A genome-wide survey reveals a diverse array of enhancers coordinates the *Drosophila* innate immune response

Lianne B. Cohen[1, 2], Tamara Hadzic[2, 3], Caitlin Sauer[1, 2], Julia R. Gibbs[1, 2], Zeba Wunderlich[1, 2, 3, 4*]

1. Department of Biology, Boston University, Boston, MA 02215, USA
2. Biological Design Center, Boston University, Boston, MA 02215, USA
3. Program in Bioinformatics, Boston University, Boston, MA 02215, USA
4. Department of Biomedical Engineering, Boston University, Boston MA 02215, USA

*Corresponding author: zeba@bu.edu

Supplemental Methods

Genome-wide reporter library construction

To generate a genome-wide reporter library, genomic DNA was isolated from both male and female *iso-1* flies, fragmented via sonication, and size selected (500-750 bps). A solution of 100 ng/µL DNA was sonicated on ice with a Sonic Dismembrator sonicator (Thermo Fisher Scientific) using an 8-tip probe for 10 cycles of 15 sec on/59 sec off at 25% Amp. Fragments of size 500-750 bp were extracted from a gel. The genomic library preparation and STARR-seq protocol was done as described in Neumayr, *et al.* and outlined briefly below. The NEBNext End Prep kit (New England Biolabs) was used to process genomic fragments before PCR amplification with 8 cycles. These genomic fragments were then cloned into the pSTARR-seq fly plasmid (AddGene, #71499) using Gibson cloning.

The genomic library was transformed into MegaX DH10B *E. coli* (Thermo Fisher Scientific) by electroporation (exponential decay, 2kv, 25 µF, 200 ohms). The culture was grown overnight shaking at 37°C in LB with ampicillin. Cells were pelleted at 3,000×g for 30 mins. The plasmid library was extracted with the ZymoPURE II plasmid Maxi prep kit (Zymo Research).

STARR-seq in S2* cells

To test the genomic library's enhancer activity, we performed STARR-seq in S2* cells. S2* cells were a gift from Steven Wasserman and grown according to standard methods in complete Schneider's media at 28°C. For three condition experiment, 2.8×10^9 S2* cells were washed with electroporation buffer (MaxCyte) before resuspension in electroporation buffer to a total volume of 6 mL with input STARR-seq library (300 µg). Cells were electroporated in a R1000 processing assembly (MaxCyte) in the Maxcyte ATX using the S2 protocol. After electroporation, cells were added to 150 µL of basal serum-free Schneider's media with a final concentration of 0.0256 U/µLof DNaseI in the center of a tissue culture dish. Cells were incubated with DNaseI for 30 mins at 28°C before complete Schneider's media [Schneider's *Drosophila* Medium (Fisher), 10% heat inactivated FBS (Fisher), 2 nM L-glutamine, 50U/mL Pen/Strep (Life Technologies.), 1% Fungizone (Life Technologies)] was added to stop the enzymatic activity.

Cells were split into three treatment groups: Control, 20E and IMD, treated with either water (Control) or a final concentration of 40 nM 20E (Sigma-Aldrich) (20E and IMD) for 24 hrs, and then PBS (Control and 20E) or heat-killed *S. marcescens* at a final concentration of 0.4 OD (IMD) for 24 hrs. Total RNA was isolated with RNAeasy prep kits (Qiagen) from which mRNA

was purified using oligo-dT Dynabeads before digestion with DNaseI. Reverse transcription, followed by junction PCR and barcoding PCR were completed as described in Neumayr et al. 2019. Paired end Illumina sequencing was performed on each sample of 3 biological replicates and the input library. 75,000,000-122,000,000 reads were collected for each sample.

Alongside the STARR-seq replicates, RNA-seq samples were also created in triplicate. Untransfected S2* cells were treated in the same manner to create the same three treatment groups: Control, 20E and IMD. Total RNA was sent to Genewiz for poly(A) mRNA selection and library preparation before sequencing. At least 36,000,000 reads were collected for each sample.

