

# SCIENTIFIC REPORTS



OPEN

## YY1 binding association with sex-biased transcription revealed through X-linked transcript levels and allelic binding analyses

Received: 04 April 2016  
Accepted: 24 October 2016  
Published: 18 November 2016

Chih-yu Chen<sup>1,2</sup>, Wenqiang Shi<sup>1,2</sup>, Bradley P. Balaton<sup>3</sup>, Allison M. Matthews<sup>1</sup>, Yifeng Li<sup>1</sup>, David J. Arenillas<sup>1</sup>, Anthony Mathelier<sup>1</sup>, Masayoshi Itoh<sup>4,5,6</sup>, Hideya Kawaji<sup>4,5,6</sup>, Timo Lassmann<sup>4,5</sup>, Yoshihide Hayashizaki<sup>4,6</sup>, Piero Carninci<sup>4,5</sup>, Alistair R. R. Forrest<sup>4,5,7</sup>, Carolyn J. Brown<sup>3</sup> & Wyeth W. Wasserman<sup>1,3</sup>

Sex differences in susceptibility and progression have been reported in numerous diseases. Female cells have two copies of the X chromosome with X-chromosome inactivation imparting mono-allelic gene silencing for dosage compensation. However, a subset of genes, named escapees, escape silencing and are transcribed bi-allelically resulting in sexual dimorphism. Here we conducted *in silico* analyses of the sexes using human datasets to gain perspectives into such regulation. We identified transcription start sites of escapees (escTSSs) based on higher transcription levels in female cells using FANTOM5 CAGE data. Significant over-representations of YY1 transcription factor binding motif and ChIP-seq peaks around escTSSs highlighted its positive association with escapees. Furthermore, YY1 occupancy is significantly biased towards the inactive X (Xi) at long non-coding RNA loci that are frequent contacts of Xi-specific superloops. Our study suggests a role for YY1 in transcriptional activity on Xi in general through sequence-specific binding, and its involvement at superloop anchors.

Sex disparities in disease progression and susceptibility for many diseases, including cancer<sup>1</sup>, autism<sup>2</sup>, cardiac<sup>3</sup> and autoimmune disorders<sup>4</sup>, have long been known. Such discrepancies likely result from a combination of the sex chromosomes, sex hormones, and environmental factors. Due to such sexual dimorphism, there has been a recent push in policy to balance sex in cell and animal studies by NIH<sup>5</sup>. The key genetic differences between the sexes are the sex chromosomes, with mammalian females being XX and males XY. Furthermore, there is an enrichment of brain-related genes on the X chromosome in mammals<sup>6</sup>. With the exception of the pseudoautosomal regions (PAR) shared with the Y-chromosome, X-linked genes are present in two copies in females and only one in males. X-chromosome inactivation (XCI) silences one copy of the X chromosome (chrX) in female cells in order to compensate for dosage between the sexes. Up-regulation of *XIST*, a long non-coding RNA (lncRNA), is known to be responsible for initiation of XCI, and this process has been reported to be mediated through recruitment of factors such as polycomb repressive complex 2, leading to tri-methylation of histone H3 at lysine 27 (H3K27me3) and silencing of the inactive X chromosome (Xi)<sup>7</sup>. XCI involves the establishment of a peripheral nuclear architecture, which includes association with the lncRNA, *FIRRE*<sup>8</sup>, anchoring the Xi near the nucleolus to preserve H3K27me3 and the silencing state<sup>9</sup>. DNA methylation (DNAm) is recruited to promoters, providing maintenance of the inactive state. As a result, the majority of chrX genes outside of PAR regions are subject to XCI, and are transcribed mono-allelically from the active X (Xa) in female somatic cells. A small subset of chrX

<sup>1</sup>Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, University of British Columbia, Vancouver, British Columbia, Canada. <sup>2</sup>Graduate Program in Bioinformatics, University of British Columbia, Vancouver, British Columbia, Canada. <sup>3</sup>Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. <sup>4</sup>RIKEN Omics Science Center, Yokohama, Japan. <sup>5</sup>RIKEN Center for Life Science Technologies, Division of Genomic Technologies, Yokohama, Japan. <sup>6</sup>RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Saitama, Japan. <sup>7</sup>Harry Perkins Institute of Medical Research, QEII Medical Centre and Centre for Medical Research, the University of Western Australia, Nedlands, Western Australia, Australia. Correspondence and requests for materials should be addressed to C.J.B. (email: carolyn.brown@ubc.ca) or W.W.W. (email: wyeth@cmmt.ubc.ca)

genes including *XIST*, known as escapees, escape from XCI and are transcribed on the Xi<sup>10</sup>. These escapees are therefore bi-allelically transcribed with the exception of *XIST*, which is solely transcribed from the Xi. Binding of the Ying-Yang 1 (YY1) transcription factor (TF) to *XIST* RNA and DNA contributes to *XIST* transcription<sup>11–13</sup>. *XIST* and *FIRRE* are among the four long non-coding RNAs (lncRNAs) previously found at frequently interacting regions of Xi-specific superloops in the GM12878 cell line<sup>14</sup>. Rao *et al.* reported tandem CTCF motifs at three (*FIRRE*, *DXZ4*, *LOC550643*) of the four lncRNAs, and suggested a role for chromatin looping through CTCF and RAD21 in shaping the chromatin structure of Xi. At the *FIRRE* locus, Yang *et al.* showed differential occupancy by CTCF and YY1 using ENCODE ChIP-seq peaks from male and female cells<sup>9</sup>.

Investigation of differences between the sexes has been tackled both through direct comparisons of male and female data and/or a focused study of XCI. Direct sex comparisons have identified male:female differences in gene expression, DNAm, accessible chromatin, and TF binding levels. At the gene expression level, microarray and RNA-seq platforms have been used to examine differential gene expression between the sexes. The Genotype-Tissue Expression (GTEx) pilot study reported genes with differential expression between male and female samples in 43 tissues using RNA-seq data<sup>15</sup>. Sex differences in brain expression have been observed repeatedly<sup>16,17</sup>. Hall *et al.* integrated RNA-seq, DNAm, and microRNA datasets in islet cells for a sex comparison to identify genes associated with insulin secretion, revealing the potential molecular mechanism for phenotypic sex differences in this tissue<sup>18</sup>. At the DNAm level, we previously focused on chrX to compare between sexes in 27 tissues using Illumina 450K arrays, from which we identified escapees and subject genes across cell types<sup>19</sup>. The assay of transposase accessible chromatin with sequencing (ATAC-seq) maps chromatin accessibility in a given cell population. A sex comparison on chrX by Qu *et al.* using ATAC-seq data in T cells revealed the highest female to male ratio of chromatin accessibility at *XIST* and *FIRRE*, followed by escapees<sup>20</sup>.

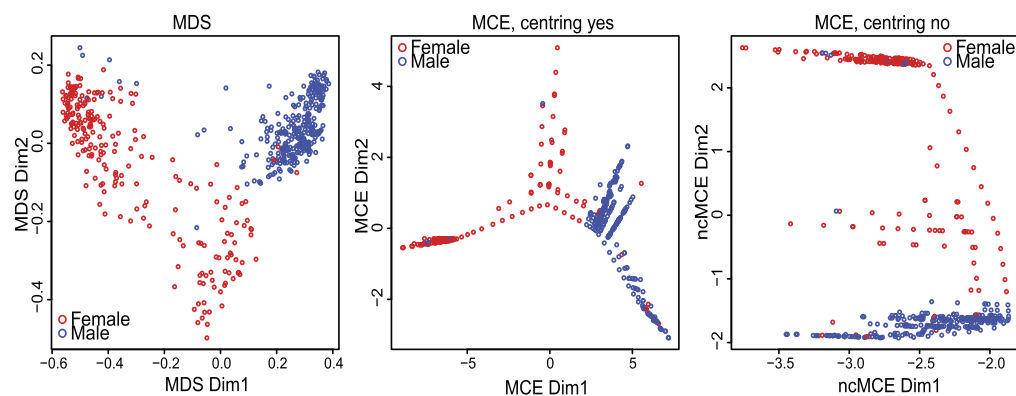
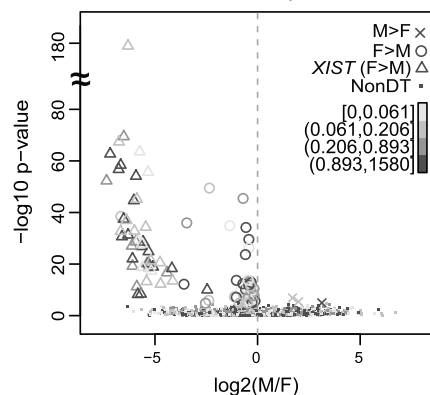
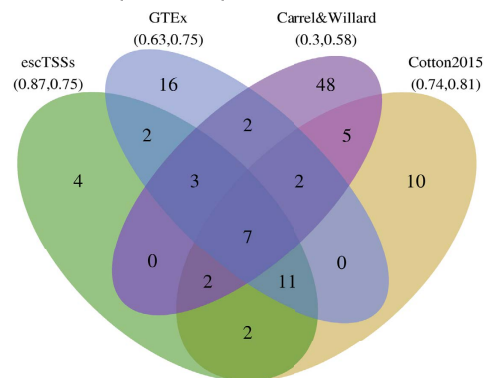
Allelic expression or allelic binding approaches are powerful in delineating activities from either Xi or Xa, although they require cells with XCI of one chrX favored over the other (skewed XCI) as well as the genotype information of the cells. These studies in general are limited by the availability of known heterozygous sites within the particular cells for allelic comparison. Experiments using F1 mouse cells containing large numbers of known heterozygous sites have therefore been favored. In human lymphoblastoid cell lines, allelic expression and histone modifications, measured using Illumina BeadChip genotyping arrays, showed the majority of genes to be subject to XCI, and a correspondence between expression and histone modifications<sup>21</sup>. Allelic RNA-seq and ChIP-seq for RNA polymerase II and various TFs have previously been conducted on the lymphoblastoid GM12878 cell line known to have skewed XCI<sup>22–24</sup>. Reddy *et al.* reported that TFs are predominately bound on the Xa in GM12878, and allelic binding of RNA polymerase II is non-differential between Xa and Xi at escapees<sup>24</sup>.

