doi:10.1038/nature19338

PionX sites mark the X chromosome for dosage compensation

Raffaella Villa[1], Tamas Schauer[1], Pawel Smialowski[2], Tobias Straub[2] & Peter B. Becker[1]

The rules defining which small fraction of related DNA sequences can be selectively bound by a transcription factor are poorly understood. One of the most challenging tasks in DNA recognition is posed by dosage compensation systems that require the distinction between sex chromosomes and autosomes. In *Drosophila melanogaster*, the male-specific lethal dosage compensation complex (MSL-DCC) doubles the level of transcription from the single male X chromosome, but the nature of this selectivity is not known[1]. Previous efforts to identify X-chromosome-specific target sequences were unsuccessful as the identified MSL recognition elements lacked discriminative power[2,3]. Therefore, additional determinants such as co-factors, chromatin features, RNA and chromosome conformation have been proposed to refine targeting further[4]. Here, using an *in vitro* genome-wide DNA binding assay, we show that recognition of the X chromosome is an intrinsic feature of the MSL-DCC. MSL2, the male-specific organizer of the complex, uses two distinct DNA interaction surfaces—the CXC and proline/basic-residue-rich domains—to identify complex DNA elements on the X chromosome. Specificity is provided by the CXC domain, which binds a novel motif defined by DNA sequence and shape. This motif characterizes a subclass of MSL2-binding sites, which we name PionX (pioneering sites on the X) as they appeared early during the recent evolution of an X chromosome in *D. miranda* and are the first chromosomal sites to be bound during *de novo* MSL-DCC assembly. Our data provide the first, to our knowledge, documented molecular mechanism through which the dosage compensation machinery distinguishes the X chromosome from an autosome. They highlight fundamental principles in the recognition of complex DNA elements by protein that will have a strong impact on many aspects of chromosome biology.

Previous work suggested that MSL2 may tether the MSL-DCC to DNA and that an intact CXC domain is required for X-chromosome discrimination[5,6]. To assess the DNA-binding specificity intrinsic to MSL2 comprehensively, we surveyed the *Drosophila* genome for MSL2-binding sites *in vitro* by DNA immunoprecipitation (DIP)[7,8]. Recombinant MSL2 was incubated with sheared genomic DNA (gDNA) purified from male *Drosophila* S2 cells. MSL2-bound DNA was recovered and sequenced.

Considering the lack of X-chromosome binding selectivity seen in previous *in vitro* studies, we did not expect to find that MSL2 preferentially retrieved DNA from distinct genomic loci, with a notable enrichment of sequences from the X chromosome (Fig. 1a). On the X chromosome, the MSL2 binding pattern was remarkably similar to the *in vivo* pattern that marks the positions of high-affinity binding sites (HAS; or chromatin entry sites) of the MSL-DCC (Fig. 1b). A total of 57 DIP sites coincided with *in vivo* HAS, although they show different signal intensities (Extended Data Fig. 1a, b). The results were similar if DIP followed by sequencing (DIP–seq) was performed with gDNA extracted from female cells or synthesized *in vitro* by whole-genome amplification (excluding the contribution of male-specific RNA contaminants or DNA modifications) (Extended Data Fig. 1c, d). It therefore appears that recombinant MSL2 has an intrinsic ability to enrich X-chromosomal sequences from complex genomic DNA.

We next assessed the contribution of the three known MSL2 domains to DNA binding (Fig. 2a and Extended Data Fig. 2a). Deletion of the RING finger domain that mediates MSL2 interaction with MSL1 (ref. 9) and contains E3 ligase activity[10] had no obvious effect (Fig. 2b, c). Unexpectedly, however, deletion of a region rich in proline and basic amino acid residues (the Pro/Bas domain) that may bind RNA[11] resulted in the complete loss of DNA binding (Fig. 2b).

Upon deletion or mutation of the CXC domain, binding to a subset of sites was much reduced (Fig. 2b, c). Statistical analyses revealed 56 regions that specifically required a functional CXC domain for binding. Notably, these ‘CXC-dependent’ sites displayed a higher enrichment on the X chromosome (Fig. 2d and Extended Data Fig. 2b). A total of 37 sites mapped to MSL2 *in vivo* peaks (HAS) on

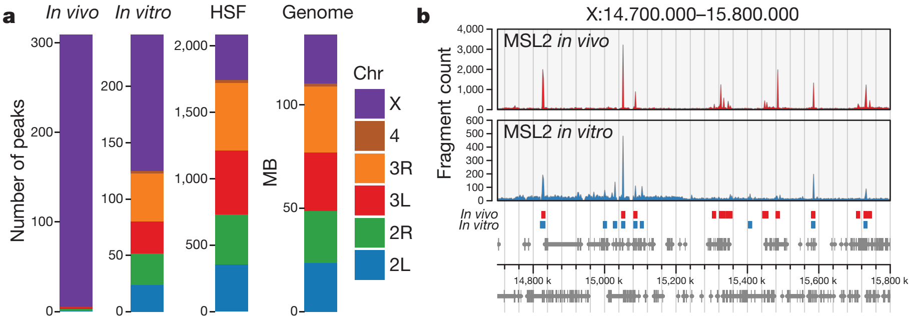


Figure 1 | Genome-wide MSL2 *in vitro* binding partially recapitulates the *in vivo* pattern. a, Chromosomal distribution of robust *in vivo* and *in vitro* MSL2 binding peaks, each determined by two independent experiments. The DIP–seq profile of heat shock factor (HSF)[8] and the relative size of the chromosomes (genome) serve as references for uniform distribution. **b**, Representative profiles of MSL2 chromatin immunoprecipitation with sequencing (ChIP–seq) and DIP–seq experiments in a 1.3-Mb window on the X chromosome. Red and blue bars indicate the positions of robust peaks. Gene models are depicted in grey at the bottom.

[1]Division of Molecular Biology, Biomedical Center and Center for Integrated Protein Science Munich, Ludwig-Maximilians-University, 82152 Planegg-Martinsried, Germany. [2]Bioinformatics Unit, Biomedical Center, Ludwig-Maximilians-University, 82152 Planegg-Martinsried, Germany.

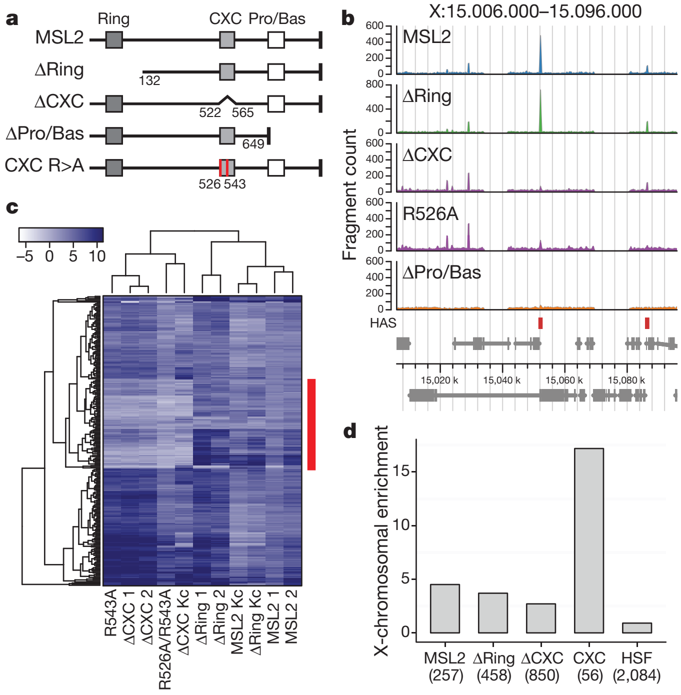


Figure 2 | The CXC domain of MSL2 increases X-chromosomal specificity. a, Linear representation of MSL2 domain organization and mutant proteins assayed in DIP. Point mutants in the CXC domain included a single (R543A) and double (R526A/R543A) mutant version. **b**, Representative DIP–seq profiles of wild-type and mutated MSL2 proteins in a region around the *hiw* gene. Red bars indicate HAS. **c**, Clustered heat map of DIP signals in all MSL2 *in vitro* peak regions. The red bar indicates a group of sites that show a prominent loss in signal upon CXC mutation. For some proteins, two independent replicates are shown. **d**, X-chromosomal enrichment over autosomes of robust wild-type and mutant MSL2 DIP–seq peaks. CXC indicates the peaks that significantly (false discovery rate (FDR) < 0.05) lose binding upon CXC depletion or mutation. The *x*-axis labels indicate the total number of peaks for each target in brackets.

the X chromosome, and 2 sites corresponded to rare cases of autosomal sites that show MSL2 enrichment *in vivo* (Extended Data Fig. 1e, Extended Data Table 1 and Supplementary Table 1). Our data suggest that MSL2 interacts with DNA via two domains, CXC and Pro/Bas, and that the CXC domain is the major determinant of the selectivity for the X chromosome. While binding-site specificity can be achieved by cooperation between different transcription factors[12], our finding suggests that cooperation between two different DNA-binding surfaces within this one protein may also refine its overall binding specificity.

