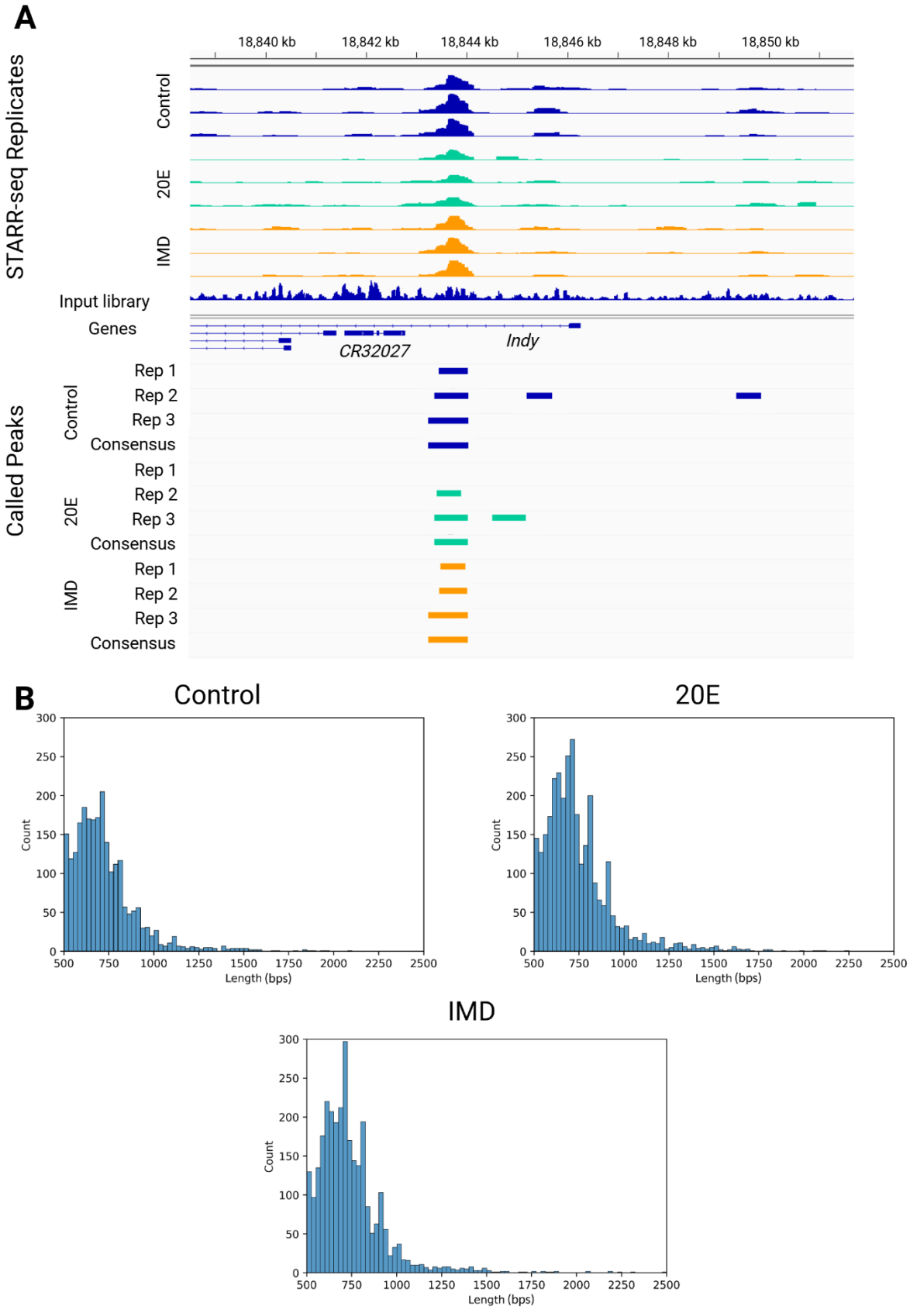
A genome-wide survey reveals a diverse array of enhancers coordinates the *Drosophila* innate immune response

