Transcriptome regulation during the X chromosome inactivation process

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Guido Pacini

Berlin, 2023

Erstgutachterin: Prof. Dr. Annalisa Marsico

Zweitgutachter: Prof. Dr. Mark Friedländer

Tag der Disputation: 05/07/2024

Declaration of authorship

Surname: Pacini
Name: Guido

I declare to the Freie Universität Berlin that I have completed the submitted dissertation independently and without the use of sources and aids other than those indicated. The present thesis is free of plagiarism. I have marked as such all statements that are taken literally or in content from other writings. This dissertation has not been submitted in the same or similar form in any previous doctoral procedure.

I agree to have my thesis examined by a plagiarism examination software.

Date: 12/12/2023

Signature:

_____

Abstract

In mammals, female cells achieve dosage compensation between the sexes randomly chosing and transcriptionally silencing one of the two X chromosomes through a process known as X-chromosome inactivation (XCI). This process is initiated during early development through up-regulation of the long non-coding RNA *Xist*, which mediates chromosome-wide gene silencing of the future inactive chromosome (Xi) in cis. Upon completion of the XCI process Xi will maintain its silenced state in all daughter cells, which results in the genetic mosaicism of female organisms. Cell differentiation, *Xist* up-regulation and gene silencing are thought to be coupled at multiple levels to ensure inactivation of exactly one out of two X chromosomes.

In this thesis I performed an integrated analysis of all three processes through the analysis of allele-specific single-cell RNA-sequencing data. Specifically, I investigated the endogenous random XCI process in hybrid mouse embryonic stem cells at different time points throughout cellular differentiation developing dedicated analysis approaches that rely on the high number of polymorphisms between the two parental strains.

Putative *Xist* regulators were identified exploiting the inter-cellular heterogeneity of XCI onset. A large fraction of cells transiently expressed *Xist* on both X chromosomes which resulted in biallelic gene silencing right before being resolved to a monoallelic state, confirming a prediction of the stochastic model of XCI. The two X chromosomes showed different gene silencing dynamics, and a number of strain-specific *escapees* (namely, genes that escape transcriptional silencing) were identified and experimentally validated. These results suggest that genetic variation modulates the XCI process at multiple levels, providing a potential explanation for the long-known X-controlling element (Xce) effect, which leads to preferential inactivation of a specific X chromosome in inter-strain crosses.

Overall, this work provides a detailed picture of the different levels of regulation that govern both *Xist* up-regulation and the initiation of XCI.

Zusammenfassung

In Säugetieren erreichen weibliche Zellen einen Dosisausgleich zwischen den Geschlechtern, indem sie eines der beiden X-Chromosomen zufällig auswählen und durch einen als X-Chromosomen-Inaktivierung (XCI) bekannten Prozess transkriptionell ausschalten. Dieser Prozess wird während der frühen Entwicklung durch die Hochregulierung der langen nichtkodierenden RNA *Xist* eingeleitet. Die *Xist* RNA vermittelt das Ausschalten des kompletten zukünftig inaktiven X-Chromosoms (Xi) in cis. Auch nach dem Abschluss des XCI-Prozesses bleibt das Xi in allen Tochterzellen ausgeschaltet, weshalb weibliche Individuen genetische Mosaike sind. Zelldifferenzierung, *Xist* Hochregulierung und Gen-Silencing sind vermutlich auf mehreren Ebenen miteinander gekoppelt, um die Inaktivierung genau eines der beiden X-Chromosomen sicherzustellen.

Diese Arbeit analysiert allelspezifische Einzelzell-RNA-Sequenzierungsdaten anhand derer eine integrierte Analyse aller drei Prozesse durchgeführt wird. Insbesondere untersucht sie den endogenen zufälligen XCI-Prozess in hybriden embryonalen Stammzellen von Mäusen zu verschiedenen Zeitpunkten während der zellulären Differenzierung. Hierzu werden spezielle Analyseansätze entwickelt, welche auf der großen Zahl an Polymorphismen zwischen den beiden Elternstämmen aufbauen.

Durch Ausnutzung der intrazellulären Heterogenität zu Beginn des XCI Prozesse können außerdem potenzielle *Xist* Regulatoren identifiziert werden. Ein großer Teil der Zellen exprimiert *Xist* zeitweise von beiden X-Chromosomen, und beginnt deshalb mit der Abschaltung beider X-Chromosomen. Kurze Zeit später wird dieser biallelische *Xist* Expressionszustand zu monoallelischer Expression aufgelöst, was eine zentrale Vorhersage des stochastischen XCI Modells bestätigt. Die beiden X-Chromosomen zeigen außerdem unterschiedliche Dynamiken der Genabschaltung, und es werden eine Reihe von stamm-spezifischen *Escapees* (Gene, die der Ausschaltung entkommen) identifiziert und experimentell validiert. Diese Ergebnisse legen nahe, dass genetische Variation den XCI-Prozess auf mehreren Ebenen moduliert. Dies könnte eine potenzielle Erklärung für den seit langem bekannten X-Kontrollelement (Xce)-Effekt sein, der zur bevorzugten Inaktivierung eines bestimmten X-Chromosoms in Kreuzungen verschiedener Stämme führt.

Insgesamt liefert diese Arbeit ein detailliertes Bild der verschiedenen Regulationsniveaus, die sowohl die Hochregulierung von *Xist* als auch den Beginn der XCI steuern.

Acknowledgements

I would like to express my gratitude to the following people for all their support throughout this project.

First and foremost I wish to thank both my supervisors. I really would like to thank Dr. Edda Schulz for funding this project, for hosting me in her lab, and for all the time and counseling that she dedicated in supervising my PhD project. In the same way I would like to thank Prof. Dr. Annalisa Marsico for motivating my interest in bioinformatics and data analysis since my first days in Berlin, and for her endless support throughout my research. I really would like to thank both of them, beyond their patience in supervising and reviewing this work, they contributed to my growth both as a researcher and as a person.

Moreover I would like to dedicate a few sentences to the wonderful people I had the pleasure to work with. First I would like to thank Ilona, the beating heart of the lab, not only for all the time and patience that she spent on the wet lab work which represents the milestone of this project, but also for the enthusiasm and happiness that she brought to the lab on a daily basis. I would like to thank my office mates Zeba and Rutger for all the time that we spent together, for all the laughter and for everything that they taught me. In the same way I would like to thank my distal-enhancers Oriana and Liat for all their support throughout this project. Finally I would like to thank Verena. During the years I spent in Berlin we got to know each other, and I enjoyed every single day that we spent together. She is a wonderful person, a caring friend and an amazing scientist. She never left me alone and did her best in supporting me even at the most difficult of times. I consider myself extremely lucky to have spent so much time with these wonderful people, who eventually became close friends.

All this work would have not been possible without the endless support of my whole family, together with Pina, Caterina, Giacomo, Francesco, Giacomino, Eugenio, Iacopo, Luigi and many other close friends. I can't express how grateful I am to each of them. Especially my father taught me to never give up and to always do my best, my mother taught me the importance of listening and learning, my sister taught me the beauty of never stop searching for answers to difficult questions, and my grandmother taught me that every little step forward is part of the journey and ultimately leads to the arrival.

This entire work is dedicated to the memory of Galileo and Fiorella.

Contents

1 Introduction

1 Historical perspective

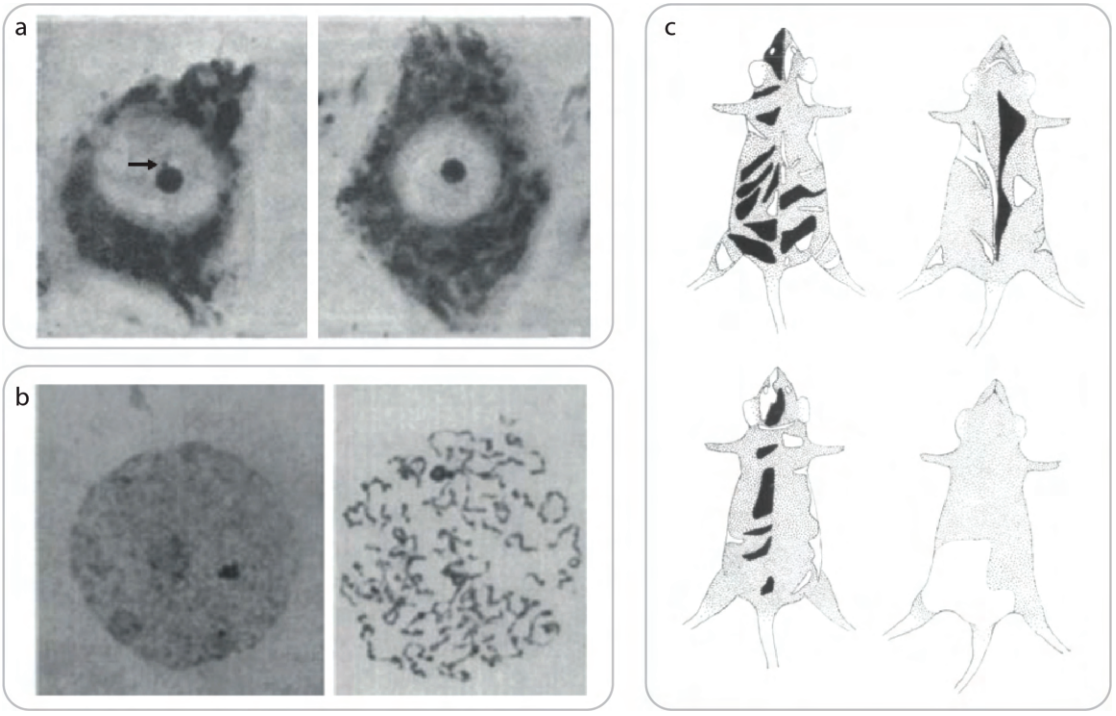


FIGURE 1: (a) Neuronal nucleus of a mature female cat (left) and a mature male cat (right). The arrow indicates the nucleolar satellite [9]. (b) Female nuclei of rat liver cells on interphase (left) and early prophase (right) [139]. (c) Dorsal and ventral sides of two dappled female mice [113].

Mary Lyon's postulation of X chromosome inactivation (XCI) as the key process that mammalian organisms use to achieve dosage compensation between the sexes was suggested by a number of previous discoveries. In 1928 Emil Heitz provided a broad classification of the genomic content into loose and condensed chromatin, respectively referred to as euchromatin and heterochromatin. Through cytological studies he characterized the latter as densely stained chromatin representing the unaccessible and inactive portion of the genome, while the former as the gene active counterpart [85]. The first evidence leading to Lyon's hypothesis can be attributed to the work of Murray Barr and Ewart Bertam [9]. In 1949 they observed that the neuronal nuclei of female cats were characterized by a nucleolar satellite located adjacent to the nucleolus, which was absent in males (Fig. 1a). This unit, which will later be referred to as *Barr body*, was hypothesized to represent a sex chromosome. In 1960 Susumu Ohno and others observed a similar heteropyknotic unit in neoplastic and normal diploid female cells of mouse and rat, which resembled a single X chromosome in both length and heterochromatic content (Fig. 1b) [141]. In 1961 Mary Lyon hypothesized that the Barr body characterizing female mammalian cells was the result of the random inactivation of either the paternal (Xp) or maternal (Xm) X chromosomes at the early stages of embryonic development [111].

This silenced state was then propagated through subsequent cellular divisions leading to cellular mosaicism of X-linked genes in female cells. This conclusion was supported by the mosaic phenotype observed in female mice carrying heterozygous sex linked genes (Fig. 1c), and by the normal development of female mice lacking a second X chromosome (XO) [113]. A further classification of heterochromatin was provided by Spencer Brown in 1966 [23]. Brown defined as *facultative* the chromatin which could take either a heterochromatic or euchromatic state across different parental chromosomes, cell types or developmental stages, and as *constitutive* the chromatin which always maintains its heterochromatic form. In mammals the X chromosome represents one of the clearest examples of facultative heterochromatin. Where the heterochromatic X chromosome is characterized by a condensed shape and late replication timing [82, 112].

2 Dosage compensation mechanisms across species

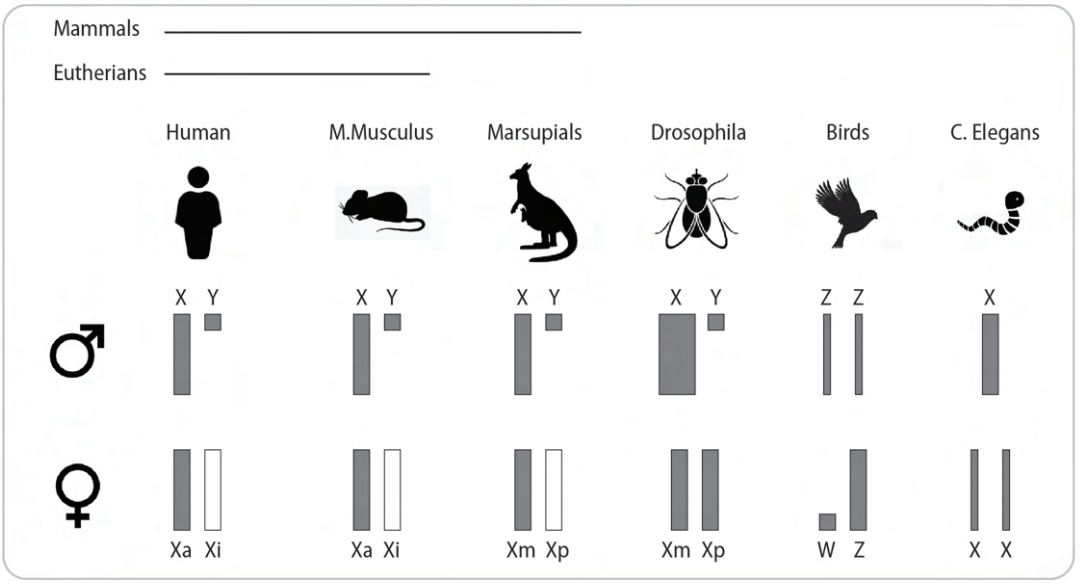


FIGURE 2: Sex chromosome dosage compensation mechanisms across species.

In mammalian diploid organisms, males represent the heterogametic sex (XY) and females the homogametic sex (XX). Present day X and Y chromosomes evolved by a combination of decay and differentiation of an ancestral homolog pair of autosomes, where the absence of recombination between the two chromosomes led to the progressive accumulation of mutations, rapid degradation and loss of many genes on the heterogametic Y-chromosome (Muller's ratchet theory). The X chromosome carries a much larger number of transcribed genes than the Y-chromosome, and is enriched in genes associated to sexual reproduction and brain functions. On the other hand, the Y-chromosome is enriched of male advantageous genes around the testis determining gene (*SRY*) and of several testis expressed genes which are essential for male fertility [54, 67, 84, 126].

Dosage compensation mechanisms aim to balance the unpaired gene expression levels between the X and Y sex chromosomes, and between these and the autosomes. X-linked genes are thought to have evolved higher expression levels than their ancestral counterparts to compensate for the two-fold reduction of X chromosome gene expression in males, and to balance their expression with respect to the autosomal pairs. In mammalian organisms the female specific XCI process balances the two-fold difference in gene expression levels between male and female organisms at the earliest stages of embryonic development. In mice, every female mouse embryonic cell undergoes two waves of chromosome-wide gene silencing throughout its early developmental stages. First, every cell composing the mouse embryo transcriptionally silences Xp at the 4-8 cell stage, through a process known as imprinted XCI (iXCI). Xp will retain its inactive state in the trophectoderm, while it will restore its gene activity in the inner cell mass (ICM) by the epiblast stage. Every cell composing the ICM will then undergo random XCI (rXCI), which leads to the silencing of either Xp or Xm. Once established, the inactive X chromosome (Xi) will irreversibly maintain its silenced state throughout cellular divisions in all daughter cells, leading to phenotypic chimaerism of X-linked genes in female organisms. On the other hand, female human embryonic cells do not undergo iXCI, and maintain both X chromosomes transcriptionally active until the occurrence of rXCI at the late blastocyst stage. Distantly related species evolved different mechanisms to compensate for the difference in genetic content between the sex chromosomes [67, 93, 108, 142, 147, 193, 203].

While eutherian mammals randomly select and inactivate either Xp or Xm independently in every cell composing the female embryo through the XCI process, marsupials specifically silence the paternal X chromosome, although this silencing process was reported as quite leaky and unstable [77, 94, 115]. Differently from mammalian organisms, Drosophila achieve dosage compensation between the sexes through the two-fold up-regulation of the X-linked genes in males, while females maintain both X chromosomes active [46, 70]. In birds, where females represent the heterogametic sex (ZW) and males the homogametic sex (ZZ), the lack of a global dosage compensation mechanism between the sexes results in significantly higher expression levels in males' sex chromosomes compared to their female counterpart [11, 89, 93]. In Caenorhabditis Elegans, where the sex is determined by the number of X chromosome copies, the difference in gene expression levels between the sexes is compensated down-regulating the expression of both X chromosomes in hermaphrodites (XX) [54].

For the aims of this project, the following sections provide a description of the random XCI process in Mus Musculus, with a particular focus on mouse embryonic stem cells (mESCs). This is an in vitro system derived from pre-implantation blastocysts from the inner cell mass of E3.5 mouse embryos, which are undifferentiated cells characterized by the expression of pluripotency-associated factors such as *Nanog* and *Oct3/4* (*Pou5f1*)

[18]. This in vitro system has been extensively used to study the murine random XCI process upon induced cellular differentiation.

3 The X-inactivation center (Xic)

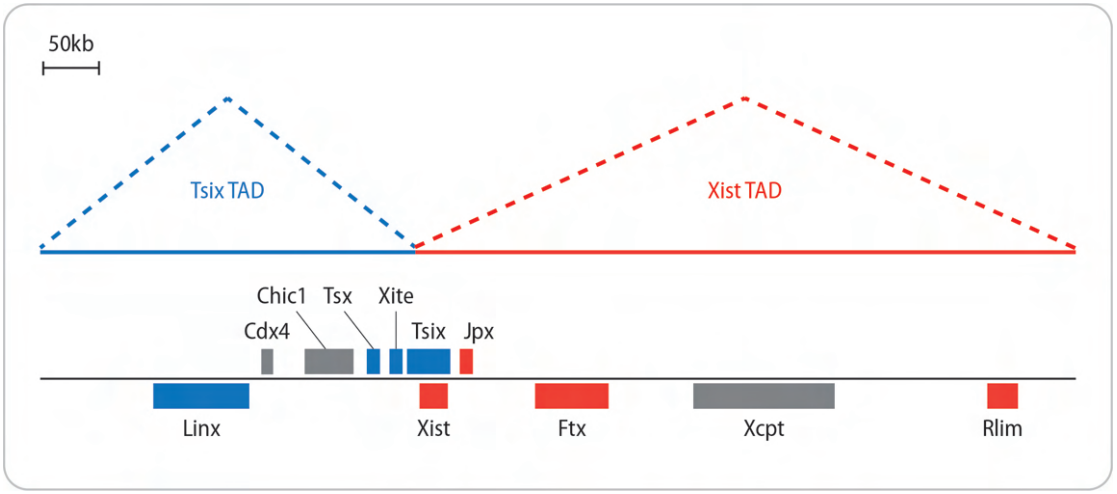


FIGURE 3: Schematic representation of the X inactivation center (Xic) and of *Xist* positive (red) and negative (blue) *cis* regulators lying within this locus.

In mammals the XCI process is controlled by a master regulatory locus named *X inactivation center* (Xic), a minimal genetic region on the X chromosome which is necessary and sufficient to initiate random XCI when present in two copies [5, 67, 152]. The Xic locus is relatively GC-poor, enriched in repetitive regions, and shows little sequence conservation across mammalian species. Many of the genes stored within the Xic code for untranslated RNAs (*Ftx*, *Jpx*, *Linx*, *Tsix*, *Xist*, *Xite*), while others are specifically expressed in the testis (*Tsx*, *Ppnx*) [5, 21, 67, 152, 153].

