ISCI, Volume *10*

Supplemental Information

CRISPR/Cas9 deletions in a conserved exon of *Distal-less* generates gains and losses in a recently acquired morphological novelty in flies

Gowri Rajaratnam, Ahiraa Supeinthiran, Rudolf Meier, and Kathy F.Y. Su

Supplemental Information

CRISPR/Cas9 deletions in a conserved exon of *Distal-less* generates gains and losses in a recently acquired morphological novelty in flies.

Authors:

Gowri Rajaratnam[1], Ahiraa Supeinthiran[3], Rudolf Meier[1,2*] and Kathy F.Y. Su[1*]

[1] Department of Biological Sciences, National University of Singapore, Singapore

[2]Lee Kong Chian Natural History Museum, Singapore

[3]University of Toronto Scarborough, Canada

[*]corresponding author

Transparent Methods

Fly cultures

Themira biloba fly cultures were kept at a constant temperature of 23 ℃ with a 12-hour light and dark cycle and maintained on a diet of duck and bovine dung.

Preparation of guide RNAs

Three CRISPR guide RNAs targeting Exon 2 and Exon 3 (homeodomain) were designed for *T. biloba*. *T. biloba* assembled transcriptome was downloaded from NCBI, Accession number PRJNA218740 (Melicher et al., 2014). Using CLC Genomics Workbench we searched for *Dll* transcript sequences within the *T. biloba* transcriptome using BLASTn against *D. melanogaster Dll.* Using CLC Genomics Workbench we then designed guides targeting Exon 2 and Exon 3 of *T. biloba Dll*. We searched for regions that contained $GGN_{19}NGG$. The guides were screened for off-target effects using a blastn query against the assembled *T. biloba* transcriptome(Melicher et al., 2014). Specifically, for sgRNA-2B, this search did not find any 100% off-target matches. However, it did identify three hits with a hit-length of 13 bp and E-values around 1E-3. Further investigation of these hits reveal that they do not flank a PAM site and so are unlikely to result in off-target effects (Supplementary table 2).

The single guide RNA (sgRNA) templates were synthesised artificially using **gBlocks® Gene Fragments (Integrated DNA Technologies).** Following *in vitro* transcription of the gblocks DNA template with a T7 polymerase (New England Biolabs), the sgRNAs were purified using the TURBO DNase kit followed by the Qiagen RNeasy kit. *In vitro* Cas9 cleavage assays were then carried out to test the guide's ability to cut the target DNA fragment in the presence of Cas9 protein (PNAbio).

Microinjection

Eggs were extracted from the egg-laying substrate after 4 hours, rinsed gently, aligned onto a coverslip, and covered with a thin layer of oil to prevent desiccation. 1ul of 1ug/ul Cas9 protein (PNAbio) was mixed with 1ul of 1ug/ul of sgRNA and injected into the posterior of the eggs using a 76mm needle. Needles were pulled fresh with a Sutter P-97 Flaming/Brown type micropipette puller. To overcome the lethality of a *Dll* knockdown in early embryogenesis (Cohen and Jürgens, 1989), the protocol was optimized to reduce the number

of mosaic mutant cells generated by injecting 4-hour old embryos with a reduced volume of the sgRNA/Cas9 complex.

The injected eggs were then rinsed thoroughly and kept in a moist chamber for 24 hours. The emerging larvae were then picked out and placed into petri dishes of bovine dung to develop. After 8-10 days, the pupae were recovered and left to develop in a moist chamber. The emerging adults were allowed to completely sclerotize for 3 days before they were screened for mutations. The mutant individuals were then preserved in 70% ethanol for imaging and DNA extraction. Images were obtained using a Visionary Digital Imaging system.

Genotyping of Mutations

QuickExtract (Bioline) solution was used to extract gDNA from mutant tissue dissected from the vicinity of affected structure. Extracted DNA was used to amplify the gene region of interest (~200 bp fragment). The tarsi of a wildtype *T. biloba* male were dissected and used as a control for this experiment.

PCR products were cleaned up with SureClean and sent for Illumina Miseq (2x301bp) sequencing at 10,000x coverage (the read count and coverage for each sample is listed in Supplementary table 3).. These mutations were mostly small deletions within the guide sequence. To ensure that the mutations observed were not an artefact of PCR or sequencing error, 3 PCR replicates (with tagged primers, Supplementary table 4) were carried out for all mutants. The reads for all three replicates were processed separately and combined after ascertaining that the three most abundant reads for each were identical. This combined dataset was then used for the quantification of the read count and proportions of mutant haplotypes for each specimen.

The DNA reads were then recovered using an in-house pipeline: the sequences were merged using PEAR(Zhang et al., 2014) and demultiplexed using an in-house script (Meier et al., 2016) to generate a fasta file for each PCR product/specimen. The reads were aligned using MAFFT v7.0 (Katoh et al., 2002). Once aligned, the sequences were submitted to the 'DNA to haplotype collapser and converter' Fabox tool.(Villesen, 2007) This generates a table of haplotypes and read counts as well as a summary of all the haplotypes observed. The three most abundant haplotypes were recorded (see Supplementary table 3). The most abundant mutant haplotype for each specimen was used for downstream alignments and comparisons. The results were also confirmed with CRISPresso (Pinello et al., 2016) (window size set to 30 bp and sequence homology for an HDR occurrence set to 98%). Bioinformatics analysis

for ESE sites were performed using ESEfinder (v 3.0)(Cartegni et al., 2003). We used the weighted matrix values for SRSF1, the human homolog of the Drosophila SF2/ASF at the threshold of 1.956(Smith et al., 2006).

Control Injections

To rule out off-target effects and injection artefacts, 1004 embryos were injected with 500ng of Cas9 alone. Note that the control injections were used to train students in the microinjection technique and as such, the mortality for the control injections is high overall due to mechanical damage. No mutants of any type were observed for the surviving adults. Adults from control injections were genotyped and no mutations were observed.

Exon-skipping

In order to detect if exon skipping was occurring when Exon 2 was targeted, RNA was extracted from two independent replicates of 160 injected embryos (C1 and C2) as well as two wildtype replicates using TRIzol (Invitrogen). The RNA was transcribed into cDNA using the ProtoScript II First Strand cDNA synthesis kit (New England Biolabs). The gene region of interest was then amplified. The same volume of product was loaded onto a 1% agarose gel for both the wildtype and mutant replicates. A shorter band (~ 400bp) was observed for only C1 and C2 and not the wildtype replicates (Fig.5A). The shorter band was then sent for Sanger sequencing, where a transcript lacking Exon 2 was found.

Protein analysis

PCR was used to generate the template needed for *in vitro* protein synthesis. Primers specific to the start and stop codons of *Dll* coding sequence were used to add specific adaptor sequences (see PURExpress manual and supplementary table 4) to the *Dll* coding sequence lacking Exon 2. A T7 promoter and ribosome binding site were added to the upstream of the start of *Dll* translation and a 35-mer loop structure was added to the 3' of *Dll* coding sequence (See PURExpress manual and supplementary table 4 for primer sequences). This PCR product was purified using the Qiagen PCR purification kit and used as a template for *in vitro* protein synthesis with the PURExpress kit (New England Biolabs). Both the Dll template as well as a control reaction with no template were set-up and processed in the same way. 8ul of this synthesised protein product was run out on an SDS-PAGE gel (10% Mini-PROTEAN® TGX™ Precast Gel) at 120V for 90 minutes, along with the control. The >30 kDa band, representing the synthesized *Dll* protein, was excised from the gel and sent for

analysis on the Triple TOF 5600 (Ab SCIEX). Unfortunately, we noticed that although the same amount of the reaction was loaded onto the gel, the control reaction appeared fainter than the reaction with the *Dll* template.

As an additional measure to rule out the possibility that the >30kDa band was simply not observed in the fainter control, we excised out the corresponding area in the control gel lane and sent it in for mass spectrometry analysis as well.

