D1 is a novel enhancer blocker protein that functions via multiple AT-rich simple sequence repeats

Kundurthi Phanindhar[1*], Fathima Athar[1], Rakesh K Mishra[1,2,3*]

[1]CSIR-Center for Cellular and Molecular Biology, Hyderabad-500007, India

[2]Tata Institute for Genetics and Society, Bengaluru-560065, India

[3]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, 201002, India

*Corresponding authors:

Rakesh K Mishra, Email: mishra@ccmb.res.in,

Telephone: +91 40 27192658

Fax: +91 40 27160310

Kundurthi Phanindhar, Email: Kundurthiphanindhar@csirccmb.org

Abstract:

Simple sequence repeats (SSRs) are tandem repeats of 1-6 bp DNA motifs with a size of at least 12 bp, constitute a significant portion of, especially, higher eukaryotes. Earlier studies have shown that length-enriched repeats have various cis-regulatory activities, however, the mechanism of these activities is not well understood. To better understand their function, here we identify proteins that bind to the repeats using protein pull-down followed by LC-MS/MS. We find that many repeats have multiple repeat-binding proteins and that some of these proteins bind to multiple repeats. We report that D1, an AT-hook domain containing protein, binds to multiple repeats, viz., AAT_{14}, $AAAT_{10}$, and $AAAAT_{8}$. As repeats are known to function as enhancer blockers in *Drosophila*, we developed a new enhancer blocker assay that uses visible wing phenotype as enhancer blocking activity and allows use of classic *Drosophila* genetic approach to explore modifiers of this activity at the *vestigial* locus. Using this assay, we show that the repeats AAT_{14}, $AAAT_{10}$, $AAAAT_{8}$ and $AATAT_{9}$ function as enhancer blockers via the protein D1. Our findings make the enhancer blocking regulatory function of a DNA sequence of interest, accessible to the powerful genetics of *Drosophila* and, thereby, presents effective approach to understand the mechanism of such chromatin domain boundary elements.

Keywords:

Simple sequence repeats, SSR, Repeat binding proteins, chromatin domain boundary, enhancer blocker, D1, Drosophila,

Introduction:

Simple sequence repeats (SSRs) or microsatellites or simple tandem repeats (STRs) are tandem repetitions of 1-6 bp DNA motifs with a minimum size of 12 bp (Subramanian et al., 2003). The total number of possible 1-6 bp unique motifs is 5356, classified into 501 SSR classes by combining a motif's cyclical variations and reverse complements into one class (Avvaru et al., 2017). SSRs from here on, referred to as repeats, are functional across organisms from bacteria to humans and are known to regulate many biological processes, including transcription (Hammock & Young, 2005; Horton et al., 2023; Staib et al., 2002; Vinces et al., 2009; Xu et al., 1997) and splicing (Hefferon et al., 2004; Hui et al., 2003, 2005). SSRs are prone to replication slippage and thus show length polymorphism. In prokaryotes, repeat polymorphism generates antigenic variation (Åberg et al., 2014; Eisenbart et al., 2020; Hood et al., 1996; Weiser et al., 1989). In eukaryotes, repeat polymorphism helps regulate gene expression by modulating various biological processes (Albanèse et al., 2001; Fotsing et al., 2019; Gebhardt et al., 1999; Meloni et al., 1998). The repeat length of particular loci increases uncontrollably, leading to repeat expansion disorders (Oberlé et al., 1991; Spada et al., 1991; Verkerk et al., 1991). Length polymorphism helped use repeats as genetic markers (Dietrich et al., 1992; Jacob et al., 1991; Litt & Luty, 1989; Weber & May, 1989). Thus, the repeat length is an essential feature of SSRs.

Generally, the repeat occurrence frequency in a genome is expected to decrease with an increase in repeat length. However, certain repeats show length-specific enrichment, i.e., increased occurrence frequency at longer repeat lengths contrary to expectations pointing to a possible function at these longer lengths (Ramamoorthy et al., 2014). When we analyzed 719 organisms, we observed that 131 repeat classes showed length preference in at least one organism. There were 16 length-enriched repeats in the human genome (Srivastava et al., 2019). However, when we previously analyzed the length preference using an earlier assembly of the human genome, we found 23 length-enriched repeats and most of these repeats showed cis-regulatory functions in K562

cells (Krishnan et al., 2017; Ramamoorthy et al., 2014). So, we used many of these 23 length-enriched repeats for our subsequent experiments.

Enhancer blockers, a boundary element, prevent the interaction between an enhancer and its promoter only when they are present between them (Geyer & Corces, 1992). Enhancer blockers facilitate differential genome packaging resulting in cell-type specific gene expression profile, which ultimately dictates the cell identity (Muravyova et al., 2001; Phillips-Cremins et al., 2013). Repeats function as enhancer blockers. For example, AGAT and GA repeat from humans and mice, respectively, function as enhancer blockers in *Drosophila* (Kumar et al., 2013; Vasanthi et al., 2010). Gypsy, a potent enhancer blocker, has 12 Su(Hw) binding sites (Spana et al., 1988). Repeats are presumed to harbor repetitive protein binding sites due to their repetitive nature. Thus, we hypothesized that enriched repeats may function as enhancer blockers in *Drosophila*.

Traditionally, enhancer blocker assays in *Drosophila* are P-element based assays. The P-elements are integrated randomly into the genome and are subject to position effects, making it difficult to compare test elements. Moreover, these P-element based assays depend on reporter genes such as mini-white or LacZ, requiring molecular biological assays for quantification, thus being low throughput (Hagstrom et al., 1996). So, we thought of developing an assay that integrates various test fragments at the identical site and utilizes endogenous enhancers, which, when blocked, results in a scorable morphological phenotype.

For repeats to function as cis-regulatory elements, they require transacting factors such as proteins to interact with them. However, the identity of many repeat binding proteins is unknown. In this study, we find the presence of repeat binding proteins in *Drosophila* and identify them by performing a DNA-based protein pull-down followed by tandem mass spectrometry. Furthermore, we developed a novel enhancer blocker assay that tests DNA fragments site specifically with improved throughput. This assay enables genetic dissection of enhancer blocker elements using visible wing phenotypes. Using

the assay, we show that some enriched repeats function as enhancer blockers and validate the in vivo role of proteins in the enhancer blocking chromatin domain boundary function in the native context of the enhancer and the target promoter.

Results:

1. **Presence of repeat binding proteins in *Drosophila*:**

Repeats show no length enrichment, but they perform biological functions in *Drosophila*. For example, $AGAT_n$ and GA_n repeats function like enhancer blockers (Kumar et al., 2013; Vasanthi et al., 2010). AAGAG repeat RNA plays a role in *Drosophila* development and male fertility (Mills et al., 2019; Pathak et al., 2013). Interestingly, the symptoms observed in human repeat expansion disorders are also replicated in *Drosophila,* suggesting that *Drosophila* is a good model system for studying repeat function (Miller et al., 2000; Warrick et al., 1998; Zhang et al., 2015). Therefore, we investigated the presence of the repeat binding proteins in *Drosophila* by performing EMSAs with multiple length-enriched repeats from the human genome.

Radiolabelled AAT_{14} showed a retardation in electrophoretic migration in the presence of nuclear extract (0-16 h *Drosophila* embryos), suggesting a DNA-protein complex formation (Figure 1A, lane 2). when a higher concentration of unlabelled AAT_{14} (10-, or 20-, or 50-fold > radiolabelled AAT_{14}) was added, the radiolabelled probe's retardation in migration was rectified, suggesting that the unlabelled AAT_{14} competed out the protein binding to radiolabelled AAT_{14} (Figure 1A, lanes 3-5). However, radiolabelled AAT_{14}'s retardation in migration was unaffected even when a 50-fold higher concentration of an unlabelled repeat dissimilar to AAT_{14}, such as $AAGG_{11}$ or $ATCC_9$, was added, indicating that even high concentrations of dissimilar repeats didn't compete out the protein binding to radiolabelled AAT_{14} (Figure 1A, lanes 8 and 11). In summary, the AAT_{14} sequence specifically binds a protein from the 0-16h *Drosophila* embryos (Figure 1A).

We performed EMSAs with 15 of the 23 repeats and observed that all repeats except $ATCC_9$ have sequence-specific repeat binding proteins in 0-16h *Drosophila* embryos (Figure 1). Also, We noticed that some repeats show multiple shifts in their EMSAs, indicating that either multiple repeat-binding proteins are present or that repeat-binding proteins form higher protein complexes (Figure 1F, 1H, 1N, 1O). We also observed that proteins bound to repeats are competed out with repeats having similar motifs. For example, $AAAAT_8$ competes out the protein binding to $AATAT_9$ and likewise (Figure 1D, 1F, 1G and 1H). The above result indicates that same protein could bind to multiple repeats. However, we could not find a sequence-specific repeat binding protein for $ATCC_9$ with the tested binding conditions (Figure 1M).

2. Proteomics of *Drosophila* repeat binding proteins:

To identify the repeat binding proteins, we chose the following eight repeats that show sequence-specific protein binding in *Drosophila*: AAT_{14}, $AAAT_{10}$ and $AAAAT_8$ repeats show robust shifts in EMSAs. The motifs of $AAAAG_{11}$ and $AAAG_{13}$ were similar to $AAAAT_8$ and $AAAT_{10}$, respectively, except for one nucleotide, and they show 2 EMSA shifts (Figure 1F and 1O). $ACATAT_8$ shows multiple shifts in EMSAs (Figure 1N). $AGAT_{10}$ is a known enhancer blocker and $AATAT_9$ was used as a positive control as earlier studies showed that it binds to the D1 protein (Kumar et al., 2013; Levinger, 1985).

Repeat binding protein identities were deciphered by performing protein pull-downs using repeat DNA as bait followed by LC-MS/MS. We did protein pull-downs in sets of 2 repeats and control in triplicates. In total, we performed eight repeats and four control pull-downs in triplicates, amounting to 36 MS-MS runs. We identified 1320 proteins in 36 MS-MS runs after removing potential contaminants, identified only by site, reverse hits and proteins identified by < 2 unique peptides. The percentage of proteins identified in duplicates across repeats varies between 80.67% to 91% (Supplementary Table 1).

Differentially enriched proteins between repeats and their controls were identified using LFQ analyst with default parameters (Shah et al., 2020). Subsequently, we used the

pairwise comparisons of LFQ analyst to plot volcano plots with log_2(LFQ intensity fold change over control) on the x-axis vs. -log_{10}(adj. p-value) on the y-axis using volcanoseR. The volcano plots show repeat enriched proteins if their log_2(LFQ intensity fold change over control) is > “1.5” and -log_{10}(adj. p-value) is > “2” (Figure 2). The following are the number of enriched proteins for each repeat: $AAAAG_{11}$ - 3, $AAAAT_8$ - 8, $AAAT_{10}$ - 3, AAT_{14} - 4, $AATAT_9$ - 9, $ACATAT_8$ - 9, $AGAT_{10}$ - 7, $AAAG_{13}$ - 10 (Supplementary Table 2). Of these enriched proteins, we observed that D1 is enriched in the pull-downs of $AATAT_9$ as expected (Figure 2D). Secondly, we observed that six proteins enriched in $AAAAT_8$ are also enriched in $AATAT_9$ (Figure 2C and 2D). This was in sync with the EMSA of radiolabelled $AATAT_9$, where $AAAAT_8$ competed out the protein binding to $AATAT_9$ (Figure 1D) and also because the motifs are almost identical. Lastly, we also observed that many of the proteins enriched in $ACATAT_8$, namely Combgap, CG7239, CG1888, prod and CG14005, were also enriched in the pull-down of Combgap. All these indicate that the repeat pull-downs have worked.

We identified 53 repeat-enriched proteins, of which 32 were known or predicted transcription factors. Of these 32 TFs, 12 are enriched in single repeats. The remaining 20 are actually 7 TFs that are enriched in more than one repeat. Of these 19 enriched TFs, 6 TFs have a MADF (myb/SANT-like domain in Adf-1) DNA binding domain and 5 TFs have a zinc finger domain. There are also TFs with ARID (AT-rich interaction domain) domains, AT-hooks, BED finger, etc. Each repeat had more than one TF enriched in their lists with 2/3 in $AAAAG_{11}$, 4/8 in $AAAAT_8$, 3/3 in $AAAT_{10}$, 4/4 in AAT_{14}, 6/9 in $AATAT_9$, 6/9 in $ACATAT_8$, 2/7 in $AGAT_{10}$ and 5/10 in $AAAG_{13}$ (Table 1).

We observed some enriched proteins in multiple repeats, such as D1, which was enriched in 6 repeats, whereas 4 repeats showed enrichment of Jigr1 and Rrp1. Phr, Phr6-4, Tet, Dip3, Jarid2, X16, Diminutive, Top1 and CG8924(CC9) were enriched in 2 repeats each. As expected, 6 of the eight enriched proteins in $AAAAT_8$ were also enriched in $AATAT_9$ because the motifs are identical except for one nucleotide(Supplementary Table 3).