The Xic locus in human and mice harbors the *Xist* (Xi-specific transcript) gene, which is exclusively expressed by the inactive X chromosome (Xi) in somatic cells and encodes an alternatively spliced long non-coding RNA (lncRNA) 17kb-long transcript which is retained in the nucleus. Embryonic cells characterized by two transcriptionally active X chromosomes, hence bi-allelic expression of *Xist*, undergo the random XCI process. This is initially triggered by the up-regulation of *Xist* on the future inactive allele (Xi), with the *Xist* lncRNA localizing within the nucleus and gradually coating the X chromosome in *cis*, which then mediates its transcriptional silencing [20, 22, 42, 128, 181, 201].

Xist RNA creates a repressed nuclear compartment depleted of transcriptional machinery on the X chromosome from which it is expressed, which recruits and transcriptionally silences genes in cis. *Xist* RNA is necessary to induce chromosome wide gene silencing, indeed its heterozygous deletion results in non-random silencing of the wild type allele,

while experiments employing *Xist* cDNA transgenes on autosomes demonstrated that *Xist* RNA is sufficient to transcriptionally repress autosomal genes during early developmental stages [101, 146, 198].

The *Xist* locus shows little sequence conservation across mammalian organisms, however it is characterized by highly conserved repetitive regions which are present at different copy numbers across eutherian mammals. Among these repeat rich regions (Rep A-F) the RepA sequence, localized at the 5' end of *Xist*, is the most conserved across different species and plays a crucial role for *Xist* silencing function. Indeed its deletion results in *Xist* RNA molecules which are still able to associate with chromatin and spread over the chromosome, while being unable to trigger its transcriptional repression [132, 199].

Both prior and after cellular differentiation, the Xic of each X chromosome is partitioned into two topologically associated domains (TADs). Notably the *Xist* gene is located at the boundary between these two regions of increased gene-to-gene interaction, and separates the regions upstream and downstream of *Xist* which respectively harbor most of its positive and negative regulators [136].

3.1 Positive *Xist* regulators

One of the most extensively studied *Xist* activators encoded within the Xic is the E3 ubiquitin ligase *Rnf12*, which is also referred to as *Rlim* [92]. *Rlim* plays a crucial role in the imprinted XCI (iXCI) process, indeed its knockout in oocytes results in a defective iXCI process which leads to embryo death. On the other hand *Rlim* is dispensable for rXCI, indeed mouse embryonic stem cells lacking *Rlim* are still capable of forming Xist clouds and of silencing some genes within the X chromosome during rXCI [173, 174]. *Rlim* is expressed at higher levels in females than in male cells, it is up-regulated in differentiating mouse embryonic stem cells and acts as a trans-activator of Xist transcription through the ubiquitination and proteasomal degradation of the pluripotency factor *Rex1*, which acts as an Xist repressor [7, 76, 131]. Previous studies showed that the enhanced expression of *Rlim* in mouse ESCs causes the ectopic initiation of the XCI process on the single X chromosome of male cells, and in both X chromosomes of female cells. On the other hand, the heterozygous deletion of *Rlim* in differentiating female mouse ES cells delays the initiation of the rXCI process [92].

Rlim was reported to activate *Xist* expression acting in concert with two lncRNAs encoded within the *Xist* TAD, namely *Jpx* and *Ftx* [8]. In mice, *Jpx* escapes X-inactivation and activates *Xist* expression at the onset of XCI by binding to and titrating away CTCF, a protein which represses *Xist* transcription, from the *Xist* promoter of a single allele [186]. While its role as an *Xist* activator is well established, it is still controversial whether *Jpx* regulates *Xist* transcription acting in *cis* or *trans* [27, 190]. Notably the

regulatory mechanism of this gene differs across eutherian mammals. Indeed in mice the *Jpx* transcript is necessary to activate *Xist* expression, while in humans the sole act of *JPX* transcription rather than its RNA product activates *XIST* expression in *cis* [164]. *Ftx* (five prime to *Xist*) is a X-linked lncRNA whose deletion results in a considerable decrease of *Xist* RNA transcripts [40]. *Ftx* expression is dispensable for imprinted XCI in preimplantation embryos, while this gene is up-regulated in female mESCs at the onset of random XCI [40, 178]. *Ftx* is required to promote *Xist* expression in *cis*, where its regulatory function depends on the gene's transcription rather than its lncRNA product [65]. Finally, *Xert* is a recently identified lncRNA which mediates *Xist* activation in cis during the random XCI stage where it is up-regulated together with *Jpx* and *Ftx* [75].

3.2 Negative *Xist* regulators

The major negative *Xist* regulator within the Xic locus is its antisense transcription unit, named *Tsix*. This gene, which entirely overlaps the *Xist* locus on the opposite strand, encodes an alternatively spliced long non-coding RNA (lncRNA) 4kb-long nuclear transcript [100]. While undifferentiated cells express this gene on both X chromosomes, *Tsix* becomes monoallelically expressed by the future active allele (Xa) at the onset of the XCI process. The heterozygous deletion of the *Tsix* locus or its promoter leads to non-random *Xist* up-regulation on the mutated allele, which precedes its transcriptional silencing through the XCI process [43, 102]. Notably, the expression of *Xist* does not significantly increase whenever the *Tsix* gene is deleted [125]. The transcription of *Tsix* prevents *Xist* up-regulation rather than repressing its expression, hence playing a crucial role in determining which allele will be silenced. *Tsix* was indeed shown to silence *Xist* through the modification of its chromatin structure, which is mediated by the establishment of a repressive chromatin configuration at the *Xist* promoter [129, 138, 166]. A number of genomic loci within the *Tsix* TAD are tightly associated with *Tsix* expression levels. The following paragraph shortly describes their role in the regulation of *Tsix* activity.

DXPas34 is a 34mer minisatellite which plays a dual role throughout the XCI process. At the onset of XCI it enhances the activity of *Tsix* promoter, while it represses *Tsix* expression once XCI has been established [44, 183]. *DXPas34*'s deletion results in the loss of *Tsix* transcription, which is followed by non-random *Xist* up-regulation and initiation of XCI on the mutated allele [52, 195]. *Xite* (X-inactivation intergenic transcription element) is an untranslated gene which enhances *Tsix* expression. Its deletion or lack of transcription at the onset of XCI results in the down-regulation of *Tsix cis* expression, leading to *Xist cis* up-regulation and non-random XCI of the mutated allele [137, 183]. *Tsx* and *Linx* are two lncRNAs located within the *Tsix* TAD, which are reported to slightly enhance *Tsix* expression. Deletion of *Tsx* prior to cellular differentiation mildly affects the expression of both *Xist* and *Tsix*, while it does not significantly

influence the choice of which X chromosome will be inactivated [3, 49]. *Linx* is an untranslated gene which is generally monoallelically expressed together with *Tsix* in the ICM of mice. The promoter of *Linx* (LinxP) represses the *cis* expression of *Xist*, which is independent of *Linx* transcription or its effects on *Tsix* expression levels [68, 73, 136].

4 Gene silencing mechanism

Differently from other eutherian mammals, mice undergo two waves of XCI. Through the iXCI process every cell composing the mouse embryo specifically silences the paternal X chromosome (Xp) at the 4-8 cell stage, while the other X chromosomes remain transcriptionally active due to the presence of a genetic imprint which is mediated by the deposition of H3K27me3 histon marks preventing the *cis* up-regulation of *Xist*. Every cell composing the ICM of blastocysts then reactivates Xp, and independently silences a single X chromosome through the rXCI process [88, 93].

The following sections describe the *Xist*-mediated gene silencing mechanism employed in the murine rXCI process, which can be broadly divided into four key steps: initiation, spreading, silencing and maintainance.

4.1 Initiation

The random XCI process initiates with the up-regulation of the *Xist* gene on the future inactive X chromosome, Xi.

The expression level of *Xist* is tightly linked to that of core pluripotency trans-acting factors such as *Oct4*, *Nanog* and *Sox2* which bind to a region within the first intron of *Xist*. While the depletion of these factors in undifferentiated cells leads to the ectopic up-regulation of *Xist* in male embryos, which would otherwise be expressed at very low levels in both male and female embryos, the deletion of their *Xist* binding site does not considerably affect *Xist* expression levels. This suggests that *Oct4*, *Nanog* and *Sox2* might indirectly regulate *Xist* expression levels. Notably, while the above core pluripotency factors maintain their expression levels unchanged upon *Xist* up-regulation, other stem cell factors such as *Rex1* and *Prdm14* are strongly down-regulated. Finally a recent work demonstrated that *GATA* transcription factors play a crucial role in the early *Xist* activation, and are essential to ensure XCI induction in mouse preimplantation embryos [7, 130, 131, 133, 154].

Throughout the last few decades several mechanisms were proposed to explain how XCI can effectively count the number of X chromosomes and ensure that a single X chromosome remains active in male and female cells. A recent work hypothesized and verified experimentally that the presence of an X-linked XCI activator (xXA) results in

the effective silencing of $(N-1)$ X chromsomes in female cells characterized by $N \geq 2$ X chromosomes. xXA would indeed initiate XCI once its expression exceeds a certain threshold, which can be set in between xXA's expression levels in males and female cells. Therefore this mechanism would result in the XCI process in female cells characterized by two or more active X chromosomes, and ensure that a single X chromosome remains active in both male and female cells [112, 127, 172].

4.2 Spreading

Xist expression leads to the formation *Xist* RNA clouds along the X chromosome, which will progressively cover the future inactive allele.

The mechanism by which *Xist* RNA spreads throughout Xi has been extensively studied through RNA crosslinking and super resolution imaging, depicting this as a two step procedure. First a high number of *Xist* molecules are synthesized, localizing around its own locus and spreading to 3D closeby gene-rich regions. Specifically, a couple of *Xist* molecules are tethered to the X chromosomes at approximately 50 loci along the future inactive allele. This stage is followed by a steady phase where a smaller number of RNA molecules are produced and the *Xist* RNA hubs progressively diffuse to gene-poor regions and almost entirely cover the X chromosome [62, 117, 162, 176].

Xist repeats play an important role for the *Xist* RNA spreading stage. The lack of *Xist* repeat A results in a defective expansion of the *Xist* RNA cloud hence leading to incomplete chromosome wide silencing, while the deletion of repeat B leads to incomplete cis coating of the X chromosome. *Xist* repeat E plays a crucial role for *Xist* RNA's tethering to the X chromosome. At early stages of differentiation repeat E interacts with *Ciz1* whose deletion leads to the diffusion of *Xist* RNA throughout the nucleus, while at later time points interacts with *Ptpb1* and *Celf1* which ensure the correct adhesion of *Xist* RNA molecules to the chromosome. Furthermore the *Xist* repeat C, the YY1 protein and *hnRNP-U* ensure *Xist* RNA tethering [45, 91, 144, 157, 187].

4.3 Silencing

Concomitantly to the cis-spreading of the *Xist* RNA clouds, the future inactive X chromosome undergoes drastic chromatin changes leading to the silencing of the majority of X-linked genes.

Upon *Xist* monoallelic up-regulation the transcriptional machinery RNA Pol II is depleted from the *Xist* RNA clouds, and a repressive nuclear compartment is formed. The genes undergoing silencing will be relocated within this compartment, while the ones escaping silencing will remain located outside where they will still be accessible by the transcriptional machinery [35, 58, 108].

One of the first events following the formation of *Xist* clouds is the loss of histone acetylation. This is mediated by *Xist* repeat A which recruits the transcriptional repressor *Spen*, whose depletion in mESCs prevents effective gene silencing. Specifically *Spen* plays a crucial role in gene silencing recruiting the Histone Deacetylase 3 (HDAC3) which leads to the loss of H3 and H4 acetylation marks. Following this Polycomb repressive complexes (PRCs) accumulate on Xi, rendering it hypomethylated with respect to its active counterpart Xa. Specifically *hnRNP-K*, which binds to *Xist* repeat B, is thought to mediate with PCGF3 and PCGF5 in the recruitment of PRC1 complex which catalyzes the monoubiquitylation of H2AK119. This mediates chromatin compaction and has the effect of gradually recruiting more and more genes within the *Xist* compartment. The deposition of the H2AK119ub1 repressive mark and the mediation of *Jarid2* in turn enable the recruitment of the PRC2 complex which is responsible for the trimethylation of H3K27 histone. These histone modifications result in a clear distinction between Xi and Xa chromatin folding profiles. Indeed upon the completion of the XCI process Xa is organized into topologically associated domains (TADs) throughout the entire X chromosome, on the other hand Xi is splitted into two heterochromatic mega-domains which are separated by the *Dxz4* locus [17, 39, 45, 47, 50, 59, 74, 134, 148, 206].

The silencing mechanism described above does not however affect each gene on the X chromosome, where the genes avoiding transcriptional silencing are generally referred to as *escapees*. The percentage of genes escaping the XCI considerably varies between different organisms and cell types within the same organism. Indeed while more than 20-30% of X-linked genes escape XCI in humans, only around 3-7% of X-linked genes remain biallelically expressed in mice. Specifically the escapees are broadly categorized as *constitutive* if they avoid XCI in most cell types or developmental stages, or as *facultative* otherwise. Notably, constitutive escapees in mice are depleted of repressive histone modifications and are not coated by Xist RNA. Furthermore some of these escapees such as *Kdm5c* and *Kdm6a* (*Jarid1c* and *Utx*, respectively) play a role in XCI establishment catalyzing the demethylation of H3K27me3 and H3K4me2 [6, 13, 25, 35, 192].

4.4 Maintenance

As previously described, upon completion of the XCI process the silenced status of Xi will be maintained in all daughter cells throughout cellular divisions.

The maintainance of the silenced status does not seem to directly depend on *Xist* transcription, indeed *Xist* deletion in somatic cells leads to the reactivation of only a small subset of genes on Xi. Nonetheless a recent experiment demonstrated that during the initiation stage *Xist* recruits a set of proteins (PTBP1, CELF1, TDP-43 and MATR3) known for their role in RNA processing, which bind to *Xist* repeat E and are crucial for the maintenance of the silenced state upon XCI completion independently of *Xist* expression. Furthermore DNA methylation seems to play a crucial role for silencing

maintanance. Indeed mice deficient for *Dnmt1* show Xi reactivation, while *Smchd1* mutant mice show H3K27me3 depletion and gene reactivation on Xi [1, 19, 48, 69, 144, 165].

5 X controlling element (Xce)

As previously described, whenever each cell undergoes the random XCI process either the maternal or the paternal X chromosomes (Xm and Xp, respectively) up-regulate the expression of *Xist* and initiate the silencing process in cis. The randomness of this cell-specific choice leads to genetic mosaicism in female somatic cells, where approximately half of the cells silenced Xp and the other half Xm.

Nonetheless more than 50 years ago the work of Cattanach and others revealed the presence of silencing skewing in mice whenever their parental X chromosomes derived by different mouse strains. Specifically they identified a macro-region on the X chromosome within the Xic with four possible alleles, referred to as the X controlling element (Xce), where each pair of alleles leads to a different extent of silencing skewing. Namely the four alleles were characterized as follows: Xce^a in (129Sv, C3H, CBA), Xce^b in (BALB/C, C57Bl6/J, DBA), Xce^c in (Pjgk1a, Cast/EiJ) and Xce^d in (Mus spretus). Where $Xce^a < Xce^b < Xce^c < Xce^d$, and whenever paired the weaker chromosome tends to be silenced in a higher fraction of cells than the stronger one [29, 31–33].

During the last decades there were multiple attempts to define the precise genomic location and key components of the Xce. The Xce, that was initially mapped to a 9Mb region, was ultimately restricted to a 80kb locus downstream of Xist lieing between the *DxPas28* and *DxPas41* microsatellite repeats. Moreover a recent paper claimed that *Linx*, which resides within the Xce locus, controls the occupancy of pluripotency factors in *Xist* Intron 1 hence indirectly modulating *Xist* expression levels. Specifically this work highlights that *Linx* expression depends on the presence of *Oct4*, and that upon *Linx*'s deletion both *Oct4* and *Rex1* show decreased binding to *Xist* Intron 1 and *DxPas34*. These results suggest that *Linx* expression might prevent the initiation of the XCI process in cis and explain the Xce effect [26, 30, 66, 87, 175].

6 Previous studies

Throughout the last two decades several studies employed sequencing technologies to study the transcriptional dynamics characterizing the murine XCI process. Recent technological advancements have considerably improved our understanding of the processes governing the first stages of mouse embryonic development.

During the first decade of the 21st century, several studies adopted microarray-based technologies to explore the global differences in expression between the autosomes and sex chromosomes across several organisms and tissues [79, 135, 155].

Microarrays technologies were however less and less adopted with the advent of RNA-Sequencing technologies (RNA-Seq). The sequencing of mRNA transcripts provided a number of advantages over array-based sequencing methods, such as a considerably higher reproducibility combined with the ability to detect lowly expressed genes, alternative splicing events and novel transcripts. A number of research groups profiled the transcriptomes of differentiating mESCs both in vivo and in vitro cultures. A number of these experiments were performed on mESCs derived by the cross of two distantly related mouse strains, where the high number of SNPs between the two parental chromosomes enabled allele specific resolution [16, 60, 71, 116, 118, 119, 171, 200, 206].

Finally, the latest technological advancements enabled the refined profiling of single cells rather than bulk population of cells. A number of research groups employed single cell assays to explore the murine XCI process with cell-specific resolution in terms of their transcriptomic profile (scRNA-Seq), chromatin accessibility (scATAC-Seq) and chromatin 3-dimensional architecture (scHiC) [14, 16, 36, 38, 53, 103, 156, 163, 189].

Notably, previous works relied on engineered systems to ensure the preferential inactivation of a single X chromosome where the non-random XCI was achieved through: allele specific deletion of *Xist*, insertion of a stop codon in *Tsix*, or through the inclusion of a Dox-inducible promoter upstream of the *Xist* locus. On the contrary, the present work explores *Xist* regulation and random XCI in endogenous conditions with single cell, allelic and temporal resolutions, which enables the analyses of both these processes in a context where each cell independently chooses and inactivates either its paternally or maternally inherited X chromosome.

7 Aim of the study

The present work relies in the analysis of strand-specific single cell and bulk RNA-Sequencing data derived from hybrid (C57BL/6NJ x CAST/EiJ) XX and XO mESCs, which were profiled throughout four days of cellular differentiation.

This project aims to explore the transcriptomic profiles of female mESCs undergoing cellular differentiation, *Xist* up-regulation and random XCI in their endogenous context with stranded, allelic and cellular resolutions.

Furthermore, the analyses which are described in the following sections aim to explore the extent of gene silencing upon monoallelic and biallelic *Xist* up-regulation, to provide a detailed picture of the differential gene silencing kinetics on the two X chromosomes, and to identify the genes involved in *Xist* up-regulation and XCI processes.

2 Methods

1 Experimental Setup

All the data and results described in this thesis derive from the analysis of in vitro experiments performed on mouse cell lines. Specifically the following sections describe the experimental design and procedures adopted to generate the transcriptomic data, whose analysis is the object of the present work.

1.1 Mouse cell lines

The TX1072 is a F1 hybrid female mouse embryonic stem cells (mESCs) line derived from a cross between the C57BL/6NJ (B6) and CAST/EiJ (Cast) mouse strains. XX cells inherited one X chromosome from each parental genome. On the other hand, mESCs with a single X chromosome (XO cells) were characterized by the loss of the Cast X chromosome. The TX1072 cell line carries a doxycycline responsive promoter in front of the *Xist* gene on the B6 allele, and a reverse tetracycline-controlled transactivator (rtTA) insertion in the Rosa26 locus [171].

The TXΔXic cell lines were derived by TX1072 XX mESCs deleting through CRISPR/Cas9-mediated genome editing a 773 kb region around the *Xist* locus on either alleles, which corresponds to the X inactivation center (Xic). Namely TXΔXic$_{B6}$ carries the deletion on the B6 chromosome (chrX:103,182,701-103,955,531, mm10), while TXΔXic$_{Cast}$ on the Cast allele (chrX:103,182,257-103,955,698, mm10).