The recent generation of high throughput datasets offers new opportunities to examine X-linked gene expression. The cap analysis of gene expression (CAGE) datasets, which measure abundance of the 5' end of transcribed RNAs, generated by the FANTOM5 consortium have facilitated the identification of transcription start sites (TSSs) and their levels of transcription in over 800 samples of non-treated human cells<sup>25</sup>. The strength of CAGE datasets compared to RNA-seq lies in the capability of pinpointing TSSs of gene promoters as well as enhancers<sup>25,26</sup>. For the ease of terminology, we refer to the reported robust CAGE peaks as TSSs in the rest of the text. In addition, The Cancer Genome Atlas (TCGA) offers a rich source of DNAm datasets of both sexes in specific cancer types from the epigenetic perspective<sup>27</sup>, while over 680 TF ChIP-seq datasets generated by the ENCODE project provide information on TF binding events in multiple cell types<sup>28,29</sup>.

Here we report a comprehensive investigation of X-chromosome gene activity between the sexes, focused on escapees through direct sex comparisons of expression, DNAm, and TF binding levels, further supported with evidence of allelic binding on chrX in the female GM12878 lymphoblastoid cell line. While XCI has previously been reported to occur at a domain level<sup>30,31</sup>, we hypothesize that there exists a common regulatory mechanism that facilitates escape from XCI. To our knowledge, this is the first study conducting large-scale analyses on TF ChIP-seq data comparing between sexes and between Xi and Xa occupancies to identify potential regulator(s) associated with escape from XCI.

## Results

**Sex classification of the FANTOM5 samples using CAGE data.** To understand how well the transcript levels at TSSs can distinguish sexes and to confirm reliable labeling of male and female samples, we built a sex classifier through a Random Forest approach<sup>32</sup> using the FANTOM5 CAGE dataset<sup>25</sup>. It is important to note that the datasets contained a mixture of samples from primary cell types, tissues and cell lines, which may have distinct properties. While biological replicates create stronger data consistency, the inconsistent availability of technical replicates creates over-emphasis on those samples with replicates. Therefore only the first replicate from each source was retained. We trained on 530 samples of known sex using the transcript levels of 5,071 TSSs in the non-pseudoautosomal region of the X (X-non-PAR) as features. Through performance assessment, the classifier performed better than a classifier based on TSSs assigned to *XIST* alone. Namely, the out of bag error rate and balanced accuracy were 6.51% and 0.908 for the X-non-PAR classifier and 11.55% and 0.825 for the *XIST* classifier. Hence, we used results from the X-non-PAR classifier in our following analyses. Linear and nonlinear dimension reduction of the sample proximity values from the Random Forest classifier consistently indicated the presence of outlying samples, as shown with multi-dimensional scaling and Minimum Curvilinear embedding<sup>33</sup> in Fig. 1A. The *XIST* TSSs of these outlying samples were differently expressed from their original sex labels (Supplementary Table S1). Of the 26 samples called female but classified male, there were 2 stem cells and 14 cancer lines, in which the *XIST* expression were zero to minimal. The latter group is consistent with reports of certain cancer cells losing the Xi<sup>34</sup>. Of the 10 samples labeled male but classified female, there were 3 testes-related tumor cell lines with strong *XIST* expression in agreement with a previous report of testicular germ cell tumors expressing *XIST*<sup>35</sup>. The remaining samples could reflect sex chromosome aneuploidy or sample mislabeling. For the latter we noted at least one case where the labeled sex differed between technical replicates from the same

**A** Classification of Sexes**B** Differential Transcription**C** Escapee comparison**Figure 1. Classification and differential expression analysis of sexes using FANTOM5 CAGE datasets.**

(A) Plots showing distances between FANTOM5 CAGE samples in the first 2 dimensions generated using multi-dimensional scaling, Minimum Curvilinear embedding, and non-centred Minimum Curvilinear embedding methods from the proximity matrix of the Random Forest sex classifier. Each circle represents a FANTOM5 sample with its labeled sex: male (blue) or female (red). (B) Scatter plot with  $\log_2$  ratio of the mean expression levels comparing male to female (with a constant of the 5<sup>th</sup> percentile expression added to avoid denominators of 0) on the x-axis, and  $-\log_{10}$  transformation of raw p-values from the differential transcription analysis between sexes on the y-axis. Each point represents a TSS in non-PAR region of chrX, and TSSs with significantly higher expression in female (escTSSs) and male cells are denoted with circles and crosses (Bonferroni-corrected p-value  $\leq 0.05$ ), with small dots for non-differentially transcribed TSSs. The escTSSs nearest to the *XIST* gene are denoted using open triangles. The grey-scale gradient represents the average expression across all samples in quartiles. The vertical dashed line represents a  $\log_2$  ratio of 0, where there is no difference between sexes. (C) Venn diagram depicting the overlapping sets of escapees from three published studies with those identified in this report. The numbers within the Venn diagram represent the overlaps between sets, and the numbers in bracket under each list name are precision and recall values where genes reported in more than one list are taken to constitute true escapees.

source. Overall, the results suggested that we can classify sexes accurately using transcript levels of TSSs in X-non-PAR regions. Thus we constructed our sex classifier using X-non-PAR TSSs training on all labeled samples, and predicted a total of 309 female and 426 male samples to be used in the subsequent section with the exclusion of outlier and mixed samples (Supplementary Table S1).

**Differentially transcribed chrX TSSs with higher expression in female reflect escapees.**

Given that there were unbalanced numbers of samples between the sexes for multiple cell types (Supplementary Table S2), we incorporated cell categories from ontologies to group multiple closely related cell types to minimize bias from cell-type specific transcription. Using the sex labels from our classifier, we assessed the differential transcription between sexes using a linear regression model for each TSS, incorporating the cell categories as covariates (see Methods for details). Parallel analysis was conducted on both the X-non-PAR region and autosomes (Supplementary Table S3). Here we focused on chrX, with information from the autosome analysis presented in Supplementary File S1. TSSs that overlap repetitive elements were excluded from the analysis. Of the 94 TSSs on chrX that were significantly differentially transcribed between sexes with Bonferroni-corrected

p-values < 0.05, two and one TSSs assigned to the *ARHGAP4* and *SH2D1A* genes, respectively, had higher expression in male cells, while 91 TSSs (referred to as 'escTSSs') corresponded to 31 unique genes with higher expression in female cells (Fig. 1B). Forty-five of the 91 escTSSs were associated to the *XIST* gene, and exhibited much stronger differences between sexes than others (Fig. 1B).

Given the higher expression in female cells, we expected the majority of our predicted escTSSs to escape from XCI (Supplementary Fig. S1A). Consistent with previously reported escape of an alternative TSS for *UBA1* gene<sup>36</sup>, we identified escTSSs overlapping the alternative TSS while other *UBA1* TSSs were not significantly different between male and female. CAGE advantageously provides precise information about individual TSSs, while other methods are constrained to gene-level resolution. We compared the associated genes of our escTSSs to previously identified escapees from the literature, in which different data and techniques were used: RNA-seq ('GTEX')<sup>37</sup>, rodent/human somatic cell hybrids ('Carrel&Willard')<sup>38</sup>, and Illumina 450k DNAm array ('Cotton2015')<sup>19</sup>. For our comparison, the escapees identified by at least two approaches were assumed to be true escapees (these are inclusive of reported brain-specific escapees<sup>17</sup>). Our analysis was geared to identify TSSs that broadly escape across cell types. As escape genes can express at different levels from Xi between cell types and samples<sup>19,38–40</sup>, the comparison included only genes that broadly escape among samples and tissues, where such information was provided (detailed in Methods). Twenty-seven of the 31 genes from our list were captured by at least one publication, and four genes (five escTSSs in total) appeared to be novel (Fig. 1C). There were 9 genes that were not detected in our set and appeared in two or more published escapee sets, of which 2 occurred in all 3 published sets (*GYG2* and *STS*). Our approach using CAGE data obtained the highest positive predictive value for escapees (precision = 0.87), and ranked second after 'Cotton2015' for the proportion of true escapees predicted (recall = 0.75).

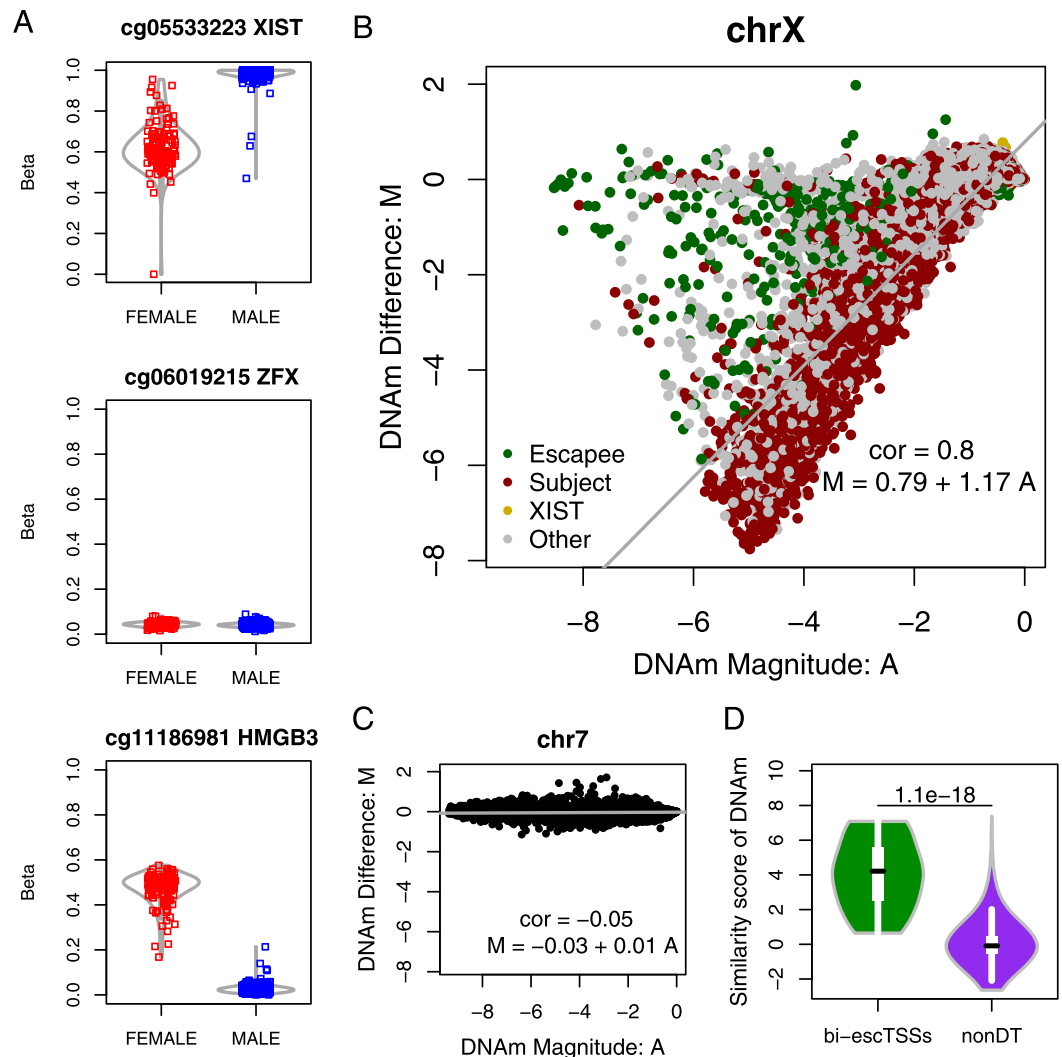
We further examined the five apparently novel escTSSs we predicted. Notably, all five escTSSs were on the opposite strand of known escapees identified in two or more studies (Intron 1 of *DDX3X*, Exon 6 of *XIST*, Intron 1 of *HDHD1* and Exon 5 of *MXRA5*). Three novel escTSSs belonged to two non-coding RNAs: *TSIX* and *MIR4767*. The other two novel escTSSs were distant from the nearest annotated genes (11 kb and 44 kb away from *RN7SL15P* and *ASS1P4*, respectively), and therefore it is unclear to which genes they are best ascribed. We successfully validated the escTSS which was antisense to *DDX3X* and had the highest significance in differential transcripts between sexes (Supplementary Fig. S2). *DDX3X* antisense transcripts can be observed from previously published GRO-seq data in IMR90, K562 and GM12878 female cell lines<sup>41,42</sup>, and interestingly, our validation with mouse-human hybrid cells indicated that the antisense transcripts were Xi-specific (Supplementary Fig. S2). While the two *MIR4767*-labelled escTSSs had high GC content making them difficult to validate, recently updated RefSeq annotation shows that *MIR4767* overlaps the 5' end of two *STS* transcripts. Therefore, it may be more accurate to assign the two escTSSs as alternative promoters of *STS* (a known escapee noted above as one of the two undetected genes). Summarizing, CAGE-based detection of escTSSs had the highest precision compared to the published sets.