D1 is an AT-hook protein with multiple AT hooks enriched in pull-downs of 6 different repeats. Along with $AATAT_9$, the protein is also enriched in the other AT-rich repeats such as AAT_{14}, $AAAT_{10}$, $AAAAT_8$, $AAAG_{13}$ and $AAAAG_{11}$. D1 protein has been shown to bind to the chromocenters and has a role in heterochromatin maintenance, exemplified by its effect on the *w[m4h]* PEV model (Aulner et al., 2002). Depletion of the D1 protein results in the declustering of the chromocenter and formation of micronuclei with increased DNA damage in the *Drosophila* spermatogonial cells (Jagannathan et al., 2018). We also find other proteins involved in heterochromatin maintenance enriched in various repeat pull-downs. Prod, a MADF domain-containing protein enriched in pull-downs of $ACATAT_8$, is responsible for chromocenter formation in imaginal discs and lymph glands, just as D1 is in spermatogonial cells (Jagannathan et al., 2019). Topoisomerase 2 enriched in $ACATAT_8$ pull-down also contributes to heterochromatin maintenance (Blattes et al., 2006; Napoletano et al., 2021). CG1815 (Zmnyd 8), a zinc finger-containing protein, a novel interactor of HP1a protein was enriched in the pull down of $AAAG_{13}$ repeat (Alekseyenko et al., 2014). Rrp1, a protein associated with both right and left telomere-associated repeats, was also enriched in repeat pull-downs of $ACATAT_8$, $AAAG_{13}$, $AAAAG_{11}$ and $AGAT_{10}$ repeats (Antão et al., 2012). Many proteins involved with heterochromatin are pulled down in various repeat pull-downs, which is expected as repeats are known to be associated with constitutive heterochromatic regions.

We also identified proteins in the repeat pull-downs that have a role in polycomb regulation. Jigr1, a MADF-BESS transcription factor, was pulled down along with Pc, suggesting a possible role in polycomb regulation (Kang et al., 2015; Sun et al., 2006). Jigr1 enhances the nicked wing phenotype caused by the RNAi of domino, a chromatin remodeler, suggesting that Jigr1 does play an essential role in chromatin regulation (Ellis et al., 2015). Another protein, Jarid2, enriched in $AAAAT_8$ and $AATAT_9$ repeats, was pulled down along with E(z), which is a part of the PRC2 complex (Kang et al., 2015). Jarid2 and Jing are purified as a part of the PRC2 complex (Herz et al., 2012; Kang et al., 2015). Jing is the *Drosophila* ortholog of AEBP2, an established component of PRC2. Jigr1 (Jing interacting gene regulatory 1) or D1, when overexpressed in

combination with Jing in the eyes of *Drosophila,* results in eye defects, suggesting that they could function as a part of the complex or that they could have redundant functions (Sun et al., 2006). Combgap(Cg), another protein known to recruit polycomb complexes to PREs, was pulled down with $ACATAT_8$ repeats (Ray et al., 2016). Chorion factor 2(Cf2), which is a paralog of Cg, was also pulled down with $ACATAT_8$ and is known to bind to a sequence of GTATATTAT where GTATAT is a reverse complement of ACATAT (Hsu et al., 1992). All these together indicate that the AT-rich repeats may have a role in polycomb-mediated gene regulation.

RNA binding proteins such as Rox8 and X16 were also pulled down. Rox8 and X16 regulate splicing (Bradley et al., 2015; Park et al., 2004). Moreover, DNA repair proteins such as Rrp1, Phr and Phr6-4 have also been pulled down along with repeats. Rrp1 (recombination repair protein 1) is essential for the DNA repair process via the base excision or recombination repair pathway. Phr and Phr6-4 proteins pulled down with $AAAAT_8$ and $AATAT_9$ repeats also help repair DNA damage caused by UV exposure (Boyd & Harris, 1987; Todo et al., 1996). Many uncharacterized proteins such as CG8924, CG8944, CG1888, CG7239, CG7878, CG14005 and CG10139 were also pulled down in various repeat-based pull-downs. However, how these RNA binding proteins, DNA repair proteins and uncharacterized proteins function via repeats needs to be explored. Altogether, proteomic studies confirm the observations seen in EMSAs. Additionally, the results also suggest that repeats could have a role in heterochromatin maintenance and that repeats may have many uncharacterized functions which need to be further investigated.

3. D1 protein binds multiple repeats:

D1 and Jigr1 together were enriched in AAT_{14}, $AAAT_{10}$, $AAAAT_8$ and $AATAT_9$ pull-downs. However, $AAAAG_{11}$ and $AAAG_{13}$ pull-downs also showed D1 enrichment. When we aligned AAT_{14}, $AAAT_{10}$, $AAAAT_8$ and $AATAT_9$, we noticed the "**AATAA"** motif recurring with AAT_{14}, $AAAT_{10}$ and $AAAAT_8$ having one ("T"), three ("ATA") or zero

nucleotides separating two AATAA motifs respectively. AATAT repeat motifs are similar to AATAA motifs except for the last nucleotide (Figure 3A).

If D1 or Jigr1 bind multiple repeats, we hypothesized that D1 or Jigr1 binding to a particular radiolabeled repeat should be competed out only by unlabeled repeats that also bind D1 or Jigr1. So, we performed EMSA with radiolabelled $AAAT_{10}$, which showed a shift in the presence of nuclear extract (0-16h *Drosophila* embryos), and this shift was competed out when unlabelled $AAAT_{10}$ was added at 10- or, 20- or 50-fold higher concentrations than radiolabelled $AAAT_{10}$. Expectedly, the shift was also competed out by higher concentrations of unlabeled AAT_{14} and $AAAAT_8$ but not by unlabeled ATC_{12} even at 50-fold higher concentration than the radiolabelled probe, suggesting that these three repeats ($AAAT_{10}$, AAT_{14}, and $AAAAT_8$) bind the same protein. Additionally, the amount of free probe seen upon competition with unlabeled $AAAT_{10}$ was higher than $AAT_{14,}$ followed by $AAAAT_8$, indicating that unlabelled $AAAT_{10}$ competes out the protein binding to radiolabelled $AAAT_{10}$ best followed by AAT_{14} and lastly by $AAAAT_8$ (Figure 3B). The competition of radiolabelled AAT_{14} or $AAAAT_8$ with unlabelled AAT_{14}, $AAAT_{10}$ and $AAAAT_8$ resulted in identical results (Figure 3C, 3D). As AAT_{14}, $AAAT_{10}$ and $AAAAT_8$ bind the same protein, we expect that the protein binding to these repeats could be either D1 or Jigr1.

Further, we tested if other repeats also bind to D1 or Jigr1. Since $AAAT_{10}$ competes better, we used radiolabelled $AAAT_{10}$ as the probe for subsequent competition experiments. Competition of radiolabeled $AAAT_{10}$ with unlabelled AAG_{19}, $AAAG_{13}$ and $AAAAG_{11}$ showed that $AAAAG_{11}$ competes out better than $AAAG_{13}$ and only a little competition by AAG_{19} (Figure 3E). However, competition with unlabelled $ACATAT_8$, $AGAT_{10}$ or A_{36} very slightly or did not compete out the protein binding to $AAAT_{10,}$ suggesting that they do not bind to these repeats (Figure 3F). Similarly, $AATAG_{11}$ and $AACAT_{10}$ barely competed out the protein binding to $AAAT_{10}$ compared to $AATAT_9$ (Figure 3G). We also plotted the LFQ intensities of D1 and Jigr1 across various repeat DNA-based protein pull-downs to observe if there is any correlation between competition EMSAs performed with radiolabeled $AAAT_{10}$ and proteomic runs. The LFQ

intensities of D1 across various repeat pull-downs were of the following order: $AAAT_{10}$~AAT_{14}~$AATAT_9$>$AAAAT_8$>$AAAAG_{11}$~$AAAG_{13}$>$ACATAT_8$>$AGAT_{10}$=controls (Figure 3H). Whereas the LFQ intensities of Jigr1 were in the following order: $AAAAT_8$>AAT_{14}~$AATAT_9$>$AAAT_{10}$>$AAAAG_{11}$~$AAAG_{13}$~$ACATAT_8$~$AGAT_{10}$~controls (Figure 3I). Since EMSA competition patterns match the LFQ intensities of D1 across repeat pull-downs, the protein binding to the $AAAT_{10}$ repeat is likely D1 rather than Jigr1.

4. Development of novel enhancer blocker assay in *Drosophila*:

We wanted to test the enhancer blocker activity of AAT_{14}, $AAAT_{10}$, $AAAAT_8$ and $AATAT_9$ as some repeats function as enhancer blockers and because these repeats bind the same protein D1 but with varying affinity. Traditionally, enhancer blocker assays in *Drosophila* are P-element transposon based which are subject to position effects due to their random integration. The assays have low throughput as they require molecular biology-based quantitations (Hagstrom et al., 1996). To address these drawbacks, we developed a novel enhancer blocker assay. The novel assay allows site-specific integration of various test fragments using the attP-attB based recombination system, thus avoiding position effects (Bateman & Wu, 2008; Groth et al., 2004). The assay utilizes the endogenous enhancers of the *vestigial (vg)* gene, which, when blocked, results in wing abnormalities, allowing direct quantification.

vestigial locus was utilized to generate attP landing sites:

vestigial gene is essential for *Drosophila* wing and haltere development and DV-enhancer and quadrant enhancer regulate *vg* gene. DV-enhancer drives the *vg* expression in the dorsoventral (DV) and anteroposterior (AP) margins of the wing pouch. In contrast, the quadrant enhancer drives the *vg* expression in the wing pouch quadrants demarcated by DV and AP margins (Figure 4A, (Kim et al., 1996)). DV-enhancer deletion results in complete loss of wings and halteres, as exemplified in

vg[83b27] (Williams et al., 1991). Thus, we hypothesized that blocking the DV-enhancer should also result in the loss of wings and halteres.

To develop the assay, we inserted an attP landing site, $attP1_{220}$-3XP3GFP-$attP2_{220}$, upstream of the DV-enhancer in the 2nd intron of the *vestigial* locus using the CRISPR-Cas9 system. The resulting transgene is referred to as DV-enhancer attP line. The attP landing site insertion in the *vg* locus results in GFP expression in the eyes (Supplementary Figure 1A). We generated another attP landing site where the insert $attP1_{220}$-3XP3GFP-$attP2_{220}$ was inserted downstream of the DV-enhancer in the 2nd intron. Both the attP lines have normal wings and halteres, suggesting that the attP insertions did not affect any of the functional elements required for wing and haltere development.

Gypsy functions like an enhancer blocker in the *vestigial* enhancer blocker assay:

To test the *vestigial* enhancer blocker assay, we used gypsy as a positive control. To accomplish recombination-mediated cassette exchange (RMCE) between the $attP_{220}$ sites of the *vestigial* locus and the $attB_{40}$ sites of the pBSKS attB1-gypsy-attB2 plasmid, the attP transgenic lines were brought in the background of *nanos*-phiC31 integrase. Upon RMCE, the 3XP3GFP in the *vestigial* locus is exchanged with gypsy from the pBSKS plasmid (Figure 4B). Thus, RMCE successful flies show loss of GFP in the eyes (Supplementary Figure 1B).

Gypsy integrated DV-enhancer attP line referred to as DV-gypsy shows a good percentage of lethality in homozygous conditions. When heterozygous DV-gypsy lines were selfed, only 13.7% of homozygous flies survived against the expected 33% (Figure 4D). Of the surviving homozygous flies, 16.67% had wings absent, around 41.67% had one wing absent and the rest 41.67% had normal wings (Figure 4C and Supplementary Figure 1C). Almost all homozygous flies also show upright post-scutellar bristles and around 80% show the absence of both halteres (Supplementary Figure 1D and Figure 4C). Moreover, around 80% of the flies that show at least one wing absent also show a

transformation to notum (Supplementary Figure 1D). All these phenotypes are characteristic of *vestigial* mutants (Williams et al., 1990). Taken together, gypsy functions like an enhancer blocker in the *vestigial* enhancer blocker assay.

A protein complex consisting of DNA binding protein Su(Hw), Mod(mdg4) and Cp190 regulates gypsy activity (Gerasimova et al., 1995; Pai et al., 2004; Spana et al., 1988). To test the gypsy enhancer blocker activity, we brought DV-gypsy lines in the mutant background of either *su(Hw)*, *mod(mdg4)*, or *Cp190*. When flies heterozygous for both DV-gypsy and *su(Hw)* mutation were selfed, we observed 13.5% flies (expected 11.11% flies) homozygous for both gypsy as well as *su(Hw)* mutation with normal wings and halteres except for mild wing notches in 5% flies (Figure 4C and 4E). Whereas homozygous DV-gypsy flies heterozygous for the *su(Hw)* mutation had no halteres in 80% of flies and defective wings in 13%. The above results indicate that the phenotype rescue depends on the *su(Hw)* mutation dosage. DV-gypsy heterozygous flies had normal wings and halteres irrespective of *su(Hw)* mutation zygosity. Moreover, homozygous *su(Hw)* mutation rescues the abnormal morphology of wing imaginal discs from the 3rd instar larvae of homozygous DV-gypsy line (Figure 4C). When DV-gypsy/CyO; *mod(mdg4)[ul]*/TM6B flies were selfed, we observed only 6% gypsy homozygous flies (Figure 4F). While DV-gypsy/CyO; *Cp190[H4-1]*/TM6B flies when selfed resulted in only 2% DV-gypsy; *Cp190[H4-1]*/TM6B flies (Figure 4G). The reason for partial rescue is that the homozygous *Cp190* mutation is lethal and also because *su(Hw)* majorly influences gypsy function. Collectively, all the above results suggest that gypsy functions like an enhancer blocker in the novel *vestigial* enhancer blocker, and the assay can be used to assay various test fragments.

