Evidence for Local Regulatory Control of Escape from Imprinted X Chromosome Inactivation

Joshua W. Mugford,[1,2] Joshua Starmer,[1] Rex L. Williams Jr., J. Mauro Calabrese,[4] Piotr Mieczkowski, Della Yee, and Terry Magnuson[3]

Department of Genetics, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

ABSTRACT X chromosome inactivation (XCI) is an epigenetic process that almost completely inactivates one of two X chromosomes in somatic cells of mammalian females. A few genes are known to escape XCI and the mechanism for this escape remains unclear. Here, using mouse trophoblast stem (TS) cells, we address whether particular chromosomal interactions facilitate escape from imprinted XCI. We demonstrate that promoters of genes escaping XCI do not congregate to any particular region of the genome in TS cells. Further, the escape status of a gene was uncorrelated with the types of genomic features and gene activity located in contacted regions. Our results suggest that genes escaping imprinted XCI do so by using the same regulatory sequences as their expressed alleles on the active X chromosome. We suggest a model where regulatory control of escape from imprinted XCI is mediated by genomic elements located in close linear proximity to escaping genes.

THE three-dimensional shape of chromosomes has a direct impact upon gene regulation, as chromatin looping mediates the interaction of enhancers with transcriptional start sites (TSSs) (Lieberman-Aiden *et al.* 2009; Li and Reinberg 2011; Krivega and Dean 2012). Analysis of genome-wide interactions suggests that chromosomes self-organize into topologically associated domains (TADs) that are ~0.8–1 Mb in linear length (Dixon *et al.* 2012). Loci within a TAD are more likely to interact with each other as opposed to forming interactions with loci residing in other TADs.

Chromosomal interactions are thought to play a pivotal role in the epigenetic process of X chromosome inactivation (XCI) (Splinter *et al.* 2011). During XCI, mammalian females transcriptionally inactivate one X chromosome (Xi) per somatic cell to balance X-linked gene dosage with males (Chow and Heard 2009). Whereas genes on the active X chromosome (Xa) are thought to form stable interactions with other loci on the Xa (*cis*) and other chromosomes (*trans*), the interactions formed by Xi-linked loci are relatively less established, suggesting that Xi chromatin folds in a random manner (Splinter *et al.* 2011).

In the mouse, two forms of XCI are observed: imprinted XCI and random XCI. Imprinted XCI occurs within extra-embryonic tissues and is characterized by the exclusive inactivation of the paternally derived X chromosome (Xp) (Takagi and Sasaki 1975). Random XCI occurs within somatic tissues of the developing embryo and adult (Lyon 1961). While imprinted and random XCI may initiate via distinct mechanisms (Kalantry *et al.* 2009), the genetic programs required for the maintenance of both forms appear similar (Marahrens *et al.* 1997; Kalantry *et al.* 2006; Jonkers *et al.* 2009; Shin *et al.* 2010).

Interestingly, a few genes are known to escape both imprinted and random XCI and are expressed from both X chromosomes (Berletch *et al.* 2011; Calabrese *et al.* 2012). Profiles of XCI escape vary among different cell types; the number of escape genes, termed escapers, ranges from 3% to 25% of all X-linked genes (Berletch *et al.* 2011). The molecular mechanism underlying escape has proven elusive. It is suggested that escapers physically associate with each other to facilitate their expression from the Xi (Splinter *et al.* 2011).

doi: 10.1534/genetics.114.162800
Manuscript received February 9, 2014; accepted for publication March 13, 2014; published Early Online March 19, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.162800/-/DC1.
All raw sequencing data and processed interaction files can be found at GEO accession GSE49111.
[1]These authors contributed equally to this work.
[2]Present address: Biogen Idec, 14 Cambridge Center Dr., Cambridge, MA 02142.
[3]Corresponding author: Department of Genetics, The University of North Carolina at Chapil Hill, 120 Mason Farm Rd., Chapel Hill, NC 25799.
E-mail: terry_magnuson@email.unc.edu
[4]Present address: Department of Pharmacology and Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599.

To date, no study has correlated escape from XCI with any genomic regulatory element. It is possible that a single genomic element could act as a locus control region (LCR) for escape. Alternatively, escape-specific sequences may be regionally scattered along the X chromosome, licensing escape on a region-by-region basis. Finally, a sequence specific to escape may not exist and mechanisms of escape may vary from gene to gene.

To distinguish among these possibilities, we used allele-specific circular chromosome conformation capture 4C-sequencing (4C-Seq) to identify genomic interactions occurring at escapers within female F_1 hybrid trophoblast stem (TS) cells. TS cells undergo imprinted XCI. Therefore, F_1 hybrid TS cells serve as an ideal system for allele-specific analysis of mechanisms underlying escape from XCI, as no mutations are necessary to bias XCI.

Our results suggest that escape from imprinted XCI happens on a gene-by-gene basis. We demonstrate that escapers do not converge upon a single LCR, and we did not identify sequences consistent with regionally dispersed escape regulatory elements. Rather, regardless of escape status, genomic regions in close linear proximity tend to share regions of contact. Furthermore, we show that unlike genes subject to XCI, escapers are located in close linear proximity to putative active enhancer elements that are also found on the active X chromosome. We suggest that genes escaping imprinted XCI utilize regulatory elements in close linear proximity and that mechanisms of escape may vary from gene to gene.

Materials and Methods

TS cell derivation and culture

F_1 hybrid TS cells were derived and cultured as described previously (Himeno *et al.* 2008; Calabrese *et al.* 2012).

Allele-specific 4C-Seq

Allele-specific 4C was based upon previously published work (Splinter *et al.* 2011), with several changes (see Figure 1A and Supporting Information, File S1). 4C anchor primer sequences were designed to capture informative single nucleotide polymorphisms (SNPs) during sequencing. To generate 3C library templates, chromatin from 3×10^7 TS cells was isolated, digested, ligated, and purified using lysis conditions adapted to TS cells. Each 4C template was generated with 12.5 μg of 3C library and was performed as previously described (Splinter *et al.* 2011). One microgram of 4C template was amplified per 4C primer pair using optimized PCR conditions. The 4C PCRs were then size selected between 150 bp and 650 bp, purified, and further amplified in a linear range with outer sequencing adapter primers. Amplified 4C libraries were purified with AmpureXP beads (BioRad) and submitted for paired-end 100-bp sequencing on Illumina HiSeq 2000 sequencers (Illumina). Biological replicates for each anchor were performed and sequenced separately.

Filtering and statistical analysis of allele-specific 4C-Seq reads

See File S1 for in-depth details of all filtering and statistical analysis. Briefly, raw sequencing reads were filtered through custom Perl and R scripts (available upon request). Anchor fragment portions of reads were filtered by known SNPs to identify the anchor point of origin for any given read pair (either Xa or Xi). The unknown portion of reads was then mapped to the B6 and Cast genomes using Bowtie (version 0.12.7) (Langmead *et al.* 2009). The Cast genome was generated by substituting identified SNPs (Yalcin *et al.* 2011) into the mm9 B6 consensus genome. Reads were paired and files were generated for statistical analysis. PCR amplification bias was not detected in any sequencing dataset, allowing the use of the full range of sequencing reads.

Statistical enrichment of sequencing reads (Williams *et al.* 2014) was performed on each biological replicate. Briefly, a sliding window analysis was performed using window size corresponding to three 3C restriction fragment lengths. Analysis was performed per chromosome and raw reads were shuffled randomly across chromosomes 1000 times to generate significance thresholds. Read-containing windows passing empirically determined thresholds were called interactions. Per anchor, interactions from each biological replicate were then compared to generate a final list of genomic coordinates. Interactions were assigned to the B6, Cast, B6/Cast (equal contribution of both alleles), or NoCall (no allelic data), based upon the presence of informative SNPs within reads contributing to interactions.

RNA-Seq, DNAse-Seq, ChIP-Seq, genomic repeat analysis

All allele-specific RNA-Seq, DNAse-Seq, and ChIP-Seq datasets in C/B TS cells were obtained (GEO accession GSE39406) and analyzed for the whole genome, based upon previously described methods (Calabrese *et al.* 2012). A table of genomic repeats and their locations in build mm9 were obtained from the University of California Santa Cruz genome browser (Meyer *et al.* 2013). Genomic feature indices were generated by dividing the total number of identified features over the total number of bases covered by all interactions in a dataset. Paired, two-tailed *t*-tests or paired, two-sample *t*-tests were used to determine *P*-values when comparing between homologous alleles or among anchors on the same chromosome, respectively. Custom Perl scripts (available upon request) were used to identify the overlap of genomic features. Enrichment of any feature was measured by calculating the ratio of random occurrences over number of permutations ($P \leq 0.05$, FDR $\leq$ 0.05, see File S1 for further details).

Binning of 4C interaction data for comparison and correlation analysis

To properly compare the genomic coordinates of interaction profiles among anchors, each anchor interaction profile was transformed into a binned profile. For Pearson correlation analysis of *cis* interactions, allelic data for each anchor was