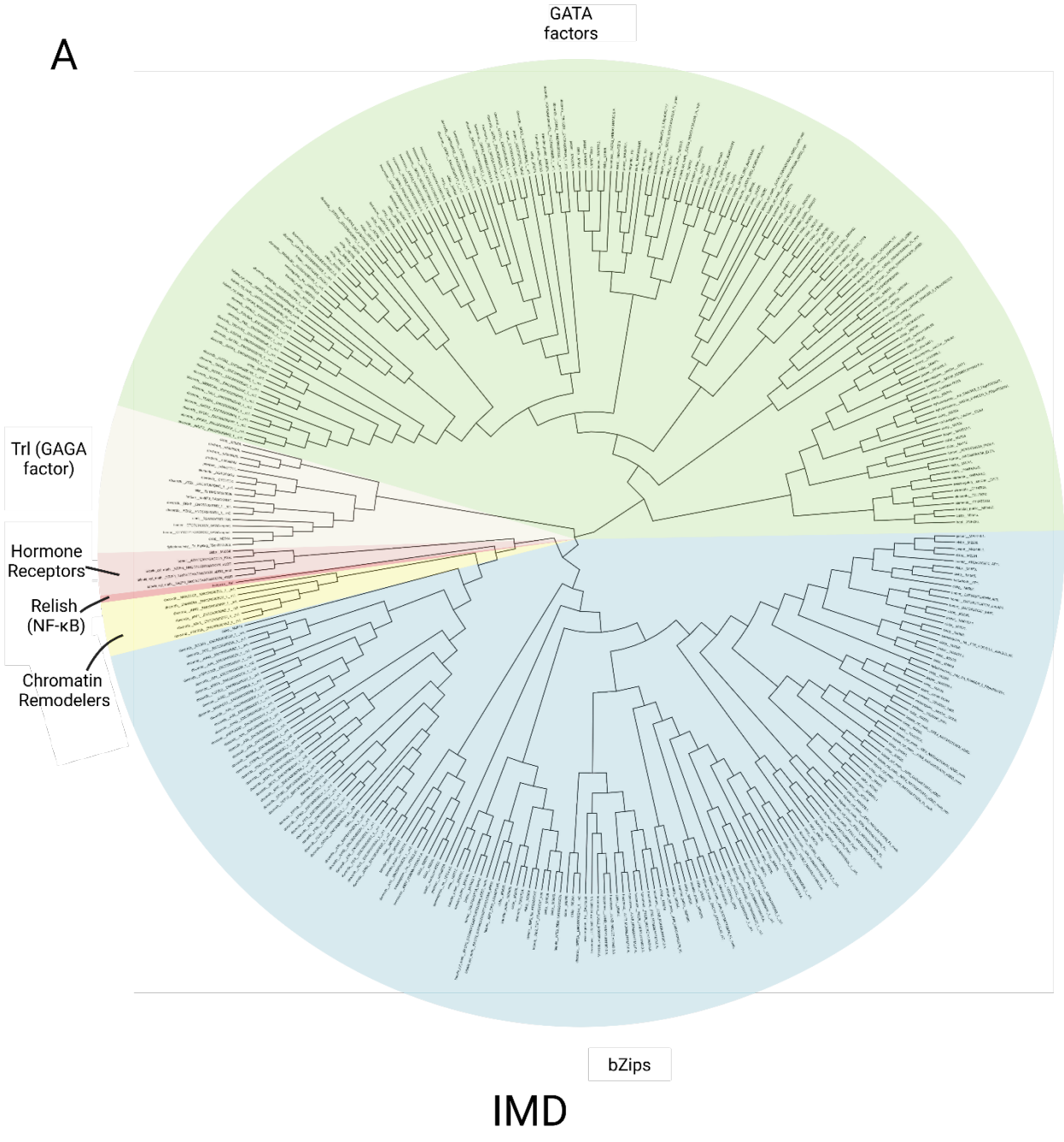
Lianne B. Cohen[1, 2], Tamara Hadzic[2, 3], Caitlin Sauer[1, 2], Julia R. Gibbs[1, 2], Zeba Wunderlich[1, 2, 3, 4]*

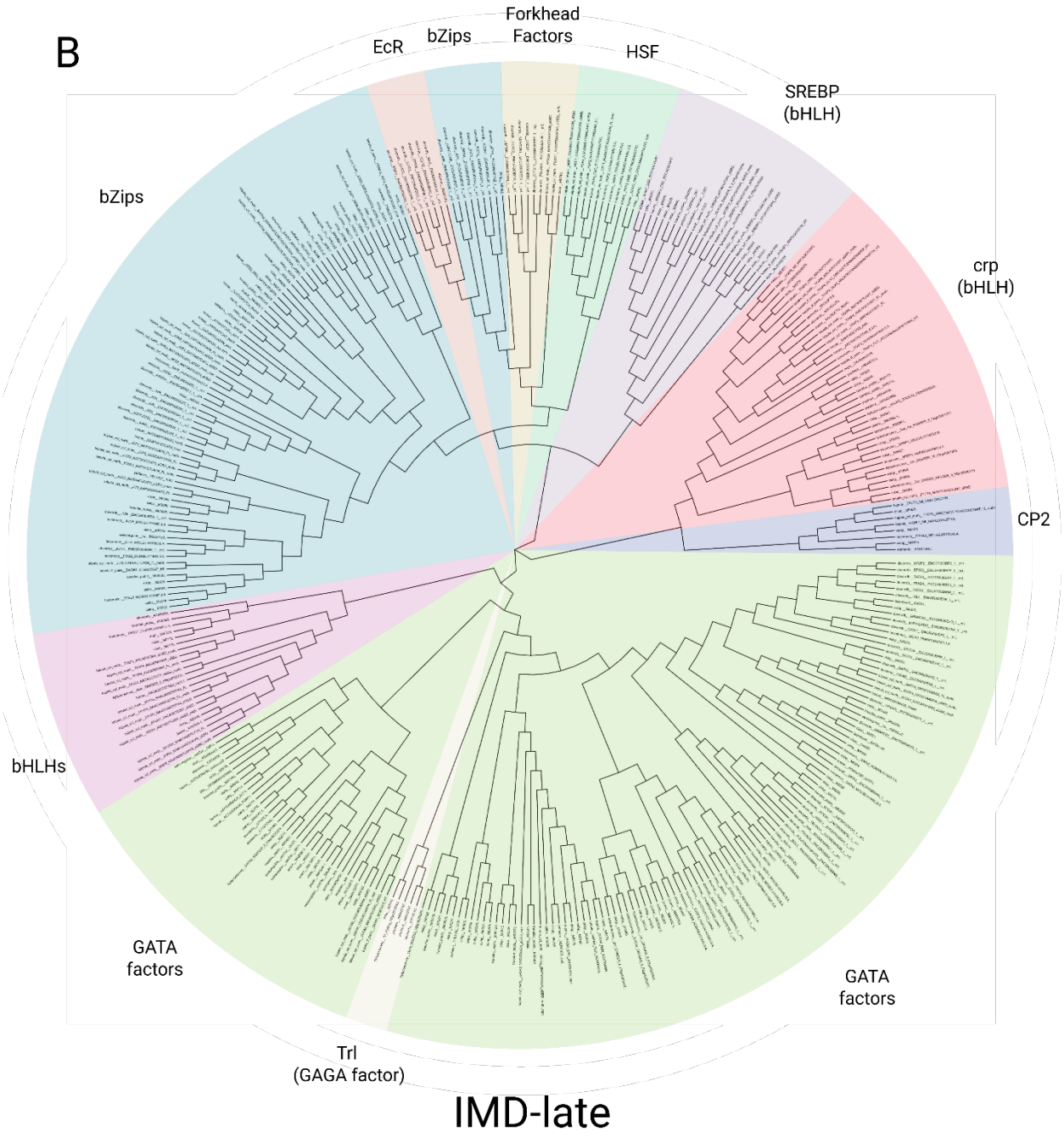
1. Department of Biology, Boston University, Boston, MA 02215, USA
2. Biological Design Center, Boston University, Boston, MA 02215, USA
3. Program in Bioinformatics, Boston University, Boston, MA 02215, USA
4. Department of Biomedical Engineering, Boston University, Boston MA 02215, USA

