Supplemental Data

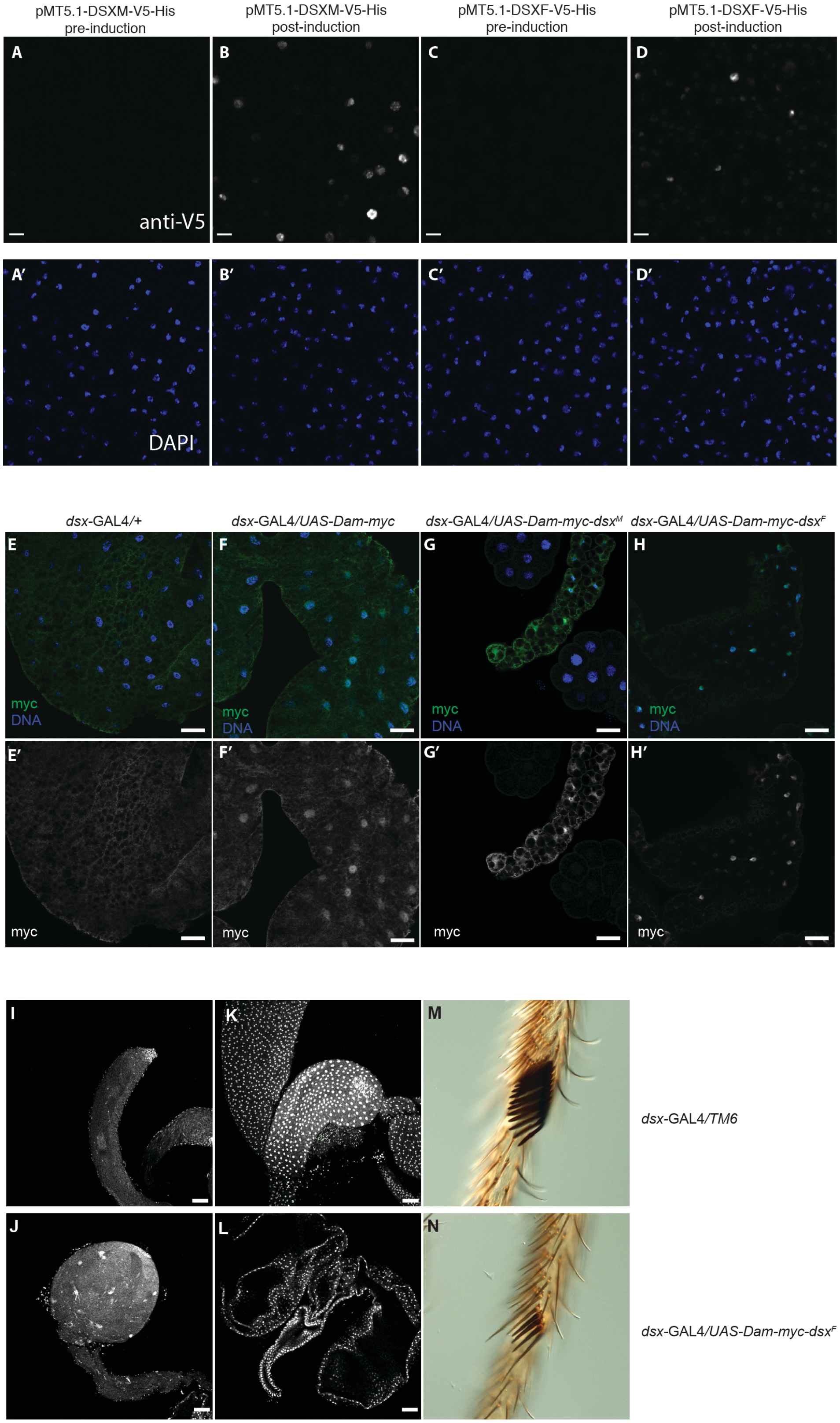

Figure S1: Expression of Dam-DSX fusion protein and resultant phenotypes, related to Figure 1. S2 cells carrying the pMT5.1-DSXM-V5-His construct before (A, A') and after (B, B') 60 hour induction. S2 cells carrying the pMT5.1-DSXF-V5-His construct before (C, C') and after (D, D') 60 hour induction. Scale bar = 10µm. A, B, C, and D are the V5 channel (white), and A', B', C', and D' are the DAPI channel (blue). Third instar larval fatbody from *dsx*-GAL4*/+* (E, E') *dsx*-GAL4*/UAS-Dam-myc* (F, F') *dsx*-GAL4*/UAS-Dam-myc-dsx*^{M} (G, G') and *dsx*-GAL4*/UAS-Dam-myc-dsx*^{F} (H, H'). E,F,G, and H are merged images of anti-myc (green) and DAPI (blue). E', F', G' and H' is a split of only the anti-myc signal (white). Testes (I,J) and ejaculatory ducts (K,L) were dissected and stained with DAPI. Light microscopy images of sex combs from control *dsx*-GAL4*/TM6* (M) and *dsx*-GAL4*/UAS-Dam-myc-dsx*^{F} (N). Scale bar = 50µm. The Dam fusion proteins include the myc epitope incorporated at the C-terminus of the Dam coding sequence such that Dam fusion proteins can be detected with anti-myc antibodies.

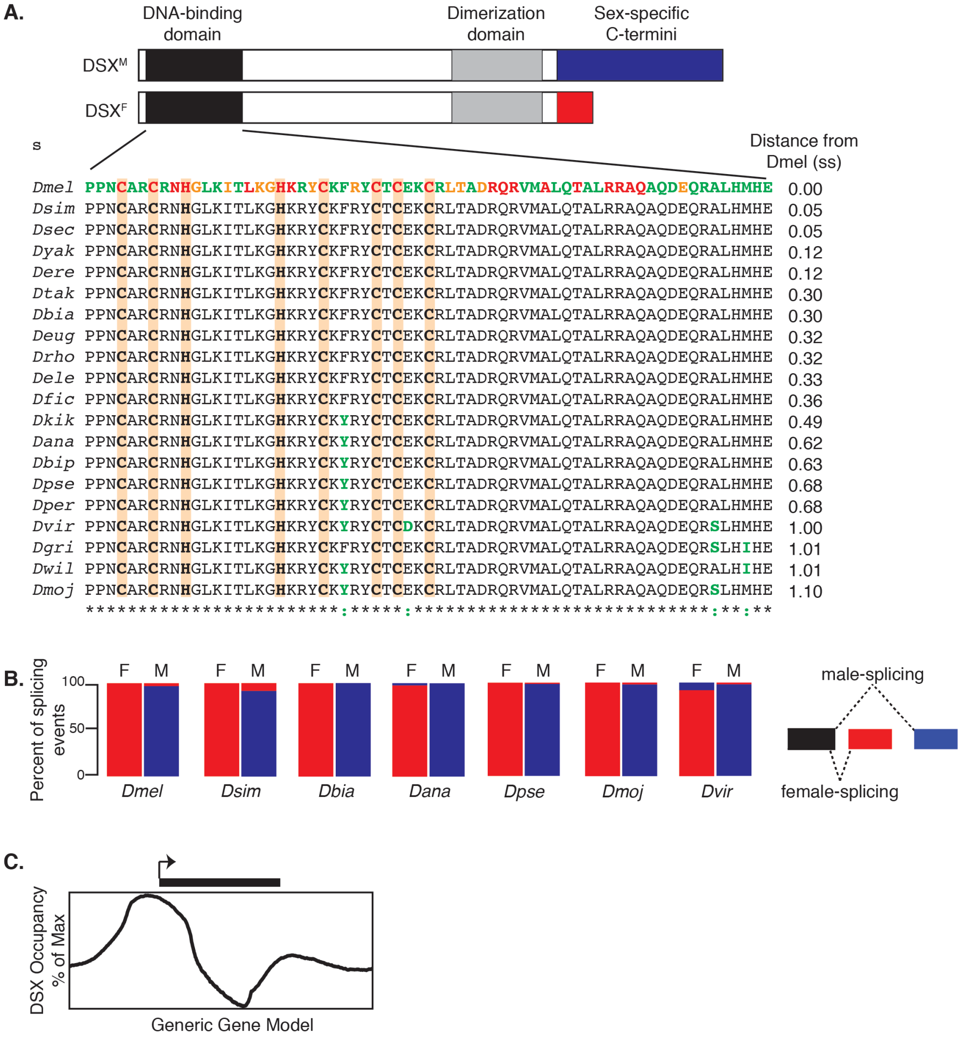
A.
DNA-binding domain
Dimerization domain
Sex-specific C-termini
DSX^M
DSX^F
s
Distance from Dmel (ss)
Dmel PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.00
Dsim PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.05
Dsec PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.05
Dyak PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.12
Dere PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.12
Dtak PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.30
Dbia PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.30
Deug PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.32
Drho PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.32
Dele PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.33
Dfic PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.36
Dkik PPNCARCRNHGLKITLKGHKRYCKYRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.49
Dana PPNCARCRNHGLKITLKGHKRYCKYRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.62
Dbip PPNCARCRNHGLKITLKGHKRYCKYRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.63
Dpse PPNCARCRNHGLKITLKGHKRYCKYRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.68
Dper PPNCARCRNHGLKITLKGHKRYCKYRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHMHE 0.68
Dvir PPNCARCRNHGLKITLKGHKRYCKYRYCTCDKCRLTADRQRVMALQTALRRAQAQDEQRSLHMHE 1.00
Dgri PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRSLHIHE 1.01
Dwil PPNCARCRNHGLKITLKGHKRYCKYRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRALHIHE 1.01
Dmoj PPNCARCRNHGLKITLKGHKRYCKYRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRSLHMHE 1.10
************************:*****:****************************:**:**
B.
Percent of splicing events
100
50
0
F
M
Dmel
Dsim
Dbia
Dana
Dpse
Dmoj
Dvir
male-splicing
female-splicing
C.
DSX Occupancy % of Max
Generic Gene Model