### DNA methylation similarity between sexes on chrX showed strong agreement with bi-allelically transcribed escapees.

We have previously used DNAm data to identify genes that were subject to or escaping from XCI<sup>19</sup>. Here we used an independent public DNAm dataset from TCGA<sup>27</sup> to examine whether the differential transcription we identified was reflected at DNAm level between the sexes. On autosomes, we expected differential DNAm between the sexes to correspond to differential transcription. On chrX, however, the overall methylation level was reported to be higher in female than in male due to the nature of the Xi<sup>19</sup>, but we expected similar DNAm levels between sexes at bi-allelically transcribed escTSSs (Supplementary Fig. S1B). For ease of reference, we refer to bi-allelically transcribed escTSSs as 'bi-escTSSs' for the rest of the text. DNAm of probes near TSSs of *XIST*, *ZFX* (a bi-allelically transcribed escapee), and *UPF3B* (a subject gene) are shown as examples in Fig. 2A using data in female and male urothelial bladder cancer samples from TCGA.

To show the differences in DNAm between sexes on chrX and autosomes, we generated microarray-inspired MA plots to compare the relationship between log<sub>2</sub> ratios of DNAm between male and female samples (i.e. the difference: M) versus log<sub>2</sub> average of DNAm between the sexes (i.e. the magnitude: A) for probes in chrX and chr7 (Fig. 2B,C; see methods for details). We discovered a drastically stronger linear correlation between difference and magnitude on chrX ( $\rho = 0.8$ ) but not on autosomes (chr7:  $\rho = -0.05$ ). On autosomes, probes that deviate from the zero difference values were differentially methylated between sexes; whereas on chrX, probes that deviate from the fitted regression line, with the log beta ratio closer to 0, had similar DNAm levels between sexes. Hence, we identified the chrX probes with similar DNAm levels between sexes by computing the residual of the regression model as the similarity score, which also gives greater weights to probes with lower DNAm levels. Indeed, the DNAm similarity scores from probes within 50 base pairs (bps) of TSSs were found to be significantly higher for bi-escTSSs than non-differentially transcribed TSSs (nonDT; p-value of Wilcoxon test:  $1.1 \times 10^{-18}$ ; Fig. 2D). We observed similar results using three other cancer datasets (Supplementary Fig. S3). Taken together, the results showed that higher DNAm similarity between sexes on chrX reflected TSSs of escapees, and further supported the bi-allelic activities of the bi-escTSSs we predicted.

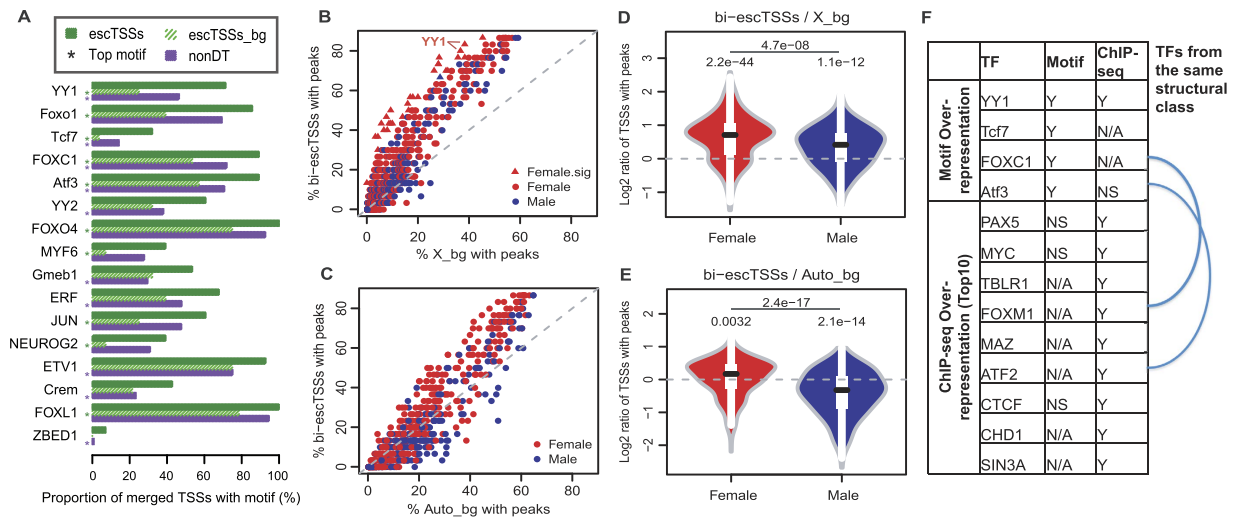
**YY1 binding motif over-representation around escTSSs.** Given that escTSSs (both bi-escTSSs as well as *XIST*) are transcribed on the Xi, we began testing our hypothesis that common regulator(s) and mechanism exist to facilitate the escape from XCI. We first probed the regulation of the escTSSs on chrX at the sequence level through enrichment testing of JASPAR motifs<sup>43</sup> using the CAGED-oPOSSUM web tool<sup>44</sup>. Merged sequences from 500 bps up- and down-stream of escTSSs were compared to two background sets: nonDT and a randomly selected %GC and length matched background set (escTSSs\_bg). The TF binding motifs associated with YY1, Tcf7, FOXC1 and Atf3 were found to be over-represented around escTSSs when compared to both background sets (Fig. 3A and Supplementary Table S4). The significance of YY1 binding motif was highest among all motifs, and it was 2.9 and 1.5 times more frequently found in escTSSs than escTSSs\_bg and nonDT sets, respectively. The



**Figure 2. DNAm comparison between sexes on chrX in urothelial bladder cancer (BLCA) samples from TCGA.** (A) DNA methylation status for positions (i.e. probes from the Illumina 450k array) near TSSs in both sexes from BLCA samples, where the  $\beta$  values (Y-axis) range from 0 (unmethylated) to 1 (fully methylated). The three TSSs are most proximal to the following genes (from top to bottom): *XIST*, an escapee (*ZFX*) and a subject gene (*HMGB3*). Each square represents a sample for the BLCA dataset. Red or blue color represents a female or male sample, respectively. Each violin plot in gray lines shows the distribution of beta values for each sex at each probe. Plots (B,C) show MA plots for chrX probes and autosomal probes on chr7 between sexes, respectively. Each dot represents a probe from the array. M (difference) on y-axis is the logged differential methylation value between sexes, and A (magnitude) on x-axis is the logged average methylation value (as indicated in Methods). The fitted robust regression line is represented in gray, with the corresponding function and correlation reported. Green and red colors in plot (B) represent probes nearest to escapees and subject genes previously reported in Cotton *et al.* 2015. Gold and gray colors represent probes nearest to *XIST* and genes not in either three categories. (D) Violin plots showing the distributions of DNA methylation similarity scores between sexes for probes within 50 bps of escTSSs and non-differentially transcribed (nonDT) TSSs on chrX. The similarity score of DNA methylation on y-axis is the residual of M as a function of A on chrX. Only TSSs with at least one probe within 50 bps were plotted, and for those TSSs within 50 bps of multiple probes, the average similarity scores of probes were obtained. The p-value from the Wilcoxon test is reported above the violin plots.

motif over-representation analyses indicated YY1 as a potential regulator of escapees through sequence specific binding.

**Over-representation of TF ChIP-seq peaks around bi-allelically transcribed escTSSs.** To explore the regulation of escapees at the experimental level, we examined TF ChIP-seq data in male and female cell lines from the ENCODE project to identify over-represented TF binding at the bi-escTSSs. TF binding on chrX in female cells can be either bi- or mono-allelic, and the measured binding degree would reflect a mixture of both X chromosomes. Therefore, given a positive-regulating TF, we expect a stronger binding pattern at bi-escTSSs than