*Corresponding author: zeba@bu.edu

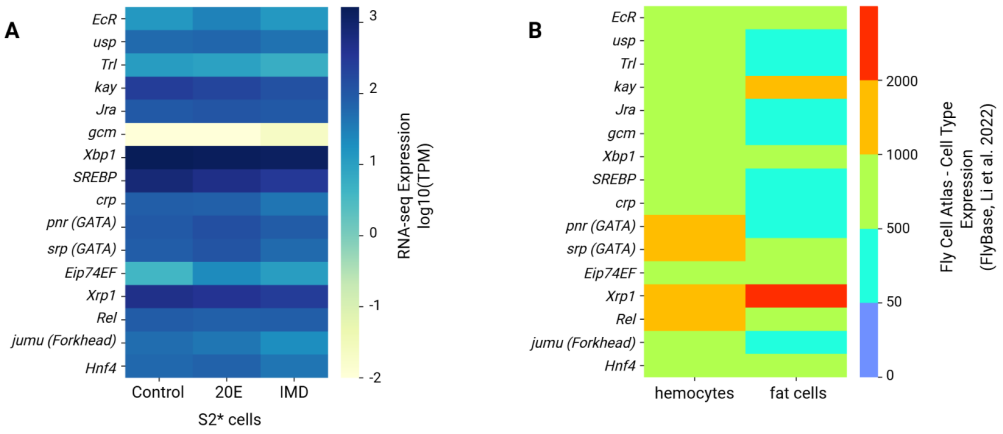

A
STARR-seq Replicates
18,840 kb
18,842 kb
18,844 kb
18,846 kb
18,848 kb
18,850 kb
Control
20E
IMD
Input library
Genes
CR32027
Indy
Called Peaks
Control
Rep 1
Rep 2
Rep 3
Consensus
20E
Rep 1
Rep 2
Rep 3
Consensus
IMD
Rep 1
Rep 2
Rep 3
Consensus
B
Control
20E
IMD
Count
Length (bps)

Supplemental Figure S1: STARR-seq enhancers have similar lengths across treatments. (A) IGV screenshot of tracks for STARR-seq replicates (output), input library, and references genes (Robinson et al. 2011). STARRPeaker called peaks are shown for each replicate in addition to the consensus peaks for each condition. Control tracks are shown in blue, 20E tracks are shown in green and IMD peaks are shown in orange. (B) Histogram of lengths for STARR-seq consensus enhancers.