Figure S2: Conservation of the DSX DNA binding domain and sex-specific splicing, related to Figures 1, 2, and 3.

(A) Diagram of the DSX^{M} and DSX^{F} proteins (above) and amino acid sequence alignment of the DSX DNA binding domain from 20 Drosophila species (below). Cysteine and histidine residues in the Zn-binding site are highlighted in tan. The evolutionary distance from *D. melanogaster* is indicated in substitutions/site (ss) (Chen et al., 2014). Color-coding of *D. melanogaster* amino acids represent mutations that do not affect DSX activity (green), partially affect activity (orange), or impair activity (red) (adapted from (Zhang et al., 2006)). (B) Bar graphs representing the percentage of *dsx* splicing events resulting in production of female (red) or male (blue) isoform from RNA-seq data obtained from adult females (F) or males (M) from 7 Drosophila species (Chen et al., 2014). (C) The normalized (% of maximum average occupancy) distribution of DSX occupancy values along a generic gene model using +1.5Kb upstream of transcription start (bent arrow), the gene body (rectangle), where the first 0.5Kb and last 0.5Kb are at base level, and the middle 0.5Kb is scaled from all gene models, and the -1.5Kb downstream region are shown.

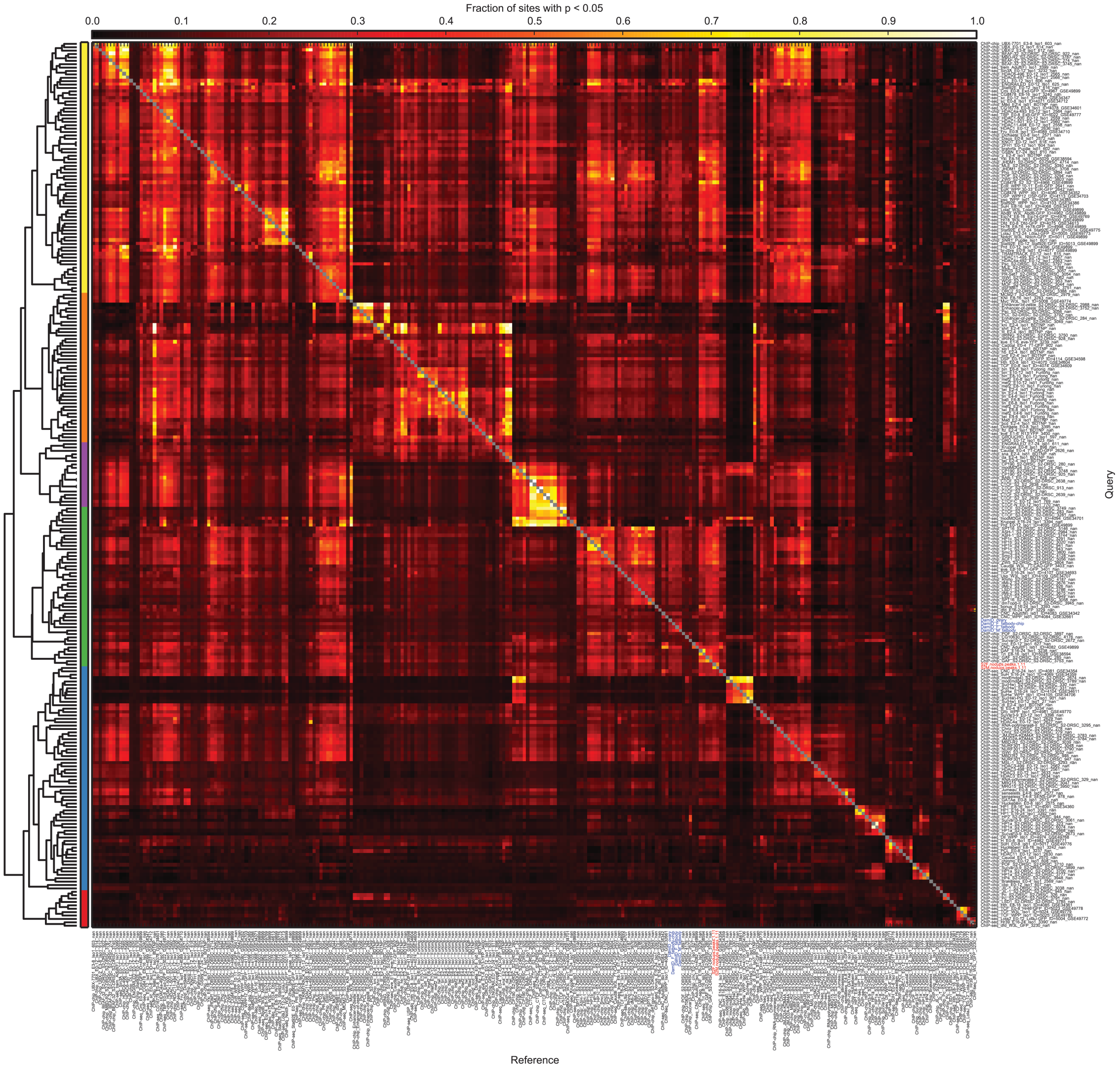
Fraction of sites with $p < 0.05$
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Query
Reference

Figure S3: **DSX^{F}- and DSX^{M}- occupied regions are not correlated with other transcription factors, related to Figure 1 and 2.** Hierarchically clustered heatmap of pairwise similarity metrics between all 255 available ChIP-chip and ChIP-seq experiments and DSX ChIP-seq (highlighted in red) as well as DSX DamID-seq/chip (highlighted in blue). Brighter colors indicate higher similarity (higher fraction of sites with $p < 0.05$); DSX^{F} and DSX^{M} (highlighted in red (ChIP-seq) and blue (DamID)) are more similar to each other than they are to any other assayed factor. Self-self comparisons along the diagonal are indicated in gray. Colored blocks along the left side indicate broad clusters. See methods for details. Table S6 contains the source and description of all occupancy data sets tested for correlation with DSX occupancy data.

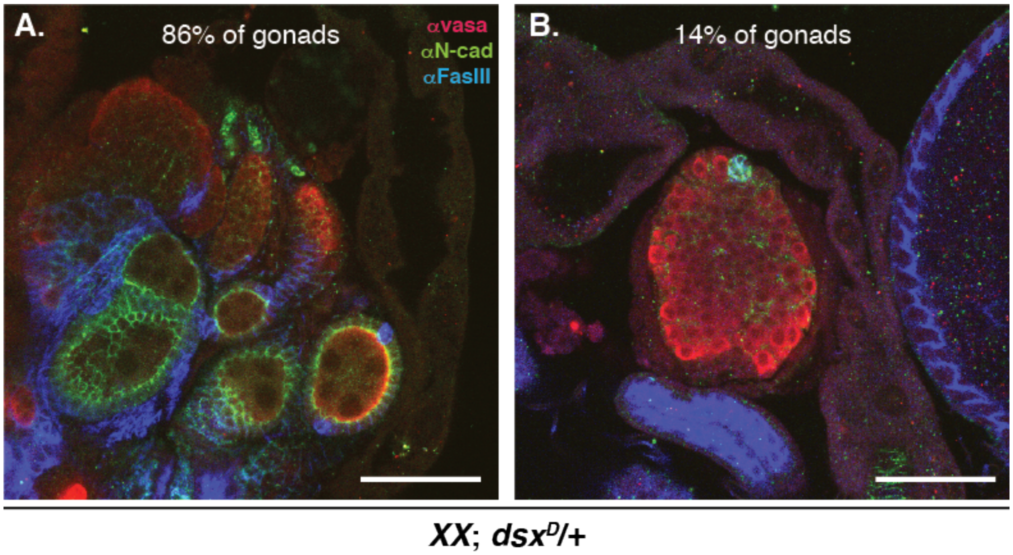
A.
86% of gonads
αvasa
αN-cad
αFasIII
B.
14% of gonads
XX; dsxD/+

Figure S4: XX; *dsx^D*/+ gonad phenotypes, related to Figure 6. Representative images of gonads dissected from XX; *dsx^D*/+ adults having either female-like terminal filaments (A) or male-like hubs (B). Terminal filaments and hubs are marked with anti-N-Cad (green), hubs are marked with anti-FasIII (blue) and anti-N-cad (green), and germ cells with anti-Vasa (red). Scale bar = 50µm.