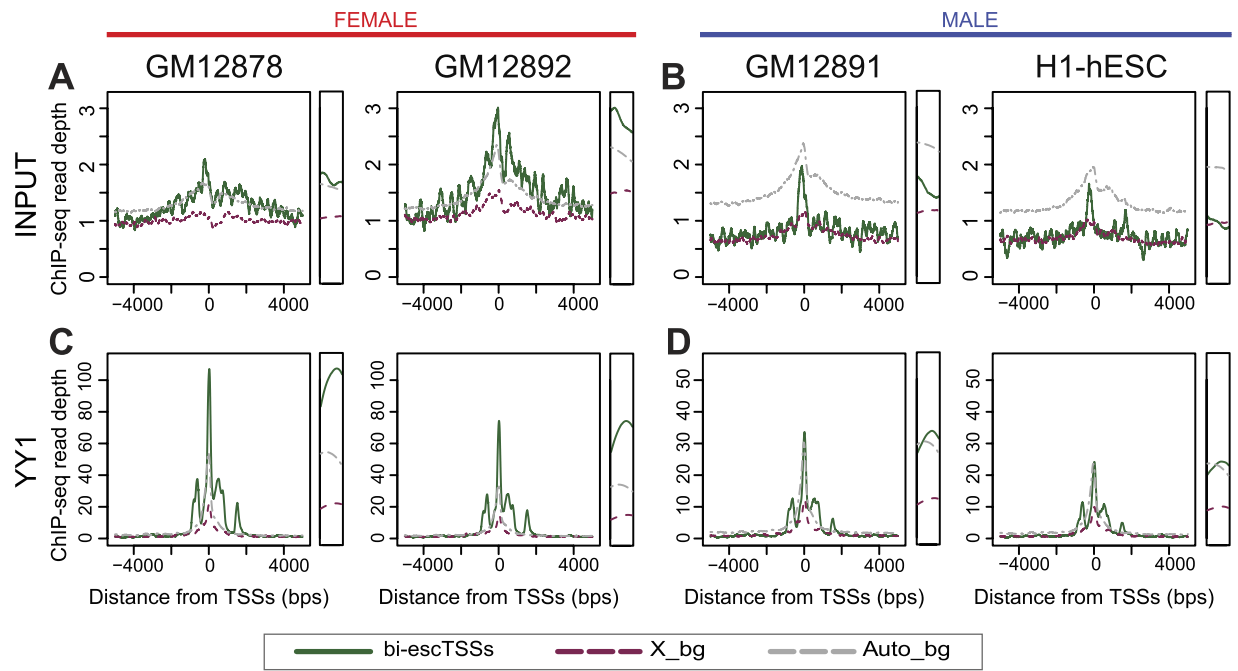


**Figure 3. Over-representation analyses of TF binding motifs and ENCODE TF ChIP-seq peaks at escape TSSs.** (A) Barplots showing the proportions of merged TSS regions (x-axis) containing the JASPAR TF motifs labeled on the y-axis for escTSSs (green) and non-differentially transcribed set on X with Bonferroni-corrected p-values equal to 1 (nonDT; purple). The proportions for %GC composition and length matched background set sampled from the genome for escTSSs (escTSSs\_bg) is shown in shaded green. The top motifs with Fisher scores greater than 95<sup>th</sup> percentile in escTSSs compared to either background set (escTSSs\_bg and nonDT) are plotted. The top motifs derived with each background set are marked with asterisks under the background bars with corresponding color. Motifs are presented in decreasing order of Fisher score sums from both comparisons. Figures (B,D) compare bi-escTSSs to matched background TSSs on chrX (X\_bg) for over-representation of TF ChIP-seq peaks, whereas figures (C,E) compare to matched background TSSs on autosomes (Auto\_bg). (B,C): Scatter plots showing the percentages of TSSs that overlap a peak comparing between bi-escTSSs (y-axis) and matched background TSSs (x-axis). Each dot represents a uniformly processed TF ChIP-seq dataset from ENCODE. Red and blue colors represent female and male cells in all plots, respectively. The dashed gray lines are the baselines reflecting no differences between proportions of escTSSs and background TSSs overlapping peaks. Datasets with significant over-representation of peaks in escTSSs compared to background TSSs are displayed as triangles (Bonferroni-corrected p-values  $\leq 0.05$ ). Significantly over-represented YY1 datasets are labeled on figure (B). (D,E) Violin plots showing the distributions of log<sub>2</sub> ratio of escTSSs to background TSSs in female and male cells. The p-value from comparing the log<sub>2</sub> ratios between male and female cells (one-sided Wilcoxon test) and the p-values of one-sample Wilcoxon tests for the distributions are shown. (F) Figure listing TF motifs over-represented compared to both background sets, and the top 10 over-represented TFs in ranking of significance. The semi-circle lines link TFs within the same structural classes. 'N/A' indicates that data is unavailable. 'Y' and 'NS' indicate significant or not significant over-representation, respectively.

mono-allelically active TSSs on chrX in female cells (Supplementary Fig. S1C). We refer to this expectation as the bi-allelic effect. As there is only one copy of the chrX in the male cells, enrichment is not expected.

We tested the over-representation of all 689 uniformly processed ENCODE TF ChIP-seq peak sets within 500 bps of bi-escTSSs compared to background TSSs on chrX (X\_bg) and autosomes (Auto\_bg) with matched average expression (see Methods for details). Thirty individual ChIP-seq datasets, corresponding to 18 unique TFs, were over-represented around bi-escTSSs when compared to X\_bg (One-sided Fisher's exact test, Bonferroni-corrected p-values  $\leq 0.05$ ; Supplementary Table S5 and Fig. 3B). The top 10 unique over-represented TFs in the order of decreasing significance were PAX5, MYC, TBLR1, FOXM1, YY1, MAZ, ATF2, CTCF, CHD1 and SIN3A. Notably, the significant TF datasets were all generated from female cells except for YY1 data in the male HepG2 cell line. By contrast, when background was changed to the autosomes (Auto\_bg), none of the TF ChIP-seq peaks were found to be significantly over-represented (Supplementary Table S5 and Fig. 3C).

Both motif and ChIP-seq analyses revealed YY1 to be over-represented around escTSSs. For a subset of the top escTSSs ( $\pm 500$  bp regions), visual inspection affirms a strong consistency between predicted YY1 motifs and ChIP-seq peaks (Supplementary Fig. S4). We sought to determine if there were additional cases of TFs, which were consistent between both analyses (Fig. 3F). TFs within the same structural class often have highly similar binding properties. For two cases, a motif was enriched for a TF (FOXC1 and Atf3), and ChIP-seq was enriched for another member of the same structural class (FOXM1 and ATF2, respectively). Four of the top 10 TFs with ChIP-seq over-representation lacked binding models. ChIP-seq peak but not motif over-representation of PAX5, c-Myc and CTCF might reflect a presence of these TFs via secondary mechanisms (not sequence-specific DNA binding). As an independent validation, we examined two published X;autosomal translocation studies<sup>30,45</sup>, which reported in aggregate 1106 and 473 autosomal genes that escaped from or were subject to the spread of heterochromatin from Xi in trisomic cells, respectively. When comparing autosomal escapees to autosomal subject



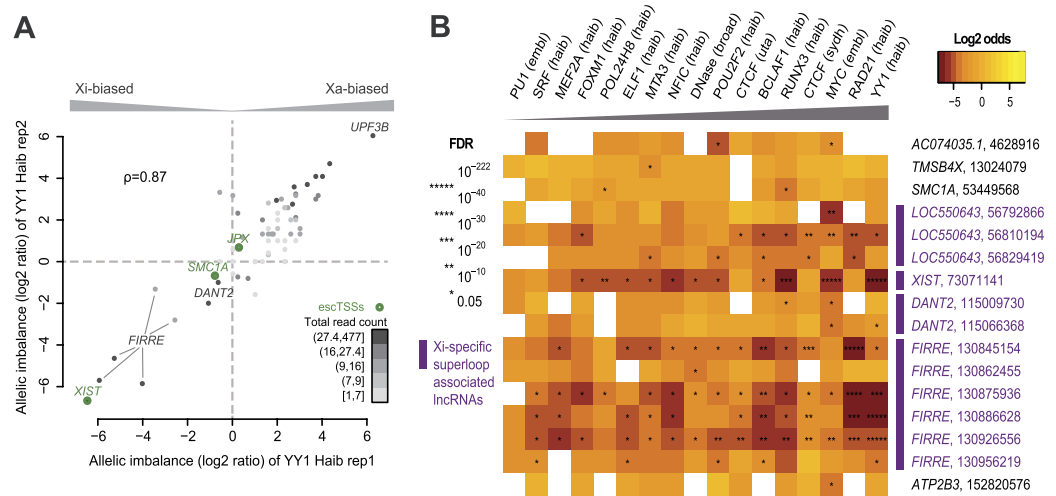
**Figure 4. Input and YY1 ChIP-seq read depths around bi-escTSSs for both sexes.** The read depth plots for ENCODE ChIP-seq input samples in two female (A) and male cell lines (B) within 5 kb of three TSS sets: a subset of bi-escTSSs in green with a filter of unique escTSS per gene symbol, background TSSs on chrX with matched averaged expression (X\_bg) in violet dashed line, and autosomal TSSs with matched averaged expression (Auto\_bg) in gray dashed line. The read depth plots for ENCODE YY1 ChIP-seq data in the same female (C) and male cells (D). The scales on y-axis of figures A and B are the same, while the scale of C is two times the scale of D to reflect the expected X-copies in female and male cells. For the ease of visualization, the narrow panel on the right of each plot displays the read depth within 50 bps of the TSSs.

genes, we found YY1 and MYC (ranked 6<sup>th</sup> and 8<sup>th</sup>) to be among the top 10 unique TFs with ChIP-seq peaks over-represented around escapee TSSs (Supplementary Table S6). Taken together, the significance of YY1 binding in multiple cellular contexts from the chrX and X;A data further highlighted the strong potential for a role for YY1 in facilitating the escape from XCI.

While the 30 ChIP-seq datasets have peaks significantly over-represented at escTSSs compared to X\_bg, many additional sets (250 of 689) were reported non-significant after the conservative Bonferroni correction was applied. We observed an overall trend of higher proportions of escTSSs containing peaks compared to that of X\_bg (Fig. 3B). Indeed, the distribution of log<sub>2</sub> ratios of percent overlaps with TF peaks (escTSSs/X\_bg) in both female and male cells were significantly greater than 0 (one-sample Wilcoxon test  $p = 2.2 \times 10^{-44}$  and  $1.1 \times 10^{-12}$ , respectively; Fig. 3D). In agreement with our expectation, the log<sub>2</sub> ratios in female cells were significantly higher than that observed in male cells (one-sided Wilcoxon test  $p = 4.7 \times 10^{-08}$ ). The log<sub>2</sub> ratios comparing escTSSs to autosomal background (Auto\_bg) was less significantly different from 0 in female cells ( $p = 0.0032$ ; Fig. 3E), whereas it was significantly lower in male cells ( $p = 2.1 \times 10^{-14}$ ), due to the single copy of chrX in male cells. The overall higher number of binding events at escTSSs stressed the importance of taking multiple TF datasets into consideration when comparing between sexes at escTSSs or between escTSSs and other chrX TSSs, otherwise the differential occupancy may merely reflect the bi-allelic effect.