The Mass Spectra raw data was searched against the NCBI *D. melanogaster* protein database and the *E. coli* database supplemented with the *Dll* predicted protein sequence for *Themira biloba* with ProteinPilot® v4.5 (Revision: 1656; Paragon Algorithm: 4.5.0.0, 1654). Searches were run as thorough identification searches, specifying tryptic digestion and cys-alkylation (Iodoacetamide). For the *Dll* protein sample, peptides were found with strong matches to an alternative initiation codon as well as to the homeodomain and Exon 4 (Supplementary Figure 3). No confident peptide matches were found for the analogous control sample (supplementary file in Mendeley resource).

Distal-less isoform characterisation

In order to qualitatively characterise the alternative splice forms of *Dll* present in both mutant and wildtype individuals, targeted long-read isoform circular consensus sequencing(Gonzalez-Garay, 2016) was conducted for 4 individuals. 405 embryos were injected with sgRNA-2A (targeting the second ESE site within Exon 2). After 7 days, total RNA and DNA was isolated from 12 injected 3rd instar larvae using TRIzol® Reagent (Invitrogen). cDNA was synthesised using the ProtoScript II First Strand cDNA synthesis kit (New England Biolabs) for each of these individuals and an in-vitro cleavage assay was performed to identify the mutant specimens. Two individuals (M4 and M6) were observed with mutations and exon skipping.

Similarly, RNA was extracted and used for cDNA synthesis for two wildtype 3rd instar larvae (WT1 and WT2). Primers were designed within the *Dll* locus to amplify and capture *Dll* specific splice forms (Supplementary Table 4). 5’ phosphorylated forward primers were designed in *Dll* exon 1 (5' UTR) and reverse 16bp-tagged primers were designed for *Dll* exon 7 (3' UTR). A 16bp tag unique to each specimen was attached to the 3’ end of the reverse primer so as to identify the two wildtype and two mutant sequences (see Supplementary table 4 for primer and tag sequences). The PCR products were purified with Sureclean and sent in for Pacbio Sequel circular consensus sequencing with a library insert size of 2kb.

Pacbio IsoSeq files were error-corrected and converted into CCS reads; We performed 10 minimum passes with a minimum predicted accuracy of 0.9 using the PacBio SMRT analysis software (v5.1.0). The analysis generated 225,740 CCS reads with a median CCS read length between 1200 – 1249 bp (see Supplementary Figure 5 for results statistics).

The reads were then demultiplexed using Geneious with a 100% stringency match to primer and tag sequences. The demultiplexed sequences were then filtered for low read length (< 500bp) yielding 18194 reads for M4, 124385 reads for M6, 13422 reads for WT1 and 18817 reads for WT2. For each individual, the reads were collapsed into haplotypes using the 'DNA to Haplotype converter' tool in FaBox (Villesen, 2007). Haplotypes with read counts below 0.5% of the total count were discarded. The remaining haplotypes were aligned using MAFFT v7.0 and then analysed in Aliview to identify alternative splice forms (alignment files in Mendeley resource).

Distal-less expression in developing histoblast clusters

To determine if *Dll* was naturally expressed in the 3rd abdominal segment where an ectopic sternite brush was observed, RNA was extracted from dissected epidermal tissues of 3rd instar larvae. The epidermis of the thoracic segment and the abdominal segments (8 abdominal segments for *D. melanogaster* and 7 abdominal segments for *T. biloba*) were dissected for 5 wildtype 3rd instar larvae for both *T. biloba* and *D. melanogaster*. RNA was extracted for each segment and cDNA was synthesised. RT PCR for *Dll* was carried out for each segment to check for *Dll* expression in the late larval stages.

The same was done for *D. melanogaster*, which served as a control for the RT-PCR experiment. RT- PCR for an additional gene, Abdominal-B, was also carried out in D. melanogaster as a control to rule out the presence of any artefacts from the epidermal dissections. All PCR products were then PCR purified and sent in for sanger sequencing to confirm the correct products were amplified (results in Supplementary Table 5, Supplementary Figure 4).

***Distal-less* alignment across Holometabola**

A protein search performed on NCBI using the following search terms "Distal-less[All Fields] AND ("Mandibulata"[Organism] OR Mandibulata[All Fields])" yielded 467 protein sequences. This dataset was filtered for sequences that: were not Distal-less, had no homeodomain, were incomplete or of poor quality. Sequences belonging to Coleoptera,

Lepidoptera, Diptera and Hymenoptera were extracted from the filtered dataset and aligned with MAFFT v7.0(Katoh et al., 2002) and visualised in Aliview (Larsson, 2014).

Quantitative PCR

We injected embryos with both Cas9 and the guide targeting Exon 2 and compared expression levels to control embryos that were injected with Cas9 alone. We let embryos develop into first instar larvae. All control larvae were individually extracted for RNA using Trizol. All treatment larvae were first genotyped to confirm presence of mutations in Exon 2 using the T7 endonuclease kit (New England Biolabs). RNA from mutant larvae was then used for cDNA synthesis to generate template for qPCR. The Forkhead transcription factor (Mnf) gene was used as a housekeeping gene. We utilised a customised TaqMan gene expression assay with primers nested within the homeodomain and with a probe complementary to the homeodomain. The expression levels were not significantly different between wildtype and mutant larvae. However, note that CRISPR/Cas9 generates mosaic mutants that consist mostly of wildtype cells and mutants that survive are likely to only have small amounts of mutant cells. This makes it difficult to disentangle expression levels between wildtype mRNA from wildtype cells and exon-skipped transcripts of mutant cells. With naturally low *Dll* expression (Ct value > 30), a lack of significant expression difference is inconclusive.

Data availability

Most of the processed sequencing data files are available on a Mendeley database (DOI: 10.17632/ps3p7jnb5t.1). However, for pre-filtered/processed files and other data, please contact Rajaratnam, G. For the in-house bioinformatics script, please contact Meier, R.

References

CARTEGNI, L., WANG, J., ZHU, Z., ZHANG, M. Q. & KRAINER, A. R. 2003. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Research,* 31**,** 3568-3571.

COHEN, S. M. & JÜRGENS, G. 1989. Proximal—distal pattern formation in Drosophila: cell autonomous requirement for Distal-less gene activity in limb development. *The EMBO Journal,* 8**,** 2045-2055.

GONZALEZ-GARAY, M. L. 2016. Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq). *In:* WU, J. (ed.) *Transcriptomics and Gene Regulation.* Dordrecht: Springer Netherlands.

KATOH, K., MISAWA, K., KUMA, K.-I. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research,* 30**,** 3059-3066.

LARSSON, A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics,* 30**,** 3276-3278.
MEIER, R., WONG, W., SRIVATHSAN, A. & FOO, M. 2016. $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. *Cladistics,* 32**,** 100-110.
MELICHER, D., TORSON, A. S., DWORKIN, I. & BOWSHER, J. H. 2014. A pipeline for the de novo assembly of the Themira biloba(Sepsidae: Diptera) transcriptome using a multiple k-mer length approach. *BMC Genomics,* 15**,** 188.
PINELLO, L., CANVER, M. C., HOBAN, M. D., ORKIN, S. H., KOHN, D. B., BAUER, D. E. & YUAN, G.-C. 2016. Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat Biotech,* 34**,** 695-697.
SMITH, P. J., ZHANG, C., WANG, J., CHEW, S. L., ZHANG, M. Q. & KRAINER, A. R. 2006. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet,* 15**,** 2490-508.
VILLESEN, P. 2007. FaBox: an online toolbox for fasta sequences. *Molecular Ecology Notes,* 7**,** 965-968.
ZHANG, J., KOBERT, K., FLOURI, T. & STAMATAKIS, A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics,* 30**,** 614-620.

