Supplemental Information

Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the *Drosophila* Dosage Compensation Complex to the Male X Chromosome

Emily G. Kaye, Matthew Booker, Jesse V. Kurland, Alexander E. Conicella, Nicolas L. Fawzi, Martha L. Bulyk, Michael Y. Tolstorukov, and Erica Larschan
SUPPLEMENTAL METHODS

Immunostaining of Polytenes:
Blocking was performed in 0.5% Bovine Serum Albumin in phosphate buffered saline (1X PBS) for 1 hr. Primary antibodies were incubated overnight at 4˚C. Slides were then washed in 1X PBS three times for 15 min each. Slides were incubated in secondary antibodies for 2 hr at 22˚C in a dark humid chamber. Secondary antibodies were donkey anti-rabbit Alexafluor 488 and donkey anti-goat Alexafluor 594 (Thermofisher Scientific). Slides were counterstained with Hoechst for 10 sec, then washed three times in 1X PBS for 10 min. Slides were mounted with Prolong Gold mounting medium and imaged on a Zeiss Axio Imager M1 Epifluorescence upright microscope using AxioVision version 4.8.2 software.

PEV Eye pigmentation assay:
Flies were put into 2.0 mL microcentrifuge tubes and flash frozen in liquid nitrogen. Frozen flies were decapitated by agitation, or razor blade if needed. Five heads were added to a fresh tube containing a steel bead and 250 μL pigment assay buffer (3% HCl in ethanol). Fly heads were then homogenized for 1 min at a frequency of 30 per second using a bead mill mixer (Retsch MM300). Heads were incubated at 50˚C for 10 minutes, then centrifuged at max speed for 10 minutes at 22˚C. ~200 μL of supernatant was transferred to a fresh tube. 150 μL was then transferred to a clear 96-well plate, absorbance at 480 nm was measured on a plate reader. A pigment assay buffer-only well absorbance value was subtracted from sample-well readings.

MBP-GAF Cloning primers and purification buffers:
For the MBP-GAF DBD protein, restriction sites were added on through PCR with the following forward primer (NdeI): 5´-cttcaggtccatatgAGTGGTAGTGTGCAGCAG-3´ and reverse primer (XhoI): 5´-gtgtgtgctgcatATGCACCACACTACCAAC-3´. For expression, cells were grown in LB incubated at 37˚C with shaking until reaching an OD of ~0.6-0.9. Expression was then induced for 4 hours by adding IPTG to a final concentration of 1 mM and incubating at 37˚C with
shaking. Cells were pelleted at 6,000 rpm and frozen in liquid nitrogen for storage at -80°C. Pellet of cells expressing MBP-GAF was resuspended in buffer A (20 mM Tris-Cl, 500 mM NaCl, 10 mM imidazole, 1 mM DTT, pH 8.0, 0.1 mM ZnCl). Cells were mechanically lysed with an EmulsiFlex C3 (Avestin) and centrifuged at 20,000 rpm for 1 hr. Lysate was then filtered through a 0.22 μm filter and run on a 5 mL HisTrap HP column (GE Healthcare). For column elution, buffer B was used (20 mM Tris-Cl, 500 mM NaCl, 500 mM imidazole, 1 mM DTT, pH 8.0, 0.1 mM ZnCl). Fractions were collected and analyzed by SDS-PAGE. Fractions containing protein were combined and further purified by size-exclusion chromatography (SEC) using a Superdex 200 (GE Healthcare) equilibrated with SEC buffer (20 mM Tris-Cl, 1 mM DTT, 150 mM NaCl, pH 8.0, 0.1 mM ZnCl). Fractions of 1 mL were collected and analyzed via SDS-PAGE. Fractions containing MBP-GAF DBD were combined and concentrated to 40 μM. Aliquots were flash frozen in liquid nitrogen and stored at -80°C until use.

**GST-GAF cloning**

The following primers were used to add attB primers for cloning the GAF DBD cDNA into the pDONR221 gateway vector (Invitrogen): GAF attB Forward primer (5’ to 3’):

GGGGACAAGTTTGTACAAAAAAGCAGGCTTCAGTGGTAGTGTGCAGCAG

GAF attB Reverse primer (5’ to 3’):

GGGGACCACTTTGTACAAAAAAGCAGGCTTCAGTGGTAGTGTCAGCAG

The Gateway cloning kit was used to perform the BP Clonase II reaction for cloning the PCR product into pDONR221, and then the LR Clonase II enzyme was used to clone into pDEST15.