Computational Analysis

Genomic bins
To determine in which genomic regions the STARR-seq enhancers are, we compared the genomic coordinates of each enhancer to the annotated FlyBase FASTA files for exons, introns, 5' UTR, 3' UTR and intergenic regions. Since the *D. melanogaster* genome is very dense in genes, many enhancers occupied multiple genomic regions. To limit the region assignment to one region, we first chose the region with the highest overlap with the enhancer. If multiple regions have identical overlap lengths, we then choose the region based on the following hierarchy: 1. intergenic regions, 2. introns, 3. 5' UTRs, 4. 3' UTRs, 5. Exons

RNA-seq analysis
RNA-seq reads were aligned with Bowtie 2 and counts were generated with Subread's featureCounts (Langmead and Salzberg 2012, Liao et al. 2014). Transcripts per million (TPM) were then calculated. To determine agreement between the triplicates and identify any outliers, principal component analysis (PCA) was completed. Sample 1, a control sample, was determined to be an outlier and removed, leaving 2 control samples, 3 20E samples, and 3 IMD samples.

Logistic regression
Logistic regression models to determine the TFBS that contribute to activity classifications were made using scikit-learn (Pedregosa et al. 2011) with 1000 maximum iterations, a test size of 0.2 and balanced class weights.

Gene assignments
Gene centric enhancer assignments (used in Fig. 1) were made by scanning 15 kb upstream and 5 kb downstream of a gene and assigning any enhancer in this window to the gene, as in Arnold, et al. Since enhancers function independent of orientation, enhancer-centric assignments (used in Fig. 3) were made by scanning 15 kb upstream to 15 kb downstream of an enhancer and assigning any genes in this window to that enhancer.

More restrictive enhancer-gene assignments were made by sorting genes into activity classes based on expression in the RNA-seq data (expression threshold: TPM > 1). The criteria for the activity classes are as follows: Control Only (gene expressed in 2/2 Control samples), 20E Only (2/3 20E samples), IMD Only (2/3 IMD samples), Control+20E (4/5 Control and 20E samples), Control+ IMD (4/5 Control and IMD samples) 20E+ IMD (5/6 20E and IMD samples), and Constitutive (7/8 all samples). For each activity class, genes were matched to the allowed set of enhancers by scanning 15 kb upstream and 5 kb downstream of the gene.

Enrichment of TFBS in pathways, time, immune role, and functional categories was determined by calculating the odds ratio of enhancers within the category having a particular motif compared to the whole IMD set. Statistical significance was determined with Fisher's exact test.

Construction and flow cytometry of enhancer reporters

STARR-seq consensus enhancers were amplified from *iso-1* flies and cloned upstream of the Drosophila Synthetic Core Promoter (DSCP) (Pfeiffer et al. 2008; Arnold et al. 2013) driving EGFP into the section of the pAc5.1/V5-His-A (Invitrogen) plasmid containing the origin of replication and ampicillin resistance gene via Gibson cloning. See Supplementary Table 1 for regions cloned. Plasmids were designed using Benchling (www.benchling.com). TSS's contained within cloned enhancers were mutated by site-directed mutagenesis. An additional plasmid, pKM7 (p-IEX-mCherry), (Russell et al. 2021) was used as a positive transfection control. Plasmids were prepared with the ZymoPURE II plasmid Midiprep kit (Zymo Research) and concentrated via ethanol precipitation.

Plasmids were electroporated using the Maxcyte ATX machine following the S2 cell protocol. To allow for either Toll or Imd stimulation, S2* cells stably transfected with pMT-Torso-pelle (Sun et al. 2002) were used for this experiment. Pelle, a key component of Toll signal transduction, is fused to the transmembrane domain of Torso and placed under control of the metallothionein promoter, which is activated upon the addition of $CuSO_4$. Ten million cells per sample were washed in 2 mL of electroporation buffer (MaxCyte), then concentrated into 100 µL of electroporation buffer. After washing and concentrating, cells were mixed with 5 µg of enhancer reporter DNA, and 5 µg of the mCherry transfection control and loaded into a well of a OC-100×2 processing assembly (MaxCyte). Directly following electroporation, transfected cells incubated at 28°C for 30 minutes in a droplet of 50 µL of basal Schneider's *Drosophila* Medium and 0.5 µLof 1000 U/mL DNaseI. 5 mL of Schneider's Complete Media were added to the cells in the T25 cell culture plate post incubation. Cells were then split across into five wells of a 6-well plate, for each of five treatment conditions: Control, 20E, IMD-induced, Toll-induced, or Toll- and IMD- (dual) induced.

