



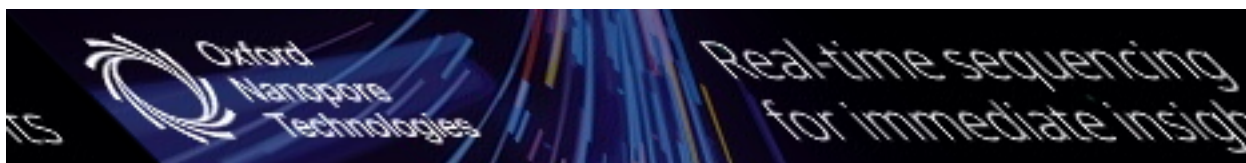
Kinetics of *Xist*-induced gene silencing can be predicted from combinations of epigenetic and genomic features

Lisa Barros de Andrade e Sousa, Iris Jonkers, Laurène Syx, et al.

Genome Res. published online June 7, 2019

Access the most recent version at doi:[10.1101/gr.245027.118](https://doi.org/10.1101/gr.245027.118)

P<P	Published online June 7, 2019 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution-NonCommercial 4.0 International license), as described at http://creativecommons.org/licenses/by-nc/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

Kinetics of *Xist*-induced gene silencing can be predicted from combinations of epigenetic and genomic features

Author Information

Lisa Barros de Andrade e Sousa^{1,11}, Iris Jonkers^{2,7,11}, Laurène Syx^{3,4}, Ilona Dunkel¹, Julie Chaumeil^{3,8}, Christel Picard³, Benjamin Foret³, Chong-Jian Chen^{3,6}, John T. Lis^{2,12} & Edith Heard^{3,12} & Edda G. Schulz^{1,12} & Annalisa Marsico^{1,5,9,12}

Author Affiliations

¹ *Otto Warburg Laboratories, Max Planck Institute for Molecular Genetics, Ihnestraße 63, 14195 Berlin, Germany*

² *Department of Molecular Biology and Genetics, Cornell University, 416 Biotechnology Building, 14853, Ithaca, New York, USA*

³ *Institut Curie, PSL Research University, CNRS UMR3215, INSERM U934, UPMC Paris-Sorbonne, 26 Rue d'Ulm, 75005 Paris, France*

⁴ *Institut Curie, PSL Research University, Mines Paris Tech, INSERM U900, 75005 Paris, France*

⁵ *Department of Mathematics and Informatics, Free University of Berlin, Arnimallee 14, 14195 Berlin, Germany*

Present address: ⁶ Annoroad Gene Technology Co., Ltd, Beijing, China; ⁷ Department of Genetics, University Medical Centre Groningen, University of Groningen, 9700 RB Groningen, The Netherlands; ⁸ Institut Cochin, Inserm U1016, CNRS UMR8104, Université Paris Descartes - Sorbonne Paris Cité, Paris, France; ⁹ Institute for Computational Biology (ICB), Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Oberschleißheim,

Germany; ¹⁰ European Molecular Biology Laboratory (EMBL), Directors' research unit,
Mayerhofstrasse 1, 69117 Heidelberg, Germany

¹¹ These authors are co-first authors and contributed equally to this work.

¹² These authors are co-senior authors and contributed equally to this work.

Corresponding authors

Correspondence to Annalisa Marsico (marsico@molgen.mpg.de), Edda G. Schulz
(edda.schulz@molgen.mpg.de), Edith Heard (Edith.Heard@embl.org) or John T. Lis
(johnlis@cornell.edu).

Running title: Modelling kinetics of *Xist*-induced gene silencing

Keywords: X Chromosome inactivation, *Xist*-mediated gene silencing kinetics, PRO-seq,
Random Forest modelling, Forest-guided clustering, machine learning

Abstract

To initiate X Chromosome inactivation (XCI), the long non-coding RNA *Xist* mediates chromosome-wide gene silencing of one X Chromosome in female mammals to equalize gene dosage between the sexes. The efficiency of gene silencing is highly variable across genes, with some genes even escaping XCI in somatic cells. A gene's susceptibility to *Xist*-mediated silencing appears to be determined by a complex interplay of epigenetic and genomic features; however, the underlying rules remain poorly understood. We have quantified chromosome-wide gene silencing kinetics at the level of the nascent transcriptome using allele-specific Precision nuclear Run-On sequencing (PRO-seq). We have developed a Random Forest machine learning model that can predict the measured silencing dynamics based on a large set of epigenetic and genomic features and tested its predictive power experimentally. The genomic distance to the *Xist* locus, followed by gene density and distance to LINE elements are the prime determinants of the speed of gene silencing. Moreover, we find two distinct gene classes associated with different silencing pathways: a class that requires *Xist* repeat A for silencing, which is known to activate the SPEN-pathway and a second class where genes are pre-marked by Polycomb complexes and tend to rely on the B repeat in *Xist* for silencing, known to recruit polycomb complexes during XCI. Moreover, a series of features associated with active transcriptional elongation and chromatin 3D structure are enriched at rapidly silenced genes. Our machine learning approach can thus uncover the complex combinatorial rules underlying gene silencing during X inactivation.

Introduction

X Chromosome inactivation (XCI) is a developmental process in mammals that ensures equal gene dosage of X-linked genes between XX and XY individuals by transcriptional inactivation of one of the two X Chromosomes in female cells. In placental mammals XCI is triggered by the long non-coding RNA (lncRNA) *Xist*, which is upregulated in a monoallelic fashion and coats the future inactive X Chromosome in *cis* leading to the recruitment of several factors involved in transcriptional inactivation and eventually converting the entire X Chromosome into silent heterochromatin (Escamilla-Del-Arenal et al. 2011; Gendrel and Heard 2014; Galupa and Heard 2015).

Early events following *Xist* coating include the depletion of RNA polymerase II from the *Xist* RNA domain and loss of active histone marks (Chaumeil et al. 2002, 2006) as well as gain of repressive chromatin modifications, such as H2AK119ub1 and H3K27me3, deposited by the polycomb repressive complexes (PRC) 1 and 2, respectively (Plath et al. 2003, 2004; Silva et al. 2003; de Napoles et al. 2004). Subsequently, additional chromatin modifications are gained such as the histone variant macroH2A and DNA methylation of gene promoters (Escamilla-Del-Arenal et al. 2011). In recent years progress has been made in identifying proteins that mediate *Xist*'s functions as well as the domains within the *Xist* RNA that recruit these proteins. *Xist* contains multiple conserved repeats, among which the A-repeat mediates gene silencing through activation of SPEN and other factors including RBM15 (Chu et al. 2015; McHugh et al. 2015) and the B-repeat recruits PRC indirectly through HNRNPK (Wutz et al. 2002; Brockdorff 2017; Pintacuda et al. 2017). The dynamics of *Xist*-mediated silencing are highly variable across genes (Chow et al. 2010; Borensztein et al. 2017; Żylicz et al. 2019), with a subset of so-called escapees remaining active even in somatic cells (Berletch et al. 2011). However, the determinants of susceptibility to XCI remain poorly understood. Since XCI is a multi-step process, local interference with any step, such as *Xist* coating or access to the silencing machinery of one or several silencing

pathways, could delay or prevent silencing of a certain gene. Defining the features that underlie differential susceptibility to XCI remains an important question, particularly as genes that are not fully silenced are implicated in diseases, such as autoimmune syndromes (Bianchi et al. 2012).