**Computational methods:**

**Sequencing data alignment and initial processing**

Sequencing reads were mapped against FlyBase reference assembly (release 5.22) of *D. melanogaster* genome with bowtie (v. 0.12.9). Only uniquely mapped reads were retained for
further analysis (option -m 1). Alignment statistics are reported in Table S1. For each replicate of each ChIP experiment, genome-wide signal tracks of enrichment values (for GAF, CLAMP, and MSL3 binding data) were generated using utility bdgcmp from MACS2 suite with -m FE option to compare treatment over input (Zhang et al., 2008).

**Peak calling and classification of binding events**
Peaks in ChIP-seq profiles were called using MACS2 version 2.1.1 (Zhang et al., 2008) with default parameters independently for each biological replicate (see Table S1 for number of replicates in each condition). Narrow peaks that met FDR threshold of 0.1 were considered for GAF and CLAMP enriched data. Since MSL occupies extended regions within X chromosome rather than punctate peaks (Alekseyenko et al. 2008), the MACS2 broad peak option (--broad) and P value threshold of 0.05 was used for MSL3 enriched data.

The list of peaks identified in all replicates of CLAMP and GAF ChIP samples was scanned and the peaks with summits within 100 bp of each other were grouped together, resulting in 46,716 bound regions. Then, GAF and CLAMP enrichment scores were computed for each bound region as a maximal value in the mean enrichment profile obtained for the corresponding IP and RNAi condition within this region. The resulting peaks were included in the final high-confidence set (used in most analyses in the paper) if the GFP RNAi condition enrichment score was >=3 for CLAMP or GAF data. Peaks meeting this threshold for both CLAMP and GAF were classified as co-bound, otherwise they were classified as CLAMP or GAF only. The enrichment threshold was selected so that reproducibility of the peaks in pairwise comparison of the replicates of the same condition is above 80% (Figure S1). The peaks were considered MSL3 peaks if they overlapped with MSL3 broad peaks in replicates 1 or 2 of GFP RNAi MSL3 ChIP.
Generation of feature enrichment profiles

Enrichment profiles around ChIP-Seq peaks and CES positions were generated using replicate set enrichment profiles. Average enrichments and 95% confidence intervals were generated at each base pair in a specified region around the feature. Heat maps were generated across +/- 500 bp regions centered at GAF and/or CLAMP peak summits. Difference enrichment heat maps were generated by subtracting the IP enrichment scores of the *gfp* treated sample from the IP enrichment scores of the *Trl* or *clamp* treated sample.

Nucleosome occupancy profiles

Nucleosome occupancy profiles were obtained using published data set (Mieczkowski et al., 2016) which comprises data produced for four MNase digestions of increasing depth (MNase titration) in CLAMP-depleted and control S2 cells. The ‘pooled’ nucleosome occupancy was calculated by averaging the fragment counts, which were normalized for sequencing library size, in 3bp bins over all titration points and replicates. The set of obsTSSs is based on Start-Seq data and was obtained from (Henriques et al., 2013). TSS-proximal profiles were generated over a +/- 1000 bp window relative to the start site and represent mean nucleosome occupancy values for all obsTSSs. The profiles were additionally smoothed in 30 bp running window.

Analysis of protein-binding microarray data

Microarray scanning, quantification, and data normalization were performed using GenePix Pro ver. 6 (Axon) and masliner software (Dudley et al., 2002) as previously described (Kuzu et al., 2016). Protein Binding Microarray z-scores were generated by subtracting standard deviation of log10 signal intensity of the array from the log10 signal intensity of each spot and dividing by the mean log10 signal intensity of the array. For GAF binding, z-scores were determined from two
replicates (this study), while for previous CLAMP data (GSM2203099) a single replicate was used. The median Z-score is reported for probes represented by multiple spots. Motif analyses on PBM probes were performed using MEME (Machanick and Bailey, 2011) with the options -dna, -revcomp. Motifs on ChIP-seq peaks were generated using MEME-ChIP (Machanick and Bailey, 2011) with default parameters.

**Prediction of peak classes using machine learning**

Peaks used for classification included 3,387 GAF peaks, 1,324 CLAMP peaks, and 2,427 co-bound peaks. For non-peaks, 2427 regions of the genome outside of GAF or CLAMP peaks were randomly sampled. Because the mean width of MACS2 peaks in the ChIP-seq data was 428 bp, peak sequence was defined as the nucleotide sequence within 214 bp of the peak summit. As a control, we included in the analysis the sets of randomly selected 214 bp sequences, which were not within CLAMP or GAF peaks. The sizes of the random sets were equal to the sizes as CLAMP and GAF peak sets. Loci with sequences containing symbols other than “A”, ”C”, ”G”, or “T” were excluded from the analysis.