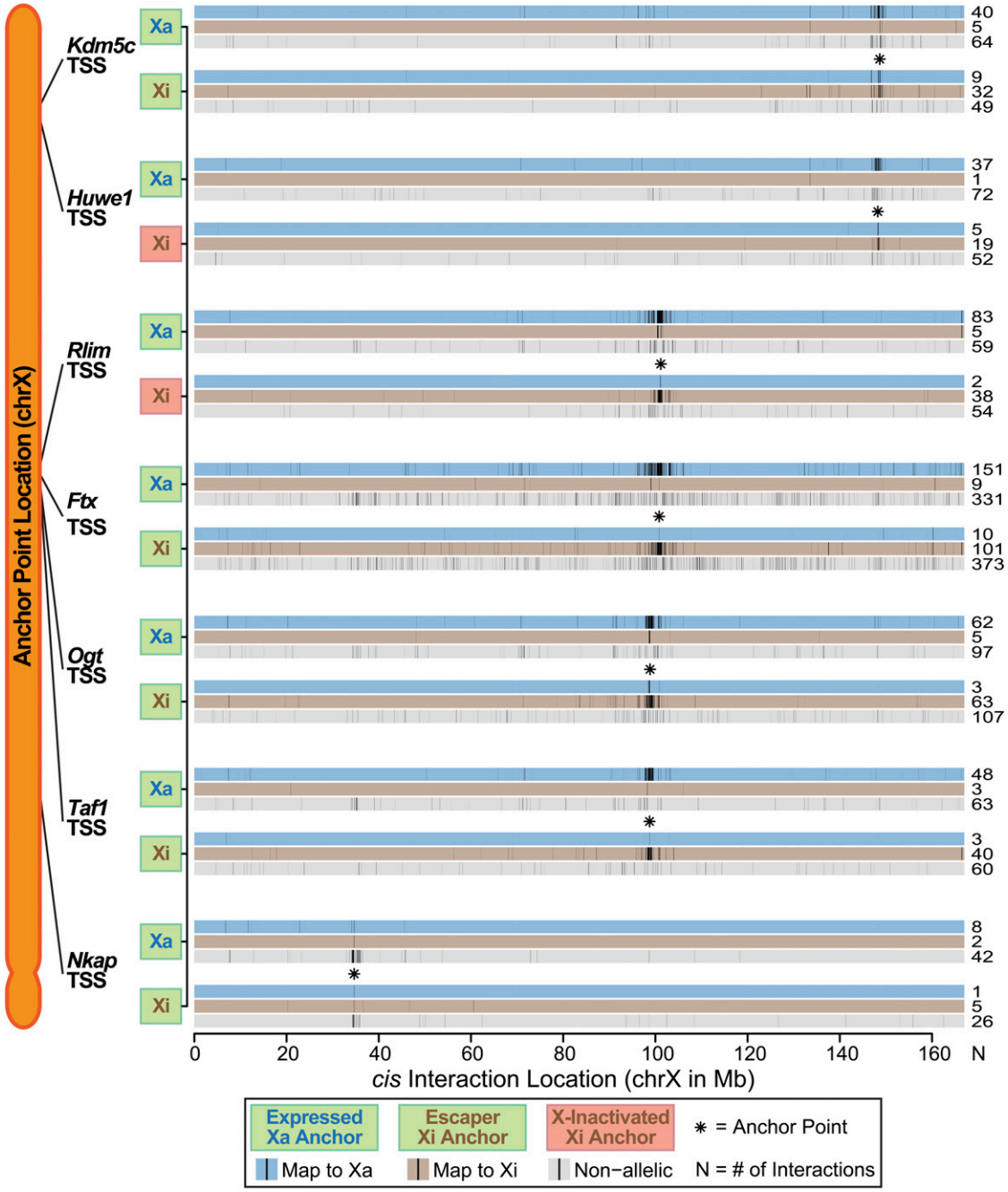


Figure 1 The active and inactive X chromosomes form distinct conformations. *Cis* interactions generated by Xa and Xi anchors are mapped to their chromosomal positions and are depicted by black vertical lines. Horizontal blue, brown, and gray bars indicate the chromosome to which interactions mapped: the Xi (B6 genome), the Xa (Cast genome), and no allelic call, respectively. The number of interactions detected for each allelic call is noted to the right. Asterisks indicate anchor point location. Anchor points are listed along the *y*-axis according to their position along the X chromosome with Xa and Xi anchors colored in blue or brown, respectively. Anchor points from genes subject to XCI or escaping XCI are highlighted in light red and light green, respectively. Genomic position along the X chromosome is listed on the *x*-axis in megabases. Also see Figure S1 and Figure S2.

divided into bins of 500 bp. For common shared genomic regions, bins were also set at 500 bp, and allelic data were analyzed separately from nonallelic data. Per anchor, reads contributing to called interactions were sorted into appropriate bins. Binned data were then binarized to 1 or 0, depending upon the presence or absence of data within a bin.

Because interaction frequency is expected to decay over linear distance, an adjustment value for each 4C pair was determined using

$$-\log(\text{abs}(a-b)/\text{chrX}),$$

where a and b are the coordinates (in base pairs) of the TSS of the genes being compared and chrX is the total length of

the X chromosome (based on mm9). Positive and negative Pearson *R*-values for each 4C anchor profile comparison were then divided or multiplied, respectively, by the corresponding adjustment value.

Identification of putative active enhancer elements near genes

Allelic peaks for H3K27Ac ChiP-Seq, H3K4me1 ChiP-Seq, and DNAse in C/B TS cells were compared to generate a conservative list of putative active enhancer elements. A putative enhancer element was called only at regions of overlap when all three features contained allelic data. If overlap occurred, but any of the features lacked an allelic assignment, the putative active enhancer was identified, but given a "no allele" assignment. When locating putative enhancer elements in close proximity to genes, we only used genes where allelic contribution to overall levels could be determined, based on RNA-Seq in C/B TS cells (Calabrese *et al.* 2012). Genes were then divided into two categories: genes subject to XCI and escaper genes. Fifty kilobases were then added to the annotated TSS and transcriptional termination of each gene. The number of putative enhancer elements located within each gene body (±50 kb) was normalized per gene by dividing the total number of enhancers found by the total kilobases searched. Mouse embryonic stem (ES) cell TAD boundaries (Dixon *et al.* 2012) were obtained (GEO accession GSE35156) and compared to the coordinates of identified enhancer elements.

Fluorescence in situ hybridization confirmation of interactions

RNA/DNA fluorescence *in situ* hybridization (FISH) was performed identically to Calabrese *et al.* (2012). See File S1 for additional information.

Results

Generation of allele-specific interactions

We performed allele-specific 4C-Seq (Splinter *et al.* 2011) at the TSS of escapers to test if the physical association of escape promoters with regulatory elements governs escape from XCI. A female F_1 hybrid TS cell line between the strains *Mus musculus castaneous* (Cast/EiJ or Cast) and *Mus musculus domesticus* (C57BL6/N or B6) were used for our analysis (Calabrese *et al.* 2012). TS cells undergo imprinted XCI (Mak *et al.* 2002), therefore our Cast/B6 (C/B) TS cells harbor a paternally inherited B6 Xi and a maternally inherited Cast Xa (Calabrese *et al.* 2012).

We modified a previously published 4C-Seq protocols and statistical analysis for our study (Splinter *et al.* 2011) (Figure S1, A and B and *Materials and Methods*). For all gene promoters (termed anchors), one primer within each 4C primer pair hybridizes upstream of a known SNP (Yalcin *et al.* 2011), allowing for the identification of the anchor allele of origin for every sequencing read (Figure S1A, red lines). Appropriate 4C-Seq anchor points were identified using allele-specific RNA-Seq datasets in C/B TS cells (Calabrese *et al.* 2012). Seven anchor points were chosen: five escapers (*Nkap*, *Taf1*, *Ogt*, *Ftx*, and *Kdm5c*) and two X-inactivated genes (*Rlim* and *Huwe1*).

Statistically significant genomic interactions were categorized by their location and the presence of informative SNPs (see *Materials and Methods*). *Cis* and *trans* interactions that could confidently be assigned to an allele were termed allelic, whereas interactions lacking sufficient informative SNPs were termed non-allelic. Depending upon the anchor, allelic interactions account for ~17–37% of all data (Table S1).

FISH of randomly selected interactions in C/B TS cells was used to confirm our 4C-Seq pipeline (Figure S1, B and C). On both the Xi and Xa, measured distances between FISH signals located within interactions were shorter than distances measured between FISH probes located outside of interactions (Figure S1C), suggesting that we reliably detected allele-specific interactions within the genome.

The Xi in TS cells generates stable interactions

A broad analysis of our 4C-Seq data was performed and we tested if the contact profiles generated by Xa anchors differed from those generated by Xi profiles. We found no difference in the overall number ($P = 0.722$) and proportion of *cis* to *trans* ($P = 0.215$) interactions generated by TS cell Xi and Xa anchors (Figure 1, Figure S2, and Table S1). Additionally, anchor origin and transcriptional activity had no effect on the linear distance bridged by *cis* interactions (Figure 1). In agreement with previous chromosome conformation studies (Lieberman-Aiden *et al.* 2009; Dixon *et al.* 2012), Xa and Xi anchors preferred to interact with their X chromosome of origin (Figure S2 and Table S1), with most interactions occurring within a few megabases of the anchor point (Figure 1). Correlation analysis of *cis* interactions demonstrated that Xa and Xi contact profiles were largely different (Figure 1 and Table S2). Additionally, Xi loci were located closer together than Xa loci (Figure S1C). Taken together, our data suggest that the Xi in TS cells behaves in a similar manner to other chromosomes, though it likely adopts a different overall structure as compared to the Xa.

Contact with transcriptionally active genes does not correlate with escape from XCI

In general, transcriptionally active genes tend to interact with each other, while silent genes interact with other silent genes (Lieberman-Aiden *et al.* 2009). Therefore, we sought to test if genes escaping XCI tended to interact with other active genes.

We classified the transcriptional activity within interacting regions using RNA-seq data from C/B TS cells (Calabrese *et al.* 2012). Normalized indices (genes/kilobase) were used to compare anchor points since the number and median width of interactions varied among anchors (Table S1). Genes found within interactions were grouped into three classes corresponding to their expression status: expressed (transcriptionally active), repressed (transcriptionally silenced due to non-XCI mechanisms), and inactivated (transcriptionally inactivated due to XCI).

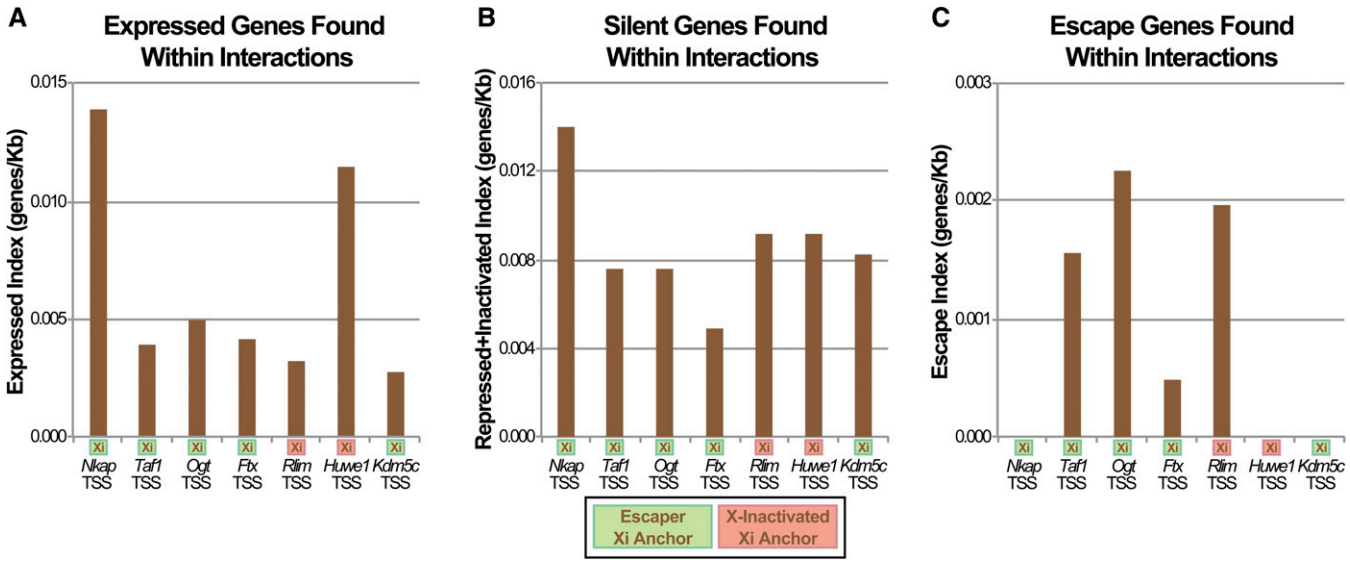


Figure 2 Interactions with active genes do not correlate with transcriptional status. Gene indices (genes/kilobase) per Xi anchor point for specific gene classes are plotted. (A) Plot of expressed gene indices. (B) Plot of silent gene indices (repressed and X-inactivated genes). (C) Plot of escape gene indices. Anchor points are listed on the *x*-axis in their order along the X chromosome. Anchor points from genes subject to XCI or escaping XCI are highlighted in light red and light green, respectively.

The transcriptional profiles found within the interactions generated by X-inactivated and escaper genes did not differ (Table S3 and Figure 2). X-inactivated loci and escaper loci were equivalent in their ability to contact active genes (P = 0.737, Figure 2A), silent genes (P = 0.788, Figure 2B), and genes escaping XCI (P = 0.913, Figure 2C). Thus, in TS cells, the association of a TSS with other actively transcribed genes is not a likely mechanism for escape from imprinted XCI.