Figures

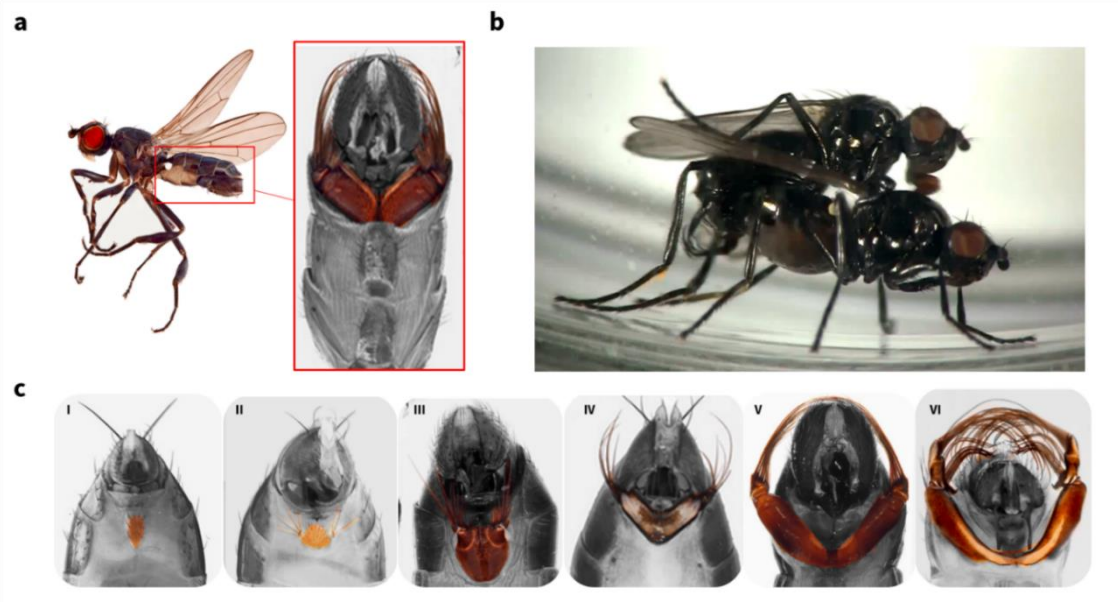


Figure 1

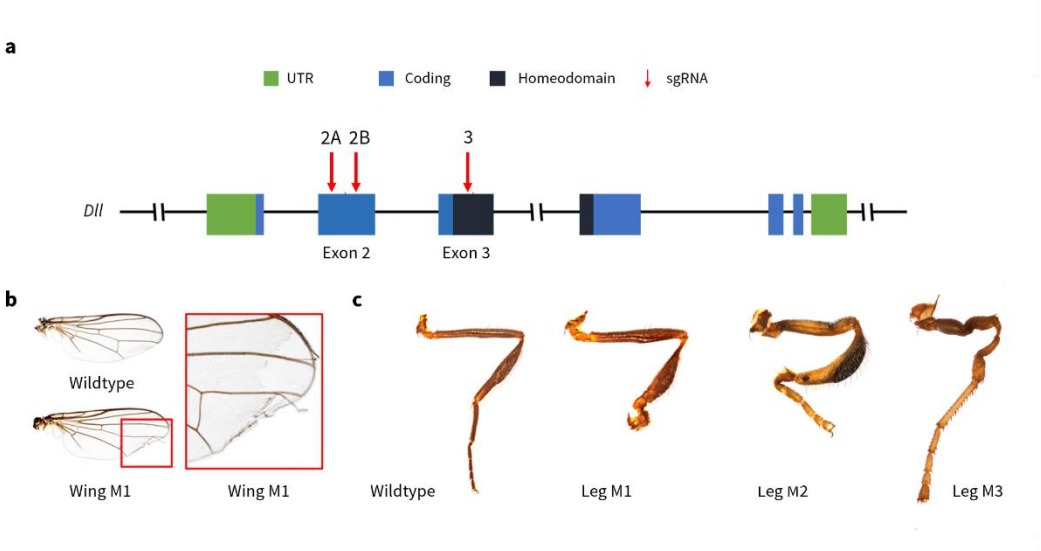


Figure 2

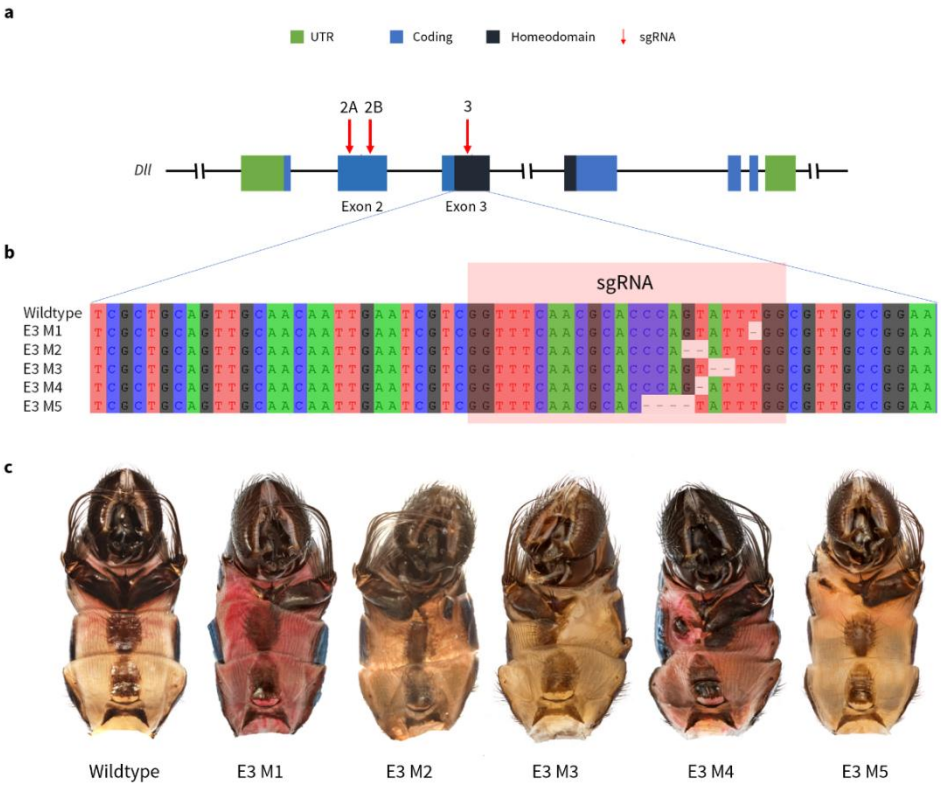

a
UTR
Coding
Homeodomain
sgRNA
2A 2B
3
Dll
Exon 2
Exon 3
b
sgRNA
Wildtype
E3 M1
E3 M2
E3 M3
E3 M4
E3 M5
c
Wildtype
E3 M1
E3 M2
E3 M3
E3 M4
E3 M5

Figure 3.