1.2 Cell culture and differentiation

Cells were maintained in a fully pluripotent state by growing them on gelatin-coated flasks in serum-containing ES cell medium (DMEM (Sigma), 15% ESC-grade FBS (Gibco), 0.1mM β-mercaptoethanol, 1000 U/ml leukemia inhibitory factor (LIF, Millipore) supplemented with 2i (3 µM Gsk3 inhibitor CT-99021, 1 µM MEK inhibitor PD0325901, Axon).

Cellular differentiation was then induced by 2i/LIF withdrawal in DMEM supplemented with 10% FBS and 0.1mM β-mercaptoethanol at a density of $1.5{\cdot}10^4$cells/cm^2 in fibronectin-coated (10 µg/ml) tissue culture plates.

Upon cellular differentiation every TX1072 XX cell undergoes random XCI leading to the silencing of either one or the other X chromosome, while its XO counterpart will not initiate this process due to the presence of a single X chromosome. The absence of the Xic in TXΔXic cells leads to non-random XCI resulting in *Xist* mono-allelic expression and gene silencing on the wild type X chromosome.

1.3 Single cell and Bulk RNA-Sequencing

In order to explore the transcriptomic profiles of murine mESCs at different stages of the X-chromosome inactivation process, RNA-Sequencing experiments were performed throughout cellular differentiation at both single cell and bulk resolutions.

1.3.1 Single cell RNA-Sequencing

The transcriptomes of female TX1072 mESCs characterized by a single (TX1072 XO) or two X chromosomes (TX1072 XX) were profiled through single cell RNA sequencing (scRNA-seq) in an undifferentiated state (day 0, 2i & Lif) and throughout four days of induced cellular differentiation (days 1-4, after 2i & Lif withdrawal). For each time point, TX1072 XX and XO mESCs were isolated and sequenced using the Fluidigm C1 system and high-throughput integrated fluidics circuits (HT IFCs) mRNA Chip. The following sections briefly describe the experimental procedure implemented to profile the transcriptome of every cell through scRNA-seq experiments.

Single cell isolation and imaging

The HT IFCs mRNA Chip microfluidics system enables the simultaneous capture and isolation of single cells from two different samples. Every chip is divided into two halves composed by 10 separate columns each, where every column can isolate up to 40 single cells. At each time point throughout cellular differentiation (days 0-4), TX1072 XX and XO cells were separately loaded onto the two halves of the chip, resulting in a maximum of 400 cells isolated for each cell population.

scRNA-seq libraries were prepared with the C1-HT mRNA-seq v2 protocol according to the manufacturer's recommendations (Fluidigm). Cells were rinsed thoroughly with PBS, trypsinized for 7 minutes and resuspended in the respective growth medium at a concentration of 400 cells/µl. 30 µl cell suspension was diluted with 20 µl of suspension reagent (Fluidigm) and 10 µl of the dilution was loaded into one compartment of a Single-Cell mRNA Seq HT IFC 10-17 µm. The loading step was repeated on the two halves of the chip, for TX1072 XX and XO cells separately.

Cell viability staining was performed on the IFC using the LIVE/DEAD viability/Cytotoxicity Kit (Thermofisher) with 1 µM Ethidium and 0.05 µM Calcein. IFC loading and life/dead staining was analyzed with automated image acquisition using a Zeiss CellDiscoverer microscope (Zeiss) with a 20x objective.

Transcriptome sequencing

During the lysis step ERCC Spike-in Mix 1 (Thermofisher) was added with a final dilution of 1:200.000. ERCC Spike-ins are artificial sequences added at known concentrations which are generally used in sequencing experiments to assess the sensitivity of the sequencing platform or to perform downstream data normalization.

Lysis, reverse transcription and cDNA amplification was performed on the C1 machine. cDNA pools were quantified by Qubit and Bioanalyzer HS. Around 2.25 ng of each pool were subjected to tagmentation and library preparation using the NexteraXT library preparation kit according to the C1-HT protocol. All pools were mixed in equal proportions and quantified with KAPA Library Quant-Kit.

The libraries were sequenced on a HiSeq2500 instrument (Illumina) with asymmetric read length, either in High Output (Read1: 13bp, Index read: 8pb, Read2: 48bp) or in Rapid Run mode (Read1: 16bp, Index read: 8pb, Read2: 36bp), with 10pM loading concentration and 5% PhiX. Where read R1 contains a custom cellular barcode (position 1-6, row barcode) and a unique molecular identifier (position 7-11, UMI molecular barcode), read R2 maps to the cDNA sequence and read R3 contains a Nextera (column) barcode.

This strategy enables the sequencing of all the transcripts derived by the cells isolated on the chip, and their following single cell demultiplexing through the row and column barcodes, which uniquely identify each cell isolated on the chip.

Read demultiplexing

For each sequenced read, the row (R1) and column (R3) barcodes are used to uniquely identify the single cell isolated on the sequencing chips which synthesized the transcript. These two barcodes were used to demultiplex the FASTQ files originating from each time point onto single-cell FASTQ files. This was achieved in BASH using the `FastqToSam` command to convert FASTQ files to a SAM format, which were then demultiplexed with `TagBamWithReadSequenceExtended` command through the Drop-seq pipeline [114].

1.3.2 Bulk RNA-Sequencing

The transcriptomes of TX1072 XX and TXΔXic mESCs were profiled through bulk RNA-sequencing in an undifferentiated state (day 0, 2i & Lif) and throughout four days of induced cellular differentiation (days 1-4, after 2i & Lif withdrawal) with three biological replicates per cell line and time point. The following subsection describes

the experimental procedure used to profile each sample through bulk RNA-sequencing experiments.

Transcriptome sequencing

For the TX1072 XX cell line, bulk RNA-Sequencing was performed in parallel to the single cell RNA-Seq experiment (replicate 1) and for two more biological replicates (replicate 2 & 3). RNA-Seq libraries were generated using the Tru-Seq Stranded Total RNA library preparation kit (Illumina) with 1 µg starting material for rRNA-depletion and amplified with 15 Cycles of PCR. For each time point, each biological replicate was sequenced 2x50bp on a HiSeq 2500 with 1% PhiX spike-in, which generated 50 Mio. fragments per sample.

For TXΔXic cell lines, libraries were generated with the KAPA-RNA Hyper Prep-Kit with RiboErase (Roche) following the protocol, with 500ng total RNA used for rRNA-depletion. For undifferentiated (TXΔXic$_{Cast}$) samples fragmentation was adjusted (85°C/5min instead of 94°C/8min) due to RNA degradation. Nextflex unique dual-index-adaptors (PerkinElmer) were used and the final library was PCR-amplified with 10 cycles. For each cell line and time point, three biological replicates were sequenced 2x100bp on a NovaSeq6000 with 1% PhiX spike-in, which generated 50 Mio. fragments per sample.

1.4 RNA FISH

RNA FISH is a cytogenic technique that relies on fluorescent probes designed to target and bind to specific RNA sequences of interest. The hybridization of these probes to the target RNAs results in a fluorescent signal which can be captured and visualized through fluorescence microscopy.

For the aims of this analysis, Stellaris RNA FISH probes were used to visualize the allele specific expression of two X-linked genes (*Xist* and *Huwe1*) in TX1072 XX mESCs throughout cellular differentiation (days 2-4).

Cells were dissociated using Accutase (Invitrogen) and adsorbed onto coverslips (#1.5, 1mm) coated with Poly-L-Lysine (Sigma) for 5 min. Cells were fixed with 3% paraformaldehyde in PBS for 10 min at room temperature (18–24°C) and permeabilized for 5 min on ice in PBS containing 0.5% Triton X-100 and 2 mM Ribonucleoside Vanadyl complex (New England Biolabs). Coverslips were preserved in 70% EtOH at -20°C. Prior to FISH, coverslips were incubated for 5 minutes in Stellaris RNA FISH Wash Buffer A (Biosearch Technologies), followed by hybridization overnight at 37°C with 250 nM of each FISH probe in 50 µl Stellaris RNA FISH Hybridization Buffer (Biosearch Technologies) containing 10% formamide. Coverslips were washed twice for 30 min at 37°C with Stellaris RNA FISH Wash Buffer A (Biosearch Technologies), with 0.2 mg/ml Dapi

being added to the second wash. Prior to mounting with Vectashield mounting medium coverslips were washed with 2xSSC at room temperature for 5 minutes.

For each time point, images of three biological replicates were acquired using a widefield Z1 Observer microscope (Zeiss) with a 100x objective. The intronic signal of *Huwe1* was used in combination with *Xist* to estimate the percentage of XO cells in the population. For each replicate sample and time point, the fraction of cells showing mono-allelic and bi-allelic gene expression was calculated over 100 individual cells.

1.5 Pyrosequencing

Pyrosequencing is a DNA-sequencing procedure where the sequence characterizing a single stranded DNA (ssDNA) target molecule is read by synthesing its complementary strand adding one nucleotide (A, T, C, G) at a time. The target DNA sequence is first denatured, then the ssDNA molecule is hybridized to a sequencing primer and mixed with DNA polymerase, ATP sulfurylase and firefly luciferase. Starting from the sequencing primer, whenever a nucleotide complementary to the target ssDNA is added to the mix the DNA polymerase hybridizes it to the ssDNA resulting in the release of pyrophosphate. The light signal emitted by the pyrophosphate release is captured by a camera, where the intensity of the recorded light signal indicates if a single or several nucleotides successfully hybridized to the ssDNA. On the other hand, the absence of light signal indicates that the nucleotide has not hybridized to the ssDNA. Adding one different nucleotide at a time to the mix, recording the light intensity upon each nucleotide addition and repeting this procedure until the synthesis of the whole complementary strand reveals the target ssDNA sequence.

For the aims of this analysis, Pyrosequencing experiments were performed at each time point throughout cellular differentiation (days 0-4) to analyse the relative allelic expression of a number of X-linked genes (*Atrx*, *Cul4b*, *Hprt*, *Klhl13*, *Pir*, *Prdx4*, *Renbp*, *Rnf12* and *Xist*). Notably the relative allelic expression of each gene was quantified across four biological replicates.

For each target gene, an amplicon containing a SNP between the two parental lines was amplified by PCR from cDNA using GoTaq Flexi G2 (Promega) with 2.5mM MgCl2 or Hot Star Taq (Qiagen) for 40 cycles. The PCR product was then sequenced using the Pyromark Q24 system (Qiagen), and allelic expression was quantified counting the number of reads with either genotypes. Table 2.1 stores the SNP loci and sequencing primers which were used to perform the Pyrosequencing experiments.

1.6 Author contributions

Ilona Dunkel (Max Planck Institute for Molecular Genetics, Schulz group) cultured and subcloned TX1072 XX and XO mESCs, and set up the differentiation experiments. Verena Mutzel (Max Planck Institute for Molecular Genetics, Schulz group) generated the TXΔXic cell lines through CRISPR-Cas9-mediated gene editing [described in Methods sections: 1.1-1.2]. Norbert Mages (Max Planck Institute for Molecular Genetics, Sequencing core facility) prepared the libraries and performed RNA sequencing experiments at single-cell and bulk resolutions with the input from Edda Schulz (Max Planck Institute for Molecular Genetics) and Bernd Timmermann (Max Planck Institute for Molecular Genetics, Head of the Sequencing core facility) [described in Methods sections: 1.3]. Ilona Dunkel performed the Pyrosequencing, qPCR and RNA FISH experiments on TXΔXic mESCs at different stages of cellular differentiation [described in Methods sections: 1.4-1.5].

The original contribution of the doctoral candidate Guido Pacini (Max Planck Institute for Molecular Genetics, Schulz group) to the thesis is the analysis and interpretation of all the experimental data described in this section. This was achieved under the supervision of Edda Schulz and Annalisa Marsico (Helmholtz Center, Institute for Computational Biology).

Table 2.1: Pyrosequencing sequencing primers and SNPs

Gene	Dispensation order	SNP position / ID
Hprt	ATCGTCTA	mm10: 53,021,504
Prdx4	GCTGTGTA	mm10: 155,330,388
Cul4b	GACTGTGA	mm10: 38,539,035
Renbp	CAGCATGA	mm10: 73,927,864
Pir	GATCGAGA	rs250477126
Klhl13	GCAGTCATTAAT	rs237072964
Rnf12	CTCGTAGCTA	rs29081561
Xist	CAGTCTCA	mm10: 103,470,658
Atrx	TCGTGCTA	rs29078297

Gene	Primer name		Sequence		
Hprt	ES705	F1, bio	[Bio]ATTCAGGAGAGAAAGATGTGATTG	TC/GGTCTAAATTAACAATATCAATCACA	
	ES706	R1	CCACTGAGCAAAACCTCTTAGAT		
	ES707	S1	AAATCGAGAGCTTCAGAC		
Prdx4	ES714	F1-bio	[Bio]TGTCCTGAGTCTTCAAGGTATACA	C/TGGTGTATACCTTGAAGACTCAGGAC	
	ES715	R1	GACTGGGGCCAATAAGGATT		
	ES716	S1	CATCAGATCTCAAAGGACTA		
Cul4b	ES717	F1-bio	[Biotin]AGTTTGTGTTTGAAATCGTCATTA	AC/TGGTGACAAATTTATTTGTAATGACG	
	ES718	R1	TCCTAAGGGCAAAGATATTGAAG		
	ES719	S1	GGGCAAAGATATTGAAG		
Renbp	ES720	F1-bio	[Biotin]TCCGCATCCTGGAAGTAGA	A/GCATGGAGGCCTCTTCTACTTCCAGG	
	ES721	R1	CTGCTCCAGTATGCCCTCAG		
	ES722	S1	TGGATGGGACCCTGA		
Pir	ID80	F1_bio	[Biotin]CCAGCACTCTAGGAGTCAGAGACA	ATC/TGGAGATACCCAATGTCTCTGACTCC	
	ID81	R1	TCAAGACAGATGGGCTTGGAA		
	ID82	S1	CAGATGGGCTTGGAA		
Klhl13	ID83	F1	AATCCCCATTTTTCATGGAAG	C/AAGTCAATTTTTAAAAATTTGTATTT	
	ID84	R1_bio	[Biotin]TTGGTTTGGGGTTTTTTTTAAG		
	ID85	S1	TCACATTTATTTGACCTGA		
Rnf12	ID004	F1	TGTTGTTCGGAGCCTGAGAT	TCTG/ACTAGCTATACT	
	ID005	R1_bio	[Biotin]GGAGAATACCGGCAGAGAGATA		
	ID006	S1	CCTGAGATCTTGATCGAGT		
Xist	ES400	F1_bio	[Biotin]AGAGAGCCCAAAGGGACAAA	A/GTCTCACATAG	
	ES401	R1	TGTATAGGCTGCTGGCAGTCC		
	ES402	S1	GCTGGCAGTCCTTGA		
Atrx		F1	CTGAATTCTGATCCCCAATCAC	G/CTTGCTTAG	
		R1_bio	[Biotin]CCTCACAAGGTACCCAAAGCTAA		
		S1	TGATCCCCAATCACTG		

2 Bulk RNA-sequencing data pre-processing

This section describes the pre-processing procedures implemented to align the sequencing reads resulting from the bulk RNA-Sequencing experiments to a custom mouse genome, to quantify the overall and allele-resolved expression of each annotated gene in every TX1072 or TXΔXic mESC sequenced sample, and to normalize the resulting count matrices.

2.1 Read alignment procedure

2.1.1 Custom mouse genome

Since all the mESCs under investigation are hybrids between the B6 and Cast mouse strains, half of their transcriptome is inherited by each parental allele. Aligning the sequencing reads of samples derived from this hybrid cell line to the B6 or Cast mouse genomes individually would bias the alignment procedure towards one or the other parental allele. Indeed the large number of single nucleotide polymorphisms (SNPs) between these two distantly related mouse strains would be erroneously regarded as alignment mismatches, hence underestimating the gene expression of either alleles.

In order to avoid biases in the read alignment and gene expression quantification procedures, a custom mouse genome was first generated using the `SNPsplit` software (v.0.3.2) [96] to mask a set of high confidence SNPs (18658116 SNPs genome-wide) between the two distantly related parental cell lines. Where the set of masked SNPs was confirmed to be present in the TX1072 cell line based on previous ChIP-seq data [60]. Following this, every sequencing read was aligned to the masked mouse genome in combination with the 96 ERCC spike-in sequences.

2.1.2 Read alignment

Paired-end reads for each TX1072 or TXΔXic sequenced sample were first aligned to the masked mouse genome combined with 96 ERCC spike-in sequences with STAR (v.2.5.2b) [56], keeping only uniquely aligned reads with a maximum of two mismatches.

For allele specific (AS) transcript quantification, each read overlapping at least one SNP position was assigned to its parental genome using the `SNPsplit` software (v.0.3.2) [96], discarding the reads overlapping multiple SNPs with discordant genotypes.

2.2 Gene expression quantification

For every sample j, the number of molecules transcribed by the g-th gene was estimated as the number of uniquely aligned reads overlapping its annotated exonic regions. Repeating this procedure for all sequenced samples and genes led to the estimation of the count matrix $\mathbf{B} = [b_{g,j}] \in \mathbb{N}^{G \times N}$. The same procedure was repeated for each allele separately, restricting the analysis to SNP-overlapping reads assigned to either parental genome. This resulted in the estimation of the AS count matrices $\mathbf{B}^{B6} = [b^{B6}_{g,j}] \in \mathbb{N}^{G \times N}$ and $\mathbf{B}^{Cast} = [b^{Cast}_{g,j}] \in \mathbb{N}^{G \times N}$.

In detail, the `Rsubread` (v.1.34.7) R package [104] was used to quantify the expression of each gene. Where *Xist* (ENSMUSG00000086503) AS gene expression was also quantified using only two SNPs on its 5'-end (namely, chrX:103,482,240 and chrX:103,482,895, mm10).

2.3 Data normalization

Count matrix normalization aims to remove technical and biological biases between samples, such as differences in capture efficiency, sequencing depth or composition bias. Where the latter refers to the presence of up-regulated or down-regulated genes in a subset of samples which results in spurious under or over representation of the remaining genes, respectively. These biases can be accounted for dividing every gene count by a sample-specific scaling factor. As a result, the normalized counts will show no difference in gene expression for the majority of genes.

The following subsections describe the methods used to normalize the previously defined not-AS and AS bulk RNA-Sequencing count matrices.

2.3.1 TMM normalization procedure

The most widely used normalization methods for bulk RNA-sequencing expression data select a sample or pseudo-sample as the reference, compute the ratio of gene expression between this and every other sample, and robustly estimate the sample-wise scaling factor across a set of putatively not differentially expressed (not-DE) genes.

In this work, bulk RNA-sequencing count data were normalized through the *Trimmed Mean of M-values (TMM)* normalization procedure [159], which is briefly described below.

The TMM method computes sample-specific scaling factors aiming to equate the overall expression of every gene ($g = 1, \ldots, G$) across all sequenced samples ($j = 1, \ldots, N$), under the hypothesis that the majority of genes G are not differentially expressed across

the N samples. This procedure involves selecting a sample r as the reference, and scaling the gene expression of every other sample j relative to r. For each gene g, the log-fold-change $M^r_{g,j}$ and absolute intensity $A^r_{g,j}$ are computed as:

$$\begin{aligned} M^r_{g,j} &= \frac{\log_2\left(b_{g,j} / \sum_{g=1}^{G} b_{g,j}\right)}{\log_2\left(b_{g,r} / \sum_{g=1}^{G} b_{g,r}\right)} \\ A^r_{g,j} &= \frac{1}{2} \cdot \log_2\left(\frac{b_{g,j}}{\sum_{g=1}^{G} b_{g,j}} \cdot \frac{b_{g,r}}{\sum_{g=1}^{G} b_{g,r}}\right) \\ &\text{such that: } b_{g,j} \neq 0 \text{ and } b_{g,r} \neq 0 \end{aligned} \qquad \text{(Eq. 1)}$$

Where, as previosuly described in Section 2.2, $b_{g,j}$ represents the observed read counts for the g-th gene and j-th sample.