Lack of evidence for an escape LCR or a common escape motif

It is possible that escape from XCI is facilitated by an LCR upon which escapers converge. Alternatively, specialized escape-specific enhancers, or other genomic features, could be dispersed across the length of the X chromosome and utilized on a regional basis. Finally, an escape-specific enhancer sequence may not exist, and mechanisms of escape from XCI may vary from gene to gene.

To test the hypothesis that escapers converge upon an LCR, we first preformed a correlation analysis of Xi *cis*-interaction profiles (Figure 3A). Xi anchors clustered according to linear position on the X chromosome, not by expression status. Next, we directly compared the genomic locations of *cis* and *trans* interactions of Xi anchors and searched for regions within the B6 and Cast genomes where Xi anchor profiles overlapped (Figure 3B and Figure S3). Escapers only converged with each other when they were located within 1 Mb of each other (Figure 3B). Consistent with our correlation analysis of interaction profiles (Figure 3A), escapers and X-inactivated genes in close linear proximity shared common regions of contact (Figure 3B).

We next tested if escapers converged on a dispersed sequence or class of sequences. In an attempt to identify a common sequence motif among escape genes, we performed *de novo* sequence analysis using multiple EM for motif elicitation (MEME) (Bailey *et al.* 2009) on ±5 kb of all 30 genes known to escape XCI in C/B TS cells (Calabrese *et al.* 2012). Identified nonrepetitive sequences were then passed to Cis-eLement OVERrepresentation (CLOVER) (Frith *et al.* 2004) to identify if these motifs were enriched in escaper interaction profiles *vs.* X-inactivated interaction profiles. We detected no enrichment of any motifs examined (data not shown). Further, CLOVER analysis of the JASPAR database of transcriptional regulators (Bryne *et al.* 2008) in escaper interaction profiles and X-inactivated interaction profiles did not associate a particular biological pathway with escape from XCI (data not shown).

In comparison to the autosomes, the X chromosome is enriched for LINE elements (Meyer *et al.* 2013) and these features may play a role in the initiation of XCI (Chow *et al.* 2010). We hypothesized that if LINEs facilitate XCI, then genes escaping XCI may form fewer contacts with repeat elements. Upon testing this possibility, we found no difference in the presence of all repetitive elements (P = 0.347) or LINE elements only (P = 0.932, Table S3) found within interactions generated by X-inactivated and escaper anchors.

Taken together, our 4C-Seq data suggest that genes escaping imprinted XCI likely do not converge upon a genomic region, or particular sequence, to facilitate escape. Rather, genes within close linear proximity have similar contact profiles (Figure 3 and Figure S3), suggesting a model whereby escape from imprinted XCI may be governed by regulatory elements found within close linear proximity of escape genes.

Evidence for active enhancers in close proximity to escapers

We next tested the possibility that escaper interaction profiles were generally enriched for genomic features

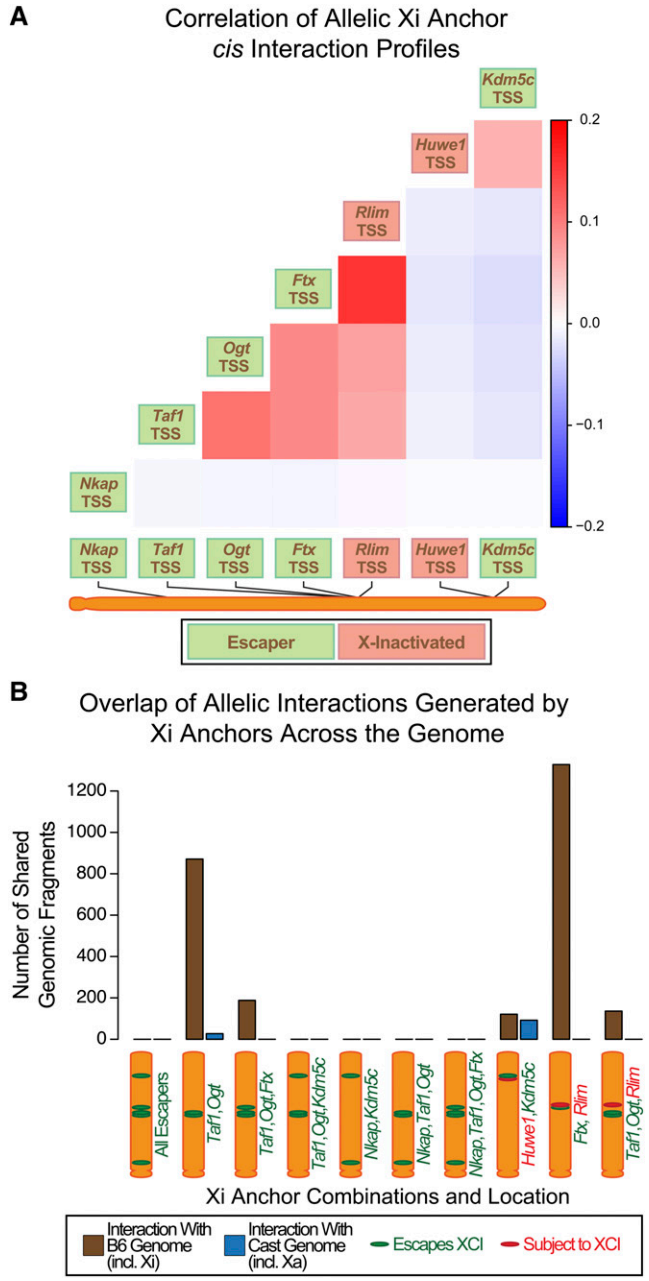


Figure 3 Escapers do not converge upon a common genomic location. (A) Scaled Pearson correlation of *cis* interactions generated by Xi anchor points. Anchor points are listed along the *x*- and *y*-axes according to their position along the X chromosome. Pearson correlations are scaled for linear distances between the anchor points. (B) Bar graph of shared genomic regions of Xi anchor interactions. The combination of anchor points tested is listed on the *x*-axis with escaper anchors in green and inactivated anchors in red. Brown and blue bars indicate shared genomic regions in the B6 or Cast genomes, respectively. Also see Figure S3.

associated with enhancer elements (H3K4me1, H3K27Ac, and DNAse) (Ernst *et al.* 2011; Shen *et al.* 2012) or formation of chromatin loops (CCCT-binding factor) (Dixon *et al.* 2012). Our analysis of datasets in C/B TS cells (Calabrese *et al.* 2012) did not detect any significant difference of these features between escapers and X-inactivated genes (Figure S4). These results suggest that a general strategy of simultaneous interaction with individual epigenetic features does not facilitate escape from XCI.

We next searched gene bodies, plus an additional 10 kb, 15 kb, and 50 kb upstream and downstream (Figure 4A and data not shown), for the presence of putative active enhancer elements to determine if they were correlated with escape from XCI. Putative active enhancers in TS cells were identified as genomic regions enriched for the combination of H3K4me1, H3K27Ac, and DNAse. We then used our RNA-Seq data (Calabrese *et al.* 2012) to generate a full set of genes subject to XCI (n = 276) and escapers (n = 30).

Our analysis demonstrates that putative active enhancers mapped to the Xi were found at higher frequencies surrounding escapers *vs.* X-inactivated genes (Figure 4A, brown bars). This difference was not observed on the Xa where all 296 genes are expressed (Figure 4A, blue bars). For any given escape gene, we noted that the Xa allele was associated with more putative enhancers as compared to the Xi allele (Figure 4). Interestingly, if a putative active enhancer was in close proximity to an escaper, the homologous region on the Xa was also identified as a putative active enhancer (Figure 4B and Table S4). This latter observation was not observed in genomic regions surrounding *Xist*, a gene exclusively expressed on the Xi (Table S4).

It is possible that identified putative enhancers license escape of all genes located within close proximity. If true, we would not expect to find inactivated genes in close proximity to putative enhancers. To that end, we assessed the transcriptional activity of genes found within ±50 kb of putative enhancers identified in close proximity to escapers. We found that, while escaping genes were found at the highest frequencies, inactivated and silenced genes were also found (Table S5). This suggests that proximity to a putative enhancer does not license escape and an additional layer of transcriptional control is likely required.

We tested to see if escapers formed contacts with putative enhancer elements. With the exception of *Nkap*, all escapers formed at least one *cis* contact with a putative enhancer element (Table S5). We note that while several putative enhancer elements surround *Nkap*, there is insufficient SNP data to make a proper call as to their location on the Xa or Xi (data not shown). Therefore, while it is likely that *Nkap* also contacts a putative enhancer on the Xi, we cannot properly demonstrate it with the available data.

In general, TADs are thought to be conserved across cell types (Dixon *et al.* 2012). Because TADs for TS cells have not been defined by any study, we used mouse ES TAD boundaries (Dixon *et al.* 2012) as a proxy to test if the identified putative active enhancers resided in the same TAD as the escape genes we tested. Our analysis revealed that 85% of such enhancers resided in the same TAD as the escape gene associated with them.

Taken together, escaper genomic interactions are not enriched for individual factors associated with active enhancers. However, putative active enhancers mapped to

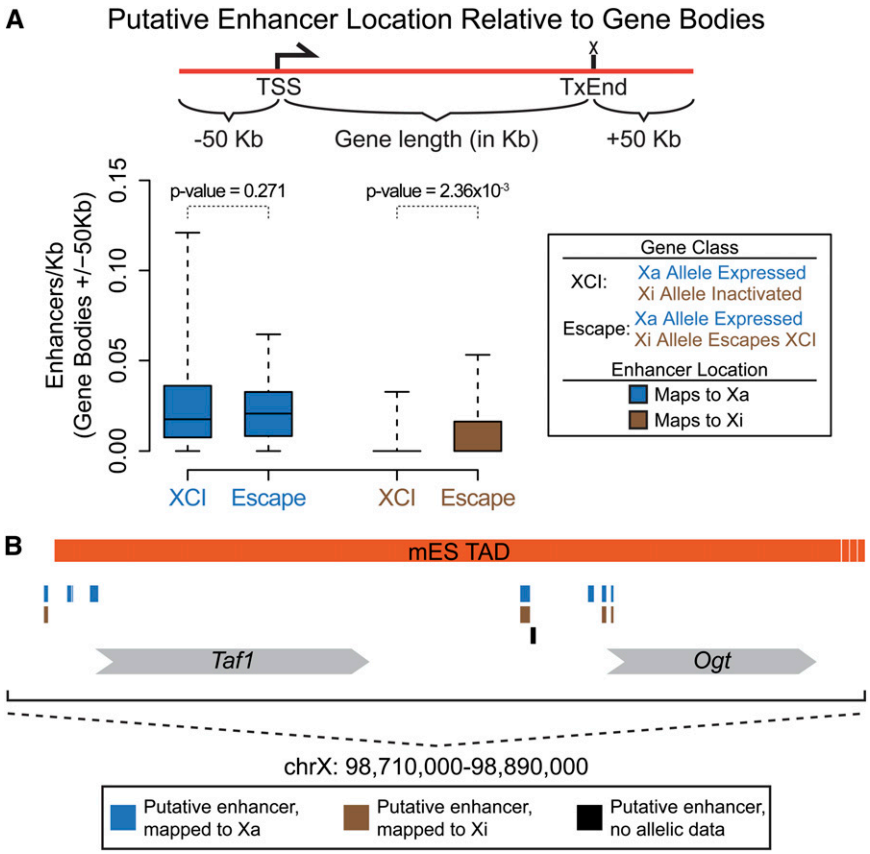


Figure 4 Putative active enhancers are found in close proximity to escaper genes. (A) Boxplots of the number of putative active enhancer elements per kilobase found within gene bodies ±50 kb. Schematic above the plot indicates generic genomic regions searched. XCI = Xa allele is expressed and the Xi allele is subject to XCI. Escape = Xa allele is expressed and the Xi allele escapes XCI. Blue (Xa) and brown (Xi) bars indicate the chromosome to which the putative active enhancer elements mapped. (B) Map of putative active enhancers located near *Taf1* and *Ogt*. Putative enhancers mapped to the Xa, Xi, or no allelic call are indicated by blue, brown, or black boxes, respectively. Gray boxes indicate *Taf1* and *Ogt* gene bodies. Orange denotes the mouse ES cell TAD for this genomic region. Note, one Xi enhancer falls outside of this TAD. Also see Figure S4.

identical genomic coordinates on the Xa and Xi are located in close proximity to escapers and are likely within the same TAD as the escape genes, suggesting that escape from imprinted XCI is facilitated by promoter proximal regulatory elements.