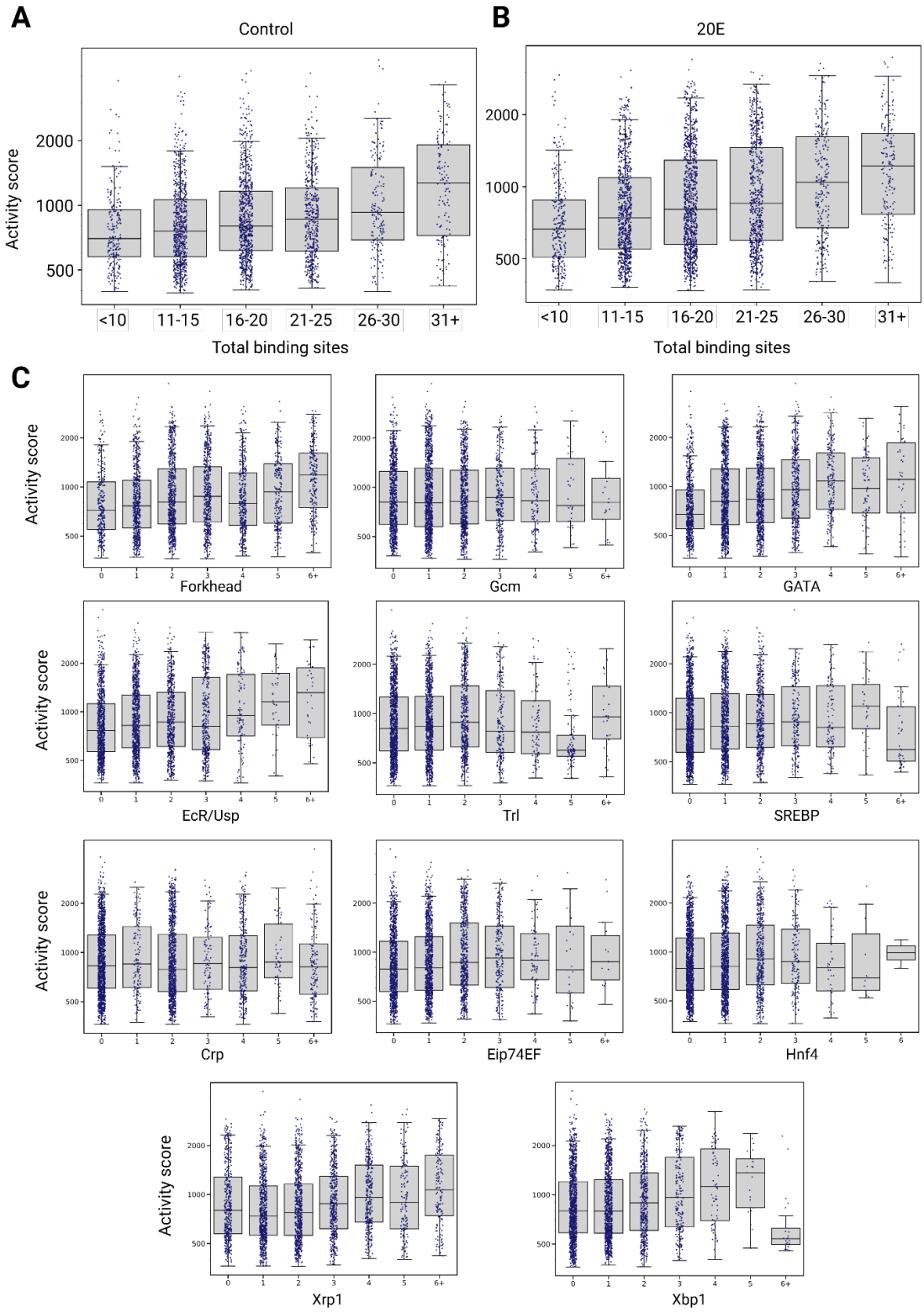




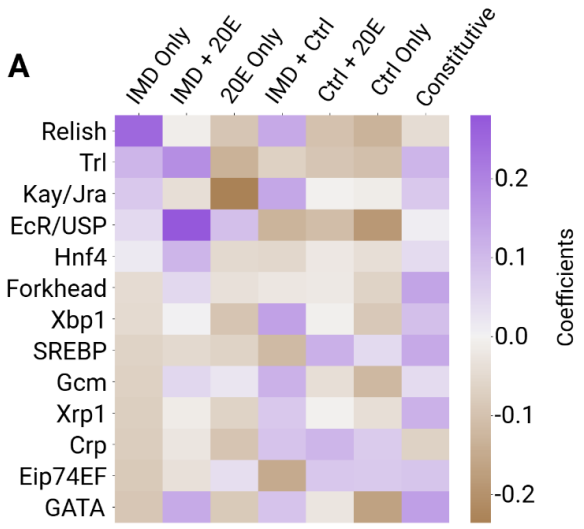
Supplemental Figure S2: Several TF families are enriched in IMD enhancers. (A) A motif tree of 360 motifs from i-cisTarget analysis enriched three-fold in all IMD enhancers compared to the genome. TF families are indicated with colored sections. (B) A motif tree of 363 motifs from i-cisTarget analysis enriched three-fold in IMD enhancers assigned to late acting genes.

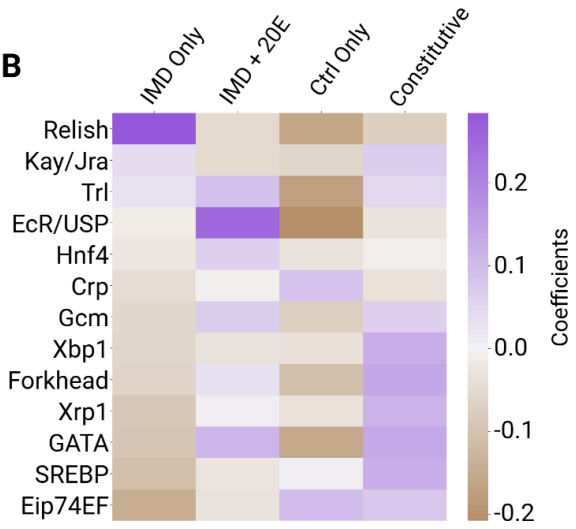


Supplemental Figure S3: Transcription factor expression in S2* cells and immune tissues. (A) Log10(TPM) of expression of transcription factors in S2* cells upon IMD treatment, measured by RNA-seq. Transcription factors for all motifs used in study are expressed in S2* cells except Gcm. Gcm was included because it is expressed in hemocytes and regulates their development. (B) Expression of transcription factors in hemocytes and fat cells from Fly Cell Atlas (Li et al. 2022) as reported in FlyBase: very low expression (0-50), low expression (50-500), moderate expression (500-1000), high expression (1000-2000), very high expression (2000+).