Xist RNA spreading occurs by proximity transfer to sites close to the *Xist* locus genomically or in 3D space ('*Xist* entry sites') (Engreitz et al. 2013). From there *Xist* has been proposed to move first into gene-dense regions and then spread to intergenic domains of the X Chromosome (Engreitz et al. 2013; Simon et al. 2013). In differentiated cells *Xist* covers the entire X Chromosome, but is reduced at escape genes (Engreitz et al. 2013; Simon et al. 2013). *Xist* coating is positively correlated with gene density and with PRC2 enrichment and negatively correlated with the density of LINE elements (Engreitz et al. 2013; Simon et al. 2013). Similarly, gene silencing tends to be slower for genes further from the *Xist* locus (Marks et al. 2015; Borensztein et al. 2017) and from LINE elements (Chow et al. 2010; Loda et al. 2017). Moreover, efficiently silenced genes tend to be enriched for polycomb complexes (RING1B, H3K27me3) and depleted for CTCF and active marks such as H3K4me3 and H3K27ac prior to *Xist*-induced silencing (Kelsey et al. 2015; Loda et al. 2017). Thus, a variety of genetic and epigenetic features have been implicated in controlling gene-specific silencing efficiency. However, none of these features alone can predict whether and to what extent a gene will be silenced upon XCI, and the associations with measured silencing efficiencies are generally weak. Since no predictive pattern of features has so far been identified, the susceptibility of genes to *Xist*-mediated silencing is likely to be controlled by a complex combination of different features.

In this study we set out to identify the genetic and chromatin features that predispose genes on the X Chromosome to be efficiently silenced or escape XCI. We measured chromosome-wide silencing dynamics of X-linked genes following induction of *Xist* expression, using allele-specific Precision nuclear Run-On sequencing (PRO-seq) (Kwak et al. 2013). We then trained two Random Forest machine-learning models to predict from 77 genomic and epigenetic features 1) whether a gene is subject to XCI and 2) whether it will be silenced

with fast or slow kinetics. Through Forest-guided gene clustering we identified feature sets that determine the silencing dynamics of sub-groups of genes, indicating that variable silencing efficiencies might be associated with distinct silencing pathways. We have thus developed a framework to comprehensively assess the relative contribution of genetic and epigenetic factors to transcriptional silencing of the X Chromosome in an unbiased and quantitative manner.

Results

Quantification of gene-specific silencing dynamics by PRO-seq

To quantify gene-silencing dynamics during XCI in a chromosome-wide manner, we performed allele-specific PRO-seq upon ectopic *Xist* induction in female murine embryonic stem cells (mESCs). Using an inducible system allowed us to overcome the asynchronous nature of XCI, which normally limits the temporal resolution of population measurements in differentiating mESCs, which are the classical model system to study XCI (Chow et al. 2010). We used the female TX1072 mESC line (Schulz et al. 2014), which was derived from a cross between two different mouse strains (C57BL/6 x CAST/EiJ) and where *Xist* up-regulation from the endogenous locus on the *B6* X Chromosome can be induced by doxycycline (Dox) treatment in undifferentiated cells. For a direct readout of gene silencing we measured the nascent transcriptome by allele-specific PRO-seq (Kwak et al. 2013) at different time points up to 24h of Dox treatment (**Figure 1A**). *Xist* started to be upregulated from the *B6* chromosome about 1h after Dox treatment and reached a plateau after 4h (**Figure 1B** and **1C**, **Supplemental Figure 1**), while global expression of the *B6* X Chromosome was gradually reduced over time due to X inactivation, starting at 4h of treatment (**Figure 1D**).

To quantify silencing dynamics, we fitted an exponential decay function to each gene and estimated gene-specific silencing half-times, i.e. the time point when transcription on the *B6* X Chromosome is reduced by 50% compared to the uninduced control (**Figure 1E** and

examples in **1F**). After several filtering steps, we had estimated reliable half-times for 280 genes, which were used for all subsequent analyses and ranged from several hours up to several days (**Figure 1G**, **Supplemental Table 1**, **Supplemental Table 2**).

Silencing dynamics are comparable *in vitro* and *in vivo*

To ensure that the relative silencing dynamics across genes, when XCI is induced in undifferentiated mESCs, are comparable to those in the cellular context where XCI occurs endogenously, we generated two additional data sets, where mRNA-seq was performed at different time points of Dox treatment in undifferentiated and differentiating mESCs (**Figure 2A**). The computed half-times were comparable between these two data sets (**Figure 2B**, Pearson correlation coefficient: $r=0.75$), suggesting that the differentiation process only has a minor impact on relative gene silencing dynamics. When comparing half-times estimated from the two different data types (mRNA-seq vs PRO-seq) correlation was generally a bit lower, independent of the cellular context (**Figure 2C** and **2D**, Pearson correlation coefficient $r=0.52$ and $r=0.51$), which would be expected given that PRO-seq measures the direct transcription dynamics, whereas mRNA-seq kinetics are modulated by transcription, RNA-processing and degradation. We also compared the estimated half-times to a previous study that had used a different, Dox-independent strategy to make XCI non-random by inserting a stop-cassette in *Xist*'s repressive antisense transcript *Tsix* (Marks *et al.* 2015). The silencing classes defined in that study (early, intermediate, late, escapee) are in good agreement with the half-times estimated from the PRO-seq data (**Figure 2E**), suggesting that Dox-induced XCI recapitulates endogenous gene silencing dynamics.

Finally, we compared our *Xist*-induced gene silencing half-times in mESCs to the dynamics of (imprinted) XCI measured in pre-implantation mouse embryos *in vivo* through single-cell RNA-seq (Borensztein *et al.* 2017). The gene classification in that study (early: 16-cell stage; intermediate: 32-cell stage; late: blastocyst stage or escapee) was once more in good agreement with the silencing half-times estimated from the PRO-seq data (**Figure 2F**).

Based on the PRO-seq derived silencing half-times, we classified all genes according to whether they are subject to XCI or escape (**Figure 1G** and **2G**, silenced / not silenced) and whether they are silenced with slow or fast kinetics (**Figure 1G** and **2G**, early / late). Genes with intermediate half-times between the classes were excluded from the analysis (see gap between groups in **Figure 1G**). The resulting classes largely agree with those previously defined in differentiating mESCs and in pre-implantation embryos (compare **Figure 2G** with **Figure 2E** and **2F**). Moreover, the 'not silenced' class contains 37 out of 50 (74%) known escapees annotated from different cell types, and is strongly enriched for escapees compared to the 'silenced' class when considering only (high confidence) escapees identified in at least two different studies (e.g. *Ddx3x*, *Taf1*, *Pdbc1*, *Kdm6a*, *Usp9x*, *Hcfc1*, *Hdac6*, *Mgmt1*, *Ftx*, *Nkap* and *Uba1*, odd ratio = 6.7, $p = 2.6 \times 10^{-5}$, Fisher's exact test, **Supplemental Table 2**) (Yang et al. 2010; Berletch et al. 2011; Splinter et al. 2011; Calabrese et al. 2012; Wu et al. 2014; Marks et al. 2015; Andergassen et al. 2017).

Identifying determinants of gene silencing dynamics with Random Forest modelling

We noted that genes close to the *Xist* locus tended to be silenced earlier than distal genes (**Figure 2H**), in agreement with a previous study (Marks et al. 2015). However, many genes did not follow this trend. To uncover additional factors that potentially determine the susceptibility to *Xist*-mediated silencing we developed a machine learning model to predict silencing dynamics based on genomic and epigenetic features.