Discussion

We have used allele-specific 4C-Seq to understand mechanisms of escape from XCI. Our data are consistent with a model where escape from imprinted XCI is facilitated by regulatory elements proximal to escaping genes.

The imprinted Xi interacts with the genome

A previous study using neural progenitor cells (NPCs), which undergo random XCI, found that the NPC Xi did not form a predictable structure and was less likely to interact with other chromosomes (Splinter *et al.* 2011). In that same work, the two escaping genes tested were found to interact with other escaping genes more frequently than X-inactivated genes (Splinter *et al.* 2011). Our 4C-Seq data in TS cells, which undergo imprinted XCI, suggest that these findings are not universally true, highlighting a potential difference between imprinted and random XCI.

Another possibility is technical differences between the two studies. Primarily, Splinter *et al.* (2011) used a median of 2.3×10^6 raw reads per anchor with an "*N*" of 1 to draw their conclusions. The small number of reads forced the authors to "binarize" their data, reducing the mapped dataset to a series of 1s and 0s, depending on the presence of mapped reads. This reduction of the data, while protecting against PCR amplification artifacts, forces the use of large window sizes (100 3C fragments) during analysis and decreases resolution.

Our study utilized biological replicates per anchor, and depending on the replicate, had a median of 6.54×10^6 and 5.87×10^6 mapped/processed reads per anchor. The barcodes included in our primer design allow for the elimination of PCR artifacts during analysis. Further, our substantial increase in mapped reads allowed for the use of the full dynamic range of each replicate, a significantly smaller window size (three 3C fragments) during analysis, and increased resolution. Finally, our ability to compare biological replicates for each anchor allowed for the exclusion of interactions generated due to random collisions in the nucleus. Regardless of the source of the differences between the two studies, our data indicate that escapers may or may not interact with other escapers. Furthermore for TS cells, genes escaping imprinted XCI do so by using local regulatory sequences that are the same as their expressed alleles on the active X chromosome (see below).

Significant *cis* and *trans* interactions between anchor points and other genomic coordinates were found for all genes examined, regardless of the location of the anchor (Xa *vs.* Xi) or the transcriptional status of the gene. While the Xa and Xi likely adopt different structures, Xi alleles

preferred to interact with other genomic loci in close linear proximity, an identical behavior noted for loci on the autosomes (Dixon *et al.* 2012). Together, our data suggest that Xi-linked loci in TS cells physically interact with the genome in a manner similar to the Xa and potentially other chromosomal regions.

Escape from XCI is likely mediated within topological domains by active enhancers located proximal to escapers

The observance of TADs within the genome suggests that the majority of regulatory elements required for the expression of any given gene are likely located in close proximity to the gene (Dixon *et al.* 2012). Our 4C-Seq data are consistent with this model. Escaper genes do not always contact other escaper genes; an observation that would be predicted if escape LCRs existed. In the same vein, overlap of interactions was independent of transcriptional activity and only occurred when genes were within a few megabases of each other. This latter result is consistent with a recent study showing that *Huwe1* and *Kdm5c*, an X-inactivated and escaper gene pair separated by ~500 kb, adopt similar positions relative to the *Xist* cloud in nuclear space (Calabrese *et al.* 2012).

Consistent with the TAD model, we find that putative active enhancer elements contact escape genes. In addition, they are found within close proximity to, and likely within the same TADS as escaper genes. In contrast, X-inactivated genes are rarely found in close proximity to putative active enhancers. Interestingly, with the exception of *Xist*, the genomic coordinates of escaper-associated putative active enhancer elements on the Xi are identical to a subset of the putative enhancers on the Xa. The larger number of enhancers found on the Xa may explain why, among escapers, the Xa-linked allele is transcribed at a higher level than the corresponding Xi-linked allele (Calabrese *et al.* 2012). That inactivated genes are found in close proximity to putative enhancer elements suggests that these sequences regulate individual escaper genes and do not license wholesale escape from XCI of any gene located within a close linear distance.

Taken together, our observations support a model where regulatory control of escape from imprinted XCI is possibly governed within TADs. While it is possible that our identified putative enhancer elements are not causing escape, our evidence strongly supports their role in at least maintaining escape from XCI. Recently, it was shown that in TS cells, XCI appears to be maintained independently of a chromosome-scale nuclear compartment dedicated to transcriptional silencing (Calabrese *et al.* 2012). Our model for escape is consistent with this conclusion, as it places the regulatory elements necessary for escape in close proximity to escaping genes. Our model also may explain cell-type-specific escape profiles. If transcription from the Xi does not require any additional regulatory elements other than those used on the Xa, then any gene is capable of escape, so long as the appropriate mechanisms are in place to license usage of the necessary regulatory elements.

Acknowledgments

The authors thank members of the Magnuson lab, particularly Andrew Fedoriw and Jesse Raab, for their many helpful comments and suggestions and for their critical reading of this manuscript. We also thank E. deWitt for providing us with code used as a basis for our statistical analysis. This work was funded by National Institutes of Health grants (R01GM101974 to T.M. and F32-CA144389 to J.W.M.).

Literature Cited

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant *et al.*, 2009 MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37: W202–W208.

Berletch, J. B., F. Yang, J. Xu, L. Carrel, and C. M. Disteche, 2011 Genes that escape from X inactivation. Hum. Genet. 130: 237–245.

Bryne, J. C., E. Valen, M. H. Tang, T. Marstrand, O. Winther *et al.*, 2008 JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 36: D102–D106.

Calabrese, J. M., W. Sun, L. Song, J. W. Mugford, L. Williams *et al.*, 2012 Site-specific silencing of regulatory elements as a mechanism of X inactivation. Cell 151: 951–963.

Chow, J., and E. Heard, 2009 X inactivation and the complexities of silencing a sex chromosome. Curr. Opin. Cell Biol. 21: 359–366.

Chow, J. C., C. Ciaudo, M. J. Fazzari, N. Mise, N. Servant *et al.*, 2010 LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell 141: 956–969.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li *et al.*, 2012 Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376–380.

Ernst, J., P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward *et al.*, 2011 Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43–49.

Frith, M. C., Y. Fu, L. Yu, J. F. Chen, U. Hansen *et al.*, 2004 Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res. 32: 1372–1381.

Himeno, E., S. Tanaka, and T. Kunath, 2008 Isolation and manipulation of mouse trophoblast stem cells. Curr. Protoc. Stem Cell Biol. Chapter 1: Unit 1E 4.

Jonkers, I., T. S. Barakat, E. M. Achame, K. Monkhorst, A. Kenter *et al.*, 2009 RNF12 is an X-Encoded dose-dependent activator of X chromosome inactivation. Cell 139: 999–1011.

Kalantry, S., K. C. Mills, D. Yee, A. P. Otte, B. Panning *et al.*, 2006 The Polycomb group protein Eed protects the inactive X-chromosome from differentiation-induced reactivation. Nat. Cell Biol. 8: 195–202.

Kalantry, S., S. Purushothaman, R. B. Bowen, J. Starmer, and T. Magnuson, 2009 Evidence of Xist RNA-independent initiation of mouse imprinted X-chromosome inactivation. Nature 460: 647–651.

Krivega, I., and A. Dean, 2012 Enhancer and promoter interactions-long distance calls. Curr. Opin. Genet. Dev. 22: 79–85.

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

Li, G., and D. Reinberg, 2011 Chromatin higher-order structures and gene regulation. Curr. Opin. Genet. Dev. 21: 175–186.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy *et al.*, 2009 Comprehensive mapping of long-range

interactions reveals folding principles of the human genome. Science 326: 289–293.

Lyon, M. F., 1961 Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature 190: 372–373.

Mak, W., J. Baxter, J. Silva, A. E. Newall, A. P. Otte *et al.*, 2002 Mitotically stable association of polycomb group proteins eed and enx1 with the inactive x chromosome in trophoblast stem cells. Curr. Biol. 12: 1016–1020.

Marahrens, Y., B. Panning, J. Dausman, W. Strauss, and R. Jaenisch, 1997 Xist-deficient mice are defective in dosage compensation but not spermatogenesis. Genes Dev. 11: 156–166.

Meyer, L. R., A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn *et al.*, 2013 The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. 41: D64–D69.

Shen, Y., F. Yue, D. F. McCleary, Z. Ye, L. Edsall *et al.*, 2012 A map of the cis-regulatory sequences in the mouse genome. Nature 488: 116–120.

Shin, J., M. Bossenz, Y. Chung, H. Ma, M. Byron *et al.*, 2010 Maternal Rnf12/RLIM is required for imprinted X-chromosome inactivation in mice. Nature 467: 977–981.

Splinter, E., E. de Wit, E. P. Nora, P. Klous, H. J. van de Werken *et al.*, 2011 The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev. 25: 1371–1383.

Takagi, N., and M. Sasaki, 1975 Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. Nature 256: 640–642.

Williams, Jr., R. L., J. Starmer, J. W. Mugford, J. M. Calabrese, P. Mieczkowski *et al.*, 2014 fourSig: A Method for Determining Chromosomal Interactions in 4C-Seq Data. Nucleic Acids Res. (in press).

Yalcin, B., K. Wong, A. Agam, M. Goodson, T. M. Keane *et al.*, 2011 Sequence-based characterization of structural variation in the mouse genome. Nature 477: 326–329.