Supplemental Figure S4: Activity score varies by number of TFBS. A. Box and whiskers plot of STARR-seq activity score by total binding sites from 13 TFs in Control enhancers (A) and 20E (B) enhancers. (C) Activity scores for IMD enhancers by number of binding sites of individual TFs.





C

Predicted Classifications

True Classifications

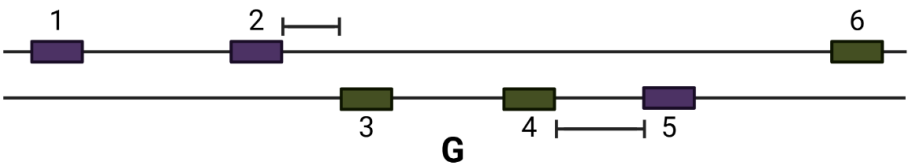
	IMD Only	IMD + 20E	20E Only	IMD + Ctrl	Ctrl + 20E	Ctrl Only	Consti-tutive
IMD Only	**27**	7	15	15	3	10	3
IMD + 20E	21	**57**	34	28	17	20	28
20E Only	14	8	**27**	7	4	15	9
IMD + Ctrl	3	0	1	**3**	3	5	3
Ctrl + 20E	6	5	6	6	**2**	8	6
Ctrl Only	17	10	29	23	10	**39**	11
Constitutive	24	50	42	46	18	31	**78**

D

Predicted Classifications

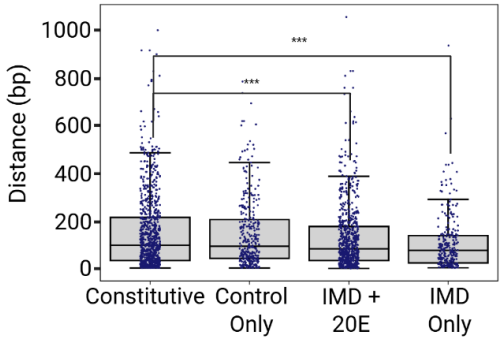
True Classifications

	IMD Only	IMD + 20E	Ctrl Only	Consti-tutive
IMD Only	**6**	23	18	21
IMD + 20E	1	**97**	23	95
Ctrl Only	3	31	**54**	57
Constitutive	3	72	46	**164**



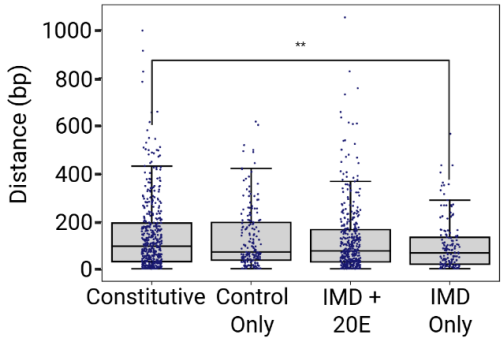
F

Median distance (bp)	100	96	85	79
# of Rel-GATA pairs	899	329	695	275
# of enhancers in class	1344	748	1106	372
average pairs per enhancer	0.669	0.440	0.628	0.739