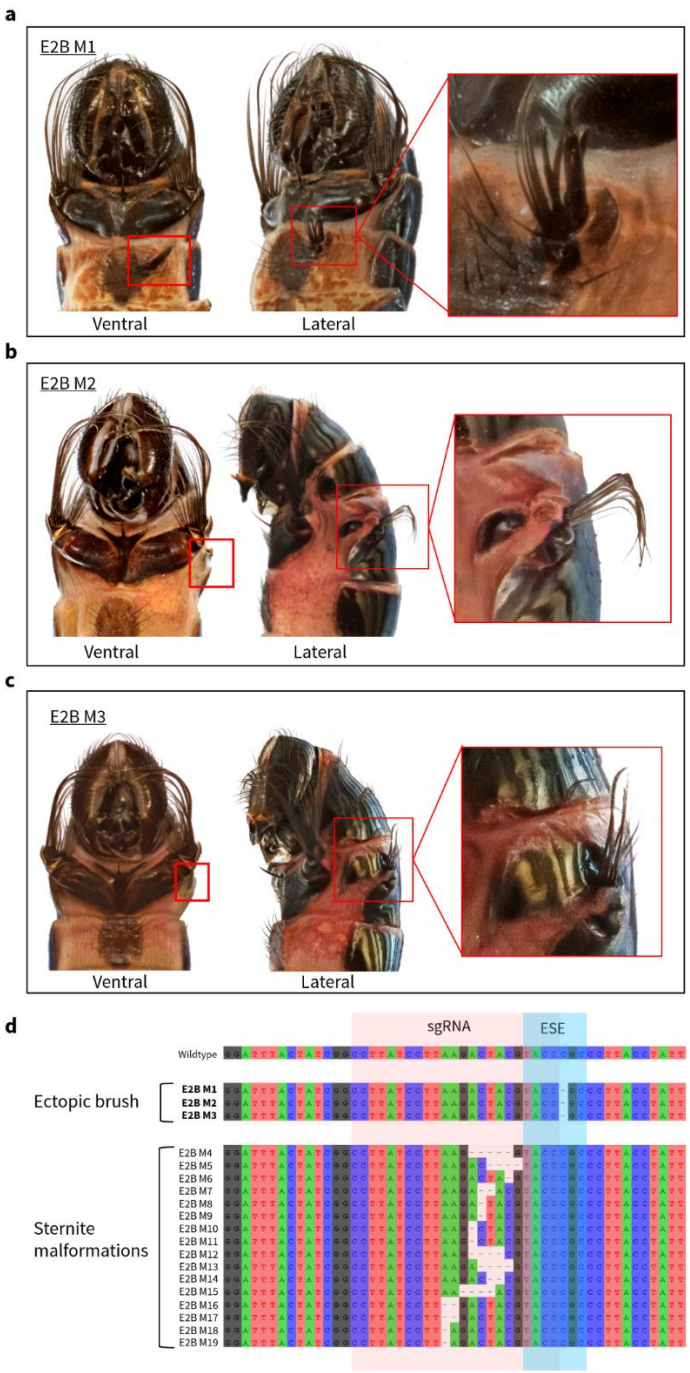

a
E2B M1
Ventral
Lateral
b
E2B M2
Ventral
Lateral
c
E2B M3
Ventral
Lateral
d
sgRNA
ESE
Wildtype
Ectopic brush
E2B M1
E2B M2
E2B M3
Sternite malformations
E2B M4
E2B M5
E2B M6
E2B M7
E2B M8
E2B M9
E2B M10
E2B M11
E2B M12
E2B M13
E2B M14
E2B M15
E2B M16
E2B M17
E2B M18
E2B M19

Figure 4

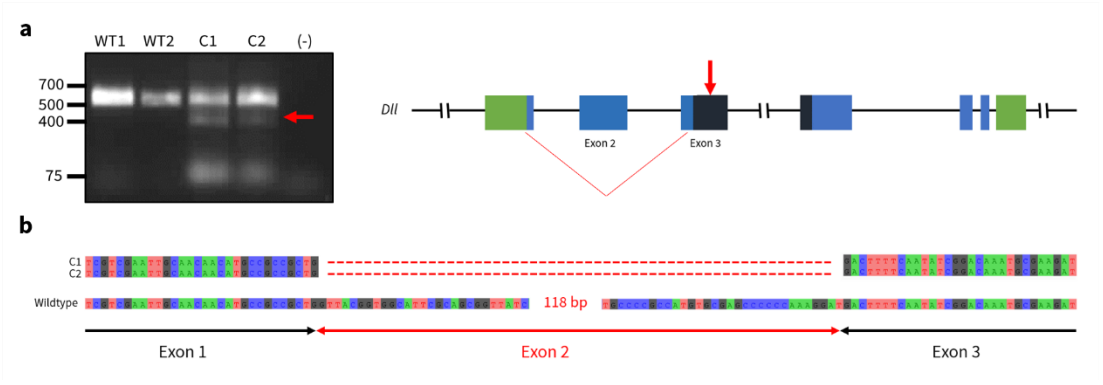


Figure 5

Figure 6

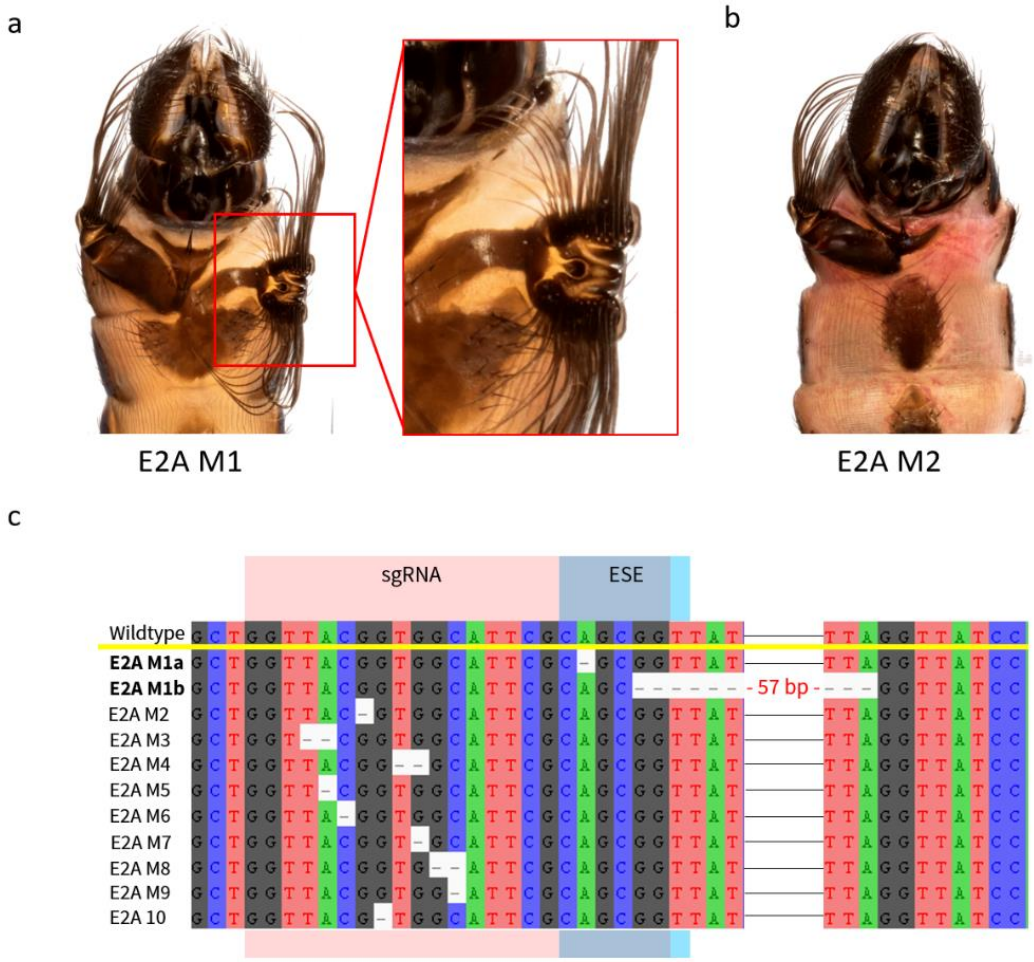

a
b
E2A M1
E2A M2
c
sgRNA
ESE
Wildtype
E2A M1a
E2A M1b
E2A M2
E2A M3
E2A M4
E2A M5
E2A M6
E2A M7
E2A M8
E2A M9
E2A 10
- 57 bp -