Communicating editor: J. Schimenti

GENETICS

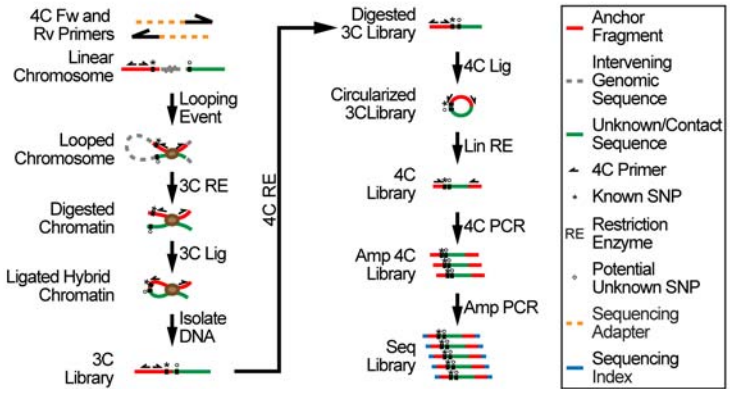
Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.162800/-/DC1

Evidence for Local Regulatory Control of Escape from Imprinted X Chromosome Inactivation

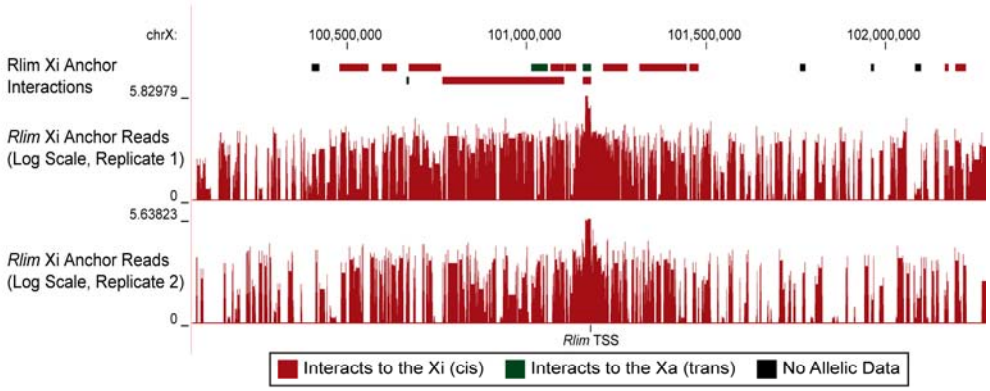
Joshua W. Mugford, Joshua Starmer, Rex L. Williams Jr., J. Mauro Calabrese, Piotr Mieczkowski, Della Yee, and Terry Magnuson

DOI: 10.1534/genetics.114.162800

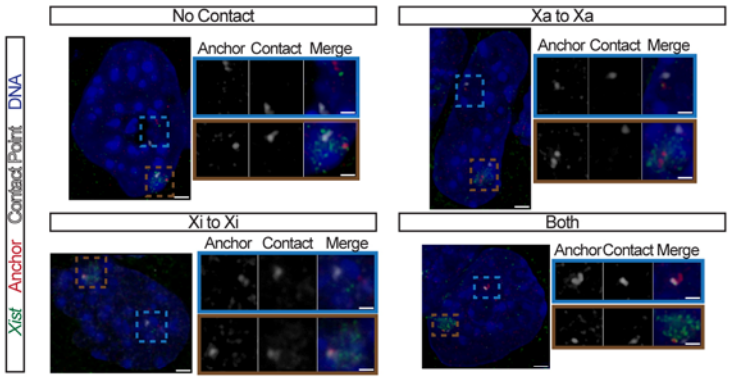
A


4C Fw and Rv Primers
Linear Chromosome
Looping Event
Looped Chromosome
3C RE
Digested Chromatin
3C Lig
Ligated Hybrid Chromatin
Isolate DNA
3C Library
4C RE
Digested 3C Library
4C Lig
Circularized 3CLibrary
Lin RE
4C Library
4C PCR
Amp 4C Library
Amp PCR
Seq Library
Anchor Fragment
Intervening Genomic Sequence
Unknown/Contact Sequence
4C Primer
Known SNP
RE Restriction Enzyme
Potential Unknown SNP
Sequencing Adapter
Sequencing Index

B


chrX:
100,500,000
101,000,000
101,500,000
102,000,000
Rlim Xi Anchor Interactions
5.82979
Rlim Xi Anchor Reads (Log Scale, Replicate 1)
0
5.63823
Rlim Xi Anchor Reads (Log Scale, Replicate 2)
0
Rlim TSS
Interacts to the Xi (cis)
Interacts to the Xa (trans)
No Allelic Data

C


No Contact
Xa to Xa
Xi to Xi
Both
Anchor Contact Merge
Xist Anchor Contact Point DNA

D

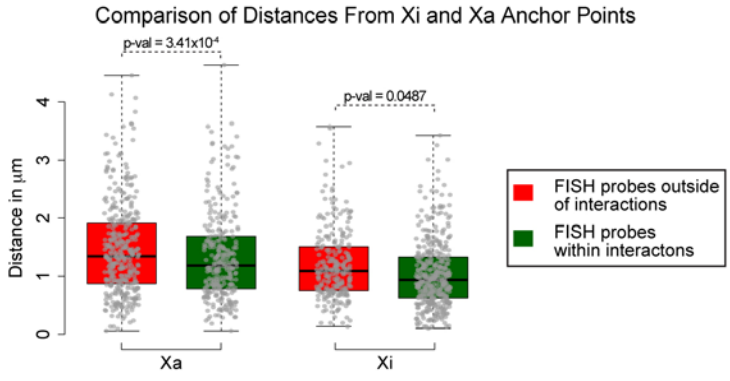

Comparison of Distances From Xi and Xa Anchor Points
p-val = 3.41x10-4
p-val = 0.0487
Distance in µm
0
1
2
3
4
Xa
Xi
FISH probes outside of interactions
FISH probes within interactons

Figure S1 Design and confirmation of allele-specific 4C-Seq. Related to Figure 1. (A) Schematic of allele-specific 4C-Seq workflow. (B) Called interactions surrounding the *Rlim* Xi Anchor with reads from each replicate displayed in log base 10 scale. (C) Representative images of DNA/RNA FISH in TS cells for indicated interactions. No Contact, Xa to Xa, Xi to Xi, and Both indicates regions were not found within interactions, a *cis* interaction identified only on the Xa, a *cis* interaction identified only on the Xi, and a *cis* interaction identified on both the Xi and Xa, respectively. Fluorescence signals for *Xist* RNA (green), anchor point DNA (red), contact point DNA (white), and genomic DNA (blue) are shown in the full sized and merged insets. Insets are zoomed gray scale images of the contact and anchor point on Xi and Xa for clarity. Scale bars, large images = 2μm, insets = 1μm. (D) Box plots of combined distances measured on the Xi or Xa between DNA FISH probes located within interactions (green), or outside of interactions (red).

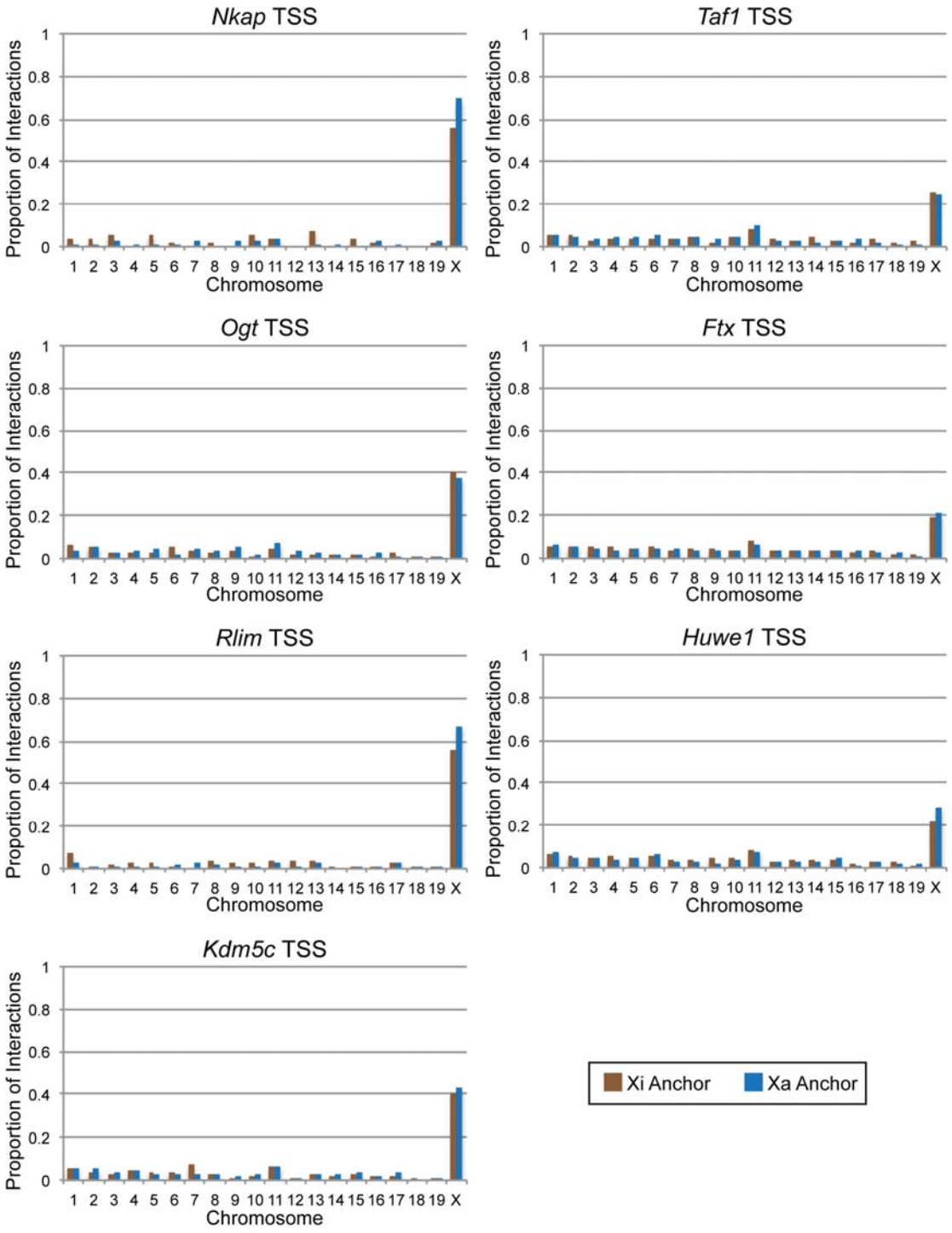


Figure S2 Genomic distribution of interactions. Related to Figure 1. Bar graphs for each anchor indicating the proportion of all interactions (allelic and non-allelic) on each chromosome across the genome. Blue and brown bars represent proportions of interactions generated by the Xa and Xi anchors, respectively.

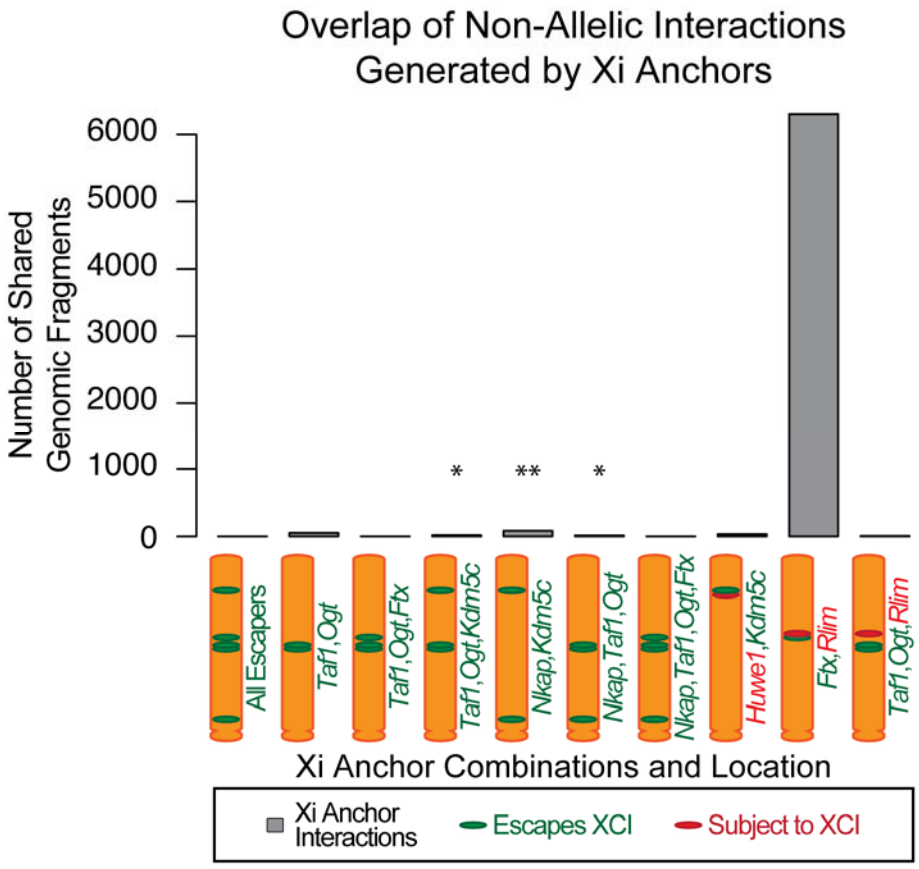


Figure S3 Xi anchor non-allelic interaction convergence. Related to Figure 3. Bar graph of shared genomic regions of Xi anchor interactions where no allelic assignment could be made. The combination of anchor points tested is listed on the x-axis with escaper anchors in green and inactivated anchors in red. * = region on chromosome 11 known to be duplicated in C/B TS cells. ** = duplicated region in chr11 + additional intractable genomic region on chrX lacking SNPs.