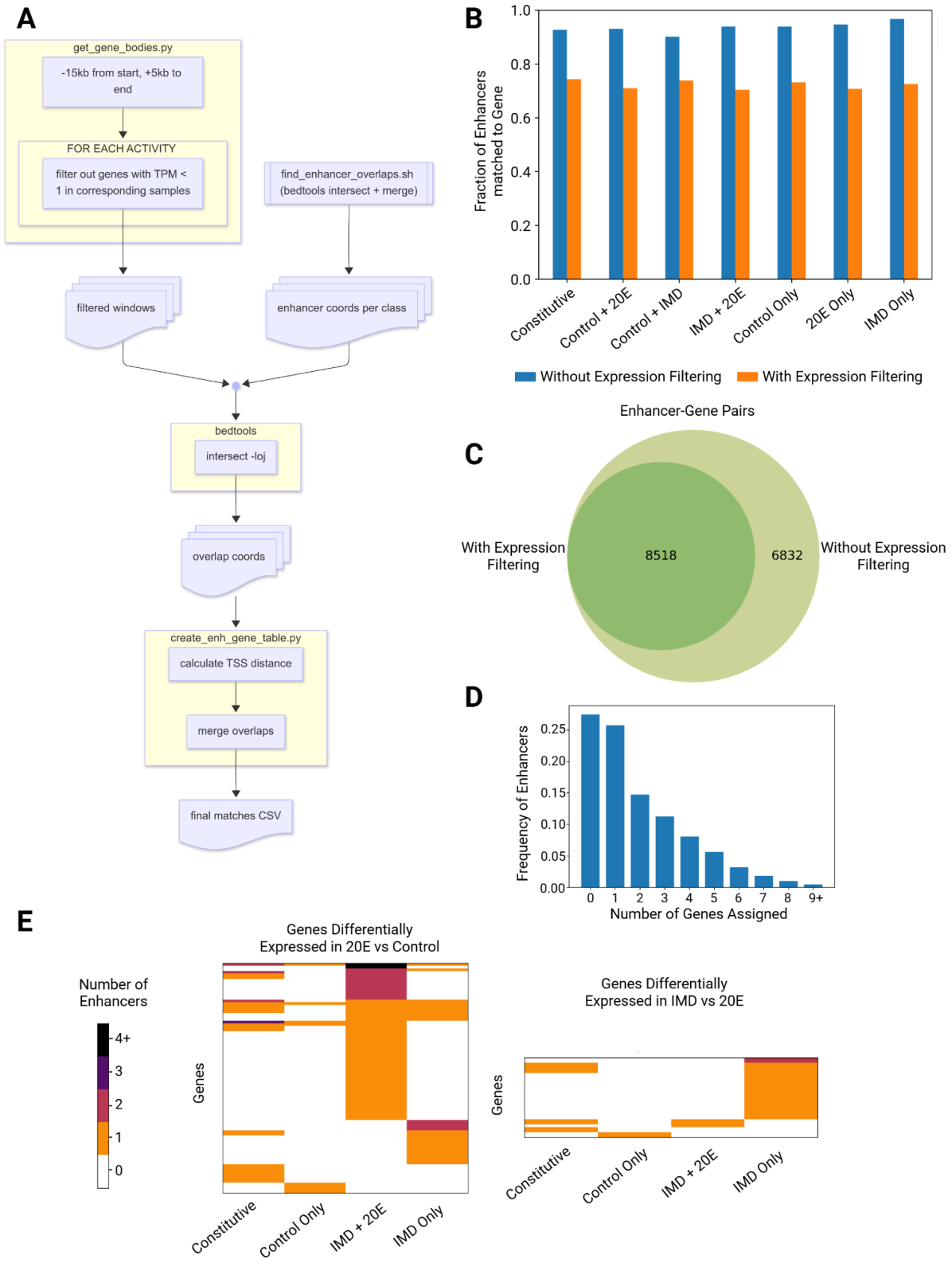
G

Median distance (bp)	99	74.5	79.5	71.5
# of Rel-GATA pairs	453	182	384	152
# of enhancers in class	1344	748	1106	372
average pairs per enhancer	0.336	0.243	0.347	0.409



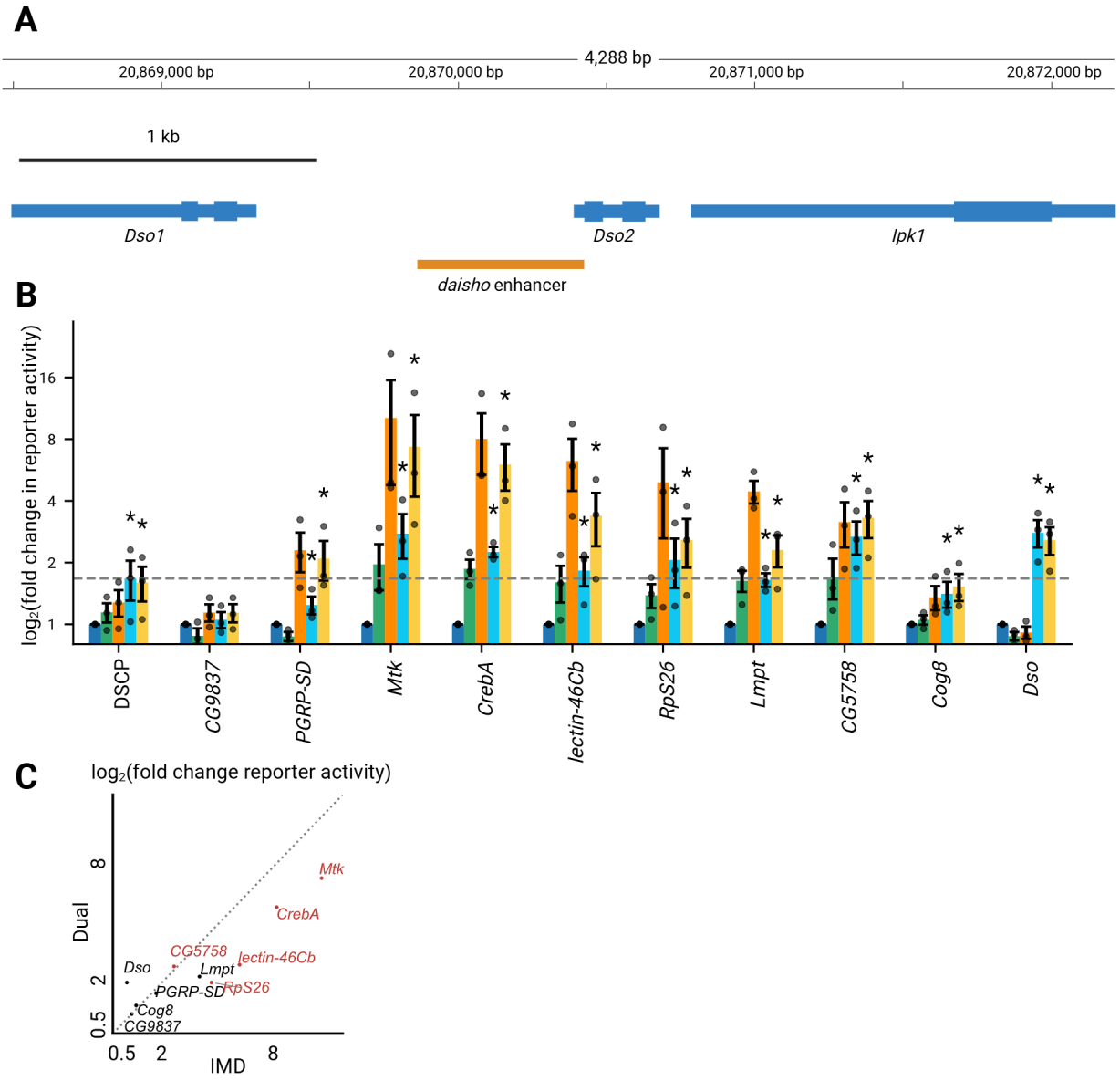
Supplemental Figure S5: Logistic Regression analysis highlights the role of some TFs in predicting activity class and analysis of Relish-GATA binding site pairs supports their correlation with immune-responsiveness. (A) Coefficients of TF motifs from logistic regression classifying enhancers by activity class. (B) Coefficients from logistic regression classifying enhancers in four most