**ChIP-seq read depth reveals overall reduction of input from heterochromatin Xi.** As TF binding peaks do not necessarily reflect linearly the magnitude of TF occupancy, we further compared the read depth at escTSSs to the same background sets using input and YY1 ChIP-seq data from two female and male cells. Although female cells have two copies of autosomes and chrX, the read depth for input within  $\pm 50$  bps of X\_bg was 0.68 fold lower than that of Auto\_bg on average, while the read depth within  $\pm 50$  bps of escTSSs was closer to Auto\_bg (1.18 fold; Fig. 4A). This indicated a discounted number of input fragments available for capture due to the compactness of Xi around X\_bg in female cells. While male cells have two copies of each autosome and one of chrX, the input read depths at chrX (X\_bg and escTSSs) and autosomes (Auto\_bg) reflected the ratio as expected (0.50 and 0.58 fold, respectively; Fig. 4B). Overall, the input reads in female cells reflect bi- and mono-allelic activities at escTSSs and X\_bg, respectively, and input reads at X\_bg in female cells, while somewhat higher than those in male cells, reflect a reduction of the heterochromatin input.

**YY1 binding at escTSSs shown by read depth and allelic analyses.** Interestingly, despite slightly lower input levels, YY1 ChIP-seq read depth at escTSSs was 2.01 times that of Auto\_bg in female cells (Fig. 4C). Furthermore, despite the 1 to 2 ratio of input levels at escTSSs and Auto\_bg in male cells, the YY1 read depths at



**Figure 5. Allelic imbalance at heterozygous sites within ChIP-seq peaks on chrX in the GM12878 cell line.** (A) Scatterplot showing the allelic imbalance of replicated YY1 ChIP-seq data sets from ENCODE (see Methods). For visualization purposes, allelic imbalance is represented by the log<sub>2</sub> ratio of (Xa + 1) to (Xi + 1). A positive log<sub>2</sub> ratio value indicates more reads on Xa, while a negative value represents more reads on Xi. Each of the 67 dots represents a heterozygous site within a YY1 binding peak. The Pearson correlation between allelic imbalance of the replicated datasets is 0.87. Dotted lines indicate the baselines for balanced allelic binding. Heterozygous sites within 50 bps of escTSSs are indicated by green circles. The intensity of shading of each dot reflects the total number of YY1 reads from both replicates at the heterozygous site (where read counts were assigned to five 20 percentile bins). (B) Heatmap showing heterozygous sites (rows) significantly Xi-biased in more than one dataset (ChIP-seq and DNase I data; column). Only datasets that are significantly Xi-biased at more than four heterozygous sites are listed. Datasets are denoted by the feature name followed by the ENCODE lab where data was generated, and heterozygous sites are denoted by the gene name of the nearest TSS followed by the chrX coordinate of the site. Colors in the heatmap represent log<sub>2</sub> odds ratio values reflecting Xa- or Xi-biased binding of the TF with positive (gold) or negative (brown) values, respectively. The log<sub>2</sub> odds ratio distinguishes Xi bias (negative) and Xa bias (positive). White boxes indicate zero read counts at the corresponding site-data pair. The degrees of significance estimated by FDR-corrected p-values are indicated on a scale of 1 to 5 asterisks with the corresponding p-value thresholds shown in the legend. The datasets from left to right are ordered in increasing counts of higher significance scales denoted by the gray triangle, and the heterozygous sites are ordered using genomic coordinates on chrX. The four lncRNAs previously reported to be associated with Xi-specific superloops are marked with purple bars and colored in purple.

escTSSs and Auto\_bg were similar (1.05 fold; Fig. 4D). In contrast, the YY1 read depth ratio between X\_bg and Auto\_bg was similar to that of input read depth in both female and male cells (Fig. 4C,D). Overall the results showed higher occupancy of YY1 at escTSSs compared to both X\_bg and Auto\_bg for both sexes, and thus enhanced YY1 binding may predispose to ongoing transcription from the inactive X.

Given that YY1 motifs and ChIP-seq peaks were over-represented at escTSSs, we next examined the allelic binding of YY1 on chrX in the female GM12878 cell line to probe its functional role, and confirm its bi-allelic binding at bi-escTSSs (see also schematic in Supplementary Fig. S1D). We extracted the YY1 ChIP-seq read counts on Xi and Xa at the 67 heterozygous sites within YY1 ChIP-seq peaks in GM12878 (Supplementary Table S7). Only two YY1 peaks containing heterozygous sites were within 500 bps to two bi-escTSSs. Indeed, YY1 was bound on both Xa and Xi at both sites, within exon 1 of *SMC1A* (7:13, representing counts of Xa:Xi) and exon 1 or exon 2 of *JPK* transcripts (12:8), confirming bi-allelic transcription and bi-allelic YY1 binding of the escapees.

Consistent with previous reports of Xa-biased binding of TFs reflecting Xa-biased transcription in female cells, the majority of heterozygous sites (52 out of 67) had more reads on Xa than Xi, while 13 heterozygous sites had more reads on Xi (3:474) within exon 1 of *XIST*. The second highest number of YY1 reads was at 215 bps upstream of *UPF3B* (140:0), and the Xa-biased binding agreed with its subject status reported previously<sup>19</sup>. The site with the third highest number of YY1 reads overlapped intron 5 of *FIRRE* and was Xi-biased (4:138). Interestingly, Xa-biased YY1 was detected at sites 477 bps upstream of *FIRRE* TSS and in intron 1 (44:5 and 20:0). Such discrepancy was consistent with the previous report of a shorter alternative *FIRRE* transcript on Xi<sup>9</sup> and the finding of female-specific enhancers in introns 2–12 from ATAC-seq<sup>20</sup>. The Xa- or Xi- specific binding of YY1 might reflect its reported capacity for methylation-sensitive binding<sup>46</sup>, and/or be due to cofactors acting at escapees in collaboration with YY1. Overall, the allelic binding reads of YY1 indicated a positive association with regulation of X-linked genes.

**Significant Xi-biased YY1 occupancy at *XIST*, *FIRRE* and two other superloop-associated lncRNAs in GM12878.** Given the important roles of *XIST* and *FIRRE* on XCI and the strong allelic



imbalance of YY1 binding at these lncRNAs, log<sub>2</sub> ratios of X<sub>a</sub> to X<sub>i</sub> counts were computed to identify strong allelic imbalance at all 67 heterozygous sites (see Methods for detail). The significant correlation between allelic imbalance scores from replicates of YY1 ChIP-seq datasets indicated strong consistency ( $\rho = 0.87$  with  $p < 2.2 \times 10^{-16}$ ; Fig. 5A). Heterozygous sites at exon 1 of *XIST* and intron 5 of *FIRRE* were shown to have the strongest Xi-biased YY1 binding. These sites with YY1 binding were found to be top 5 in significance of allelic imbalance when we extended the allelic analysis to all available ChIP-seq and DNase I datasets in GM12878 as well as 1,321 heterozygous sites, and tested each heterozygous site and data pairs for imbalance significance (see Methods for details and Supplementary Table S8 for raw read counts). We identified 389 significant imbalance site-data pairs comparing to the overall X<sub>a</sub>-biased norm using False Discovery Rate (FDR) correction (discussed in Supplementary File S1 and Fig. S5), which corresponded to 178 unique heterozygous sites (FDR corrected p-values  $\leq 0.05$ ; Supplementary Table S9). Thirty-eight out of 178 sites exhibited imbalance in more than one dataset, within which 16 and 18 sites were biased consistently across data sets towards X<sub>i</sub> and X<sub>a</sub>, respectively. Such consistency in allelic imbalance across TFs indicated broader influence such as open or closed chromatin rather than individual TF binding affinity of alleles.

Interestingly, the majority of significantly Xi-biased imbalances mapped to the vicinity of the four lncRNAs associated with Xi-specific superloops (Fig. 5B and Supplementary Table S9). This finding is complementary and consistent to previous reports on Xi-specific binding of YY1 and CTCF to two of the four lncRNAs, *DXZ4* and *FIRRE*, in mouse and human<sup>9,47–49</sup>. YY1 binding at the heterozygous site within *XIST* was the most significantly Xi-biased across all data sets and sites (FDR-corrected p-value =  $7.12 \times 10^{-221}$ ), while 21 other datasets were also significantly Xi-biased at the same site to a lesser degree. The Xi-biased YY1 binding at heterozygous sites around intron 5 of *FIRRE* were ranked fourth and fifth in overall significance, whereas RAD21 and CTCF were preferentially Xi-bound with lower ranks. While no heterozygous sites overlapped *DXZ4*, significant Xi-biased sites overlapped the *DXZ4* associated non-coding transcript 2, *DANT2*. Notably, *DANT2* is located within the previously reported boundary of two Xi superdomains, which were not associated with X<sub>a</sub> or in male cells<sup>14</sup>. Overall, significant Xi-biased bindings of YY1 and MYC were found in all four lncRNAs, while RAD21 and CTCF were significantly Xi-biased at *FIRRE* and *LOC550643*, but not *XIST* and *DANT2*. The allelic analysis revealed Xi binding of YY1 at Xi-transcribed genes in general, including escTSSs and key lncRNAs involved with XCI, Xi-specific superloops and the superdomain boundary.

## Discussion

Here we analyzed CAGE datasets reporting the transcription levels at chrX TSSs to identify escTSSs through differential transcription analysis between male and female samples matched for cell categories. Predicted escTSSs yielded the highest precision and the second highest recall for escapee gene identification compared to three reports using different techniques. In addition, we experimentally validated a novel escapee, which was antisense to an escapee, *DDX3X*. Analyses of experimental data revealed unique properties of escapees and the resulting bias of bi-allelic activity in data. Over-representation analyses on motifs and ChIP-seq data suggested involvement of YY1 sequence-specific binding at escTSSs. Consistently in an independent context (X<sub>a</sub>/autosomal translocation study), over-representation of YY1 ChIP-seq peaks around autosomal escapee TSSs supported the link between YY1 and escapees. Allelic binding analysis in GM12878 cells further indicated a role for YY1 proximal to lncRNAs involved with Xi-specific superloops. In aggregate, the analyses showed that YY1 is associated with Xi expression of not only *XIST* but more broadly of escapees.

From a bioinformatics perspective, the analysis of chrX is challenging. In female datasets, measurements captured are a combination of properties from X<sub>a</sub> and X<sub>i</sub>. Genes subject to XCI are under a mixture of positive (X<sub>a</sub>) and negative (X<sub>i</sub>) regulation, whereas bi-allelically transcribed escapees are positively regulated on both X chromosomes. Direct use of tools and approaches designed for autosomes can confound interpretation. Our analyses explored properties of bi-escTSSs both in DNAm and TF binding perspectives. Despite having higher expression in females, bi-escTSSs have similar total DNAm levels between sexes consistent with bi-allelic activity in females. In ChIP-seq datasets, there are more bi-escTSSs with peaks than the mono-allelically active background TSSs on chrX for both sexes (recall Fig. 3D). Such over-representation in female datasets resulted from a combination of the bi-allelic effect, and predispositions of TFs to bind at escapees. We have revealed properties of escapees by addressing chrX with targeted analyses of genome-scale data.

