https://doi.org/10.1093/nargab/lqad052

Refining the genomic determinants underlying escape from X-chromosome inactivation

Samantha Peeters[1], Tiffany Leung[1,2], Oriol Fornes [1,2], Rachelle A. Farkas[1,2], Wyeth W. Wasserman [1,2] and Carolyn J. Brown [1,*]

[1]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada and [2]Centre for Molecular Medicine and Therapeutics at British Columbia Children's Hospital, University of British Columbia, Vancouver, British Columbia, Canada

Received February 01, 2023; Revised May 01, 2023; Editorial Decision May 08, 2023; Accepted May 19, 2023

ABSTRACT

X-chromosome inactivation (XCI) epigenetically silences one X chromosome in every cell in female mammals. Although the majority of X-linked genes are silenced, in humans 20% or more are able to escape inactivation and continue to be expressed. Such escape genes are important contributors to sex differences in gene expression, and may impact the phenotypes of X aneuploidies; yet the mechanisms regulating escape from XCI are not understood. We have performed an enrichment analysis of transcription factor binding on the X chromosome, providing new evidence for enriched factors at the transcription start sites of escape genes. The top escape-enriched transcription factors were detected at the *RPS4X* promoter, a well-described human escape gene previously demonstrated to escape from XCI in a transgenic mouse model. Using a cell line model system that allows for targeted integration and inactivation of transgenes on the mouse X chromosome, we further assessed combinations of *RPS4X* promoter and genic elements for their ability to drive escape from XCI. We identified a small transgenic construct of only 6 kb capable of robust escape from XCI, establishing that gene-proximal elements are sufficient to permit escape, and highlighting the additive effect of multiple elements that work together in a context-specific fashion.

INTRODUCTION

The mammalian sex chromosomes (X and Y) are derived from an ancestral pair of autosomes, but have diverged significantly throughout evolution in order to suppress recombination and conserve sex-determining genes. As females generally have two copies of the more gene-rich X chromosome compared to one in males, it is hypothesized that X-chromosome inactivation (XCI) evolved to compensate for the difference in dosage between the sexes by silencing all but one X chromosome in every cell in female mammals. Despite the physical condensation and heterochromatic environment of the inactive X (Xi), a substantial number of genes are able to escape the silencing and continue to be expressed from both human X chromosomes, albeit at lower levels from the Xi than the active X (Xa) copy (1). The threshold to call such escape has historically been expression from the Xi of at least 10% the level of expression from the Xa (1,2), although escape definitions continue to expand as models systems and new statistical methods evolve (e.g. (3–5)). Comparing expression levels between males and females (and X aneuploidies) can also provide evidence suggestive of escape (6). In addition to gene expression, epigenomic features that differentiate active and inactive genes, such as the inverse correlation between DNA methylation (DNAm) of X-linked gene promoters and gene activity, have been established as being predictive of the inactivation status of X-linked genes (7–9).

An aggregation of multiple datasets in humans determined that about 12% of X-linked genes (~80 genes) consistently escape inactivation, while another 15% (~93 genes) are variable across tissues and/or individuals (10). The list of escapees includes all characterized pseudoautosomal region (PAR) 1 genes (~25) (10), two (of four) PAR2 genes, and 12 (of 14 informative, 17 total) genes with functional X-Y gametologues outside the PAR (11), leaving many additional genes that escape inactivation but lack expressed Y gametologues. The two more centromeric PAR2 genes are silenced on both the Xi and Y (12,13), demonstrating that evading inactivation is not a basic characteristic of genes with Y gametologues.

The extent to which XCI is shared between cells and tissues was further characterized in an extensive survey of GTEx data that found evidence for escape from XCI for 23% of X-chromosomal genes with expression heterogeneity between tissues, individuals and cells, resulting in a range of sex biases in gene expression for 29 tissues ((6), reviewed

[*]To whom correspondence should be addressed. Tel: +1 604 822 0908; Email: carolyn.brown@ubc.ca

in (14)). The differential expression of escapees in females can manifest in profound impacts on health, such as offering a protective effect against *de novo* and inherited X-linked mutations like those found in certain types of cancer (15). However, escapees have also been proposed to contribute to the over-representation of females for some complex traits such as autoimmune disorders (reviewed in (16)). Determining which genes escape from inactivation reveals an important source of sexually dimorphic gene expression, and identifying the mechanism by which these exceptions occur will inform our understanding of XCI and broader questions of selective epigenetic repression of genes.

A recent study using DNAm of X-linked genes with CpG islands expanded our knowledge of X-inactivation status across 12 different species, and found that most species had 10–20% of genes (excluding PARs) escape from XCI (17). In contrast to humans and most other species examined, mouse is an outlier with considerably fewer genes escaping inactivation. Data suggests only 3–7% of mouse genes are expressed from the Xi, including 16 constitutive escape genes, and approximately 10–20 variable escape genes depending on origin of the cell and threshold used to call escape (4,17). As approximately one half of the mouse escape genes also escape in humans, there is likely to be some conservation of the elements and mechanisms involved between species. While human genes often escape in larger blocks clustered on the p arm of the X, mouse escapees are predominantly singletons, suggestive of local regulatory elements driving expression (reviewed in (18)). A subset of escape genes which are conserved across all species also suggests more proximal or gene-specific regulation as a factor. These conserved escape genes are distributed along the p arm in humans and most have either a conserved Y homolog or Y pseudogene, suggesting relatively recent loss of Y homology (19). However, the escape genes that are discordant in XCI status between species often 'switch' status as a block, suggesting some domain regulation is also involved in their expression (17).

Some of the strongest evidence for the existence of an intrinsic 'escape element' in the DNA sequence in or near an escape gene comes from a series of random X-linked BAC integrations containing the mouse escape gene *Kdm5c*, where it was demonstrated that escape from XCI was an intrinsic property of the locus (20). Indeed, other mouse studies have suggested a model of regulatory control of escape that is mediated by genomic elements lying in close linear proximity to escaping genes (21). We recently expanded the use of BAC integrations to establish that mouse cells have functional capacity to support escape of the human gene *RPS4X* (and variable escape gene *CITED1*) across several tissues and developmental time points (22,23). Transgenic mice were generated with a human BAC containing the escape gene *RPS4X*, variable gene *CITED1* and subject gene *ERCC6L* (RP11-1145H7, ∼158 kb), which was integrated at a docking site 5' of the *Hprt* locus on the mouse X chromosome. *Hprt* is normally subject to inactivation in mouse, and many transgenes integrated at this locus have been subject to XCI when on the Xi (24). Analyses of Xi gene expression and promoter DNAm demonstrated that mouse was able to correctly recapitulate the escape and subject statuses of all three intact human genes on the BAC (22,23). This suggested that intrinsic escape elements within the *RPS4X* BAC share recognizable properties between mouse and human, despite the majority of mouse genes being subject to inactivation, including the mouse ortholog *Rps4x*.

The variability in escape from XCI for *RPS4X* across species can be seen as a microcosm for evolution of the sex chromosomes. While the ancestral mammalian sex chromosomes contained a Y-linked copy of *RPS4X*, primates appear to be the only lineage retaining (and duplicating) this gene, as well as retaining ongoing expression of *RPS4X* from the Xi, as in other mammals lacking this gametologue, including mouse, *RPS4X* is subject to XCI (11,17,25,26). The transition from an autosomal pair of chromosomes to an X-linked gene subject to inactivation (without a Y homolog), has been proposed to begin with Y decay, followed by upregulation of X-linked expression, and then spread of XCI across the gene (25). According to such a model, escape genes may lack features that enable silencing. However, the vast majority of transgenes with autosomal promoters become subject to XCI when integrated onto the X chromosome, implicating further unique features allowing ongoing expression from the Xi. A variety of epigenetic and genetic features have been implicated in controlling whether a gene escapes from XCI, yet none of these features alone have had the power to predict a gene's XCI status. In general, bioinformatics analyses have identified enrichment of sequence motifs and transcription factor (TF) peaks such as YY1 and CTCF near escape genes on the X (27,28), with CTCF binding thought to play a role in chromatin loops and boundary formation between subject and escape genes (4,29,30). A positive correlation has also been observed between SINEs and escape from inactivation in both mouse (28) and human (Alu elements (8,31)), in both a promoter-centric context as well as larger domains of multiple escape genes. Alu elements have a high potential to modulate gene transcription by binding several TFs (32), and can also contribute CTCF binding sites (33,34), which may explain why they are found in close vicinity to escape genes. The X chromosome as a whole is enriched for L1 repetitive elements compared to autosomes; however, there is a reduction in LINEs around genes that escape inactivation (8,31,35). Combinatory models in both mouse and human have established some predictive models for XCI status (5,9,31), however, they cannot correctly predict all classes of genes, highlighting the complexity of escape regulation, and that unique combinations of elements have yet to be identified for some genes.

Many bioinformatics studies have relied upon motif and sequence enrichment analyses; however, such analyses are limited as they do not consider experimentally supported binding events at their locations, and have a large number of incorrectly predicted binding events that increases with the length of the region analyzed (36). Additionally, some TFs will avoid capture through these methods as they do not have a determined DNA-binding motif, or they act in cooperation with other factors to uniquely bind to DNA. To reduce these limitations, we have undertaken a new enrichment analysis of TF binding on the X utilizing the ReMap database, which has compiled and uniformly reprocessed thousands of public DNA-binding experiments of transcriptional regulators (37). With new evidence for escape-enriched TFs as well as hypothesized elements from previous literature, we revisit the *RPS4X* gene region to further refine specific regions that are necessary for escape from XCI in a cell line model.

MATERIALS AND METHODS

Data preparation for TF analyses

The inactivation status of X-linked genes was obtained from a previous meta-analysis (10), in which inactivation status categories were assigned to genes based on expression level and methylation status. We combined the two categories with strongly reproducible evidence for escape (29 'Escape' and 26 'Mostly Escape') for a total of 55 escape genes. Subject genes consist of 331 'Subject' and 131 'Mostly Subject' genes for a total of 462 genes. Variable escape genes as well as those in the PAR were excluded. The transcription start sites (TSSs) of human genes were obtained from NCBI RefSeq Select, which provides a single representative transcript for every protein-coding gene (38). For genes not annotated in NCBI RefSeq Select, the TSS coordinates were obtained from the meta-analysis (10) and then converted from hg19 to hg38 build using LiftOver (39). The TSS region of each gene was denoted as the 500 bp upstream and downstream of the TSS. The 'upshell' region was the 10kb upstream of the TSS region (i.e. opposite of the direction of transcription), whereas the 'downshell' region was the 10 kb downstream of the TSS region. All analyzed X-linked gene and region of interest (ROI) coordinates are listed in Supplementary Table 1.