Figure 7

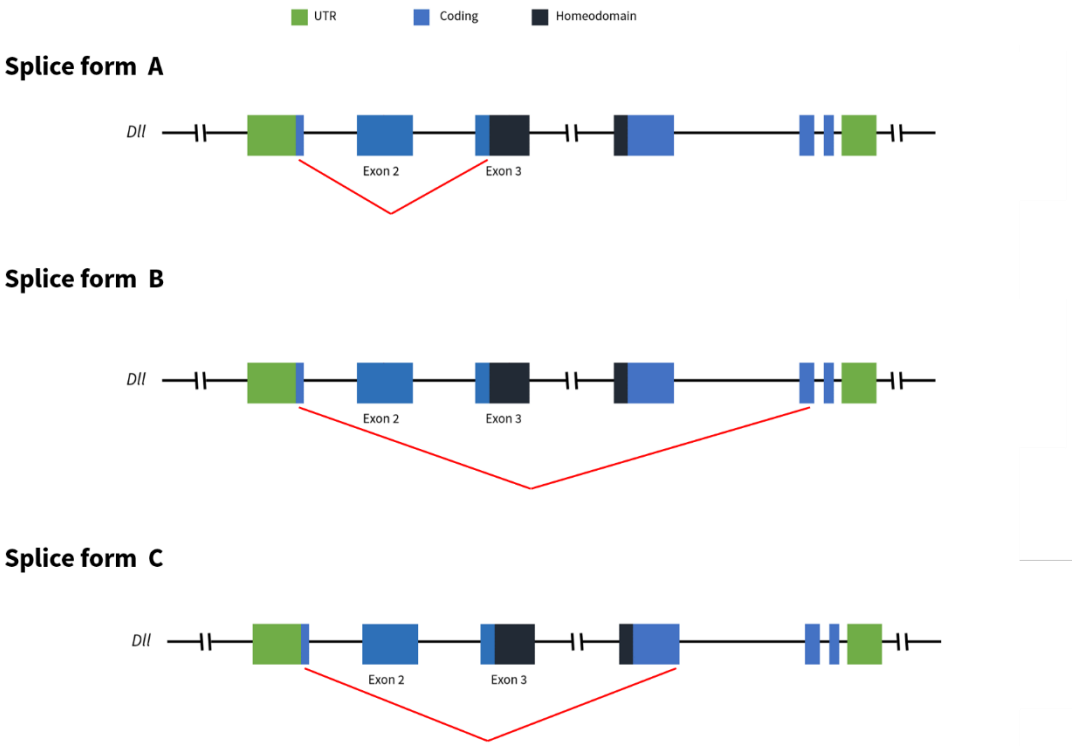


Figure 8

Tables

Table 1.

	Injected eggs	Larvae	Adults	Mutants			
				Leg/Wing	Sternite malformations	Ectopic structures	Others (e.g. clasper malformations)
Exon 2 (sgRNA 2A)	1715	605	151	1	8	1	0
Exon 2 (sgRNA 2B)	4631	957	363	8	29	3	6
Exon 3 (sgRNA3)	4003	1187	261	1	21	0	2
Control	1004	153	109	0	0	0	0

Supplementary information

Supplementary Table 1. Summary of ESEfinder search with Matrix values for SRSF1 (human homolog for SF2/ASF) with threshold of 1.956. The ESE motif present in Exon 2 sgRNA-2B is highlighted in bold while the ESE motif used to design sgRNA-2A is italicized. Related to Figure 4.

Sequence ID	Motif	Position on Exon 2	Motif	Score
gi|654231031|gb|GBGG01003309.1|	**SRSF1**	**100**	**cgcccat**	**2.21734**
gi|654231031|gb|GBGG01003309.1|	*SRSF1*	*17*	*cagcggt*	*4.39195*
gi|654231031|gb|GBGG01003309.1|	SRSF1	49	gtcagga	3.28514
gi|654231031|gb|GBGG01003309.1|	SRSF1	75	cgcagtg	2.69094
gi|654231031|gb|GBGG01003309.1|	SRSF1	15	cgcagcg	2.23323
gi|654231031|gb|GBGG01003309.1|	SRSF1	169	caaagga	3.67496

Supplementary Table 2. Summary of blastn results. Related to Figure 4.

Query	Number of HSPs	Lowest E-value	Accession (E-value)	Greatest identity %	Greatest HSP length	Greatest bit score	Sequence overlap with guide %
gi|654231031|gb|GBGG01003309.1| (Themira biloba Dll)	1	1.38661E-06	sgRNA-2B	100	20	32.7626	100
gi|654228759|gb|GBGG01004492.1|	1	0.0066504	sgRNA-2B	100	13	21.9569	65
gi|654228161|gb|GBGG01004814.1|	1	0.00242708	sgRNA-2B	100	13	21.9569	65
gi|654223893|gb|GBGG01007092.1|	1	0.00259143	sgRNA-2B	100	13	21.9569	65

Supplementary Table 3. Summary of read counts for mutant haplotypes. The 3 most dominant mutant haplotypes per individual are recorded here. The dominant haplotype (in bold) per individual is used for downstream analyses. See separate excel document. Related to Figure 2, 3, 4 and 7.

Supplementary Table 4. Primer sequences. Tags are in lower case. Related to Figure 2, 3, 4, 5, 7 and 8.

Purpose	Name	Primer sequence (5' → 3')
Characterisation of homeodomain mutants	Dll_sgrna12_F1	cagtctgGACAAATGCGAAGATTCTGG
Characterisation of homeodomain mutants	Dll_sgrna12_F2	catgggaGACAAATGCGAAGATTCTGG
Characterisation of homeodomain mutants	Dll_sgrna12_F3	tcacgtaGACAAATGCGAAGATTCTGG
Characterisation of homeodomain mutants	Dll_sgrna12_F4	tggtccaGACAAATGCGAAGATTCTGG
Characterisation of homeodomain mutants	Dll_sgrna12_F5	aacctgtGACAAATGCGAAGATTCTGG
Characterisation of homeodomain mutants	Dll_sgrna12_F6	cttggttGACAAATGCGAAGATTCTGG
Characterisation of homeodomain mutants	Dll_sgrna12_R1	cagtctgCTGCGTTTGTGTTAGGCCCA
Characterisation of homeodomain mutants	Dll_sgrna12_R2	catgggaCTGCGTTTGTGTTAGGCCCA
Characterisation of homeodomain mutants	Dll_sgrna12_R3	tcacgtaCTGCGTTTGTGTTAGGCCCA
Characterisation of homeodomain mutants	Dll_sgrna12_R4	tggtccaCTGCGTTTGTGTTAGGCCCA
Characterisation of homeodomain mutants	Dll_sgrna12_R5	aacctgtCTGCGTTTGTGTTAGGCCCA
Characterisation of homeodomain mutants	Dll_sgrna12_R6	cttggttCTGCGTTTGTGTTAGGCCCA
Characterisation of Exon 2 mutants	Dll_sgrna4_F1	aacctgtTACGGTGGCATTCGCAGCGGTTATC
Characterisation of Exon 2 mutants	Dll_sgrna4_F2	agaagtgTACGGTGGCATTCGCAGCGGTTATC
Characterisation of Exon 2 mutants	Dll_sgrna4_F3	cggttatTACGGTGGCATTCGCAGCGGTTATC
Characterisation of Exon 2 mutants	Dll_sgrna4_F4	gtacactTACGGTGGCATTCGCAGCGGTTATC
Characterisation of Exon 2 mutants	Dll_sgrna4_F5	gtgatgaTACGGTGGCATTCGCAGCGGTTATC
Characterisation of Exon 2 mutants	Dll_sgrna4_F6	cttggttTACGGTGGCATTCGCAGCGGTTATC
Characterisation of Exon 2 mutants	Dll_sgrna4_R1	aacctgtGGGCTCGCACATGGCGGG
Characterisation of Exon 2 mutants	Dll_sgrna4_R2	agaagtgGGGCTCGCACATGGCGGG
Characterisation of Exon 2 mutants	Dll_sgrna4_R3	cggttatGGGCTCGCACATGGCGGG
Characterisation of Exon 2 mutants	Dll_sgrna4_R4	gtacactGGGCTCGCACATGGCGGG