5. Repeats function as enhancer blockers:

We integrated four different AT-rich repeats known to bind D1 into the DV enhancer attP line individually, and generated four different lines referred to as DV-"repeat" line. We assayed the homozygous DV-repeat lines for wing phenotypes and observed a meager percentage of flies having a mild phenotype (T2 phenotype in Figure 5B): DV-AAT$_{14}$ -

0.88%, DV-$AAAT_{10}$-0.43%, DV-$AAAAT_8$-1.15% and DV-$AATAT_9$-5.26%. The percentages are an aggregate of phenotyping multiple lines for each repeat with the repeat in forward orientation. One possible reason for the mild phenotype could be that the repeat binding protein, D1, could be absent in the DV border cells, or its expression could be low. So, we overexpressed D1 protein using C96Gal4. C96Gal4 expresses the protein in the DV border cells, where the DV-enhancer is active (Supplementary Figure 2B). Overexpression of D1 in the homozygous DV-repeat lines resulted in a range of phenotypes: T1 - wild type, T2 - has less than three notches or loss of bristles or both, T3 - has three or more notches, T4 - decreased wing length with disorganized arrangement of cells and T5 - small sized wings (Figure 5B). The percentage of T1 wings we observed upon D1 overexpression in the homozygous repeat flies was highest for $AAAAT_8$ at 10%, followed by AAT_{14} and $AATAT_9$, and the least for $AAAT_{10}$ at 0.38%. The T4 and T5 phenotypes were highest for $AAAT_{10}$ and the lowest for AAT_{14}, with the order being $AAAT_{10}$>$AATAT_9$>$AAAAT_8$>AAT_{14} (Figure 5C). At the same time, The T2 phenotype, which was the milder phenotype, was the highest for $AAT_{14,}$ with the order being AAT_{14}>$AAAAT_8$>$AATAT_9$>$AAAT_{10}$, which is the reverse of the order of T4 and T5 (Figure 5C). Overexpression of D1 without the repeat background (C96Gal4>UASD1) majorly resulted in only two phenotypes, T2 and T3, with the percentage of T3 being at least 75%. We did observe T4 and T5, but they were less than 1% where ever observed (Figure 5C). These results indicate that $AAAT_{10}$ and $AATAT_9$ function as strong enhancer blockers, whereas AAT_{14} functions the weakest among the four AT-rich repeats.

Discussion:

SSRs regulate multiple biological processes and sometimes result in repeat expansion disorders when they expand uncontrollably. However, their normal physiological roles are still poorly understood. Distributional analysis of SSRs across various organisms found some SSRs enriched at longer repeat lengths contrary to expectations. 23 length-enriched repeats in the human genome perform cis-regulatory functions (Krishnan et al., 2017; Ramamoorthy et al., 2014). To understand their function better, We performed

EMSAs with repeats and identified sequence-specific repeat-binding proteins in *Drosophila*. We report that many sequence-specific repeat binding proteins probably bind multiple repeats. For example, $AAAAG_{11}$ competes out protein binding to radiolabelled $AAGAG_{12}$ and likewise (Figure 1D, 1F, 1G and 1H). Proteomic analysis of repeat enriched proteins showed many commonalities between similar repeats $AATAT_9$ and $AAAAT_8$ (Figure 2C and 2D). We also identified a list of proteins that are enriched in at least two repeats (Supplementary Table 3). In 6 repeat pull downs, we observed D1 protein enrichment and this was also confirmed by competition EMSAs (Figure 3). We also show that D1 is responsible for the enhancer blocker activity of 4 different repeats: AAT_{14}, $AAAT_{10}$, $AAAAT_8$ and $AATAT_{9,}$ but with differing activity (Figure 5). A protein binding multiple repeats with differing affinity can coregulate genes having these repeats in their regulatory regions to varying extents depending on the protein binding affinity, and the regulation can be altered depending on the protein concentration.

Another aspect evident from proteomics is that some repeats bind more than one DNA-binding protein and is not unusual as GA_n repeats are known to be bound by at least two different Zn finger proteins, such as GAGA-associated factor (GAF) and Clamp. GAF has one C2H2 zinc finger and binds both long and short GA repeats, whereas Clamp binds only to long GA repeats because of its six C2H2 zinc fingers (Kaye et al., 2018). In our data, D1 and Jigr1 are enriched in 4 different repeats but with differing intensities. However, in the EMSAs of AAT_{14}, $AAAT_{10}$ and $AAAAT_8$, we don't find two different shifts as the size of D1 and Jigr1 is quite similar, around 37kDa (Figure 1A-1C). D1 or Jigr1 (Jing interacting gene regulatory 1), when overexpressed in combination with Jing in the eyes of *Drosophila,* results in eye defects suggesting that they could function as a part of the complex or that they could have redundant functions (Sun et al., 2006). Additionally, D1 but not Jigr1 is enriched in the repeats of $AAAG_{13}$ and $AAAAG_{11}$. Thus, D1 may have a redundant role and some context-specific functions. It would be interesting to understand how D1 and Jigr1 regulate these repeats when both proteins express in the same cell type. We observe multiple transcription factors that bind each of the repeats (Table 1).

Of the 19 TFs uniquely enriched across various repeats, 6 have MADF domain. In humans, telomere repeat binding proteins, TRF1 and TRF2, have a myb DNA binding domain (Broccoli et al., 1997). MADF domain containing proteins such as Dip3, CG3163, CG4004 and Adf1 are enriched in the telomere repeat regions of *Drosophila* (Antão et al., 2012). Since TFs enriched in our pull-downs majorly have MADF domains, it's tempting to speculate that MADF domains may have a propensity to bind repeats.

Many proteins involved in heterochromatin maintenance are enriched across various repeat pull-downs. However, the same set of heterochromatin maintenance proteins are not enriched in every repeat. The reason for this could be that different proteins sometimes perform the same function in distinct cell types. For example, D1 is responsible for chromocenter maintenance in the spermatogonial cells whereas Prod is responsible for chromocenter maintenance in the imaginal discs and lymph glands. Another possibility could be that repeat binding proteins bind sequences other than repeats and therefore, it is important to ascertain whether the phenotypes observed upon repeat protein deletion actually function via repeat sequences. We also observe that many AT-rich repeat binding proteins have a role in polycomb regulation. It would be interesting to study how the AT-rich repeats contribute to polycomb regulation.

Repeats function as enhancer blockers. To test the length enriched repeats, we developed a novel enhancer blocker assay. Traditional P-element-based enhancer blocker assays are subject to position effects and have low throughput (Hagstrom et al., 1996). We developed a novel assay by inserting the attP1-3XP3GFP-attP2 into the *vestigial* locus. One transgenic line had the attP landing site between the promoter and the DV-enhancer (DV-enhancer attP line). In contrast, we inserted the second attP landing site after the DV-enhancer (Quadrant enhancer attP line). Both the transgenic lines had normal wings and halteres, suggesting that the attP landing site insertion did not affect any functional elements. When we integrated gypsy into the DV-enhancer attP line, we observed greater than 50% lethality in the DV-gypsy homozygous flies. Of the surviving homozygous flies, most had both halteres absent and more than half had at least one wing missing (Figure 4C and 4D). The wing and haltere effects are

expected as deletion of the DV-enhancer is known to affect the wing and haltere formation as in *vg[83b27]* (Williams et al., 1991). We also observed flies with upright postscutellar bristles and transformation to notum in flies that have wings absent. All these phenotypes are observed in some of the extreme *vestigial* mutants suggesting that gypsy blocks not only the DV-enhancer but the other enhancers of the *Vg* gene (Williams et al., 1990). All these phenotypes were rescued upon homozygous *su(Hw)* mutation suggesting that gypsy functions as an enhancer blocker in the DV-enhancer attP line (Figure 4C and 4E). We did not observe any phenotypes when we blocked the QE enhancer with gypsy. To test the enhancer blocker activity of repeats, we integrated AAT_{14}, $AAAT_{10}$, $AAAAT_{8}$ or $AATAT_{9}$ into the DV-enhancer attP line. Integration of repeats only showed a mild phenotype that too, in less than 5% of the flies. We assumed that one possibility could be the absence of the protein that binds these repeats. So, we overexpressed D1 using C96Gal4 in the DV border cells of the wing pouch of the homozygous DV-repeat lines, and we observed that all of them show small wings (T5), but the percentage of small wings varies. The order of enhancer blocking activity was $AAAT_{10}$~$AATAT_{9}$>$AAAAT_{8}$>AAT_{14} and doesn't match the binding affinities observed in EMSAs (Figure 5C). One possible reason for the mismatch could be the difference in the DNA interacting partners other than D1. The difference in the interacting partners could be because EMSAs were performed with nuclear extract from 0-16h embryos, whereas the enhancer blocker assay works in the wing imaginal discs.

In summary, we identify repeat binding proteins and demonstrate their role in the enhancer blocking activity of chromatin domain boundary activity of repeats. The novel assay developed in the study enables use of classical fly genetics to identify modifiers of enhancer blocking activity and thereby opens this unique function of chromatin structure level regulation, for the first time, to genetic analysis. Furthermore, as the assay uses visible wing phenotype as the reporter of the test element in vivo, it also allows to setup genetic screens to uncover the molecular mechanisms of in vivo boundary function.

Materials and methods:

Preparation of nuclear extract from *Drosophila* embryos:

10 g of dechorionated 0-16 h *Drosophila* embryos were suspended in 20 ml of nuclei isolation buffer I (NIB I) (15 mM HEPES pH 7.6, 40 mM KCl, 1mM EDTA, 0.25 M sucrose, 0.1 mM spermine, 0.125 mM Spermidine, PMSF, 1X protease inhibitor cocktail), homogenized using five strokes. Homogenized embryos were filtered through Mira cloth and washed with 5 ml of NIB I. The filterate was centrifuged at 600g for 30 s at 4^0 C, and the supernatant was mixed with an equal volume of NIB II (15 mM HEPES pH 7.6, 40 mM KCl, 1 mM EDTA, 1.8 M sucrose, 0.1 mM spermine, 0.125 mM Spermidine, PMSF, 1X protease inhibitor cocktail), centrifuged at 8000 rpm at 4^0 C. The obtained pellet was submerged in NIB I for 5 min on ice, resuspended and centrifuged at 1000g for 10 minutes at 4^0 C. The above step was repeated and the pellet was resuspended in 2 ml of NEB 20 (10 mM HEPES pH 7.6, 20 mM KCl, 3 mM $MgCl_2$, 1 mM EDTA, 1M Sucrose, 10% glycerol, 1X PIC, PMSF). Then an equal volume of NEB 700 (10 mM HEPES pH 7.6, 700 mM KCl, 3 mM $MgCl_2$, 1 mM EDTA, 1M Sucrose, 10% glycerol, 1X PIC, PMSF) was added slowly with mixing. The solution was rocked gently for 1 h at 4^0 C. Subsequently, it was centrifuged at 42000 rpm for an hour, the supernatant was collected and stored as aliquots at -80^0 C.

Electrophoretic Mobility Shift Assay (EMSA):

Synthesized oligos were annealed in an annealing buffer (10mM Tris pH 8, 50mM NaCl, 1mM EDTA) by heating them to 95^0 C for 2 min followed by decreasing 2^0 C/s to 84^0 C. Subsequently, the oligos were cooled to 25^0 C/s by decrementing 0.1^0 C/s. 15 pmol of annealed oligos were end-labeled using γ^{32}P-ATP and T4 Polynucleotide kinase (New England Biolabs) for 1 h at 37^0C. The radiolabelled oligos were purified from the leftover γ^{32}P-ATP by passing it through the Sephadex G-50 (Sigma) column. The eluate was adjusted to 150 ul so that the final concentration of the oligos was 0.1 pmol/ul.

Binding reactions were performed in a reaction volume of 15 ul containing 25 mM HEPES pH 7.6, 10% glycerol, 1.5 ug tRNA, 1 mM EDTA, 150 mM NaCl (this includes the salt contributed by nuclear extract as well), 0.3 pmol of radiolabelled oligos and 4 ug of nuclear extract from 0-16 h *Drosophila* embryos. The binding reactions were incubated at RT for 15 mins, followed by incubation at 4^0 C for another 15 mins. Whenever competition reactions were performed, the unlabelled oligos were added to the above binding reactions at 10-fold, 20-fold, or 50-fold higher concentrations compared to the radiolabelled oligos. However, in some cases, competitions were performed at 10-fold, or 50-fold higher concentrations owing to constraints in the number of wells.

The reaction mixtures were loaded onto a 4% acrylamide-bisacrylamide gels (39:1 acrylamide: bisacrylamide) in 0.5X Tris-borate-EDTA (TBE) and were run at 15 mA at 4^0C for 3 h in 0.5X TBE buffer. The gels were subsequently dried and exposed on a phosphor screen, which was imaged using a phosphorimager.

DNA-based protein pulldowns:

Repeat DNA was amplified from repeat cloned pgl3 vectors (Krishnan et al., 2017) using a 5' biotinylated forward primer and an unmodified reverse primer. The control DNA was amplified using the same set of primers from the pgl3 vector without repeats. 5 ug of purified biotinylated DNA was bound to 20ul of M280 streptavidin beads in 1X binding buffer (5 mM Tris pH 7.5, 0.5 mM EDTA, 1 M NaCl) at 4^0 C. Subsequently, the beads were washed twice with 1X binding buffer and collected using a magnetic stand.