Sequence analyses within CXC-dependent and CXC-independent binding sites for MSL2 yielded two distinct motifs. Whereas the CXC-independent binding sites shared low-complexity GA repeats (Extended Data Fig. 3a), the CXC-dependent peaks centre around a more complex variation of the MSL response element (MRE), with a notable 5′ extension (Figs 3a, 4c and Extended Data Fig. 3c). Remarkably, this novel motif can predict *in vivo* MSL2 binding (HAS) better than the MRE, as its position weight matrix (PWM) is superior in classifying whether MRE hit regions overlap HAS (Fig. 3b and Extended Data Fig. 3b, d). Applying low thresholds ($q \leq 0.2$) we found 2,667 instances of this motif throughout the genome (Supplementary Table 1), with an approximately twofold enrichment on the X chromosome. Higher-scoring matches to the consensus sequence tend to be more strongly enriched on the X chromosome. For example, the 34 best matches are 9.8-fold enriched on the X chromosome (Extended Data Fig. 3e, f). However, 18 of those instances were not bound *in vitro* by MSL2 in a CXC-dependent manner, indicating that the recognition sequence represented by a PWM cannot fully explain this binding mode.

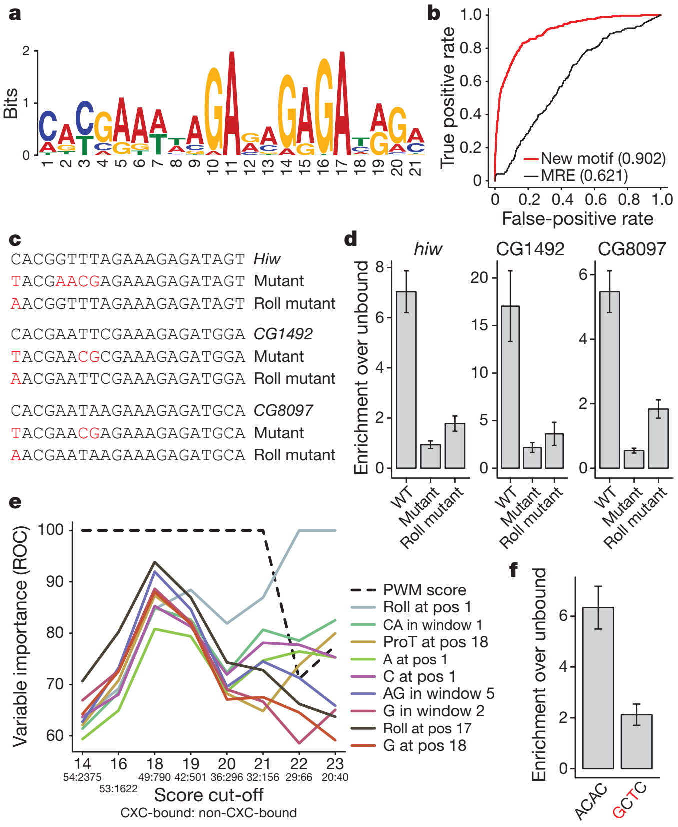


Figure 3 | The CXC domain reads out nucleotide sequence and additional features. a, Motif discovered by the MEME motif-discovery tool in CXC-dependent binding regions (E-value $= 3.9 \times 10^{-158}$). **b**, Receiver operating characteristic (ROC) curves representing the performance of MRE and the new motif PWMs in predicting whether genomic MRE instances (35,659) overlap with a HAS (266). Areas under the curves (AUCs) are provided in brackets. **c**, A list of oligonucleotides used in DIP experiments. Nucleotides highlighted in red are mutations introduced based on the predictions of our classification model. **d**, DIP experiments using synthetic DNA representing wild-type or mutated binding sites in the genomic context. Results from qPCR amplification were normalized for their input and shown as enrichment over an unbound fragment. Data are mean ± s.e.m. for 3 biological replicates. **e**, Individual feature importance evaluated on sets of CXC-dependent motif instances defined by increasing score thresholds. For each feature a ROC analysis on CXC-dependent binding was performed. The AUCs of all features were scaled from 0 to 100 at each threshold level. Only features which ranked at least twice among the top five are reported. Numbers of instances (CXC-bound, non-CXC-bound) are provided underneath the *x*-axis tick labels. **f**, DIP experiments using the wild-type CG1492 sequence and a mutant in which the DNA roll at position +1 was reduced by mutating positions −1 and +2. Results from qPCR amplification were normalized for their input and shown as enrichment over an unbound fragment. Data are mean ± s.e.m. for 4 biological replicates.

PWMs model the base readout of DNA sequences with the implicit assumption that each nucleotide at a given position contributes to binding independently of other positions. Physical interactions of neighbouring base pairs, however, alter the structural conformation of the DNA double helix (often referred to as the DNA shape), which may manifest as variations in the minor groove width, roll, helix twist, and propeller twist. Many proteins depend on both base identity and localized helix shape to recognize their binding site[13–15]. Using a pentamer-based model built from all-atom Monte Carlo simulations of DNA structures[16], we calculated DNA shape parameters at each base position of the low-stringency motif hits, with 20-base-pair (bp) extensions on either side. To complement these position-centred features we also calculated regional mono- and dinucleodide frequencies (*k*-mers) in 4-bp windows along the hit sequences. Principal component analysis (PCA) revealed that a combination of DNA shape and k-mer features was able to separate the two classes of sequences: those that were bound in a CXC-dependent manner (CXC-bound) and those that were not bound in a CXC-dependent manner (non-CXC-bound). Sequences in the latter group were either not bound at all (2,502) or were bound independently of the CXC domain (111) (Extended Data Fig. 4a and Supplementary Table 2). This suggested that at least some of the DNA features might improve binding prediction. Indeed, classification models constructed with our additional feature sets performed much better than a PWM-score model in predicting CXC-dependent binding sites on all motif hits (Extended Data Fig. 4b).

Guided by the good performance of our classification model using both the PWM-hit-score and k-mer features, we predicted mutations that would convert robust CXC-bound sites to non-CXC-bound sites. The model suggested that the best discriminating residues would localize to the 5′ part of the motif and not to the GA-rich region (Fig. 3c). To test these predictions, we modified the DIP experiment by mixing appropriately diluted DNA oligonucleotides, representing either a native site or its mutated version, into the genomic DNA. The efficiency of DNA retrieval of experimental oligonucleotides and control genomic loci was quantified by quantitative PCR (qPCR). The results confirmed our predictions (Fig. 3d), leading us to conclude that the main determinants for CXC-dependent binding reside within the first eight bases at the 5′ end of the consensus motif. Notably, this is the part of the motif that diverges most from the MRE.

To achieve a switch in the predicted class from CXC-bound to non-bound in the context of the unbalanced data set of low-threshold instances (54 bound sites, 2,613 non-bound sites) required at least three mutations. This inevitably affected the motif score, making it difficult to distinguish the effects of base and shape readout. To reduce class imbalance and to evaluate the contribution of shape features to CXC-dependent binding of sequences with high similarity to the motif consensus, we limited the analysis to fewer sites through the stepwise increase of motif score thresholds. Figure 3e reveals the relative success of the PWM score compared with a selection of additional features in predicting CXC-dependent binding. In a more balanced data set of motifs consisting of the 95 best motif hits (29 sequences CXC-bound, 66 non-bound) the PWM score was no longer a good predictor and DNA shape features became increasingly relevant. In particular, ‘roll at position +1’ (that is, the roll between the first two base pairs of the motif) turned into the best-performing predictor when sequences with PWM scores higher than 21 were considered. We therefore focused on the 34 highest-scoring motif sequences ($q < 0.05$), which are highly enriched on the X chromosome; however, only 16 of them are bound in a CXC-dependent manner by MSL2. We systematically scanned these high-scoring sequences for statistically significant shape differences between the CXC-bound and non-bound classes at any nucleotide position (Extended Data Fig. 4d). The results confirmed that ‘roll at position +1’ was appreciably different between the two classes. To test experimentally the importance of this feature we changed the degree of roll at position +1 from $>4°$ to $< -2°$ by either replacing cytosine at position +1 or the two adenines at positions −1 and +2. These mutations led to a clear reduction in MSL2 binding (Fig. 3c, d, f). We were also able to convert a sequence that was not efficiently bound by MSL2 into one that was by changing the roll at position +1 from $< -1.9°$ to $>4°$ (Extended Data Fig. 4c). Adding the DNA shape feature ‘roll at position +1’ to our PWM-hit-score classification model resulted in substantially improved performance when applied to the complete list of 2,667 motif hits (Extended Data Fig. 4b). We therefore conclude that the ability of MSL2 to distinguish true binding sites from a large collection of irrelevant elements with highly related sequences also relies on structural features.

To investigate further the role of MSL2-binding sites in X chromosome dosage compensation, we first attempted to monitor the interactions of MSL2 with HAS *in vivo*, with minimal contributions from other DCC subunits. Genetic studies had shown that the assembly of a mature MSL-DCC bound to the non-coding *roX* RNA in male flies is compromised by inactivating the RNA helicase maleless (MLE). Under those circumstances, the remaining MSL2–MSL1 sub-complex is bound to a small subset of HAS[17]. We recreated this scenario in S2 cells by using RNA interference (RNAi) against *mle* expression, and found that MSL2 binding was preferentially retained at HAS, corresponding to CXC-dependent binding sites (Extended Data Fig. 5b). The 25 HAS that were most resistant to MLE depletion (Fig. 4a) revealed a shared sequence, bearing a strong resemblance to the CXC-dependent motif (Fig. 4c). By contrast, the 25 sites most sensitive to MLE depletion (bound only by the complete DCC) shared a GA motif similar to the one found in CXC-independent *in vitro* binding sites (Extended Data Fig. 5a). This suggests that under physiological conditions, the MSL2–MSL1 sub-complex directly contacts a subset of HAS in a CXC-dependent manner in the absence of associated protein and RNA subunits.