Machine learning was performed with the gradient boosting algorithm as implemented in R package XGBoost (Chen and Guestrin, 2016). For training, 500 peaks in each category were sampled with the remaining peaks reserved for testing. Training was performed using the “xgboost” function with parameters max.depth = 5, eta = 0.1, nrounds = 5000, subsample = 1, objective= ”multi:softprob”, and early_stopping_rounds=100. For testing, the “predict” function was used with the parameter reshape set to TRUE. All other parameters were set to their default values. The sequence features used for machine learning are described in Table S2. Statistical analyses and figure generation were performed in the R computational environment (http://r-project.org).
**Figure S1. ChIP-seq peak calling and overlap in binding sites. Related to Figure 1.** Plots of replicate 1 versus replicates 2, 3, and 4 compare the fraction of peaks replicating versus enrichment level for CLAMP (A), GAF (B), and MSL3 (C) ChIP-seq under control (gfp) RNAi. From these data, peaks were called for enrichment scores matching a fraction replicating of 0.8 or higher. D. and E. Comparing overlap of binding sites between GAF and CLAMP. Venn diagram showing number of unique GAF (red) and CLAMP (blue) peaks or overlapping peaks (purple). Peaks are separated based by location on autosomes (D) or the X chromosome (E). F. Scatter plot of maximum GAF enrichment scores (x-axis) and maximum MSL3 enrichment scores (y-axis) in the regions depicted in panel Figure 1E. MSL3 and GAF exhibit a negative correlation with a Pearson’s correlation coefficient of -0.39 with a significance of p = 3.43 × 10^{-127} using a t-distribution.
Figure S2. Binding of GAF and CLAMP across genes and different chromatin states. Related to Figure 4. A. Average enrichment profiles over gene bodies of GAF and CLAMP. Shading around lines represents 95% confidence interval. B. Chromatin state enrichment for CLAMP (left) and GAF (right). Enrichment presented as frequency compared to abundance of that state in the genome. Enriched states are highlighted in green (CLAMP) or pink (GAF). State summaries, which are based on previous state definitions (Kharchenko et al. 2011), are also listed.
Figure S3. RNAi treatment of clamp and Trl alters the binding of the CLAMP, GAF, and MSL3 proteins. Related to Figure 5. A. Western blotting for CLAMP and GAF. Three replicates of protein samples prepared from S2 cells treated with RNAi targeting gfp, clamp, or Trl. Protein was detected with either rabbit anti-CLAMP or rabbit anti-GAF antibody, as well as a mouse anti-actin loading control. * indicate specific bands. Note that this GAF antibody detects both isoforms of GAF. C-F. Screen shots from genome browser showing CLAMP and GAF ChIP-seq enrichment under control and reciprocal RNAi conditions. Binding is interdependent (C), independent (D), competitive (E), or partially dependent (F) at different sites throughout the genome. G. Boxplot of the number of occurrences of the pentamer “GAGAG” within 214 bp of GAF and CLAMP peaks under control (dark blue) or Trl RNAi (pink) conditions. H. ChiP-qPCR for GAF, CLAMP, and MSL at the CES 3′ of the roX2 gene. Samples are as in the ChIP-seq experiment. One-way ANOVA followed by Tukey multiple comparison of means was performed to test statistical differences between RNAi treatments within each IP performed. If p > .05, p-value is listed. p < 0.05 is indicated by *.
Table S1. Aligned reads from each analyzed S2 ChIP-seq sample. Related to Figure 1.

| Sample  | Reads     | Alignments   |
|---------|-----------|--------------|
| gfp Input R1 | 19,218,421 | 13,630,217   |
| gfp CIP R1    | 25,437,957 | 17,703,752   |
| gfp GIP R1    | 19,411,300 | 14,265,721   |
| gfp MIP R1    | 25,376,516 | 18,146,111   |
| gfp Input R2  | 4,468,053  | 1,500,751    |
| gfp CIP R2    | 27,442,936 | 18,934,987   |
| gfp GIP R2    | 25,924,790 | 14,426,424   |
| gfp MIP R2    | 18,009,319 | 13,458,480   |
| gfp Input R3  | 6,702,486  | 2,087,432    |
| gfp GIP R3    | 2,860,2646 | 19,344,367   |
| gfp MIP R3    | 26,288,689 | 18,174,554   |
| gfp Input R4  | 15,415,451 | 6,228,407    |
| gfp CIP R4    | 28,739,336 | 23,000,396   |
| Trl Input R1  | 23,397,826 | 17,282,427   |
| Trl CIP R1    | 17,696,632 | 12,269,533   |
| Trl GIP R1    | 27,868,552 | 19,730,344   |
| Trl MIP R1    | 29,437,758 | 21,835,078   |
| Trl Input R2  | 8,780,963  | 5,904,853    |
| Trl CIP R2    | 30,860,817 | 21,463,377   |
| Trl GIP R2    | 16,204,495 | 12,548,896   |
| Trl MIP R2    | 29,461,218 | 23,146,958   |
| Trl Input R3  | 4,636,303  | 1,357,768    |
| Trl MIP R3    | 24,357,846 | 17,369,589   |
| clamp Input R1 | 6,525,567  | 1,651,091    |
| clamp CIP R1  | 21,356,281 | 14,590,584   |
| clamp GIP R1  | 18,987,942 | 13,255,475   |
| clamp MIP R1  | 19,585,467 | 13,347,245   |
| clamp Input R2 | 2,033,630  | 732,493      |
| clamp CIP R2  | 29,502,913 | 19,919,643   |
| clamp GIP R2  | 22,527,570 | 15,683,734   |
| clamp MIP R2  | 22,696,988 | 15,898,513   |
| clamp Input R3 | 17,892,526 | 12,574,408   |
| clamp CIP R3  | 12,835,320 | 9,734,810    |
| clamp GIP R3  | 28,263,295 | 19,607,145   |