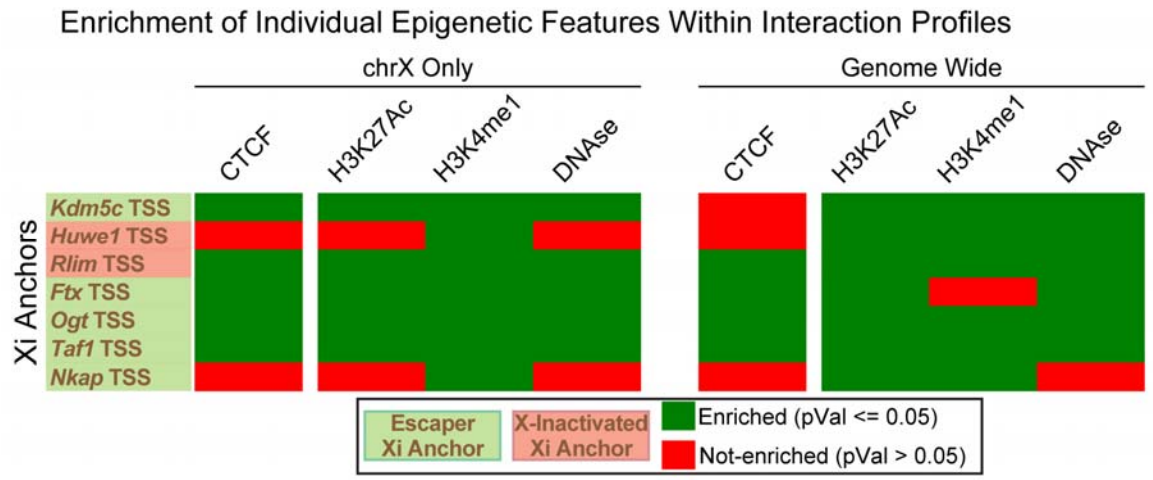


Figure S4 Epigenetic feature enrichment within Xi anchor interactions. Related to Figure 4. Plots of statistical enrichment for epigenetic features within Xi anchor interactions. Green and red boxes indicate enrichment or lack of enrichment, respectively, among all interactions for a given anchor. Anchor points from genes subject to XCI and escaping XCI are highlighted in light red or light green, respectively.

Table S1 Summary of 4C interaction data. Various properties of the interactions generated by Xa and Xi anchors are provided.

Table S1 is available for download as an Excel file at
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.162800/-/DC1.

Table S2 Lack of correlation between Xa and Xi anchor interaction profiles. *Cis* allelic interaction profiles of Xa and Xi alleles of the same gene were compared by Pearson correlation to assess their similarity.

Anchor	R-value
Nkap TSS	0.694
Taf1 TSS	0.095
Ogt TSS	0.345
Rlim TSS	0.078
Ftx TSS	0.223
Huwe1 TSS	0.548
Kdm5c TSS	0.535

Tables S3-S4 are available for download as Excel files at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.162800/-/DC1.

Table S3 Summary of genomic features within allelic Xi 4C interactions. Gene and repetitive element content per Xi anchor is provided. Indexes for each class is determined by dividing the number of features by the total number of bases covered by all interactions (expressed in features/Kb).

Table S4 Putative active enhancers found within gene bodies of escape genes. The normalized number of putative active enhancers (enhancer/Kb) is provided for each allele of each escaping gene. Genomic coordinates of enhancers mapped to the Xi and Xa are provided.

Table S5 Putative escape enhancers contact escape genes in *cis* and are located in close proximity to silenced, inactivated, and escaper genes. The number of genes and their expression status within +/- 50Kb of identified putative escape enhancers are shown. In addition, the escapers that form *cis* contacts with putative enhancer elements are listed.

Enhancer Location	# Inactivated Genes	# Escape Genes	# Silent Genes	Allelic *cis* 4C Contact
chrX:91422447-91424223	0	1	0	No
chrX:91456926-91457542	1	1	0	No
chrX:98717539-98718402	0	1	0	*Taf1, Ogt*
chrX:98817383-98819454	0	2	0	*Taf1, Ogt*
chrX:98834591-98835556	0	2	0	*Taf1, Ogt*
chrX:98836535-98837131	0	2	0	*Taf1, Ogt*
chrX:99314968-99315948	1	1	1	*Ogt, Ftx*
chrX:100678670-100679056	0	2	1	*Ogt, Ftx*
chrX:100703338-100703902	0	2	1	*Ogt, Ftx*
chrX:100709107-100712186	0	3	1	*Ogt, Ftx*
chrX:100721054-100722212	0	3	1	*Ogt, Ftx*
chrX:100724150-100725414	0	3	1	*Ogt, Ftx*
chrX:100745873-100748453	0	2	1	*Taf1, Ogt, Ftx*
chrX:100779106-100780492	0	1	2	*Ftx*
chrX:100812383-100813228	0	1	1	*Taf1, Ogt, Ftx*
chrX:100846965-100849232	0	1	2	*Taf1, Ogt*
chrX:146892930-146894384	0	1	0	No
chrX:146912033-146913502	0	1	0	No
chrX:148667959-148668911	0	1	1	*Kdm5c*
chrX:159325350-159326205	1	1	0	No
chrX:160394995-160396099	1	2	0	No
chrX:160513140-160515277	1	1	1	No
chrX:166423320-166425970	0	1	0	*Taf1, Ftx*

File S1

Supplemental Materials and Methods

Allele Specific 4C-Seq

This is a combination of the 3C/HiC and 4C protocols found in Liberman-Aiden et al., 2009 and Simonis et al., 2006, respectively, adapted for use in undifferentiated TS cells.

The mouse genome build throughout the analysis is NCBI37/mm9.

Restriction enzyme choice and primer design

In order to pass as a 4C anchor, the following criteria were met:

1. As we used paired-end 100bp sequencing, we required that at least 20bp of obtained sequence should be within unknown/contact point sequence. Therefore, the total number of base pairs of known anchor fragment between the 5' end of the primer and the restriction enzyme site could not be greater than 80bp.

2. A SNP must be located between the 3' end of at least one primer and its associated restriction enzyme site. SNPs could be associated with either the 3C or 4C restriction enzyme.

3. The distance between the 3C restriction enzyme site and the 4C restriction enzyme site must be at least 140bp. Anything less runs the risk of inefficient circular ligation.

4. A rare linearization restriction enzyme site must be located between the 5' ends of both primers.

The following criteria were not required, but were met whenever possible:

1. We attempted to avoid 4C enzymes (4bp recognition sites) containing internal CpG dinucleotides, as these cut less frequently than expected. However, this was not always possible due to SNP locations. In this case, 2 4C restriction enzymes could be used.

Primers to 4C anchor points were designed such that one primer bound 5 prime of an annotated SNP associated with either the 3C or 4C restriction enzyme recognition site. A partial sequence corresponding to the former Nextera (now defunct, but can still hybridize to current Illumina flowcells) V1 Universal sequencing adapter was added to the 5' end of the 4C primer associated with the SNP. Between this adapter sequence and the 4C primer sequence, four random nucleotides were added. Therefore, any given 4C primer associated with a SNP would have the following structure:

5’ Partial_Nextera_Fw_Seq-NNNN-4C_SNP_Primer 3’

Similarly, the non-SNP associated primer was constructed such that a partial sequence of the TruSeq Indexed adapter (Illumina reverse adapter) was added 5’ to the 4C primer sequence. Again, 4 random nucleotides were added between the TruSeq and 4C primer sequences:

5’ Partial_TruSeq_Rv_Seq-NNNN-4C_nonSNP_Primer 3’

The random 4-mers serve a dual purpose. The first aids the Illumina software in distinguishing different clusters on a flowcell. The second barcodes each paired read with a random 4-mer on the forward read and an additional 4-mer on the reverse end. 65,536 possible barcode combinations are therefore possible. These were used to identify potential PCR amplification biases (see below).

See Figure 1A for location of generic primers relative to restriction enzyme sites and informative SNPs. Primer sequences, enzymes, and SNP locations are given following the description of the protocol.

Adapted 4C Protocol

We found that standard 3C lysis conditions did not work well for TS cells. We therefore optimized lysis and downstream conditions such that our modifications did not impair enzymatic reactions. ~$120x10^6$ TS cells were harvested for each fixation condition. For fixation, cells were gently trypsinized in 0.25% trypsin, broken up to single cell suspension, spun, and resuspended in 11.2ml of 10% FBS+RPMI. 37% Formadehyde (Sigma) was added to 1% and cells were fixed at RT for 10min on a nutator. 2.5M glycine was added to a final concentration of 130mM, cells were incubated for 5min at RT on a nutator, and then for 15min on ice. Cells were spun, washed 2x in fresh ice-cold PBS, and aliquoted into aliquots of $3x10^7$ cells. Cell pellets were flash frozen on liquid nitrogen and stored at -80˚C. Cell aliquots from separate harvests (separate initial thaws) are considered biological replicates.

$3x10^7$ cells were used per 3C library. Cell pellets were thawed on ice and then resuspended into 660µl ice-cold lysis buffer (10mM Tris-HCl pH8.0, 10mM NaCl, 0.2% Igepal CA-630, 1/10 dilution of Sigma protease inhibitors). Cells were dounced in a 2ml dounce (pestle B) 10 times, placed on ice for 1min, and dounced 10 more times. The suspension was moved into a new Eppendorf tube and spun at 5000rpm for 5min. Nuclei pellets were washed 2x with ice-cold 1X NEB buffer appropriate for the 3C enzyme. Pellets were spun at 5000rpm between each wash. After the

final wash, cell pellets were resuspended in 600µl of ice-cold 1x NEB Buffer. Samples were split into 12 fresh Eppendorf tubes. A mixture of 1X NEB buffer + SDS was added to each tube, such that the final volume for each sample was 400µl with a concentration of 0.2% SDS. Samples were placed at 37˚C for 30min (rocking), and then at 65˚C for 10min (agitated every 2-3min). Cells were then immediately placed on ice and TritonX100 was added to a final concentration of 2%. Tubes were placed at RT for 10min. 400U of the appropriate restriction enzyme, BSA, and 1X NEB buffer were added to each tube such that the final volume was 500µl with 1mg/ml BSA. 3C digestion was carried out O/N at 37˚C on a vortex at 950rpm.