It is possible that these chromosomal interactions represent an intermediate of MSL-DCC assembly. To test this hypothesis, we initiated

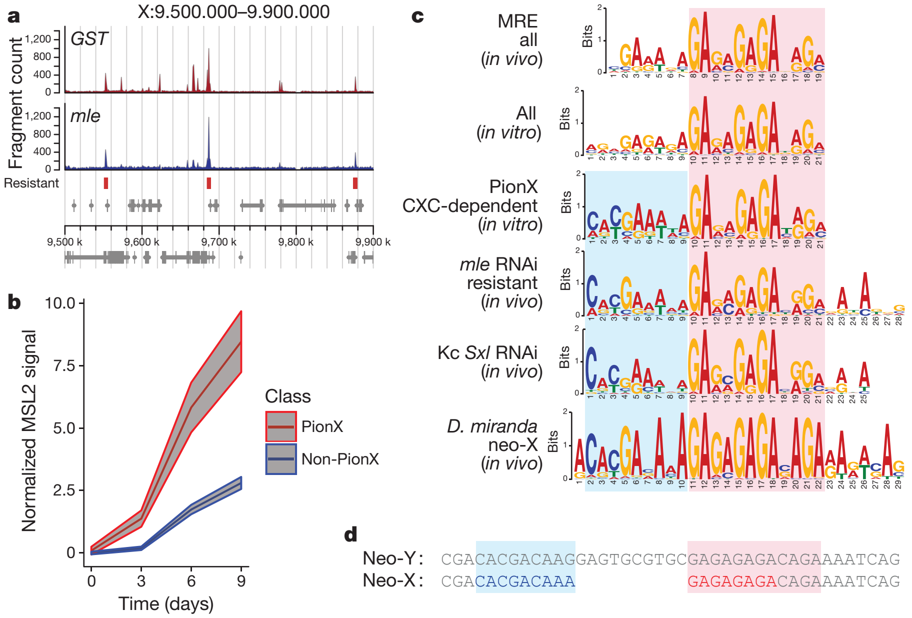


Figure 4 | The CXC-dependent sites are pioneer HAS. a, Representative profiles of MSL2 ChIP–seq from S2 cells treated with RNAi against *GST* (control) or *mle*. Red bars indicate binding sites that are maintained in the absence of MLE. **b**, MSL2 signal on 37 HAS matching CXC-dependent *in vitro* binding sites (PionX) or 272 non-matching ones (non-PionX) during SXL knockdown in Kc cells. Signals were averaged across 4 biological replicates and normalized to the mean signal at time point 0. Curves depict mean and s.e.m. across all sites within one class. **c**, Comparison of motifs found in MSL2-bound regions using different experimental approaches. See main text for details. Shown are the top scoring motifs except for ‘all (*in vitro*)’ which places second after a low-complexity GA-repeat similar to Extended Data Fig. 3a. **d**, Schematic representation of the 10-bp deletion that generated PionX motifs on the *D. miranda* neo-X chromosome[21].

de novo MSL-DCC assembly in female Kc cells by reducing the expression of the sex-lethal gene *Sxl*. The SXL protein prevents MSL2 expression and thus the dosage compensation program in female cells. Upon depletion of SXL (Extended Data Fig. 5c, d), binding of newly expressed MSL2 to CXC-dependent HAS was stronger and occurred earlier when compared to CXC-independent ones (Fig. 4b). Consistent with this finding, hierarchical clustering of MSL2 signals from the common set of Kc and S2 peak regions revealed 30 sites that acquire strong MSL2 binding ability 3 days after SXL depletion (Extended Data Fig. 5d). *De novo* motif discovery on these sites revealed a consensus sequence that resembles the one in the CXC-dependent sites (Fig. 4c). Our data strongly suggest that those sites identified *in vitro* as CXC-dependent are pioneering binding sites for MSL2 *in vivo*. We therefore refer to them as PionX sites, and to their defining motif as the PionX motif.

The notion that PionX sites are important for dosage compensation is further supported by evolutionary considerations. *Drosophila miranda* represents a unique system to study how newly evolving X chromosomes acquired dosage compensation. The *D. miranda* neo-X chromosome is a sex chromosome that began to evolve just 1 million–2 million years ago[18]. Owing to the relatively short evolutionary time span, the neo-X chromosome still retains many autosomal features, but has already acquired partial dosage compensation. Recent work has identified the MSL-DCC-binding sites on all *D. miranda* X chromosomes[19]. *De novo* motif analysis yielded the typical GA-rich MREs for the older, fully compensated X-chromosomal arms XL and XR. Notably, though, the consensus sequence derived from the neo-X chromosome clearly resembled the PionX signature[19] (Fig. 4c).

The neo-Y chromosome originated from the fusion of one Müller-C chromosome to the Y chromosome, resulting in evolutionary pressure on the second Müller-C chromosome to become the neo-X chromosome. We found the PionX motif (but not the MRE) to be particularly enriched on the *D. miranda* neo-X chromosome but not on the related *Drosophlia pseodoobscura* Müller C autosome, supporting the idea that this motif represents a new X-chromosome-specific feature (Extended Data Fig. 5e). Careful comparison of neo-X-chromosome sequences with the homologous regions in *D. pseodoscura* revealed that the novel MSL-DCC-binding sites were acquired by diverse molecular mechanisms, including point mutations and short insertions/deletions of precursor sequences[20]. About half of them originated from precursor sequences contained in a *D. miranda*-specific helitron transposon[21]. The homologous neo-Y helitron does not contain PionX motifs—only precursor sequences in which the 5′ CAC motif and the 3′ GA-rich element are separated. On the neo-X chromosome these two parts are fused by a 10-bp deletion to form PionX consensus motifs[19,21] (Fig. 4d). The insertion of a PionX consensus motif derived from the *D. miranda* neo-X chromosome into an autosome of *D. melanogaster* led to strong, ectopic binding of the MSL-DCC. By contrast, the corresponding homologous neo-Y-chromosome sequence, in which the 5′ and GA sequences are split by a 10-bp insertion, did not recruit the complex[19,21]. A similar experiment used the strongest DCC-binding site from the neo-X chromosome and the corresponding neo-Y-chromosomal fragment, showing MSL-DCC recruitment to the former, but not to the latter[19]. While the neo-Y-chromosomal fragment does not contain a PionX motif, the evolved neo-X chromosome contains nine of them (Extended Data Fig. 5f). Collectively, these observations suggest that PionX motifs play an important role in *de novo* acquisition of dosage compensation.

In summary, we provide three lines of argument suggesting that PionX sites are X-chromosome-specific determinants that function early in the series of events that lead to exclusive targeting of the X chromosome and correct dosage compensation. First, PionX sites are bound by an MSL2–MSL1 sub-complex in the absence of all other subunits, a state that may reflect an early intermediate of MSL-DCC assembly at HAS. Second, PionX sites are the first to be occupied during *de novo* establishment of dosage compensation. Finally, PionX motifs arose during the early phase of neo-X-chromosome evolution in *D. miranda*.

A pertinent conceptual advance from our study is the understanding that not all HAS contain the same amount of information. The subset of PionX sites are not necessarily sites of highest MSL2 occupancy *in vivo* (Extended Data Fig. 1b), but contribute an important qualitative element of X-chromosomal discrimination. This discrimination is not wholly apparent from the consensus motif as it also relies on the shape of the DNA at the MSL2-binding site.

The initial recruitment of MSL2 to PionX sites on the X chromosome may trigger the distribution of the complex to nearby non-PionX HAS within the chromosomal territory, thereby further amplifying the difference in MSL2 occupancy between the X chromosome and the autosomes. It is likely that other factors contribute to the stability of the targeting system *in vivo*, such as the cooperativity of MSL2 domains within what is presumed to be a dimeric complex[22]; the assembly of functional complexes within the X-chromosomal territory owing to transcription of *roX* RNA from the X chromosome[23]; synergistic interactions between different MSL-DCC complexes and with the CLAMP protein at clustered MREs[24]; and a supportive organization of the conformation of the X chromosome[25].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 December 2015; accepted 27 July 2016.
Published online 31 August 2016.