The non-redundant ChIP-seq peaks were obtained from ReMap 2022, in which clusters of duplicate peaks were used to determine an average start, end and peak summit of coordinates with peaks trimmed to the median size of peaks for each specific TF (37). Peaks were obtained for 154 TFs characterized on the X chromosome in GM12878 cells (Supplementary Table 2). As the GC percentage varies regionally in genomes and is correlated with functionality, we matched GC composition for the ROI in order to reduce confounding influence by the GC frequency on the enrichment analysis. A background was created by matching each escape gene with 5 random genes subject to XCI having similar GC composition (within 3%) in the ROI to create a full dataset to be used for enrichment analysis. Furthermore, to allow proper comparison within the set of annotated ROI and avoid counting multiple overlaps of ChIP-seq peaks within a single region, the ChIP-seq peaks were intersected with the GC-matched full dataset to retain only peaks that overlap the ROI. Only one instance of overlap per TF was counted, regardless of the number of overlapping Chip-seq peaks.

TF enrichment analysis

GIGGLE (40), version 0.6.3 obtained from https://github.com/ryanlayer/giggle, is a package that identifies and ranks the significance of overlaps between provided genomic regions of a query and features of interest. Background sets were processed via the GIGGLE index command with default parameters. For enrichment analysis, GIGGLE was executed with additional options: -s -g < background size > . -s was added to output significance per feature analyzed and -g was added to provide a more accurate genome size for significance testing, and the input genome size was calculated as the size of the background set multiplied by the length of TSS region or shells, 1000 or 10 000 bp, respectively.

GIGGLE outputs several metrics for the enrichment analysis, including odds ratio, Fisher's tail p-values and a GIGGLE score. The GIGGLE score is the product of $-\log_{10}(P\text{-value})$ and $\log_2(\text{odds ratio})$, which can help with interpretation as the two metrics reflect different related but complementary properties (40). Overlap ratio, the number of overlapped TF peaks divided by the number of TFs in the background set, was also calculated for each TF. The thresholds for defining enrichment of a TF in escape gene regions were: (a) odds ratio >1; (b) top 15% of TFs based on Fisher's left tail *P*-value rank in an ascending manner; (c) top 15% of TFs based on GIGGLE score rank in a descending manner; and d) overlap ratio >0.167 ($\frac{1}{6}$ as the ratio of number of escape genes within the background set).

TF co-binding analysis

To measure the co-binding of two TFs in a region, the Dice similarity coefficient (DSC) was used to compare the agreement between the gene set overlapped by the TFs (41). To determine TF pairs that are specifically enriched in escape regions, the DSCs obtained for the overlapped escape gene set were divided by the DSCs obtained for the overlapped background gene set, similar to the concept of an expected-over-observed ratio. Only TF pairs with a DSC ratio over 1.5 were considered. Fisher's exact test was performed to evaluate the significance of co-binding. The *P*-values obtained from the Fisher's exact test were corrected for multiple testing by the Benjamini–Hochberg method for a false discovery rate of 0.1 (Supplementary Table 3).

TF distribution analysis

Escape and subject gene TSS regions were obtained as described above. Autosomal and remaining X gene TSSs were also obtained from NCBI RefSeq Select (38), and the TSS region of each gene was denoted as the 500bp upstream and downstream of the TSS. A Chi-Square Goodness of Fit Test was used to compare TF distributions of the 154 TFs in GM12878 between different categories of genes. Autosomal and escape genes were used as expected distributions for comparison to X chromosome and subject genes, respectively. Each distribution was divided into bin sizes of 10, e.g. 0–9, 10–19. The frequency of each bin was determined by dividing the count by the sum of all counts in the corresponding distribution. The test statistic was compared to a Chi-square critical value at a significance level of 0.05 with 8 degrees of freedom. The distributions were determined to be significantly different if the chi-square test value was larger than the chi-square critical value (Supplementary Table 3).

TF distribution analysis with matched gene expression

A subset of autosomal genes that matched the expression level of the escape and subject genes were selected and

TF distribution at the TSS regions was performed as described above. The expression data was obtained from Expression Atlas under the experiment 'RNA-seq of long poly-adenylated RNA and long non poly-adenylated RNA from ENCODE cell lines' (GEO GSE26284). Specifically, we used the processed whole-cell long polyA RNA expression data for GM12878. To match expression for comparison, each escape or subject gene was matched with a single autosomal gene with the same TPM, or the closest TPM if there was not a gene with the same expression.

X-escape construct (XEC) design and generation

Bioinformatics protocols for the *RPS4X* XEC designs were adapted from previous works (42,43), with slight modifications. Identification of promoters and other regulatory regions was limited to within the RP11-1145H7 BAC (hg38:chrX:72193017–72351571), previously shown to recapitulate human escape of *RPS4X* from XCI in mouse (22). Designs relied on the integration of multiple sources of evidence: CpG islands (44), candidate *cis*-regulatory elements from ENCODE (45), chromatin accessibility (DNase I hypersensitivity) and histone modifications (H3K4me1 and H3K4me3), also from ENCODE, TF-bound regions in GM12878 cells from ReMap 2022 (37) for the top escape TSS-enriched TFs (EE-TFs) plus CTCF and YY1, multi-species conservation (46), and SINE and LINE repetitive elements.

The *PGK1* promoter was cloned from p5E-hPGK, a gift from Dr. Kryn Stankunas (Addgene plasmid #82579, (47)). All *RPS4X* XEC sequences were synthesized by GenScript and cloned into a modified *Hprt* homologous recombination targeting plasmid, designed to integrate constructs 5' of the *Hprt* gene on the mouse 129 X chromosome. pEMS2001 (48) was a gift from Dr Elizabeth M. Simpson (Addgene plasmid # 105871) and was modified to reduce homology arm length and change the reporter to *EmGFP* for XECs 1–3, and 5. pEMS2001 also contains a complementary sequence that rescues *HPRT1* activity through creation of a chimaeric locus consisting of the human *HPRT1* promoter and exon 1 and mouse *Hprt* exons 2–9 (49). Constructs were sequence-verified across all junctions after final stages of cloning. All XEC regulatory region coordinates are listed in Supplementary Table 4.

Modification of embryonic stem cells (ESCs) for use as experimental model

The inducible Xist (iXist) cell line is a female F1 2–1 XX ESC line (129/Sv-Cast/Ei) with an endogenous *Xist* allele driven by a tetracycline inducible promoter on the 129-X, gifted from Dr. Neil Brockdorff (50). During modification of the *Xist* promoter, a recombination event occurred between the 129 and Cast X chromosomes, limiting SNPs to the 103 Mb proximal to *Xist*. The iXist cell line was further modified in our lab through CRISPR-Cas9 mutations at each *Hprt* allele in order to render the gene nonfunctional, with a deletion on the 129 allele similar to the one used previously for *Hprt* targeting (51). Lack of functional HPRT was verified by 6-thioguanine selection (6-TG, Sigma-Aldrich) as it is toxic to cells still producing the HPRT protein. gRNA sequences for *Hprt* deletion are listed in Supplementary Table 5.

ESCs were cultured without feeders on 0.1% gelatin (Fisher Chemical) coated plates at 37°C in a humid atmosphere with 5% CO_2. ESCs were grown in Dulbecco's modified Eagle's medium (DMEM; Life Technologies) supplemented with 15% fetal bovine serum (FBS, Wisent Bioproducts), 2 mM L-glutamine (Invitrogen), 0.1 mM MEM nonessential amino acid solution (Invitrogen), 1000 U/ml LIF, 3 μm GSK3 Inhibitor (CHIR99021, Millipore Sigma), 1 μm MEK Inhibitor (PD0325901, Millipore Sigma) and 0.01% β-mercaptoethanol (Sigma-Aldrich). Cells were continuously sampled for retention of two X chromosomes by testing genomic DNA by pyrosequencing for X-linked allelic ratios of *Zfx* and *Taf1* genes (Supplementary Table 5 for primer information). *Xist* expression was driven by a TetOn promoter induced by addition of 1.5 μg/ml of doxycycline (dox, Sigma) for 6 days. ESC differentiation was achieved by LIF and 2i withdrawal from the medium and low-density cell plating after 1 full day of Xist induction.

Generation of transgenic ESC lines

To increase homologous recombination at *Hprt*, a guide RNA sequence targeting 5' of *Hprt* was designed (E-CRISP online tool (52)) and cloned into pSpCas9(BB)-2A-Puro(PX459) V2.0 gifted from Feng Zhang (Addgene plasmid #62988 (53)). After cloning each candidate escape construct into the modified *Hprt* homology plasmid, these plasmids were linearized with *I-SceI* enzyme and transfected alongside the gRNA-Cas9 plasmid into ESCs at a 1:1 ratio using Lipofectamine 3000 Transfection Reagent (Invitrogen) according to manufacturer's protocol. After 24 h, transfected cells were passaged to a 100 mm gelatinized Petri dish, and media was supplemented with HAT (Gibco, 50×) 24 h later to select for reconstitution of the *HPRT1*/*Hprt* locus. Cells were grown under HAT selection for 10–12 days until colonies were picked.

Selected clones were screened for retention of two X chromosomes by SNP pyrosequencing, as well as PCRs to confirm 129-allelic integration of the escape construct. Clones were assayed for evidence of random integration and copy number by qPCR, as well as proper expression of either the *EmGFP* reporter or *RPS4X* gene. After validation, three clones were chosen as biological replicates for each escape construct.

DNA and RNA extraction

DNA and RNA extractions were performed using DNAzol and TRIzol Reagents (Invitrogen), according to the manufacturer's protocols. Nucleic acids were quantified by UV spectrophotometry (Ultraspec 2000, Pharmacia Biotech). RNA extractions were diluted to concentrations of 1 μg/μl and treated with 1 μl DNase I with 10 μl buffer (Roche) and 1 μl Ribolock (Thermo Fisher Scientific) in a volume of 50 μl at 37°C for 1 h followed by heat inactivation at 75°C for 10 min.

Gene expression analyses

For analysis of transcription, 2 μg of DNase-treated RNA was converted to cDNA using standard reverse transcription conditions with Random Hexamer Primers (ThermoFisher Scientific) and 200 U M-MLV Reverse Transcriptase (Invitrogen) in a 20 μl reaction. Reactions were carried out at 42°C for 2 h followed by 5 min incubation at 95°C. RT-qPCR was used to determine relative transcription levels of transgenes compared to the endogenous control gene *Abl1* (Supplementary Table 5). 1.5 μl of each sample was added to a master mix containing 0.1 μl GoTaq G2 Hot Start Polymerase (Promega), 4 μl 5× buffer, 1.6 μl 25 mM $MgCl_2$, 1 μl EvaGreen dye (Biotium), 0.16 μl 25 mM NTPs, 0.2 μl of each 25 μM forward and reverse primers, and sterile dH_2O to 20 μl.