biologically relevant activity classes. Positive coefficients are shown in purple, while negative coefficients are shown in brown. (C) Confusion matrix of logistic regression with a 20% test set size. Correctly called activity classifications are bolded and shaded in green. (D) Confusion matrix of four class logistic regression with a 20% test set size. Correctly called activity classifications are bolded and shaded in green. (E) To investigate the possible Relish-GATA binding site grammar, we identified pairs of Relish-GATA sites in enhancers. Shown is an example of Relish-GATA site pair matching in a double-stranded enhancer. Relish sites are shown in purple, GATA sites are shown in green, and sites are annotated in a strand-aware fashion to allow for the analysis of the relative orientation of Relish-GATA site pairs. First Relish sites are matched to the closest GATA site, i.e. 1 and 3, 2 and 3, 4 and 5, 5 and 6. Then, only the closest pair for each site was kept, i.e. 2 and 3, 4 and 5. Sites with the same orientation are on the same strand, and sites with opposite orientations are on different strands. (F) Distance between Rel and GATA pairs, regardless of orientation, in enhancers across activity classes. Median distance, number of Relish-GATA pairs, number of enhancers in the activity class, and the density (pairs per enhancer) are presented in the boxes above the graph for each activity class. *** indicates $p<0.001$, two-sided Mann-Whitney *U* test and (G) Distance between Rel and GATA pairs only in the same orientation (i.e. sites 4 and 5 in E) across activity classes. Median distance, number of Relish-GATA pairs, number of enhancers in the activity class, and the density (pairs per enhancer). ** indicates $p< 0.01$, two-sided Mann-Whitney *U* test.



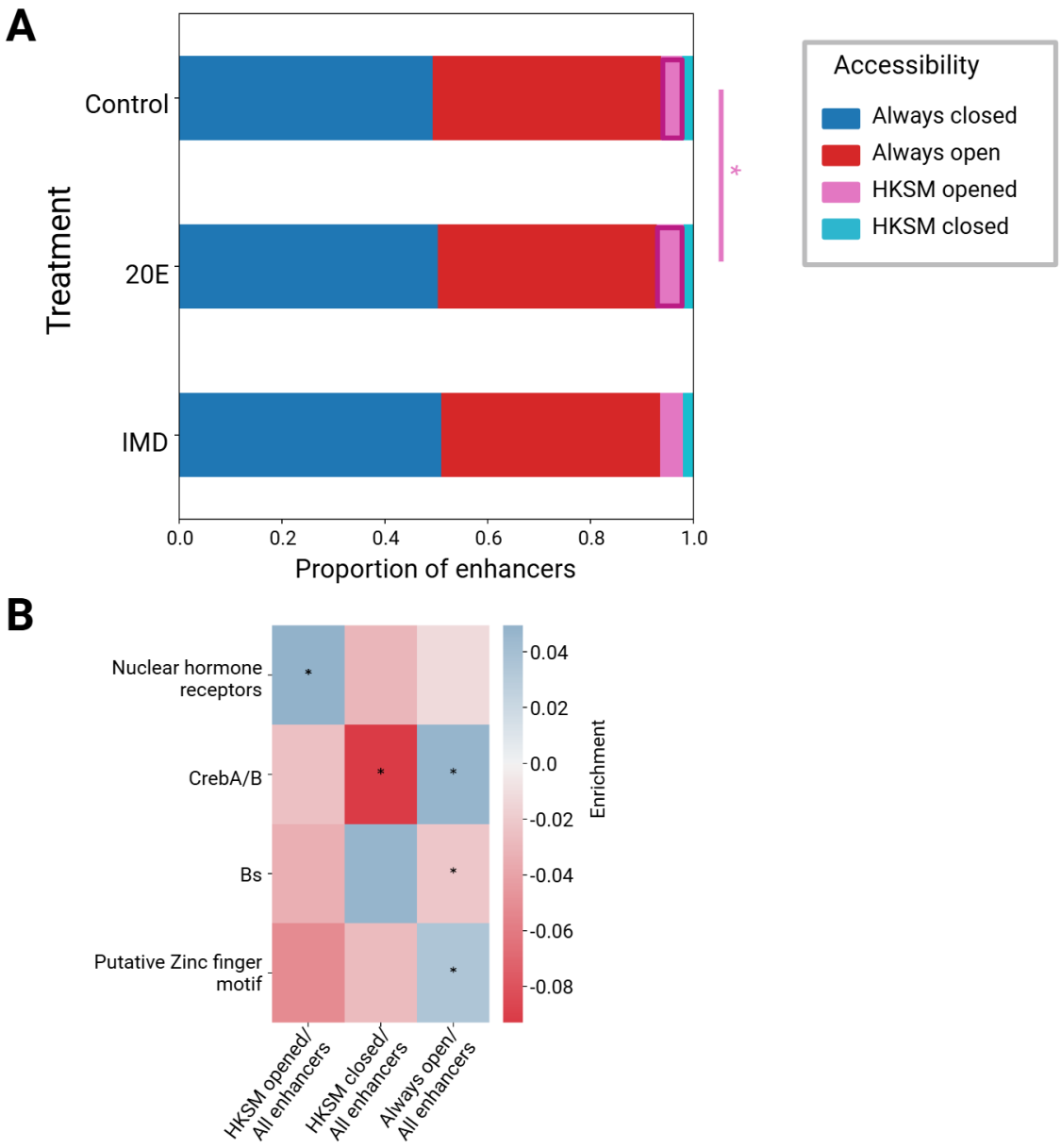
Supplemental Figure S6: Expression filtering refines enhancer-gene assignments. (A) Overview of gene-enhancer matching workflow. (B) Fraction of enhancers matched to genes with and without expression filtering for all seven activity classes. (C) Venn Diagram of the total number of enhancer-gene pairs with and without expression filtering. Expression filtering removes a little less than half of the enhancer-gene pairs. (D) Frequency of enhancers for number of putative gene targets after RNA-seq expression filtering (E) To test for the accuracy of these gene-enhancer assignments, we also identified