1. Lucchesi, J. C. & Kuroda, M. I. Dosage compensation in *Drosophila*. *Cold Spring Harb. Perspect. Biol*. **7,** a019398 (2015).
2. Alekseyenko, A. A. *et al.* A sequence motif within chromatin entry sites directs MSL establishment on the *Drosophila* X chromosome. *Cell*. **134,** 599–609 (2008).
3. Straub, T., Grimaud, C., Gilfillan, G. D., Mitterweger, A. & Becker, P. B. The chromosomal high-affinity binding sites for the *Drosophila* dosage compensation complex. *PLoS Genet*. **4,** e1000302 (2008).
4. McElroy, K. A., Kang, H. & Kuroda, M. I. Are we there yet? Initial targeting of the Male-Specific Lethal and Polycomb group chromatin complexes in *Drosophila*. *Open Biol*. **4,** 140006 (2014).
5. Fauth, T., Müller-Planitz, F., König, C., Straub, T. & Becker, P. B. The DNA binding CXC domain of MSL2 is required for faithful targeting the Dosage Compensation Complex to the X chromosome. *Nucleic Acids Res*. **38,** 3209–3221 (2010).
6. Zheng, S. *et al.* Structural basis of X chromosome DNA recognition by the MSL2 CXC domain during *Drosophila* dosage compensation. *Genes Dev*. **28,** 2652–2662 (2014).114
7. Gossett, A. J. & Lieb, J. D. DNA immunoprecipitation (DIP) for the determination of DNA-binding specificity. *CSH Protoc*. http://dx.doi.org/10.1101/pdb.prot4972 (2008).
8. Guertin, M. J., Martins, A. L., Siepel, A. & Lis, J. T. Accurate prediction of inducible transcription factor binding intensities *in vivo*. *PLoS Genet*. **8,** e1002610 (2012).
9. Copps, K. *et al.* Complex formation by the *Drosophila* MSL proteins: role of the MSL2 RING finger in protein complex assembly. *EMBO J*. **17,** 5409–5417 (1998).
10. Villa, R. *et al.* MSL2 combines sensor and effector functions in homeostatic control of the *Drosophila* dosage compensation machinery. *Mol. Cell* **48,** 647–654 (2012).
11. Li, F., Schiemann, A. H. & Scott, M. J. Incorporation of the noncoding *roX* RNAs alters the chromatin-binding specificity of the *Drosophila* MSL1/MSL2 complex. *Mol. Cell. Biol.* **28,** 1252–1264 (2008).
12. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527,** 384–388 (2015).
13. Abe, N. *et al.* Deconvolving the recognition of DNA shape from sequence. *Cell* **161,** 307–318 (2015).
14. Joshi, R. *et al.* Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **131,** 530–543 (2007).
15. Zhou, T. *et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl Acad. Sci. USA* **112,** 4654–4659 (2015).
16. Zhou, T. *et al.* DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res*. **41,** W56–62 (2013).
17. Dahlsveen, I. K., Gilfillan, G. D., Shelest, V. I., Lamm, R. & Becker, P. B. Targeting determinants of dosage compensation in *Drosophila*. *PLoS Genet*. **2,** e5 (2006).
18. Lucchesi, J. C. Gene dosage compensation and the evolution of sex chromosomes. *Science* **202,** 711–716 (1978).
19. Alekseyenko, A. A. *et al.* Conservation and de novo acquisition of dosage compensation on newly evolved sex chromosomes in *Drosophila*. *Genes Dev*. **27,** 853–858 (2013).

20. Zhou, Q. *et al.* The epigenome of evolving *Drosophila* neo-sex chromosomes: dosage compensation and heterochromatin formation. *PLoS Biol.* **11,** e1001711 (2013).
21. Ellison, C. E. & Bachtrog, D. Dosage compensation via transposable element mediated rewiring of a regulatory network. *Science* **342,** 846–850 (2013).
22. Hallacli, E. *et al.* Msl1-mediated dimerization of the dosage compensation complex is essential for male X-chromosome regulation in *Drosophila*. *Mol. Cell* **48,** 587–600 (2012).
23. Park, Y., Kelley, R. L., Oh, H., Kuroda, M. I. & Meller, V. H. Extent of chromatin spreading determined by roX RNA recruitment of MSL proteins. *Science* **298,** 1620–1623 (2002).
24. Soruco, M. M. *et al.* The CLAMP protein links the MSL complex to the X chromosome during *Drosophila* dosage compensation. *Genes Dev.* **27,** 1551–1556 (2013).
25. Ramírez, F. *et al.* High-affinity sites form an interaction network to facilitate spreading of the MSL complex across the X chromosome in *Drosophila*. *Mol. Cell* **60,** 146–162 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement number 293948 (PBB) and the German Research Council (CRC 1064, T.St.). We thank P. Korber for suggesting the DIP experiment, F. Gebauer for sharing her antibody against SXL, and S. Schunter, V. Flynn, S. Krause and A. Zabel for technical assistance. We thank N. Gompel for critical reading of the manuscript and S. Krebs and H. Blum for their sequencing service.

Author Contributions R.V. and T.St. conceived the project. R.V. conducted all the experiments except for the ones in Kc cells that were performed by T.Sc. All bioinformatics analyses were conducted by T.St. with the exception of machine learning procedures that were performed by P.S. P.B.B. supervised the experiments and provided intellectual support toward design and interpretation of the results. R.V., T.St. and P.B.B. wrote the manuscript.

Author Information The next-generation sequencing data have been deposited at the Gene Expression Omnibus (GEO) under accession number GSE75033. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.B.B. (pbecker@med.uni-muenchen.de) or T.St. (tstraub@med.uni-muenchen.de).

Reviewer Information *Nature* thanks J. Larsson, R. Rohs and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Protein purification. MSL2 proteins were expressed in Sf21 cells and purified by Flag affinity chromatography as described[5].

Genomic DNA preparation. The pellet from 6×10^7 S2 or Kc cells was suspended in 1.2 ml of lysis buffer (10 mM Tris pH 8, 100 mM NaCl, 25 mM EDTA pH 8, 0.5% SDS, 0.15 mg ml^{-1} of proteinase K) and incubated at 56 °C overnight. After addition of sodium acetate to a final concentration of 0.3 M, the nucleic acids were extracted with phenol–chloroform and precipitated with an equal volume of isopropanol at −20 °C for 1 h. Precipitated nucleic acids were centrifuged and washed with 70% ethanol. Dried pellets were resuspended in TE buffer and sonicated with Covaris AFA S220 (microTUBEs, peak incident power 175 W, duty factor 10%, cycles per burst 200, 430 s) to generate 200-bp fragments. After RNase digestion (0.1 mg ml^{-1}, 1 h at 37 °C), DNA was purified with the GenElute kit (Sigma). Synthetic DNA was produced using the Repli-g kit (Qiagen) with 20 ng of gDNA as starting material.

DIP–seq. DIP–seq experiments were performed as in ref. 7 with few modifications. In brief, 400 ng of gDNA was incubated with either 80 nM of MSL2–Flag or mutated recombinant protein at 26 °C for 30 min in 100 μl of binding buffer (100 mM KCl, 2 mM $MgCl_2$, 2 mM Tris-HCl pH 7.5, 10% glycerol, 10 μM $ZnCl_2$). For DIP experiments in the presence of synthetic DNA, 10 pM of the specified synthetic DNA was added to the reaction. 10% of the reactions was taken as input material and subjected to quantitative PCR and/or deep sequencing. DNA–protein complexes were immunoprecipitated using 15 μl of Flag bead slurry (M2, Sigma) for 15 min at room temperature and washed twice with 100 μl of binding buffer to eliminate unbound DNA. After digestion with proteinase K (0.5 mg ml^{-1}, 1 h at 56 °C), DNA was purified with the GenElute kit (Sigma) and subjected to qPCR and/or deep sequencing. The DIP experiments in presence of synthetic DNA were performed using the deltaRING construct (three different protein preps).

Cells, RNAi, ChIP–seq. All cells used in this study were authenticated performing karyotyping and staining for the MSL-DCC and regularly tested for mycoplasma contamination.

Double-stranded RNAi fragments were generated from PCR products obtained using the following oligonucleotides: *mle* RNAi: 5′-TTAATACGACTCACTATAGGGAGAATGGATATAAAATCTTTTTTGTACCAATTTTG-3′; 5′-TTAATACGACTCACTATAGGGAGAACAGGGCGCATGACTTGCT-3′. *Sxl* RNAi-1: 5′-TAATACGACTCACTATAGGGAGAGATCACAGCCGCTGTCC-3′; 5′-TAATACGACTCACTATAGGGAGATACCGAATTAAGAGCAAATAATAA-3′. *Sxl* RNAi-2: 5′-TAATACGACTCACTATAGGGAGACCCTATTCAGAGCCATTGGA-3′; 5′-TAATACGACTCACTATAGGGAGAGTTATGGTACGCGGCAGATT-3′.

The culture of *Drosophila* male S2 (subclone L2-4, provided by P. Heun), female Kc cells and RNAi against *mle* and *Sxl* were performed as previously described[3] with modifications. At days 3, 6 and 9 after the initial treatment with dsRNA, *Sxl* RNAi cells were split and either collected for ChIP experiments and western blot analyses or treated again with *Sxl* dsRNA. For S2 cells, ChIP experiments were performed using a Covaris AFA S220 (PIP 100 W, DF 20%, 200 CB for 30 min) to generate chromatin fragments of sizes averaging 180 bp. For Kc cells, ChIP experiments were performed as before[26] with modifications. In brief, about 4×10^7 cells were suspended in ice-cold homogenization buffer and fixed with 1% formaldehyde for 10 min at room temperature. After quenching with 125 mM glycine, the cells were collected and washed three times with ice-cold RIPA buffer (1% Triton X-100, 0,1% Na deoxycholate, 0,1% SDS, 140 mM NaCl, 10 mM Tris-HCl ph 8, 1 mM EDTA). Fixed nuclei were sonicated in RIPA buffer with a Covaris sonifier (PIP: 140, DF 20%, CB: 200) for 30 min.