Samples were run in triplicate using a QuantStudio 3 Real-Time PCR System (ThermoFisher Scientific) with conditions as follows for all primer sets: 95°C for 2 min; followed by 40 cycles of 95°C for 30 s, 60°C for 30 s and 72°C for 1 min; and a melt curve stage of 95°C for 15 s, 60°C for 1 min and an increase of 0.3°C until 95°C. Testing for multiple Tm peaks for primer specificity, as well as removal of outliers from triplicate samples were performed using QuantStudio Design and Analysis software. Negative controls of RNA without reverse transcriptase were also run to ensure that the samples contained no genomic DNA contamination. Expression levels were quantified using the comparative CT method and tested for significant differences in percent escape between groups using either one-way ANOVA or unpaired *t*-tests with Welch's correction in GraphPad Prism 5.

RT-PCR to detect splicing patterns in XEC1 was performed on cDNA with Taq DNA polymerase (Invitrogen) and conditions as follows for all primer sets: 95°C for 3 min; followed by 35 cycles of 95°C for 30 s, 56°C for 30 s and 72°C for 1 min, with a 5 min extension time at 72°C. Products were run on a 2.0% agarose gel stained with SYBR safe (Invitrogen) with the 100 bp plus GeneRuler DNA ladder (ThermoFisher Scientific).

DNAm and SNP pyrosequencing

Using the EZ DNA Methylation-Gold Kit (Zymo Research), 500 ng of DNA was bisulphite converted following the manufacturer's protocol. Internal bisulphite conversion controls were included in the pyrosequencing assays to monitor complete conversion of genomic DNA. Each 25 μl pyrosequencing PCR was performed with 2.5 μl 10× PCR buffer (Qiagen), 0.2μl 25mM dNTPs, 0.125 μl Hot Start Taq DNA polymerase (Qiagen), 0.25 μl each of 25 μM forward and reverse primers and 12–35 ng bisulfite-converted DNA. Conditions for PCR were 95°C for 15 min, 50 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 1 min and finally 72°C for 10 min. One forward or reverse primer was biotinylated, depending on which strand contained the target region to be sequenced, to subsequently isolate the strand of interest for pyrosequencing. Template preparation for pyrosequencing was done according to the manufacturer's protocol, using 10 μl of PCR products.

Runs were performed on either the PyroMark Q96 MD machine, or the PyroMark Q48 Autoprep (Qiagen). Each human promoter assay was tested in at least one mouse sample without the target transgene to ensure the specificity of the human primers. At least three CpGs in an island were evaluated and averaged per assay. SNP pyrosequencing was performed as above in both genomic and cDNA (with annealing conditions of 58.3°C) using primers that amplify a single-nucleotide polymorphism (Supplementary Table 5 for primer information). *t*-tests and one-way ANOVAs were performed using GraphPad Prism 5.

Flow cytometry

Flow cytometry of cell lines was performed on a BD LSR II Cell Analyzer with downstream analysis in FlowJo software (BD).

RESULTS

A subset of TFs show enrichment at the TSSs of escape genes

Given the evidence for gene-proximal regulatory elements in enabling escape from inactivation, we sought to assess whether there are TFs that may contribute to the ability of certain genes to escape from XCI. We used the curated and uniformly processed ChIP-seq data available in ReMap (37) for human GM12878, a female lymphoblastoid cell line with extensive data. We performed enrichment analyses for binding of 154 TFs on a curated list of 55 human escape genes (see Methods, (10)), against a background set of genes that are known to be subject to inactivation. Enrichment analyses were performed for a 1 kb region around each gene's TSS (±500 bp), as well as two 10 kb regions either upstream (upshell) or downstream (downshell) of the TSS region. In order to reduce the confounding influence of GC frequency on the enrichment analysis, each escape gene was matched with 5 random subject genes with similar GC composition, and 20 iterations of GC-matching and enrichment analysis were performed.

In the TSS region, the total number of TFs binding at each escape gene is shown in Supplementary Figure 1A. A subset of 19 TFs were found to be enriched with a threshold of 50% of iterations (Figure 1A, purple and green circles) out of 43 TFs showing any enrichment in analysis of at least one of the regions of interest (TSS, upshell or downshell). Notably, none of the 19 enriched TFs bind to all 55 escape genes; indeed, 11 of them bind to <20% of escape genes. Although some TFs bind to a high percentage of escape genes, such as ARID3A and CREM, they also show an overall high binding frequency in background (subject) genes and thus were not enriched, indicating that the analysis is not biased towards TFs that have many binding sites on the X. To increase stringency, the number of escape genes (out of 55) overlapped by a TF was considered, and the minimum percentage of enriched iterations was increased to 75% (Figure 1A, green circles). Of the TFs significantly enriched in 75% of iterations, five top escape-enriched TFs (EE-TFs), ZFP36, NIPBL, MYB, STAT1 and HSF1, were found at >20% of escape gene TSSs (Figure 1A, marked with asterisks).

TF enrichment was similarly analyzed in a larger shell around the TSS. A subset of 18 TFs in the upshell region

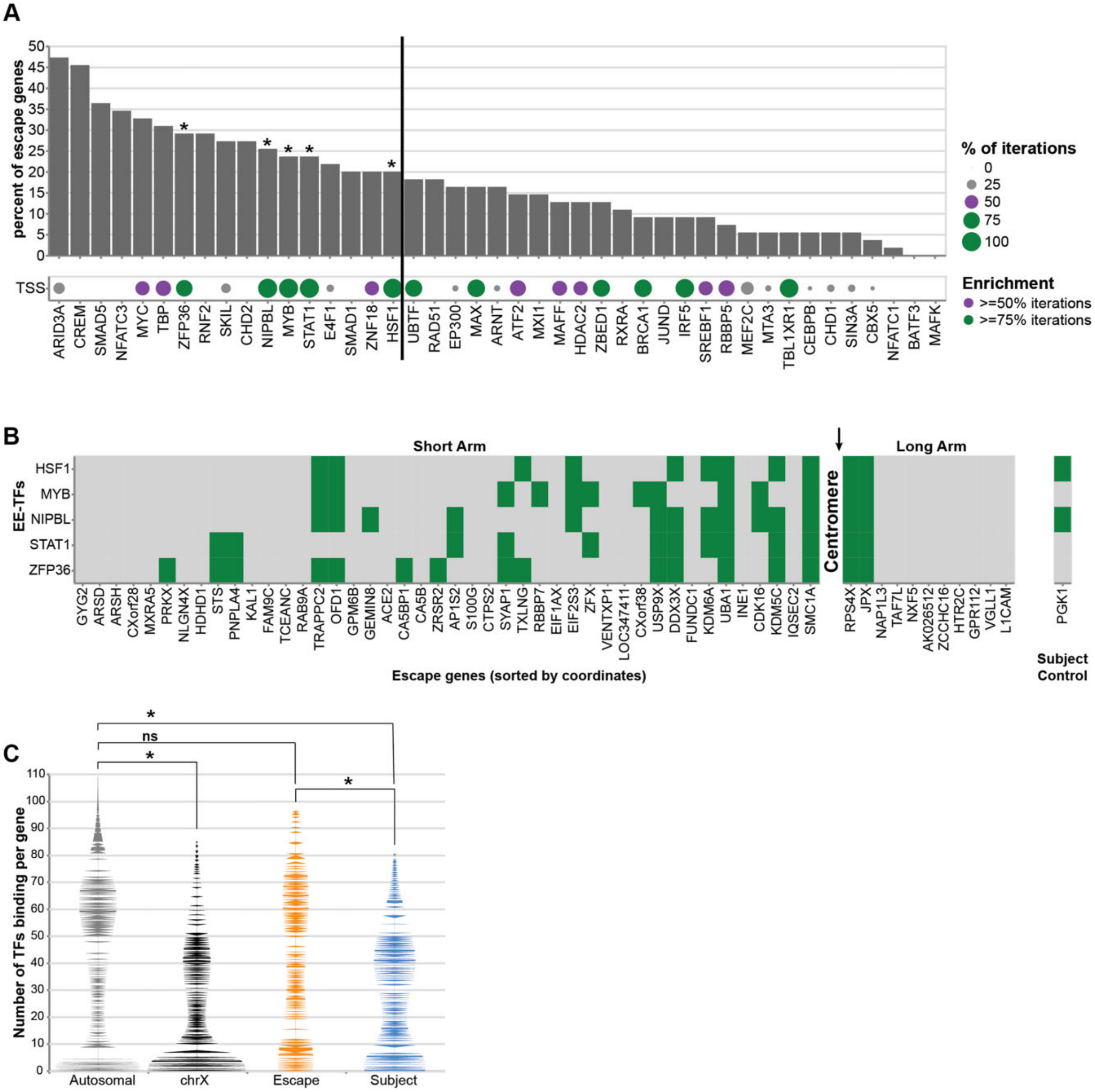

Figure 1. Transcription factor enrichment and binding distributions. (**A**) The number of escape genes overlapped by each TF is represented by a percentage, calculated out of 55 escape genes used in the analysis (y-axis). As enrichment analysis was performed for 20 iterations, the size of the circle (x-axis) indicates the number of iterations that a specific TF was considered enriched. The colour of the circle indicates that the TF has met enrichment thresholds (odds ratio ≥ 1, top 15% ranked by on Fisher's left tail p-value, top 15% ranked on GIGGLE score, overlap ratio ≥0.167) in at least 50% of the iterations (purple) or in at least 75% of the iterations (green). A subset of five TFs enriched in at least 75% of the iterations and overlapping at least 20% of escape genes (shown to the left of vertical bar) are starred. (**B**) Top five enriched TFs are plotted against 55 escape genes arranged horizontally by location on the X chromosome. Subject gene *PGK1*, used as a control for X-inactivation in the XEC experiments, is shown on the end for comparison, with an arrow indicating its location on the X. (**C**) TF binding frequencies are shown for autosomal genes compared to X-linked genes, as well as escape and subject genes on the X. Chi-square tests comparing TF distributions in bin sizes of 10 show a significant difference between autosomes and the X, as well as escape and subject genes on the X.

and 19 TFs in the downshell region were identified as enriched using the lower threshold of 50% of iterations (Supplementary Figure 1B and C). While four of the top five TSS EE-TFs were enriched in either the upshell or downshell regions at this threshold, increasing the enrichment stringency to greater than 75% of iterations with an overlap of at least 20% of escape genes resulted in only 1 TF in the downshell region, SKIL, being labelled as enriched. Distribution of SKIL across the 55 escape genes is shown in Supplementary Figure 1D.