We collected 138 publicly available high-throughput datasets (ChIP-seq and Bisulfite-seq) measuring chromatin modifications and DNA binding factors, mostly in male mESCs. As these data sets had been generated in undifferentiated mESCs, they correspond to the chromatin state before *Xist* induction. After stringent filtering on data quality, we computed the enrichment for 59 of these epigenomic features at promoters or gene body regions as appropriate (**Table 1**, **Supplemental Text 1**). In addition, we included a series of genomic

and structural features, such as gene density, the frequency of 3D chromatin interactions with different genomic elements and the linear distance to other genomic features, such as the distance to the *Xist* locus, the next TAD boundary, lamin-associated domain (LAD) or full-length LINE element (**Table 1, Supplemental Text 1**).

A linear model to predict a gene's susceptibility to *Xist* mediated silencing from the collected epigenetic and genomic features had little predictive power. This indicates that no single linear combination of features or rules could be defined to discriminate, for example, silenced from not silenced genes. The different functional domains of *Xist* might recruit distinct silencing complexes (e.g. PRC1 and SPEN/HDAC3) and susceptibility to each silencing pathway might be determined by distinct feature patterns. We expected to identify such feature patterns from our data set, since genes that require the *Xist* A-repeat for silencing, which activates SPEN, exhibit longer silencing half-times compared to A-repeat independent genes that might be preferentially targeted by polycomb complexes (**Figure 2I**, Kolmogorov-Smirnov (KS) test, $p = 2.2 \times 10^{-6}$) (Sakata et al. 2017).

To identify combinatorial rule sets that could predict silencing susceptibility, we used Random Forest, a non-parametric machine learning method that combines an ensemble of single classification trees, which successively split the feature input space in a non-linear fashion, to predict the value of a discrete binary variable (**Figure 3A** and **Figure 3B**). We built two binary classification models, to predict from a total of 77 epigenetic and genomic features whether a gene would be silenced or not (XCI/escape model), and whether it would be silenced early or late (silencing dynamics model) using the classification described above (**Table 1, Figure 1G**). The half-time thresholds were selected such that they maximize the model classification accuracy (**Supplemental Text 2, Supplemental Table 3**). The XCI/escape model would identify combinations of factors, that are important for silencing in general, and the silencing dynamics model would find those that influence the kinetics of gene silencing.

The two Random Forest models predict gene silencing dynamics with error rates of 28-29% meaning that 71-72% of genes are classified correctly. We assessed the individual

contribution of each feature to the classification accuracy via Random Forest variable importance analysis for each class in the two models by the Mean Decrease in Accuracy (MDA), which indicates the importance of the feature for the classification performance (**Figure 4, Supplemental Figure 2, Supplemental Text 2**). We also trained our models on a set of only 10 (XCI/escape model) and 8 (silencing dynamics model) top features, which greatly improved the prediction error rate to 22.5% and 21.5%, respectively (**Supplemental Figure 3**). The most important feature associated with silencing in both models was close genomic proximity of a gene's transcriptional start site (TSS) to the *Xist* locus (MDA 15-16% in the XCI/escape model, MDA 14-22% in the silencing dynamic model, **Figure 4, Supplemental Figure 2, Supplemental Figure 4, Supplemental Figure 5**). Also a close proximity to LINE elements, low gene density and enrichment for PRC1 (RING1B, H2AK119ub1, RYBP) and PRC2 (EZH2, H3K27me3) are associated with (early) silenced genes in both models, with PRC1 playing a more prominent role in the XCI/escape model (MDA 8% for not silenced class). (Early) silenced genes are enriched for histone deacetylases (HDAC), involved in gene repression (MDA 1.6-2.3% for early silenced class and 3.1-3.8% for not silenced class) and depleted for features associated with active transcription, such as H3K27ac and RNAPII S5P (MDA 1.8%/1.7% for silenced class) and the general transcription factor TAF1 (MDA 6.3% for not silenced class). Moreover, also sequence-specific transcription factors, such as ESRRB (MDA 2.2% for not silenced class) and SOX2 (MDA 1.6%-5.5% in silencing dynamics model) contribute to model performance. Among the top features specific for the XCI/escape model we found binding of the TET1, implicated in DNA demethylation, enriched at silenced genes (MDA 2.1% for the not silenced class). In the silencing dynamics model by contrast, several features related to 3D chromosome organisation seem to be important. While genes located in close proximity to a TAD border tend to be silenced late, genes that are close to a LAD or highly connected to other genomic regions based on Hi-C/HiCap data tend to be silenced earlier. In summary, we have identified different feature sets that appear to influence whether or not a gene is subject to XCI and also, whether silencing occurs with slow or fast dynamics.

Forest-guided clustering of X-linked genes uncovers combinatorial rules of gene silencing

The variable importance analysis above pinpoints the individual contribution of each feature to the classification problem but cannot identify the role of correlated features and of feature combinations associated with different silencing pathways, which ultimately determine the silencing class of each gene. We therefore implemented a Forest-guided clustering approach to stratify the genes into subgroups according to different combinations of rules. We used the proximity between genes within the Random Forest model to group genes that are regulated by the same set of genomic and epigenetic features. The number of clusters is chosen such that each cluster has a low degree of class mixture (containing mainly genes from one class and none or only few genes from the other class) while maintaining a small number of clusters in total (**Figure 3C, Supplemental Figure 6, Supplemental Text 3**). The results are visualized in a heatmap showing the genes (columns), grouped by cluster, and a subset of features (rows) selected based on whether they were significantly different across clusters (p-value from an ANOVA test, **Figure 5A, Figure 6A**).

For the XCI/escape model three clusters are found (**Figure 5A and 5B**). Genes in clusters 1 and 2 are mainly predicted as silenced, and those in cluster 3 as not silenced (**Figure 5C**). Genes tend to escape when they are far from the *Xist* locus, from LINE elements and from LADs; they are found in gene dense regions and are enriched for transcription elongation marks such as RNAPII S2P and H3K36me3 and never overlap with LADs. A cluster of silenced genes (cluster 1) is already marked by a repressive chromatin state (PRC1/2, HDAC1) and bound by TET1, while genes in the other silenced cluster (cluster 2) are depleted for those marks (**Figure 5A, Supplemental Figure 7, Supplemental Figure 8**).

To test whether these two clusters might be associated with different silencing pathways, we analyzed how they were affected in *Xist* mutants lacking either the A or the B and C repeats (**Supplemental Text 4**). We used data from a previous study that had analyzed both mutations in mESCs and from another study that had characterized the A-repeat mutant in trophoblasts *in vivo* (Sakata et al. 2017; Bousard et al. 2018). Cluster 2 was enriched for genes that could still be silenced by a BC repeat mutant compared to cluster 1 (repeat BC independent, odd ratio = 1.6, $p = 0.19$, Fisher's exact test), while cluster 1 was enriched for genes still silenced in the A repeat mutant in either data set (mESC data: odd ratio = 2.3, $p = 0.09$ and trophoblast data: odd ratio = 2.76, $p = 0.003$, Fisher's exact test; **Figure 5D**, **Supplemental Figure 9**). These findings are consistent with the idea that genes that require polycomb for silencing (repeat BC dependent) are already pre-marked by PRC1/2 and the associated histone modifications (cluster 1), in contrast to genes that require that A repeat, which activates the SPEN/HDAC3 silencing pathway (cluster 2).