SDS was added to 1.47% and samples were incubated at 65˚C for 30min. Samples were then pooled on ice into one 50ml conical tube containing 41.5ml ligation mix buffer (1X NEB Ligation Buffer, 1mg/ml BSA, 2% TritonX100). 5 aliquots of 8.8ml were split into pre-cooled 15ml conical tubes. The remaining sample was placed into a separate pre-cooled 15ml conical tube and sealed. This final sample served as a –Ligase control. To the other 5 aliquots, 15U of T4 DNA ligase (Life Sciences) was added. Ligations took place for 4hrs at 16˚C. 50µl of 10mg/ml ProteinaseK was added to each tube and all were placed at 55˚C O/N to reverse crosslinks and break down protein.

An additional 50µl of 10mg/ml ProteinaseK was added to each tube and all were placed at 55˚C for at least 2hrs. Samples were phenol extracted using standard procedures. The aqueous layers was then extracted with phenol:chlorophorm:isoamyl alcohol (1:1:24), using standard procedures. All large extractions were carried out using phase lock tubes (5 Prime). After the second extraction, aqueous layers were transferred to fresh 50ml conical tubes and brought to 13ml with water. Samples were ethanol precipitated at -80˚C for 1.5hrs. Samples were spun at 10,000xg for 20min at 4˚C. Pellets were brought up in 450µl of TE and extracted 2X with phenol:chrolorphorm:isoamyl alcohol (1:1:24) using standard procedures. After the final extraction, the aqueous layer was ethanol precipitated at -80˚C for 1hr. Samples were spun at high speed, and washed 4x in ice-cold 70% ethanol. Pellets were brought up in 50µl of TE and 3C samples were combined (3C volume = 250µl, -ligase sample = 50µl). 10µl and 60µl of 10mg/ml RNaseA was added to the –ligase and 3C samples, respectively. These were placed at 37˚C for 2.5hrs, and extracted with phenol:chrolorphorm:isoamyl alcohol (1:1:24) using standard procedures. The organic layer was back-extracted with water and combined with the aqueous layer from the first extraction. Samples were ethanol precipitated O/N at -20˚C. Samples were spun at high speed and washed with ice-cold 70% ethanol 2x. 3C pellets were brought up in a total of 300µl of TE and –ligase sample were brought up in 50µl of TE. The concentration of 3C and –ligase samples were determined by a qBit fluorometer (Life Sciences).

To assess the quality of the 3C library, 500ng of 3C and –ligase sample was run on a 0.8% agarose gel. If a high molecular weight band was not observed in the 3C library, it was discarded and 3C was started again. If the 3C library was of good quality, 4C restriction enzyme digestions were carried out. For each anchor point, 12.5µg of 3C library was used and digested in a total volume of 100µl with the appropriate enzyme in its appropriate buffer/temperature for 4hrs. Reactions were phenol:chrolorphorm:isoamyl alcohol (1:1:24) extracted using standard procedures and ethanol precipitated O/N at -20˚C. Samples were spun, washed in 70% ethanol, and brought up in 50µl of water. The concentration of DNA was determined with the qBit fluorometer and adjusted to 100ng/µl. Anything in excess of 10µg total was removed as a 4C -ligase control. 500ng of 3C library, 3C –ligase, and 4C –ligase control were run out on a gel to ensure that the 4C enzyme efficiently cut. If the 4C -ligase samples did not migrate at an expected molecular weight (as determined by generating the theoretical average size of a genomic fragment digested by the enzyme), the restriction digestion was repeated.

4C ligations were then set up using 12.5µg of digested 3C library as template. 12.5µg of digested 3C DNA was placed in a total volume of 4ml of 1X NEB Ligation Buffer and 6.25U of T4 DNA Ligase (Life Sciences). Samples were incubated at 16˚C for 4 hours, phenol:chrolorphorm:isoamyl alcohol (1:1:24) extracted using standard procedures, and ethanol precipitated O/N at -20˚C. Samples were spun and washed in 70% ice-cold ethanol. Pellets were resuspended in 200µl of water. 4C templates were then linearized with the appropriate linearization enzyme in its appropriate buffer/temperature in a total of 300µl for 4hrs. Samples were phenol:chrolorphorm:isoamyl alcohol (1:1:24) extracted using standard procedures, and ethanol precipitated O/N at -20˚C. Linearized samples were spun, washed with ice-cold 70% ethanol, and brought up in 25µl of water. Concentrations were determined with the qBit fluorometer and adjusted to 100ng/µl (or 50ng/µl, if 100ng/µl was not possible).

PCR conditions at 35 PCR cycles using the Platinum Pfx polymerase (Life Sciences) were optimized by testing various melting temperatures, Mg^{2+} concentrations, and Platinum Pfx Enhancer solution concentrations. If the stereotypical 4C bands of ~300bp and ~500bp were observed in a manner that depended upon the amount of 4C library input (1ng, 25ng, 50ng, 100ng, and 200ng), but not in 3C (200ng input), 3C –ligase (200ng input), and 4C –ligase (200ng input) controls, then the PCR primers were considered optimized. 5 PCR reactions were then set up to amplify 200ng of 4C template per reaction at 20 PCR cycles (1µg of total 4C template used). As paired-end Illumina sequencing fails for large DNA fragments, 4C PCRs were combined, ethanol precipitated at -80˚C for 1hr, washed with ice-cold 70% ethanol, brought up in 25µl of TE, and size selected on 2% agarose gels (BioRad low range agarose) between 150bp and 650bp. Size selected 4C libraries were then amplified in a linear range (as determined by qPCR

per library) with outer primers containing the remaining TruSeq or Nextera sequences (including TruSeq Indexes). At least three amplification PCRs per 4C library were combined and purified with AmpureXP beads (BioRad). A small aliquot of each purified 4C library was run on a gel to ensure the proper banding pattern was observed (as compared to the initial PCR optimization steps). Each amplified 4C library was adjusted to 15nM and multiple 4C libraries (up to 12) were mixed for paired-end 100bp sequencing on an Illumina HiSeq 2000 (Illumina).

4C Restriction Enzymes, Primers, and SNP Information

4C Anchor	3C Fragment Coordinates	3C Enzyme	4C Enzyme	Linearizing Enzyme
Nkap TSS	chrX:34666470-34674269	HindIII	HpyCH4V	PsiI
Taf1 TSS	chrX:98727675-98729180	HindIII	TaqI	NcoI
Ogt TSS	chrX:98834142-98837509	HindIII	CviQ1	AseI
Ftx TSS	chrX:100817221-100820507	BglII	MluCI	DraI
Rlim TSS	chrX:101169720-101178474	HindIII	MluCI	NdeI
Huwe1 TSS	chrX:148237739-148240822	BglII	CviQ1	SphI
Kdm5c TSS	chrX:148667130-148670431	HindIII	TaqI/MspI	DraI

4C Anchor	Primer 1 Sequence	Primer 2 Sequence
Nkap TSS	GGAACTGTTTCCAAGTTGGTTT	GCAAACAAAAGATCATGATTAGGG
Taf1 TSS	CAGTCACTGATGCTGAGGTGTT	TTCCCATGGTAAAATGTTAAGC
Ogt TSS	GGAGACTGTCCGTTTTTCTCAT	TTGATAATATTGTTCATTTCTTCTGC
Ftx TSS	GATTTGAGCGAAGGACAACTTA	AAGTGCTTCTACATTGGTTGAAA
Rlim TSS	AGAGGGATTACACTCCCATGTC	AAACTTTTTCCCCATGATTGAA
Huwe1 TSS	GTCGGGCCGCCTGTTAAGAT	TCGCCTTTAGGAAACATGAGAT
Kdm5c TSS	AGCAGTAGACACGCGGAATG	CTAGAGAATGTGGAGTTTTAGAAGC

4C Anchor	Informative SNP (B6->Cast)	SNP Coordinate
Nkap TSS	T->C	chrX:34666869
Taf1 TSS	G->T	chrX:98728468
Ogt TSS	G->T	chrX:98834179
Ftx	C->A	chrX:100817663
Rlim TSS	G->A	chrX:101169907
Huwe1 TSS	G->T	chrX:148238234
Kdm5c TSS	G->A	chrX:148667265

Determining Allele Specific 4C-Seq Contact Probabilities

Filtering of raw data and allelic assignment of reads

Custom Perl scripts were used to filter raw sequencing reads based upon the primer sequenced used and any expected SNPs on the known anchor fragment. For any given anchor, the number of sequencing reads originating from the Xi and Xa anchor was equivalent, demonstrating that primer pairs hybridized equally well to both the B6 and Cast alleles (data not shown). The portion of unknown/contact sequence between the restriction enzyme site associated with the primer and the next used restriction enzyme site was then mapped to both the B6 and Cast (KEANE

et al. 2011; YALCIN *et al.* 2011) genomes using Bowtie (version 0.12..7,(LANGMEAD *et al.* 2009)), allowing for 2 mismatches, and only uniquely mapping sequences (settings: -n 2 -l 100 -m 1 --best --strata). Reads were then paired. Based on our primer design strategy, the sequencing read containing the informative SNP that determines the anchor allele will always be the forward read, while the reverse read will not contain allelic information for the anchor point. Therefore, a forward read where the anchor allele could be identified and the unknown/contact sequence is mappable was always included in the analysis. Reverse reads served to increase the chance of finding an informative SNP within the unknown/contact sequence. Only mappable reverse reads were used. If a forward read was not mappable due to non-unique sequence, but the reverse read was unique, then this read pair was informative. Any sequence that mapped to a unique region of the B6 genome, but also mapped to a different unique region of the Cast genome was discarded. Pairs where the forward read was mapped to a different chromosome than the reverse read were also discarded. The ability for a read to pair did not increment its value in statistical analysis.

Two sets of data were generated for analysis. One set was blind to allelic assignment and is referred to as "all" or "non-allelic" data. The other set only contained data where an allele could be assigned to the unknown/contact sequence. In order to assign unknown/contact sequences to either the B6 or Cast genome, we searched through each sequence looking for mismatches identified by Bowtie. Because each read was matched to both the B6 and Cast genomes, if a mismatch was in the same position as a SNP, the read was not useful for allelic analysis and was assigned to "all" data for that anchor/genome file. If there were no mismatches or a mismatch did not align with a known SNP, the read was considered "allelic" for that particular genome. Therefore, any primer pair would give rise to 8 files used for statistical analysis:

Xi anchor to B6 genome (all)	Xa anchor to B6 genome (all).
Xi anchor to Cast genome (all)	Xa anchor to Cast genome (all).
Xi anchor to B6 genome (allelic)	Xa anchor to B6 genome (allelic)
Xi anchor to Cast genome (allelic)	Xa anchor to Cast genome (allelic)

In general, the B6 and Cast genome (all) files for a given anchor were extremely similar, with differences attributed to allele-specific reads.

Statistical analysis of mapped reads

The methods used for statistical analysis are detailed elsewhere (WILLIAMS JR. *et al.* 2014), though they are based upon previously published methods (SPLINTER *et al.* 2011). What follows is a basic outline of our method.

Previous 4C-Seq analysis of X-linked anchors used a statistical method that reduces the number of reads mapped to any given genomic fragment to a value of 1, while genomic fragments lacking mapped reads are assigned a value of 0 (SPLINTER *et al.* 2011). This transformation guards against the detection of false positives due to technical problems such as PCR amplification bias. However, the reduction of all fragments to either 1 or 0 comes at a cost of losing statistical power and, consequently, resolution (THONGJUEA *et al.* 2013).