We also examined the distribution of the top five TSS EE-TFs across the 55 escape genes, as each was seen to be enriched at just over 20% of escape genes (Figure 1B). Given that many escape genes bind either many or none of the EE-TFs, the relative incidences of co-binding events was calculated with the DSC of the escape gene TSS regions overlapped by each pair of TFs, restricting reported results to pairs including at least one of the top five EE-TFs. Ranking these pairs based on their escape-over-background ratio of DSC and corrected *P*-values (see Materials and Methods), 36 pairs of TFs met the thresholds (Supplementary Table 3). Remarkably, 6 of the top 10 ranked pairs involved ZFP36, highlighting its strong presence at escape regions. Interestingly, over a half of the escape genes do not bind any of the TSS EE-TFs at all (Figure 1B), consistent with previous hypotheses that there is not a universal mechanism for escape from XCI (e.g (16). Only four escape genes, *UBA1*, *SMC1A*, *RPS4X* and *JPX*, bind all five EE-TFs in their TSS regions (Figure 1B).

The dichotomous distribution of our EE-TFs in the escape genes led us to examine the genome-wide distribution of TF binding events. The average number of EE-TFs binding at autosomal TSSs is 1.1, similar to the escape average binding of 1.2, both standing in contrast to the binding average of 0.4 for subject genes (Supplementary Figure 2A). To explore if this TF binding distribution replicated beyond the EE-TFs, we used all 154 TFs available in ReMAP for GM12878 and plotted the number of TFs binding per gene in four different categories: autosomal, X chromosome, X-escape, and X-subject (Figure 1C). TF binding distribution for individual autosomes is shown in Supplementary Figure 2B. Compared to autosomal genes, the X chromosome has a significantly different distribution (Supplementary Table 3), with a larger proportion of genes having low numbers of TFs bound. Breaking down the X into escape and subject genes, subject genes appear to be driving the low TF-binding on the X, with the escape gene distribution more closely resembling the properties of autosomal genes. Similar to the autosome to X difference, the escape gene TF distribution is significantly different compared to the subject gene distribution (Supplementary Table 3). The apparent depletion of TF-binding could be on the Xa and/or the Xi, as the analysis could not distinguish allele-specific binding. As the X chromosome has been noted to have unique expression patterns, we wished to determine if the differences in TF-binding were driven by differences in expression. Therefore, the TF-binding enrichments were repeated using a set of autosomal genes with expression matched to either the escape or subject gene sets. The distributions (Supplementary Figure 2C,D) continued to show that escape genes were similar to autosomal genes while subject genes had substantially less TF-binding.

Overall, these findings suggest that the promoters of genes that escape XCI differ from those subject to XCI, so we wished to test whether a promoter region could provide sufficient information to drive expression from the Xi. We previously showed *RPS4X* escape from a transgenic BAC integration, and as it has binding of all 5 EE-TFs in the TSS, we reasoned that *RPS4X* makes an excellent candidate for identifying gene-proximal escape elements. We therefore hypothesized that local elements at the *RPS4X* promoter itself might be able to drive escape from XCI, and developed three *RPS4X* promoter-based constructs to test this theory.

RPS4X region includes features enriched at escape genes

Studies using a transgenic *RPS4X* BAC demonstrated the ability of mouse to recognize human elements regulating escape from XCI and stably express human escapees throughout development (22,23), yet the nature and location of the elements themselves remained elusive. Reviewing the region between subject and escape genes on the BAC, regulatory regions (RR) containing putative escape elements for *RPS4X* (and *CITED1*) must lie within ∼112 kb from the subject *ERCC6L* promoter to the 3' end of the BAC (Figure 2A). A boundary element such as CTCF binding between *ERCC6L* and *RPS4X* could help to maintain the open escape domain (Figure 2, RR4), or a distal enhancer with primate-specific features (Figure 2, RR2/RR3) could conserve activity at *RPS4X* despite the heterochromatic environment of the Xi. Additionally, the TF enrichment study strongly suggested that the promoter proximal region of *RPS4X* contained unique binding sites for escape-specific factors.

To characterize the minimal region necessary for escape from XCI, we focused our initial X-escape constructs (XEC) on the promoter region of *RPS4X* (Figures 2B,C). We designed three *RPS4X* promoter variations driving an *EmGFP* reporter (Figure 3B) to test first if they were sufficient to drive expression on an Xa, followed by analysis of their functionality on an Xi. Based on previously described design pipelines (see Methods), the first construct XEC1 includes the classical 5'UTR, first exon, and majority of the first intron (RR1) of *RPS4X* in order to capture potential enhancer elements in this area. In XEC1, the canonical *RPS4X* ATG was mutated to ATC, ensuring that translation would begin at the *EmGFP* reporter. As XEC1 retains the splice donor site at the beginning of *RPS4X* intron 1, but not the acceptor, it could splice from *RPS4X* intron 1 to the end of the synthetic intron before the *EmGFP* reporter, or leave *RPS4X* intron 1 intact as part of the 5'UTR, splicing only at the synthetic intron (Supplementary Figure 4).

The second construct, XEC2, placed RR1 (most of intron 1) upstream of the short promoter and exon 1, conserving the overall sequence of XEC1, but testing if the context of the sequences was important. As the *RPS4X* intronic piece was moved in front of the TSS, XEC2 will only splice at the synthetic intron before *EmGFP* (Supplementary Figure 4). Of note, this construct sub-divided the *RPS4X* promoter CpG island (moving 31/48 CpGs),

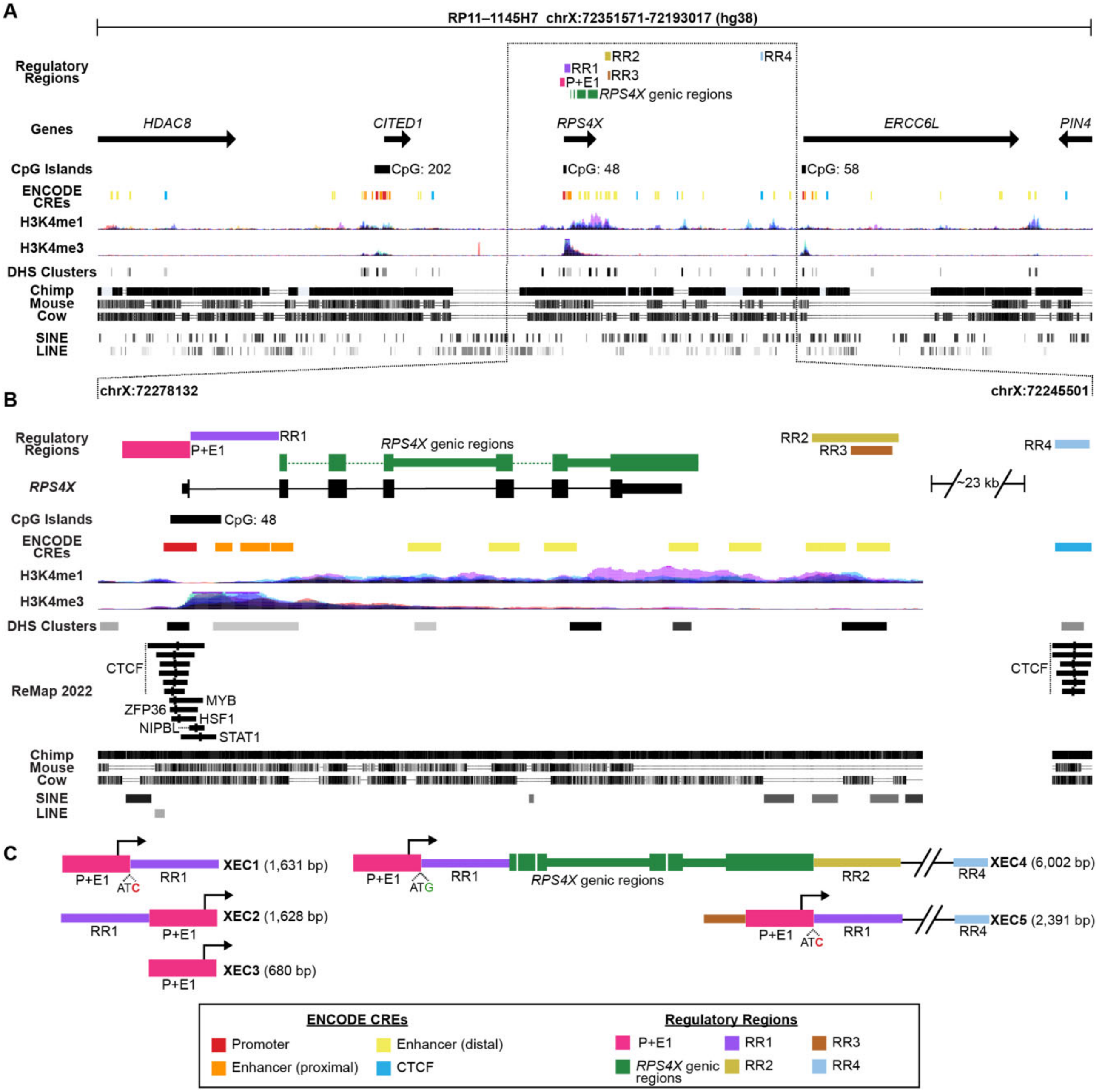

Figure 2. Bioinformatics design of five X-Escape Constructs (XECs) based on promoter (P) and regulatory regions (RRs) potentially regulating the expression of *RPS4X*. (**A**) Schematic overview of the BAC RP11–1145H7 (hg38:chrX:72193017–72351571, reversed), in which *RPS4X* and *CITED1* escape from X-chromosome inactivation (22). Regulatory regions, gene definitions (60), CpG islands (44), ENCODE *cis*-regulatory elements (CREs) (45), histone modifications H3K4me1 (an enhancer mark) and H3K4me3 (a promoter mark), chromatin accessibility state (DHS clusters; (61)), the pairwise conservation for three model organisms (i.e. chimp, mouse and cow), and repeat elements (SINEs and LINEs) in the region are highlighted (46). (**B**) Zoom-in overview of the regulatory regions included in the XECs, including tracks listed above as well as the peak coordinates of five EE-TFs, and CTCF in GM12878 cells from ReMap 2022 (37). (**C**) XEC designs 1–3 and 5 are promoter constructs driving expression of an *EmGFP* reporter. The ATG start codon in XEC1 and XEC5 has been mutated to ATC, and is not included in XEC2 and XEC3. XEC4 begins with the XEC1 sequence (ATG intact) and drives expression of the added *RPS4X* gene regions (minus introns 2, 3 and 5). Total base pairs (bp) of potential *RPS4X* regulatory sequence listed next to each construct, hg38 coordinates listed in Supplementary Table 4.