Antibodies. MSL2, MLE and Lamin antibodies were previously described[10]. The SXL antibody was obtained from F. Gebauer.

Library preparation and sequencing. The Diagenode MicroPlex library kit was used to prepare libraries from 1–2 ng of input, DIP or ChIP DNA quantified using the Qubit dsDNA HS Assay kit (Life Technologies Q32851). The libraries were sequenced on a HighSeq 1500 (Illumina) instrument to yield roughly 15 million–25 million reads of 50-bp single-end sequences per sample.

Oligonucleotides. Double-stranded synthetic DNA fragments were obtained by annealing equimolar concentrations (10 μM) of complementary oligonucleotides. All oligonucleotides used in the DIP studies are listed in Extended Data Table 2.

Data analysis. If not indicated otherwise, data were processed using R (http://www.r-project.org) or Bioconductor (http://www.bioconductor.org) and function calls with default parameters. For hierarchical clustering of binding sites based on MSL2 signals, we applied the complete method on Euclidean distances.

Read processing, coverage and normalized coverage. Sequence reads were aligned to the *D. melanogaster* release 6 reference genome using Bowtie[27] version 1.1.1 allowing only for single matches to the reference (parameter –m 1). We extended the matched reads to a total of 200 bp and calculated for each sample a per-base genomic coverage vector by cumulating the total spans of all sequenced fragments.

We defined target signal enrichment as the standardized difference between normalized immunoprecipitate and corresponding normalized input coverage using:

$$\text{normalized coverage}_i = \arcsin\left(\sqrt{\frac{\text{coverage}_i}{\sum_{i=1}^{n} \text{coverage}_i}}\right)$$

in which i denotes genomic position, and coverage denotes number of fragments covering i.

Peak calling, definition of robust peak sets and chromosomal enrichment. Peaks were called using Homer[28] findPeaks version 4.7.2 with the parameters: style = factor; size = 200; fragLength = 200; inputFragLength = 200 and C = 0. All peaks were called using the corresponding input samples as controls. We defined peaks as robust if the region was called in at least two biologically replicated samples. X-chromosomal enrichment describes the ratio of X-chromosomal peak density to autosomal peak density, with density being the number of peaks divided by length of chromosome(s).

Definition of CXC-dependent sites (PionX). We first defined a robust set of *in vitro* MSL2-binding regions by combining the peaks called in the two MSL2 DIP–seq experiments performed on S2 gDNA and the one performed on Kc gDNA. We calculated the average signal enrichment over the input for the profiles described in Fig. 2c. We then tested for signal differences in samples with intact CXC domain against the ones with a deleted or mutated CXC domain using a linear model (R package, limma) including the cell type (origin of gDNA) as random effect. CXC-dependent sites were defined with an FDR threshold of < 0.05 and a fold change < 0.

***De novo* motif discovery and genome-wide motif searches.** We searched for enriched motifs in peak regions using MEME[29], with the zero or one occurrence per sequence (zoops) model, except for searches in *D. Miranda*. Here we applied the any number of repetitions (anr) model, given the extreme amplification of motifs in some of the peak regions. Genome-wide searches were performed with FIMO[29] version 4.10.0 and an initial *P*-value cutoff of 1×10^{-5}.

Definition of HAS and MRE. MSL2 *in vivo* peaks were called on two published high-quality profiles (GEO accession codes GSM929148 and GSM929149). A total of 309 overlapping peak regions (304 on the X chromosome, 5 on the autosomes) were defined as HAS. MEME-based *de novo* motif discovery using the zoops model yielded a consensus sequence that we refer to as MRE. This MRE closely matches the original definition[2,3] (Extended Data Fig. 3c).

Performance comparison of MRE and PionX PWM. On each of the genome-wide 36,410 MRE hits we calculated the score of the PionX PWM after extending the hit region by 2 bp 5′ to the MRE consensus. We then determined the overlap of the hit regions with the 309 HAS. If more than one MRE hit matched to the same HAS, we kept the hit with the highest PionX score. We determined the ROC of each PWM by continuous thresholding of the respective scores using the match to a HAS as response. The analysis comprises 35,659 instances mapping to 266 HAS.

Definition of *mle* RNAi-resistant sites. Average MSL2 enrichment was calculated on all 309 robust MSL2 *in vivo* binding sites (HAS) in control and MLE RNAi ChIP–seq samples (3 biological replicates each). We then tested for difference in signals between the two experimental groups using limma (R). We defined the most resistant as the 25 sites with the highest moderated *t* values. Accordingly, the 25 sites with the lowest *t* values were defined as most sensitive sites.

Definition of strong MSL2-binding sites upon *Sxl* RNAi in Kc cells. Average MSL2 enrichment was calculated for all 309 HAS including the 90 robust peak regions arising in Kc cells (4 biological replicates for the *Sxl* RNAi at each time point, 2 for the controls). The signals were clustered hierarchically. Two clusters with the strongest gain were combined to constitute a set of 30 sites.

DNA shape calculation and extended feature description. The initial set of regions subjected to extended feature analysis was defined by applying low thresholds ($q \leq 0.2$) on FIMO motif searches for the PionX motif. We obtained 2,667 hits (Supplementary Table 1), 54 of which were bound in a CXC-dependent manner by MSL2, 111 bound in a CXC-independent manner by MSL2 and 2,502 were unbound in our *in vitro* DIP–seq experiments. We refer to unbound instances as well as CXC-independently bound ones as non-CXC-bound.

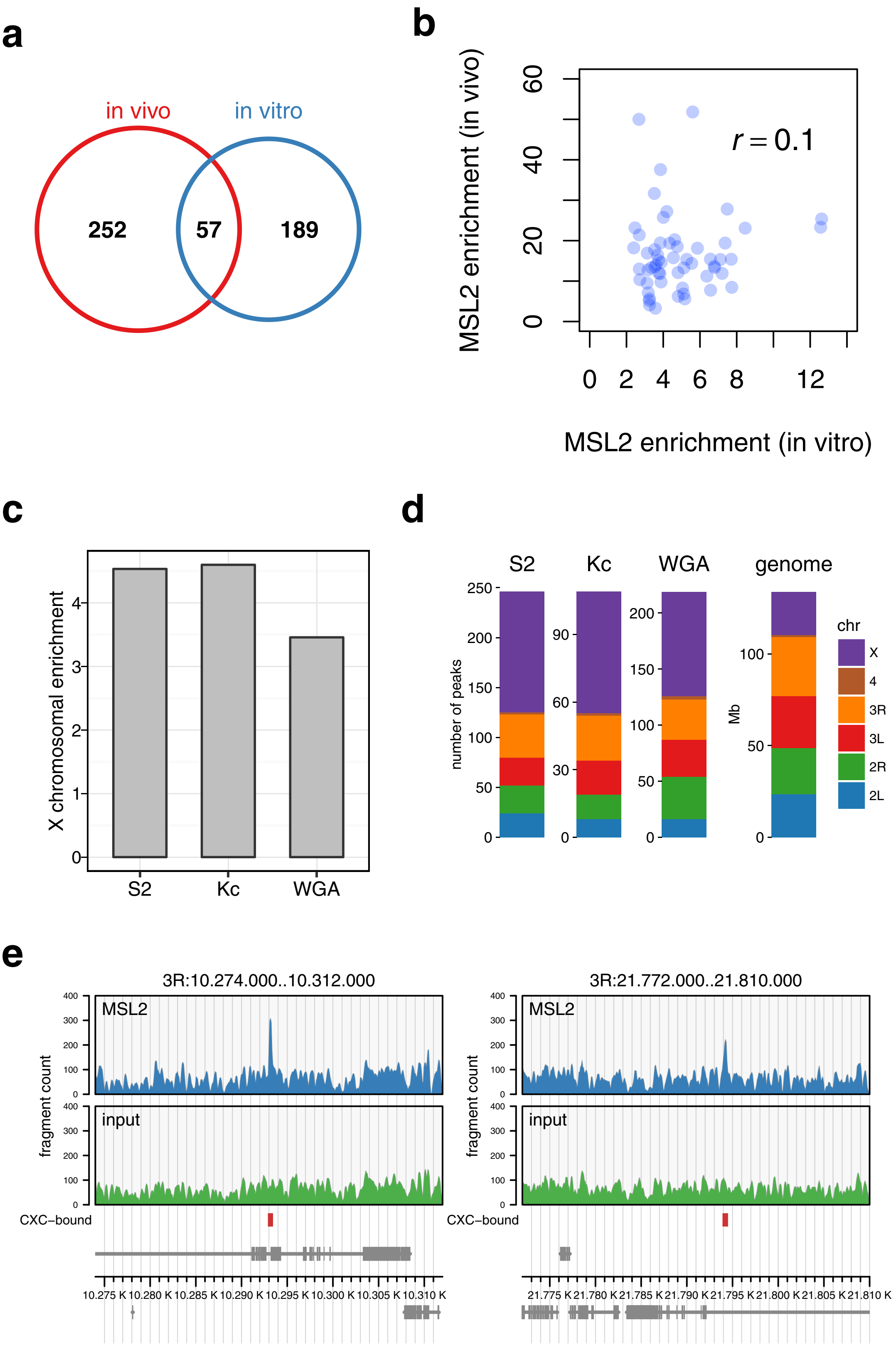
DNA shape parameters were calculated with the DNAshape program[16]. While minor groove width and propeller twist refer to specific nucleotide positions, roll and helix twist specify structural parameters between adjacent bases. For the sake of simplicity, we assigned the values of roll and helical twist to the preceding base ('roll at position 1' actually specifies the roll between the bases of nucleotides at position 1 and 2).