A robust estimate of the scaling factor sf^r_j for sample j relative to sample r is then computed as a weighted average of the $M^r_{g,j}$ values across a set of putatively not-DE genes G^*. Under the assumption that genes with extremely high or low M and A values are more likely to be DE across samples, the subset of genes G^* is identified trimming both the M and A values (by default by 30% and 5%, respectively). The scaling factors sf^r_j are then estimated taking a weighted average of the M values, with weights equal to the inverse of the approximate asymptotic variance:

$$\begin{aligned} \log_2\left(sf^r_j\right) &= \frac{\sum\limits_{g \in G^*} w^r_{g,j} \cdot M^r_{g,j}}{\sum\limits_{g \in G^*} w_{g,j}} \\ \text{where: } w^r_{g,j} &= \frac{\left(\sum\limits_{g \in G^*} b_{g,j}\right) - b_{g,j}}{\left(\sum\limits_{g \in G^*} b_{g,j}\right) \cdot b_{g,j}} + \frac{\left(\sum\limits_{g \in G^*} b_{g,r}\right) - b_{g,r}}{\left(\sum\limits_{g \in G^*} b_{g,r}\right) \cdot b_{g,r}} \\ &\text{such that: } sf^r_r = 1 \end{aligned} \qquad \text{(Eq. 2)}$$

This factor sf^r_j can then be used to scale the total number of counts for sample j, such that every sample is normalized to the same number of counts observed in reference sample r across the not-DE genes G^*.

In detail, the TMM normalization procedure was implemented through the `edgeR` (v.3.26.8) R package [161], where the scaling factor of every sequenced sample was computed through the `calcNormFactors` R function.

2.3.2 Normalized Counts Per Million (CPM) values

Let sf_j^r represent the scaling factor for sample j which was estimated by the TMM method for bulk RNA-sequencing data. Since XX cells undergo XCI throughout cellular development, X-linked genes are more likely to be differentially expressed across samples with respect to autosomal genes. For this reason the scaling factors estimate was restricted to autosomal genes. Then the normalized Counts Per Million (CPM) values for each gene g and sample j ($CPM_{g,j}$) of both the AS and not-AS count matrices were computed as:

$$CPM_{g,j}^{\mathbf{X}} = \frac{x_{g,j}}{sf_j \cdot \sum_g x_{g,j}} \cdot 10^6 \qquad \text{(Eq. 3)}$$

Where sf_j is the scaling factor for sample j, $\mathbf{X} = [x_{g,j}]$ is a random variable which represents the AS or not-AS count matrices relative to the bulk RNA-sequencing data $\mathbf{X} = \{\mathbf{B}, \mathbf{B}^{B6}, \mathbf{B}^{Cast}\}$, and $\sum_g x_{g,j}$ represents the overall gene expression across all genes profiled for sample j.

3 Single cell RNA-sequencing data pre-processing

Similarly to the previous section, the following sub-sections describe the pre-processing procedures implemented to align the sequencing reads resulting from the single cell RNA-sequencing experiments to the previously defined custom mouse genome, to quantify the overall and allele-resolved expression of each annotated gene in every TX1072 mESC sequenced cell, and to normalize the resulting count matrices.

3.1 Read alignment

Hybrid mESCs were profiled throughout cellular differentiation with single cell paired-end RNA sequencing, where sequencing read 1 (R1) carries both the cellular and UMI molecular barcodes, while the 3'-biased cDNA sequence of every transcribed gene is encoded in read 2 (R2).

The Drop-seq pipeline [114] was first used to extract from R1 both the cellular and UMI molecular barcodes (respectively bps 1-6 and 7-11). All the R2 sequences were then aligned to the custom mouse genome using STAR (v.2.5.2b) [56]. Downstream analyses were restricted to uniquely aligned reads with a maximum of two mismatches.

For allele specific (AS) gene expression quantification, every R2 sequence which was aligned onto one or more high confidence SNP loci was assigned to its parental allele

according to the SNP genotypes. This was achieved using the SNPsplit software (v.0.3.2) [96]. Notably we removed from all downstream analyses the reads overlapping multiple SNP loci which were characterized by discordant genotypes, as well as the sequencing reads assigned to the paternal and maternal genomes carrying identical UMI barcode sequences.

3.2 Gene expression quantification

For any cell j, the amount of molecules transcribed by the g-th gene can be estimated as the number of unique UMI barcodes across all the reads which align over the g-th annotated exonic regions. The use of random UMI barcodes, annealed to every cDNA molecule prior library amplification, reduces the bias in gene expression quantification caused by differences in amplification efficiency across transcripts.

Repeating the above UMI counting procedure for all N cells and G annotated genes, results in the estimation of the UMI count matrix: $\mathbf{Y} = [y_{g,j}] \in \mathbb{N}^{G \times N}$, which quantifies the expression of every annotated gene across every sequenced mESC. The same procedure was then repeated for each allele separately, restricting the analysis to SNP-overlapping reads which were previously assigned to either parental genomes. This resulted in the estimation of allele-specific UMI count matrices, one for the B6 and one for the Cast allele, namely: $\mathbf{Y}^{B6} = [y^{B6}_{g,j}]$, $\mathbf{Y}^{Cast} = [y^{Cast}_{g,j}] \in \mathbb{N}^{G \times N}$. These two counting procedures are referred to as not allele specific (not-AS) and allele specific (AS) gene expression quantifications throughout the text.

In order to estimate the number of spliced and unspliced mRNA molecules, the not-AS and AS quantifications described above were repeated separately for the subset of reads completely aligned to an exonic region, or to those overlapping an intronic region. This resulted in the estimation of the spliced (not-AS: $\mathbf{S} = [s_{g,j}]$; AS: $\mathbf{S}^{B6} = [s^{B6}_{g,j}]$ and $\mathbf{S}^{Cast} = [s^{Cast}_{g,j}]$) and unspliced (not-AS: $\mathbf{U} = [u_{g,j}]$; AS: $\mathbf{U}^{B6} = [u^{B6}_{g,j}]$ and $\mathbf{U}^{Cast} = [u^{Cast}_{g,j}]$) UMI count matrices, such that: $\mathbf{S}, \mathbf{S}^{B6}, \mathbf{S}^{Cast}, \mathbf{U}, \mathbf{U}^{B6}, \mathbf{U}^{Cast} \in \mathbb{N}^{G \times N}$.

In detail, the Drop-seq pipeline [114] was first used to tag the uniquely aligned reads overlapping annotated genes. Then the amount of molecules transcribed by each gene was estimated as the number of unique UMI barcodes across all tagged reads.

3.3 Data normalization

A strong limitation of the previously described TMM normalization procedure is that zero counts can't be included in the calculation of the sample-specific scaling factors (Eq. 1). Therefore, given the characteristic zero-inflation observed in single cell RNA-sequencing count data, the TMM procedure might fail to accurately estimate the normalization factors. When dealing with zero-inflated data, meaningful scaling factors can

be obtained grouping cells with similar transcriptomic profiles, pooling their counts to reduce the number of gene dropouts, computing a scaling factor for each pooled sample, and estimating single cell scaling factors through linear deconvolution.

In this work, single cell RNA-sequencing count data were normalized using the pooling-clustering normalization procedure [2], which is briefly described below.

3.3.1 Pooling-clustering normalization procedure

Let $y_{g,j}$ represent the observed read counts for the g-th gene and j-th cell, as previosuly described in Section 3.2. Then the expected value of the random variable $Y_{g,j}$ can be written as: $E(Y_{g,j}) = sf_j \cdot \lambda_g$, where sf_j is the cell-specific bias factor and λ_g is the expected gene count.

Let T_j represent the library size of the j-th cell (i.e. the total number of gene counts across all profiled genes, $\sum_g Y_{g,j}$), then the random variable for the relative gene expression can be written as: $Z_{g,j} = Y_{g,j} \cdot T_j^{-1}$, where $E(Z_{g,j}) = sf_j \cdot \lambda_g \cdot T_j^{-1}$.

Let $V_{g,k}$ be the random variable representing the sum of the relative gene expression values observed for all cells belonging to a set S_k, and let the random variable A_g represent the average relative gene expression values across all N cells. These two random variables have expected values: $E(V_{g,k}) = \lambda_g \cdot \sum_{j \in S_k} sf_j \cdot T_j^{-1}$ and $E(A_g) = \lambda_g \cdot N^{-1} \cdot \sum_{j=1}^{N} sf_j \cdot T_j^{-1}$.

Assuming that most genes are not-DE between the k-th cell pool $V_{g,k}$ and the overall average A_g, the scaling factor for the k-th cell pool can be estimated as the median of the ratios $R_{g,k} = V_{g,k}/A_g$ across all genes. This can be written as the linear combination of the cell-specific scaling factors relative to the cells composing the pool S_k, as:

$$sf_{S_k} = \underset{g=1,\dots,G}{\text{median}} \frac{V_{g,k}}{A_g} = \sum_{j \in S_k} \alpha_j \cdot sf_j \qquad \text{(Eq. 4)}$$

Repeating the above procedure for several cell pools (S_k) results in a linear system of equations which can be solved by weighted least squares to estimate the cell-wise scaling factors (sf_j, with $j = 1, \dots, N$). Where every cell pool of size w is defined grouping cells with similar library sizes, such that every cell is represented in w equations and the total number of equations equals the number of cells N. Finally the cell cluster with the most non-zero counts is defined as the reference, and used to re-scale the size factors from all the others.

The above procedure can also be performed for separate clusters of cells, in order to further reduce the strength of the non differential expression assumption. In this case, each nested cell pool is normalized towards its cluster's average resulting in cluster

specific scaling factors. The inferred scaling factors are only comparable between clusters upon rescaling, which is achieved by selecting a cluster's average as the reference and normalizing each scaling factor with respect to it [2].

In detail, the above method was implemented through the `scran` (v.1.12.1) R package [2]. Cell clusters were defined independently for each time point through the `computeSumFactors` function and the *clusters* parameter, restricting the cell pooling step to cells sequenced at the same time point of cellular differentiation. Cluster-based scaling factors were then deconvoluted into cell-specific factors (sf_j; for $j = 1, \ldots, N$).

3.3.2 Normalized Counts Per Million (CPM) values

Let sf_j represent the scaling factor for cell j which was estimated by the pooling-clustering method. Similarly to the bulk RNA-sequencing data analysis, the estimation of cell-specific scaling factors was restricted to autosomal genes. Indeed on one hand XX cells undergo XCI throughout cellular development while on the other XO cells have a single copy of the X chromosome, hence X-linked genes are more likely to be differentially expressed across cells with respect to autosomal genes.

The normalized Counts Per Million (CPM) values for each gene g and cell j ($CPM_{g,j}$) of both the AS and not-AS count matrices were computed in the same way as described for bulk RNA-sequencing data (Eq. 3).

Where sf_j is the scaling factor for cell j, $\mathbf{X} = [x_{g,j}]$ is a random variable which represents the AS or not-AS count matrices relative to single cell RNA-sequencing data $\mathbf{X} = \{\mathbf{Y}, \mathbf{Y}^{B6}, \mathbf{Y}^{Cast}\}$, and $\sum_g x_{g,j}$ represents the overall gene expression across all genes profiled for cell j.

4 Data filtering

Data filtering is a crucial step in the analysis of transcriptomic data, since the presence of outlying samples and genes could bias downstream analyses.

The following sections describe the pre-processing procedures which were applied to scRNA-seq not-AS and AS count matrices aiming to identify and remove from the analysis putatively dead or poorly sequenced cells, as well as lowly expressed or poorly annotated genes.

4.1 Cell filtering

Problematic cells could be identified combining imaging and gene expression data.

4.1.1 Image-based filtering

Capture sites isolating multiple cells or no cells were identified through manual inspection of their brightfield images and removed from subsequent analyses. Furthermore, the dead (Eithidium) and life (Calcein) stain fluorescence levels of every capture sites were quantified as the average intensity signal measured within a rectangle of constant size which was manually centered at each capture site.

In detail, the fluorophores' intensities of each capture site were quantified through the `ZEN` (Zeiss v2.3) software.

4.1.2 Transcriptome-based filtering

Putatively dead cells were identified as the samples characterized by extremely low sequencing depth, total number of transcripts or number of expressed genes, as well as cells with extremely high dead stain fluorescence intensities, percentages of mithochondrial DNA or ERCC spike-in reads.

Let $\boldsymbol{x} \in \mathbb{R}^N$ be a random variable representing the values observed for any of the above features measured across all the N sequenced cells. Then the k-fold median-absolute-deviation (k·MAD) upper and lower thresholds can be computed as:

$$median(\mathbf{x}) \pm k \cdot median(\left|\mathbf{x} - median(\mathbf{x})\right|) \qquad \text{(Eq. 5)}$$

The 3-fold median-absolute-deviation ($k = 3$) thresholds were used to identify problematic capture sites, and remove from the analysis putatively dead cells. Specifically capture sites showing extreme values for any of the above variables were removed from all the downstream analyses.

4.1.3 Remove cells losing one X chromosome

Aiming to remove XO cells within the XX population, any sequenced cell that didn't express the *Xist* gene and with more than 80% of its X-linked AS UMI counts assigned to a single allele, was assumed to have lost one X chromosome.

To this end, the X-chromosomal ratio of the j-th XX cell (namely, XR_j) was computed as the fraction of X-linked AS UMI counts assigned to the B6 allele:

$$XR_j = \frac{\sum\limits_{g \in chrX} y_{g,j}^{B6}}{\sum\limits_{g \in chrX} \left(y_{g,j}^{B6} + y_{g,j}^{Cast}\right)} \qquad \text{(Eq. 6)}$$

Any XX cell j such that $XR_j \geq 0.8$ or $XR_j \leq 0.2$ and $y^{B6}_{Xist,j} + y^{Cast}_{Xist,j} = 0$ was assumed to be XO and removed from all the downstream analyses.

4.2 Gene filtering

In order to get robust results in downstream data analyses, the gene filtering procedures described below aim to identify and remove from the analysis poorly detected genes as well as those genes characterized by skewed detection biases.

4.2.1 Dropout rate

Let the dropout rate of a gene g represent the proportion of profiled cells that do not transcribe the gene.

For every gene g, we can estimate its not-AS and AS dropout rates ($DropRate^{notAS}_g$ and $DropRate^{AS}_g$, respectively) as:

$$\begin{aligned} DropRate^{notAS}_g &= \sum_j I(y_{g,j} = 0) \Big/ \sum_j 1 \\ DropRate^{AS}_g &= \sum_j I(y^{B6}_{g,j} + y^{Cast}_{g,j} = 0) \Big/ \sum_j 1 \end{aligned} \qquad \text{(Eq. 7)}$$

Where $I(\,)$ represents the indicator function, j denotes each sequenced XX cell, and $\sum_j 1$ quantifies the number of profiled cells.

Aiming to restrict the analysis only to genes with sufficiently high detection rates, we excluded from downstream analyses any gene with a dropout rate higher than 0.8. Importantly, since a much smaller set of genes could be profiled with allelic resolution relative to the not allele specific one, this filtering procedure was performed separately for the not-AS ($DropRate^{notAS}_g \geq 0.8$) and AS ($DropRate^{AS}_g \geq 0.8$) downstream data analyses.

4.2.2 Allelic rate

Let the allelic rate of a gene g represent the fraction of molecules transcribed by the B6 allele across all sequenced cells.

For every gene g, we can estimate the allelic rate ($AllelicRate_g$) as:

$$AllelicRate_g = \sum_j y^{B6}_{g,j} \Big/ \sum_j \left(y^{B6}_{g,j} + y^{Cast}_{g,j}\right) \qquad \text{(Eq. 8)}$$

Aiming to exclude any gene showing strong detection skewing towards a single allele, we removed from both AS and not-AS downstream data analyses every gene such that over 90% of its transcripts derived from a single allele ($AllelicRate_g \geq 0.9$ or $AllelicRate_g \leq 0.1$).

5 Global and gene-wise X-linked silencing measures

This section describes a number of measures which were used to characterize the extent of global and gene-wise X-linked expression and silencing for each single cell throughout cellular differentiation.

Specifically the *Xist AS ratio* estimates the extent of mono or bi-allelic expression of the *Xist* gene for each single cell. The *X:A ratio* compares the average gene expression on the X-chromosome and autosomes, which is a commonly used measure to estimate the silencing of X-linked genes. Finally the *Xi:Xa ratio* and *X chromosome inactivation progress (XP)* compare the expression levels of the *Xist* positive and negative X-chromosomes, which are used to estimate the extent of *Xist*-mediated chromosome-wide gene silencing for each sequenced XX cell.

5.1 *Xist* AS ratio

Every XX cell j was classified based on its allelic expression of the *Xist* gene (ENSMUSG00000086503). Where the *Xist* allele specific ratio for the j-th XX cell was computed as:

$$ratio_{Xist,j} = \frac{y^{B6}_{Xist,j}}{y^{B6}_{Xist,j} + y^{Cast}_{Xist,j}} \quad \text{(Eq. 9)}$$

Cells with more than 5 *Xist* AS counts were classified as: *Monoallelic* (MA-B6 or MA-Cast) if all AS counts mapped to the same allele; *Skewed* if more than 80% of the AS counts were assigned to a single allele; *Biallelic* (BA) if at least 20% of the AS counts were assigned to each allele; and *Undetected* if the cell didn't express *Xist* at the not-AS level:

- Undetected: if $y_{Xist,j} = 0$
- Low: if $(y^{B6}_{Xist,j} + y^{Cast}_{Xist,j}) \leq 5$
- MA-B6 (Xi = B6): if $(y^{B6}_{Xist,j} + y^{Cast}_{Xist,j}) > 5$ and $ratio_{Xist,j} = 1$
- MA-Cast (Xi = Cast): if $(y^{B6}_{Xist,j} + y^{Cast}_{Xist,j}) > 5$ and $ratio_{Xist,j} = 0$

- BA: if $(y^{B6}_{Xist,j} + y^{Cast}_{Xist,j}) > 5$ and $ratio_{Xist,j} \in [0.2, 0.8]$
- Skewed: if $(y^{B6}_{Xist,j} + y^{Cast}_{Xist,j}) > 5$ and $ratio_{Xist,j} \in \{(0, 0.2) \cup (0.8, 1)\}$

5.2 X:A ratio

The X-to-Autosome (X:A) ratio compares the average expression observed across the genes on the X chromosome and autosomes. Similarly to a previous study [16], a bootstrapping approach was used to account for the larger number of autosomal genes compared to X-linked genes. For each cell, a set of autosomal genes (g'_b) of the same size of the X-linked gene set was randomly sampled with reintroduction, and the ratio between the average X-linked expression and the average expression across the sampled autosomal genes was computed. This step was repeated B=1000 times, and the X:A ratio for each cell was estimated as the median across all B bootstrap ratios. Only genes retained through the gene filtering step were included in this analysis. For each cell j, the not-AS and AS X:A ratios were computed as:

$$\begin{aligned}
\text{X:A}_j^{\text{not-AS}} &= \underset{b=1,\dots,B}{\text{median}} \Big(\sum_{g \in chrX} y_{g,j} \Big/ \sum_{g \in g'_b} y_{g,j} \Big) \\
\text{X:A}_j^{\text{B6}} &= \underset{b=1,\dots,B}{\text{median}} \Big(\sum_{g \in chrX} y^{B6}_{g,j} \Big/ \sum_{g \in g'_b} y^{B6}_{g,j} \Big) \\
\text{X:A}_j^{\text{Cast}} &= \underset{b=1,\dots,B}{\text{median}} \Big(\sum_{g \in chrX} y^{Cast}_{g,j} \Big/ \sum_{g \in g'_b} y^{Cast}_{g,j} \Big)
\end{aligned} \qquad \text{(Eq. 10)}$$

5.3 Xi:Xa ratio

The Inactive-to-Active X chromosome expression ratio (Xi:Xa) is a measure which compares the global X-linked expression of the two alleles. This ratio was estimated as the fraction between the number of X-linked molecules transcribed by the future inactive allele (Xi) and the active allele (Xa), excluding the UMI counts assigned to the *Xist* gene. For each XX cell with *Xist* monoallelic expression (namely, MA-B6 and MA-Cast), this measure was computed as:

$$\begin{aligned}
&\text{Xi:Xa}_j = \sum_{\substack{g \in chrX \\ g \neq Xist}} y^{Xi}_{g,j} \Big/ \sum_{\substack{g \in chrX \\ g \neq Xist}} y^{Xa}_{g,j} \\
&\text{if } j \in \text{MA-B6} : Xi = B6 \text{ and } Xa = Cast \\
&\text{if } j \in \text{MA-Cast} : Xi = Cast \text{ and } Xa = B6
\end{aligned} \qquad \text{(Eq. 11)}$$

Where for each cell j classified as monoallelically expressing *Xist*, $y_{g,j}^{Xi}$ and $y_{g,j}^{Xa}$ represent the observed read counts for the g-th gene on the *Xist* expressing and negative alleles, respectively.