Table S1: A summary of gene-level DSX occupancy, PWM score, CI score, orthology, and expression data (TableS1.xlsx), related to Figures 1, 2, and 3. An .xlsx workbook file containing all gene-level occupancy scores, binary scores for occupied versus unoccupied, gene-level conservation index (CI) scores, gene-level DSX position weight matrix scores, k-means cluster ID for occupancy score clustering, binary scores for mouse DMRT1 target orthology, and RNA-seq FPKM values for each RNA-seq sample. See sheet titled "README" for more information and definition of column headings.

Table S2: Site-level position weight matrix scores, conservation index scores, and PhastCons scores, all position weight matrices (TableS2.xlsx), related to Figure 1. An .xlsx workbook file containing DSX position weight matrix (PWM) scores for all 566,628 DSX motif-related sequences (sheet "PWM_scores"), conservation index scores, and PhastCons scores for all 173,775 sites with a positive PWM scores located within a *D. melanogaster* gene body plus 1 kb upstream excepting those in coding sequence (sheet "CI_scores"), conservation index (CI) scores for the 17,380 sites in the top 10% of all sites with any relationship to the DSX PWM that occur within a gene excluding coding sequence or 1Kb upstream of a gene (sheet "top_10_percent_PWM_CI_scores"), and the DSX position weight matrix along with all

shuffled position weight matricies ("PWMs"). See sheet titled "README" for more information and definition of column headings.

Table S3: 2nd chromosome deficiencies tested for genetic interaction with *dsx^D* (TableS3.xlsx)*,* related to Figure 4. A .xlsx workbook file containing results from the unbiased genetic interaction tests between *dsx^D* and 101 deficiencies of the 2nd chromosome, 17 genomic intervals interacting with *dsx^D*, and the tested genomic intervals that did not interact with *dsx^D*. For each deficiency tested for genetic interaction with *dsx^D*, a description of the genetic interaction phenotype is provided ("Phenotype Description") along with the FlyBase deficiency name and aberration ID ("FlyBase ID"). The known or estimated cytological location ("Deleted Segment") as well as chromosome base position (FlyBase release 5) for the left and right breakpoints are also provided.

Table S4: A summary of RNAi data for putative DSX target genes (TableS4.xlsx), related to Figures 5 and 6. For each gene tested by RNAi knockdown, the FlyBase gene name and ID ("FlyBase Identifier") are provided along with a text description summarizing phenotypic findings after RNAi knockdown ("Phenotype Descriptions"). There is also a text summary of DSX occupancy, occupancy clustering, gene-level PWM score, gene-level conservation index, and other criteria used to select the gene for RNAi test ("Criteria for Selection").

Table S5: Gene ontology term enrichment in occupied genes (TableS5.xlsx), related to Figure 1 and 2.

An .xlsx workbook file containing enriched gene ontology terms in the 3,717 occupied genes identified in this work. The ontology domains molecular function and biological process were examined, with output from each domain occupying a separate sheet in this file. See sheet titled “README” for more information and definition of column headings.

Table S6: Occupancy data sets used for comparison to DSX occupancy data, related to Figure 1 and Figure S3.

An .xlsx workbook file containing a source and description for all occupancy data sets tested for correlation with DSX occupancy data. See sheet titled “README” for more information and definition of columns headings.

Supplemental Experimental Procedures

Fly stocks. Fly stocks were obtained from the Bloomington Drosophila Stock Center (Cook et al., 2010), the Transgenic RNAi Project (Ni et al., 2011), and from the B.S. Baker lab and other generous members of the Drosophila community. See FlyBase for gene and allele descriptions (Marygold et al., 2013) for *tra2*^{ts2} (FBal0017028), *tra2*^{ts1} (FBal0017027), *P{UAS-tra.F}20J7* (FBti0010566), *P{tubP-GAL80ts}7* (FBti0027798), *P{tubP-*GAL4*}LL7* (FBti0012687), *dsx*^{D} (FBal0003200), GAL4$^{dsx.KI}$ (FBal0277019), dsx^{GAL4} (FBal0244772), *gpp*^{X} (FBal0175658), lilli^{A17-2} (FBal0103689), *w*^{1118}*,* and *Oregon R*. Bloomington Deficiency Kit II stocks used can be found in Table S3. Alleles tested for genetic interaction with *dsx* and TRiP RNAi lines can be found in Table S4. Information on *UAS-Dam-myc* (*Dam*), *UAS-Dam-myc-dsx*^{F} (*Dam-dsx*^{F}*)*, and *UAS-Dam-myc-dsx*^{M} (*Dam-dsx*^{M}) can be found below. Flies were grown on standard Bloomington Drosophila Stock Center (Bloomington, IN, USA) or Drosophila Species Stock Center (San Diego, CA, USA) medium at 25°C unless otherwise noted.

Immunohistochemistry and all sample imaging. Tissue was dissected from adult flies aged 1 to 3 days in PBS followed by fixation for 12-25 minutes in PBS containing 0.1% Triton X-100 (PBTx) with 4.0-4.5 % formaldehyde. Samples were blocked in PBTx with 0.1 or 1.0% BSA (BBTx) with or without 2% normal goat serum (NGS) for at least 1 hour and then incubated in BBTx with primary antibody 1-2 hours at room temperature or overnight at 4 °C. Following 3X 10 minute washes in PBTx, samples were incubated in BBTx with or without 2% NGS plus secondary antibody for 1-2 hours at room temperature. Following 3X 10 minute washes in PBTx, samples were mounted on slides in 2.5% DABCO (Sigma-Aldrich, St. Louis, MO, USA) or Fluoromount G (Southern Biotech, Birmingham, AL, USA).

The following primary antibodies were used: chicken anti-VASA (K. Howard) at 1:10,000; rabbit anti-VASA (R. Lehmann) at 1:1000; rat anti-DN-cadherin Ex#8 (Developmental Studies Hybridoma Bank, (Iwai et al., 1997)) at 1:20-50; mouse anti-FAS3 7G10 (Developmental Studies Hybridoma Bank, (Patel et al., 1987)) at 1:30; guinea pig anti-TJ at 1:1000 (Jemc et al., 2012); mouse anti-myc9E10 (Roche); mouse anti-V5 (Invitrogen) at 1:200. The following secondary antibodies were used: Alexa 546 goat anti-chicken at 1:500; Alexa 633 goat anti-chicken at 1:500; Alexa 488 goat anti-rat at 1:300-500; Alexa 633 goat anti-mouse at 1:500; Alexa 546 goat anti-mouse at 1:500; Alexa 488 goat anti-mouse at 1:300; Alexa 488 goat anti-guinea pig at 1:500, Alexa 555 donkey anti-rabbit at 1:300 (Invitrogen, Carlsbad, CA, USA), and Cy5 goat anti-guinea pig (Jackson ImmunoResearch, West Grove, PA, USA) at 1:300. We stained DNA with DAPI (Sigma-Aldrich, St. Louis, MO, USA) at 10 µg/ml for 10 minutes and then rinsed 3 X 5 minutes in PBTx. All immunohistochemistry samples were imaged on a LSM 510 Meta confocal microscope (Zeiss, Jena, Thuringia, Germany**)**.

For electron miscroscopy, adult flies aged 2 days were mounted (without desiccation or other treatment) on aluminum pedestals and imaged in a FEI Quanta 200 ESEM at 80Pa (FEI, Hillsboro, Oregon, USA).