20 ul of DNA-bound beads were incubated with 200 ug of 0-16 h *Drosophila* embryonic nuclear extract for 1 h at RT in 25 mM HEPES pH 7.6, 10% glycerol, 1.5 ug tRNA, 1 mM EDTA, 150 mM NaCl (including the salt contribution by nuclear extract), 1X protease inhibitor cocktail. Then, the protein-bound beads were separated and washed for 10 min at 4^0C with wash buffer (25 mM HEPES pH 7.6, 10% glycerol, 1.5 ug tRNA, 1 mM EDTA, 150 mM NaCl, 1X protease inhibitor cocktail) thrice. The separated beads

are then suspended in 1X laemlli buffer and heated for 5 min at 95^0 C. Finally, the beads are separated, and the extract is stored.

In-Gel digestion for LC-MS/MS:

The pulled-down protein mixture was resolved for 1 cm on 12% SDS-PAGE. The gel was cut into 1 mm^3 pieces and washed thrice in 50% ACN in 25 mM NH_4HCO_3, followed by dehydration in 100% ACN with vigorous vortexing. Dehydrated gel pieces were speed vacuumed for 20 min followed by a reduction of disulfide bonds in 10 mM DTT in 25 mM NH_4HCO_3 at 55^0 C for 45 min. Subsequently, the gel pieces were incubated in 55 mM Iodoacetamide in 25 mM NH_4HCO_3 at RT in the dark for 45 min. Then the gel pieces were dehydrated in 100% ACN and speed vacuumed later for 10 min. These dried gel pieces were incubated with trypsin gold at 37^0 C for 16 h. The tryptic peptides were eluted twice in 2.5% TFA in 50% ACN by incubating the gel pieces at RT with vigorous vortexing for 30 mins. The supernatant containing peptides was quickly vacuumed to dryness and resuspended in 2% ACN in 0.1% formic acid. The tryptic peptides were desalted using p10 Ziptip C18 columns and the desalted peptides were resuspended in 2% ACN in 0.1% formic acid.

Proteomics and data analysis:

The spectra were collected by running on Thermoscientific Q-Exactive HF using a 60 min acetonitrile gradient. The spectra obtained were analyzed using max quant version 2.1.0.0 with default parameters, except that “Label-free quantification” and “match between runs” were applied (Cox & Mann, 2008). The protein groups file obtained from maxquant was then further analyzed using LFQ analyst at default parameters (Shah et al., 2020). The volcano plots were plotted using VolcaNoseR with the output table obtained from LFQ analyst (Goedhart & Luijsterburg, 2020).

Molecular Cloning:

Generation of the gRNA plasmids: Annealed gRNA oligos were ligated into BbsI digested pCFD3.

Generation of the donor plasmids: SmaI digested pBSKS was ligated with attP1_3XP3GFP. The resultant plasmid was digested with SacI and SpeI and ligated with left homology arms. The left homology arms for the DV-enhancer attP line were amplified by LfDV_F and LfDV_R, and the QE-enhancer attP line was amplified by LfQE_F and LFQE_R. The resultant plasmid was subsequently digested with SalI and KpnI and ligated with right homology arms. The right homology arms were amplified with RtDV_F and RtDV_R for the DV-enhancer attP line and RtQE_F and RtQE_R for the QE-enhancer attP line. The resultant plasmid was digested with EcoRV and ligated with attP2 (orientation opposite to attP1) for both vectors.

Generation of pBSKS_attB1$_{40}$_attB2$_{40}$ plasmid: attB$_{40}$ oligos were annealed and ligated into SacI digested pBSKS and the resulting plasmid was digested with KpnI and ligated with another set of attB$_{40}$ oligos in opposite orientation to generate pBSKS_attB1$_{40}$_attB2$_{40}$ plasmid.

Cloning of test fragments into pBSKS_attB1$_{40}$_attB2$_{40}$ plasmid:
Gypsy: pCFD3 plasmid was digested with StuI and BglII and the resulting 411bp gypsy insert was ligated into SmaI and BglII digested LML (LoxP1-MCS-LoxP2) vector (Vasanthi et al., 2010).

Repeats: SmaI digested LML vector was ligated with annealed oligos (AAT$_{14}$ or AAAT$_{10}$ or AAAAT$_{8}$ or AATAT$_{9}$).

Gypsy or repeats ligated LML plasmids were digested with NotIHF and the resulting inserts were ligated into NotI digested pBSKS_attB1$_{40}$_attB2$_{40}$.

Screening of attP lines:

Injected *nanos-Cas9* (X-chromosome) flies (Bloomington St. No. 54591) were mated to Pin/CyO flies in single fly crosses. The GFP-positive and CyO flies were subsequently crossed to Pin/CyO flies again in single fly crosses. The CyO-only flies in the next generation were selfed and the lines balanced. Eventually, the *nanos-Cas9* containing X-chromosome was crossed out if present.

Screening of Gypsy or repeat integrated flies:

Injected nanos-PhiC31 integrase; attP1-3XP3GFP-attP2 male flies were crossed to Pin/CyO virgins in single fly crosses, and their male progeny with loss of GFP and CyO were subsequently crossed to Pin/CyO virgins in single fly crosses in the next generation. The resulting CyO-only flies were selfed, maintained and sequence confirmed. The integration of the test elements in the RMCE flies can be in the forward or reverse orientation. The repeat integrated flies that had forward orientation were phenotyped and the line that showed the least phenotype was used for overexpression of D1.

Genetic Crosses:

<u>Generation of DV-enhancer attP line or QE enhancer attP line with PhiC31 integrase background:</u>

nanos-PhiC31 integrase; attP40 (Bloomington stock No. 79604) females were crossed to FM7a/Y; Sp/CyO flies to obtain nanos-PhiC31 int./Fm7a; attP40/CyO females, which were crossed to w[1118]/y; attP1-3XP3GFP-attP2/CyO flies. nanos-PhiC31 int./y; attP1-3XP3GFP-attP2/CyO males were mated to nanos-PhiC31 int./Fm7a; attP40/CyO females. nanos-PhiC31 int.; attP1-3XP3GFP-attP2/CyO were selfed and balanced.

<u>Generation of DV-Gypsy/CyO; Cp190[H4-1]/TM6B and DV-Gypsy/CyO; mod(mdg4)[ul]/TM6B:</u>

We crossed males of y[2] w[1] ct[6]; Cp190[H4-1]/TM6B, Tb1 (Bl. Stock No. 59965) with db virgins and collected male progeny with Pin and TM2 only. The Pin and TM2 only males were crossed to db virgins again. In the next generation, w[1118]; +/CyO; Cp190[H4-1]/(TM2 or TM6B) virgins were mated to w[1118]/y; DV-Gypsy/pin;

TM2/TM6B. The resulting progeny of w[1118]; DV-Gypsy/CyO; Cp190[H4-1]/TM6B were selfed, and the stock was maintained. The same strategy was followed for the mod(mdg4)[ul] (Bl. Stock No. 59960) stock as well.

DV-Gypsy/CyO male flies were mated to db virgins to obtain w1118/y; DV-gypsy/Pin; +/TM2, which were then mated again to have w1118/y; DV-gypsy/Pin; TM2/TM6B males.

Generation of DV-gypsy/CyO; su(hw)[8]/TM6B:

y[2] sc[1] v[1]; su(Hw)[8] (Bloomington stock no. 1053) males were mated to db virgins and y[2] sc[1] v[1]/w[1118]; +/CyO; su[Hw]8/TM2 progeny were collected and mated to w1118/y; DV-gypsy/Pin; +/TM6B progeny of DV-Gypsy/CyO and db virgins. In the subsequent generation, y[2] sc[1] v[1]/w1118; DV-gypsy/CyO; Su[Hw]8/TM6B flies were selfed and balanced.

Balancing of C96Gal4 or UASD1-3XHA:

Males of C96Gal4 (Bloomington stock No. 25757) were mated to db virgins, and the Pin and TM2 only male progeny were mated to db virgins again. The Pin, CyO and TM2 (or TM6B) only flies were collected and selfed. The same procedure was used to balance the UASD1-3XHA (F000455 from the flyORF collection). The resulting flies were w/y;Pin/CyO; C96Gal4 (or UAS-D1-3XHA)/TM2 (or TM6B).

DV-Repeat flies:

Repeat integrated (RMCE) males were mated to db virgins. The Pin (or CyO) and TM2 (or TM6B) only males were collected and mated to db virgins in the next generation. The resulting DV-repeat/CyO; TM2/TM6B flies were selfed and maintained.

The w1118; DV-repeat/CyO; TM2/TM6B were crossed to w/y;Pin/CyO;UASD1/TM2 and in the subsequent generation w; DV-repeat/CyO; UASD1/TM2 (or TM6B) were selfed. Similarly, w; DV-repeat/CyO; C96Gal4/TM2 (or TM6B) were generated.

DV-repeat; UASD1/C96Gal4 were obtained by crossing DV-repeat; UASD1/UASD1 with DV-repeat; C96Gal4/TM2. The wings of the controls and test flies were phenotyped according to the phenotyping scale given in Figure 5.

References:

Åberg, A., Gideonsson, P., Vallström, A., Olofsson, A., Öhman, C., Rakhimova, L., Borén, T., Engstrand, L., Brännström, K., & Arnqvist, A. (2014). A Repetitive DNA Element Regulates Expression of the Helicobacter pylori Sialic Acid Binding Adhesin by a Rheostat-like Mechanism. *PLOS Pathogens*, *10*(7), e1004234. https://doi.org/10.1371/journal.ppat.1004234

Albanèse, V., Biguet, N. F., Kiefer, H., Bayard, E., Mallet, J., & Meloni, R. (2001). Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite. *Human Molecular Genetics*, *10*(17), 1785–1792. https://doi.org/10.1093/hmg/10.17.1785

Alekseyenko, A. A., Gorchakov, A. A., Zee, B. M., Fuchs, S. M., Kharchenko, P. V., & Kuroda, M. I. (2014). Heterochromatin-associated interactions of Drosophila HP1a with dADD1, HIPP1, and repetitive RNAs. *Genes & Development*, *28*(13), 1445–1460. https://doi.org/10.1101/gad.241950.114

Antão, J. M., Mason, J. M., Déjardin, J., & Kingston, R. E. (2012). Protein Landscape at Drosophila melanogaster Telomere-Associated Sequence Repeats. *Molecular and Cellular Biology*, *32*(12), 2170–2182. https://doi.org/10.1128/MCB.00010-12

Aulner, N., Monod, C., Mandicourt, G., Jullien, D., Cuvier, O., Sall, A., Janssen, S., Laemmli, U. K., & Käs, E. (2002). The AT-Hook Protein D1 Is Essential for Drosophila melanogaster Development and Is Implicated in Position-Effect Variegation. *Molecular and Cellular Biology*, *22*(4), 1218–1232. https://doi.org/10.1128/MCB.22.4.1218-1232.2002

Avvaru, A. K., Saxena, S., Sowpati, D. T., & Mishra, R. K. (2017). MSDB: A Comprehensive Database of Simple Sequence Repeats. *Genome Biology and Evolution*, *9*(6), 1797–1802. https://doi.org/10.1093/gbe/evx132

Bateman, J. R., & Wu, C. (2008). A Simple Polymerase Chain Reaction-Based Method for the Construction of Recombinase-Mediated Cassette Exchange Donor Vectors. *Genetics*, *180*(3), 1763–1766.

https://doi.org/10.1534/genetics.108.094508

Blattes, R., Monod, C., Susbielle, G., Cuvier, O., Wu, J., Hsieh, T., Laemmli, U. K., & Käs, E. (2006). Displacement of D1, HP1 and topoisomerase II from satellite heterochromatin by a specific polyamide. *The EMBO Journal*, *25*(11), 2397–2408. https://doi.org/10.1038/sj.emboj.7601125

Boyd, J. B., & Harris, P. V. (1987). Isolation and Characterization of a Photorepair-Deficient Mutant in Drosophila melanogaster. *Genetics*, *116*(2), 233–239. https://doi.org/10.1093/genetics/116.2.233

Bradley, T., Cook, M. E., & Blanchette, M. (2015). SR proteins control a complex network of RNA-processing events. *RNA*, *21*(1), 75–92. https://doi.org/10.1261/rna.043893.113

Broccoli, D., Smogorzewska, A., Chong, L., & de Lange, T. (1997). Human telomeres contain two distinct Myb–related proteins, TRF1 and TRF2. *Nature Genetics*, *17*(2), Article 2. https://doi.org/10.1038/ng1097-231

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, *26*(12), Article 12. https://doi.org/10.1038/nbt.1511

Dietrich, W., Katz, H., Lincoln, S. E., Shin, H. S., Friedman, J., Dracopoli, N. C., & Lander, E. S. (1992). A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics*, *131*(2), 423–447. https://doi.org/10.1093/genetics/131.2.423

Eisenbart, S. K., Alzheimer, M., Pernitzsch, S. R., Dietrich, S., Stahl, S., & Sharma, C. M. (2020). A Repeat-Associated Small RNA Controls the Major Virulence Factors of Helicobacter pylori. *Molecular Cell*, *80*(2), 210-226.e7. https://doi.org/10.1016/j.molcel.2020.09.009

Ellis, K., Friedman, C., & Yedvobnick, B. (2015). Drosophila domino Exhibits Genetic Interactions with a Wide Spectrum of Chromatin Protein-Encoding Loci. *PLOS ONE*, *10*(11), e0142635. https://doi.org/10.1371/journal.pone.0142635

Fotsing, S. F., Margoliash, J., Wang, C., Saini, S., Yanicky, R., Shleizer-Burko, S., Goren, A., & Gymrek, M. (2019). The impact of short tandem repeat variation on gene expression. *Nature Genetics*, *51*(11), Article 11.

https://doi.org/10.1038/s41588-019-0521-9

Gebhardt, F., Zänker, K. S., & Brandt, B. (1999). Modulation of Epidermal Growth Factor Receptor Gene Transcription by a Polymorphic Dinucleotide Repeat in Intron 1 *. *Journal of Biological Chemistry*, *274*(19), 13176–13180. https://doi.org/10.1074/jbc.274.19.13176

Gerasimova, T. I., Gdula, D. A., Gerasimov, D. V., Simonova, O., & Corces, V. G. (1995). A Drosophila protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation. *Cell*, *82*(4), 587–597. https://doi.org/10.1016/0092-8674(95)90031-4

Geyer, P. K., & Corces, V. G. (1992). DNA position-specific repression of transcription by a Drosophila zinc finger protein. *Genes & Development*, *6*(10), 1865–1873. https://doi.org/10.1101/gad.6.10.1865

Goedhart, J., & Luijsterburg, M. S. (2020). VolcaNoseR is a web app for creating, exploring, labeling and sharing volcano plots. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-76603-3

Groth, A. C., Fish, M., Nusse, R., & Calos, M. P. (2004). Construction of Transgenic Drosophila by Using the Site-Specific Integrase From Phage φC31. *Genetics*, *166*(4), 1775–1782. https://doi.org/10.1093/genetics/166.4.1775

Hagstrom, K., Muller, M., & Schedl, P. (1996). Fab-7 functions as a chromatin domain boundary to ensure proper segment specification by the Drosophila bithorax complex. *Genes & Development*, *10*(24), 3202–3215.