Characterisation of Exon 2 mutants	Dll_sgrna4_R5	gtgatgaGGGCTCGCACATGGCGGG
Characterisation of Exon 2 mutants	Dll_sgrna4_R6	cttggttGGGCTCGCACATGGCGGG
Characterisation of ESE mutants	Dll_intron1_F1	aacctgtGTTGTGCCTCCAAGGATTTCATAC
Characterisation of ESE mutants	Dll_intron1_F2	agaagtg GTTGTGCCTCCAAGGATTTCATAC
Characterisation of ESE mutants	Dll_intron1_F3	cggttat GTTGTGCCTCCAAGGATTTCATAC
Characterisation of ESE mutants	Dll_intron1_F4	gtacact GTTGTGCCTCCAAGGATTTCATAC
Characterisation of ESE mutants	Dll_intron1_F5	gtgatga GTTGTGCCTCCAAGGATTTCATAC
Characterisation of ESE mutants	Dll_intron1_F6	cttggtt GTTGTGCCTCCAAGGATTTCATAC
Characterisation of ESE mutants	Dll_sgrna4_R1	aacctgtGGGCTCGCACATGGCGGG
Characterisation of ESE mutants	Dll_sgrna4_R2	agaagtgGGGCTCGCACATGGCGGG
Characterisation of ESE mutants	Dll_sgrna4_R3	cggttatGGGCTCGCACATGGCGGG
Characterisation of ESE mutants	Dll_sgrna4_R4	gtacactGGGCTCGCACATGGCGGG
Characterisation of ESE mutants	Dll_sgrna4_R5	gtgatgaGGGCTCGCACATGGCGGG
Characterisation of ESE mutants	Dll_sgrna4_R6	cttggttGGGCTCGCACATGGCGGG
Amplification of *Dll* in Drosophila melanogaster	Dmel_Dll_F2	CCGATAAGTGCGAGGACTCCGG
Amplification of *Dll* in Drosophila melanogaster	Dmel_Dll_R	CTGCGTTTGCGTGAGGCCCA
Amplification of *Abd-B* in Drosophila melanogaster	Dmel_ABDB_442_F	CCCACCTACTCCTCGCCAGGCGG
Amplification of *Abd-B* in Drosophila melanogaster	Dmel_ABDB_571_R	TCCACTCGTGCAGTCCGGGATTGGGC
Shortened reverse primer with lower Tm for cycle sequencing of Drosophila melanogaster Abd-B amplicon	Dmel_ABDB_571_SEQ_R	TCCACTCGTGCAGTCCGGGATT
Amplification of *Dll* in Themira biloba	DLL Exon 2 sgRNA4 F	TAC GGT GGC ATT CGC AGC GGT TAT C
Amplification of Dll in Themira biloba	DLL Exon 2 sgRNA4 R	CAC ATG GCG GGG CAT AAC TGC CTA AAT G

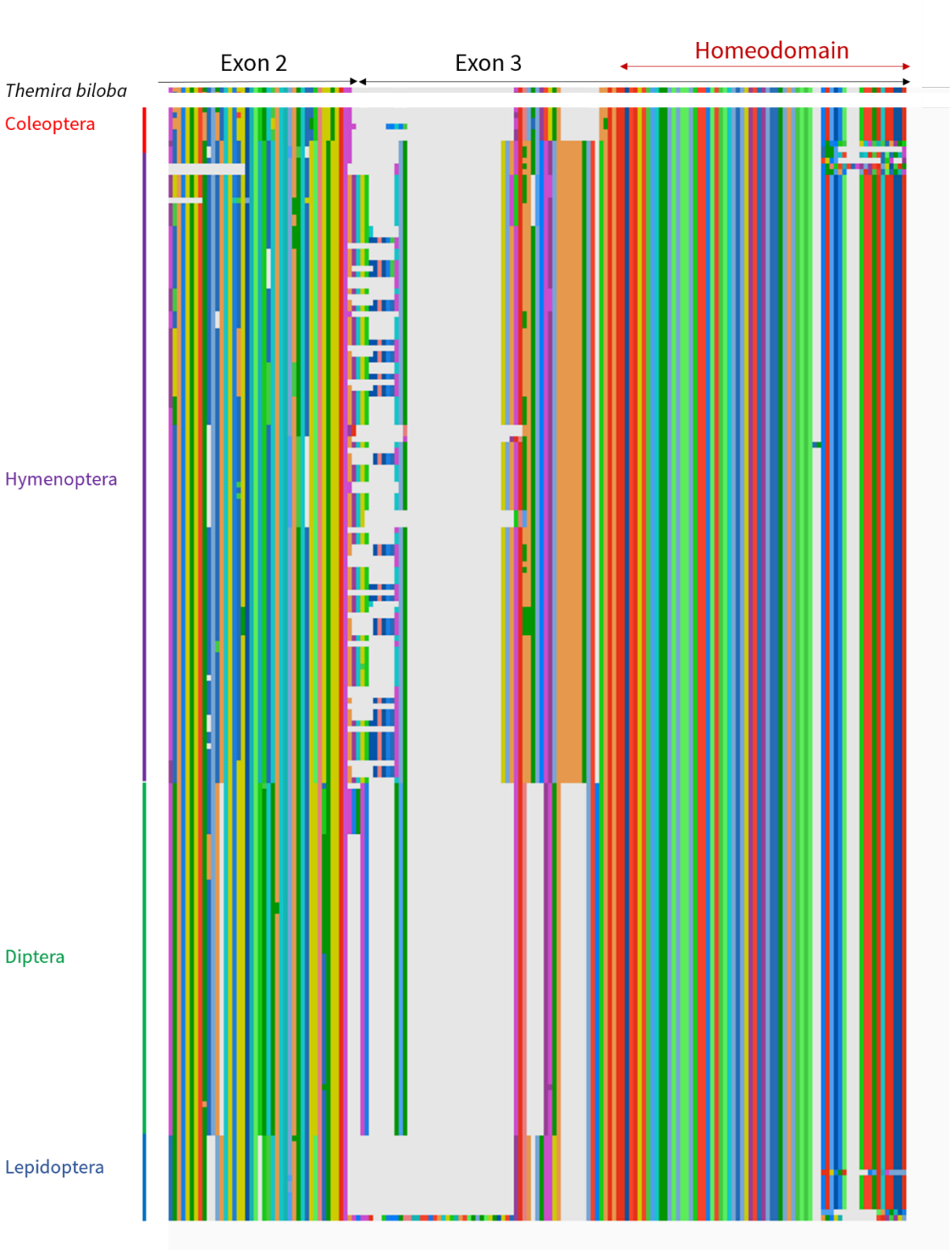
Amplification of Dll for isoform sequencing	DLL 5'UTR Pacbio F	/5Phos/CAC GCT TAA TTT CAC AGA TTT AGG GAG CCC C
Amplification of Dll for isoform sequencing	DLL exon 7 Pacbio R1	/5Phos/tca gac gat gcg tca tAG GTG GAG GTG GTA ATT GCG GCG AAT GG
Amplification of Dll for isoform sequencing	DLL exon 7 Pacbio R2	/5Phos/cta tac atg act ctg cAG GTG GAG GTG GTA ATT GCG GCG AAT GG
Amplification of Dll for isoform sequencing	DLL exon 7 Pacbio R3	/5Phos/tac tag agt agc act cAG GTG GAG GTG GTA ATT GCG GCG AAT GG
Amplification of Dll for isoform sequencing	DLL exon 7 Pacbio R4	/5Phos/tgt gta tca gta cat gAG GTG GAG GTG GTA ATT GCG GCG AAT GG
Generating exon-skipped template for protein in vitro synthesis	Tbil Dll start codon F	GCGAATTAATACGACTCACTATAGG GCTTAAGTATAAGGAGGAAAAAAT ATGGATGCGCCCGATGCACCGCATA C
Generating exon-skipped template for protein in vitro synthesis	Tbil Dll stop codon R	AAACCCCTCCGTTTAGAGAGGGGTT ATGCTAGTTATTAGTGGTGGTGGTG GTGGTGAACTGCCGGCCACACCGTT AATAGCGATGG

Supplementary Table 5. Sanger sequencing results for the RT-PCR to determine *Distal-less* expression in developing histoblast clusters. See separate excel document. Related to Figure 7.