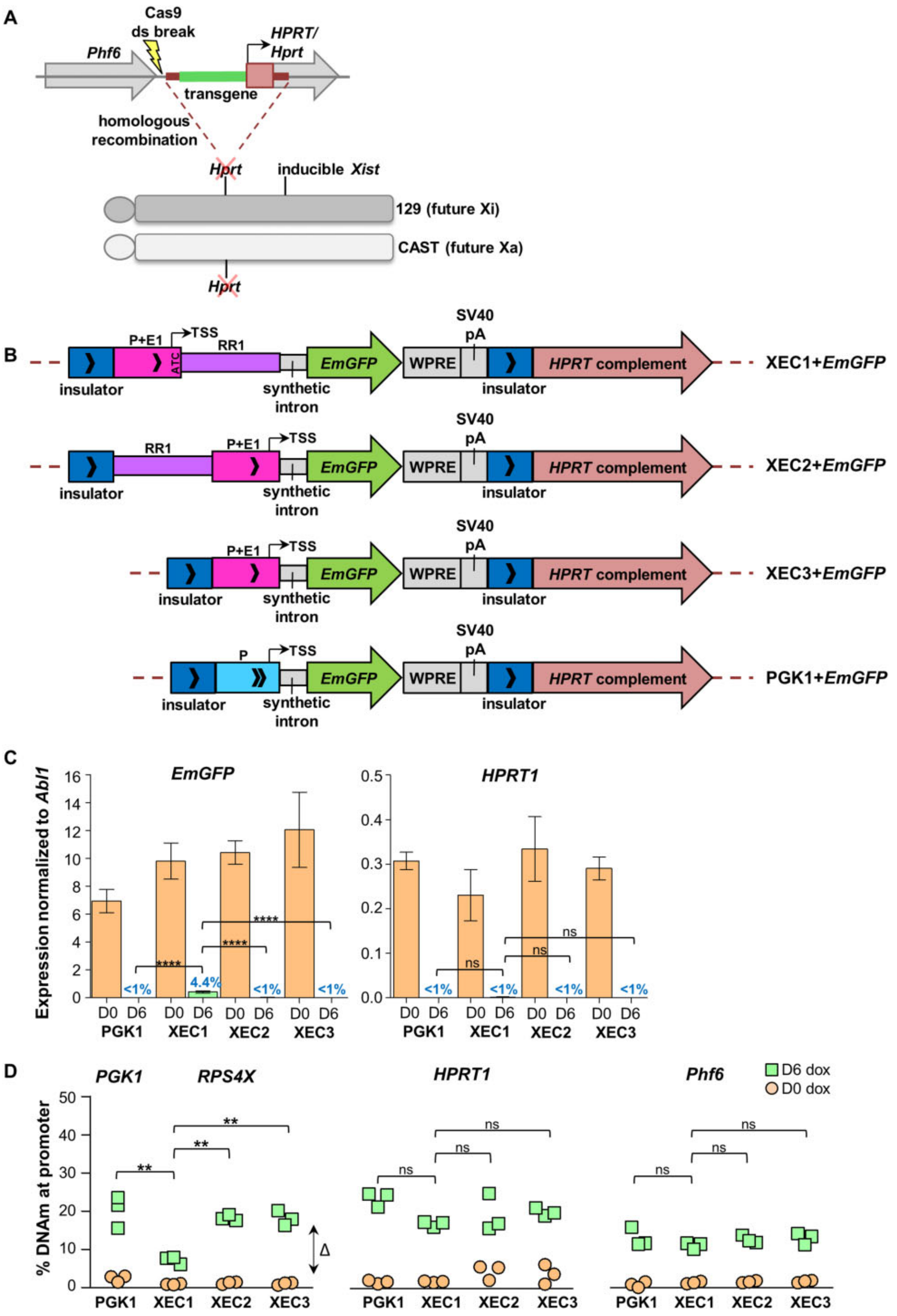

Figure 3. *RPS4X* promoter constructs demonstrate different potential for escape from XCI. (**A**) Schematic of homologous recombination to dock transgenes 5' of the X-linked *Hprt* gene on the 129-X, which contains an inducible *Xist*. Proper integration creates a chimaeric locus consisting of the human *HPRT1* promoter and exon 1, and mouse *Hprt* exons 2–9. (**B**) *RPS4X* promoter XECs 1–3, as well as subject gene promoter *PGK1* each drive an *EmGFP* reporter and are flanked by two insulator sequences in the *HPRT1*/*Hprt* correction plasmid. Only plasmid sequences between 5' and 3' *Hprt* homology arms are shown. Arrowheads within boxes denote directionality of CTCF motifs. (**C**) Normalized to endogenous control *Abl1*, RT-qPCR of *EmGFP* expression at D0 (Xa, orange bars) shows all three *RPS4X* promoter constructs drive similar expression of the reporter gene (averaged for three biological triplicates each). At D6 (Xi, green bars) of *Xist* induction and differentiation, only XEC1 shows minimal escape from XCI (D6/D0 expression shown in blue) and is significantly different from XEC2, XEC3 and *PGK1* subject control. *HPRT1* (human component of the fusion *HPRT1*/*Hprt* gene) is expressed from the Xi at less than 1% of the Xa consistent with being normally subject to inactivation. There is no significant difference in *HPRT1* percent escape between all four promoter constructs. (**D**) Average DNAm in promoter CpG islands of *PGK1* and *RPS4X* shows a gain in DNAm for all promoters from D0 to D6. The change in promoter DNAm at XEC1 is significantly lower than the other three constructs as it only gains about 6% (while XEC2, XEC3 and PGK1 gain upwards of 15%) supporting XEC1's unique expression from the Xi. *HPRT1* and mouse gene *Phf6* both show an increase in promoter DNAm from D0 to D6 in all promoter construct lines, as expected for genes that are being silenced on the Xi. All statistical comparisons are one-way ANOVA with Bonferroni's Multiple Comparison Test, significance denoted by asterisks; *P*-value <0.001 ***, 0.001 to 0.01 **, 0.01 to 0.05 *, >0.05 ns.

although all CpGs are retained in the sequence. The third promoter construct, XEC3, contained the smallest region posited to drive transcription and so was our most minimal promoter design. As it contains no intronic sequences from *RPS4X*, XEC3 will only splice at the synthetic intron before *EmGFP*. The promoter CpG island is truncated in this construct (removing 31/48 CpGs). We also included a control promoter from a gene that is normally subject to XCI, to validate that any expression observed from the Xi was due to the *RPS4X* sequences rather than the mouse integration site. We chose a commonly available *PGK1* promoter (47) known to drive expression, and cloned it into the same reporter gene construct.

RPS4X promoter constructs demonstrate different potential for escape from XCI

To screen our promoter constructs for escape from XCI at the same genomic location as our previous BAC integrations, we used a homologous recombination and complementation system to dock transgenes 5' of the X-linked *Hprt* gene. Using CRISPR-Cas9 technology, we modified an existing female 129 × Cast XX ESC model containing an inducible endogenous *Xist* on the 129-X (iXist, (50)) by mutating each *Hprt* allele in order to render the gene nonfunctional. Our XEC constructs were cloned into a plasmid containing 129-derived homology arms flanking the *Hprt* deletion on the 129-X, as well as a human *HPRT1* complementary sequence (48). The transgenes recombine just 5' of *Hprt*, creating a chimaeric locus consisting of the human *HPRT1* promoter and exon 1, and mouse *Hprt* exons 2–9. Recombination of our constructs at the *Hprt* locus reconstitutes HPRT activity, and thus correctly targeted clones can be selected with media containing hypoxanthine aminopterin thymidine (HAT). To increase recombination at *Hprt*, we co-transfected constructs with a plasmid expressing Cas9 and a gRNA targeting adjacent to the 5' homology arm (Figure 3A). Figure 3B shows all three *RPS4X* promoter escape constructs (XECs1-3), as well as control promoter PGK1, driving the *EmGFP* reporter, up until the *HPRT* complementary sequence. Of note, the promoter sequences are followed by a synthetic intron for splicing before the *EmGFP* reporter gene, as well as two insulators in tandem flanking the promoter and reporter. CTCF motifs are present in the same orientation in both insulators as well as in all four promoter constructs.

Three ESC clones were chosen for each construct based on screening for single-copy, 129-X integration of the escape construct, in cell lines with stable retention of both X chromosomes (Supplementary Table 5 for primer information). Interestingly, all three *RPS4X* promoters drove *EmGFP* expression to a similar extent when on the undifferentiated active X (Xa) in ESCs (Figure 3C, orange bars). *Xist* was then induced with doxycycline (dox) for 6 days under differentiating conditions to drive inactivation of the 129-X, and reporter expression was measured again, this time from the inactive X (Xi). We generally define escape from XCI as Xi expression being at least 10% of the expression from an Xa, and while not reaching that threshold to call escape, XEC1 D6/D0 expression was significantly different from the other *RPS4X* promoter constructs as well as the subject control promoter *PGK1* (Figure 3C, green bars). To confirm that this expression is not due to incomplete XCI, adjacent subject gene *HPRT1* with a human promoter was also analyzed. *HPRT1* has less than 1% expression from the Xi, as expected being normally subject to XCI in both mouse and humans, and has no significant difference between promoter constructs. Additionally, *Xist* upregulation as well as downregulation of pluripotency marker *Rex1* were examined at D6 to verify proper *Xist* induction and differentiation of cells (Supplementary Figure 3A). In comparison to non-induced differentiated controls, the inducible *Xist* system demonstrates a large increase in *Xist* expression, prompting us to question whether the overproduction of Xist RNA influenced the ability of any gene, including endogenous escapees, to escape from XCI. We examined the allelic ratios of mouse *Kdm6a*, a well-established and consistent escape gene (4), after 6 days of dox-induction with differentiating conditions. Despite increased presence of Xist in this system, mouse *Kdm6a* escapes from XCI at ∼22% demonstrating that a higher level of escape is achievable than what we saw with our XEC1 transgene (Supplementary Figure 3B).

For genes that have promoter CpG islands, measuring DNAm of the region is an indirect approach to examine XCI status, as genes that escape from XCI generally have less than 10% DNAm (7). To reinforce our expression studies, DNAm was analyzed in the promoter CpG islands for *RPS4X* or *PGK1*, as well as the human component of the fusion *HPRT1*/*Hprt* gene, and mouse gene *Phf6*, which is the closest endogenous mouse gene to the integration site (Figure 3D). We see a modest increase in *RPS4X* and *PGK1* DNAm from D0 to D6 except for XEC1. In agreement with the gene expression data, the change in DNAm at the XEC1 promoter is significantly different than the other three constructs as it only gains about 6%, remaining under 10% methylated at D6 (Supplementary Table 5). For the *HPRT1* and *Phf6*, genes, which are normally subject to inactivation, there is a gain in methylation to greater than 10% at D6 that is not significantly influenced by the promoter construct in each cell line.