In the next step, we investigated the factors that would distinguish early and late silenced genes (silencing dynamics model). Again, the Forest-guided clustering approach produced two early silenced clusters (1, 2) with lower half-times and one late silenced cluster (3) with higher half-times (**Figure 6A** and **6B** and **6C**). Again, one early silenced cluster (1) is pre-marked by polycomb repressed chromatin (H2AK119ub1, RING1B, EZH2, SUZ12, H3K27me3) and also H3K4me1. The other early silenced cluster (2) is mainly characterized by a preferential location of genes in LINE-dense regions, an enrichment of features related to transcriptional elongation, such as E2F1 subunit and H3K79me2 and the transcription factor YY1. Genes in both early silenced clusters tend to be far away from TAD borders, to overlap with *Xist* entry sites and to exhibit strong 3D contacts with the *Xist* locus. The late silenced genes in cluster 3 are mainly characterized by genomic features; they are located in gene dense regions, far from the *Xist* locus, from LINE elements and from LADs (**Figure 6A**, **Supplemental Figure 10**, **Supplemental Figure 11**). We again analyzed the repeat A and repeat BC dependency in the two early silenced clusters. Again, the polycomb pre-marked cluster 1 tends to be enriched for repeat A-independent genes, albeit this effect was

not statistically significant, and no difference was found for repeat BC dependent genes (**Figure 6D**, **Supplemental Figure 9**).

Experimental testing of model predictions

To validate our machine learning model we used the trained XCI/escape Random Forest model to predict the silencing class for X-chromosomal genes that had not been measured in the PRO-seq experiment due to insufficient coverage and had therefore not been used for model training (**Figure 3B**, **Supplemental Table 4**). We performed an independent Dox induction time course experiment and used pyrosequencing to assess the allele-specific expression of 6 genes predicted to be silenced (**Figure 7A**, top) and 5 genes predicted to be not silenced (**Figure 7A**, bottom). The half-times of the silenced genes ranged from 0.45-0.87 days, and those of the not silenced genes lay between 0.99 and 3.01 days. The difference was highly significant ($p = 0.0043$, Wilcoxon Rank Sum Test, **Figure 7B**) and the half-time of all silenced genes fell in the silenced category (<0.9 days, **Figure 1G**), while 3 out of 5 not silenced genes also fell in the respective category (>1.6 days, **Figure 1G**). To further validate the model, we compared all model predictions to the silencing half-times estimated from the mRNA-seq time course in undifferentiated mESCs (**Figure 2A**, replicate A). Genes predicted as not silenced exhibited a much longer silencing half-times than gene predicted as silenced ($p = 1.7 \times 10^{-5}$, Wilcoxon Rank Sum Test, **Figure 7C**). Genes that have not previously been reported as escapees, but are either measured or predicted from the PRO-seq model to be not silenced and are also measured as not silenced in the mRNA-seq data are potentially novel escape genes, such as *B630019k06Rik*, *Porcn*, *Ssr4*, *Gm14820*, *Ppp1r3f* (**Supplemental Table 2**, **Supplemental Table 4**).

We next tested whether the model could predict silencing susceptibility to *Xist* transgenes located on an autosome. We used published allele-specific mRNA-seq data for a series of mESC clones that had integrated doxycycline-inducible *Xist* transgenes in different locations

on the X Chromosome and on Chromosome 12 (**Supplemental Table 5**) (Loda et al. 2017). The cells were trisomic for Chromosome 12 such that silencing of one copy should not affect cell viability. We adapted the model feature 'distance to *Xist*' to account for the different locations of the transgene and calculated the fraction of genes predicted to be silenced by our model within all genes that were silenced, for each *Xist* transgene (**Figure 7D**, red lines). These values varied considerably between clones depending on the size of the chromosomes and the location of the transgene. For 5 out of 6 clones the percentage predicted to be silenced was significantly higher than expected for a random sample, as estimated through a bootstrapping approach (**Figure 7D**, cp. red line to background distribution, **Supplemental Text 5**). Although potentially limited by the efficiency of the transgenes (**Supplemental Figure 12**), this analysis shows that our model can to some extent be generalized even to other chromosomes. Taken together, these results confirm that our machine learning model can predict X-chromosomal gene silencing based solely on epigenetic and genomic features.

Discussion

In this study we have developed a machine learning model that can predict a gene's susceptibility to *Xist*-mediated silencing from a combination of epigenetic and genomic features. To train the model we measured silencing kinetics with high temporal resolution through allele-specific PRO-seq. Compared to previous studies (Marks et al. 2015; Borensztein et al. 2017), we assess silencing dynamics by measuring nascent transcription, therefore observing instantaneous changes in transcription by transcriptionally engaged RNAPII allowing a more direct quantification of silencing compared to mRNA-seq which has previously been used. Moreover, the use of an inducible system allows us to uncouple XCI from differentiation and to avoid the use of mutations to ensure non-random XCI. Unlike previous studies that focus on just a few genes and/or investigate a few selected promoter features which are potentially linked to the XCI (Kelsey et al. 2015; Marks et al. 2015; Loda

et al. 2017), we set out to identify silencing determinants in an unbiased manner based on a large number of epigenetic and genomic features. To uncover the combinatorial rules that control silencing dynamics we go one step beyond classical variable importance analysis in Random Forests and introduce a Random Forest-guided visualization scheme. The determinants of silencing for groups of clustered genes recapitulate previous observations but also shed light on novel players or combination of features potentially controlling a gene's susceptibility to *Xist*-mediated inactivation.

Linear distance and 3D interactions with the *Xist* locus (leading to fast *Xist* deposition at so-called 'entry sites') are thought to be the prime determinants of early *Xist* spreading (Engreitz et al. 2013). Our model finds the same features to be highly predictive of gene silencing dynamics, an association that has been described previously (Marks et al. 2015; Borensztein et al. 2017) and suggests that efficient *Xist* coating is required for fast silencing. However, previous studies had shown that *Xist* RNA initially tends to spread to gene-dense and LINE-poor regions (Engreitz et al. 2013; Simon et al. 2013), but in our analysis, gene density is associated with reduced silencing and LINE elements are found in proximity to rapidly silenced genes. A similar association has been reported previously (Chow et al. 2010; Loda et al. 2017) and suggests that *Xist* coating is not the only determinant of silencing.

The *Xist* RNA recruits several protein complexes that mediate gene silencing, such as SPEN, which binds directly to the repeat A element and PRC complexes, which are indirectly recruited by the repeat B (Wutz et al. 2002; Chu et al. 2015; Monfort et al. 2015; Brockdorff 2017). Our model identifies groups of genes associated with each of these silencing pathways. Repeat B/PRC-associated genes are already enriched for PRC1 and PRC2 prior to the onset of XCI, suggesting that polycomb pre-marking might promote and even accelerate gene silencing and/or reinforce *Xist* spreading, as suggested by a recent study (Colognori et al. 2019). A similar enrichment of PRC components has previously been found at genes susceptible to ectopic silencing by *Xist* transgenes (Kelsey et al. 2015; Loda et al. 2017). Although we do not find a clear signature at repeat A-associated silenced genes in the XCI/escape model, early silenced genes in the silencing dynamics model that are not

enriched for PRC, are located in particularly LINE-dense regions, suggesting that LINE elements might promote silencing of polycomb-independent genes.

Previous studies looking at post-XCI cells have proposed a role of CTCF in XCI (Filippova et al. 2005; Berletch et al. 2015) and have found a moderate enrichment of CTCF prior to XCI at promoters of escapees compared to promoter of silenced genes (Loda et al. 2017). While we do not find CTCF as one of the discriminating promoter features in our XCI/escape model, we observe a significant enrichment of CTCF signal at enhancers of not silenced X-linked genes (**Supplemental Text 6, Supplemental Figure 13**), suggesting a potential role of CTCF in gene silencing mediated by chromatin looping between enhancers and promoters.