Supplementary Table 6. Target region sequences for single guide RNA for CRISPR/Cas9 genome editing. Related to Figure 2, 3, 4 and 7.

sgRNA	Target sequence (5' → 3')
sgRNA-2A	GGTTACGGTGGCATTCGCAGCGG
sgRNA-2B	GGAATAGGAATTCTGATGCATGG
sgRNA-3	GGTTTCAACGCACCCAGTATTTGG

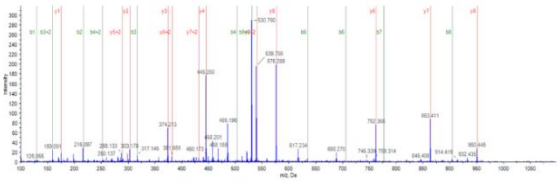
Supplementary Figure 1. Deletions in *Dll* Exon 2 produce mosaic mutant (E2B M8) with loss of the 4th sternite brush. Related to Figure 4.



Supplementary Figure 2. Alignment of *Distal-less* protein sequences across holometabola. Alignment in fasta format is provided in the Mendeley resource. Related to Figure 4.

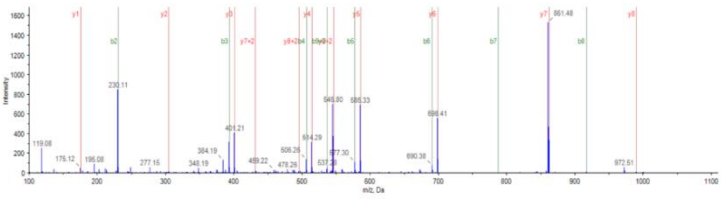
a

Residue	b	b+2	y	y+2
Q	129.0659	66.0366	1078.5098	539.7586
S	216.0979	108.5626	950.4513	475.7293
T	317.1456	159.0764	863.4192	432.2133
W	503.2249	252.1161	762.3716	381.6894
M	634.2654	317.6363	576.2922	288.6498
A	705.3025	353.1549	445.2518	223.1296
A	776.3396	388.6734	374.2146	187.6110
Q	904.3982	452.7027	303.1775	152.0924
R	1060.4993	530.7533	175.1190	88.0631



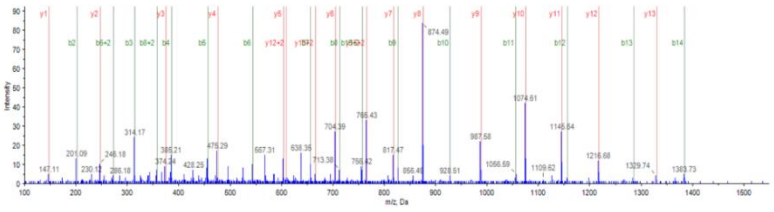
b

Residue	b	b+2	y	y+2
T	102.0550	51.5311	1090.5891	545.7982
Q	230.1135	115.5604	989.5415	495.2744
Y	393.1769	197.0921	861.4829	431.2451
L	506.2609	253.6341	698.4196	349.7134
A	577.2980	289.1527	585.3355	293.1714
L	690.3821	345.6947	514.2984	257.6528
P	787.4349	394.2211	401.2143	201.1108
E	916.4775	458.7424	304.1615	152.5844
R	1072.5786	536.7929	175.1190	88.0631



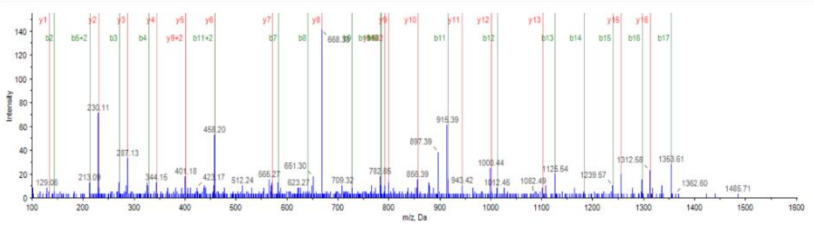
c

Residue	b	b+2	y	y+2
A	72.0444	36.5258	1529.8633	765.4363
E	201.0870	101.0471	1458.8162	729.9118
L	314.1710	157.5892	1329.7736	665.3905
A	385.2082	193.1077	1216.6896	608.8484
A	456.2453	228.6263	1145.6525	573.3299
S	543.2773	272.1423	1074.6154	537.8113
L	656.3614	328.6843	987.5833	494.2953
G	713.3828	357.1951	874.4993	437.7533
L	826.4669	413.7371	817.4778	409.2426
T	927.5146	464.2609	704.3937	352.7005
Q	1055.5732	528.2902	603.3461	302.1767
T	1156.6208	578.8141	475.2875	238.1474
Q	1284.6794	642.8433	374.2398	187.6236
V	1383.7478	692.3775	246.1812	123.6942
K	1511.8428	756.4250	147.1128	74.0600



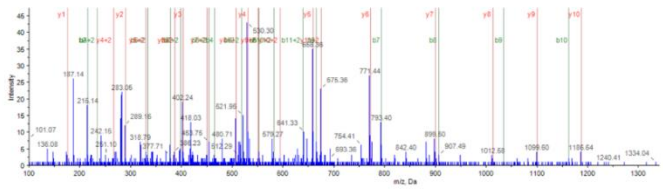
d

Residue	b	b+2	y	y+2
A	72.0444	36.5258	1582.7278	791.8676
A	143.0815	72.0444	1511.6907	756.3490
Q	271.1401	136.0737	1440.6536	720.8304
G	328.1615	164.5844	1312.5950	656.8012
P	425.2143	213.1108	1255.5736	628.2904
G	482.2368	241.6215	1158.5208	579.7640
T	583.2835	292.1454	1101.4993	551.2533
G	640.3049	320.6561	1000.4517	500.7296
S	727.3369	364.1721	943.4302	472.2187
G	784.3584	392.6828	856.3982	428.7027
M	915.3989	458.2031	799.3767	400.1920
P	1012.4517	506.7295	668.3362	334.6717
L	1125.5357	563.2715	571.2835	286.1454
G	1182.5572	591.7822	458.1994	229.6033
G	1239.5786	620.2930	401.1779	201.0926
G	1296.6001	648.8037	344.1565	172.5819
G	1353.6216	677.3144	287.1350	144.0711
P	1450.6743	725.8408	230.1135	115.5604
N	1564.7173	782.8623	133.0608	67.0340