The total set of features considered in this study (with the number of variables in brackets) comprise: the PWM-hit-score (1), nucleotide composition at each position from −20 bp to +20 bp around the motif (244), minor groove width from −18 bp to +18 bp around the motif (57), roll from −19 bp to +18 bp around the motif (58), helix twist from −19 bp to +18 bp around the motif (58), propeller twist from −18 bp to +18 bp around the motif (57), nucleotide frequencies in six consecutive 4-bp windows starting at position 1 of the motif (24), dinucleotide frequencies in six consecutive 4-bp windows starting at position 1 of the motif (96). Minor groove width, roll, twist and propeller twist constitute the shape features (230). Mono- and dinucleotide frequencies constitute the *k*-mer features (120). The total number of features was 595.

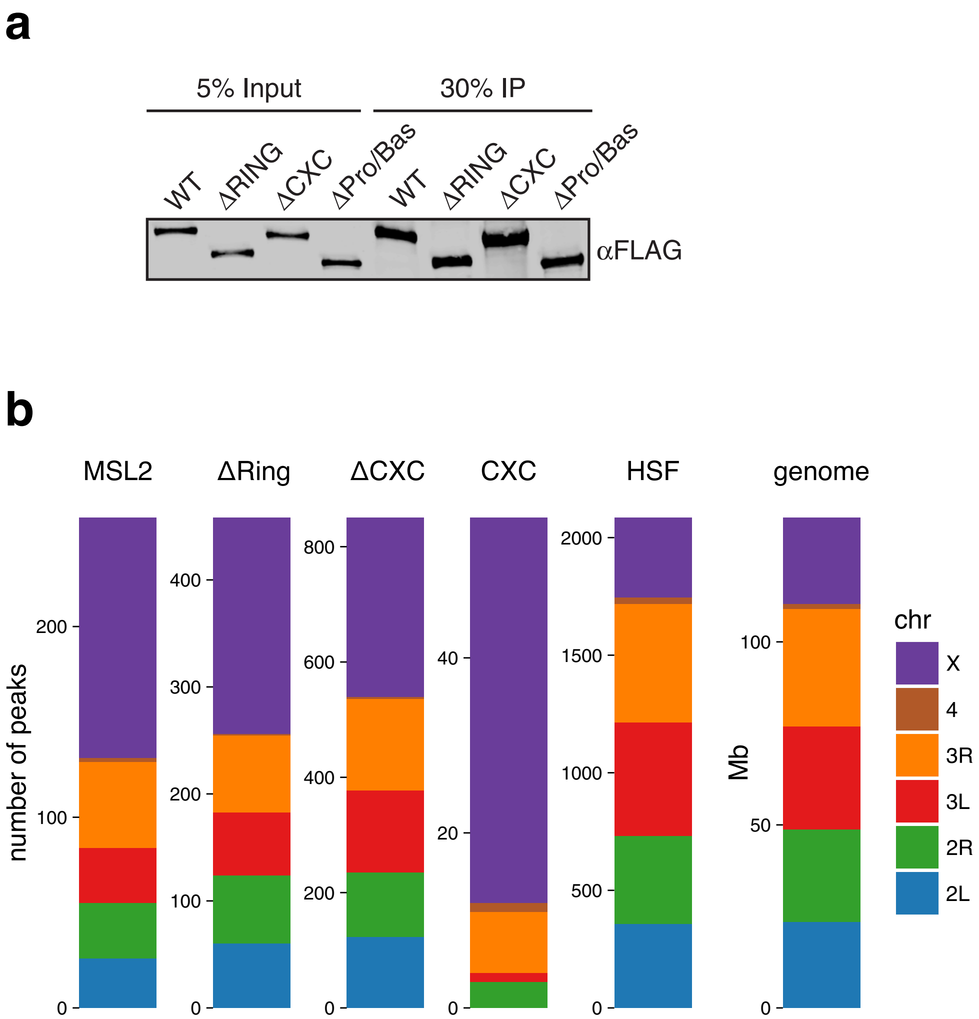
Machine learning. Classification models for feature evaluation were built using simple logistic classifier[30]. ROC curves of the classifiers were based on tenfold cross-validation.

The importance of all features were ranked by measuring their correlation (Pearson's) with the class label on the whole set of PionX PWM hits with a $q \leq 0.2$. Features selected as relevant for and present at CXC-dependent binding of the *hiw*, CG8097 and CG1492 genes were: 'CA in window 1', 'TT at window 2' and 'T at window 2'. Mutations were proposed based on the results of feature selection and the presence of respective *k*-mers in the sites selected for mutation. The proposed mutations were evaluated by simple logistic classifier trained using the PWM score and *k*-mers on the full data set. Modified sites were designed to result in the switch of the predicted class from CXC-bound to not CXC-bound.

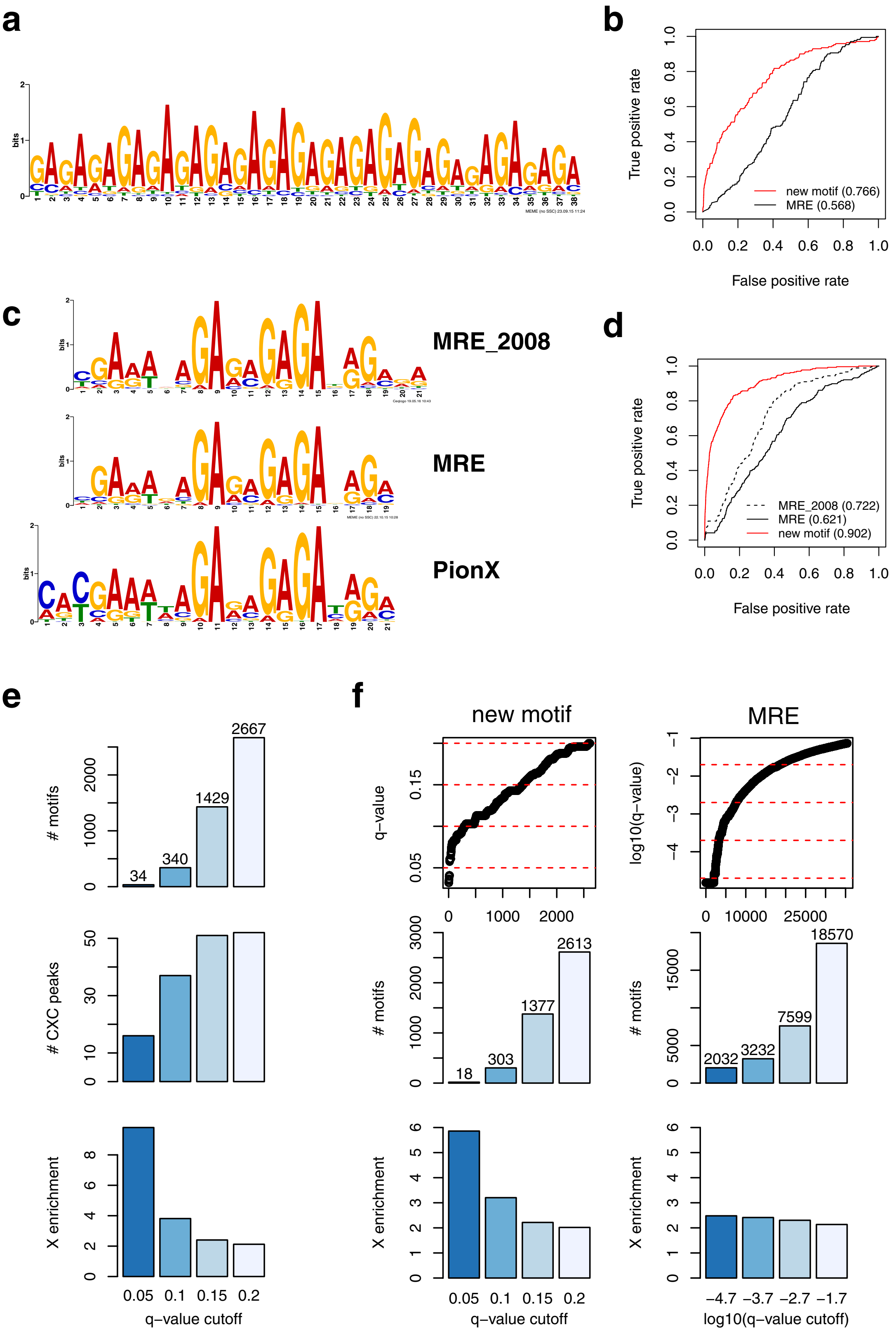
26. Schauer, T. *et al.* CAST-ChIP maps cell-type-specific chromatin states in the Drosophila central nervous system. *Cell Reports* **5,** 271–282 (2013).
27. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. **10,** R25 (2009).
28. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).
29. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. **37,** W202–W208 (2009).
30. Summer, M., Frank, E. & Hall, M. *Speeding Up Logistic Model Tree Induction* 675–683 (Springer, 2005).