***dsx^D* genetic interaction and RNAi of putative target genes**. *dsx^D*, *e^1*, *Sb^1*/TM6B males were crossed to virgin females carrying alleles or deficiencies being tested for genetic interaction with *dsx* (Tables S3). Female offspring of this cross carrying both *dsx^D* and the allele/deficiency being tested were scored under dissecting microscope for novel phenotypes in abdominal pigmentation, sex comb structure, or genitalia when compared to their female siblings heterozygous for the *dsx^D* allele. Genitalia in XX; *dsx^D/+* flies had female genital structures including a small vaginal plate with fewer teeth than wildtype females (not shown). Failed vulva closure and hemolymph clotting occurred over the vaginal plate. The male genital arch was thin dorsally and spread apart ventrally, and had small lateral and posterior lobes. Male clasper teeth were reduced in number, and the penis apparatus was usually missing.

For the deficiency screen, 19 of 101 tested deficiencies shifted some aspect of the external sexual morphology relative to XX; *dsx^D*/+. Overlapping and adjacent deficiencies with similar effects on the *dsx^D/+* phenotype were merged into new shifting regions. When 2 deficiencies overlapped but did not show coordinate genetic interactions, it was assumed that the overlapping region did not contain the gene(s) in each deficiency that interacted with *dsx^D*; therefore, the overlap was removed. Shifting regions overlapping with non-shifting deficiencies were removed to define 17 unique genomic intervals (Table S3) that shift some aspect of the *dsx^D/+* phenotype.

For gonad genetic interaction tests with *dsx*^{D}, gonads of female offspring carrying both *dsx*^{D} and the allele being tested were dissected. The morphology of mutant gonads was scored by the presence of male or female germline stem cell niches when compared to their female siblings carrying only the *dsx*^{D} allele. A gonad was scored as positive for a male germline niche (i.e. hub) if a structure with morphology similar to a normal hub (Figure 5B) also double-labeled with N-cadherin (N-cad) and Fasciclin III (FasIII). A gonad was scored as positive for a female germline stem cell niche if a structure with morphology similar to terminal filaments was labeled with N-cad but was not labeled with FasIII (FasIII does not label wildtype terminal filaments). At least 35 gonads were scored for each tested allele.

For RNAi tests of genes, *dsx*-GAL4*,* (Rideout et al., 2010) and/or (Robinett et al., 2010), virgin females were crossed to males carrying various individual RNAi constructs and raised at 29°C or 25°C. 10 male and 10 female day 3-5 progeny carrying both *dsx*-GAL4 and *UAS-RNAi* were screened under dissecting microscope for phenotypes in the following sexually dimorphic structures: genitalia, gonad, sex comb, abdominal pigmentation, male reproductive tract (accessory gland, ejaculatory duct, ejaculatory bulb), and female reproductive tract (oviduct, spermathecae, parovaria). Flies carrying *dsx*-GAL4 and *UAS-RNAi* were compared to control flies carrying *dsx*-GAL4 alone and Oregon R. For a full description of RNAi alleles and phenotypes see Table S4.

DamID-seq and DamID-array. The *UAS-Dam-myc-dsx*$^{M/F}$ DNA constructs were made by ligating PCR amplified sex-specific *dsx* cDNA (gift from Gyunghee Lee, University of Tennessee, Knoxville, USA) into the pUASt-attB-NDam-myc plasmid (gift from Tony Southall, University of Cambridge, Cambridge, UK). UASt-attB-Dam-myc (Dam), UASt-attB-Dam-myc-dsx^{F} (*Dam-dsx^{F})*, and UASt-attB-Dam-myc-dsx^{M} (Dam-dsx^{M}) constructs were independently integrated (Genetic Services, Inc., Cambridge, MA, USA) into the attP2 site on chromosome 3L using ϕC31 site-directed integration (Bischof et al., 2007). In accordance with DamID protocols, a GAL4 driver was not used in order to keep Dam, Dam-*dsx^{F}* and Dam-*dsx^{M}* expression low to prevent lethality and saturation (Greil et al., 2006; Southall and Brand, 2007). S2 cells were transfected with pMT5.1-DSXM-V5-HisB or pMT5.1-DSXF-V5-HisB (Garrett-Engele et al., 2002) and pCoBlast (Invitrogen, Carlsbad, CA, USA) as the selection plasmid using Effectene (Qiagen, Valencia, CA, USA). Expression was induced using Cu^{+} and presence of fusion proteins was confirmed by immunostaining (Figure S1) and western blot (data not shown). Chromatin immunoprecipitation was performed with anti-V5 tag monoclonal antibody (Invitrogen, Carlsbad, CA, USA) on Protein G coupled Dynabeads (Invitrogen, Carlsbad, CA, USA) followed 1% formaldehyde and shearing chromatin to 200-1000 bp. Adult fatbody and ovaries were dissected from 20-70 flies in PBS at room temperature from day 5 adult flies heterozygous for one of Dam insertions. Samples were transferred to ice after 30 minutes.

Genomic DNA was extracted using components of Qiagen's DNEasy Blood and Tissue kit (Qiagen, Valencia, CA, USA). Samples processed for DamID-seq were homogenized in 175 µl PBS and incubated with 200 mg of RNAse A for 2 minutes at room temperature. Tissue was lysed with 20 µl of proteinase K and 200 µl of buffer AL for ten minutes at 70 °C. 200 µl of ethanol were added to each sample and they were transferred to the spin columns after which genomic DNA extraction continued following manufacturer's instructions (Qiagen, Valencia, CA, USA) with the exception of a 30 minute incubation prior to first elution and a second elution step after a 10 minute incubation. Genomic DNA for DamID-array was extracted following manufacturer's protocol except for the following modifications: a 1.5 hr incubation with lysis buffer prior to the addition of proteinase K, addition of 400 µl of Buffer AL and 300 µl of 100% ethanol, two rounds of both AW1 and AW2 and an incubation with the elution buffer for 30 minutes prior to two rounds of elution. 2.5 - 3 µg of fatbody genomic DNA and 0.3 µg of ovary genomic DNA was used for selective PCR amplification of methylated DNA. DNA was incubated with 10-30 units of *Dpn*I in 50-100 ul of Buffer 4 (New England Biolabs, Ipswich, MA, USA). *Dpn*I (New England Biolabs, Ipswich, MA, USA) was inactivated at 80 °C for 20 minutes and digested DNA was purified through a Qiaquick PCR Purification column (Qiagen, Valencia, CA, USA) following manufacturer's protocol and eluted in 30 µl ddH20. One-half of the *DpnI* reaction products were ligated to 40 pmol of the doublestranded DamID adaptors (top strand : 5'-CTAATACGACTCACTATAGGGCAGCGTGGTCGCGGCCGAGGA-3'; bottom strand: 5'-TCCTCGGCCG-3') for 2 hours at 16 °C with 400 units of T4 ligase (New England Biolabs, Ipswich, MA, USA) or 5 units of T4 ligase (Roche, Indianapolis, IN, USA) in a 20 µl reaction volume. All 20 µl of the adapter-ligated DNA were then subjected to *Dpn*II digestion with 10 units of *Dpn*II (New England Biolabs, Ipswich, MA, USA) in a 80 µl reaction volume for at least one hour. PCR amplification was performed with 20 µl of the *Dpn*II digested DNA in an 80 µl volume with 100 pmol PCR primer (5′-TCCTCGGCCG-3′), 16 nmol of each dNTP and 1.6 µl PCR Advantage enzyme mix in 1X PCR Advantage Reaction Buffer (Clontech, Mountain View, CA, USA) or 62.5 pmol PCR primer (5′-TCCTCGGCCG-3′), 16 nmol of each dNTP, 80 nmol $MgCl_2$ in 1X buffer with 8 units of *taq* polymerase (Fermentas, Pittsburgh, PA, USA). DNA was amplified with the following program: 10 minutes at 68 °C, 1 minute at 94 °C, 5 minutes at 65 °C and 15 minutes at 68 °C, followed by 3 cycles of 1 minute at 94 °C, 1 minute at 65 °C and 10 minutes at 68 °C and then 17 cycles of 1 minute at 94 °C, 1 minute at 65 °C and 2 minutes at 68 °C. DNA was purified through a Qiaquick column (Qiagen, Valencia, CA, USA).