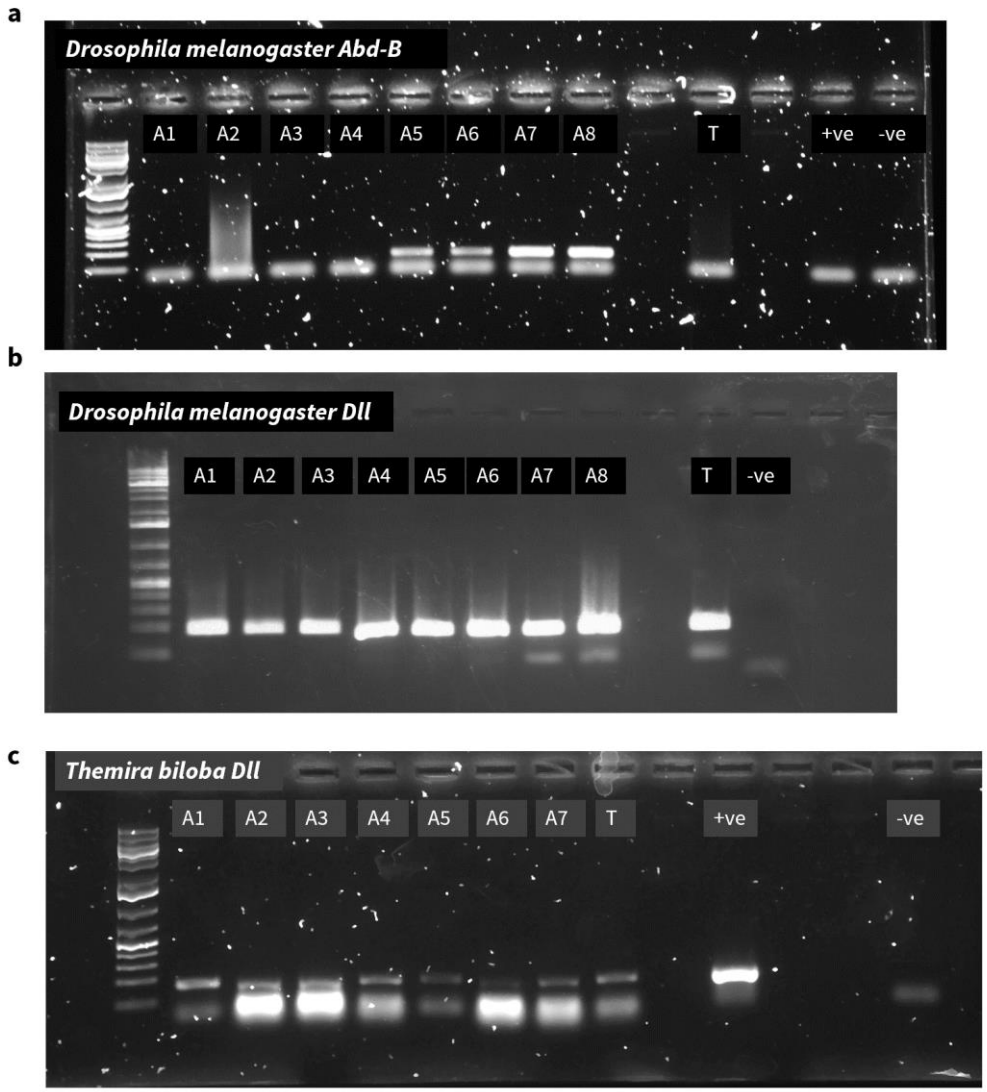


e

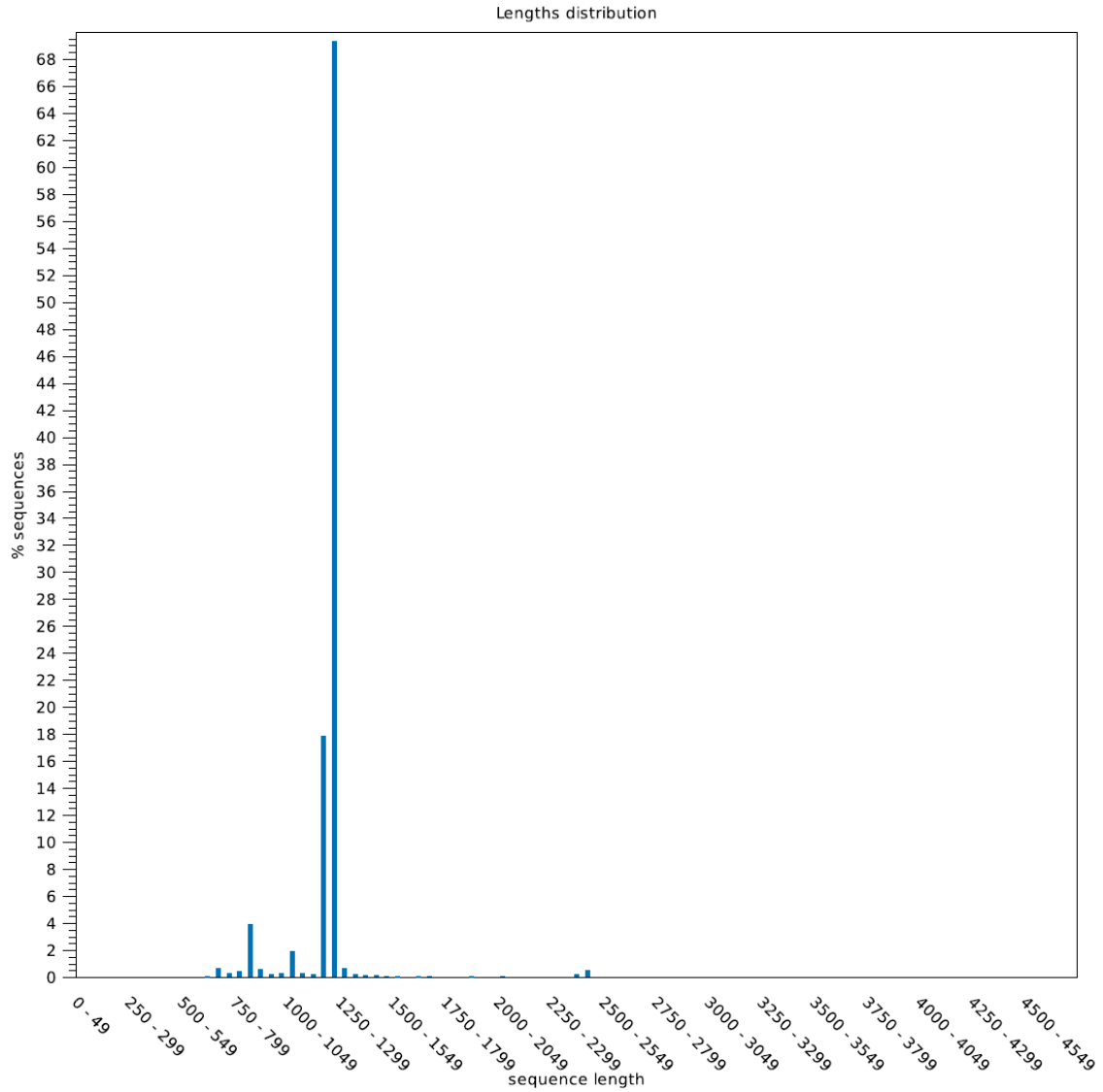
Residue	b	b+2	y	y+2
T	102.0550	51.5311	1663.8489	782.4281
I	215.1390	108.0731	1462.8013	731.9043
Y	378.2023	189.6048	1349.7172	675.3622
S	465.2344	233.1208	1186.6539	593.8306
S	552.2664	276.6368	1099.6218	550.3146
L	665.3505	333.1789	1012.5898	506.7985
Q	793.4090	397.2082	899.5057	450.2565
L	906.4931	453.7502	771.4472	386.2272
Q	1034.5517	517.7795	658.3631	329.6852
Q	1162.6103	581.8088	530.3045	265.6559
L	1276.6943	638.3508	402.2459	201.6266
N	1389.7373	695.3723	289.1619	146.0846
R	1545.8384	773.4228	175.1190	88.0631



Supplementary Figure 3. Fragmentation evidence for *Dll* peptides recovered from ProteinPilot. (a) Fragment ion mass values and peptide peak intensity for 'QSTWMAAQR', a peptide matching to an alternative initiation codon (b), (c), (d) and (e) Fragment ion mass values and peptide peak intensity for peptides matching to *Dll* homeodomain. Related to Figure 6.



Supplementary Figure 4. RT-PCR results to determine *Dll* expression in late larval epidermal tissues. *Drosophila melanogaster* was used as a control. Primers for this experiment are found in Supplementary Table 5. Identity of amplified products were confirmed with sanger sequencing (results in Supplementary Table 6). A1-A8: Tissue from the respective abdominal epidermal segments. T: Tissue from the thoracic epidermal segment. (a) In the control, *D.melanogaster*, Abd-B is present and amplified in the 5^{th} – 8^{th} abdominal segments as expected. (b) *Dll* is present and amplified in all 8 abdominal segments in *D.melanogaster* (c) *Dll* is present and amplified in all 7 abdominal segments in *Themira biloba.* Related to Figure 7.



Supplementary Figure 5. Read length distribution of Pacbio Isoseq error-corrected reads (10 minimum passes with a minimum predicted accuracy of 0.9) using the PacBio SMRT analysis software (v5.1.0). The analysis generated 225,740 CCS reads with a median CCS read length between 1200 – 1249 bp. Related to Figure 8.

361

362

363

364

365

366

367