Extended Data Figure 1 | Analysis of *in vitro* versus *in vivo* MSL2-binding sites. a, Venn diagram showing the genome-wide overlap of robust MSL2 *in vivo* and *in vitro* DNA binding peaks. **b**, MSL2 enrichment (immunoprecipitate (IP) over input) of all 57 overlapping peaks from *in vitro* DIP–seq and *in vivo* ChIP–seq experiments. The average of two biological replicates is shown, and the Pearson correlation coefficient is indicated. **c**, X-chromosomal enrichment over autosomes of MSL2 DIP–seq peaks using genomic DNA from S2 cells, Kc cells or synthetic gDNA (whole-genome amplified). S2 peaks correspond to an overlapping set of two biological replicate experiments; Kc cell and whole-genome amplification experiments were performed once. **d**, Chromosomal distribution of MSL2 DIP–seq peaks of experiments shown in **c**. The relative size of chromosomes and the genome serve as a reference for uniform distribution. **e**, Representative profiles of *in vivo* MSL2 ChIP–seq and the corresponding chromatin input on chromosome 3R. Red bars indicate the positions of CXC-dependent *in vitro* binding sites. Gene models are depicted in grey at the bottom.

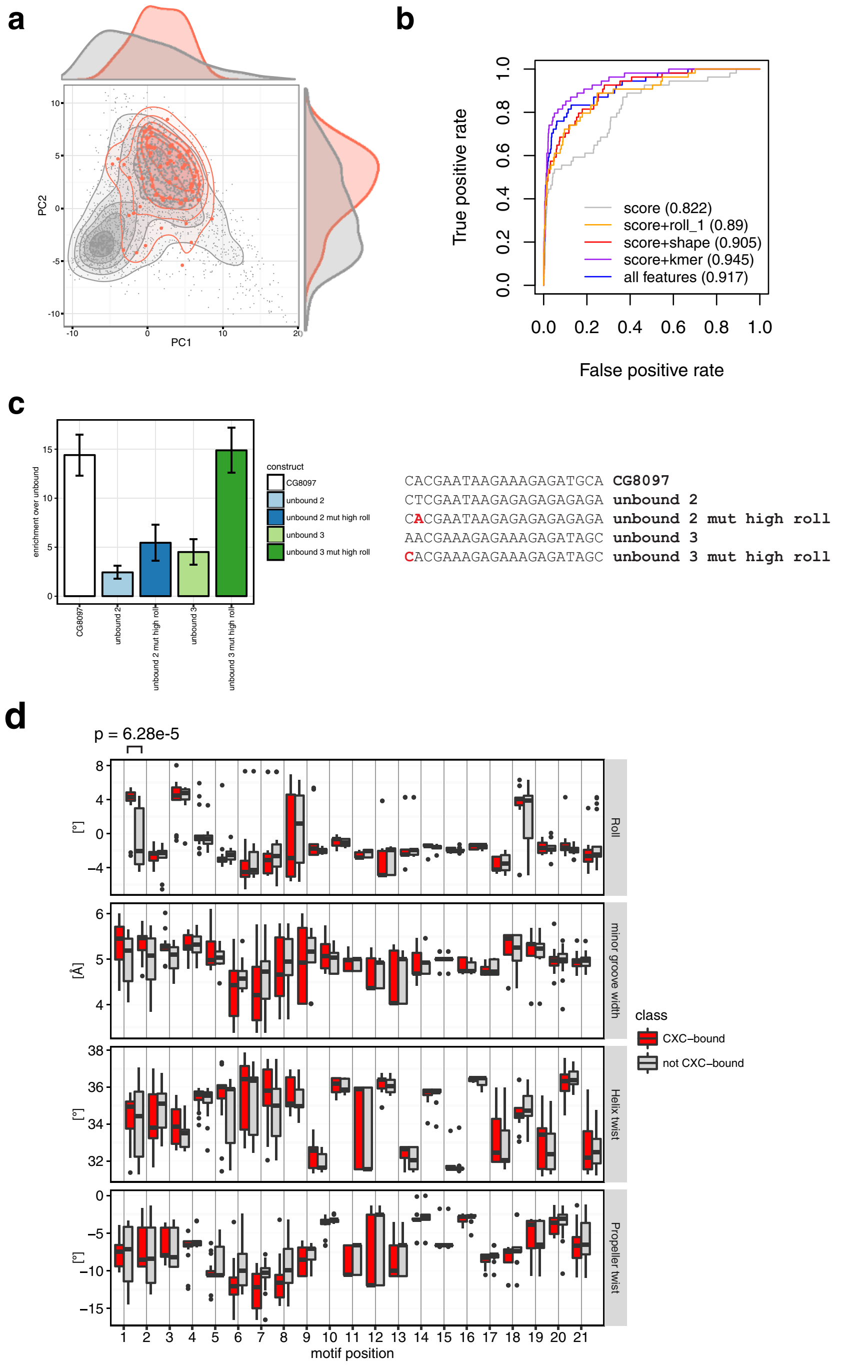


Extended Data Figure 2 | Analysis of MSL2 mutants in DIP–seq assays. **a**, Western blots showing input and anti-Flag immunoprecipitated MSL2 proteins from a representative DIP experiment (for gel source data see Supplementary Fig. 1). **b**, Chromosomal distribution of DIP–seq peaks obtained with MSL2, MSL2 mutants and HSF[8] (see Fig. 2d). The chromosomal size distribution (genome) is provided for reference.