5.4 X chromosome inactivation progress (XP)

The XCI progress (XP) is a measure which quantifies the percentage of global X-linked silencing of the future inactive allele (Xi) relative to the active allele (Xa). For each XX cell with *Xist* monoallelic expression (namely, MA-B6 and MA-Cast), this measure was computed as:

$$\begin{aligned} XP_j &= 100 \cdot \left(1 - \frac{\sum\limits_{g \in chrX} y_{g,j}^{Xi} + 0.01}{\sum\limits_{g \in chrX} y_{g,j}^{Xa} + 0.01}\right) \\ &\text{if } j \in \text{MA-B6} : Xi = B6 \text{ and } Xa = Cast \\ &\text{if } j \in \text{MA-Cast} : Xi = Cast \text{ and } Xa = B6 \end{aligned} \qquad \text{(Eq. 12)}$$

Intuitively, a high XP_j value indicates that the cell has already substantially reduced the expression of genes on Xi, while a value proximal to zero implies that the two alleles have similar gene expression levels. The XP_c value was set to 0 for those cells showing higher X-linked gene expression on Xi compared to Xa.

5.5 Gene silencing progress (Xi/Xa)

Similarly to the previous measure, the extent of silencing of every X-linked gene g and *Xist* monoallelic XX cell j was measured as:

$$\begin{aligned} \text{Xi:Xa}_{g,j} &= \frac{y_{g,j}^{Xi} + 0.01}{y_{g,j}^{Xa} + 0.01} \in [0,1];\ g \in chrX \\ &\text{if } j \in \text{MA-B6} : Xi = B6 \text{ and } Xa = Cast \\ &\text{if } j \in \text{MA-Cast} : Xi = Cast \text{ and } Xa = B6 \end{aligned} \qquad \text{(Eq. 13)}$$

Intuitively, a value proximal to 0 indicates that the gene is silenced on the Xi allele, while a value proximal to 1 indicates that the gene is equally expressed by the two alleles.

6 Trajectory inference and RNA-velocity methods

6.1 Dimensionality reduction techniques

RNA-sequencing experiments profile each sample measuring its expression over a large number of genes. Dimensionality reduction techniques are crucial tools for data exploration and clustering. These procedures enable the visual inspection of the sequenced samples into a lower dimensional space, where the closer the samples the more similar their transcriptomic profiles.

The following sections describe the most widely adopted dimensionality reduction methods to visualize bulk and single cell RNA-sequencing data.

6.1.1 Bulk RNA-Sequencing

In this work, the differences between the sequenced bulk RNA-sequencing samples were explored through the multi-dimensional-scaling (MDS) plot, which can be used to verify the similarity between the transcriptomic profiles of the sequenced samples together with the presence of batch effects.

Shortly, the MDS method [110, 158] plots the differences between samples such that the distance between each pair of samples is computed as the root-mean-square of their largest 500 $\log_2$FCs, which is commonly referred to as the *leading fold change*.

In detail, following the TMM gene count normalization procedure, the MDS procedure was implemented with the `limma` (v.3.52.4) R package [158] through the `plotMDS` R function with default parameters.

6.1.2 Single cell RNA-Sequencing

The Uniform Manifold Approximation and Projection (UMAP) [121] is a non-linear dimensionality reduction procedure which is commonly used for the graphical representation of high dimensional data, such as single cell RNA-sequencing data.

Shortly, UMAP first defines a high-dimensional graph representation of the data and then optimizes a low-dimensional graph to minimize the cross-entropy between the two topological representations.

In detail, following the log-transformation of the gene and cell filtered not-AS CPM count matrix, the analysis was restricted to the 500 most variable genes across all cells and time points. Principal Component Analysis (PCA) was performed on this matrix, which was implemented with the `pcaMethods` (v.1.88.0) R package [182] through the

`pca` R function. The top 50 principal components were then provided as the input to the UMAP dimensionality reduction method, which resulted in a two-dimensional cell embedding.

Specifically, the UMAP procedure was implemented with the `umap` (v.0.2.10.0) R package [121] through the `umap` R function, using 20 nearest neighbors and a minimum distance of 0.5 (respectively, `n_neighbors` and `min_dist` parameters).

6.2 Trajectory inference

When single cell measurements are collected to study a continuous process, every cell can be thought of as an observation along a trajectory defined by the expression similarity between cells. The aim of trajectory inference methods is to assign every single cell to a continuous measure, referred to as *pseudotime*, which measures its distance from the origin of the process along the estimated trajectory.

Most of trajectory inference methods developed for single cell data select a number of highly variable genes across cells and project every observation onto a lower dimensional space that accounts for a large portion of between cell variability, define a manifold structure that connects the most similar cells within the lower dimensional space, identify a set of cells representing the origin of the biological process under investigation, and finally estimate the pseudotime of each cell as the distance between the origin and the projection of the cell onto the estimated manifold [81, 150, 151, 184, 191].

A recent work provides an extensive comparison of several single-cell trajectory inference methods [167]. This analysis highlights that the methods under investigation mostly differ on whether the manifold topology is fixed, and on the type of topologies that each method is able to identify. Based on this method comparison, one of the best performing unsupervised trajectory inference methods (namely `Monocle-DDRTree`) was applied to the time series gene expression data of XX cells, aiming to identify a smooth trajectory and estimate cellular pseudotimes. This method is briefly described in the following section.

6.2.1 Trajectory and pseudotime estimation with Monocle

Monocle is a fully unsupervised trajectory inference method, meaning that it doesn't require any a priori information about which genes drive the continuous biological process under investigation, nor about the topology of the inferred manifold structure. Briefly, this method can be summarized into three major steps: identification of the ordering genes, dimensionality reduction, and pseudotimes calculation [150, 151, 191].

Let $\mathbf{Y}$ be the filtered and normalized expression count matrix which quantifies the expression of G genes across N single cells resulting from a single-cell RNA-seq experiment.

The *ordering genes* are defined as the n genes showing the highest variability across all cells or time points. In the latter case, the normalized expression vector for gene g across all cells ($\mathbf{y}_g \in \mathbb{R}^N$) can be modeled as the response variable of a Generalized Linear Model (GLM) where the time measurement is defined as the only independent variable. The significance of each gene g is then evaluated through a likelihood ratio test (LRT), which compares the full model to an intercept-only model. The ordering genes are then selected as the ones showing the n smallest significance values.

Monocle uses the *Reversed Graph Embedding (RGE)* method to learn a function which projects the data onto a lower dimensional space, while simultaneously learning a graph structure into this space that can be projected back to the higher dimensional space. Let $f_G : \mathbb{R}^d \to \mathbb{R}^G$ be a function which projects the latent points within a lower d-dimensional space ($Z = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$; where $\mathbf{z}_j \in \mathbb{R}^d$) back to the original observed gene expression space ($Y = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$; where $\mathbf{y}_j \in \mathbb{R}^G$), and let $G = (Z, \varepsilon)$ be the graph structure where the edges ε connect the latent vertices $\mathbf{z}_{j=1,\ldots,N}$ [151]. Given a set of graph structures G_b, the RGE method optimizes the objective function:

$$\min_{G \in G_b} \min_{f_G \in F} \min_{Z} \sum_{j=1}^{N} ||\mathbf{y}_j - f_G(\mathbf{z}_j)||^2 + \frac{\lambda}{2} \sum_{(V_i, V_j) \in \varepsilon} b_{ij} ||f_G(\mathbf{z}_i) - f_G(\mathbf{z}_j)||^2 \qquad \text{(Eq. 14)}$$

Where the λ parameter controls the degree of minimization between the first term, which minimizes the distance between observed data and the image associated to their latent values, and the second term, which minimizes the distance in original space between neighboring observations. This might however encounter scalability problems as the number of cells increases. This problem is accounted for by the *DDRTree* method, which solves the above minimization problem on a smaller set of latent points defined as the K centroids of the N observed latent points. Where in the first run of the minimization problem, the K centroids are set to the K-means of the N latent points [151].

Once that an optimal solution is found for the above minimization problem, the *root cell* of the graph is identified as the node with the highest number of cells sequenced at the earliest time point. The latent point of every cell $\mathbf{z}_{j=1,\ldots,N}$ is then projected to the inferred graph, distances are computed between every projection, and a minimum spanning tree (MST) is constructed. The pseudotime of every cell $pdt_{j=1,\ldots,N}$ is then computed as the geodesic distance on the MST from the root cell [151]. Finally, the pseudotime estimate for each cell j is divided by the maximum observed value across all N cells:

$$pdt_j^* = \frac{pdt_j}{\max\limits_{j=1,\ldots,N} pdt_j} \qquad \text{(Eq. 15)}$$

In detail, single cell pseudotime trajectories were constructed using the `monocle` (v.2.12.0) R package, restricting the analysis to XX cells only. Similarly to a previous study [24], a set of ordering genes was defined as the 500 most differentially expressed genes over time points (`differentialGeneTest` R function). The DDRTree method was then used to project the cells into a two-dimensional space based on the expression of the selected genes, and simultaneously learn a graph structure into this space (`reduceDimension(method = "DDRTree")` R function). Pseudotime values were then estimated as the distance of each cell from the root of the graph. Where the root was defined as the state with the highest number of undifferentiated XX cells (`orderCells` R function). The scaled pseudotime values were computed dividing the estimated pseudotimes by the maximum value observed across all XX cells. Finally the pseudotimes estimated for each single cell were visualized onto the first two UMAP cellular embeddings.

6.3 RNA velocity of single cells

One of the central dogma of biology is that upon gene transcription a pre-mRNA transcript is synthesised from the gene's DNA sequence by RNA polymerase enzymes. Notably the pre-mRNA molecules consist of both coding and non-coding RNA-sequences, namely referred to as exons and introns respectively. These molecules are then processed into mature mRNA transcripts in a process known as RNA splicing, where introns are removed and exons are joined together.

A recent work suggests that the high number of intronic reads detected in single cell RNA-seq experiments could be used to quantify the number of unspliced mRNA transcripts. Modeling together the number of spliced and unspliced mRNA molecules observed for each gene and cell would then enable the prediction of the future transcriptomic profile of every sequenced cell [97]. The RNA velocity method is briefly described in the following sections.

6.3.1 Modeling spliced and unspliced mRNA transcripts

Suppose that a single cell RNA-seq experiment results in the quantification of spliced and unspliced transcripts for G genes and N cells, which are respectively estimated by the count matrices $\mathbf{S} = [s_{g,j}], \mathbf{U} = [u_{g,j}] \in \mathbb{N}^{GxN}$.

Intuitively, the number of unspliced molecules increases at every transcription event and decreases at every splicing event. On the other hand, the number of spliced molecules increases at every splicing event and decreases anytime a spliced molecule is degraded. This can be described by the following differential equations:

$$\begin{aligned} \frac{du}{dt} &= \alpha(t) - \beta(t)u_t \\ \frac{ds}{dt} &= \beta(t)u(t) - \gamma(t)s(t) \end{aligned} \qquad \text{(Eq. 16)}$$

Where ($\alpha(t)$, $\beta(t)$, $\gamma(t)$) represent the time-dependent transcription, splicing, and degradation rates respectively. While $u(t)$ and $s(t)$ represent the expected number of unspliced and spliced mRNA transcripts at time point t, respectively.

The above system of differential equations can be simplified assuming constant transcription and degradation rates over time ($\alpha(t) = \alpha$ and $\gamma(t) = \gamma$), and setting the splicing rate to 1 ($\beta(t) = 1$). Whenever the number of spliced reads are in steady state ($\frac{ds}{dt} = 0$), the number of unspliced molecules equals the number of degraded transcripts: $u(t) = \gamma s(t)$. Therefore, the degradation rate of each gene g, can be estimated by ordinary least squares as the slope coefficient of a linear model.

Assuming that the number of spliced reads varies at a fixed rate ($\frac{ds}{dt} = v$), for each cell j and gene g, the predicted number of spliced molecules after a time interval t and the velocities can be estimated as:

$$\begin{aligned} v_{g,j} &= u_{g,j} - \hat{\gamma}_g \cdot s_{g,j} \\ s_{g,j}(t) &= s_{g,j} + v_{g,j} \cdot t \end{aligned} \qquad \text{(Eq. 17)}$$

Where $s_{g,j}$ and $u_{g,j}$ respectively represent the observed number of spliced and unspliced transcripts, $s_{g,j}(t)$ the predicted number of spliced molecules after a time interval t, and $v_{g,j}$ the velocity for gene g and cell j, which is derived by the estimated gene-specific degradation rate $\hat{\gamma}_g$.

6.3.2 Degradation rate estimation in real data analysis

In real data analysis, the gene-specific degradation rate γ_g is estimated as the slope of a linear model with the cell-specific relative unspliced transcripts as response variable ($\mathbf{u}_g = [u_{g,j} \cdot (\sum_{g=1}^{G} u_{g,j})^{-1}] \in [0,1]^N$) and the relative spliced transcripts as the only independent variable ($\mathbf{s}_g = [s_{g,j} \cdot (\sum_{g=1}^{G} s_{g,j})^{-1}] \in [0,1]^N$):

$$\begin{aligned} &\mathbf{u}_g = o + \gamma_g \cdot \mathbf{s}_g + \boldsymbol{\varepsilon} \\ &\text{where: } \boldsymbol{\varepsilon} \sim N(0, \sigma^2) \end{aligned} \qquad \text{(Eq. 18)}$$

Where o represents an optional offset value which accounts for any baseline skewing, and the error term ε is a normally distributed random variable.

The above linear model provides an unbiased estimator for γ_g if and only if the steady state assumption is met. This limitation is addressed through a quantile fit, which restricts the linear fit to the cells resulting in the highest or lowest number of transcripts (by default selecting cells up to the 2.5th percentile, and above the 97.5th percentile), under the assumption that these cells are more likely to be in a steady state.

Notably, the high dropout rates which affects both the spliced and unspliced transcript quantification might bias the linear model fit. This problem is accounted for by pooling the spliced and unspliced transcripts of cells showing similar transcriptomic profiles. For each cell j, a k-Nearest-Neighborhood (kNN) is defined based on the Pearson linear correlation distances computed across all G genes. First the spliced and unspliced transcripts of all cells assigned to each kNN are added up, then a robust estimate of γ_g can be obtained fitting the quantile linear model to the pooled counts.

The velocity estimation is restricted to genes G^* with sufficiently high degradation rates and Pearson's correlations coefficients (by default $\hat{\gamma}_g \geq 0.05$ and $cor(\mathbf{u}_g, \mathbf{s}_g) \geq 0.05$, respectively). Following the gene-specific model fitting, the velocities $v_{g,j}$ and predicted spliced transcripts $s_{g,j}(t)$ can be estimated as described above (Eq. 17).

6.3.3 Projection into a lower dimensional space

The vector of predicted spliced transcripts after a time interval t for every cell j (namely, $\mathbf{s}_j(t) \in \mathbb{R}^{G^*}$) can be visualized onto a linear or non-linear cell embedding.

Suppose that the principal component analysis (PCA) method was applied to the original spliced count matrix $\mathbf{S}$, and that every cell was then projected into the lower dimensional space defined by the first two principal components. The matrix of predicted spliced transcripts after a time interval t (namely, $\mathbf{S}(t) = (\mathbf{s}_1(t), \ldots, \mathbf{s}_N(t))$) can then be projected onto the same two-dimensional PCA plot. Where the coordinates of each predicted cell state $\mathbf{s}_j(t)$ are derived by the eigenvectors of the $\mathbf{S}$ matrix associated to the first two principal components.

For non-linear embeddings (such as UMAP [121]) the position of the predicted cell state is estimated as the one which maximizes the Pearson correlation coefficient between the estimated velocity vector $\mathbf{v}_j$ and the expression difference between every pair of cells $\mathbf{r}_j^i$, with $j, i = 1, \ldots, N$. Let $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_N)$ represent the coordinates of the N cells on a non-linear embedding space. Upon estimating the transition probability matrix between each pair of cells (j, i), namely $\mathbf{P} = [p_{j,i}]$, the projection of the j-th predicted cell state on this embedding is derived by its predicted velocity displacement $\Delta\mathbf{e}_j$, which can be computed as:

$$\Delta \mathbf{e}_j = \sum_{i=1}^{N} (p_{j,i} - 1/N) \cdot \frac{\mathbf{e}_i - \mathbf{e}_j}{||\mathbf{e}_i - \mathbf{e}_j||} \quad \text{(Eq. 19)}$$

For large values of N, individual cell velocities can be summarized by a vector field representation. First the embedding space is partitioned through a grid. Then, for each position on the grid, the grid vector field is estimated applying a Gaussian kernel to cell-specific velocity vectors.

6.3.4 not-AS RNA velocity model fit

The RNA velocity method was applied to the not-AS spliced and unspliced count matrices, removing the genes with low average spliced ($\bar{s}_g \leq 1$) or unspliced ($\bar{u}_g \leq 0.5$) expression across all time points and XX cells. RNA velocities were calculated (`velocyto.R` (v.0.6) R package) setting the cell neighbourhood size to `kCells = 20`, performing the gene-wise fit on the top and bottom 2.5% quantiles (`fit.quantile = 0.025`), and setting the remaining parameters to their default values.

Every XX cell was projected onto a UMAP cell embedding. This was estimated based on the expression of the 500 genes showing the highest variance across all XX cells. The number of variables was further reduced through a principal component analysis based on the mean-centered expression levels of the selected genes (`pca` function from the `pcaMethods` (v.1.76.0) R package). The top-50 PCs were provided as an input to UMAP dimensionality reduction method (`umap` function from the `umap` (v.0.2.3.1) R package) to further reduce dimensionality to two variables. The estimated velocities were then projected onto the UMAP embedding, and locally summarized through a vector field representation of single cell velocities (using the `show.velocity.on.embedding.cor` function from the `velocyto.R` (v.0.6) R package).

6.3.5 Visualization of X-linked RNA velocities

The RNA velocities estimated for X-linked genes were then used to predict the future transcriptional state of the X chromosome for each sequenced XX cell. Aiming to visualize the predicted transcriptomic profile of every XX cell on an embedding space which separates cells silencing one or the other allele, the fraction of spliced UMI counts assigned to the B6 allele was computed for each X-linked gene g and XX cell j as:

$$l_{g,j} = s_{g,j}^{B6} / (s_{g,j}^{B6} + s_{g,j}^{Cast}) \in [0, 1] \quad \text{(Eq. 20)}$$

Where g represents any X-linked gene for which the RNA velocity model could be fitted.

The embedding space which separates XX cells silencing either alleles was then obtained applying the PCA method to the matrix of X-linked allelic ratios, $\mathbf{L} = [l_{g,j}]$, and projecting every XX cell to the space defined by the first two principal components. For every XX cell, the vector of predicted spliced X-linked transcripts (Eq. 17) was first projected onto the PCA embedding space, and then summarized by a vector field representation of single cell velocities, as described above.

Notably, the PCA loadings of the first two components correspond to those X-linked genes which account for most of the variance in the X-linked allelic ratios, and therefore are the most important in determining differential silencing between the two alleles.