Finally, our analysis identified several structural features that appear to modulate the dynamics of silencing. A high 'connectivity' of some genes, i.e. how much the gene is involved in 3D interactions with other genomic elements, is associated with faster silencing, maybe because *Xist* RNA can spread more easily to these genes through proximity transfer. Moreover, early silencing preferentially occurs at genes that are close to a LADs, which generally contain repressed genes (van Steensel and Belmont 2017), while genes close to TAD boundaries tend to be silenced late.

In conclusion, we have developed two Random Forest models that can accurately predict silenced and not silenced/escape genes, but also classes of early versus late silenced genes, constituting the first chromosome-wide predictive models of gene silencing from a very large set of features. We confirmed the predictive nature of our models by experimental testing of model predictions. The Random Forest approach allows us to quantify the relative contribution of several features that have previously been associated with XCI (e.g. linear distance to *Xist*, LINE elements, enrichment for PRC1 and PRC2, etc) and suggested new features, which can be tested in more detail in future studies, such as TET1 and some pluripotency factors, such as ESRRB and SOX2, which have recently been implicated in reactivation of the X chromosome during reprogramming (Janiszewski et al. 2019). Additional features could be included in the model in the future to further improve our ability

to predict silencing susceptibility and a detailed experimental investigation of the different silencing pathways elicited by *Xist* will facilitate the interpretation of the features that predict silencing dynamics as well as escape from XCI.

Methods

ES cell culture

The female TX1072 mESC line was grown in serum-containing medium, supplemented with LIF and 2i, as previously described (Schulz et al. 2014). Details on the cell line and culture conditions are given in **Supplemental Text 7**.

PRO-seq

PRO-seq was performed as described previously with some modifications (Mahat et al. 2016). For details about the experiment and the allele-specific bioinformatics analysis see **Supplemental Text 7**.

mRNA-seq and Pyrosequencing

RNA was isolated and converted to cDNA using standard procedures. For mRNA-seq libraries were prepared using the TruSeq Stranded mRNA LT Sample Prep kit (Illumina) and sequenced and analyzed using standard procedures (**Supplemental Text 7**). For pyrosequencing, an amplicon containing a SNP is amplified by PCR from cDNA using GoTaq Flexi G2 (Promega) with 2.5 mM MgCl₂ or HotStarTaq (Qiagen) for 40 cycles. The PCR product was sequenced using the Pyromark Q24 system (Qiagen). Assay details are given in **Supplemental Table 6**.

Silencing half-times

To normalize for sequencing depth, the reads mapping on *B6* genome were divided by the total number of allele-specific reads for each gene:

$$f_{B6}^t = \frac{reads_{B6}^t}{reads_{B6}^t + reads_{Cast}^t} \quad (1)$$

f_{B6}^t was averaged across replicates (0, 24h), which showed good replicate agreement (**Supplemental Figure 14**), resulting in a total of eight time points ($t = 0, 0.5, 1, 2, 4, 8, 12, 24h$). The allelic ratio was calculated as follows

$$ratio^t = \frac{f_{B6}^t}{f_{Cast}^t} = \frac{f_{B6}^t}{1 - f_{B6}^t} \quad (2)$$

and normalized to the uninduced control ($t=0$) to correct for basal skewing (different transcriptional activity at the two alleles in the absence of Dox):

$$norm^t = \frac{ratio^t}{ratio^0} = \frac{f_{B6}^t}{1 - f_{B6}^t} \times \frac{1 - f_{B6}^0}{f_{B6}^0} \quad (3)$$

To estimate gene-specific silencing half-times an exponential decay function was used

$$N(t) = e^{-kt} \quad (4)$$

where k represents the silencing rate that was fitted to $norm^t$ using the `nls` function (`stats` R package) and the half-time was calculated as:

$$halftime\ t_{1/2} = \frac{\ln(2)}{k} \quad (5)$$

A maximum value of $k = 5$, corresponding to a half-time of 3.5 days was set, as higher half-times cannot be reliably estimated from our data, due to the limited range of time points from 0 to 24h. The goodness of fit was evaluated via the square root of the sum of squared residuals $sqrRSS$ defined as:

$$sqrRSS = \sqrt{\sum_t (norm^t - N(t))^2} \quad (6)$$

After filtering out genes without SNPs, with < 10 reads for at least one time point and strong allelic skewing ($f_{B6}^0 < 0.2$ or $f_{B6}^0 > 0.8$) and $\text{sqr}t\text{RSS} > 1.5$, we obtained reliable half-times for 296 X-chromosomal genes on mouse genome mm10 (**Supplemental Table 1, Supplemental Table 2**). Those genes were mapped to the mouse genome mm9 with the `liftOver` tool from UCSC Genome Browser (Kuhn et al. 2007).

Gene half-times from both differentiated and undifferentiated mRNA-seq time series were computed in the same way as described above. For the undifferentiated mRNA-seq data set we discarded replicate B due to insufficient read coverage and only used replicate A (**Figure 2A**), which resulted in computing half-times for 346 genes. For the differentiated mRNA-seq data set we averaged replicate A and B for each time point and computed half-times for 379 genes. For 233 genes, half-time could be estimated from all 3 data sets.

Definition of model features

The epigenetic and genomic features used for the Random Forest models are summarized in **Table 1** and listed in detail in **Supplemental Table 7**. In total, 138 ChIP-seq libraries and one bisulfite sequencing experiment on undifferentiated male mESCs (except HDAC3) were collected from various sources. After performing ChIP-seq library quality control with the `deepTools` package (Ramírez et al. 2014), 58 ChIP-seq libraries and the bisulfite sequencing experiment were used in the model (**Supplemental Text 1, Supplemental Figure 15, Supplemental Figure 16, Supplemental Table 8**). Epigenetic features are defined as the average ChIP-seq signal in a pre-defined genomic region, normalized to the signal of a matched control experiment in the same region. Read counts of each feature were normalized to the control with the R package `normR` (Helmuth et al. 2016; Kinkley et al. 2016) (**Supplemental Figure 17**). While for most features the signal is extracted in a genomic region around the active gene TSS, for broader features such as elongation marks H3K36me3, RNAPII S2P, and H3K79me2 the signal was averaged over the entire gene body (**Supplemental Table 7**). The active TSS for each gene was identified based on the

PRO-seq data at time point $t=0$ through the dREG method, which finds regions which harbor bidirectional transcription (**Supplemental Figure 18**) (Danko et al. 2015). In addition to epigenetic features, we defined 18 genomic features including: distance of each gene's TSS to the *Xist* locus, the next TAD border or the next full-length LINE element; gene overlap with LADs, LINEs and CpG islands and strength and number of 3D chromatin interactions of gene promoters with other genomic elements. For details see **Supplemental Text 1**.

Random Forest classification models

Two statistical models were developed to distinguish 1) 'silenced' from 'not silenced' genes, referred to as 'XCI/escape model' and 2) 'early' from 'late' silenced genes, referred to as 'silencing dynamics model'. The continuous half-time values were therefore assigned to discrete classes in both models, according to fixed thresholds, which were chosen such that the error rate from the classification model (described below) would be minimized (**Figure 1G, Figure 2G, Supplemental Table 3**). Genes were defined as 'silenced' for $t_{1/2} < 0.9$, as 'not silenced' for $t_{1/2} > 1.6$ and as 'early silenced' for $t_{1/2} < 0.5$, as 'late silenced' for $0.9 < t_{1/2} < 1.3$. The XCI/escape model was trained on 218 genes (168 from the 'silenced' and 50 from the 'not silenced' class) and the silencing dynamics model on 114 genes (74 from the 'early' silenced and 40 from the 'late' silenced class).