Hammock, E. A. D., & Young, L. J. (2005). Microsatellite Instability Generates Diversity in Brain and Sociobehavioral Traits. *Science*, *308*(5728), 1630–1634. https://doi.org/10.1126/science.1111427

Hefferon, T. W., Groman, J. D., Yurk, C. E., & Cutting, G. R. (2004). A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proceedings of the National Academy of Sciences*, *101*(10), 3504–3509. https://doi.org/10.1073/pnas.0400182101

Herz, H.-M., Mohan, M., Garruss, A. S., Liang, K., Takahashi, Y., Mickey, K., Voets, O., Verrijzer, C. P., & Shilatifard, A. (2012). Enhancer-associated H3K4

monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes & Development*, *26*(23), 2604–2620. https://doi.org/10.1101/gad.201327.112

Hood, D. W., Deadman, M. E., Jennings, M. P., Bisercic, M., Fleischmann, R. D., Venter, J. C., & Moxon, E. R. (1996). DNA repeats identify novel virulence genes in Haemophilus influenzae. *Proceedings of the National Academy of Sciences*, *93*(20), 11121–11125. https://doi.org/10.1073/pnas.93.20.11121

Horton, C. A., Alexandari, A. M., Hayes, M. G. B., Marklund, E., Schaepe, J. M., Aditham, A. K., Shah, N., Suzuki, P. H., Shrikumar, A., Afek, A., Greenleaf, W. J., Gordân, R., Zeitlinger, J., Kundaje, A., & Fordyce, P. M. (2023). Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science*, *381*(6664), eadd1250. https://doi.org/10.1126/science.add1250

Hsu, T., Gogos, J. A., Kirsh, S. A., & Kafatos, F. C. (1992). Multiple Zinc Finger Forms Resulting from Developmentally Regulated Alternative Splicing of a Transcription Factor Gene. *Science*, *257*(5078), 1946–1950. https://doi.org/10.1126/science.1411512

Hui, J., Hung, L.-H., Heiner, M., Schreiner, S., Neumüller, N., Reither, G., Haas, S. A., & Bindereif, A. (2005). Intronic CA-repeat and CA-rich elements: A new class of regulators of mammalian alternative splicing. *The EMBO Journal*, *24*(11), 1988–1998. https://doi.org/10.1038/sj.emboj.7600677

Hui, J., Stangl, K., Lane, W. S., & Bindereif, A. (2003). HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nature Structural Biology*, *10*(1), Article 1. https://doi.org/10.1038/nsb875

Jacob, H. J., Lindpaintner, K., Lincoln, S. E., Kusumi, K., Bunker, R. K., Mao, Y.-P., Ganten, D., Dzau, V. J., & Lander, E. S. (1991). Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. *Cell*, *67*(1), 213–224. https://doi.org/10.1016/0092-8674(91)90584-L

Jagannathan, M., Cummings, R., & Yamashita, Y. M. (2018). A conserved function for pericentromeric satellite DNA. *eLife*, *7*, e34122. https://doi.org/10.7554/eLife.34122

Jagannathan, M., Cummings, R., & Yamashita, Y. M. (2019). The modular mechanism

of chromocenter formation in Drosophila. *eLife*, *8*, e43938. https://doi.org/10.7554/eLife.43938

Kang, H., McElroy, K. A., Jung, Y. L., Alekseyenko, A. A., Zee, B. M., Park, P. J., & Kuroda, M. I. (2015). Sex comb on midleg (Scm) is a functional link between PcG-repressive complexes in Drosophila. *Genes & Development*, *29*(11), 1136–1150. https://doi.org/10.1101/gad.260562.115

Kaye, E. G., Booker, M., Kurland, J. V., Conicella, A. E., Fawzi, N. L., Bulyk, M. L., Tolstorukov, M. Y., & Larschan, E. (2018). Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the Drosophila Dosage Compensation Complex to the Male X Chromosome. *Cell Reports*, *22*(12), 3227–3239. https://doi.org/10.1016/j.celrep.2018.02.098

Kim, J., Sebring, A., Esch, J. J., Kraus, M. E., Vorwerk, K., Magee, J., & Carroll, S. B. (1996). Integration of positional signals and regulation of wing formation and identity by Drosophila vestigial gene. *Nature*, *382*(6587), 133–138. https://doi.org/10.1038/382133a0

Krishnan, J., Athar, F., Rani, T. S., & Mishra, R. K. (2017). Simple sequence repeats showing “length preference” have regulatory functions in humans. *Gene*, *628*, 156–161. https://doi.org/10.1016/j.gene.2017.07.022

Kumar, R. P., Krishnan, J., Pratap Singh, N., Singh, L., & Mishra, R. K. (2013). GATA simple sequence repeats function as enhancer blocker boundaries. *Nature Communications*, *4*(1), Article 1. https://doi.org/10.1038/ncomms2872

Levinger, L. F. (1985). D1 protein of Drosophila melanogaster. Purification and AT-DNA binding properties. *Journal of Biological Chemistry*, *260*(26), 14311–14318. https://doi.org/10.1016/S0021-9258(17)38719-7

Litt, M., & Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics*, *44*(3), 397–401.

Meloni, R., Albanèse, V., Ravassard, P., Treilhou, F., & Mallet, J. (1998). A Tetranucleotide Polymorphic Microsatellite, Located in the First Intron of the Tyrosine Hydroxylase Gene, Acts as a Transcription Regulatory Element in Vitro. *Human Molecular Genetics*, *7*(3), 423–428. https://doi.org/10.1093/hmg/7.3.423

Miller, J. W., Urbinati, C. R., Teng-umnuay, P., Stenberg, M. G., Byrne, B. J., Thornton, C. A., & Swanson, M. S. (2000). Recruitment of human muscleblind proteins to (CUG)n expansions associated with myotonic dystrophy. *The EMBO Journal*, *19*(17), 4439–4448. https://doi.org/10.1093/emboj/19.17.4439

Mills, W. K., Lee, Y. C. G., Kochendoerfer, A. M., Dunleavy, E. M., & Karpen, G. H. (2019). RNA from a simple-tandem repeat is required for sperm maturation and male fertility in Drosophila melanogaster. *eLife*, *8*, e48940. https://doi.org/10.7554/eLife.48940

Muravyova, E., Golovnin, A., Gracheva, E., Parshikov, A., Belenkaya, T., Pirrotta, V., & Georgiev, P. (2001). Loss of Insulator Activity by Paired Su(Hw) Chromatin Insulators. *Science*, *291*(5503), 495–498. https://doi.org/10.1126/science.291.5503.495

Napoletano, F., Ferrari Bravo, G., Voto, I. A. P., Santin, A., Celora, L., Campaner, E., Dezi, C., Bertossi, A., Valentino, E., Santorsola, M., Rustighi, A., Fajner, V., Maspero, E., Ansaloni, F., Cancila, V., Valenti, C. F., Santo, M., Artimagnella, O. B., Finaurini, S., … Del Sal, G. (2021). The prolyl-isomerase PIN1 is essential for nuclear Lamin-B structure and function and protects heterochromatin under mechanical stress. *Cell Reports*, *36*(11), 109694. https://doi.org/10.1016/j.celrep.2021.109694

Oberlé, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boué, J., Bertheas, M. F., & Mandel, J. L. (1991). Instability of a 550-Base Pair DNA Segment and Abnormal Methylation in Fragile X Syndrome. *Science*, *252*(5009), 1097–1102. https://doi.org/10.1126/science.252.5009.1097

Pai, C.-Y., Lei, E. P., Ghosh, D., & Corces, V. G. (2004). The Centrosomal Protein CP190 Is a Component of the gypsy Chromatin Insulator. *Molecular Cell*, *16*(5), 737–748. https://doi.org/10.1016/j.molcel.2004.11.004

Park, J. W., Parisky, K., Celotto, A. M., Reenan, R. A., & Graveley, B. R. (2004). Identification of alternative splicing regulators by RNA interference in Drosophila. *Proceedings of the National Academy of Sciences*, *101*(45), 15974–15979. https://doi.org/10.1073/pnas.0407004101

Pathak, R., Mamillapalli, A., Rangaraj, N., Kumar, R., Vasanthi, D., Mishra, K., &

Mishra, R. K. (2013). AAGAG repeat RNA is an essential component of nuclear matrix in Drosophila. *RNA Biology*, *10*(4), 564–571. https://doi.org/10.4161/rna.24326

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., Ong, C.-T., Hookway, T. A., Guo, C., Sun, Y., Bland, M. J., Wagstaff, W., Dalton, S., McDevitt, T. C., Sen, R., Dekker, J., Taylor, J., & Corces, V. G. (2013). Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell*, *153*(6), 1281–1295. https://doi.org/10.1016/j.cell.2013.04.053

Ramamoorthy, S., Garapati, H. S., & Mishra, R. K. (2014). Length and sequence dependent accumulation of simple sequence repeats in vertebrates: Potential role in genome organization and regulation. *Gene*, *551*(2), 167–175. https://doi.org/10.1016/j.gene.2014.08.052

Ray, P., De, S., Mitra, A., Bezstarosti, K., Demmers, J. A. A., Pfeifer, K., & Kassis, J. A. (2016). Combgap contributes to recruitment of Polycomb group proteins in Drosophila. *Proceedings of the National Academy of Sciences*, *113*(14), 3826–3831. https://doi.org/10.1073/pnas.1520926113

Shah, A. D., Goode, R. J. A., Huang, C., Powell, D. R., & Schittenhelm, R. B. (2020). LFQ-Analyst: An Easy-To-Use Interactive Web Platform To Analyze and Visualize Label-Free Proteomics Data Preprocessed with MaxQuant. *Journal of Proteome Research*, *19*(1), 204–211. https://doi.org/10.1021/acs.jproteome.9b00496

Spada, A. R. L., Wilson, E. M., Lubahn, D. B., Harding, A. E., & Fischbeck, K. H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, *352*(6330), Article 6330. https://doi.org/10.1038/352077a0

Spana, C., Harrison, D. A., & Corces, V. G. (1988). The Drosophila melanogaster suppressor of Hairy-wing protein binds to specific sequences of the gypsy retrotransposon. *Genes & Development*, *2*(11), 1414–1423. https://doi.org/10.1101/gad.2.11.1414

Srivastava, S., Avvaru, A. K., Sowpati, D. T., & Mishra, R. K. (2019). Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics*, *20*(1),

153. https://doi.org/10.1186/s12864-019-5516-5

Staib, P., Kretschmar, M., Nichterlein, T., Hof, H., & Morschhäuser, J. (2002). Host versus in vitro signals and intrastrain allelic differences in the expression of a Candida albicans virulence gene. *Molecular Microbiology*, *44*(5), 1351–1366. https://doi.org/10.1046/j.1365-2958.2002.02967.x

Subramanian, S., Mishra, R. K., & Singh, L. (2003). Genome-wide analysis of microsatellite repeats in humans: Their abundance and density in specific genomic regions. *Genome Biology*, *4*(2), R13. https://doi.org/10.1186/gb-2003-4-2-r13

Sun, X., Morozova, T., & Sonnenfeld, M. (2006). Glial and Neuronal Functions of the Drosophila Homolog of the Human SWI/SNF Gene ATR-X (DATR-X) and the jing Zinc-Finger Gene Specify the Lateral Positioning of Longitudinal Glia and Axons. *Genetics*, *173*(3), 1397–1415. https://doi.org/10.1534/genetics.106.057893

Todo, T., Ryo, H., Yamamoto, K., Toh, H., Inui, T., Ayaki, H., Nomura, T., & Ikenaga, M. (1996). Similarity Among the Drosophila (6-4)Photolyase, a Human Photolyase Homolog, and the DNA Photolyase-Blue-Light Photoreceptor Family. *Science*, *272*(5258), 109–112. https://doi.org/10.1126/science.272.5258.109

Vasanthi, D., Anant, M., Srivastava, S., & Mishra, R. K. (2010). A functionally conserved boundary element from the mouse HoxD locus requires GAGA factor in Drosophila. *Development*, *137*(24), 4239–4247. https://doi.org/10.1242/dev.058701

Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S., Fu, Y.-H., Kuhl, D. P. A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., Zhang, F., Eussen, B. E., van Ommen, G.-J. B., Blonden, L. A. J., Riggins, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Thomas Caskey, C., Nelson, D. L., … Warren, S. T. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, *65*(5), 905–914. https://doi.org/10.1016/0092-8674(91)90397-H

Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., & Verstrepen, K. J. (2009). Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability. *Science*, *324*(5931), 1213–1216. https://doi.org/10.1126/science.1170097

Warrick, J. M., Paulson, H. L., Gray-Board, G. L., Bui, Q. T., Fischbeck, K. H., Pittman, R. N., & Bonini, N. M. (1998). Expanded Polyglutamine Protein Forms Nuclear Inclusions and Causes Neural Degeneration in Drosophila. *Cell*, *93*(6), 939–949. https://doi.org/10.1016/S0092-8674(00)81200-3

Weber, J. L., & May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics*, *44*(3), 388–396.