RNAi treatment, followed by input or IP antibody: “CIP” = CLAMP IP, “GIP” = GAF IP, “MIP” = MSL3 IP. Replicates are indicated as Replicate 1 = R1, 2 = R2, 3 = R3, and 4 = R4.
Table S2. Feature Importance for xgboost Classification of GAF and CLAMP Peaks. Related to Figure 4.

| Feature                                         | Gain  |
|-------------------------------------------------|-------|
| CLAMP unique motif PWM (PBM)                    | 0.197 |
| Number of “GAGAG” sequences                     | 0.187 |
| CLAMP unique motif PWM (ChIP-Seq)               | 0.154 |
| Chromatin State                                 | 0.114 |
| Distance to nearest TSS                         | 0.106 |
| GAF Motif PWM (PBM)                             | 0.095 |
| GAF Motif PWM (ChIP-Seq)                        | 0.075 |
| GAF & CLAMP Motif PWM (ChIP-Seq)                | 0.071 |

Features used to predict GAF and CLAMP peaks are listed in descending order of importance as reported by the xgboost-generated model. Gain indicates the relative contribution of the feature to the model.
Table S3. Feature Importance for xgboost Classification of GAF, CLAMP, GAF and CLAMP Peaks and sampled unenriched regions of the genome. Related to Figure 4.

| Feature                                         | Gain |
|-------------------------------------------------|------|
| GAF & CLAMP Motif PWM (ChIP-Seq)                | 0.181|
| CLAMP unique motif PWM (ChIP-Seq)               | 0.176|
| Distance to nearest TSS                         | 0.176|
| CLAMP unique motif PWM (PBM)                    | 0.111|
| Number of “GAGAG” sequences                     | 0.109|
| Chromatin State                                 | 0.087|
| GAF unique motif PWM (ChIP-Seq)                 | 0.084|
| GAF Motif PWM (PBM)                             | 0.075|

Features used to predict GAF and CLAMP peaks are listed in descending order of importance as reported by the xgboost-generated model. Gain indicates the relative contribution of the feature to the model.
| Feature                                                                 | Gain  |
|------------------------------------------------------------------------|-------|
| Number of “GA” dinucleotides                                           | 0.046 |
| Maximum number of consecutive “GN” dinucleotides                       | 0.029 |
| Maximum number of consecutive “GB” dinucleotides                       | 0.020 |
| Number of “GANNGAGA” sequences                                        | 0.012 |
| Number of “GAGAGNG” sequences                                         | 0.012 |
| Number of “GA not(GA) GAGA” sequences                                 | 0.009 |
| Number of “GAGAGAG” sequences                                         | 0.009 |
| Maximum number of consecutive “GA” dinucleotides                       | 0.006 |
| Number of “GAAAGAGA” sequences                                        | 0.005 |
| Number of “GACAGAGA” sequences                                        | 0.005 |
| Number of “GAGNGAGA” sequences                                        | 0.004 |
| Number GAF ChIP-Seq MEME regular expression matches                   | 0.002 |
| Number of “GNGGANNGAGANRG” matches                                    | 0.002 |
| Number CLAMP PBM MEME regular expression matches                      | 0.001 |
| Number of “GAGAGAGA” sequences                                        | < 0.001 |
| Number of co-bound ChIP-Seq MEME regular expression matches           | < 0.001 |
| Number CLAMP ChIP-Seq MEME regular expression matches                 | < 0.001 |
| Number GAF PBM MEME regular expression matches                        | < 0.001 |

Features used to predict GAF and CLAMP peaks are listed in descending order of importance as indicated by the xgboost-generated model. Gain indicates the relative contribution of the feature to the model.