The two Random Forest classification models were implemented with the randomForest R package and use 77 predictor variables (epigenetic and genomic features), which show variable degree of correlation between each other (**Supplemental Figure 19**) (R Core Team 2009). The error rate of the models is computed based on the out-of-bag (OOB) error, which is the mean prediction error over all the trees of the Random Forest. Importance of each feature is computed as 'mean decrease in accuracy' (MDA) (**Supplemental Text 2**). Variables with large positive values of the MDA correspond to important features for the classification, while variables with MDA close to zero or negative correspond to noise.

Feature importance (MDA) and classification performance (OOB error) measures were further averaged over a collection of five hundred Random Forests to obtain stable results.

Simple feature selection was performed to improve the model performance by removing weaker or redundant features, which potentially introduce noise. We retain only the top 10 features from each class of the XCI/escape model and the top 8 features from each class of the silencing dynamics model which yield the model minimal error rate (**Supplemental Text 2**). For both models, the classification performance on the top features is reported as average of five hundred Random Forests.

Given the trained XCI/escape model, we predicted the silencing class of all X-linked genes, which were not included in the training set and chose few genes for experimental validation with pyrosequencing (**Supplemental Text 2**).

Forest-guided clustering for model interpretation

We can extract a *proximity matrix* from the trained Random Forest, which is a rough estimate of the distance between genes based on the proportion of times the genes are found in the same leaf node of a tree (**Figure 3C, Supplemental Text 3**).

Based on this *proximity matrix*, genes were grouped into 3 clusters with the *k*-medoids algorithm for both models (Reynolds et al. 2006). Genes of the same silencing class (e.g. not silenced) are largely expected to cluster together according to a certain combination of epigenetic and genomic features. Given the non-linear nature of the classification problem modeled here, we also expect, to some extent, genes from the same silencing class to be grouped in different clusters according to different combinations of features.

Similarly to *k*-means clustering, *k*-medoids clustering requires setting in advance the number of clusters *k* (**Supplemental Text 3**). The results of the *k*-medoids clustering are visualized for both models as heatmaps, displaying the top 10 features which have a significant variation across clusters according to the p-value of an ANOVA test and few others which are also important for classification. Compared to classical Random Forest feature

importance, the outcome of the Forest-guided clustering enables an alternative interpretation of the Random Forest predictions in terms of combinatorial rules which determine the silencing state of groups of genes.

Statistical tests

Kolmogorov-Smirnov (KS) test was performed to test differences in silencing dynamics between A-repeat dependent and A-repeat independent genes. A Fisher's exact test was performed to test for enrichment of escapees in the silenced class and for A or BC-repeat dependent genes in the clusters of both models. A Wilcoxon Rank Sum Test was used to test for differences in half-times between silenced and not silenced genes from the pyrosequencing experiment and for comparison with mRNA-seq data. An Analysis of Variance (ANOVA) test was performed to find significantly different features across clusters in both models. All statistical tests were performed in R with the *base* statistical functions package.

Data Access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE121144.

The Random Forest pipeline is available as **Supplemental Code** and the scripts for all additional analysis on GitHub, https://github.com/marsicoLab/xist_mediated_gene_silencing.

Acknowledgements

We thank the NGS platform of the Curie Institute sequencing, J. Helmuth and Q. Deletang for the help with data analysis, S. Rocha and A. Bousard for providing RNA-seq of *Xist* mutants associated with a recent preprint. This work was supported by NIH grant GM025332

to JTL; Center for Vertebrate Genomics (Cornell University, Ithaca, NY, USA) stipend to JTL and IJ; Helen Hay Whitney Postdoctoral Fellowship and Rosalind Franklin Fellowship (University of Groningen) to IJ; ERC Advanced Investigator award (ERC-2014-AdG no. 671027), Labelisation La Ligue, FRM (grant DEI20151234398), ANR DoseX 2017, Labex DEEP (ANR-11-LBX-0044), part of the IDEX Idex PSL (ANR-10-IDEX-0001-02 PSL), and ABS4NGS (ANR-11-BINF-0001) funding to EH; ARC Foundation for Cancer Research (France; post-doctoral fellowship and grant R17016KK) and the ATIP-AVENIR program (INSERM / CNRS, France; grant C16002KS) to JC; IMPRS for Computational Biology and Scientific Computing funding to LBAS; HFSP long-term fellowship (LT000597/2010-L), Max-Planck Research Group Leader program and E:bio Module III - Xnet grant (BMBF 031L0072) to EGS.

Author contributions

JTL and EH conceived the PRO-seq experiment and strategy; AM, EGS and EH conceived the main idea of the computational analysis; EGS, IJ and CJC performed the PRO-seq experiment and preprocessing of the data; EGS generated the mESC; AM, EGS and LBAS conducted the investigation; LBAS and AM performed statistical analysis and data modeling; JC and CP generated the mRNA-seq data, ID and BF performed the Pyrosequencing experiment. LBAS, AM and EGS wrote the original draft; IJ, JTL and EH reviewed and edited the paper; JTL and EH financed the study; JTL, EH and AM provided resources.

Disclosure Declaration

The authors declare no competing financial interests.

References

- Andergassen D, Dotter CP, Wenzel D, Sigl V, Bammer PC, Muckenhuber M, Mayer D, Kulinski TM, Theussl H-C, Penninger JM, et al. 2017. Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression. *elife* **6**.
- Berletch JB, Ma W, Yang F, Shendure J, Noble WS, Disteché CM, Deng X. 2015. Escape from X inactivation varies in mouse tissues. *PLoS Genet* **11**: e1005079.
- Berletch JB, Yang F, Xu J, Carrel L, Disteché CM. 2011. Genes that escape from X inactivation. *Hum Genet* **130**: 237–245.
- Bianchi I, Lleo A, Gershwin ME, Invernizzi P. 2012. The X chromosome and immune associated genes. *J Autoimmun* **38**: J187-92.
- Borensztein M, Syx L, Ancelin K, Diabangouaya P, Picard C, Liu T, Liang J-B, Vassilev I, Galupa R, Servant N, et al. 2017. Xist-dependent imprinted X inactivation and the early developmental consequences of its failure. *Nat Struct Mol Biol* **24**: 226–233.
- Bousard A, Raposo AC, Zylitz JJ, Picard C, Pires VB, Qi Y, Syx L, Chang HY, Heard E, da Rocha ST. 2018. Exploring the role of Polycomb recruitment in Xist-mediated silencing of the X chromosome in ES cells. *BioRxiv*.
- Brockdorff N. 2017. Polycomb complexes in X chromosome inactivation. *Philos Trans R Soc Lond B Biol Sci* **372**.
- Calabrese JM, Sun W, Song L, Mugford JW, Williams L, Yee D, Starmer J, Mieczkowski P, Crawford GE, Magnuson T. 2012. Site-specific silencing of regulatory elements as a mechanism of X inactivation. *Cell* **151**: 951–963.
- Chaumeil J, Le Baccon P, Wutz A, Heard E. 2006. A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev* **20**: 2223–2237.
- Chaumeil J, Okamoto I, Guggiari M, Heard E. 2002. Integrated kinetics of X chromosome inactivation in differentiating embryonic stem cells. *Cytogenet Genome Res* **99**: 75–84.
- Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, Attreed M, Avner P, Wutz A, Barillot E, et al. 2010. LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**: 956–969.
- Chu C, Zhang QC, da Rocha ST, Flynn RA, Bharadwaj M, Calabrese JM, Magnuson T, Heard E, Chang HY. 2015. Systematic discovery of Xist RNA binding proteins. *Cell* **161**:

404–416.

