs and in the luciferase assay amplicon it has thirteen.

A.



B.



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

ISEN2L_21.9MB A. Predicted transcription factor binding sites over inversion specific variation in ISEN2L_21.9MB. An alignment of the MOPTI and PEST wild type alleles along with the experimentally generated mutations differing in 13 SNP positions and one 8bp indel. Blue shaded regions of the alignment highlight sequences with significant similarity to annotated transcription factor binding sites ($p > 0.0005$). Here seven of the thirteen inversion specific SNPs lie within predicted transcription factor binding sites (red highlighted stars beneath the alignment) and the remaining six lie in regions lacking a predicted transcription factor binding site (highlighted yellow stars). For inversion specific SNPs occurring over transcription factor binding sites are linked (by an arrow) to the seqlogo for the transcription factor binding site with the variable position(s) noted above the seqlogo with an asterisks. Note, in a seq logo, the most conserved positions are those with the greatest information content as measured in bits and as indicated by the height of the letter. Where available, the *Anopheles gambiae* PEST orthologues are given. **B.** All variable sites between the MOPTI and PEST alleles of the ISEN2L_25.8MB enhancer including those that are inversion-specific (shaded blue), inversion-specific like (see Methods, Additional File 1 Figure S2C, bar graph in green) , those that are variable in 2L⁺^a but fixed in 2La (shaded yellow), and those that are variable in 2La but fixed in 2L⁺^a (shaded orange). The 8bp inversion-specific indel is shown in dark gray. Additional indels are shown in light gray. Open black rectangles indicate positions that are considered fixed either because they are not within the Ag1000G variant data or because they have a maximum allele frequency > 0.95 (see Methods). Open purple rectangles indicate positions that are not considered SNPs between 2La and 2L⁺^a individuals either because

$2L^a/2L^a$ and $2L^{+^a}/2L^{+^a}$ mosquitoes have the same max allele at that position or have the same mode genotype.

>Ag55_ISEN2L_25.8MB_Indel

AGTTCCGGGGGTGTATGTGTGTGTGCCGTGGC

```
>PEST_ISEN2L_25.8MB_Indel
```

AGTTCCGGGGGTGTATGTGTGTGTGCCGTGGC

Additional File 1. Figure S5. Predicted transcription factor binding sites within the indel in ISEN2L_25.8MB. Using the JASPAR available catalog of Drosophila transcription factors and the FIMO tool within MEME Suite transcription factor binding sites were predicted for the 32bp and 40bp insertions in PEST and Ag55 sequences, respectively. The variation in site is due to the difference in length of a TG microsatellite repeat. Ets98B (MA2231.1) is a tryptophan cluster factor (PEST ortholog AGAP011341), cg (MA2107.1), ci (MA2207.1) and BtbVII (MA2297.1) are all C2H2 zinc finger factors (PEST ortholog AGAP029534, AGAP004637, and AGAP006453, respectively)



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*. While *A. gambiae* and *A. coluzzii* freely segregate for the 2La inversion in nature, the *A. gambiae* PEST and *A. coluzzii* MOPTI genomes are, 2L+a/2L+a and 2La/2La respectively, *A. arabiensis* and *A. merus* are fixed for the 2La form of the 2La inversion while *A. melas* and *A. quadriannulatus* are fixed 2⁺_a. Note for 10 of the 13 inversion specific SNPs (indicated by yellow stars) *A. arabiensis* and *A. coluzzii* MOPTI have the same nucleotide and the rest of the *A. gambiae* species complex has an alternate nucleotide. For 2 sites, all the 2La members of the *A. gambiae* complex have the same nucleotide and the 2L⁺_a members share a different variant. This pattern is also true for the 8bp indel (highlighted by a horizontal bracket in the top row of sequence). For 1 inversion specific SNP position, all 2La genomes share the same nucleotide while there is variation among the 2L⁺_a genomes. Though evolutionarily more distance, *A. epiroticus* and *A. christyi*, the closest relatives of the *A. gambiae* s.l. are included for comparison.



Additional File 1 Figure S7

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*. The 32-40bp indel present in 2L⁺ genomes and absent in 2La genomes is shown in the black box. Note—the variation in the indel length is due to differences in the length of the TG microsatellite.