genes that were differentially expressed either between the Control and 20E condition, the Control and IMD condition, or the IMD and 20E condition. Notably, this is distinct from our filtering process, which only requires a gene to be expressed in a condition, not necessarily differentially expressed between conditions. In each case, we found the majority of gene-enhancer assignments for these differentially expressed genes were to enhancers in the appropriate activity class. For example, 77% of genes that show a greater than two-fold change in expression between the 20E and IMD conditions have at least one IMD Only enhancer. Enhancers assigned to genes upregulated (RNA-seq, $\log_2$ fold change > 2) in 20E vs Control (88 genes) and IMD vs 20E (31 genes) by activity class. Each row represents one gene and the number of enhancers in each activity class are indicated by the cell color.



Supplemental Figure S7: Enhancer reporters are differentially responsive to 20E, IMD, Toll and Dual Induction. (A) Genomic locus surrounding *daisho* enhancer. We chose to refer to this enhancer as *Dso*, since it lies between *Dso1* and *Dso2* and the closest gene highly expressed in the IMD condition, *Ipk1*, does not have an immune function. Genes are represented in blue, while the *daisho* STARR-seq enhancer is shown in orange. (B) The $\log_2$(fold change) of each reporter's activity, normalized to the control condition, was measured by flow cytometry. The bar shows the mean $\log_2$(fold change), the dots represent three biological replicates, and the error bars are the standard error of the mean. Each replicate measurement included at least 10,000 cells. Stars indicate a significant increase in activity between the control and starred conditions (one-sided Mann-Whitney *U* test; $p<0.05$). The dashed horizontal line is the mean fold change of the DSCP control construct in the Toll condition. The values for the 20E and IMD conditions for all reporters and the Toll and Dual conditions for the *daisho* reporter are also displayed in Figure 5, but are replicated here for ease of comparison to other conditions. (C) To test if there is a tradeoff or synergy in enhancer function upon stimulation of the Imd and Toll pathways, we compare enhancer function in the IMD and Dual induction conditions. Log_2 of GFP expression levels for reporters comparing Dual treatment to IMD treatment. Both treatment conditions are normalized to Control treatment. Dashed line is y=x, denoting an equal response to IMD and Dual treatment. Enhancers highlighted in red are both Toll and IMD responsive. We found that gene expression between IMD and

Dual was only statistically significant in the *Dso* and *Lmpt* reporters, with *Dso* being more active in Dual and *Lmpt* more active in IMD (p=0.1, two-sided Mann-Whitney *U* test).



Supplemental Figure S8: Most enhancers maintain chromatin structure upon immune stimulation. (A) Fraction of enhancers of each accessibility group by treatment groups. * indicates p = 0.055, z-test (B) Log odds ratio of motifs enriched in accessibility grouped enhancers. We identified four distinct motifs: 1. A motif (GGTCA) assigned to several nuclear hormone receptors 2. A motif for either CrebA or CrebB. CrebA has been shown to act downstream of the Toll and IMD pathways to regulate tolerance to infection (Troha et al. 2018). 3. A motif for Blistered, known to play a role in wing development. 4.“GGTGTGTAT”, a highly enriched motif in HKSM opened enhancers compared to the genome. While no *Drosophila* TF was a clear match to this motif, we found the mouse C2H2 zinc finger Prdm14 had the closest known motif. We have called this sequence a putative zinc finger motif. Positive enrichment values are in blue and negative enrichment values are in red, and significance is indicated by * (p <0.05, Fisher’s test).

Enhancer Name	Coordinates (dm6)	Mutated?
PGRP-SD	3L:7650744-7651565	start codon mutated to ATT
Cog8	2L:11445934-11446855	no
Mtk	2R:15408446-15409027	start codon mutated to ATA
CG5758	2L:18051506-18052253	no
Lmpt	3L:16878911-16879751	no
CrebA	3L:15539043-15539751	no
dso	2R:20869864-20870426	no
CG9837	3R:8813425-8814751	no
lectin-46Cb	2R:9808943-9809693	no
RpS26	2L:18335434-18336088	no

Supplemental Table 1: Enhancer reporters