YY1 was previously implicated to trans-activate *XIST* expression through tethering *XIST* RNA<sup>12,13</sup>, and knocking-down YY1 significantly affected *XIST* expression<sup>50</sup>. Complementary to previous studies, our report of motif and ChIP-seq peak enrichment at bi-escTSSs and allelic binding suggested that YY1 plays a broader and positive role in the regulation of escapees through sequence-specific binding. Furthermore, adding to previous reports of Xi-specific binding of YY1 and CTCF at *DXZ4* and *FIRRE*<sup>9,47–49</sup>, our analysis identified significant Xi-specific binding of YY1 and MYC at all four lncRNAs frequently associated with superloops in GM12878 cells. A plausible mechanism for the observed YY1 allelic binding preference toward X<sub>i</sub> is its binding sensitivity to DNAm<sup>12,46</sup>.

Combining our results with the literature, we postulate two non-exclusive mechanistic models for YY1's involvement in the regulation of escapees: trapping by RNA and chromatin looping. A RNA trapping model is based in part on the capability of YY1 to bind both DNA and RNA<sup>11,28,51</sup>, localizing *XIST* RNA to the X nucleation center<sup>11</sup>. Supporting the trapping model, a recent report showed enhanced YY1 occupancy on DNA leads to higher stability in gene expression when YY1 interacts with nearby RNAs<sup>51</sup>. Alternatively, the chromatin looping model is supported by the observation that YY1 participates in controlling long distance DNA interactions during B cell development<sup>52,53</sup>, and the enrichment of YY1 ChIP-seq peaks at boundaries of large chromatin domains<sup>54</sup>. As active regions on X<sub>i</sub> loop away from the condensed X<sub>i</sub><sup>55–57</sup>, the significant YY1 Xi-biased binding at the four lncRNA loci, where Xi-specific superloops are highly interactive, indicated a potential involvement of YY1 to chromatin looping.

Overall the results support a model in which YY1 facilitates escape from XCI, but the analyses are generally limited by the availability of data. At the transcript level, our approach in identifying differentially transcribed TSSs from FANTOM5 CAGE datasets was limited to broadly escape genes, as there are few precisely matched male and female samples. Variability of escape from XCI between tissues has been observed<sup>19,38–40</sup>, therefore the role and the regulation of tissue-specific escapees are important future directions. Further investigation of relationships between cell categories and sexes could enable the identification of cell type-specific escapees, but will require sufficient numbers of samples from both sexes in the same cell type. At TF binding level, both the JASPAR motif collection and the ENCODE ChIP-seq datasets are far from complete, covering only a subset of TFs. Furthermore, the allelic imbalance analysis in human is particularly constrained to a small set of heterozygous sites within TF peaks and within the female cell lines that are skewed in XCI. Lastly, our allelic analysis on chrX did not take into account the TF binding affinities towards different alleles, as we intended to capture the epigenetic influence from XCI, which was supported by the observation of multiple TFs sharing significant Xi-biased binding at individual sites. Despite the limitations, we were able to reveal the intrinsic data properties of X-linked genes in males and females, which in turn allowed the identification of the potential regulator of escapees, YY1, and suggested its functional role on the inactive X.

## Methods

All analyses were conducted in R (3.0.2)<sup>58</sup> and Bioconductor (2.13)<sup>59</sup> unless otherwise stated.

**Public datasets.** For expression analysis and sex classification, the RLE-normalized expression table of robust CAGE peaks for human samples was retrieved from the FANTOM5 consortium<sup>25</sup>. For the DNA methylation analysis, raw data files in idat format from Illumina 450 k array generated by TCGA were retrieved<sup>27</sup>, and the following datasets for four cancer types were used: Bladder Urothelial Cancer (BLCA), Colon Adenocarcinoma (COAD), Head and Neck Squamous Cell Carcinoma (HNSC) and Lung Adenocarcinoma (LUAD). For the ChIP-seq and DNase I analysis, uniformly processed ChIP-seq peaks and DNase I peaks were retrieved from the ENCODE<sup>28</sup> page on UCSC genome browser<sup>29</sup>. For generating read depth plots, normalized bigwig files from ENCODE were retrieved from <https://sites.google.com/site/anshulkundaje/projects/wiggler>. For allelic binding analysis, a total of 213 files of ChIP-seq, input and DNase I data from GM12878 cells were retrieved from the ENCODE<sup>28</sup> page on UCSC genome browser<sup>29</sup> and from Kilpinen *et al.*<sup>60</sup>. The ChIP-seq peaks of non-ENCODE data were called using MACS<sup>61</sup> with default parameters.

**Classification and differential expression of the sexes using FANTOM5 CAGE data.** All CAGE TSSs were labeled with the gene name corresponding to the nearest TSSs of Ensembl transcripts from ENSEMBL GENES 75<sup>62</sup>. A supervised Random Forest classifier using randomForest R package<sup>63</sup> was implemented to classify the sexes for the FANTOM5 CAGE data sets. Data from samples involving treatments of the cells or tissues and multiple donations except the first from the same individual were excluded, resulting in a total of 809 datasets. Of these, 296 male and 234 female labeled samples were used as training data. Two classifiers were constructed, one using 5071 TSSs in the non-PAR portions of chrX as features, and the other using the 56 *XIST*-associated TSSs. Performance of the classifier was assessed using 10-fold cross validation as well as out of bag error estimates. For the selected 5,071 TSS-based classifier, sample outliers were identified based on out-of-bag error estimation. The classifier was used to predict samples labeled unknown.

Ontologies associated with the FANTOM5 samples were retrieved from FANTOM5, including cell types (CL), anatomical systems (Uberon), and diseases (DOID)<sup>64,65</sup>. One hundred and fifty-three cell category terms with association to at least 30 samples were extracted. To take cell categories into account when assessing differential transcription between sexes, the first 29 principal components of cell categories explaining over 90% variance in addition to sexes were used as covariates of the linear regression model:  $y_k = a_{k0} + \sum_{t=1}^{29} (a_{kt} C_t) + a_{k30} S + \varepsilon_k$ , where  $y_k$  is the vector of log10 (tags per million counts at TSS  $k + 5$ ) values in all samples reflecting the level of transcription,  $C_t$  is the vector of rotated cell category relevance values of each sample for the  $t^{\text{th}}$  principal component, and  $S$  is the vector of sexes of samples from the sex classifier. The fitted coefficient of the sex variable,  $a_{k30}$ , and its p-value were used to assess differential transcription between sexes of TSS  $k$ . To reduce the multiple hypothesis testing penalty and avoid bias from non-unique alignment to the genome, 4045 X-non-PAR TSSs that do not overlap repetitive elements were retained for analysis. TSSs with Bonferroni-corrected p-values less or equal to 0.05, and negative  $a_{k30}$  values (higher expression in female samples) were defined as escTSSs.

The precision and recall values compared to escapee lists from the literature were computed using the count of true predicted escapees over the count of all predicted escapees, and the true escapee count, respectively. Genes reported in more than one list were taken to constitute true escapees. Ubiquitous escapees with respect to the samples conducted in the literature were used for comparison: escaping “9/9”, “8/9” in ‘Carrel&Willard’<sup>38</sup>, and “escape from XCI in 27 tissues” in ‘Cotton2015’<sup>19</sup>. Figure 1C is generated using the VennDiagram R package<sup>66</sup>.

**Experimental validation of escape from XCI.** Cell culture: The somatic mouse-human hybrid fibroblast cell lines AHA-11aB1, t60-12 (Xa), t11-4Aaz5, t75-2maz34-4a and t86-B1maz1b-3a (Xi) along with the mouse line tsA1S9-az31b were obtained from the Willard lab<sup>67</sup>. Cells were grown in MEM media (Gibco) supplemented with 7.5% Fetal Bovine Serum, L-glutamine and penicillin/streptomycin at 37 °C, 5% CO<sub>2</sub>. Cells were seeded in 100 mm dishes and grown to confluence, harvested using trypsin and stored at –70 °C until RNA extraction.

RNA extraction: RNA was extracted using Trizol (Invitrogen) as per the manufacturer’s instructions. DNA contamination was removed using DNase I recombinant (Roche) as per the manufacturer’s instructions.

Quantitative strand-specific reverse transcription and PCR: The strand specific reverse transcription protocol was performed according to Chapman, *et al.*<sup>12</sup>. Briefly, RNA was reverse transcribed individually for each assay

as follows: 13  $\mu\text{L}$  of RNA was incubated with 1  $\mu\text{L}$  0.50  $\mu\text{M}$  dNTPs and 2 pmol of T7-tagged strand-specific primers at 70  $^{\circ}\text{C}$  for 5 min, then placed on ice for 1 min. Superscript IV buffer (ThermoFisher) was then added to a concentration of 1x, along with 0.005 mol DTT, 1  $\mu\text{L}$  of RiboLock RNase Inhibitor (ThermoFisher) and 1  $\mu\text{L}$  of SuperScript IV enzyme (ThermoFisher), all to a volume of 20  $\mu\text{L}$ . The mixture was incubated at 55  $^{\circ}\text{C}$  for one hour then heat inactivated at 95  $^{\circ}\text{C}$  for 5 min. Reverse transcription for the antisense of *DDX3X* and *ACTIN* were performed independently.

Quantitative PCR (qPCR) was performed by adding 1.5  $\mu\text{L}$  cDNA to 0.2 mM dNTPs, 2.5 mM  $\text{MgCl}_2$ , 1x Hotstart reaction buffer, 1x EvaGreen dye (Biotum), 0.2  $\mu\text{M}$  each of T7 and amplification primer, and 0.8 U Maxima Hot Start Taq (Fermentas). The qPCR was run on a StepOnePlus<sup>TM</sup> Real-Time PCR System (Applied Biosystems) for 95  $^{\circ}\text{C}$  for 5 min followed by 40 cycles of 20 s at 95  $^{\circ}\text{C}$ , 20 s at 59  $^{\circ}\text{C}$  and 20 s at 72  $^{\circ}\text{C}$ . A melt curve from 52  $^{\circ}\text{C}$  to 95  $^{\circ}\text{C}$  was run to ensure only one product was being amplified. All qPCR runs were done in triplicate. A primer negative control with no primer during the reverse transcription step failed to amplify during qPCR, demonstrating specificity of the qPCR.