Weiser, J. N., Love, J. M., & Moxon, E. R. (1989). The molecular mechanism of phase variation of H. influenzae lipopolysaccharide. *Cell*, *59*(4), 657–665. https://doi.org/10.1016/0092-8674(89)90011-1

Williams, J. A., Atkin, A. L., & Bell, J. B. (1990). The functional organization of the vestigial locus in Drosophila melanogaster. *Molecular and General Genetics MGG*, *221*(1), 8–16. https://doi.org/10.1007/BF00280361

Williams, J. A., Bell, J. B., & Carroll, S. B. (1991). Control of Drosophila wing and haltere development by the nuclear vestigial gene product. *Genes & Development*, *5*(12b), 2481–2495. https://doi.org/10.1101/gad.5.12b.2481

Xu, W., Liu, L., Emson, P. C., Harrington, C. R., & Charles, I. G. (1997). Evolution of a homopurine-homopyrimidine pentanucleotide repeat sequence upstream of the human inducible nitric oxide synthase gene. *Gene*, *204*(1–2), 165–170. https://doi.org/10.1016/s0378-1119(97)00538-6

Zhang, K., Donnelly, C. J., Haeusler, A. R., Grima, J. C., Machamer, J. B., Steinwald, P., Daley, E. L., Miller, S. J., Cunningham, K. M., Vidensky, S., Gupta, S., Thomas, M. A., Hong, I., Chiu, S.-L., Huganir, R. L., Ostrow, L. W., Matunis, M. J., Wang, J., Sattler, R., … Rothstein, J. D. (2015). The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. *Nature*, *525*(7567), Article 7567. https://doi.org/10.1038/nature14973

Table 1: List of enriched Transcription factors enriched across various repeats along with their DNA binding domain indicated in brackets

Repeat	**Transcription factors enriched across repeats**					
$AAAAG_{11}$	D1 (AT-hook)	Dip3 (MADF-BESS)				
$AAAAT_{8}$	D1 (AT-hook)	Jigr1 (MADF-BESS)	Jarid2 (ARID)	Tet (CXXC-Zn)		
$AAAT_{10}$	D1 (AT-hook)	Jigr1 (MADF-BESS)	CC9/CG8924/Q9VXL5 (Pipsqueak TF)			
AAT_{14}	D1 (AT-hook)	Jigr1 (MADF-BESS)	Mef2-RC (MADS-BOX)	CC9/CG8924/Q9VXL5 (Pipsqueak TF)		
$AATAT_{9}$	D1(AT-hook)	Jigr1 (MADF-BESS)	Jarid2 (ARID)	Tet (CXXC-Zn)	Q9VXQ6/CG8944 (C2H2 Zinc finger and MADF domain)	Abd-B(homeobox TF)
$ACATAT_{8}$	Cf2 (C2H2 Zn-finger)	Q9VMM3/CG7239 (Myb-SANT like DNA binding domain/MADF)	Prod (Myb-SANT like DNA binding domain)	Cg (C2H2 Zn finger)	Ibf1 (BED finger)	M9PB45/CG14005 (Myb-SANT like DNA binding domain)
$AGAT_{10}$	arm	Diminutive(bHLH TF)				
$AAAG_{13}$	D1(AT-hook)	CG1815/Zmynd 8 (Zinc finger)	Diminutive(bHLH TF)	Lid (ARID)	Dip3 (MADF-BESS)	

Figure legends:

Figure 1: Electrophoretic mobility shift assay (EMSA) of length enriched repeats - Radiolabelled repeats (mentioned below their respective gels) showed a shift in the presence of nuclear extract from 0-16 h *Drosophila* embryos (lane 2, all gels), these shifts were competed out when unlabelled repeats identical to radiolabelled probes were added at 10- or 20- or 50-fold higher concentration than radiolabelled probe (lanes 3-5, all gels). However, the shifts were either slightly or not competed at all even when 10- or 20- or 50-fold higher concentration than radiolabelled probe of unlabelled repeats of a dissimilar kind was added (lanes 6-8 of all gels except M, lanes 9-11 of all gels except D, E, F, G, H, and M) whereas the shifts were competed out when 10- or 20- or 50-fold higher concentration than radiolabelled probe of unlabelled repeats of a similar kind was added (lanes 9-11 of gels D, F, G and H) suggesting that all these radiolabelled probes except $ATCC_9$ show sequence-specific DNA binding activity.

Figure 2: Proteomics of repeat associated proteins - Volcano plots depicting the enriched proteins (dots represented by gene names) in the repeat pull-down. The criteria for enrichment: $\log_2$(LFQ intensity fold change over control) > 1.5 and $-\log_{10}$(adj. p-value) > 2.

Figure 3: Competition EMSA of $AAAT_{10}$ radiolabelled probe-

A. Alignment of AAT_{14}, $AAAT_{10}$, $AAAAT_8$ and $AATAT_9$ shows the recurrence of AATAA motifs in various repeats. Alternate motifs of $AAAAT_8$ and $AATAT_9$ have been bolded to differentiate between consecutive motifs. AATAA motifs have been underlined. AATAT motifs of $AATAT_9$ differ from AATAA motifs by just one nucleotide (T) represented in red color.

B-G. Competition EMSA of $AAAT_{10}$ radiolabelled probe with higher concentrations of unlabelled probe (10 or 20 or 50-fold), showed that the amount of free probe released when 50-fold higher concentration of unlabelled repeats was added was high for $AAAT_{10}$, moderate for AAT_{14} and least for $AAAAT_8$ (lanes 5,8 and 11 of B). The same result was obtained when radiolabelled AAT_{14} or $AAAAT_8$ was used (C and D).

However, in all the cases we observed that 50-fold higher concentration of ATC_{12} did not compete out the shift (lane 13, all gels). Competition of radiolabelled $AAAT_{10}$ with 50-fold higher concentration unlabelled probes of various kinds showed that $AATAT_9$ (G, lane 5) completely competed out the shift whereas $AAAAG_{11}$ (E, lane 11), $AAAG_{13}$ (E, lane 8) were able to compete the shift only moderately. The shifts were either slightly or not competed out by $AATAG_{11}$ (G, lane 8), $AACAT_{10}$ (G, lane 11), AAG_{19} (E, lane 5), $ACATAT_8$ (F, lane 5), A_{36} (F, lane 8) and $AGAT_{10}$ (F, lane 11).

H & I. Plots showing LFQ intensities of D1 and Jigr1 across various repeat-based proteomic pull-downs.

Figure 4: **Development of novel enhancer blocker assay:**

A. Line diagram depicting the *vestigial* locus with its well-characterized DV-enhancer (red box) and quadrant enhancer (yellow box). The wing imaginal disc on the right shows the areas of the wing pouch where the DV-enhancer (red lines) and quadrant enhancer (yellow dots) are active.
B. Schematic of recombination between the attP sites of DV-enhancer attP line and the attB sites of pBSKS-attB1-gypsy-attB2 plasmid. The DV-enhancer attP line has attP1-3XP3GFP-attP2 insertion upstream of the DV-enhancer (red box). The attP sites and 3XP3GFP are represented by orange triangles and green rectangle respectively. The pBSKS-attB1-gypsy-attB2 plasmid is depicted with gypsy and attBs as black rectangle and blue triangles respectively. The plasmid also harbors loxP sites (violet triangles) for gypsy excision. The recombination between attP and attB sites results in RMCE, which ultimately leads to GFP loss in the DV-gypsy line.
C. Homozygous DV-gypsy fly shows absence of wings and halteres (top left) and DAPI stained wing imaginal discs from the 3rd instar larvae of the same genotype showing abnormal morphology (bottom left). DV-gypsy homozygous fly homozygous for *su(Hw)* mutation shows the rescue of wings (top right) and DAPI stained wing imaginal discs from the 3rd instar larvae of the same genotype showing normal morphology (bottom right).

D. Percentage of progenies of *DV-gypsy/CyO* and *DV-gypsy/DV-gypsy* when *DV-gypsy/CyO* flies were selfed.
E. Percentage of progenies of (i) *DV-gypsy/CyO; su(Hw)[8]/TM6B* (ii) *DV-gypsy/CyO; su(Hw)[8]/su(Hw)[8]* (iii) *DV-gypsy/DV-gypsy; su(Hw)[8]/TM6B* (iv) *DV-gypsy/DV-gypsy; su(Hw)[8]/su(Hw)[8]* when *DV-gypsy/CyO; su(Hw)[8]/TM6B* flies were selfed.
F. Percentage of progenies of (i) *DV-gypsy/CyO; mod(mdg4)[ul]/TM6B* (ii) *DV-gypsy/CyO; mod(mdg4)[ul]/mod(mdg4)[ul]* (iii) *DV-gypsy/DV-gypsy; mod(mdg4)[ul]/TM6B* (iv) *DV-gypsy/DV-gypsy; mod(mdg4)[ul]/mod(mdg4)[ul]* when *DV-gypsy/CyO; mod(mdg4)[ul]/TM6B* flies were selfed.
G. Percentage of progenies of (i) *DV-gypsy/CyO; Cp190[H4-1]/TM6B* (ii) *DV-gypsy/DV-gypsy; Cp190[H4-1]/TM6B* when *DV-gypsy/CyO; Cp190[H4-1]/TM6B* flies were selfed.

Figure 5: Repeats function as enhancer blockers upon overexpression of D1 in the DV border cells using C96Gal4:

A. Line diagram showing the repeat integrated DV-enhancer attP line (DV-repeat line).
B. Wing phenotypes observed upon overexpression of D1 in the homozygous repeat background: T1- wildtype; T2- loss of bristles or <3 notches or both; T3- 3 or more notches; T4- decreased wing length with disorganized arrangement of cells; T5- small sized wings
C. Percentage of different wing phenotypes across various genotypes: homozygous DV-repeat line, C96Gal4>UASD1 was the control and D1 overexpression in homozygous DV-repeat lines (DV-repeat; C96Gal4>UASD1).

Acknowledgments: We thank Dr. V. Bharathi for wing imaginal disc dissections and immunostaining. We would also like to thank Dr. V. Bharathi, Dr. Sonal Nagarkar Jaiswal for their inputs and suggestions. We would like to thank present and past members of RKM lab for their inputs and suggestions. Fly embryonic injections were performed at C-CAMP, Bengaluru.

Data Availability: "The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD058086".