To establish whether the Xi expression from XEC1 is a result of many cells with low expression, or few highly expressing cells, one transgenic line from each of PGK1 and XEC1 were FLOW sorted based on EmGFP. Results showed that the small amount of expression driven by the XEC1 promoter at D6 is due to a normal population of cells expressing a small amount, rather than a small population of outlier cells expressing at high levels (Supplementary Figure 3C). We further addressed whether the intron 1 region RR1 retained in XEC1 was spliced out (as it does in the full gene), or is retained as a longer 5'UTR with splicing occurring only at the synthetic intron as used in XEC2 and XEC3. RT-PCR results suggest that XEC1 uses multiple splice sites within RR1 as well as the synthetic intron at both D0 and D6 (Supplementary Figure 4), thus isoforms containing either an extended 5'UTR or the splicing pattern normally observed for *RPS4X* show escape from XCI. In summary, XEC1 had detectable Xi *EmGFP* expression and lower promoter DNAm than the *PGK1* control, and both XEC2 and XEC3; however, it failed to cross the standard threshold for escape. Despite all three *RPS4X* promoters driving similar

EmGFP expression in ESCs, and XEC2 having the same sequences as XEC1 but in a different order, XEC1 stands out with the most potential for escape, and so was chosen to further develop to better define the elements involved in escape from XCI.

RPS4X gene construct escapes from XCI

Starting with the promoter region of XEC1, we included the remainder of the *RPS4X* gene sequence minus the less-conserved introns 2, 3 and 5 to create a mini gene escape construct termed XEC4 (Figure 2C, and Figure 4A). Given that Alu elements have been previously identified as enriched near genes that escape from XCI, an additional region containing two primate-specific Alu elements with potential enhancer activity was added to the end of the gene region (Figure 2, RR2). As all CTCF motifs in the XEC1-3 promoter constructs had been in the same orientation, we decided to also include a small region from the original BAC construct containing a potential boundary element between *RPS4X* and subject gene *ERCC6L* (Figure 2, RR4). This places a divergent CTCF after the Alu elements, but before the second insulator to potentially interact with the 5'CTCFs either in the 5'insulator or the *RPS4X* promoter region.

As with our promoter constructs, XEC4 was cloned into a plasmid containing 129 homology arms, and a complementary sequence that rescues *HPRT1* activity, allowing correctly targeted clones to be selected for with HAT media. Three ESC clones were chosen as biological replicates based on screening for single-copy, 129-X integration of the escape construct, in cell lines with stable retention of both X chromosomes (Supplementary Table 5 for primer information). *RPS4X* gene expression was tested in ESCs at D0 and *Xist* was then dox-induced for 6 days under differentiating conditions to force inactivation of the 129-X, and *RPS4X* gene expression was measured again from the Xi.

Impressively, the *RPS4X* gene construct XEC4 robustly escaped from XCI with an Xi/Xa expression ratio of about 26% after dox-induction and differentiation (Figure 4B). *HPRT1* remained subject to XCI, as expected, with <1% expression from the Xi. *Xist* upregulation as well as downregulation of pluripotency marker *Rex1* were also verified at D6 (Supplementary Figure 5). The Xi expression is corroborated by hypomethylation of the *RPS4X* promoter in all three clones with no significant gain in DNAm from D0 to D6. At *HPRT1* there was only a slight gain in DNAm, not as high as seen with XECs 1–3 (Figure 4C). Previously we have noted that while *HPRT1* remains subject to XCI, lower DNAm could be a consequence of open chromatin from an escape gene in close proximity to the *HPRT1* promoter island (24). *Phf6* also gained DNAm, to a similar extent as what has been seen previously.

To test whether it was the *RPS4X* gene itself, or the putative boundary and enhancer elements that contributed to escape from XCI, we modified our original *PGK1* and *RPS4X* XEC1 promoters to include variations of these elements (Figure 2, XEC5), and examined whether or not *EmGFP* reporter activity could be detected from the Xi (Supplementary Figure 6A). Additional potential enhancer and boundary elements do not increase XEC1 *EmGFP* escape from XCI nor do they affect silencing of the *PGK1* reporter (Supplementary Figure 6B). Both promoters gained DNAm in the same manner as the original constructs reflecting the consistency of the previous data (Supplementary Figure 6C). Overall we refined the sequence sufficient for *RPS4X* to escape from XCI from a nearly 160 kb BAC to a minimal region of ∼6 kb.

DISCUSSION

Up to a quarter of X-linked genes evade complete silencing from the Xi and are still expressed in some cells or individuals, despite the stable epigenetic silencing of the majority of the X across the lifespan of the individual. To understand the DNA elements that allow these escapees to avoid silencing, we have combined bioinformatic studies with a mouse ESC-based model to test the ability of human transgenes to escape XCI. Analysis of ChIP-seq TF binding data highlighted a set of five TFs with enriched binding observed in the vicinity of escape gene TSSs. Multiple TFs may be involved in regulating escape, as no single TF exhibited binding across all of the diverse escape genes. Comparison of binding at escape, subject and autosomal TSSss indicate that subject genes have lower observed TF binding, while escape genes are similar to autosomal gene properties. All five of the EE-TFs were observed to bind to the *RPS4X* TSS, motivating continued focus on how the compact gene can escape silencing when introduced transgenically to a distinct location on the X chromosome. Design of an *RPS4X* mini gene was able to reproduce escape with only 6 kb of endogenous DNA sequence, providing the most compact DNA sequence sufficient for mediating escape.

Access to the curated ChIP-seq data from ReMap allowed us to bypass reliance on motif enrichment, instead providing empirical data of TF binding at TSSs. Our enrichment analysis of TF peaks along the X chromosome highlighted five TFs as being enriched at escape gene TSSs relative to subject genes, thus they may be contributors to escape gene regulation. None of the EE-TFs overlap all annotated escape genes, reinforcing current hypotheses that there are multiple possible pathways to escape from XCI. For our enrichment analysis we chose to focus on ChIP-seq data from GM12878 cells as it is a well-studied human female cell line; however, there are limitations that come with this choice. The 154 TFs for which there is data in ReMap for GM12878 still encompass less than 10% of the TFs in the human genome. As this is a female cell line, the ChIP data will be derived from binding on both the active and inactive X. When a gene escapes inactivation it will have two instances of binding vs a subject gene which will have only one (on the Xa), and thus stringent parameters in some data sets might bias towards those TFs that bind escape genes. However, since expression from the Xi is generally less than from the Xa, we considered it worth using a female (Xa + Xi) line rather than a male (Xa only) line which might miss Xi-specific binding. Furthermore, ChIP-seq data for some available TFs, including YY1, was sparse, resulting in an inability to assess YY1 enrichment, despite the motif having been highlighted in previous studies of escape (27). Additionally, as GM12878 is a somatic cell line, TFs that may only bind early in development to establish escape,

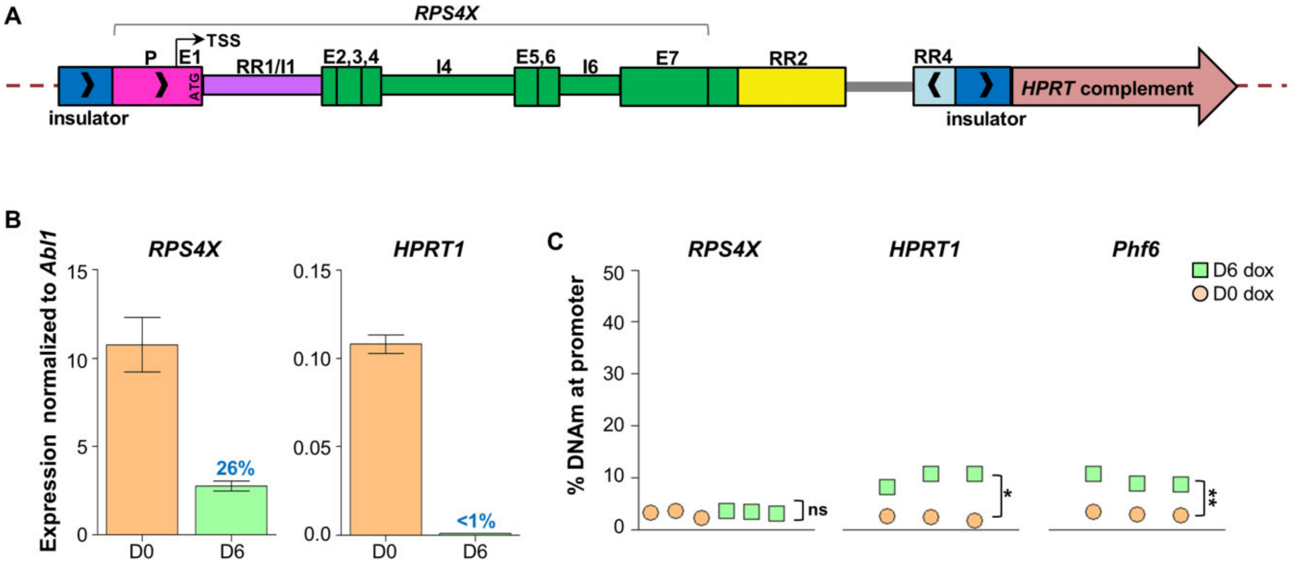

Figure 4. *RPS4X* gene construct escapes from XCI at *Hprt*. (**A**) XEC4 design begins with the sequence of XEC1 (now with intact ATG), and includes the remainder of the *RPS4X* gene sequence minus introns 2, 3 and 5. Additional regulatory regions with potential enhancer (RR2) and boundary (RR4) activity are included 3' of the gene. Arrowheads within boxes denote directionality of CTCF motifs. (**B**) Normalized to endogenous control *Abl1*, RT-qPCR of *RPS4X* expression at D0 (Xa) and D6 (Xi) of *Xist* induction and differentiation shows robust escape from XCI of ∼26% averaged for three biological triplicates. *HPRT1* (human component of the fusion *HPRT1*/*Hprt* gene) is expressed from the Xi at less than 1% of the Xa consistent with it being normally subject to inactivation. **(C)** Average DNAm of *RPS4X* shows promoter hypomethylation at D6, which supports its expression from the Xi. *HPRT1* and *Phf6* show a significant increase in promoter DNAm from D0 to D6 as expected for genes that are being silenced on the Xi (paired t-test, significance denoted by asterisks; *P*-value < 0.001 ***, 0.001 to 0.01 **, 0.01 to 0.05 *, >0.05 ns).

rather than maintain, would also have been missed. Lymphocytes have been shown to have weaker maintenance of XCI, and GM12878 is a lymphoblastoid cell line, which has been shown to require distinct XIST-interacting proteins from other cell types examined (54,55). While the mouse provides an opportunity to examine XCI dynamics early in development, enrichment studies are underpowered by the limited number of constitutive escape genes in mouse, as well as less comprehensive ChIP data sets. With the growth in genomics data and supporting database resources, it will be important to continue to analyze new data on the X to fill in missing important information regarding TFs that could be involved in the regulation of escape from XCI.