Mapped reads were "mapped" to a "genome" containing only coordinates of the 3C enzyme used to generate the 3C library template. Various factors, including the length of the primers, mappability of the 3C fragment, differential 3C restriction sites in the Cast and B6 genomes, the length of the 4C fragment generated by the digestion scheme, the ability for the fragment to be cut by the 4C enzyme, and the ability for the 3C fragment to be linearized were taken into account. Reads mapping to the anchor 3C fragment were removed from allelic files, as these represent self-ligation products. Additionally, if a forward and reverse pair did not map to the same expected restriction fragment, they were eliminated from the analysis.

PCR amplification bias was assessed using the 4-mer barcodes previously described. For each sequencing ID, the barcode was determined. Then, for each genomic position, the total number of reads and barcodes were determined. If a the ratio of barcodes:reads was less than 0.85 at any position, each barcode was examined to determine its contribution to the total reads. The read count of the barcode was then reduced to 1. In no instance did this reduction alter the statistical call of the interaction. We therefore determined that PCR amplification bias was not a problem in our datasets.

We also compared the profiles of mapped raw reads to profiles of "binarized" reads across chromosomes (SPLINTER *et al.* 2011). "Binarizing" data reduces each genomic position to a value of 1 or 0, depending on the presence of reads. Profiles of raw and "binarized" data drawn at windows of 100 consecutive restriction sites were essentially identical (data not shown), further strengthening our confidence that our data lacked PCR amplification bias. We therefore used the full raw reads for statistical analysis.

A sliding window approach was utilized to detect genomic regions enriched for sequencing reads. A window consists of x number of consecutive 3C restriction fragments. To generate a background model for significance, the experimental reads on each chromosome were shuffled randomly across the chromosome. A sliding window was used to determine the number of reads (Mrand) needed to surpass a FDR of 0.01. This random shuffling was done 1000 times. After 1000 permutations, the significance threshold for the chromosome was defined as the minimum Mrand value required to exceed the top 5% of all Mrand values. This threshold was then applied to the actual experimental data. In this way, the experimental data is used to empirically derive the background expectations. We tested window sizes of 200, 100, 20, 5, and 3. We found that a size of 3 gave the best results based upon comparing the positions of called interactions vs. where the actual reads underneath each interaction mapped.

Because anchors tend to generate more contacts within their local vicinity, there is a higher likelihood that contacts generated within +/- 35Kb of the anchor point will be detected due to random chance vs. a functional contact. When analyzing chrX, we performed two analyses. In the first, any reads outside of +/- 35Kb from the anchor point were masked out. This essentially created a higher threshold for significance in 70Kb surrounding the anchor. In the second analysis, reads mapping to +/- 35Kb from the anchor were masked out, preventing reads within an area known to have higher background from inflating the significance threshold for the entire chromosome. Results from both analyses were combined to form the final set of contacts. Because allelic data only represented a small amount of our data (no more than ~35%), we initially called interactions per anchor, per replicate, using the B6 to B6 and Cast to Cast (all) files.

Allelic assignment of interactions

After interactions were initially called, we assigned alleles to each called interaction. Reads contributing to each interaction were analyzed for allelic content using the corresponding anchor to B6 and anchor to Cast allelic files. Only interactions containing at least 20 allelic reads (Cast+B6) and at least 5 allelic reads of either Cast or B6 were analyzed. If an interaction did not meet these criteria, it was given a NoCall designation. NoCall interactions on the autosomes were designated as *trans*, however, NoCall interactions on the X could not be called as *cis* to either the Xi or Xa.

For each interaction containing the minimum number of allelic reads, we determined the probability of randomly detecting the observed number of B6 and Cast reads within the observed total number of N reads within the interaction. To this end, we determined the total number of 3C and associated (closest) 4C restriction enzyme

sites in the genome that contained a detectable SNP. The definition of a detectable SNP differed per anchor due to the length of sequence available once the known anchor primer/anchor sequence had been removed from raw reads. 3C and 4C ends falling within interactions were excluded from the total pool of available 3C and 4C ends. We then picked N available 3C and 4C fragments at random and calculated the ratio of SNP-containing ends to non-SNP-containing ends. This was done 1000 times per interaction. If the probability of finding the observed B6:N ratio at random was less than or equal to 0.05, but the probability of randomly finding the observed ratio of Cast:N was greater than 0.05, the interaction was assigned to B6. The converse was true for making a Cast call. If the probability of detecting both B6:N and Cast:N ratios was less than or equal to 0.05, the call was made as B6 and Cast (the interaction is on both the B6 and Cast allele of that genomic location). *Trans* and *cis* calls were then made accordingly, depending upon the anchor point and the contact point.

Generation of final interactions from biological replicates

False positive interactions called by 4C-Seq can be due to either random collisions between two DNA molecules at the time of fixation or PCR bias during PCR amplifications. To minimize both of these possibilities, 4C-Seq was performed in biological replicate. After initial calling and allele assignment, interactions generated by each anchor in each replicate were compared. If the genomic coordinates and allelic calls (including non-allelic) were identical, the interaction was kept. If interactions from different replicates had identical allelic calls, but overlapping (not exact) genomic coordinates, the boundaries of the interaction were expanded to the most 5' and most 3' boundaries of the interactions in each replicate. Interactions that did not overlap were not retained.

Analysis of genomic features within interactions

Datasets of RNA-Seq, ChIP-Seq, and DNAse-Seq in C/B cells were obtained from Calabrese et al., 2012 (GEO accession GSE39406). The allelic calls originally made in these datasets only took into account sequences on the X chromosome. A re-analysis of the same raw data, but taking the whole genome into account, was done in order to compare it to the 4C-Seq data. The methodology was identical to that described in Calabrese et al., 2012, but background models were based upon genome wide data, not just the X chromosome. Datasets for annotated repeats were downloaded from the UCSC Genome Browser tracks (MEYER *et al.* 2013).

Allelic interactions were correlated with these datasets by comparing the genomic coordinates and allelic calls (if appropriate) for any given feature and comparing it to the genomic coordinates and allelic calls for allelic interactions. Total occurrences of features falling within X-linked or genome-wide datasets were calculated. In some cases, due to the windowing method used to call interactions, a 3C fragment containing no read data was included between 3C fragments containing read data. In such a situation, features found within this region were not counted.

For ChIP-Seq data, enrichment was calculated by randomizing interaction peaks per chromosome and comparing the observed number of occurrences of a given ChIP-Seq allelic peak within 1000 random permutations. Random "interactions" were of the same size and similar mappability to the original interaction. Enrichment was considered at a p-value of <= 0.05.

To properly compare gene and repeat content among interactions generated by different anchor points, we generated normalized indices. These indices represent the total number of features found within all allelic interactions of a given anchor divided by the total number of nucleotide bases covered by all allelic interactions of that anchor. Therefore, for genes, the index is in genes per kilobase, while repeats are in repeats per kilobase. Where pairs of anchor points were compared (e.g. Xi anchor vs. Xa anchor), p-Values generated via a Paired, Two-tailed t-Test. Otherwise, p-Values were generated with a Two sample, Two-tailed t-Test.

All scripts used for 4C-Seq analysis are available upon request.

FISH Confirmation of Interactions Identified by 4C-Seq

DNA/RNA Fluorescent *in situ* hybridization (FISH) in TS cells cultured on glass coverslips was used to confirm interactions as described previously (CALABRESE *et al.* 2012). RNA FISH against the *Xist* RNA was used to identify the Xi vs. the Xa. 7 pairs of bacterial artificial chromosomes (BACs) were used to generate FISH probes pairs (see below). Grayscale z-stacks were obtained on a Zeiss AxioImager M2 (Carl Zeiss) and deconvolved using Axiovision software (Carl Zeiss). 3D distances were measured between the centers of DNA FISH signals using ZEN software (Carl Zeiss). For Xi to Xi measurements, the corresponding Xa to Xa measurement constituted a no contact control. Similarly for, Xa to Xa measurements, the corresponding Xi to Xi measurement constituted a no contact control. 2 pairs of no contact controls were also chosen. At least 80 nuclei were counted per probe pair. Representative images are z-projections of each channel imaged, pseudocolored and merged in Photoshop CS4 (Adobe Systems). The R Statistical Software

Package (version 2.15.2) was used to generate box plots of combined measurements. Paired, Two Sample t-Tests were used to determine p-values.

FISH BAC and Fosmid Probe Information

BAC/FOS#	Clone	Coordinates	Detects
FOS1	WI1-1863K22	chrX:34653654-34695245	*Nkap* Anchor
BAC2	RP23-81P18	chrX:46494480-46742488	Contact
BAC3	RP23-224F24	chrX:148201629-148415678	*Huwe1* Anchor
BAC4	RP23-99I10	chrX:91400085-91627729	Contact
BAC5	RP23-272J22	chrX:148569748-148807377	*Kdm5c* Anchor
BAC6	RP24-164M7	chrX:100721189-100887145	*Ftx* Anchor
BAC7	RP23-133E13	chrX:159658600-159867586	Contact
FOS8	WI1-2704K12	chrX:101145921-101187238	*Rlim* Anchor

Contact Tested	Anchor Probe	Contact Probe
Xi to Xi1	FOS1	BAC2
Xi to Xi2	BAC3	BAC4
Xa to Xa	BAC5	BAC4
Both1	FOS8	BAC4
Both2	BAC6	BAC4
No Contact1	FOS1	BAC4
No Contact2	BAC3	BAC7

All BAC/FOS probe mixes also contained a probe for the *Xist* RNA made from a region spanning Exon 1 of the *Xist* transcript (MUGFORD *et al.* 2012).

Supplemental References

CALABRESE, J. M., W. SUN, L. SONG, J. W. MUGFORD, L. WILLIAMS *et al.*, 2012 Site-specific silencing of regulatory elements as a mechanism of X inactivation. Cell **151:** 951-963.

KEANE, T. M., L. GOODSTADT, P. DANECEK, M. A. WHITE, K. WONG *et al.*, 2011 Mouse genomic variation and its effect on phenotypes and gene regulation. Nature **477:** 289-294.

LANGMEAD, B., C. TRAPNELL, M. POP and S. L. SALZBERG, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10:** R25.

MEYER, L. R., A. S. ZWEIG, A. S. HINRICHS, D. KAROLCHIK, R. M. KUHN *et al.*, 2013 The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res **41:** D64-69.

MUGFORD, J. W., D. YEE and T. MAGNUSON, 2012 Failure of extra-embryonic progenitor maintenance in the absence of dosage compensation. Development **139:** 2130-2138.

SPLINTER, E., E. DE WIT, E. P. NORA, P. KLOUS, H. J. VAN DE WERKEN *et al.*, 2011 The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes & development **25:** 1371-1383.

THONGJUEA, S., R. STADHOUDERS, F. G. GROSVELD, E. SOLER and B. LENHARD, 2013 r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. Nucleic Acids Res **41:** e132.

WILLIAMS JR., R. L., J. STARMER, J. W. MUGFORD, J. M. CALABRESE, P. MIECZKOWSKI *et al.*, 2014 fourSig: A Method for Determining Chromosomal Interactions in 4C-Seq Data. Nucleic Acids Res **In Press**.

YALCIN, B., K. WONG, A. AGAM, M. GOODSON, T. M. KEANE *et al.*, 2011 Sequence-based characterization of structural variation in the mouse genome. Nature **477:** 326-329.