DamID-array samples were analyzed at Nimblegen where Cy3- and Cy5-fluorescently labeled DamID-prepared DNA was hybridized to the DM_5_Catalog_tiling_HX1 whole genome tiling array and fluorescence data was collected by Roche NimbleGen (Madison, WI, USA). Probe sequence, probe position information and array details are available under GEO accession GPL10639. Three independent biological replicates were collected for DamID-array samples. The raw probe intensity data for each DamID-array experiment was accessed using the DM_5_Catalog_tiling_HXI_pair.txt file provided by Nimblegen (Madison, WI, USA). One dye-flip was performed for each sex and tissue. Arrays were quantile normalized with the R package *preprocessCore* (http://www.bioconductor.org/packages/release/bioc/html/preprocessCore.html). In order to define a lower limit for detection of hybridization, we calculated the mean fluorescent intensity of the 15,758 random sequence probes on the array. The 95th percentile of the mean for random probes was used as a cut-off for hybridization detection. When the replicate means of both Dam-DSX treatment and Dam-only control probes were at or below this value, the data from that probe was removed from the analysis. In order to identify probes with significantly different levels of fluorescent intensity, modified two-sided t-tests were performed assuming unequal variance. p-values were adjusted for multiple testing using the FDR method of Benjamini and Hochberg (Benjamini and Hochberg, 1995). Results of the statistical test for all probes is available at GEO (GSE49480). All calculations and statistical tests were performed in R (R Core Team, 2013). Probes that displayed a FDR < 0.01 and log2Fold-change > 0 were selected for further analysis. When the chromosomal positions of the selected probes positions occurred within 1000 bp of one another they were merged into features to form peaks using BedTools v2.16.2 (Quinlan and Hall, 2010).

For DamID-seq samples, PCR-amplified DNA was sonicated in a 200 µl volume of Qiagen's EB buffer in a BioRuptor Sonicator (Diagenode, Denville, NJ, USA) set on high for 3 X 15 minutes in a 4 °C water bath. Following sonication, DNA was purified through a Qiaquick column (Qiagen, Valencia, CA, USA). Two independent biological replicates were collected for DamID-seq samples. 20 ng of sonicated DamID-prepared DNA were used to make libraries following the protocol in the Illumina ChIP-seq Sample Preparation Kit (Illumina, San Diego, CA, USA). A gel slice of 250-350 bp was excised from the gel prior to PCR amplification. Library concentration was measured on a Nanodrop (Thermo Scientific, Waltham, MA, USA) and size distribution was assessed on a Bioanalyzer (Agilent, Santa Clara, CA, USA). DamID-seq samples were sequenced on a GAIIx or HighSeq 2000 instrument with 76 bp read lengths.

DamID-seq reads were generated using the Illumina pipeline 1.6.47.1 (Male fat body Dam-DSX^{M}), 1.8.70.0 (male fat body Dam-alone, female fat body Dam-alone, female fat body Dam-DSX^{F}), or 1.12.4 (ovary Dam-alone and ovary Dam-DSX^{F}). Reads were trimmed by 17 bp on each end to remove primer sequence and mapped to the *D. melanogaster* genome (FlyBase release 5 with no Uextra) using Bowtie 0.12.7 (Langmead et al., 2009) accepting only uniquely mapped reads with no more than 2 mismatches (-m1 –v2). Duplicate reads were removed from the libraries before peak calling with the Picard tool MarkDuplicates v1.95 (http://picard.sourceforge.net). In order to identify regions of the genome enriched for DSX occupancy, the number of reads occurring in non-overlapping consecutive 500 bp intervals across the genome were counted with HTSeq v0.5.1p2 (Anders et al., 2014). DESeq v1.12.0 (Anders and Huber, 2010) was used for library size normalization and identification of bins significantly enriched for Dam-DSX reads compared to Dam-Only reads (method adapted from (Ross-Innes et al., 2012)). The depth-normalized occupancy signal averaged between replicates, fold changes and associated p-values and Benjamini-Hochberg FDR-adjusted p-values for each bin are available on the GEO record GSE49480. Bins that contained no reads in either control or treatment samples were removed from the analysis. Bins selected for further analysis for all samples were those that displayed differential read counts with a FDR (Benjamini and Hochberg, 1995) < 0.01 and a log2 Fold-Change > 0 where the number of Dam-DSX reads was the numerator and number of Dam-Only Control reads was the denominator. Adjacent selected bins were combined into features to produce peaks using BEDTools v2.16.2 (Quinlan and Hall, 2010). In order to calculate genome-wide DSX DamID signal (used to create gene level occupancy scores; see below), the log2 Fold Change [(DamDSX +1)/(DamOnly +1)] was calculated for all 500 bp bins across the genome.

ChIP-seq. Schneider Drosophila line 2 cells (S2) were maintained at 25°C in Schneider Drosophila medium (Invitrogen, Carlsbad, CA, USA) containing 10% heat-inactivated Fetal Bovine Serum (JRH Biosciences, Lenexa, KS, USA) and antibiotics (0.5 U/ml penicillin and 0.5 µg/ml streptomycin, Invitrogen, Carlsbad, CA, USA). Cells were transfected with 1 µg expression plasmids (pMT5.1-DSXM-V5-His B and pMT5.1-DSXF-V5-His B (Garrett-Engele et al., 2002) using Effectene Transfection Reagent (Qiagen, Valencia, CA, USA) with 50 ng pCoBlast (Invitrogen, Carlsbad, CA, USA) as the selection plasmid. Following transfection, cells were grown in Schneider Drosophila medium for 60 hours prior to selection with 30 µg/ml blasticidin (Invitrogen, Carlsbad, CA, USA). After 5 weeks of selection, blasticidin-resistant cells were maintained in complete Schneider cell medium containing 25 µg/ml blasticidin. Expression of the recombinant proteins from the MT promoter was induced by adding copper sulfate to the medium to a final concentration of 500 μM. Presence of the DSX fusion proteins was confirmed by immunostaining (Figure S1) and western blot (data not shown).

~2.7 X 10^8 cells were fixed in 1% formaldehyde for 10 minutes at room temperature. The reaction was quenched by adding glycine to a final concentration of 125 mM and a 5-minute incubation on a shaker at room temperature. Subsequently, the cells were washed twice with ice-cold PBS. After centrifugation at 500 x g (1680 rpm) for 5 min at 4 °C the cell pellet was resuspended in 10 ml ice-cold cell lysis buffer (5 mM pH8.0 PIPES buffer, 85 mM potassium chloride, 0.5% Nonidet P40) containing protease inhibitors (cOmplete, EDTA-free, Roche, Indianapolis, IN, USA) for 10 minutes at 4 °C. Nuclei were released by douncing with a Wheaton homogenizer pestle B. The crude nuclear extract was collected by centrifugation at 500 x g (1680 rpm) for 5 min at 4 °C, resuspended in 2 ml ice–cold nuclear lysis buffer (50 mM pH8.1 Tris.HCl, 10 mM EDTA, 1 % SDS with protease inhibitors) and incubated for 20 minutes at 4 °C. After adding 1 ml ice-cold IP dilution buffer (0.01 % SDS, 1.1 % TritonX-100, 1.2 mM pH 8 EDTA.Na2, 16.7 mM pH8 Tris.HCl, 167 mM NaCl and protease inhibitors) and 0.3 g acid-washed glass beads (Sigma-Aldrich, St. Louis, MO, USA) to the nuclear extract, the chromatin was sheared to 200-1000 bp using a Misonix Sonicator 3000 (Misonix, Inc. Farmingdale, NY, USA). Sonication was performed on ice water with 8 pulses of 30 seconds at 30 second intervals. Thereafter, the cell debris were removed by centrifugation at 16000 x g (13000 rpm) for 10 min at 4 °C. Input DNA was prepared in an identical manner from non-transfected cells.

The sonicated, fixed chromatin was precleared by incubation with preblocked magnetic Protein G coupled Dynabeads (Invitrogen, Carlsbad, CA, USA) overnight at 4 °C on a rotating wheel. Subsequently, the chromatin was divided into three aliquots of 850 μg. IP samples were incubated with 8.5 μg anti-V5 tag monoclonal antibody (Inivitrogen, Carlsbad, CA, USA) prebound to Dynabeads overnight at 4 °C on a rotating wheel. The beads were washed three times with low salt buffer (0.1 % SDS, 1 % Trition, 2 mM EDTA, 20 mM pH 8 Tris, 150 mM NaCl) three times with high salt buffer (0.1 % SDS, 1 % Trition, 2 mM EDTA, 20 mM pH 8 Tris, 500 mM NaCl) and finally twice with LiCl buffer (10 mM pH 8.1 Tris, 1mM EDTA, 0.25M LiCl, 1 % NP40, 1 % sodium deoxycholate) with incubation at room temperature on a rotating wheel for 5 min respectively. The beads were incubated twice on a rotating wheel at room temperature for 20 min in 200 μl elution buffer (0.1 M $NaHCO_3$ and 1% SDS) to recover the immunoprecipitated DNA. Cross-links were dissociated by incubation at 65 °C overnight. DNA was purified by phenol-chloroform extraction and ethanol precipitation.