**Similarity and differential analysis of DNA methylation between sexes.** Due to the complication of XCI in female samples, a basic preprocessing approach using background subtraction of negative control probes and internal control normalization was conducted on all Illumina 450 k DNA methylation data from TCGA using the Minfi R package<sup>68</sup>. The beta value,  $\beta_i$ , representing the degree of DNAm was computed for each probe  $i$  as  $\text{Meth}_i / (\text{Meth}_i + \text{Unmeth}_i + 100)$  by default. We adopted the analysis of microarray studies by computing the logged differential methylation value between sexes ( $M$ ) and the logged average methylation value ( $A$ ) for each probe  $i$ :  $M_i = \log_2(\beta_{Mi}) - \log_2(\beta_{Fi})$  and  $A_i = (\log_2(\beta_{Mi}) + \log_2(\beta_{Fi})) / 2$ , where  $\beta_{Mi}$  and  $\beta_{Fi}$  were the average  $\beta$  values of probe  $i$  for the male and female samples, respectively. For probes on chrX, we fitted a linear model,  $M = d_0 + d_1A$ , by robust regression using the estimating function of Tukey's bisquare, which gave extreme observations zero weights. We then obtained the residual of each probe to be the DNAm similarity score between sexes. For assessing the overall correlation between DNAm similarity scores and differential transcription of CAGE TSSs, we assigned probes within 50 bps to each TSS. When multiple DNAm probes are within 50 bps of a TSS, the average of the similarity scores was computed. For probes on autosomes, differential methylation analysis was performed, including age and cancer stage as covariates (adapted from the dmpFinder method of Minfi). F statistics were computed by comparing the goodness of fit of the models using age and cancer stage as covariates with and without the addition of sex variable. Bonferroni-corrected p-values were obtained for each autosomal probe.

**Motif over-representation tests of escTSSs using CAGED-oPOSSUM.** The escTSSs were subjected to motif over-representation analysis using our CAGED-oPOSSUM web tool<sup>44</sup>. A total of 478 JASPAR2016 motifs with a minimum specificity of 8 bits for vertebrates<sup>69</sup> were used for the prediction of binding sites. The "Use only FANTOM5 CAGE peaks identified as true TSSs by the TSS classifier" option was selected to filter out CAGE peaks that were less likely to be TSSs. The remaining CAGE peaks were extended 500 bps up- and down-stream, and merged. As background property matching is important for motif over-representation analyses to avoid biases, two tests were conducted differing in background sets: (i) random CAGE TSSs with %GC composition and length matched to escTSSs sampled by CAGED-oPOSSUM, and (ii) 4,939 non-differentially transcribed chrX TSSs with Bonferroni-corrected p-values of 1 from the differential transcription analysis. The TF binding score threshold was set to 85% by default, and the Fisher scores for both tests were retrieved to assess the over-representation significance of each motif around escTSSs.

**TF ChIP-seq peak over-representation testing and read depth plots.** Over-representation testing of TF ChIP-seq peaks within 500 bps of bi-escTSSs was conducted using the one-sided Fisher's exact test. To avoid double counting, only one TSS with the strongest differential expression between sexes was selected per escape gene. As a result, we conducted the analysis on 30 bi-escTSSs, and randomly selected 300 background TSSs from chromosome X ( $X_{bg}$ ) and 3,000 from autosomes ( $Auto_{bg}$ ) with matched average transcription levels. An additional criterion of Bonferroni p-value equal to 1 (from differential transcription analysis) was used for the  $X_{bg}$  set to avoid selecting borderline escapees. We assigned all CAGE TSSs to 5 equally sized bins according to average expression percentile ranks. The percentage of bi-escTSSs assigned to each bin was determined and a background set was selected to match the distribution across bins. We extracted the counts of bi-escTSSs,  $X_{bg}$  and  $Auto_{bg}$  sets having at least one peak within 500 bps for each ChIP-seq data. Given the overlapping and non-overlapping counts of bi-escTSSs and background TSSs, one-sided Fisher's exact tests were conducted to test for positive association between the peaks of each ChIP-seq data and the bi-escTSSs. The significance was adjusted for multiple testing using Bonferroni correction. The log2 ratios of bi-escTSSs to each background set with peaks were computed as  $\log_2(\% \text{ bi-escTSSs with peaks} / \% \text{ background with peaks})$ . For the read depth plots, Bwtool<sup>70</sup> was used to extract the YY1 ChIP-seq and input read depth within 5 kb of the three TSS sets from big wig files (bi-escTSSs,  $X_{bg}$  and  $Auto_{bg}$ ). Read depth ratio between any two TSS sets were computed by comparing the average read depths within  $\pm 50$  bps of the TSSs. The ratios from cells of the same sex were averaged.

**Allelic ChIP-seq and DNase I data analysis of the female GM12878 cell line.** Our in-house allelic binding pipeline was used to extract reads at heterozygous sites from GM12878 datasets and assess mapability for filtering<sup>71</sup>. We obtained the genotype data of GM12878 from the 1000 Genomes Project<sup>72</sup>, and a personalized hg19 genome for GM12878 was built by representing single nucleotide variations as degenerate IUPAC codes (eg. Heterozygous alleles G and A are represented by 'R'). The IUPAC codes allow an equal alignment of reads from both alleles at each heterozygous site without considering the non-reference allele(s) as a mismatch. Raw reads of 214 ChIP-seq, input and DNase I datasets were aligned to this personalized genome, using Novoalign (version

3.01.00: <http://www.novocraft.com>) with default parameters. The phased information of GM12878 was obtained from Illumina (<http://www.illumina.com/platinumgenomes/>). As allelic mapping bias may exist at certain locations within the personalized genome, a set of simulated reads were generated for mappability assessment. We first merged all GM12878 TF ChIP-seq binding peaks plus 100 bps flanking regions, and then simulated all the possible 36-bp reads overlapping with the heterozygous sites for each allele and each strand. After mapping the simulated reads to the personalized genome, only 1,321 out of 1,497 heterozygous sites with balanced simulated read counts between two alleles, calculated by requiring the read count of either allele divided by the sum of two alleles to be between 0.6 and 0.4, were kept for further analysis. As paternal X chromosome corresponds to the Xi in GM12878, we assigned the allelic read counts to Xi or Xa at each heterozygous site accordingly.

All replicated data sets generated by the same lab were merged through summing Xi and Xa reads separately at each site, resulting in 101 merged data sets. The only exception was for Fig. 5A, where replicated YY1 datasets generated by the HudsonAlpha Institute for Biotechnology lab (Haib) were displayed separately to show the correlation of allelic imbalance between replicates. For the individual allelic YY1 binding analysis, only 67 heterozygous sites within uniformly processed YY1 ChIP-seq peaks were examined to avoid noise from low occupancy. Allelic imbalance score at site  $h$  was computed for visualization purposes using the  $\log_2$  ratio of reads at Xa over Xi with a constant of 1 added to avoid zero denominators:  $\log_2((Ra_h + 1)/(Ri_h + 1))$ , where  $Ra_h$  and  $Ri_h$  represented read counts on Xa and Xi at site  $h$ , respectively.

For the batch analysis of allelic imbalance, all 1321 sites were used. The significance of allelic imbalance was assessed using Fisher's exact test comparing the Xa and Xi counts at each heterozygous site to the total Xa and Xi counts within the corresponding data. For each data  $g$  at site  $h$ , the 4 values in  $2 \times 2$  table were  $Ra_h$ ,  $Ri_h$ ,  $(\sum_{m=1}^{1321} Ra_m) - Ra_h$ ,  $(\sum_{m=1}^{1321} Ri_m) - Ri_h$ , where  $Ra_h$  and  $Ri_h$  represented read counts on Xa and Xi at site  $h$  from data  $g$ , respectively. The significance reflected the extremity of imbalanced site-data pairs compared to the overall Xa-biased total read counts. With the 1321 sites and 101 merged data sets, a total of 61,470 site-data pairs with non-zero read counts from both Xi and Xa were tested. FDR p-value correction was used here because we expected multiple hits of the same TF to multiple heterozygous sites overlapping the same gene. The  $\log_2$  odds ratio was computed to reflect the direction of imbalance bias with a positive or negative value reflecting bias towards Xa or Xi, respectively:  $\log_2((Ra_h + 1) * ((\sum_{m=1}^{1321} Ri_m) - Ri_h)/(Ri_h + 1)/((\sum_{m=1}^{1321} Ra_m) - Ra_h))$ .

## References

- Dorak, M. T. & Karpuzoglu, E. Gender differences in cancer susceptibility: an inadequately addressed issue. *Front Genet* **3**, 268, doi: 10.3389/fgene.2012.00268 (2012).
- Werling, D. M. & Geschwind, D. H. Sex differences in autism spectrum disorders. *Curr Opin Neurol* **26**, 146–153, doi: 10.1097/WCO.0b013e32835ee548 (2013).
- Roeters van Lennep, J. E., Westerveld, H. T., Erkelens, D. W. & van der Wall, E. E. Risk factors for coronary heart disease: implications of gender. *Cardiovasc Res* **53**, 538–549 (2002).
- Whitacre, C. C. Sex differences in autoimmune disease. *Nat Immunol* **2**, 777–780, doi: 10.1038/ni0901-777 (2001).
- Clayton, J. A. & Collins, F. S. Policy: NIH to balance sex in cell and animal studies. *Nature* **509**, 282–283 (2014).
- Kemkemer, C., Kohn, M., Kehrer-Sawatzki, H., Fundele, R. H. & Hameister, H. Enrichment of brain-related genes on the mammalian X chromosome is ancient and predates the divergence of synapsid and sauropsid lineages. *Chromosome Res* **17**, 811–820, doi: 10.1007/s10577-009-9072-8 (2009).
- Cerese, A., Pintacuda, G., Tattermusch, A. & Avner, P. Xist localization and function: new insights from multiple levels. *Genome Biol* **16**, 166, doi: 10.1186/s13059-015-0733-y (2015).
- Hacisuleyman, E. *et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* **21**, 198–206, doi: 10.1038/nsmb.2764 (2014).
- Yang, F. *et al.* The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biol* **16**, 52, doi: 10.1186/s13059-015-0618-0 (2015).
- Balaton, B. P. & Brown, C. J. Escape Artists of the X Chromosome. *Trends Genet*, doi: 10.1016/j.tig.2016.03.007 (2016).
- Jeon, Y. & Lee, J. T. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* **146**, 119–133, doi: 10.1016/j.cell.2011.06.026 (2011).
- Chapman, A. G., Cotton, A. M., Kelsey, A. D. & Brown, C. J. Differentially methylated CpG island within human XIST mediates alternative P2 transcription and YY1 binding. *BMC Genet* **15**, 89, doi: 10.1186/s12863-014-0089-4 (2014).
- Makhlouf, M. *et al.* A prominent and conserved role for YY1 in Xist transcriptional activation. *Nat Commun* **5**, 4878, doi: 10.1038/ncomms5878 (2014).
- Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680, doi: 10.1016/j.cell.2014.11.021 (2014).
- Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660, doi: 10.1126/science.1262110 (2015).
- Trabzuni, D. *et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nat Commun* **4**, 2771, doi: 10.1038/ncomms3771 (2013).
- Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489, doi: 10.1038/nature10523 (2011).
- Hall, E. *et al.* Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets. *Genome Biol* **15**, 522, doi: 10.1186/s13059-014-0522-z (2014).
- Cotton, A. M. *et al.* Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum Mol Genet* **24**, 1528–1539, doi: 10.1093/hmg/ddu564 (2015).
- Qu, K. *et al.* Individuality and variation of personal regulomes in primary human T cells. *Cell Syst* **1**, 51–61, doi: 10.1016/j.cels.2015.06.003 (2015).
- Cotton, A. M. *et al.* Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol* **14**, R122, doi: 10.1186/gb-2013-14-11-r122 (2013).
- Ding, Z. *et al.* Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet* **10**, e1004798, doi: 10.1371/journal.pgen.1004798 (2014).
- Kucera, K. S. *et al.* Allele-specific distribution of RNA polymerase II on female X chromosomes. *Hum Mol Genet* **20**, 3964–3973, doi: 10.1093/hmg/ddr315 (2011).
- Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* **22**, 860–869, doi: 10.1101/gr.131201.111 (2012).
- Forrest, A. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470, doi: 10.1038/nature13182 (2014).

26. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461, doi: 10.1038/nature12787 (2014).
27. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120, doi: 10.1038/ng.2764 (2013).
28. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, doi: 10.1038/nature11247 (2012).
29. Rosenbloom, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res* **40**, D912–D917, doi: gkr1012 [pii] 10.1093/nar/gkr1012 (2012).
30. Cotton, A. M. *et al.* Spread of X-chromosome inactivation into autosomal sequences: role for DNA elements, chromatin features and chromosomal domains. *Hum Mol Genet* **23**, 1211–1223, doi: 10.1093/hmg/ddt513 (2014).
31. Marks, H. *et al.* Dynamics of gene silencing during X inactivation using allele-specific RNA-seq. *Genome Biol* **16**, 149, doi: 10.1186/s13059-015-0698-x (2015).
32. Breiman, L. Random Forests. *Machine learning* **45**, 5–32 (2001).
33. Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics* **29**, i199–i209, doi: 10.1093/bioinformatics/btt208 (2013).
34. Sirchia, S. M. *et al.* Loss of the inactive X chromosome and replication of the active X in BRCA1-defective and wild-type breast cancer cells. *Cancer Res* **65**, 2139–2146, doi: 10.1158/0008-5472.CAN-04-3465 (2005).
35. Kawakami, T. *et al.* The roles of supernumerical X chromosomes and XIST expression in testicular germ cell tumors. *J Urol* **169**, 1546–1552, doi: 10.1097/01.ju.0000044927.23323.5a (2003).
36. Goto, Y. & Kimura, H. Inactive X chromosome-specific histone H3 modifications and CpG hypomethylation flank a chromatin boundary between an X-inactivated and an escape gene. *Nucleic Acids Res* **37**, 7416–7428, doi: 10.1093/nar/gkp860 (2009).
37. Mele, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665, doi: 10.1126/science.aaa0355 (2015).
38. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400–404, doi: 10.1038/nature03479 (2005).
39. Peeters, S. B., Cotton, A. M. & Brown, C. J. Variable escape from X-chromosome inactivation: identifying factors that tip the scales towards expression. *Bioessays* **36**, 746–756, doi: 10.1002/bies.201400032 (2014).
40. Berletch, J. B. *et al.* Escape from x inactivation varies in mouse tissues. *PLoS Genet* **11**, e1005079, doi: 10.1371/journal.pgen.1005079 (2015).
41. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311–1320, doi: 10.1038/ng.3142 (2014).
42. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods* **12**, 433–438, doi: 10.1038/nmeth.3329 (2015).
43. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**, D142–D147, doi: 10.1093/nar/gkt997 (2014).
44. Arenillas, D. J. *et al.* CAGED-oPOSSUM: motif enrichment analysis from CAGE-derived TSSs. *bioRxiv*, doi: 10.1101/040667 (2016).
45. Bala Tannan, N. *et al.* DNA methylation profiling in X;autosome translocations supports a role for L1 repeats in the spread of X chromosome inactivation. *Hum Mol Genet* **23**, 1224–1236, doi: 10.1093/hmg/ddt553 (2014).
46. Kim, J., Kollhoff, A., Bergmann, A. & Stubbs, L. Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. *Hum Mol Genet* **12**, 233–245 (2003).
47. Moseley, S. C. *et al.* YY1 associates with the macrosatellite DXZ4 on the inactive X chromosome and binds with CTCF to a hypomethylated form in some male carcinomas. *Nucleic Acids Res* **40**, 1596–1608, doi: 10.1093/nar/gkr964 (2012).
48. Chadwick, B. P. DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. *Genome Res* **18**, 1259–1269, doi: 10.1101/gr.075713.107 (2008).
49. Hacisuleyman, E., Shukla, C. J., Weiner, C. L. & Rinn, J. L. Function and evolution of local repeats in the Firre locus. *Nat Commun* **7**, 11021, doi: 10.1038/ncomms11021 (2016).
50. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The functional consequences of variation in transcription factor binding. *PLoS Genet* **10**, e1004226, doi: 10.1371/journal.pgen.1004226 (2014).
51. Sigova, A. A. *et al.* Transcription factor trapping by RNA in gene regulatory elements. *Science* **350**, 978–981, doi: 10.1126/science.aad3346 (2015).
52. Thorvaldsen, J. L., Weaver, J. R. & Bartolomei, M. S. A YY1 bridge for X inactivation. *Cell* **146**, 11–13, doi: 10.1016/j.cell.2011.06.029 (2011).
53. Atchison, M. L. Function of YY1 in Long-Distance DNA Interactions. *Front Immunol* **5**, 45, doi: 10.3389/fimmu.2014.00045 (2014).
54. Moore, B. L., Aitken, S. & Semple, C. A. Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. *Genome Biol* **16**, 110, doi: 10.1186/s13059-015-0661-x (2015).
55. Minajigi, A. *et al.* Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* **349**, doi: 10.1126/science.aab2276 (2015).
56. Heard, E. & Bickmore, W. The ins and outs of gene regulation and chromosome territory organisation. *Curr Opin Cell Biol* **19**, 311–316, doi: 10.1016/j.ceb.2007.04.016 (2007).
57. Deng, X. *et al.* Bipartite structure of the inactive mouse X chromosome. *Genome Biol* **16**, 152, doi: 10.1186/s13059-015-0728-8 (2015).
58. Team, R. C. R.: A Language and Environment for Statistical Computing, <http://www.R-project.org/> (2015).
59. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115–121, doi: 10.1038/nmeth.3252 (2015).
60. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–747, doi: 10.1126/science.1242463 (2013).
61. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728–1740, doi: 10.1038/nprot.2012.101 (2012).
62. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res* **43**, D662–669, doi: 10.1093/nar/gku1010 (2015).
63. Wiener, A. L. a. M. Classification and Regression by RandomForest. *R News* **2**, 18–22, <http://CRAN.R-project.org/doc/Rnews/%3E> (2002).
64. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol* **13**, R5, doi: 10.1186/gb-2012-13-1-r5 (2012).
65. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* **25**, 1251–1255, doi: 10.1038/nbt1346 (2007).
66. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**, 35, doi: 10.1186/1471-2105-12-35 (2011).
67. Willard, H. F., Brown, C. J., Carrel, L., Hendrich, B. & Miller, A. P. Epigenetic and chromosomal control of gene expression: molecular and genetic analysis of X chromosome inactivation. *Cold Spring Harb Symp Quant Biol* **58**, 315–322 (1993).

68. Fortin, J. P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol* **15**, 503, doi: 10.1186/s13059-014-0503-2 (2014).
69. Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*, doi: 10.1093/nar/gkv1176 (2015).
70. Pohl, A. & Beato, M. Bwtool: a tool for bigWig files. *Bioinformatics* **30**, 1618–1619, doi: 10.1093/bioinformatics/btu056 (2014).
71. Shi, W., Fornes, O., Mathelier, A. & Wasserman, W. W. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res*, doi: 10.1093/nar/gkw691 (2016).
72. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65, doi: 10.1038/nature11632 (2012).

## Acknowledgements

We thank members of the Wasserman and Brown labs for helpful comments and discussion, Dora Pak for management support, and Miroslav Hatas for systems support. We would like to thank all members of the FANTOM5 consortium for contributing to generation of samples and analysis of the dataset and thank GeNAS for data production. This research was supported by the Canadian Institutes of Health Research Operating Grant (MOP-119586; www.cihr-irsc.gc.ca) to the Wasserman and Brown labs; the Genome Canada Large Scaled Applied Research Grant (174CDE; www.genomecanada.ca) to the Wasserman lab; the Postgraduate Scholarships-Doctoral Program from Natural Sciences and Engineering Research of Canada (PGS D; www.nserc-crsng.gc.ca), and the Four Year Doctoral Fellowship from the University of British Columbia (www.ubc.ca) to CYC; a PhD fellowship from China Scholar Council (201206110038; www.csc.edu.cn) to WS; the Four Year Doctoral Fellowship from the University of British Columbia (www.ubc.ca) to BPB; the Child and Family Research Institute, and the British Columbia Children's Hospital Foundation to AM. FANTOM5 was made possible by a Research Grant for RIKEN OSC from MEXT (www.mext.go.jp) to Yoshihide Hayashizaki, Grant from MEXT for the RIKEN PMI to YH, Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT to YH and Grant from MEXT to the RIKEN CLST.

## Author Contributions

C.Y.C., W.W.W. and C.J.B. designed and conceived experiments. Y.L. (sex classification), A.M.M. (DNAm and XCI), A.M. (motif and ChIP-seq over-representation), and W.S. (allelic binding) contributed intellectually within their areas of expertise. C.Y.C. conducted all analyses. W.S. provided the allelic and simulated binding reads. B.B. conducted the experimental validation of escape from X.C.I. A.M.M. provided the escapees and subjects from previously published X; autosome studies. A.M. and D.J.A. developed and provided support for the CAGED-oPOSSUM web tool. The FANTOM5 CAGE data sets: M.I. (data production); H.K. (data handling management); T.L. (tag mapping); P.C., Y.H. and A.R.R.F. (FANTOM5 management and concept). C.Y.C., W.W.W. and C.J.B. wrote the manuscript. All authors read and approved the final manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Chen, C.- *et al.* YY1 binding association with sex-biased transcription revealed through X-linked transcript levels and allelic binding analyses. *Sci. Rep.* **6**, 37324; doi: 10.1038/srep37324 (2016).

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016