6.3.6 Predicted change in X chromosome expression

Based on the previously described not-AS RNA velocity fit, the predicted change in X-linked gene expression for every XX cell j was measured as:

$$\Delta X_j = \frac{\sum\limits_{g \in chrX} \left(s_{g,j} \Big/ \sum\limits_{g} s_{g,j} \right)}{\sum\limits_{g \in chrX} \left(s_{g,j}(t) \Big/ \sum\limits_{g} s_{g,j}(t) \right)} \qquad \text{(Eq. 21)}$$

Where g represents any X-linked gene for which the RNA velocity model could be fitted; and $s_{g,j}(t)$ represents the predicted expression for the j-th sample after a time interval t (Eq. 17).

Intuitively, high values indicate that the cell is predicted to undergo XCI, while low values indicate that the X chromosome is predicted to increase its global gene expression.

7 Differential expression analysis

Suppose that the samples under investigation can be partitioned into two or more groups representing different experimental conditions or cell sub-populations. The aim of differential expression analysis is to identify a subset of genes showing significantly different expression levels between the groups, therefore indicating a causal relation between their expression values and the phenotype of interest.

Suppose that half of the samples sequenced through a bulk RNA-sequencing experiment belong to group A and the other half to group B, and let $\bar{y}_{g,A}$ and $\bar{y}_{g,B}$ denote the average $\log_2$-expression values of the g-th gene observed on samples from group A and B respectively. The $\log_2$-Fold Change ($\log_2 FC$) for the g-th gene can then be defined as:

$$\log_2 FC_g = \bar{y}_{g,A} - \bar{y}_{g,B} \qquad \text{(Eq. 22)}$$

In order to assess whether the g-th gene is differentially expressed between two conditions, the null hypothesis $H_0 : \log_2 FC_g = 0$ can be tested. If this hypothesis is not accepted then the g-th gene is said to be differentially expressed (DE) between the two groups, or otherwise not differentially expressed (not-DE). The DE genes may be further divided into up-regulated or down-regulated genes, meaning that the average $\log_2$-expression values observed for group A is significantly higher or lower than the one for group B, respectively.

7.1 Negative Binomial regression

Any experimental design can be represented in terms of a gene-wise linear model. Let $\mathbf{y}_g = (y_{g,1}, \ldots, y_{g,N})^T$ be the vector of $\log_2$-expression values observed for the gene g on N samples. Then for each gene, a linear model can be defined as [161, 177]:

$$\mathbf{y}_g = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{(Eq. 23)}$$

Where $X \in \mathbb{R}^{N \times p}$ is a full column rank design matrix, which describes the study design specifying how the different treatment or group labels are assigned to the samples. Moreover $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown coefficients whose components represent the treatment effects or contrasts associated with the gene g. Finally $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ is the vector of N errors or residuals, which are assumed to be multivariate normals with an unknown correlation structure between the genes [177]. Suppose for example that in a case-control experiment G genes have been analyzed on four samples. Where the first two samples are controls and the latter two are cases. Then the matrices defining each of the G gene-wise linear models may be written as:

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & y_{1,3} & y_{1,4} \\ y_{2,1} & y_{2,2} & y_{2,3} & y_{2,4} \\ \ldots & \ldots & \ldots & \ldots \\ y_{G,1} & y_{G,2} & y_{G,3} & y_{G,4} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \ldots \\ \mathbf{y}_G^T \end{bmatrix} ; \; X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} ; \boldsymbol{\beta} = \begin{bmatrix} \beta_{g,0} \\ \beta_{g,1} \end{bmatrix} \qquad \text{(Eq. 24)}$$

Suppose that the contrast between cases and controls $\beta_{g,1}$ is of biological interest. Then we can define the vector $\mathbf{c} = (0, 1)^T$ such that $\mathbf{c}^T\boldsymbol{\beta}_g = \beta_{g,1}$. The coefficient $\beta_{g,1}$ estimates the treatment effect associated with the $\log_2$-expression values of gene g such that $\hat{\beta}_{g,1} = \bar{y}_{g,cases} - \bar{y}_{g,controls}$. Importantly it can be noticed that this coefficient estimates the g-th

$\log_2$-fold change (log_2FC_g) relative to that contrast. Hence the null hypothesis for the g-th gene can be tested estimating the coefficients of the gene-wise linear models [177].

Given the count nature of bulk RNA-sequencing data, the expression of the g-th gene is usually assumed to follow a Negative Binomial (NB) distribution [160, 161, 177]:

$$\begin{aligned} \mathbf{y}_g &\sim NB(\mu_g, \phi_g) \\ E(\mathbf{y}_g) &= \mu_g \\ V(\mathbf{y}_g) &= \mu_g + \phi_g \cdot \mu_g^2 \end{aligned} \qquad \text{(Eq. 25)}$$

Where μ_g and ϕ_g are respectively the average and dispersion parameters for the g-th gene. Such that the Negative Binomial reduces to a Poisson distribution if $\phi_g = 0$. The log-transformed normalized expression of each gene can then be fitted through a Generalized Linear Model (GLM), and the gene-wise null hypotheses $H_0 : \log_2 FC_g = \hat{\beta}_{g,1} = 0$ be tested through a likelihood ratio test (LRT) or an F-test [109, 160, 161].

7.2 Model-based Analysis of Single Cell Transcriptomics (MAST)

Similarly to the count data normalization procedures, issues might arise when applying methods developed for bulk RNA-sequencing data to zero-inflated counts such as the ones resulting from a single cell RNA-seq experiment. Although the performance of bulk methods applied to single-cell RNA-Seq data is still very much debated, a large number of approaches accounting for the bi-modal distribution of single cell gene count data have been recently proposed. Specifically, some recent DE analysis method comparisons highlighted that while bulk methods have the tendency to result in a large number of false discoveries when dealing with lowly expressed genes with high dropout rates, the majority of genes don't need the inclusion of an extra component to separately model the dropout events [37, 51, 179].

Among the available methods for differential expression analysis, MAST (Model-based Analysis of Single Cell Transcriptomics) [64] was pointed out as one of the approaches showing the best performance for a number of real and simulated single cell RNA-seq datasets. Briefly, this method fits a two-part GLM Hurdle model which separately models a discrete variable representing the presence of zero counts and a continuous one which refers to strictly positive gene expression values, where the continuous model parameters are estimated through an Empirical Bayes method.

Let $CPM_{g,j}$ represent the normalized CPM values observed for gene g and cell j. Let the *Cellular Detection Rate (CDR)* for cell j represent the proportion of expressed genes, such that:

$$
\begin{aligned}
&Z_{g,j} = 0, \quad if \quad CPM_{g,j} = 0 \\
&Z_{g,j} = 1, \quad if \quad CPM_{g,j} > 0 \\
&CDR_j = \frac{1}{N} \sum_{g=1}^{G} Z_{g,j}
\end{aligned}
\qquad \text{(Eq. 26)}
$$

Therefore the random variables CPM and Z are conditionally independent. A Hurdle model can then be used to fit a two part GLM which separately models the zero counts through a logistic model, and the strictly positive expression values through a Gaussian linear model, with mean $X_g \beta_g^C$ and variance equal to σ_g^2 [64]:

$$
\begin{aligned}
&\text{logit}\big(P(Z_{g,j} = 1)\big) = X_g \beta_g^D \\
&P\big(Y_{g,j} = y \mid Z_{g,j} = 1\big) = N\big(X_g \beta_g^C,\ \sigma_g^2\big)
\end{aligned}
\qquad \text{(Eq. 27)}
$$

Since the cellular detection rate accounts for a large proportion of gene expression variability between cells, this nuisance parameter is usually included as a column of the design matrix X to account for differences in dropout rates across cells.

Suppose that a contrast between two groups of cells β_g is of biological interest, and that we aim to identify the subset of genes showing the most significant difference in gene expression levels between these two groups. Then, gene-wise differential expression can be tested summing up χ^2 distributed statistics (such as LRT or Wald tests) computed on the discrete ($Z_{g,j}$) and continuous ($Y_{g,j}$) components. Given the gene-wise conditional independence between these two components, the resulting statistic will still follow a χ^2 distribution, with a number of degrees of freedom (dof) equal to the sum of the dof of the discrete and continuous components [64].

Parameter estimation of the discrete component (β_g^D) is achieved through a Bayesian approach, setting its prior to a Cauchy distribution centered at zero. On the other hand, the continuous variance (σ_g^2) is regularized through an Empirical Bayes' method by setting a Gamma prior to the precision parameter ($\frac{1}{\sigma_g^2} \sim Gamma(\alpha_0, \beta_0)$), obtaining maximum likelihood estimates (MLE) for the two hyper-parameters (α_0, β_0), and estimating the posterior precision as a weighted average between its MLE and its prior estimate.

7.3 Identification of putative *Xist* and XCI regulators

Aiming to identify candidate regulators of *Xist* expression and XCI initiation, we used two different approaches based on differential expression and correlation analyses to

ensure the robustness of our results. These two approaches are described in the following sections.

7.3.1 Differential expression analysis

MAST differential expression analysis method was applied to the not-AS count matrix $\mathbf{Y}$ in order to identify a set of putative regulators of the *Xist* gene and of the XCI process. This set of genes was identified by two separate DE analyses.

In the first analysis, putative *Xist* regulators were identified comparing the gene-wise expression levels of cells showing high and low *Xist* expression. First, XX cells across all time points were clustered with K-means algorithm ($K = 7$) based on the logarithm of *Xist* not-AS CPM expression values, namely $\log_{10}(CPM^{\mathbf{Y}}_{Xist,j} + 1)$. Cells belonging to the top and bottom 3 K-means groups were classified as *Xist-high* and *Xist-low*, respectively. Where the number of clusters K was set in a way to minimize the within-cluster sum of squares value, while ensuring a minimum number of 50 cells in the two groups at each time point of differentiation (days 1-4). The set of differentially expressed genes between these two groups of cells were identified at each time point fitting gene-wise Hurdle GLMs (`zlm` function from the MAST (v.1.10.0) R package), including as model predictors a dummy variable representing the two cell groups and the cellular detection rate (Eq. 26). The significance of each gene was then assessed through a χ^2 likelihood ratio test, and corrected for the multiple hypotheses testing issue through the Benjamini-Hochberg (BH) procedure [12, 64].

The second analysis aims to identify genes regulating the XCI process comparing the gene-wise expression levels of cells predicted to decrease or increase the expression of chromosome X by the previously described RNA velocity method. First, XX cells at each time point were clustered with K-means algorithm ($K = 3$) based on their predicted change in X-linked expression $\log_2(\Delta X_j)$ values (Eq. 21). Separately for each time point, the set of differentially expressed genes between the cells in the top and bottom clusters were identified through the MAST differential expression analysis method, as described above. Where the cells' K-means clusters and cellular detection rate were again included as independent variables.

7.3.2 Correlation analysis

Similarly to the previous analyses, a set of putative *Xist* and XCI regulators were identified performing gene-wise correlation analyses with respect to *Xist* not-AS CPM expression and the RNA velocity predicted change in X-linked expression, respectively.

For each time point throughout cellular differentiation and gene g, the Spearman's correlation coefficient (ρ_g^{Xist}) between the not-AS CPM expression values observed for *Xist*

and for the g-th gene was computed. Similarly, the Spearman's correlation coefficient ($\rho_g^{\Delta X}$) was computed between the not-AS CPM expression of each gene g and the RNA velocity predicted change in X-linked gene expression.

For each time point and gene g, the null hypotheses $H_0 : \rho_g^{Xist} = 0$ and $H_0 : \rho_g^{\Delta X} = 0$ were tested through a Student's T test, and the significance values were corrected for the multiple hypotheses testing issue through the Benjamini-Hochberg (BH) procedure [12].

8 Differential silencing analysis

This analysis accounts for the global silencing differences between the two parental alleles aiming to classify X-linked genes based on their allele specific silencing kinetics, and to identify the subset of genes showing significantly different silencing dynamics between the two alleles.

8.1 Robust measures of silencing progress

Since analyzing the allelic expression of individual genes in single cells tends to be noisy, the cells with similar extent of global X-silencing (XP_j, Eq. 12) were grouped and the overall and gene-wise silencing progress measures were robustly estimated by aggregating the UMI counts of cells in each group. The following steps were performed separately for *Xist* monoallelic XX cells silencing the B6 (MA-B6) and Cast (MA-Cast) X chromosome.

First, the analysis was restricted to cells which had already initiated the XCI process (namely, $j : XP_g > 10\%$). Then, these cells were divided into $B = 10$ equally sized bins based on the XP observed across the two *Xist* monoallelic populations. For each *Xist* monoallelic population and b-th bin (where $b = 1, \ldots, B$), the binned overall and gene-wise silencing progress (XP_b and $\text{Xi:Xa}_{g,b}$, respectively) were estimated aggregating the allele specific UMI counts observed across all cells assigned to the same bin, as:

$$
\begin{aligned}
XP_b &= 100 \cdot \left(1 - \frac{\sum_{j \in b} \sum_{g \in chrX} y_{g,j}^{Xi} + 0.01}{\sum_{j \in b} \sum_{g \in chrX} y_{g,j}^{Xa} + 0.01}\right) \\
\text{Xi:Xa}_{g,b} &= \frac{\left(\sum_{j \in b} y_{g,j}^{Xi}\right) + 0.01}{\left(\sum_{j \in b} y_{g,j}^{Xa}\right) + 0.01} \\
&\text{if } j \in \text{MA-B6} : Xi = B6 \text{ and } Xa = Cast \\
&\text{if } j \in \text{MA-Cast} : Xi = Cast \text{ and } Xa = B6
\end{aligned}
\qquad \text{(Eq. 28)}
$$

Where XP_b quantifies the extent of X inactivation across all cells assigned to the b-th bin; while $\text{Xi:Xa}_{g,b}$ is a proxy for the extent of inactivation of a specific gene in that bin. Intuitively, a value of $\text{Xi:Xa}_{g,b}$ close to zero indicates that the X-linked gene has been completely silenced on the Xi allele, while a value proximal to one indicates that the two alleles have similar gene expression levels. For each gene and allele, the above measures were computed only for bins containing a minimum of 5 cells and a total of at least 25 AS counts, and to genes with a minimum of 5 such bins.

In order to account for basal expression skewing due to genetic variations between the two alleles, the $\text{Xi:Xa}_{g,b}$ values were then scaled by the allelic gene ratios computed by aggregating the allele specific UMI counts of all undifferentiated *Xist*-negative cells with similar expression levels of the two X chromosomes (bc_g). The normalized binned gene silencing measures ($\text{Xi:Xa}^*_{g,b}$) were then computed as:

$$
\begin{aligned}
bc_g &= \frac{\big(\sum_{j \in J} y^{B6}_{g,j}\big) + 0.01}{\big(\sum_{j \in J} y^{Cast}_{g,j}\big) + 0.01} \\
\text{Xi:Xa}^*_{g,b} &= \text{Xi:Xa}_{g,b} / \text{Xi:Xa}_{g,baseline} \\
\text{where: } \text{Xi:Xa}_{g,baseline} &= \begin{cases} bc_g, & \text{if Xi = B6} \\ bc_g^{-1}, & \text{if Xi = Cast} \end{cases}
\end{aligned}
\qquad \text{(Eq. 29)}
$$

Where J represents the set of undifferentiated (day 0) *Xist*-negative cells with $XR_j \in [0.4, 0.6]$.

8.2 Silencing halftimes

The silencing halftime of a gene ($XP_{50,g}$) represents the extent of global X chromosome silencing at which the inactive allele (Xi) reduces its expression of the g-th gene by half relative to the active allele (Xa).

As the silencing halftime is expected to follow an exponential decay distribution, this measure was estimated modeling the log-transformed normalized gene silencing rates for gene g on chromosome X ($\text{Xi:Xa}^*_{g,b}$) as linear function of the X chromosome silencing process (XP_b) separately for the two alleles, as:

$$E[\log_2(\text{Xi:Xa}^*_{g,b})] = \beta_{g,1} \cdot XP_b + \beta_{g,2} \cdot XP_b \cdot A$$
$$\text{where: } A = \begin{cases} 0, & \text{if Xi = B6} \\ 1, & \text{if Xi = Cast} \end{cases} \qquad \text{(Eq. 30)}$$

Where A represents a dummy variable, which enables the above model to estimate the gene silencing rates for *Xist* monoallelic cells silencing the B6 allele ($\hat{\beta}_g^{B6} = \hat{\beta}_{g,1}$) or silencing the Cast allele ($\hat{\beta}_g^{Cast} = \hat{\beta}_{g,1} + \hat{\beta}_{g,2}$). The gene-specific silencing halftimes for the two alleles were then computed as the XP_b value corresponding to a Xi:Xa$^*_{g,b}$ ratio of 0.5, as:

$$XP_{50,g} = \begin{cases} -1/\hat{\beta}_{g,1}, & \text{if Xi = B6} \\ -1/\left(\hat{\beta}_{g,1} + \hat{\beta}_{g,2}\right), & \text{if Xi = Cast} \end{cases} \qquad \text{(Eq. 31)}$$

Where the $XP_{50,g}$ values greater than 100 were set equal to 100.

Intuitively, a $XP_{50,g}$ value close to 0 indicates that the g-th gene is silenced at the earliest stages of XCI, while increasingly higher values identify genes silenced at a later stage of XCI or escaping the silencing process.

For each allele, the K-means clustering algorithm ($K = 4$) was applied to the estimated $XP_{50,g}$ values in order to assign every gene to a silencing dynamics class (fast, intermediate, slow, escape).

8.3 Identification of differentially silenced genes

This analysis aims to identify a subset of genes showing significantly different silencing kinetics on the two X chromosomes.

In order to identify such genes, the fit of the allele specific linear model (m_1, Eq. 30) was compared to the one of a simpler linear model which fits both *Xist* monoallelic populations with a single slope (m_0):

$$
\begin{aligned}
&H_0 : \beta_{g,2} = 0 \\
&m_1 : E[\log_2(\text{Xi:Xa}^*_{g,b})] = \beta_{g,1} \cdot XP_b + \beta_{g,2} \cdot XP_b \cdot A \\
&m_0 : E[\log_2(\text{Xi:Xa}^*_{g,b})] = \beta_{g,1} \cdot XP_b \\
&\text{where: } A = \begin{cases} 0, & \text{if Xi = B6} \\ 1, & \text{if Xi = Cast} \end{cases}
\end{aligned}
\qquad \text{(Eq. 32)}
$$

For every gene g, an ANOVA F test was used to assess whether the $\beta_{g,2}$ parameter was significantly different from zero. Intuitively, a significant result suggests considerably different silencing kinetics on the two alleles. Otherwise the two alleles are deemed to have similar gene silencing trends, which can be recapitulated by a single silencing rate $\beta_{g,1}$. Finally, any gene with a Benjamini-Hochberg adjusted p-value [12] smaller or equal than the nominal FDR=0.05 was deemed as differentially silenced between the two alleles.

9 TXΔXic data analyses

Upon differentiation both the TXΔXic$_{B6}$ and TXΔXic$_{Cast}$ mESCs underwent non-random XCI, expressing *Xist* and silencing the genes on the wild type X chromosome. Gene expression was measured every 24 hours throughout four days of cellular differentiation through bulk RNA-sequencing and pyrosequencing. These data were analyzed to validate the global and gene-specific allelic differences in gene silencing dynamics revealed by the analyses of single cell RNA-seq data.

9.1 Pyrosequencing data analysis

Let $p_{g,j}$ be a random variable representing fraction of B6 molecules observed for the X-linked gene g on replicate sample j at a specific time point.

For each day of cellular differentiation, the gene-specific Xi:Xa ratio was computed as:

$$
\begin{aligned}
&\text{Xi:Xa}_{g,j} = (1 - p_{g,j})/p_{g,j}, \text{ if j} \in TX\Delta Xic_{B6} \\
&\text{Xi:Xa}_{g,j} = p_{g,j}/(1 - p_{g,j}), \text{ if j} \in TX\Delta Xic_{Cast}
\end{aligned}
\qquad \text{(Eq. 33)}
$$

In order to account for basal expression skewing due to genetic variations between the two alleles, the above ratios were divided by the average ratio observed across the undifferentiated (day 0) biological replicates.