Cells were then treated with either water (Control) or 40 nM 20E for 24 hrs (20E, IMD-, Toll- and dual-induced. Cells were treated with either 100 µL of PBS (Control and 20E conditions), 100 µL of OD=10 Heat-killed *Serritia marcescen*s (HKSM) to induce IMD, or a final concentration of 500 µM $CuSO_4$ to induce Toll, or both HKSM and $CuSO_4$ for dual induction.

Following a 24-hour induction period, cells were centrifuged for 5 minutes at 100×g to pellet the cells. Each pellet was resuspended in 1 mL of FACS buffer (1× Ca/Mg++ free PBS, 1mM EDTA, 25mM HEPES pH 7.0, 1% FBS). After sample preparation, cells were analyzed using either a FACSCalibur or CytoFLEX flow cytometer. Fluorescence signals from GFP and mCherry reporter constructs were collected using appropriate filter channels. Instrument settings and gating thresholds were established using non-fluorescent and untransfected control samples. After gating for live, transfected cells, the geometric mean of GFP signal in the treatment conditions (20E, IMD, Toll and Dual) was normalized to the geometric mean of GFP signal in the Control for each transfected plasmid, which is the fold change in reporter activity. To estimate the absolute reporter activity across different cytometers, we normalized the geometric mean for GFP signal in each condition to the geometric mean of GFP signal for untransfected cells. At least 10,000 cells were collected for each sample. Each plasmid was tested across three biological replicates.

ATAC-seq with adult fly hemocytes

We injected 60-70 2-7 day old male OregonR flies with ~50 nL of heat-killed *Serratia marcescens* at OD=0.5 in PBS and incubated them at 25°C for 24 hrs, alongside uninjected age matched controls. Hemolymph including hemocytes were collected by placing 20-25 anesthetized flies in 0.5 mL tubes with 3 20G needle holes at the bottom. Flies were spun at 6,000×g at 4°C for 1 min into ATAC-seq lysis buffer (Grandi et al. 2022).

Tagmentation and library prep protocols were adapted from Grandi et al. 2022*,* as follows. After centrifugation, hemolymph was combined within the treatment samples (injected and uninjected) respectively and mixed by pipetting 5×, followed by 3 min incubation on ice. Samples were washed with 1 mL of wash buffer before centrifugation at 500×g for 10 mins at 4°C. Supernatant was removed and ATAC-seq Tn5 mix (1× TD buffer, 0.33× PBS, 0.01% digitonin, 0.1% Tween-20, 0.05 µL of Tn5/µL of mix) was added to the pellet. Samples were placed in a 37°C shaking incubator for 30 mins, with brief vortexing half way through. The Tn5 reaction was stopped with the addition of DNA binding buffer from the Monarch PCR clean up kit (New England Biolabs) and samples were frozen before the rest of the cleanup kit was completed. Barcoding PCR was completed with Illumina primers and qPCR quantification with NEBNext Library Quant Kit (New England Biolabs). Paired end Illumina sequencing was completed on three biological replicates per condition and at least 100 million reads were collected per sample.

ATAC-seq analysis

ATAC-seq reads were trimmed using cutadapt (Martin 2011). Reads were aligned to the genome using Bowtie 2 (Langmead and Salzberg 2012) and filtered for duplicate reads using Picard MarkDuplicates. Reads were adjusted for Tn5 offset using bamtools, and peaks were called using MACS2 (Zhang et al. 2008), using narrow peak calling and a p-value of 0.01. FrIP scores were calculated using BEDTools intersect (Quinlan and Hall 2010). Consensus ATAC-seq peaks were built from the widest margins of peaks present in at least 2 replicates. STARR-seq enhancers were labeled open if >50 bps of enhancer were in consensus ATAC-seq peaks.