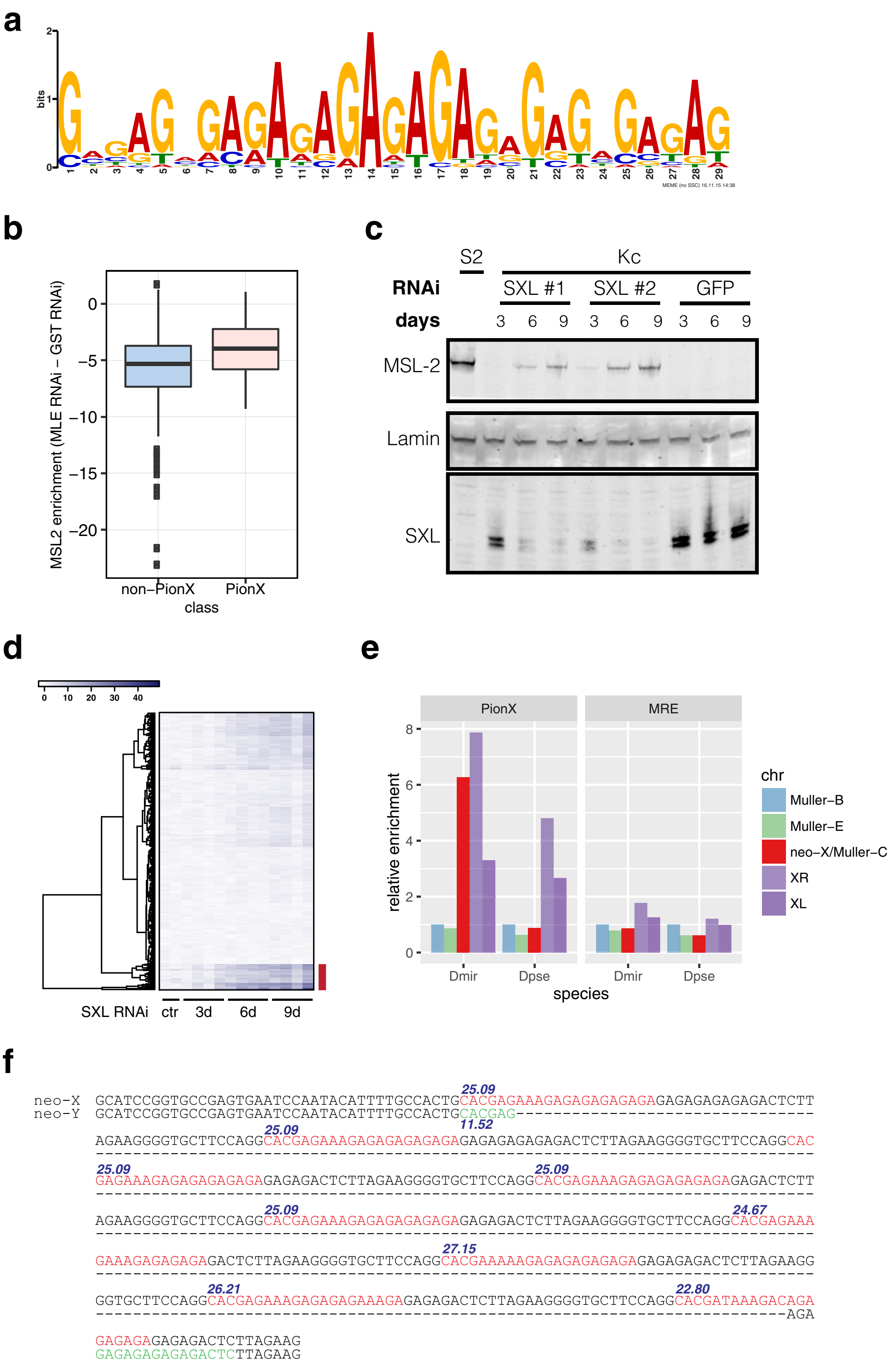


Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Comparison between the CXC-dependent motif and the MRE. a, Consensus motif in CXC-independent binding regions (present in 164 out of 201 regions; $E = 2.0 \times 10^{-1,191}$). **b**, ROC curves representing the PWM performances of MRE and the new motif in predicting whether an instance of the new motif ($n = 2{,}651$) will overlap with HAS (170). AUCs are provided in brackets. As our method slightly penalizes the MRE performance estimation (see Methods), this figure represents a symmetrical analysis of the new motif hits of Fig. 3b. **c**, Top, motif logos of MRE as reported previously[2]. Middle, MRE as reported in this study (see also Fig. 4c, top). Bottom, PionX motif as reported in this study (Fig. 3a). **d**, ROC curves representing the PWM performance comparison analogous to the result presented in Fig. 3b, including the MRE as reported previously[2] (labelled MRE 2008), the MRE as reported in this study (labelled MRE) and the PionX motif (labelled new motif) in classifying MRE instances (35,659) within HAS (266) or not. AUCs are provided in brackets. **e**, Genome-wide search with the PWM of the new motif using FIMO. q-value cut-off relation with the total number of genomic hits (top), the number of CXC-dependent *in vitro* binding sites (middle) and the X-chromosomal enrichment of motif hits (bottom). **f**, To ensure that the enrichment is not solely due to performing *de novo* motif discovery on mainly X-chromosomal sequences, we performed the analysis as presented in **e** excluding the training regions. We conducted the same analysis for the new motif (left) as well as the MRE (right). Top panels depict the q-value distribution and the cut-offs used. The total numbers of genomic hits are displayed in the centre panels, with the corresponding X-chromosomal enrichments displayed at the bottom.



Extended Data Figure 4 | Importance of *k*-mer frequencies and DNA shape for CXC-dependent MSL2 *in vitro* binding. **a**, PCA on the set of all extended features in 2,667 genomic hit regions of the new motif ($q \leq 0.2$). Scatter plots and corresponding scaled density plots of PC1 versus PC2. 2,613 sites not bound *in vitro* in a CXC-dependent manner and 54 bound in a CXC-dependent manner are coloured grey and red, respectively. **b**, ROC curves depicting the performance of simple logistic classifiers for CXC-dependent binding on 2,667 low-stringency motif hits ($q \leq 0.2$; 54 sites CXC-bound, 2,613 sites non-CXC-bound) based on different combinations of motif PWM scores and extended features. AUCs are provided in brackets. **c**, DIP experiments testing the binding affinities of DNA oligonucleotides representing two unbound sites (unbound 2 and 3) and their respective mutated sites (unbound 2 mut and unbound 3 mut) to increase the roll at position +1. Results from qPCR amplification were normalized for their input and shown as enrichment over an unbound fragment. Data are mean ± s.e.m for 4 biological replicates. **d**, DNA shape features at each base position comparing CXC-bound motifs ($n = 16$) to non-CXC-bound ones ($n = 18$) in the highest-scoring hit regions of the new motif ($q < 0.05$). Differences of shape features at all positions were evaluated by applying Wilcoxon exact rank tests with two-sided alternatives. Only roll at position +1 had $P < 0.001$. As roll and helix twist specify inter-base structural features, the corresponding bar graph representations have been centred between the respective nucleotide positions.



Extended Data Figure 5 | *In vivo* analysis of PionX sites. a, Consensus motif found in the 25 regions where MSL2 binding is most sensitive to depletion of MLE. **b**, MSL2 signal changes on 37 HAS matching CXC-dependent *in vitro* binding sites or 272 non-matching ones during MLE knockdown in S2 cells. Displayed are the mean differences of three biological replicates. **c**, Western blot analysis of whole-cell extracts from S2 and Kc cells treated with either RNAi against *Sxl* (two different double-stranded RNAs) or control RNAi directed against irrelevant *Gfp* sequences at different time points (for gel source data see Supplementary Fig. 1). **d**, Clustered heat map of MSL2 peaks from ChIP–seq experiments in female Kc cells treated with RNAi against *Sxl* for 3, 6 and 9 days. Red bar indicates 30 sites characterized by strong MSL2 recruitment. **e**, Enrichment of PionX motif hits (score > 22) and MRE motif hits (score > 27) on *D. miranda* and *D. pseudoobscura* chromosomes relative to Müller-B, normalized for chromosome length. The analysis included 225 and 400 PionX hits in *D. miranda* and *D. pseudoobscura*, respectively. A total of 784 and 755 MRE hits were considered in *D. miranda* and *D. pseudoobscura*, respectively. **f**, Sequence from the neo-X chromosome chromatin entry sites compared to its counterpart on the neo-Y chromosome as in supplementary fig. 2 of ref. 19. Motifs are highlighted in green (neo-Y-chromosomal) and in red (neo-X-chromosomal) with their corresponding PionX motif score in blue.

Extended Data Table 1 | CXC-dependent sites (PionX)

Release Dm6 coordinates		
Chromosome	Start	End
X	253666	253909
X	762263	762467
X	798142	798353
X	2024836	2025054
X	2599191	2599424
X	4628553	4628776
X	5759739	5759964
X	6083151	6083363
X	6370871	6371095
X	7288123	7288377
X	8248450	8248658
X	8689224	8689451
X	10347710	10347917
X	11580085	11580334
X	11703168	11703386
X	12010823	12011036
X	12649470	12649672
X	12715985	12716233
X	13200918	13201126
X	13263388	13263594
X	13420994	13421248
X	14103895	14104106
X	14117342	14117604
X	14585023	14585225
X	14828410	14828624
X	15052054	15052262
X	15086257	15086486
X	15584177	15584395
X	15730624	15730842
X	15875723	15875949
X	15996469	15996701
X	17655454	17655699
X	17821583	17821832
X	17926738	17926952
X	18495583	18495801
X	18849100	18849328
X	19489957	19490201
X	19730579	19730791
X	19968088	19968315
X	20024562	20024819
X	20794352	20794555
X	21317751	21317980
X	22535550	22535752
X	23100209	23100436
2R	3129113	3129339
2R	4970222	4970435
2R	13913696	13913908
3R	2556412	2556628
3R	4052724	4052974
3R	7075728	7075953
3R	10293064	10293265
3R	10516992	10517205
3R	21794080	21794326
3R	30080869	30081093
3L	3762195	3762399
4	292171	292388

Extended Data Table 2 | List of oligonucleotides used in the DIP experiments

Name	Sequence
Hiw wt	ATACGGCGACCACCGAGATAAGAACACGGTTTAGAAAGAGATAGTATTACACTCGTATGCCGTCTTCTGCTTG
Hiw mut	ATACGGCGACCACCGAGATAAGAATACGAACGAGAAAGAGATAGTATTACAC TCGTATGCCGTCTTCTGCTTG
Hiw Roll mut	ATACGGCGACCACCGAGATAAGAAAACGGTTTAGAAAGGATAGTATTACACTCGTATGCCGTCTTCTGCTTG
CG8097 wt	ATACGGCGACCACCGAGATAGAAAACACGAATAAGAAAGAGATGCAAAACATG TCGTATGCCGTCTTCTGCTTG
CG8097 mut	ATACGGCGACCACCGAGATAGAAAATACGAACGAGAAAGAGATGCAAAACATG TCGTATGCCGTCTTCTGCTTG
CG8097 Roll mut	ATACGGCGACCACCGAGATAGAAAAAACGAATAAGAAAGAGATGCAAAACATG TCGTATGCCGTCTTCTGCTTG
CG1492 wt	ATACGGCGACCACCGAGATATTTCACACGAATTCGAAAGAGATGGAAATATGC TCGTATGCCGTCTTCTGCTTG
CG1492 mut	ATACGGCGACCACCGAGATATTTCATACGAACGCGAAAGAGATGGAAATATGC TCGTATGCCGTCTTCTGCTTG
CG1492 roll mut	ATACGGCGACCACCGAGATATTTCAAACGAATTCGAAAGAGATGGAAATATGC TCGTATGCCGTCTTCTGCTTG
CG1492 Roll mut2	ATACGGCGACCACCGAGATATTTCGCTCGAATTCGAAAGAGATGGAAATATGC TCGTATGCCGTCTTCTGCTTG
Unbound	ATACGGCGACCACCGAGATAATAAAATGAAAAAGAAAAAGAAAAAGAAACACTCGTATGCCGTCTTCTGCTTG
Unbound 2	ATACGGCGACCACCGAGATATTGCACTCGAATAAGAGAGAGAGAGAGCCACCT TCGTATGCCGTCTTCTGCTTG
Unbound 2mut	ATACGGCGACCACCGAGATATTGCACACGAATAAGAGAGAGAGAGAGCCACCT TCGTATGCCGTCTTCTGCTTG
Unbound 3	ATACGGCGACCACCGAGATACAGAAAACGAAAGAGAAAGAGATAGCGTTAG TCGTATGCCGTCTTCTGCTTG
Unbound 3mut	ATACGGCGACCACCGAGATACAGAACACGAAAGAGAAAGAGATAGCGTTAG TCGTATGCCGTCTTCTGCTTG

Adapters are highlighted in yellow, mutations in blue.