- Colognori D, Sunwoo H, Kriz AJ, Wang C-Y, Lee JT. 2019. Xist Deletional Analysis Reveals an Interdependency between Xist RNA and Polycomb Complexes for Spreading along the Inactive X. *Mol Cell* **74**: 101–117.e10.
- Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, Cheung VG, Kraus WL, Lis JT, Siepel A. 2015. Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods* **12**: 433–438.
- de Napoles M, Mermoud JE, Wakao R, Tang YA, Endoh M, Appanah R, Nesterova TB, Silva J, Otte AP, Vidal M, et al. 2004. Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. *Dev Cell* **7**: 663–676.
- Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES, et al. 2013. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**: 1237973.
- Escamilla-Del-Arenal M, da Rocha ST, Heard E. 2011. Evolutionary diversity and developmental regulation of X-chromosome inactivation. *Hum Genet* **130**: 307–327.
- Filippova GN, Cheng MK, Moore JM, Truong J-P, Hu YJ, Nguyen DK, Tsuchiya KD, Disteche CM. 2005. Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev Cell* **8**: 31–42.
- Galupa R, Heard E. 2015. X-chromosome inactivation: new insights into cis and trans regulation. *Curr Opin Genet Dev* **31**: 57–66.
- Gendrel A-V, Heard E. 2014. Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation. *Annu Rev Cell Dev Biol* **30**: 561–580.
- Helmuth J, Li N, Arrigoni L, Gianmoena K, Cadenas C, Gasparoni G, Sinha A, Rosenstiel P, Walter J, Hengstler JG, et al. 2016. normR: Regime enrichment calling for ChIP-seq data. *BioRxiv*.
- Janiszewski A, Talon I, Song J, De Geest N, To SK, Bervoets G, Marine J-C, Rambow F, Pasque V. 2019. Dynamic Erasure of Random X-Chromosome Inactivation during iPSC Reprogramming. *BioRxiv*.
- Kelsey AD, Yang C, Leung D, Minks J, Dixon-McDougall T, Baldry SEL, Bogutz AB, Lefebvre L, Brown CJ. 2015. Impact of flanking chromosomal sequences on localization and silencing by the human non-coding RNA XIST. *Genome Biol* **16**: 208.
- Kinkley S, Helmuth J, Polansky JK, Dunkel I, Gasparoni G, Fröhler S, Chen W, Walter J, Hamann A, Chung H-R. 2016. reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4(+) memory T cells. *Nat Commun* **7**: 12514.
- Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, et al. 2007. The UCSC genome browser database: update 2007. *Nucleic Acids Res* **35**: D668–73.
- Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**: 950–953.
- Loda A, Brandsma JH, Vassilev I, Servant N, Loos F, Amirasr A, Splinter E, Barillot E, Poot RA, 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.
- Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, Waters CT, Munson K, Core LJ, Lis JT. 2016. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat Protoc* **11**: 1455–1476.
- Marks H, Kerstens HHD, Barakat TS, Splinter E, Dirks RAM, van Mierlo G, Joshi O, Wang S-Y, Babak T, Albers CA, et al. 2015. Dynamics of gene silencing during X inactivation using allele-specific RNA-seq. *Genome Biol* **16**: 149.
- McHugh CA, Chen C-K, Chow A, Surka CF, Tran C, McDonel P, Pandya-Jones A, Blanco M, Burghard C, Moradian A, et al. 2015. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**: 232–236.
- Monfort A, Di Minin G, Postlmayr A, Freimann R, Arieti F, Thore S, Wutz A. 2015. Identification of Spen as a Crucial Factor for Xist Function through Forward Genetic Screening in Haploid Embryonic Stem Cells. *Cell Rep* **12**: 554–561.
- Pintacuda G, Wei G, Roustan C, Kirmizitas BA, Solcan N, Cerase A, Castello A, Mohammed S, Moindrot B, Nesterova TB, et al. 2017. hnRNPK Recruits PCGF3/5-PRC1 to the Xist RNA B-Repeat to Establish Polycomb-Mediated Chromosomal Silencing. *Mol Cell* **68**: 955-969.e10.
- Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, de la Cruz CC, Otte AP, Panning B, Zhang Y. 2003. Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300**: 131–135.
- Plath K, Talbot D, Hamer KM, Otte AP, Yang TP, Jaenisch R, Panning B. 2004. Developmentally regulated alterations in Polycomb repressive complex 1 proteins on the inactive X chromosome. *J Cell Biol* **167**: 1025–1035.
- Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. 2006. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *J Math Model Algor* **5**: 475–504.
- R Core Team. 2009. R: A Language and Environment for Statistical Computing.
- Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**: W187-91.
- Sakata Y, Nagao K, Hoki Y, Sasaki H, Obuse C, Sado T. 2017. Defects in dosage compensation impact global gene regulation in the mouse trophoblast. *Development* **144**: 2784–2797.
- Schulz EG, Meisig J, Nakamura T, Okamoto I, Sieber A, Picard C, Borensztein M, Saitou M, Blüthgen N, Heard E. 2014. The two active X chromosomes in female ESCs block exit from the pluripotent state by modulating the ESC signaling network. *Cell Stem Cell* **14**: 203–216.
- Silva J, Mak W, Zvetkova I, Appanah R, Nesterova TB, Webster Z, Peters AHFM, Jenuwein T, Otte AP, Brockdorff N. 2003. Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Dev Cell* **4**: 481–495.
- Simon MD, Pinter SF, Fang R, Sarma K, Rutenberg-Schoenberg M, Bowman SK, Kesner

- BA, Maier VK, Kingston RE, Lee JT. 2013. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature* **504**: 465–469.
- Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJG, Zhu Y, Kaaij LJT, van Ijcken W, Gribnau J, Heard E, et al. 2011. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev* **25**: 1371–1383.
- van Steensel B, Belmont AS. 2017. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**: 780–791.
- Wutz A, Rasmussen TP, Jaenisch R. 2002. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet* **30**: 167–174.
- Wu H, Luo J, Yu H, Rattner A, Mo A, Wang Y, Smallwood PM, Erlanger B, Wheelan SJ, Nathans J. 2014. Cellular resolution maps of X chromosome inactivation: implications for neural development, function, and disease. *Neuron* **81**: 103–119.
- Yang F, Babak T, Shendure J, Disteche CM. 2010. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res* **20**: 614–622.
- Żylicz JJ, Bousard A, Žumer K, Dossin F, Mohammad E, da Rocha ST, Schwalb B, Syx L, Dingli F, Loew D, et al. 2019. The implication of early chromatin changes in X chromosome inactivation. *Cell* **176**: 182-197.e23.