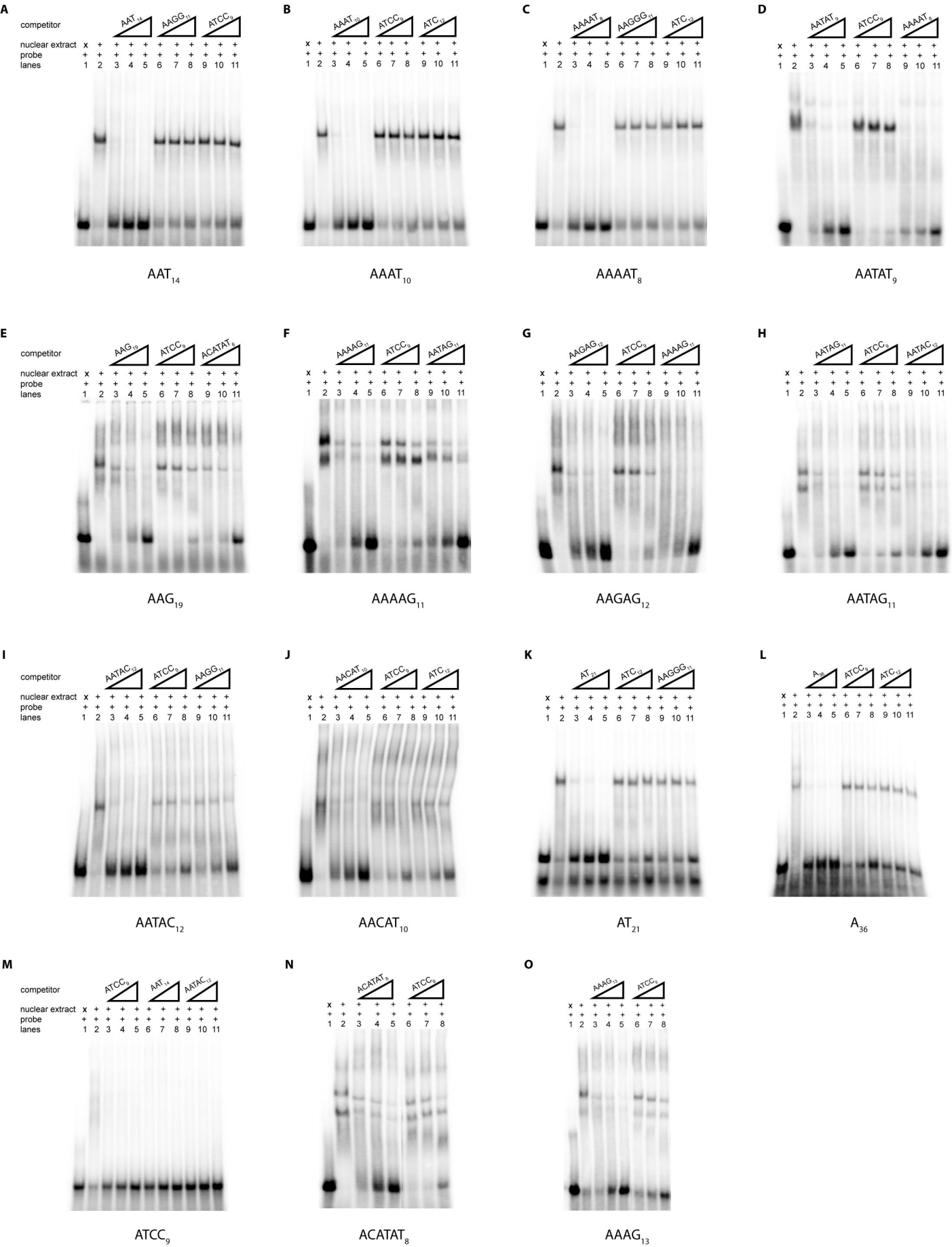

A
competitor
AAT14
AAGG11
ATCC9
nuclear extract x + + + + + + + + + +
probe + + + + + + + + + +
lanes 1 2 3 4 5 6 7 8 9 10 11
AAT14
B
AAAT10
ATCC9
ATC12
x + + + + + + + + + +
+ + + + + + + + + +
1 2 3 4 5 6 7 8 9 10 11
AAAT10
C
AAAAT8
AAGGG11
ATC12
x + + + + + + + + + +
+ + + + + + + + + +
1 2 3 4 5 6 7 8 9 10 11
AAAAT8
D
AATAT9
ATCC9
AAAAT8
x + + + + + + + + + +
+ + + + + + + + + +
1 2 3 4 5 6 7 8 9 10 11
AATAT9
E
competitor
AAG19
ATCC9
ACATAT8
nuclear extract x + + + + + + + + + +
probe + + + + + + + + + +
lanes 1 2 3 4 5 6 7 8 9 10 11
AAG19
F
AAAAG11
ATCC9
AATAG11
x + + + + + + + + + +
+ + + + + + + + + +
1 2 3 4 5 6 7 8 9 10 11
AAAAG11
G
AAGAG12
ATCC9
AAAAG11
x + + + + + + + + + +
+ + + + + + + + + +
1 2 3 4 5 6 7 8 9 10 11
AAGAG12
H
AATAG11
ATCC9
AATAC12
x + + + + + + + + + +
+ + + + + + + + + +
1 2 3 4 5 6 7 8 9 10 11
AATAG11
I
competitor
AATAC12
ATCC9
AAGG11
nuclear extract x + + + + + + + + + +
probe + + + + + + + + + +
lanes 1 2 3 4 5 6 7 8 9 10 11
AATAC12
J
AACAT10
ATCC9
ATC12
x + + + + + + + + + +
+ + + + + + + + + +
1 2 3 4 5 6 7 8 9 10 11
AACAT10
K
AT21
ATC12
AAGGG11
x + + + + + + + + + +
+ + + + + + + + + +
1 2 3 4 5 6 7 8 9 10 11
AT21
L
A36
ATCC9
ATC12
x + + + + + + + + + +
+ + + + + + + + + +
1 2 3 4 5 6 7 8 9 10 11
A36
M
competitor
ATCC9
AAT14
AATAC12
nuclear extract x + + + + + + + + + +
probe + + + + + + + + + +
lanes 1 2 3 4 5 6 7 8 9 10 11
ATCC9
N
ACATAT8
ATCC9
x + + + + + + + + +
+ + + + + + + + +
1 2 3 4 5 6 7 8
ACATAT8
O
AAAG13
ATCC9
x + + + + + + + +
+ + + + + + + + +
1 2 3 4 5 6 7 8
AAAG13

Figure 1

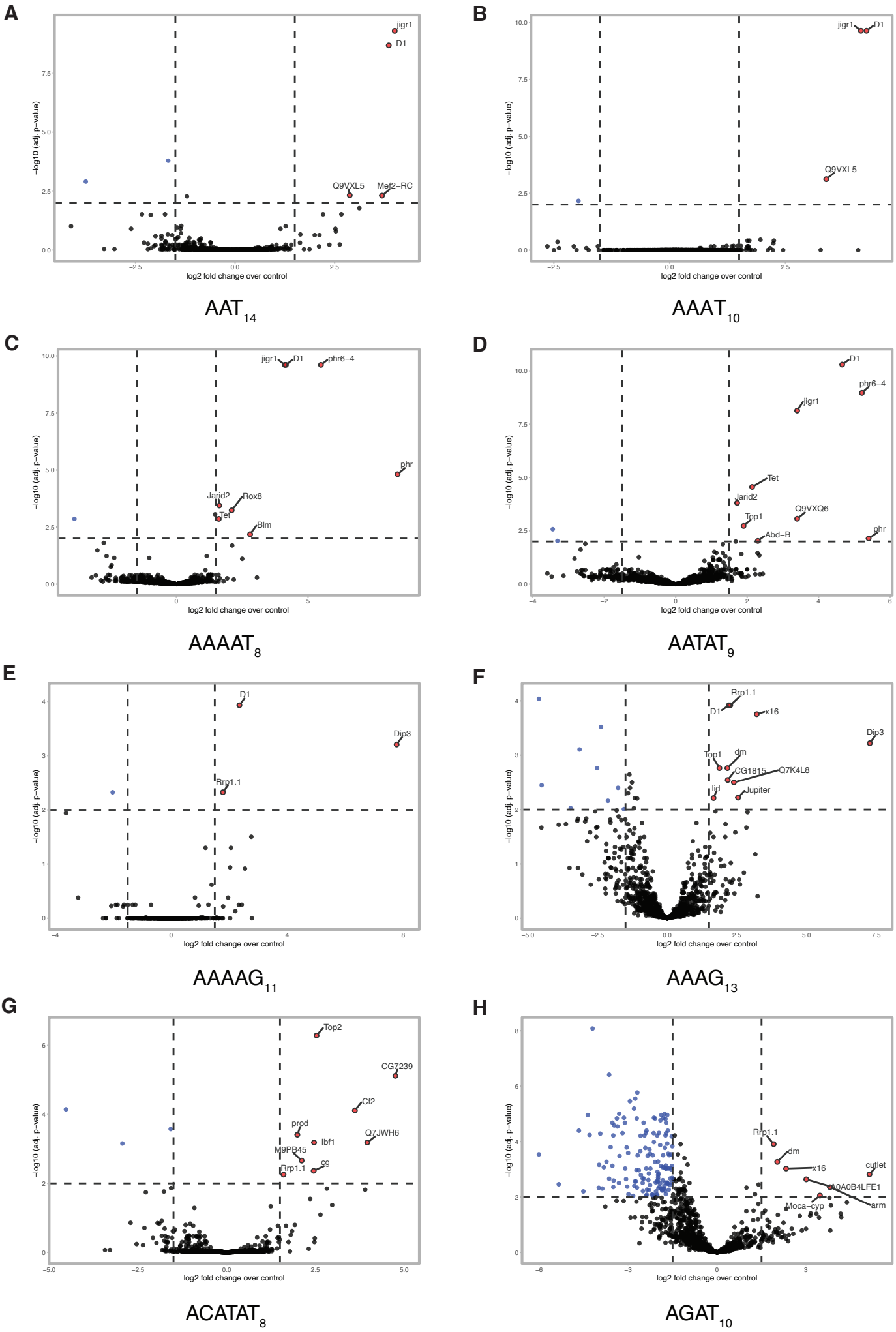

A
jigr1
D1
Q9VXL5
Mef2−RC
−log10 (adj. p−value)
log2 fold change over control
AAT14
B
jigr1
D1
Q9VXL5
−log10 (adj. p−value)
log2 fold change over control
AAAT10
C
jigr1
D1
phr6−4
phr
Jarid2
Rox8
Tet
Blm
−log10 (adj. p−value)
log2 fold change over control
AAAAT8
D
D1
phr6−4
jigr1
Tet
Jarid2
Q9VXQ6
Top1
Abd−B
phr
−log10 (adj. p−value)
log2 fold change over control
AATAT9
E
D1
Dip3
Rrp1.1
−log10 (adj. p−value)
log2 fold change over control
AAAAG11
F
Rrp1.1
D1
x16
Dip3
Top1
dm
CG1815
Q7K4L8
lid
Jupiter
−log10 (adj. p−value)
log2 fold change over control
AAAG13
G
Top2
CG7239
Cf2
prod
Q7JWH6
Ibf1
M9PB45
cg
Rrp1.1
−log10 (adj. p−value)
log2 fold change over control
ACATAT8
H
Rrp1.1
dm
x16
cutlet
A0A0B4LFE1
Moca−cyp
arm
−log10 (adj. p−value)
log2 fold change over control
AGAT10

Figure 2

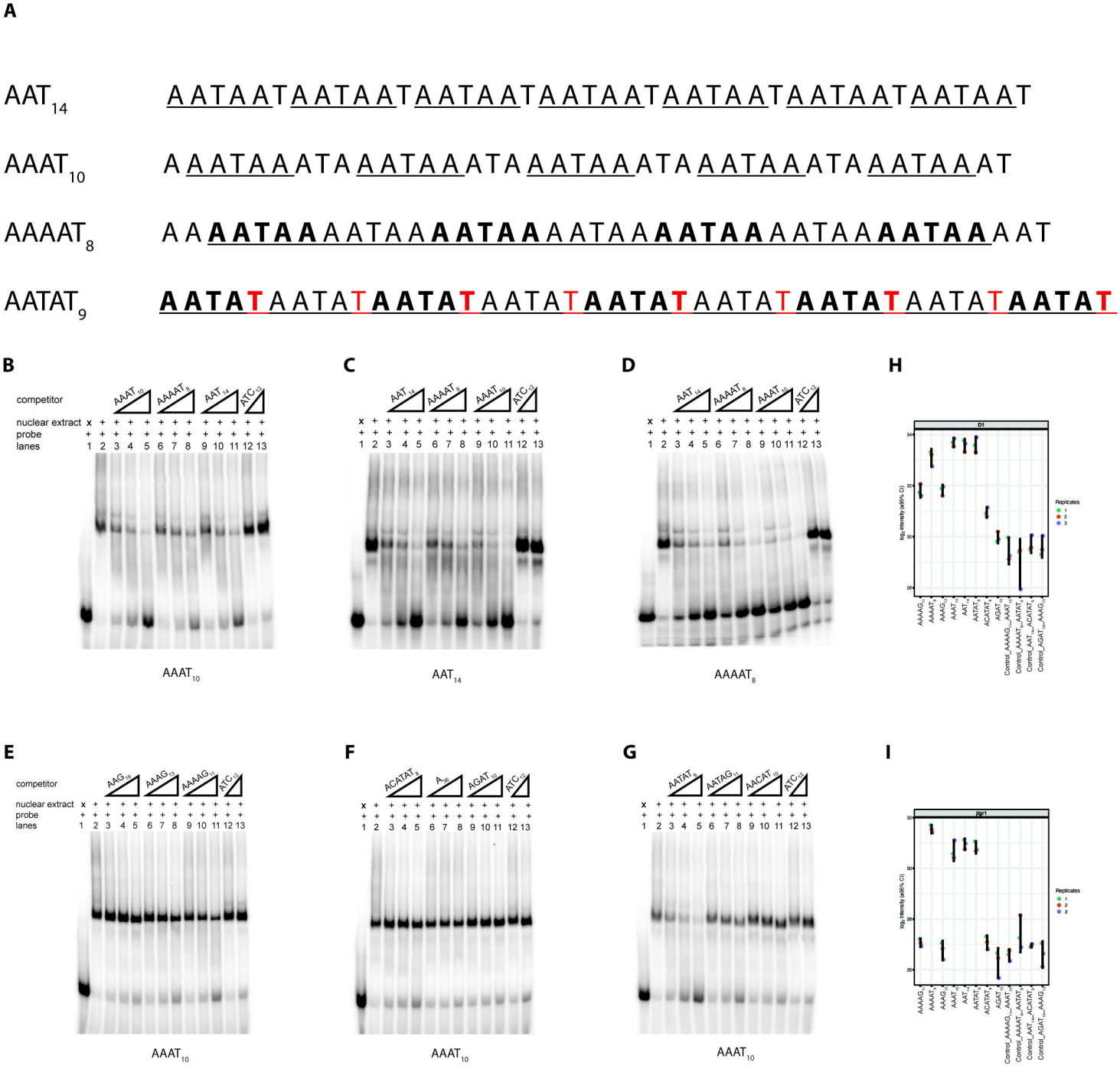

A
AAT14
AAAT10
AAAAT8
AATAT9
B
C
D
H
E
F
G
I
competitor
nuclear extract
probe
lanes
AAAT10
AAT14
AAAAT8
AAAT10
AAAT10
AAAT10
AAAT10
D1
jigr1
Replicates