100 ng immunoprecipitated DNA and 300 ng of input DNA were used to make libraries with the Genomic DNA sample preparation kit (Illumina, San Diego, CA, USA) according to the manufacturer's protocol. Adapter-ligated DNA of 200 ± 25 bp range was excised from the gel before PCR amplification. Input chromatin was prepared from two biological replicates and IP samples were prepared from three biological replicates. ChIP-seq libraries were sequenced on an Illumina GA1 instrument with either 25 or 36 bp read lengths. Reads were generated using the Illumina pipeline software 0.3.0. All ChIP-seq

reads were trimmed to 25 bp prior to mapping to the *D. melanogaster* genome (FlyBase release 5 with no Uextra). The sequence reads from all biological replicates were pooled prior to mapping using Bowtie 0.12.7 (Langmead et al., 2009) accepting only uniquely mapped reads with no more than 2 mismatches (-m1 –v2). Duplicate reads were removed from the libraries before peak calling with the Picard tool MarkDuplicates v1.95 (http://picard.sourceforge.net). The WTD method of peak calling from the ChIP-seq analysis program SPP v1.11 was used to call peaks with an FDR of 0.01 (Kharchenko et al., 2008). IP and input reads were loaded into SPP and anomalous reads due to localized regions of extremely high read count were removed with the command *remove.local.tag.anomalies*. Broader peak regions of enrichment surrounding the predicted binding site were added to create the final peak coordinates using the command *add.broad.peak.regions*. In order to calculate ChIP-seq signal across the genome for use in producing a gene-level occupancy score (see below), IP or input reads were counted in non-overlapping consecutive 500 bp intervals across the genome with HTSeq v0.5.1p2 (Anders et al., 2014). Read counts were depth-normalized and the log2 Fold Change [(IP + 1)/(input +1)] was calculated for all bins.

In all occupancy experiments, replicate preparations of a given sample type showed excellent reproducibility (spearman rho > 0.9, data not shown).

***de novo* motif analysis**. The binding positions reported for the DSX^M and DSX^F proteins from the SPP ChIP-seq analysis (above) were sorted by descending binding scores and the top 1000 scoring sites were selected for further analysis. 200 bp of DNA on either side of the identified binding position were used to search for enriched DNA sites using MEME-ChIP (Machanick and Bailey, 2011). Comparison of the position weight matrix for the biochemically determined DSX binding sequence (Yi and Zarkower, 1999) to the *de novo* site analysis from MEME for either DSX isoform was performed with TOMTOM (Gupta et al., 2007).

Gene-level occupancy scores. Gene level occupancy scores were calculated by summing the log2 fold change of (DSX occupancy signal)/(Control signal) in 500 bp bins under all called peak regions within the gene body plus 1 kb upstream of the TSS (adapted from (Ouyang et al., 2009)). We picked the window by analyzing the relationship between occupied regions and gene features (Figure S2C) and the analysis showed that the DSX-bound sites were primarily observed within 1 kb preceding the TSS (with maximum at 0.6 kb before TSS) and decreased in density throughout the gene body. For genes with no called peaks, the gene level occupancy score was computed as the average log2 Fold Change of (DSX occupancy signal)/(Control signal) over 500 bp bins in the gene body plus 1 kb upstream. Genes were sorted in non-increasing order of gene level occupancy scores (genes with peaks ranked first and then genes without peaks followed). The ranks were normalized by dividing the gene ranks by the total number of genes.

Gene-Level DSX PWM Score. The position weight matrix (PWM) for DSX sequence binding was composed of the position nucleotide percentages reported for DSX protein (Yi and Zarkower, 1999). The PWM was converted into the JASPAR format and was used to search the *D. melanogaster* genome (FlyBase release 5) for sequences matching the PWM with the Bio.Site.search_pwm method module in BioPython (Cock et al., 2009). 556,628 sequences with any relationship to the DSX PWM were identified using this method. Each sequence was assigned a score calculated by summing the log odds for each position. Scores ranged from 0.000103 to 18.84745 (Table S2).

The gene-level DSX PWM score (Table S1) was based on the number of DSX binding sequences at a gene as well as the PWM score of each site. For a gene g, let S(g) be the set of DSX binding sequences within gene body plus 1kb upstream. We computed the probability that at least one binding event occurs in S(g) and used this as the gene level DSX PWM score, assuming that binding events are independent and that the probability of binding to a sequence is $(1+\varepsilon)^{wi-W}$ where w_i is the PWM score of i-th sequence in S(g) multiplied by 10, W = 189 (rounded up from the maximum PWM score of a sequence multiplied by 10) and an adjusting parameter $\varepsilon = 0.03$. Gene level DSX PWM score is then defined as follows.

$$PWMscore(g) = 1 - \prod_{m_i \in S(g)} (1-(1+\varepsilon)^{w_i - W})$$

where m_i is i-th sequence in S(g).

Conservation of DSX binding sequences and gene-level DSX conservation index (CI) score. The genomes of *D. simulans, D. sechellia, D. yakuba, D. erecta, D. ficusphila, D. eugracilis, D. biarmipes, D. takahashii, D.elegans, D. rhopaloa, D. kikkawai, D. ananassae, D. bipectinata, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis* and *D. grimshawi* (Adams et al., 2000; Chen et al., 2014; Drosophila 12 Genomes et al., 2007; Richards et al., 2005) were searched for sequences that relate to DSX's binding sequence position weight matrix as described above for *D. melanogaster*. All identified sequences were associated with genes according to the identity of the nearest first coding exon using BEDTools v. 2.16.2

(Quinlan and Hall, 2010). For genes with multiple transcripts with different first protein-coding exons, only the most proximal first-coding exon was used. The positions of all first coding exons in each species were identified by aligning first coding exons from *D. melanogaster* (FlyBase annotation version 5.46) using liftover chain files. To create liftover chain files, whole-genome alignments between *D. melanogaster* and each other Drosophila species were performed using lastz (Harris, 2007) and executables from the UCSC Genome Browser (Meyer et al., 2013) according to a protocol on the UCSC user guide. Briefly, genomic sequences from each non-melanogaster species were split into 5 MB segments with the faSplit executable (parameters: size -oneFile 5000000 -extra=10000), and pairwise alignment was performed against *D. melanogaster* with lastz (parameters: --masking=50 --hspthresh=2200 --ydrop=3400 --gappedthresh=4000 --inner=2000). These alignments were converted to Pattern Space Layout (PSL) format and lifted to chromosomes with the lavToPsl and liftUp executables. Then, these PSL alignments were chained with the axtChain executable (parameters: -linearGap=medium -psl), combined with the chainMergeSort and chainSplit executables, and converted to alignment nets with the chainNet executable. Based on alignment nets, liftOver chain files that convert annotations from *D. melanogaster* to other species were created with the netChainSubset executable.

D. melanogaster sequences with positive PWM scores located within a protein-coding *D. melanogaster* gene body plus 1 kb upstream excepting those in coding sequence or located on chrU, chrUextra or chrM (173,775 in total, Table S2) were used to search for orthologous sequences among those sequences that were associated with the same gene in each of the remaining 19 genomes. A sequence was considered orthologous if the edit distance of the largely invariant nucleotides at position 4-10 was ≤ 1 and the position difference relative to the first coding exon was less than 2 kb. A conservation index (CI) score for each sequence (Table S2) was computed by summing the substitution/site distance associated with each species in which the sequence was identified. The evolutionary distances between *D. melanogaster* and 19 other Drosophila species (Figure S1) are expressed in units of substitutions per synonymous site (ss) as defined in (Chen et al., 2014).

A CI score for each gene (Table S1) was computed either by summation of CI scores >90th percentile for all sequences associated to the gene (gene body plus 1 kb upstream excluding coding sequence) or by taking the maximum CI score associated with a gene. The two methods yielded similar results (Spearman's rho 0.7861143). The summation method was chosen as it provided higher resolution gene-level CI scores since many genes could have identical maximum CI values. The 90 percentile threshold chosen for the summation method was chosen based on the observation that the break point in normalized CI score is present in the range of 80-90 percentile.

As a null model of DSX conservation, 100 random motifs were generated by randomly shuffling the 13 positions of the DSX PWM. For each of 100 shuffled PWMs, we identified sites in all species with positive PWM scores and calculated the site-level and gene-level CI scores using the same method by restricting the edit distance of the corresponding invariant positions in shuffled motifs ≤ 1.

The normalized site level CI scores (Figure 2A) were calculated by subtracting median CI score of shuffled motifs from DSX CI score as follows: DSX sites are sorted and divided into 1000 bins of equal number of sites and moving median of CI scores is calculated in each window of all 10 consecutive bins. The moving median of CI scores of all 100 shuffled motifs is also obtained by considering the same number of sites as DSX motifs in each window. The median of CI scores of the 100 shuffled motifs is then computed in each window and used to subtract from DSX CI scores.