Overall TF binding was seen to be dichotomous with some genes binding many TFs while other genes were bound by few or none. This dichotomy was seen for both the EE-TFs and also for all TFs across the genome. However, the pattern on autosomes was significantly different from that of the X, with reduced TF binding for the X chromosome, although the escape genes retained more similarity in TF binding distribution to the autosomal pattern. There was some correlation between expression and TF binding; however, expression-matched autosomal genes revealed this same disparity with X-subject genes generally having fewer TF bindings (Supplemental Figure S2). This may reflect that as genes became responsive to XCI they tended to lose TF binding, and that the EE-TFs have more consistently lost their binding to genes subject to XCI, but are retained on the Xi. It is likely that the EE-TFs bind both the Xa and Xi, but future studies will be required to assess the allelic binding specificity. In general, there is a large overlap of TFs at regulatory regions, making dissection by deletion of a single motif challenging.

Of particular interest is ZFP36, also known as Tristetraprolin (TTP), an RNA-binding protein involved in RNA degradation (56). The term TF is at times perceived to be restricted to DNA binding proteins, but formally is inclusive of all proteins involved in transcription and thus the presence of an RNA binding protein in the ReMap resource is fully appropriate. As XCI is initiated by the lncRNA XIST, the binding of ZFP36 could potentially direct the local degradation of XIST or other lncRNAs, thereby allowing a region to remain more accessible to the transcription machinery and escape from XCI. Both enrichment and co-binding analysis highlighted that ZFP36 has a strong presence at escape regions; ZFP36 was one of the five EE-TFs that passed the stringent thresholds, while in the co-binding analysis, pairs that involve ZFP36 were observed more often than pairs for the other four EE-TFs. It is noteworthy that previous motif enrichment analyses could not consider the presence of RNA binding proteins, and thus the potential role for ZFP36 was enabled by the focus on experimental binding data. *ZFP36* is a member of the conserved ZFP36 family of RNA binding proteins, which have been shown to have direct and indirect roles in transcription, RNA stability and translation, particularly in T cells (57). *Zfp36* knockout mice are viable, with no reported sex biases. However, as mice have fewer escape genes, impact on escapee expression might not result in a distinctive phenotype, given that 39,X mice have less of a phenotype than 45,X females (58).

While the role for these EE-TFs remains speculative, all were seen enriched at the *RPS4X* TSS, lending further support to use of the *RPS4X* gene as a model to experimentally test and characterize the DNA elements involved in escape from XCI. Starting with several promoter-based constructs, we identified a minimal promoter region (XEC3) of

680 bp sufficient for expression on the Xa. Extending this promoter to include regulatory elements in intron 1 (RR1, see Figure 2) demonstrated surprising context-dependent effects. XEC1 (with RR1 following the minimal promoter) had low promoter DNAm and slight expression from the Xi, while XEC2 (with RR1 preceding the minimal promoter) had increased DNAm and no expression. This could reflect that transcription of RR1 is required for it to function as a component of escape gene regulation. Alternatively, a critical site may have been broken in XEC2. In this regard it is notable that the five EE-TFs do not have well-established recognition motifs. Furthermore, in XEC2 the CpG island is split; however, having an intact CpG island is not a requirement for all genes that escape from XCI. There is an ongoing correlation between low promoter DNAm and gene expression, but it is important to note that DNAm thresholds for calling escape from XCI have a wide 'uncallable' zone between hypo- and hyper-methylated islands (7), that could be dependent on individual genes, tissues or developmental time point being examined. Indeed overall DNAm accumulation across all promoters in this study was not as high as seen in adult tissues (22), with the differentiated ESC model more closely resembling DNAm at embryonic day 9.5 (23), and so using DNAm to call escape in early developmental time points is more challenging as the gain in DNAm is unlikely to be complete for most genes. Promoter DNAm at *HPRT1* specifically appears to be mildly influenced by presence of an escape gene in its immediate 5'region as it gained less DNAm in XEC4 (remaining under 10% methylated) than the other constructs with less Xi expression. This has been previously documented with another escape transgene at *Hprt* (24), yet in both cases *HPRT1* expression remains subject to XCI.

While the XEC1 promoter seemed primed for Xi expression, increasing the size of our escape construct to encompass more of the *RPS4X* gene and surrounding elements (XEC4), including a CTCF-binding boundary element (RR4) consistently gave us robust escape from XCI. CTCF has shown enrichment at promoters (28) and enhancers (5) of escape genes, and has also been suggested to serve as a boundary element between subject and escape promoters. The original chicken hypersensitive site-4 (cHS4) insulators contained in our homology plasmid contain CTCF motifs and have been shown to have protective effects against transgene silencing on an active X chromosome; however, they were not able to block XCI or prevent DNAm on the inactive X on their own (24,59). A second set of promoter constructs (XEC5 and *PGK1*) testing the putative boundary (RR4) and enhancer elements (RR2/3) failed to increase escape from the Xi, again suggesting that CTCF insulator regions are insufficient to enable escape from XCI. However, the differing distances between CTCF motifs within our XECs could have impacted their ability to establish chromatin loops. We further observed a small impact on *HPRT1* promoter DNAm with escape from XCI of XEC4. Thus, we consider it is unlikely that the RR4 element is functioning as a boundary between these genes. Furthermore, the failure of our putative regulatory elements to augment escape suggests that the additional elements enabling the ability to escape lie within the *RPS4X* gene itself, consistent with previous hypotheses of proximal regulatory elements (21). Our downshell analysis identified only one TF, SKIL, as being consistently enriched at the 10 kb downstream of the TSS (Supplementary Figure 1) which was not observed at *RPS4X*. For *RPS4X* the 10 kb downshell would include the whole gene; however, for other genes intragenic enhancer elements could be missed if they lay further away. Previous studies have used different window sizes for their analysis of enriched motifs, and thus are difficult to compare to our analysis using ReMap data.

The variation in EE-TF binding between escape genes, and indeed global TF binding between the X and autosomes, highlights some of the considerable differences amongst X-linked promoters and emphasizes the idea that there will also be considerable differences in mechanisms of escape between genes. This work with the *RPS4X* gene has demonstrated that gene-proximal elements are sufficient to permit its escape from XCI, but there is likely an additive effect from multiple elements that need to reside in a specific position in order to function. The inclusion and placement of RR1 (most of *RPS4X* intron1) in XEC1 appears to be responsible for minimal escape, but an additional factor contained in the rest of the *RPS4X* sequence in XEC4 is needed to boost expression to meet established thresholds for escape. The context-dependence of elements and their interactions hints at ultrastructure effects, which could be mediated by DNA or RNA. Future studies will be informative in investigating whether an escape-specific enhancer element(s) was added in the sequence after the XEC1 promoter, or if transcription through the rest of the gene, or potential secondary structure formation, is responsible for resisting the effects of silencing in the region. Despite lack of conservation in number and distribution of escapees between species, experiments such as these demonstrate the utility of a transgenic human-in-mouse model and have added several important considerations as to what elements promote a gene to escape from XCI. To the best of our knowledge, we have synthesized the smallest transgenic escape construct to date and described a functional model system for further characterization of regulatory elements and testing of their applicability across other genes.

DATA AVAILABILITY

The data underlying this article are available in ReMap 2022, at https://remap2022.univ-amu.fr/. No new datasets were generated and data for the transgene studies are provided in text or supplemental data.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank past and present members of the Brown and Wasserman labs, with specific appreciation to Bianca Kirsh for early work on TF binding assessment, as well as Dr Elizabeth M. Simpson and her lab members for helpful discussion. This work was supported by the Canadian Institutes of Health Research (Grant PJT-16120). O.F. and W.W.W. were supported by grants from the Natural Sciences and

Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2017–06824), and the BC Children's Hospital Foundation and Research Institute. We also thank Dr Neil Brockdorff and lab for sharing the inducible Xist mESC line, as well as the LSI Flow Cytometry core facility at UBC.

FUNDING

Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2017-06824]; Canadian Institutes of Health Research [PJT-16120].
Conflict of Interest statement. None declared.