For each gene deemed as significant in the single cell differential silencing analysis and time point, the normalized allelic ratios for the two deletion lines were compared through an unpaired t-test statistic. Where significant results identify genes showing differential silencing on the two alleles at specific time points throughout cellular differentiation.

For each gene deemed as not significant in the single cell differential silencing analysis and time point, the normalized allelic ratios were averaged across replicates and compared between the two deletion lines through a Wilcoxon signed-rank test. Where a significant result at a specific time point of differentiation highlights the global difference in X chromosome silencing kinetics between the two alleles.

9.2 Bulk RNA-sequencing data analysis

For each time point and deletion cell line (namely, TXΔXic$_{B6}$ and TXΔXic$_{Cast}$), the extent of silencing of each X-linked gene g on the wild type X chromosome (Xi) relative to the allele carrying the deletion (Xa) was measured summing up the counts observed across replicate samples j:

$$\begin{aligned} \text{Xi:Xa}_g &= \frac{\sum_j b_{g,j}^{Xi} + 1}{\sum_j b_{g,j}^{Xa} + 1} \\ &\text{if j} \in TX\Delta Xic_{Cast}: \text{ Xi = B6 and Xa = Cast} \\ &\text{if j} \in TX\Delta Xic_{B6}: \text{ Xi = Cast and Xa = B6} \end{aligned} \tag{Eq. 34}$$

Where X-linked genes g within the deleted region (chrX: 103,182,257 - 103,955,531, mm10) or with less than 50 allelic counts across replicates were excluded from the analysis. For each time point, the gene-specific allelic ratios observed in the two deletion lines were compared through a paired t-test statistic. Where significant results highlight differences in global silencing dynamics of the two X chromosomes at specific time points throughout cellular differentiation.

Furthermore, the differential silencing of individual X-linked genes g across the two deletion lines was inspected by measuring for each replicate sample j:

$$\begin{aligned} &\text{Xi:Xa}_{g,j} = b_{g,j}^{Xi} / b_{g,j}^{Xa} \\ &\text{if j} \in TX\Delta Xic_{Cast}: \text{ Xi = B6 and Xa = Cast} \\ &\text{if j} \in TX\Delta Xic_{B6}: \text{ Xi = Cast and Xa = B6} \end{aligned} \tag{Eq. 35}$$

Basal genetic skewing was again accounted for on each deletion line dividing the above ratios by the average allelic ratio across undifferentiated replicates. For each time point and gene g, the normalized allelic ratios of the two deletion lines were compared through an unpaired t-test statistic. Where significant results identify genes showing differential silencing on the two alleles at specific time points of cellular differentiation.

Furthermore, a differential silencing analysis similar to the one described for single cell RNA-seq data was performed. This analysis was restricted to X-linked genes with a minimum of 500 allele specific counts across biological replicates at each time point of cellular differentiation (days 1-4) in both deletion cell lines. Similarly to the procedure described to identify differentially silenced genes for single cell RNA-seq data, basal genetic skewing was accounted for dividing the gene-wise Xi:Xa ratios (Eq. 34) by the value observed across undifferentiated replicates (day 0), and the XCI progress (Eq. 12) was measured at each time point across replicate samples excluding the genes within the deleted region. For each X-linked gene included in this analysis, the allele specific silencing halftime was measured fitting an intercept free gene-wise linear model to the Xi:Xa and XP values observed on each cell line (Eq. 30, Eq. 31). Finally, a set of putative differentially silenced genes was identified comparing the fit of the gene-wise allele specific linear model to a simpler model fitting the data from both deletion cell lines with a single slope (Eq. 32). Where the significance of each gene was again inspected through an ANOVA F test.

3 Results

1 Experimental design and single cell RNA sequencing

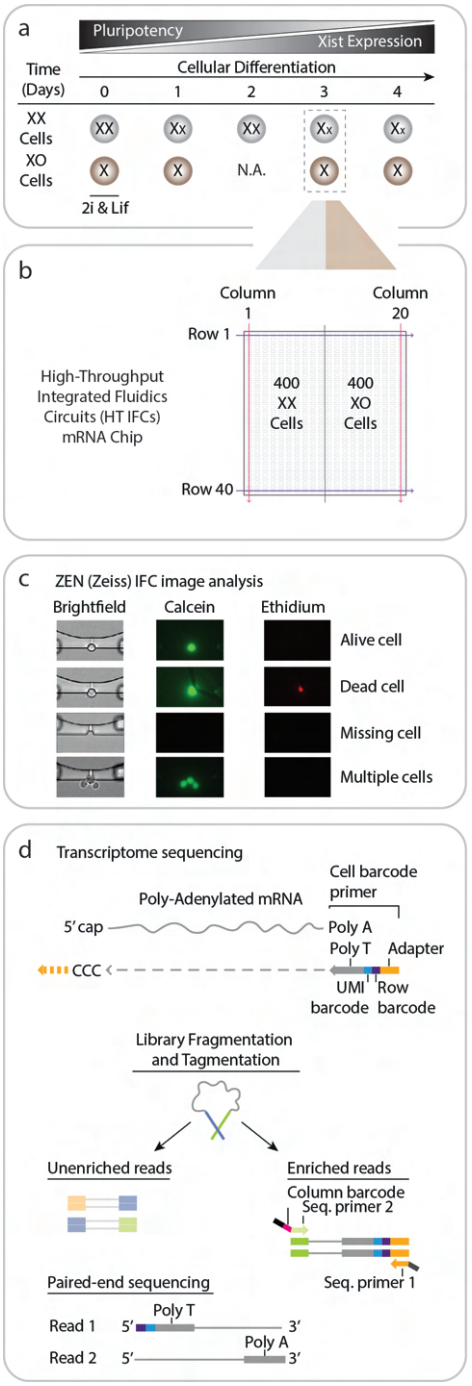


FIGURE 1: Experimental design (a); Sequencing chip (b); Image analysis (c); RNA sequencing (d)

This section describes the experimental design and single cell RNA-Sequencing which was performed on the TX1072 cell line aiming to study the transcriptomic profiles of female XX and XO mouse embryonic stem cells (mESCs) throughout cellular differentiation.

XX and XO cells were cultured in 2i&LIF conditions, which preserves their fully pluripotent and undifferentiated cellular profiles (day 0). Differentiation was then induced by 2i&LIF withdrawal, and the transcriptomic profiles of samples from the XX and XO populations were collected and measured every 24 hours throughout four days of induced cellular differentiation (days 1-4). Notably the XO mESCs after two days of induced cellular differentiation could not be sequenced due to a mistake in the cell loading procedure. Loss of pluripotency leads to up-regulation of the *Xist* gene which initiates the XCI process in the XX population, while their XO counterparts do not undergo XCI given the absence of a second X chromosome (Fig. 1a).

Xist up-regulation and XCI are very asynchronous processes especially when culturing mESCs in vitro, moreover every XX cell independently chooses and transcriptionally silences one of the two X chromosomes. This leads to high levels of heterogeneity in gene expression levels and in X chromosome silencing extent across cells sequenced at the same time point. For these reasons we opted for a single cell RNA sequencing approach rather than a bulk assay.

Indeed, while the latter estimates the average gene expression profile across an entire population of heterogeneous cells, the former enables to define the transcriptomic profile of each individual cell.

The Single-Cell mRNA Seq HT integrated fluidic circuit (IFC) microfluidics system enables the simultaneous capture and isolation of single cells from two different samples. Every sequencing chip is divided into two sections, composed by 10 separate columns with 40 IFCs each. At each time point throughout cellular differentiation (days 0-4), XX and XO cells were separately loaded onto the two halves the sequencing chip, and a maximum of 400 cells could be isolated and sequenced on each half. Every cell isolated on the 800 IFCs is uniquely identified by a combination of row and column barcodes, which is essential for the following sequencing reads demultiplexing procedure (Fig. 1b).

Prior to mRNA sequencing, every IFC was inspected through a Zeiss CellDiscoverer microscope (Zeiss) with a 20x objective. Capture sites without a cell or with multiple cells were identified by manual inspection based on brightfield imaging, while the intensity signals of the life and dead stain fluorophores (Calcein and Ethidium, respectively) were quantified using ZEN v2.3 software (Zeiss) (Fig. 1c).

All the cells isolated within the same column were jointly harvested, and their cellular membranes sheared to proceed with RNA extraction and sequencing. The polyadenylated (polyA) mRNA molecules were enriched through a polyT primer which was ligated to: a random 5-mer nucleotide sequence (UMI barcode), a cell-specific 6-mer row barcode, and a sequencing primer. Single stranded cDNA libraries were then synthesized by reverse transcription. Each of the 20 cDNA libraries was then amplified, fragmented and tagmented. The fragments that contained the polyT primer were amplified by PCR. Every enriched molecule was then paired-end sequenced, where read 1 (R1) stores both the row and UMI barcodes, while read 2 (R2) the 3'-end biased cDNA sequence (Fig. 1d).

2 Read alignment and gene quantification

For each time point throughout cellular differentiation, the sequencing reads deriving by the single cell RNA-seq experiment were demultiplexed and assigned to their cell of origin using the row and column barcodes. The following subsections describe the results of the sequencing, reads alignment and gene counting procedures.

2.1 Sequencing, alignment and gene counting throughput

Figure 2a summarizes the read alignment and gene expression quantification procedures for an example gene. The sequencing reads are first aligned to the custom SNP-masked

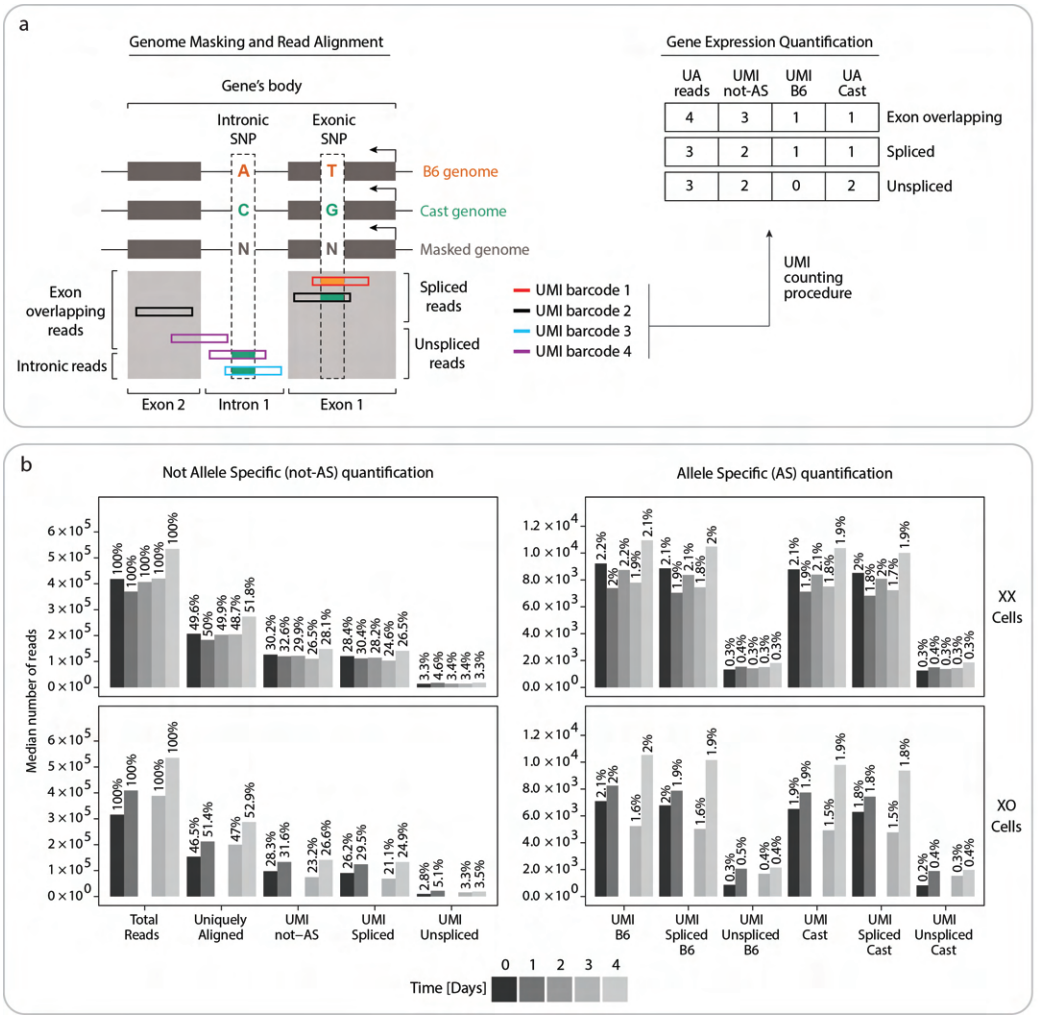


FIGURE 2: (a) Read alignment and gene expression quantification scheme for an example gene, defined by two exons and one intron. Uniquely aligned reads are portayed as rectangles, while their color identifies their UMI barcode sequence. (b) Single cell RNA-seq sequencing, alignment, not-AS and AS gene expression quantification throughput throughout cell lines and cellular differentiation.

mouse genome. Where this procedure aims to account for the presence of polymorphisms between the two distantly related parental genomes. Importantly, for each gene under investigation, the reads carrying an identical UMI barcode sequence can be assumed to derive from the same mRNA molecule. The overall number of transcribed mRNA molecules is quantified restricting the UMI counting procedure to the uniquely aligned (UA) reads with at least one nucleotide overlapping any of the gene's annotated exons (Exon overlapping). The number of spliced and unspliced mRNA molecules are estimated restricting the analysis to the reads entirely aligned to exonic regions (Spliced reads) and to the ones with at least one nucleotide overlapping any intronic region (Unspliced reads), respectively (Fig. 2a, left). For each subset of UA reads, the number of transcribed mRNA molecules is estimated as the number of unique UMI barcodes (UMI not-AS). Similarly, allelic gene expression quantification is performed assigning SNP-spanning reads to their parental genome according to the observed SNP-genotype,

and repeating the above UMI counting procedure for each subset of allele-specific reads (UMI B6/Cast). The number of not-AS and AS UMI counts relative to the exon-overlapping, spliced and unspliced gene expression quantification for this example gene are summarized by the table (Fig. 2a, right).

The bar plots (Fig. 2b) summarize the throughput of the single cell sequencing, alignment, not-AS and AS gene quantification for XX and XO cells sequenced throughout cellular differentiation. The sequencing procedure resulted in a median of around 400,000 reads per cell, where around half of these were uniquely aligned to the masked mouse genome. Around 30% of all the sequenced reads were exon-overlapping and carried a unique UMI barcode, hence were used to quantify the not-AS expression of annotated mouse genes. Almost all the exon overlapping reads were completely aligned within exonic regions and therefore used to quantify the number of spliced mRNA molecules, while the number of unspliced molecules was quantified based on around 3.5% of the total number of sequenced reads. On the other hand allele-specific gene expression was quantified by around 4% of the total number of reads, which were almost equally divided between the two parental alleles (UMI B6 and UMI Cast). Spliced AS quantification was carried out based on a slightly reduced fraction of total reads, while around 0.6% of total reads were used for unspliced AS gene quantification. These results highlight that the sequencing of XX and XO cells showed similar throughput throughout cellular differentiation.

2.2 Xist 5'-biased read coverage

The C1-HT paired-end sequencing protocol enriches and sequences poly-adenylated (polyA) mRNA molecules, resulting in sequencing reads which align to the 3' end of the gene's transcript. The expected 3'-end read alignment bias can be observed looking at the read coverage track across all sequenced cells of a number control of genes (Fig. 3a).

Notably, the expected 3' bias was not observed for the *Xist* gene (Fig. 3b). Rather its reads aligned to *Xist*'s first annotated exon, and were almost completely absent at the 3' end of its transcripts. Interestingly, this unexpected read coverage is not specific to the protocol being used, as it was also observed in a previous study which used the CEL-Seq2 3'-end biased scRNA-sequencing protocol [83]. Inspection of the sequence upstream of the *Xist* 5'-alignment peak revealed a 25 base pair (bp) long genomically encoded polyA sequence. Therefore, it can be hypothesized that the internal polyA sequence serves as a template to prime reverse transcription, and that *Xist* mRNA's polyA tail is likely inaccessible to the reverse transcription reaction.

Xist's read coverage was summarized through a heatmap, which shows the 100bp-binned *Xist* minus strand read coverage in XX cells throughout cellular differentiation (Fig. 3c).

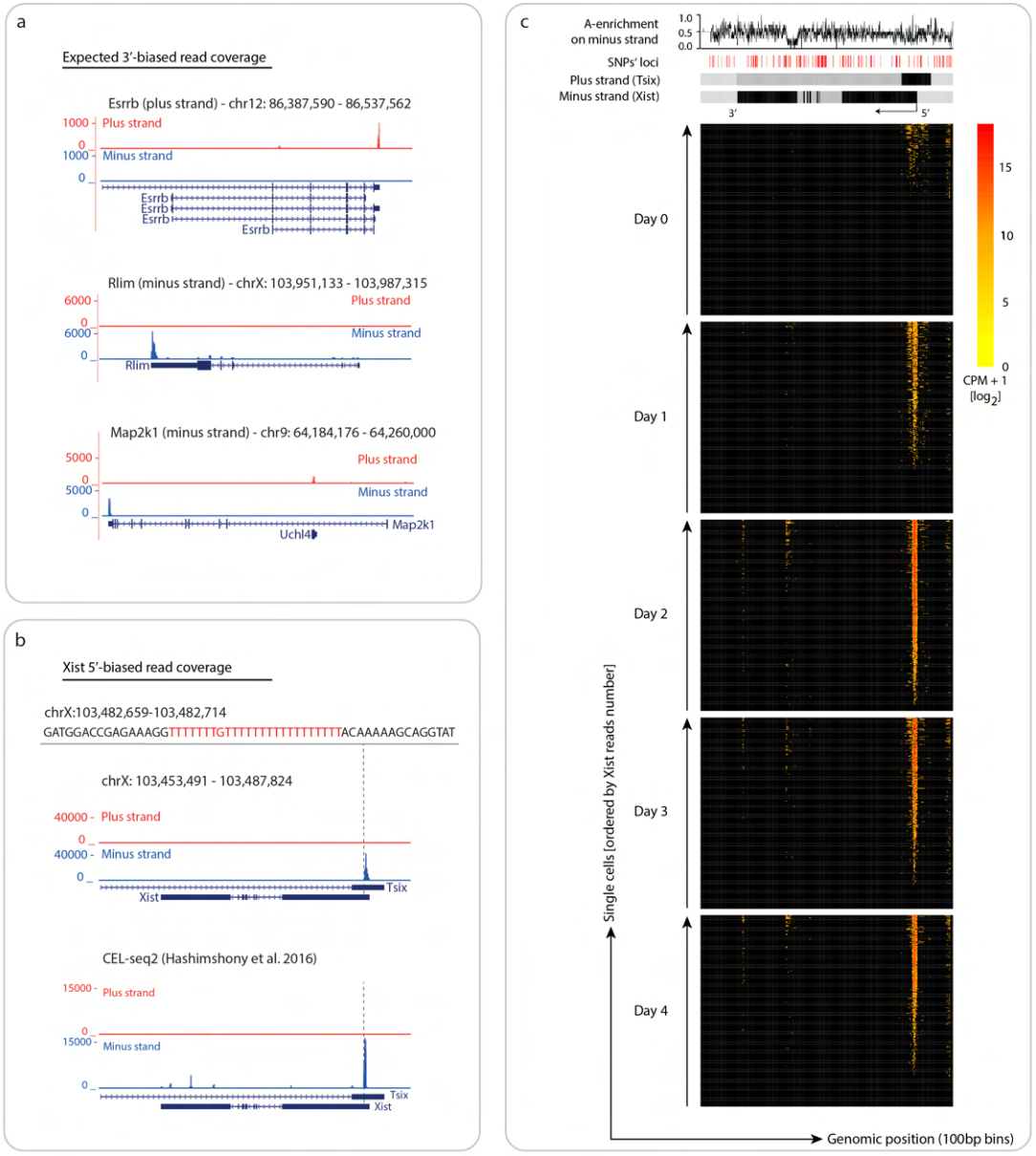


FIGURE 3: (a) 3'-end reads coverage for three control genes (Esrrb, Rlim, Map2k1). (b) Composite track showing *Xist* 5'-biased read coverage and a 25bp internal polyA sequence on *Xist* locus. (c) *Xist* (100bp-binned) read coverage in XX cells over time. Every row of the heatmap represents a single XX cell, colored and ordered by *Xist* expression. Every column of the heatmap represents a 100bp bin across the *Xist* gene locus. The plots above the heatmap highlight *Xist* and Tsix gene annotations, B6/Cast SNPs loci, and a 10bp running mean showing the Adenine enrichment on the minus strand.