Figure 3

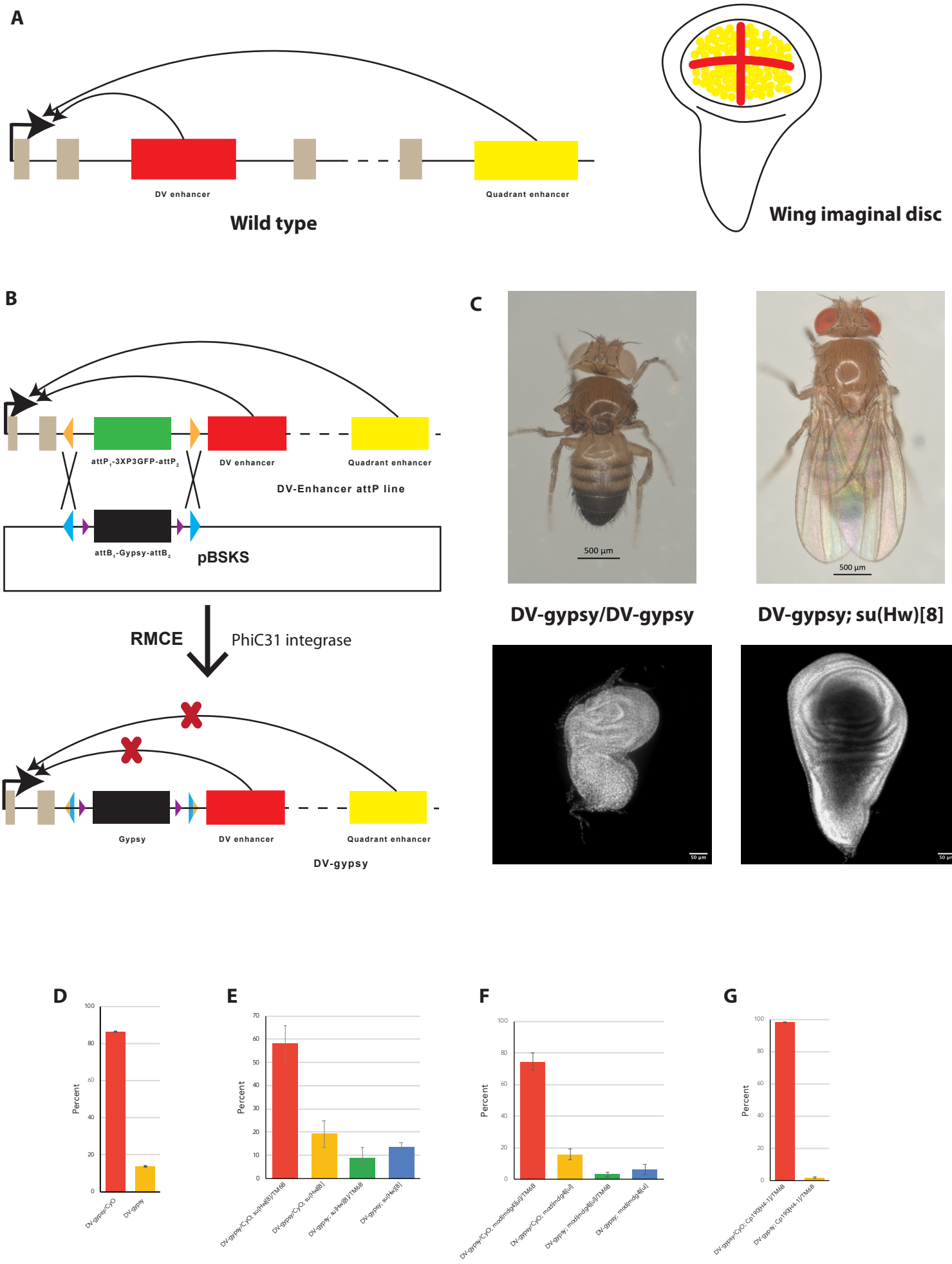
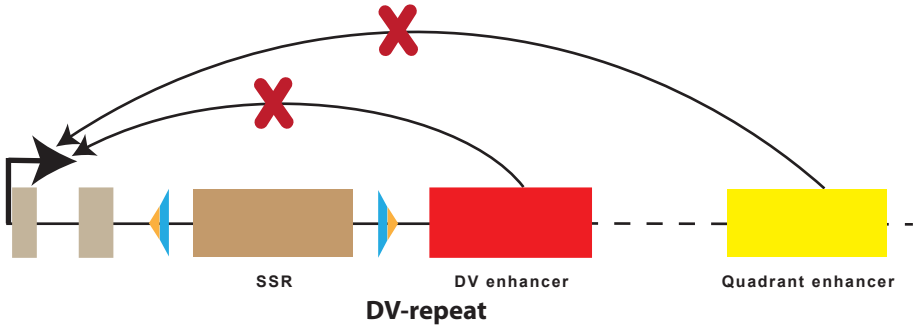

A
DV enhancer
Quadrant enhancer
Wild type
Wing imaginal disc
B
attP1-3XP3GFP-attP2
DV enhancer
Quadrant enhancer
DV-Enhancer attP line
attB1-Gypsy-attB2
pBSKS
RMCE
PhiC31 integrase
Gypsy
DV enhancer
Quadrant enhancer
DV-gypsy
C
500 µm
500 µm
DV-gypsy/DV-gypsy
DV-gypsy; su(Hw)[8]
D
Percent
DV-gypsy/CyO
DV-gypsy
E
Percent
DV-gypsy/CyO, su(Hw)[8]/TM6B
DV-gypsy/CyO, su(Hw)[8]
DV-gypsy; su(Hw)[8]/TM6B
DV-gypsy; su(Hw)[8]
F
Percent
DV-gypsy/CyO; mod(mdg4)[u]/TM6B
DV-gypsy/CyO; mod(mdg4)[u]
DV-gypsy; mod(mdg4)[u]/TM6B
DV-gypsy; mod(mdg4)[u]
G
Percent
DV-gypsy/CyO; Cp190[H4-1]/TM6B
DV-gypsy; Cp190[H4-1]/TM6B

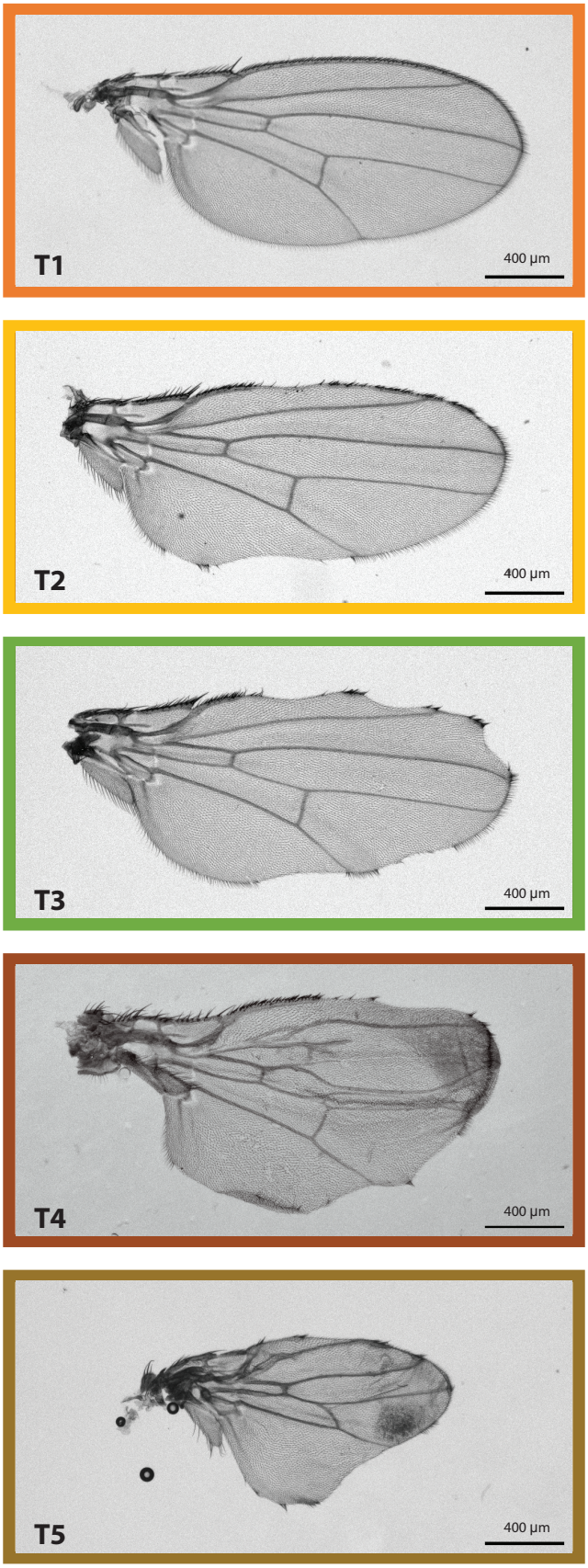
Figure 4

Figure 5

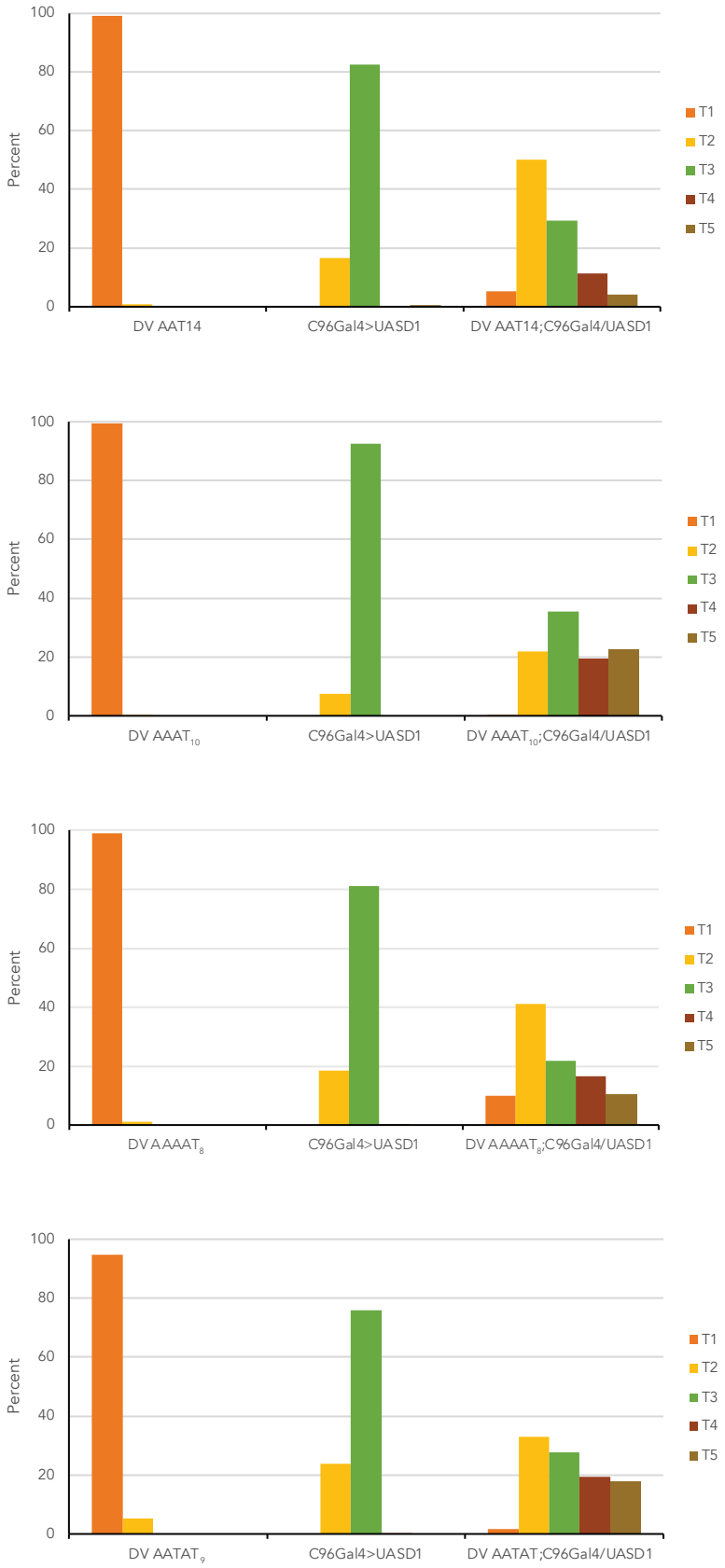
A


SSR
DV enhancer
Quadrant enhancer
DV-repeat

B


T1
400 µm
T2
400 µm
T3
400 µm
T4
400 µm
T5
400 µm

C


Percent
100
80
60
40
20
0
DV AAT14
C96Gal4>UASD1
DV AAT14;C96Gal4/UASD1
DV AAAT10
DV AAAT10;C96Gal4/UASD1
DV AAAAT8
DV AAAAT8;C96Gal4/UASD1
DV AATAT9
DV AATAT;C96Gal4/UASD1
T1
T2
T3
T4
T5