REFERENCES

1. Carrel,L. and Willard,H.F. (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, **434**, 400–404.
2. Yang,F., Babak,T., Shendure,J. and Disteche,C.M. (2010) Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.*, **20**, 614–622.
3. Calabrese,J.M., Sun,W., Song,L., Mugford,J.W., Williams,L., Yee,D., Starmer,J., Mieczkowski,P., Crawford,G.E. and Magnuson,T. (2012) Site-specific silencing of regulatory elements as a mechanism of X inactivation. *Cell*, **151**, 951–963.
4. Berletch,J.B., Ma,W., Yang,F., Shendure,J., Noble,W.S., Disteche,C.M. and Deng,X. (2015) Escape from X inactivation varies in mouse tissues. *PLoS Genet.*, **11**, e1005079.
5. Barros de Andrade e Sousa,L., Jonkers,I., Syx,L., Dunkel,I., Chaumeil,J., Picard,C., Foret,B., Chen,C.-J., Lis,J.T., Heard,E. *et al.* (2019) Kinetics of *Xist* -induced gene silencing can be predicted from combinations of epigenetic and genomic features. *Genome Res.*, **29**, 1087–1099.
6. Tukiainen,T., Villani,A.-C., Yen,A., Rivas,M.A., Marshall,J.L., Satija,R., Aguirre,M., Gauthier,L., Fleharty,M., Kirby,A. *et al.* (2017) Landscape of X chromosome inactivation across human tissues. *Nature*, **550**, 244–248.
7. Cotton,A.M., Lam,L., Affleck,J.G., Wilson,I.M., Peñaherrera,M.S., McFadden,D.E., Kobor,M.S., Lam,W.L., Robinson,W.P. and Brown,C.J. (2011) Chromosome-wide DNA methylation analysis predicts human tissue-specific X inactivation. *Hum. Genet.*, **130**, 187–201.
8. Cotton,A.M., Chen,C.-Y., Lam,L.L., Wasserman,W.W., Kobor,M.S. and Brown,C.J. (2014) Spread of X-chromosome inactivation into autosomal sequences: role for DNA elements, chromatin features and chromosomal domains. *Hum. Mol. Genet.*, **23**, 1211–1223.
9. Balaton,B.P. and Brown,C.J. (2021) Contribution of genetic and epigenetic changes to escape from X-chromosome inactivation. *Epigenetics Chromatin*, **14**, 30.
10. Balaton,B.P., Cotton,A.M. and Brown,C.J. (2015) Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol. Sex Differ.*, **6**, 35.
11. Bellott,D.W., Hughes,J.F., Skaletsky,H., Brown,L.G., Pyntikova,T., Cho,T.-J., Koutseva,N., Zaghlul,S., Graves,T., Rock,S. *et al.* (2014) Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature*, **508**, 494–499.
12. Ciccodicola,A., D'Esposito,M., Esposito,T., Gianfrancesco,F., Migliaccio,C., Miano,M.G., Matarazzo,M.R., Vacca,M., Franzè,A., Cuccurese,M. *et al.* (2000) Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum. Mol. Genet.*, **9**, 395–401.
13. De Bonis,M.L., Cerase,A., Matarazzo,M.R., Ferraro,M., Strazzullo,M., Hansen,R.S., Chiurazzi,P., Neri,G. and D'Esposito,M. (2006) Maintenance of X- and Y-inactivation of the pseudoautosomal (PAR2) gene SPRY3 is independent from DNA methylation and associated to multiple layers of epigenetic modifications. *Hum. Mol. Genet.*, **15**, 1123–1132.
14. Navarro-Cobos,M.J., Balaton,B.P. and Brown,C.J. (2020) Genes that escape from X-chromosome inactivation: potential contributors to Klinefelter syndrome. *Am. J. Med. Genet. C Semin. Med. Genet.*, **184**, 226–238.
15. Dunford,A., Weinstock,D.M., Savova,V., Schumacher,S.E., Cleary,J.P., Yoda,A., Sullivan,T.J., Hess,J.M., Gimelbrant,A.A., Beroukhim,R. *et al.* (2017) Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. *Nat. Genet.*, **49**, 10–16.
16. Carrel,L. and Brown,C.J. (2017) When the Lyon(ized chromosome) roars: ongoing expression from an inactive X chromosome. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **372**, 20160355.
17. Balaton,B.P., Fornes,O., Wasserman,W.W. and Brown,C.J. (2021) Cross-species examination of X-chromosome inactivation highlights domains of escape from silencing. *Epigenetics Chromatin*, **14**, 12.
18. Balaton,B.P. and Brown,C.J. (2016) Escape artists of the X chromosome. *Trends Genet.*, **32**, 348–359.
19. Wilson Sayres,M.A. and Makova,K.D. (2013) Gene survival and death on the human Y chromosome. *Mol. Biol. Evol.*, **30**, 781–787.
20. Li,N. and Carrel,L. (2008) Escape from X chromosome inactivation is an intrinsic property of the *Jarid1c* locus. *Proc. Natl. Acad. Sci*, **105**, 17055–17060.
21. Mugford,J.W., Starmer,J., Williams,R.L., Calabrese,J.M., Mieczkowski,P., Yee,D. and Magnuson,T. (2014) Evidence for local regulatory control of escape from imprinted X chromosome inactivation. *Genetics*, **197**, 715–723.
22. Peeters,S.B., Korecki,A.J., Simpson,E.M. and Brown,C.J. (2018) Human cis-acting elements regulating escape from X-chromosome inactivation function in mouse. *Hum. Mol. Genet.*, **27**, 1252–1262.
23. Peeters,S.B., Korecki,A.J., Baldry,S.E.L., Yang,C., Tosefsky,K., Balaton,B.P., Simpson,E.M. and Brown,C.J. (2019) How do genes that escape from X-chromosome inactivation contribute to Turner syndrome?*Am. J. Med. Genet. C Semin. Med. Genet.*, **181**, 28–35.
24. Yang,C., McLeod,A.J., Cotton,A.M., de Leeuw,C.N., Laprise,S., Banks,K.G., Simpson,E.M. and Brown,C.J. (2012) Targeting of <1.5 Mb of human DNA into the mouse X chromosome reveals presence of *cis*-acting regulators of epigenetic silencing. *Genetics*, **192**, 1281–1293.
25. Jegalian,K. and Page,D.C. (1998) A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature*, **394**, 776–780.
26. Andrés,O., Kellermann,T., López-Giráldez,F., Rozas,J., Domingo-Roura,X. and Bosch,M. (2008) RPS4Y gene family evolution in primates. *BMC Evol. Biol.*, **8**, 142.
27. Chen,C., Shi,W., Balaton,B.P., Matthews,A.M., Li,Y., Arenillas,D.J., Mathelier,A., Itoh,M., Kawaji,H., Lassmann,T. *et al.* (2016) YY1 binding association with sex-biased transcription revealed through X-linked transcript levels and allelic binding analyses. *Sci. Rep.*, **6**, 37324.
28. Loda,A., Brandsma,J.H., Vassilev,I., Servant,N., Loos,F., Amirnasr,A., Splinter,E., Barillot,E., Poot,R.A., Heard,E. *et al.* (2017) Genetic and epigenetic features direct differential efficiency of Xist-mediated silencing at X-chromosomal and autosomal locations. *Nat. Commun.*, **8**, 690.
29. Filippova,G.N., Cheng,M.K., Moore,J.M., Truong,J.-P., Hu,Y.J., Nguyen,D.K., Tsuchiya,K.D. and Disteche,C.M. (2005) Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev. Cell*, **8**, 31–42.
30. Horvath,L.M., Li,N. and Carrel,L. (2013) Deletion of an X-inactivation boundary disrupts adjacent gene silencing. *PLos Genet.*, **9**, e1003952.
31. Wang,Z., Willard,H.F., Mukherjee,S. and Furey,T.S. (2006) Evidence of influence of genomic DNA sequence on human X chromosome inactivation. *PLoS Comput. Biol.*, **2**, e113.
32. Polak,P. and Domany,E. (2006) Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics [Electronic Resource]*, **7**, 133.
33. Bourque,G., Leong,B., Vega,V.B., Chen,X., Lee,Y.L., Srinivasan,K.G., Chew,J.-L., Ruan,Y., Wei,C.-L., Ng,H.H. *et al.* (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.*, **18**, 1752–1762.
34. Schmidt,D., Schwalie,P.C., Wilson,M.D., Ballester,B., Gonçalves,Â., Kutter,C., Brown,G.D., Marshall,A., Flicek,P. and Odom,D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335–348.
35. Bailey,J.A., Carrel,L., Chakravarti,A. and Eichler,E.E. (2000) Molecular evidence for a relationship between LINE-1 elements and

X chromosome inactivation: the Lyon repeat hypothesis. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 6634–6639.

36. Boeva,V. (2016) Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Front. Genet.*, **7**, 24.
37. Hammal,F., de Langen,P., Bergon,A., Lopez,F. and Ballester,B. (2022) ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.*, **50**, D316–D325.
38. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M. *et al.* (2021) The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.
39. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.
40. Layer,R.M., Pedersen,B.S., DiSera,T., Marth,G.T., Gertz,J. and Quinlan,A.R. (2018) GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods*, **15**, 123–126.
41. Dice,L.R. (1945) Measures of the amount of ecologic association between species. *Ecology*, **26**, 297–302.
42. Simpson,E.M., Korecki,A.J., Fornes,O., McGill,T.J., Cueva-Vargas,J.L., Agostinone,J., Farkas,R.A., Hickmott,J.W., Lam,S.L., Mathelier,A. *et al.* (2019) New MiniPromoter Ple345 (NEFL) drives strong and specific expression in retinal ganglion cells of mouse and primate retina. *Hum. Gene Ther*, **30**, 257–272.
43. Korecki,A.J., Cueva-Vargas,J.L., Fornes,O., Agostinone,J., Farkas,R.A., Hickmott,J.W., Lam,S.L., Mathelier,A., Zhou,M., Wasserman,W.W. *et al.* (2021) Human MiniPromoters for ocular-rAAV expression in ON bipolar, cone, corneal, endothelial, Müller glial, and PAX6 cells. *Gene Ther.*, **28**, 351–372.
44. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
45. Project Consortium,E.N.C.O.D.E., Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
46. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F.A., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
47. Fowler,D.K., Stewart,S., Seredick,S., Eisen,J.S., Stankunas,K. and Washbourne,P. (2016) A MultiSite gateway toolkit for rapid cloning of vertebrate expression constructs with diverse research applications. *PLoS One*, **11**, e0159277.
48. Korecki,A.J., Hickmott,J.W., Lam,S.L., Dreolini,L., Mathelier,A., Baker,O., Kuehne,C., Bonaguro,R.J., Smith,J., Tan,C.-V. *et al.* (2019) Twenty-seven tamoxifen-inducible iCre-driver mouse strains for eye and brain, including seventeen carrying a new inducible-first constitutive-ready allele. *Genetics*, **211**, 1155–1177.
49. Bronson,S.K., Plaehn,E.G., Kluckman,K.D., Hagaman,J.R., Maeda,N. and Smithies,O. (1996) Single-copy transgenic mice with chosen-site integration. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 9067–9072.
50. Nesterova,T.B., Wei,G., Coker,H., Pintacuda,G., Bowness,J.S., Zhang,T., Almeida,M., Bloechl,B., Moindrot,B., Carter,E.J. *et al.* (2019) Systematic allelic analysis defines the interplay of key pathways in X chromosome inactivation. *Nat. Commun.*, **10**, 3129.
51. Yang,G.S., Banks,K.G., Bonaguro,R.J., Wilson,G., Dreolini,L., de Leeuw,C.N., Liu,L., Swanson,D.J., Goldowitz,D., Holt,R.A. *et al.* (2009) Next generation tools for high-throughput promoter and expression analysis employing single-copy knock-ins at the Hprt1 locus. *Genomics*, **93**, 196–204.
52. Heigwer,F., Kerr,G. and Boutros,M. (2014) E-CRISP: fast CRISPR target site identification. *Nat. Methods*, **11**, 122–123.
53. Ran,F.A., Hsu,P.D., Wright,J., Agarwala,V., Scott,D.A. and Zhang,F. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.*, **8**, 2281–2308.
54. Wang,J., Syrett,C.M., Kramer,M.C., Basu,A., Atchison,M.L. and Anguera,M.C. (2016) Unusual maintenance of X chromosome inactivation predisposes female lymphocytes for increased expression from the inactive X. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E2029–E2038.
55. Yu,B., Qi,Y., Li,R., Shi,Q., Satpathy,A.T. and Chang,H.Y. (2021) B cell-specific XIST complex enforces X-inactivation and restrains atypical B cells. *Cell*, **184**, 1790–1803.
56. Brooks,S.A. and Blackshear,P.J. (2013) Tristetraprolin (TTP): interactions with mRNA and proteins, and current thoughts on mechanisms of action. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.*, **1829**, 666–679.
57. Matheson,L.S., Petkau,G., Sáenz-Narciso,B., D'Angeli,V., McHugh,J., Newman,R., Munford,H., West,J., Chakraborty,K., Roberts,J. *et al.* (2022) Multiomics analysis couples mRNA turnover and translational control of glutamine metabolism to the differentiation of the activated CD4+ T cell. *Sci. Rep.*, **12**, 19657.
58. Berletch,J.B., Yang,F. and Disteche,C.M. (2010) Escape from X inactivation in mice and humans. *Genome Biol.*, **11**, 213.
59. Ciavatta,D., Kalantry,S., Magnuson,T. and Smithies,O. (2006) A DNA insulator prevents repression of a targeted X-linked transgene but not its random or imprinted X inactivation. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 9958–9963.
60. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
61. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.