Table 1: Epigenetic and genomic features used for modeling.

epigenetic features	sequence-specific transcription factors	MYC, ESRRB, KLF4, MAFK, NANOG, MYCN, OCT4, SOX2, TCF3, TCF3P2L1, YY1, ZNF384
	structural proteins	CTCF, SMC1, SMC3
	general transcription regulators	CDK9, E2F1, HCFC1, MAX, MED1, MED12, NIPBL, RNAPII (S2P, S5P, S7P, unphosphorylated), SIN3A, TAF1, TAF3, TBP
	chromatin modifications (activation)	H3K27ac, H3K9ac, H3K4me1, H3K4me3, H3K36me3, H3K79me2, KMT2B/MLL2
	chromatin modifications (repression)	H2AK119ub1, H3K27me3, RING1B (PRC1), CBX7 (PRC1) RYBP (PRC1), KMT6/EZH2 (PRC2), SUZ12 (PRC2), KDM1A/LSD1, KDM2A, KDM2B, HDAC1, HDAC2, HDAC3, DNA methylation (BS-seq), 5fC (MeDip), 5hmC (MeDip), TET1
	others	H2A.Z, OGT, BRG1, CBX3
genomic features	genomic elements	distance to the <i>Xist</i> locus, TAD borders, LADs, full-length LINEs overlap with Xist entry sites, LADs, CpG islands full-length LINE density (700kb) gene density (1Mb) CpG content

	3D structure	number Hi-C all, strength Hi-C all, number Hi-C promoter, strength Hi-C promoter, strength Hi-C <i>Xist</i> , number HiCap promoter, number HiCap enhancer, number HiCap all
--	--------------	--

Figure Legends:

Figure 1: Measuring gene silencing dynamics. (A) Schematic of the experimental setup used in B-F. Using a hybrid female mESC line (*B6xCast*) carrying a Dox-responsive promoter in front of the endogenous *Xist* gene on the *B6* allele Pol II activity was measured by allele-specific PRO-seq over a 24h time course of Dox treatment. (B) Strand-specific read density at the *Tsix-Xist* locus. Plus-strand is shown in red, minus strand in blue; the y-axis indicates reads per million. (C) *Xist* expression from the two alleles. (D) Distribution of the fraction of *B6* reads for autosomal and X-linked genes over time. (E-F) Schematic (E) and three examples (F) of how gene silencing half-times (in parentheses) were estimated from the allele-specific PRO-seq time course data through fitting an exponential decay function. (G) Distribution of estimated half-times for 280 X-linked genes with an assigned active TSS. The half-time ranges used to define the model classes, as well as the number of genes falling in each category are indicated.

Figure 2: Comparison of PRO-seq based silencing half-times to other data sets. (A) Comparison of PRO-seq (undifferentiated mESC, upper panel) and mRNA-seq data (undifferentiated mESCs, middle panel and differentiated mESCs, lower panel). Fraction of *B6* reads are shown for all genes covered in all three datasets, ordered by genomic position. (B - D) Comparison of estimated half-times (in days) between the data sets shown in (A) with fitted regression lines (red). Pearson correlation coefficient and significance (p-value)

are indicated. **(E - G)** Distribution of half-times within silencing classes defined previously in mESCs (E) (Marks et al. 2015), in pre-implantation mouse embryos (F) (Borensztein et al. 2017) and the classes used for Random Forest modelling (G) (blue: XCI/escape model; red: silencing dynamics model). **(H)** Estimated half-times (black circles) for all genes in the PRO-Seq data set along the X Chromosome. A fitted smooth curve of the half-times is displayed as blue line and the *Xist* locus is marked with a grey line. **(I)** Cumulative distribution of half-times of genes silenced (independent, light grey) or not silenced (dependent, dark grey) by *Xist* lacking the repeat A element (Sakata et al. 2017).

Figure 3: Schematic overview of our modeling approach. **(A)** Epigenetic and genomic input data for the model are collected and feature matrices are computed for all X-linked genes with estimated half-times (labeled) and without estimated half-times (unlabeled). **(B)** After model training, the XCI/escape model is then used to predict the silencing class of all unlabeled X-linked genes given the same set of input features. The predictions are validated by comparing them to measured half-times from undifferentiated mRNA-seq data, with pyrosequencing experiments (few selected genes) and with measured silencing dynamics of genes in six transgenic mESCs clones. **(C)** A forest-guided clustering approach was developed for model interpretation. A proximity matrix between genes is computed from the trained model and converted into a distance matrix. Clusters of genes and their most significant associated features are displayed as a heatmap.

Figure 4: Feature importance for XCI/escape and silencing dynamics model. For each model, features are ranked class-wise according to their importance for the classification, quantified by the Mean Decrease in Accuracy (MDA) (see Methods section). The top features of each class (10 for XCI/escape model; 8 for silencing dynamics model) which are used to build the final model are marked with *. For more details refer to **Supplemental Figure 3**. Similar results are obtained from the XCI/escape model trained on undifferentiated mRNA-seq data (**Supplemental Text 8, Supplemental Figure 20**).

Figure 5: Classification rules for the XCI/escape model. (A) Results from the forest-guided clustering of the XCI/escape model visualized as a heatmap. Columns indicate the genes grouped by cluster, rows correspond to features with significant differences among clusters (ANOVA test). The top 10 most significant features from the ANOVA test are marked with §. Distributions of features in each cluster are shown in the box plots next to the heatmap. (B) Schematic view of the feature combinations promoting gene silencing (cluster 1 and 2) or escape (cluster 3). (C) Silencing half-time distribution in each cluster. (D) Proportion of genes in each cluster that undergo silencing in mouse trophoblasts, independent or dependent of the *Xist* repeat A element (Sakata et al. 2017). ‘Repeat A dependent genes’ indicates genes with abrogated silencing in *Xist*-repeat A-mutant-cells, while ‘repeat A independent genes’ indicates genes which still undergo silencing in the same cells and ‘not covered’ refers to our genes which were not covered in that data set. The numbers in each box indicate the number of genes that fall into each category. Similar results are obtained from the XCI/escape model trained on undifferentiated mRNA-seq data (Supplemental Text 8, Supplemental Figure 20).

Figure 6: Classification rules for the silencing dynamics model. (A) Results from the forest-guided clustering of the silencing dynamics model visualized as a heatmap. Columns indicate the genes grouped by cluster, rows correspond to features with significant differences among clusters (ANOVA test). The top 10 most significant features from the ANOVA test are marked with §. Differences in the distributions of features between clusters are highlighted in the box plots next to the heatmap. (B) Schematic view of the features associated with early (cluster 1 and 2) and late gene silencing (cluster 3). (C) Silencing half-time distribution for each cluster. (D) The proportion of genes which undergo silencing in mouse trophoblasts, independent or dependent of the *Xist* repeat A element, is shown for each cluster, similar to **Figure 5D**. The numbers in each box indicate the number of genes that fall into each category.

Figure 7: Experimental validation of model predictions. (A) Half-times of 6 candidate genes predicted as 'silenced' (top) and 5 candidate genes predicted as 'not silenced' (bottom) were estimated experimentally through allele-specific quantification by pyrosequencing at different time points during 24h of doxycycline treatment in TX1072 cells, in 3 independent experiments. Individual data points (dots), the fitted exponential decay function (line) and the estimated silencing half-times are shown. (B) Dotplot of the silencing half-times ($t_{1/2}$) estimated in (A). (C) Dotplot of undifferentiated mRNA-seq half-times for genes predicted as silenced and not silenced by our XCI/escape model. The gray line in (B-C) indicates the mean and the p-value (Wilcoxon Rank Sum Test) indicates a significant difference between the mean of the two distributions. (D) Fraction of genes correctly predicted as silenced by the XCI/escape model (red lines) for 6 cell lines where an inducible *Xist* transgene was integrated in different chromosomal locations (orange, cartoon on the right, **Supplemental Table 5**). The background distributions of silenced predictions used to estimate empirical p-values is also shown (histogram, mean black dashed line).