Conservation Analysis using PhastCons

PhastCons first performs multiple alignments over 15 species and uses two-state phylogenetic hidden Markov model (phylo-HMM) to predict conserved elements. PhastCons scores (Felsenstein and Churchill, 1996) were downloaded from UCSC (WIB files from http://hgdownload.cse.ucsc.edu/gbdb/dm3/multiz15way/wib; SQL table dump from http://hgdownload.cse.ucsc.edu/goldenPath/dm3/database/phastCons15way.txt.gz). The UCSC program hgWiggle (from http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64) was used to convert each WIB file into a WIG file for each chromosome containing phastCons scores for the 15-way multiple alignment performed by UCSC (see http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=multiz15way for processing details). WIG files were manually edited to remove duplicate lines that prevented conversion to bigWig format, and the edited files were converted to bigWig using the UCSC program wigToBigWig. Data from all chromosomes were then concatenated into a single file with the UCSC program bigWigCat. A site Level PhastCons score is obtained by calculating the average of PhastCons scores of all positions with the UCSC program bigWigAverageOverBed.

Conservation of DSX DNA binding domain and splicing patterns. Exons coding for DSX were identified in 19 non-melanogaster species by liftover of the *Drosophila* *melanogaster* DSX-encoding exons as described above. Exons were translated *in silico* using ExPASy (Artimo et al., 2012). The DM DNA binding domain was defined as *D. melanogaster* DSX amino acids 39-105 (Zhang et al., 2006). Multiple species alignment of DSX DM domain protein sequence was performed with Clustal Omega 1.1.0 (Sievers et al., 2011). A single nucleotide (G) deletion was identified in the DM domain of the DSX coding sequence in the genome of *D. mojavensis* between nucleotide positions 23339379 - 2333980 of scaffold_6540 (Genbank accession NW_001979112.1). This deletion may be due to an assembly artifact as the deleted G was present in RNA-seq data from *D. mojavensis* (Chen et al., 2014).

DMRT1 Orthologs. 952 *D. melanogaster* orthologs of mouse DMRT1 targets (Murphy et al., 2010) were identified by converting the 1,439 gene names to Ensembl IDs using the Jackson Labs conversion tool (Blake et al., 2014). These Ensembl IDs were uploaded into Ensembl Biomart (Flicek et al., 2013), and orthologous *D. melanogaster* genes were obtained using the multispecies comparison tool with Ensembl 73 Genes.

Occupancy score clustering and statistical analysis. The ranked occupancy scores for all genes annotated in FlyBase 5.46 in all six occupancy experiments were clustered using the kmeans package in R (R Core Team, 2013)**.** The Kruskal-Wallis test was performed in R (kruskal.test function) and used to test for significant differences in the distributions of gene-level DSX PWM scores and gene-level conservation for genes in each cluster. The hypergeometric test was used to test for significant enrichment of *D. melanogaster* orthologs of mouse DMRT1 targets among the genes in each occupancy cluster.

ChIP-seq correlations with multiple independent occupancy data sets

Called peaks for 255 available ChIP-chip and ChIP-seq experiments annotated in Slattery et al's supplemental table S1 (Slattery et al., 2014) as well as those available in other modENCODE accessions were downloaded in GFF, BED, plain text tables, or tarballs from either individual GEO entries, authors' websites, or the modENCODE FTP site (Table S6). Files for called peaks were converted to a uniform BED3 format, and peaks from the Furlong group and BDTNP project were lifted over to the dm3 assembly. Peaks from the occupancy experiments from this study were also included for a total of 261 sets of called peaks. In each file, features that overlapped by at least 1 bp were merged using BEDTools v2.19.0. Since some experiments only included euchromatic chromosomes, for consistency all files were filtered to only retain peaks from chromosomes 2L, 2R, 3L, 3R, 4, and X. Details on data acquisition and processing can be found in Table S6.

For each pairwise comparison, one file was arbitrarily set as the query and one file as the reference. P-values for each peak in the query were calculated following the IntervalStats method of (Chikina and Troyanskaya, 2012), representing that peak's overall proximity to the reference. The similarity of the query to the reference was then summarized in a single number by taking the fraction of all features in the query with p-values < 0.05. Since this metric is not symmetric, the fraction of features with p-values < 0.05 was also calculated after swapping the query and reference. The final result is a 261 x 261 similarity matrix of pairwise comparisons with each value representing the fraction of all peaks with P < 0.05 in the query (Figure S3).

Defining DSX-occupied genes. 3,717 genes were defined as being occupied by DSX in all experimental data sets by taking the union of all genes in the 90th percentile of gene-level occupancy scores from each individual occupancy data set (Table S1). The 90th percentile cutoff was selected for use in the analysis following examining the relationship between gene-level occupancy score and gene-level PWM score. The best break point in these plots were where CI score was 80-90 percentile. 2,668 genes were defined as being occupied by DSX in fat body samples by taking the union of the genes in the 90th percentile of the three fat body occupancy data sets.

Analysis of occupancy and conserved motifs relative to gene features. Using all genes > 1 kb annotated in FlyBase 5.46, we binned loci into five regions: upstream (1.5 kb upstream of 5'-most promoter, 1-bp bins); 5' (500 bp downstream of promoters, 1-bp bins); gene body (1000 bins; bin size varies); 3' (500 bp, 1-bp bins); and downstream (1.5kb, 1-bp bins). The number of DSX peaks (union of peaks from all occupancy experiments; median size 1kb) were enumerated in each bin using metaseq v0.5 (Dale et al., 2014) and averaged across all genes. Values were then normalized by subtracting the minimum and dividing by the maximum.

GOTerm Analyses. Enrichment of gene ontology terms (Table S5) was identified using the Cytoscape app BINGO 3.0.2 (Maere et al., 2005). The genes from each individual occupancy cluster were used as the input dataset, and the total *D. melanogaster* gene set was used as the background file. p-values returned by BINGO are corrected for multiple testing using the Bonferroni method. We considered adjusted p-values < 0.001 as a significant enrichment.

RNA-seq. Fat body tissue was dissected from age-matched adult flies of the genotypes w^{1118}; tra2^{ts2}/tra2^{ts1} (experimental) or w^{1118} (control for *dsx*^{F}->dsx^{M} experiments); for *dsx*^{M}->*dsx*^{F} experiments the genotypes were: *y*^{1} *w***; P{w*$^{+mc}$*=UAS-Tra.F}20J7; P{w*$^{+mc}$*=tubP-GAL80*^{ts}*}7/P{w*$^{+mc}$*=tubP-*GAL4*}LL7* (experimental) or *P{w*$^{+mc}$*=tubP-GAL80*^{ts}*}7/P{w*$^{+mc}$*=tubP-*GAL4*}LL7* (control). XX; *tra2*^{ts} flies were morphologically female and fertile while maintained at 18°C indicating that sufficient DSX^{F} activity existed to support female-specific development and physiology. Similarly, XY; UAS-*tra*^{F}*/+*; *tub*-GAL4/*tub*-GAL80^{ts} flies are phenotypically male and fertile when grown at 18°C. All samples were raised at 18°C until 5 days after eclosion when adults were shifted to either 29°C (for *tra2*^{ts}) or 30°C (for *UAS-TraF*) for 0, 12, or 24 hours. Total RNA was extracted from fat body dissected at room temperature (placed on ice after 30 minutes) using TRIzol Reagent following manufacturer's protocol (Ambion Life Technologies, Carlsbad, CA, USA). Purified RNA was treated with DNAse I following manufacturer's protocol (New England Biolabs, Ipswich, MA, USA) and purified again using phenol:chloroform extraction followed by ethanol precipitation. Duplicate RNA-seq libraries were constructed from 200ng total RNA from independent dissection of each sample using the TruSeq RNA Sample Preparation v2 high-throughput (HT) protocol (Illumina, San Diego, CA, USA, 2011). Libraries were sequenced on the HiSeq 2000 machine following a 76 bp single-end protocol (Illumina, San Diego, CA, USA).

Reads were generated using the Illumina pipeline software 12.4.2 for all samples excluding control male t=24hr replicate 1 which used pipeline 1.13.48 (re-sequenced due to poor initial sequence quality). Reads passing the Illumina chastity filter were mapped to the *D. melanogaster* genome and assigned to gene models using Tophat 1.4.1 (Trapnell et al., 2009) with a gtf file provided (-G, FlyBase r5.46, see below) and default settings except for the following; minimum intron length was set to 42bp (-i 42) and the maximum multihits was set to 1 (-g 1). Transcript abundance was determined using Cufflinks 2.1.1 (Trapnell et al., 2013) with maximum bundle fragments set to 10,000,000 (--max-bundle-frags 10000000) due to high read density at the *Yp* loci, and upper quartile normalization was used (-N).

To generate a gtf file for Tophat and Cufflinks analyses, the FlyBase GFF annotations (release 5.46) were downloaded from FlyBase as a GFF3 format file. This file was filtered to remove any features on chromosomes Uextra or dmel_mitochondrion_genome as well as the following feature types: enhancer, regulatory_region, exon_junction, rescue_fragment, sequence_variant, pcr_product, point_mutation, orthologous_region, TF_binding_site, protein, chromosome, uncharacterized_change_in_nucleotide_sequence, origin_of_replication, chromosome_band, tandem_repeat, insulator, polyA_site, deletion, BAC_cloned_genomic_insert, complex_substitution, RNAi_reagent, transposable_element_insertion_site, repeat_region, oligonucleotide, breakpoint, transposable_element, chromosome_arm, protein_binding_site, orthologous_to, silencer, region, insertion_site, mature_peptide, DNA_motif, syntenic_region. A leading "chr" was prepended to each chromosome name for consistency with the genomic assembly sequence files used. The filtered GFF file was imported into a sqlite3 database using gffutils (https://github.com/daler/gffutils), which represents the hierarchical relationships between features as defined in GFF files. For each gene, the "child" transcripts were retrieved from the database, and for each transcript, each child that was either an exon or CDS was retrieved. For each of these exon and CDS features, the gene ID, gene name, transcript ID, and transcript type information were attached to the feature, and it was exported as a GTF format line. The resulting GTF file of exon and CDS features was then run through the gffread program (part of the cufflinks suite) as the command "gffread -E $infile -T -F -o- > $outfile" in order to confirm that the file contained no errors that would prevent downstream use by Cufflinks (Trapnell et al., 2012).

Background expression levels were estimated based on reads in intergenic space (Zhang et al., 2010). Genomic regions that are not located within an annotated gene (FlyBase 5.46), nor within +/- 500 bp flanking an annotated gene, were binned into 199 bp windows (= median of all *D. melanogaster* exons), and FPKM values for these intergenic bins were calculated using the Tophat/Cufflinks parameters used for genes. To prevent loss of mapping between bins, the original intergenic bins were shifted by 100bp and any bin entering a non-intergenic space was removed. The median expression value for all intergenic bins was 1.84839375, and all experimental FPKM values at or below this cutoff were converted to zero. Further, all genes with FPKM=0 in all experimental and control conditions were removed from further analyses. After background correction, k-means clustering of FPKM values was performed using the kmeans package in R. The optimal k value (k=9) was determined by maximizing both k and average silhouette width. Counts of *dsx*^{M} and *dsx*^{F} splice junctions were obtained using Spanki 0.4.2 (Sturgill et al., 2013).

GEO Accession Numbers. DSX occupancy data (ChIP-seq, DamID-seq, DamID-array) and RNA-seq data are available under GEO series accession GSE49480. Probe sequence, probe position information, and array details are available under GEO accession GPL10639.

Supplemental References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., *et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology *11*, R106.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq – A Python framework to work with high-throughput sequencing data. bioRxiv preprint.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., *et al.* (2012). ExPASy: SIB bioinformatics resource portal. Nucleic acids research *40*, W597-603.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B (Methodological) *57*, 289-300.

Bischof, J., Maeda, R.K., Hediger, M., Karch, F., and Basler, K. (2007). An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. Proceedings of the National Academy of Sciences of the United States of America *104*, 3312-3317.

Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., and Mouse Genome Database, G. (2014). The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic acids research *42*, D810-817.

Chen, Z.X., Sturgill, D., Qu, J., Jiang, H., Park, S., Boley, N., Suzuki, A.M., Fletcher, A.R., Plachetzki, D.C., FitzGerald, P.C., *et al.* (2014). Comparative validation of the D. melanogaster modENCODE transcriptome annotation. Genome research *24*, 1209-1223.

Chikina, M.D., and Troyanskaya, O.G. (2012). An effective statistical evaluation of ChIPseq dataset similarity. Bioinformatics *28*, 607-613.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422-1423.

Dale, R.K., Matzat, L.H., and Lei, E.P. (2014). metaseq: a Python package for integrative genome-wide analysis reveals relationships between chromatin insulators and associated nuclear mRNA. Nucleic acids research *42*, 9158-9170.

Drosophila 12 Genomes, C., Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., *et al.* (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203-218.

Felsenstein, J., and Churchill, G.A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. Molecular biology and evolution *13*, 93-104.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., *et al.* (2013). Ensembl 2013. Nucleic acids research *41*, D48-55.

Garrett-Engele, C.M., Siegal, M.L., Manoli, D.S., Williams, B.C., Li, H., and Baker, B.S. (2002). intersex, a gene required for female sexual development in Drosophila, is expressed in both sexes and functions together with doublesex to regulate terminal differentiation. Development *129*, 4661-4675.

Greil, F., Moorman, C., and van Steensel, B. (2006). DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase. Methods in enzymology *410*, 342-359.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome biology *8*, R24.

Harris, R. (2007). Improved pairwise alignment of genomic DNA. Ph.D. Thesis. The Pennsylvania State University.

Jemc, J.C., Milutinovich, A.B., Weyers, J.J., Takeda, Y., and Van Doren, M. (2012). raw Functions through JNK signaling and cadherin-based adhesion to regulate Drosophila gonad morphogenesis. Developmental biology *367*, 114-125.

Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature biotechnology *26*, 1351-1359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696-1697.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics *21*, 3448-3449.

Marygold, S.J., Leyland, P.C., Seal, R.L., Goodman, J.L., Thurmond, J., Strelets, V.B., and Wilson, R.J. (2013). FlyBase: improvements to the bibliography. Nucleic acids research *41*, D751-757.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B.*, et al.* (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic acids research *41*, D64-69.

Murphy, M.W., Sarver, A.L., Rice, D., Hatzi, K., Ye, K., Melnick, A., Heckert, L.L., Zarkower, D., and Bardwell, V.J. (2010). Genome-wide analysis of DNA binding and transcriptional regulation by the mammalian Doublesex homolog DMRT1 in the juvenile testis. Proceedings of the National Academy of Sciences of the United States of America *107*, 13360-13365.

Ni, J.Q., Zhou, R., Czech, B., Liu, L.P., Holderbaum, L., Yang-Zhou, D., Shim, H.S., Tao, R., Handler, D., Karpowicz, P.*, et al.* (2011). A genome-scale shRNA resource for transgenic RNAi in Drosophila. Nat Methods *8*, 405-407.

Ouyang, Z., Zhou, Q., and Wong, W.H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America *106*, 21521-21526.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) *26*, 841-842.

R Core Team (2014). R: A language and environment for statistical computing. http://www.R-project.org

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P.*, et al.* (2005). Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome research *15*, 1-18.

Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R.*, et al.* (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature *481*, 389-393.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology *7*, 539.

Slattery, M., Ma, L., Spokony, R.F., Arthur, R.K., Kheradpour, P., Kundaje, A., Negre, N., Crofts, A., Ptashkin, R., Zieba, J.*, et al.* (2014). Diverse patterns of genomic targeting by transcriptional regulators in Drosophila melanogaster. Genome research *24*, 1224-1235.

Southall, T.D., and Brand, A.H. (2007). Chromatin profiling in model organisms. Briefings in functional genomics & proteomics *6*, 133-140.

Sturgill, D., Malone, J.H., Sun, X., Smith, H.E., Rabinow, L., Samson, M.L., and Oliver, B. (2013). Design of RNA splicing analysis null models for post hoc filtering of Drosophila head RNA-Seq data with the splicing analysis kit (Spanki). BMC bioinformatics *14*, 320.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology *31*, 46-53.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.

Yi, W., and Zarkower, D. (1999). Similarity of DNA binding and transcriptional regulation by Caenorhabditis elegans MAB-3 and Drosophila melanogaster DSX suggests conservation of sex determining mechanisms. Development (Cambridge, England) *126*, 873-881.

Zhang, W., Li, B., Singh, R., Narendra, U., Zhu, L., and Weiss, M.A. (2006). Regulation of sexual dimorphism: mutational and chemogenetic analysis of the doublesex DM domain. Molecular and cellular biology *26*, 535-547.

Zhang, Y., Malone, J.H., Powell, S.K., Periwal, V., Spana, E., Macalpine, D.M., and Oliver, B. (2010). Expression in aneuploid Drosophila S2 cells. PLoS biology *8*, e1000320.