Regardless of the unexpected reads' alignment *Xist* expression showed the expected up-regulation throughout cellular differentiation, characterized by almost none of undifferentiated mESC expressing the gene and increasing expression throughout differentiation. This led us to conclude that the expression of *Xist* can correctly quantified by its 5'-biased read coverage. Furthermore the plot highlights that the Xist expression is very heterogeneous and asynchronous throughout differentiation, indeed we can observed high

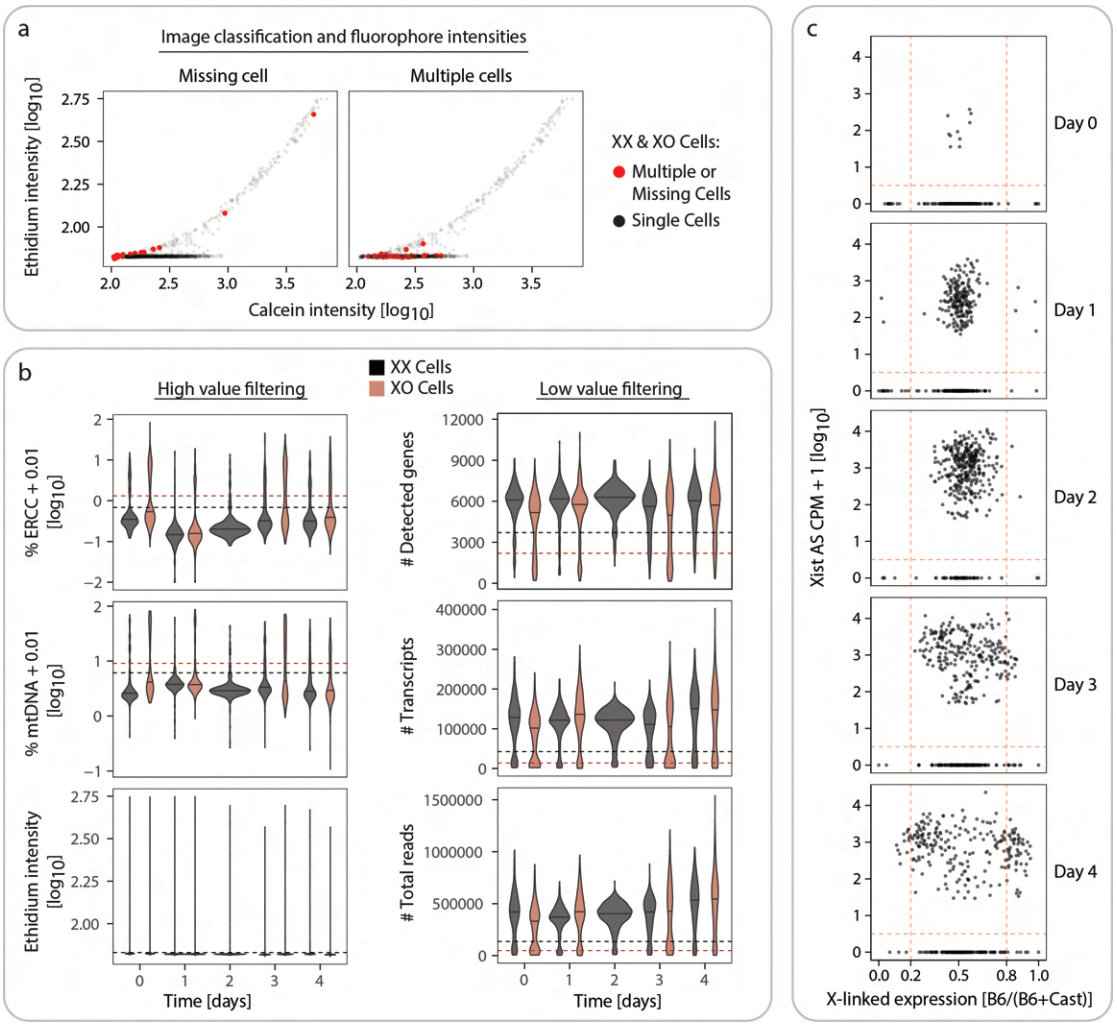


FIGURE 4: (a) Fluorophore intensities and image classification. Capture sites which isolated multiple or no cells are highlighted in red. (b) Violin plots of the observed filtering measures, and 3-MAD cell filtering thresholds for XX and XO cells (black and brown dashed lines, respectively). (c) XO cells filtering in XX population

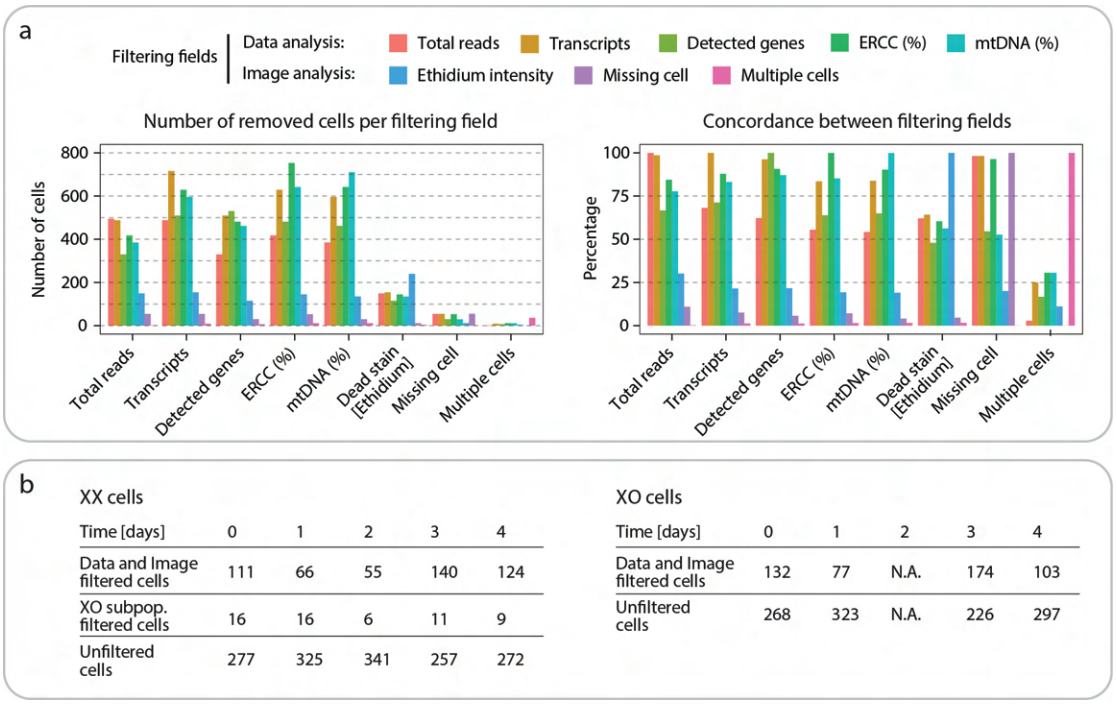
number of Xist-negative cells even after four days of induced cellular differentiation.

3 Data filtering

This section describes the cell and gene filtering procedures, which aim to remove from the analysis the cells and genes which might bias any of the downstream analyses.

3.1 Cell filtering

The aim of cell filtering is to identify and discard empty wells, dead and low-quality cells. This was achieved by combining gene expression and imaging data analyses.



b

XX cells

Time [days]	0	1	2	3	4
Data and Image filtered cells	111	66	55	140	124
XO subpop. filtered cells	16	16	6	11	9
Unfiltered cells	277	325	341	257	272

XO cells

Time [days]	0	1	2	3	4
Data and Image filtered cells	132	77	N.A.	174	103
Unfiltered cells	268	323	N.A.	226	297

FIGURE 5: (a) Concordance between gene expression and imaging filtering measures. Left panel: number of low-quality cells identified by each image or trascriptome-based measure (x-axis) and any other measure (color). Right panel: percentage of low-quality cells identified by each image or trascriptome-based measure (x-axis) which were also identified by a second variable (color). (b) Table summarizing the XX and XO cell filtering steps

The intensity signals of the Calcein (life stain) and Ethidium (dead stain) fluorophores recorded for each cell are represented through a scatter plot (Fig. 4a). The capture sites which isolated none (left panel) or multiple cells (right panel) were identified through visual inspection of each brightfield image. The scatter plots highlight an unexpected correlation between the intensities of the two fluorophores, with the live stain signal spikes up whenever the dead stain is detected. For this reason we used only the dead stain fluorophore to identify cells with a broken cellular membrane, while the signal associated to the life stain was not taken into consideration.

In addition to the dead stain, several transcriptome-based measures were used to identify low quality cells (Fig. 4b). These were identified as the cells resulting in low number of reads, low number of UMI transcripts or low number of expressed genes, as well as cells with high percentage of mitochondrial DNA or ERCC spike-in reads and high Ethidium fluorophore intensity. The violin plots (Fig. 4b) show the distribution of these measures across all cells and time points. For each measure, a 3·MAD threshold was used to identify and remove from the analysis outlying XX and XO cells (Eq. 5). This filtering procedure was performed separately for the two cell lines due to their marked differences in sequencing throughput and gene expression quantification. Indeed XX cells resulted in a higher number of detected genes and total UMI transcripts, and lower percentages of spike-in and mitochondrial DNA reads compared to their XO counterpart. Where

the black and brown horizontal dashed lines represent the thresholds computed for the XX and XO populations, respectively.

Moreover, some XX cells might lose one of the two X chromosomes either before or throughout cellular differentiation. XX cells not expressing *Xist* and with more than 80% of their X-linked transcripts mapping to a single allele (Eq. 6) were assumed to have lost one X chromosome and were removed from the analysis (Fig. 4c).

The concordance between the above filtering criteria can be represented with bar plots (Fig. 5a). The above filtering criteria showed high degree of concordance. Indeed transcriptome-based measured identified the same cells as problematic (Fig. 5a, right). As expected, the vast majority of sites classified as empty were characterized by extremely few reads or UMI transcripts, and high percentages of ERCC transcripts. On the other hand, approximately half of the site with extremely high dead stain intensities and around 75% of the sites with multiple cells were not pointed out as problematic by any other transcriptome-based measure. This reflects the effectiveness of combining image and gene expression measurements to perform an optimal cell filtering procedure. The above filtering criteria were combined, and any cell with at least one problematic measurement was removed from the analysis (Fig. 5b).

3.2 Gene filtering

The aim of gene filtering procedures is to identify and remove from the analysis lowly expressed or poorly annotated genes. This pre-processing step was performed separately for the AS and not-AS gene expression quantifications.

The scatter plots (Fig. 6a) represents the relationship between gene expression level and dropout rate, where these two measures were computed separately with respect to the not-AS and AS UMI count matrices (top and bottom panels, respectively) across all XX cells and time points. As expected, the gene-wise dropout rate decreases as the average gene expression increases. Since the presence of lowly expressed genes might bias downstream analyses such as the data normalization, cell clustering and differential expression analyses, we kept only those genes which were detected (UMI > 0) by at least 20% of cells throughout all sequenced time points. Notably, given the different detection rate and genes which could be investigated by the not-AS and AS quantifications, this filtering procedure was performed separately for the two analyses.

The expression of some genes detected across a large set of cells might however be biased by their poor SNP annotations. Indeed the scatter plots (Fig. 6b) show that a subset of the genes detected by more that 20% of the cells, hence passing the previous filtering step, showed expression bias towards a single allele as more than 90% of their AS counts were derived by a single allele. These genes, 95 autosomal and 4 X-linked genes, were

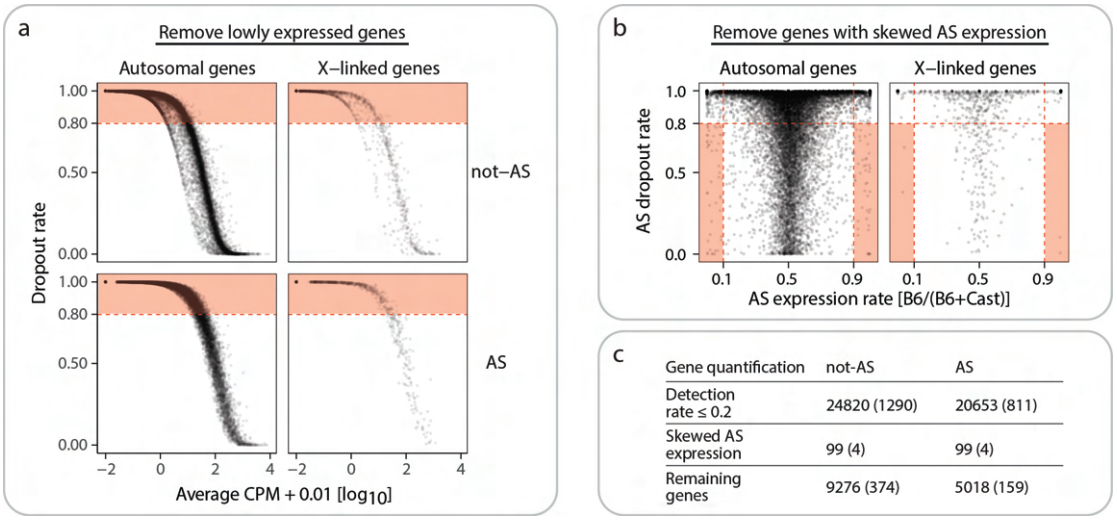


FIGURE 6: (a) Scatter plots showing the relationship between the gene-wise average normalized expression (x-axis) and the fraction of zero UMI counts (y-axis) for the not-AS and AS expression quantifications, separately for X-linked and autosomal genes. The shaded area represents the genes removed from the downstream analyses. (b) Scatter plots showing the gene-wise relationship between the fraction of UMI counts between the two alleles (x-axis) and the fraction of zero UMI counts (y-axis). The shaded area identifies genes with putatively wrong SNP annotation, which have been removed from downstream analyses. (c) Table summarizing the number of genes which have been removed from the analysis with respect to the not-AS and AS quantifications. The number of X-linked genes is represented in brackets.

deemed to have a low quality SNP annotation and removed from both the not-AS and AS downstream analyses.

The table (Fig. 6c) shows that these two gene filtering steps resulted in 9276 and 5018 with high detection rate and no allelic skewing which have been included in downstream not-AS and AS analyses. This filtering procedure highlights that a larger percentage of genes which could be quantified with allele-specificity showed extremely low detection rates with respect to the notAS quantification, respectively around 80% and 70% of the detected genes. This difference is even more pronounced on the X chromosome indeed the genes passing the filtering step at the notAS level were more than double relative to the ones with AS quantifications, respectively 374 and 159.

4 Cell clustering

The complex transcriptomic profiles of single cells measured by the expression of thousands of genes can be visually inspected through dimensionality reduction algorithms which project every cell into a lower dimensional space, where the closer the cells the more similar their transcriptomes. Furthermore the expression of the most variable genes can be modeled to define a data-driven ordering of cells, which is commonly referred to

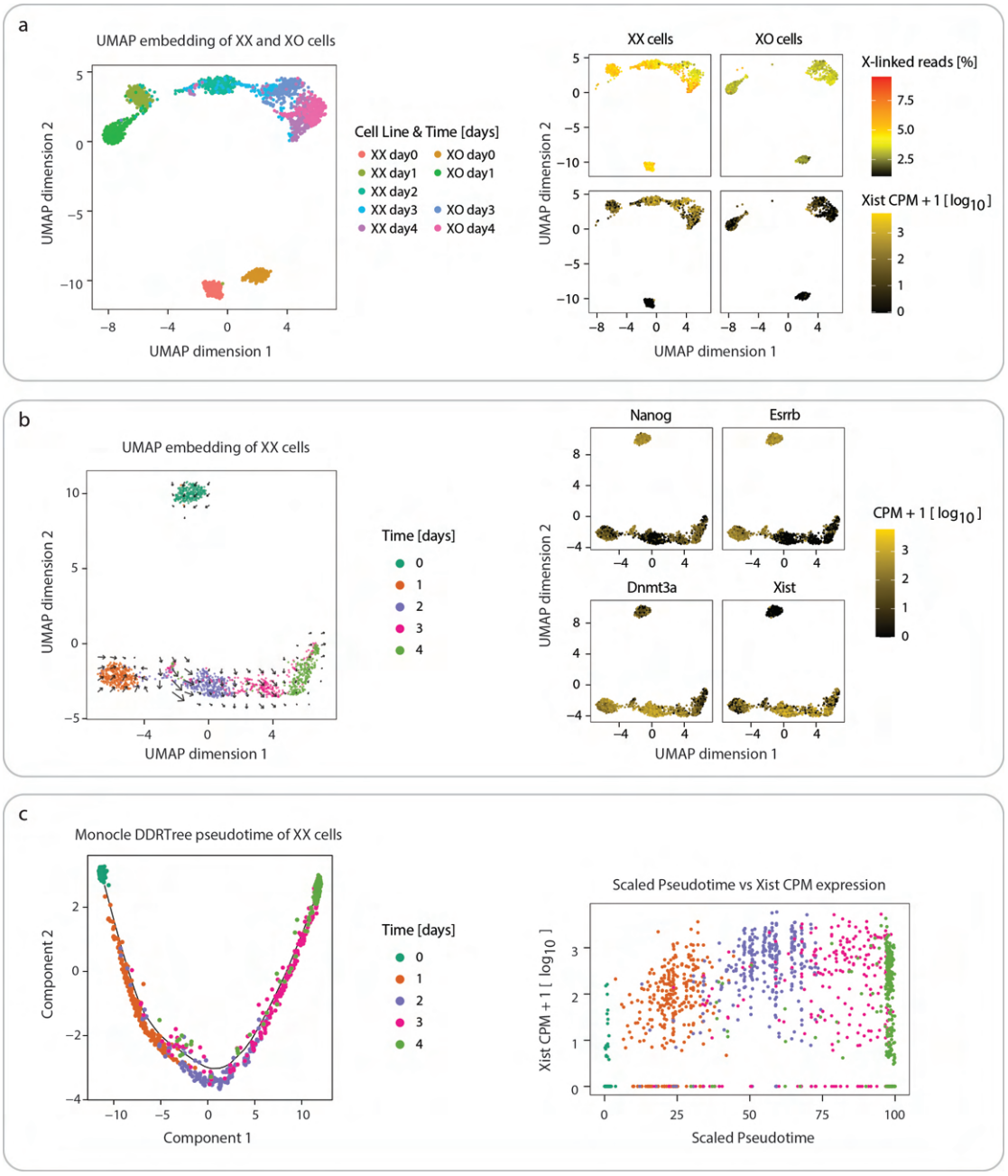


FIGURE 7: (a) UMAP embedding of XX and XO cells colored with respect to their sequencing time point (left). UMAP embedding with cells colored by notAS percentage of X-linked reads and *Xist* normalized expression levels (right). (b) UMAP embedding of XX cells where the arrows represent the predicted transcriptome change estimated through RNA velocity analysis (left). UMAP embedding with cells colored according to markers' normalized gene expression levels (right). (c) Pseudotime estimation based on the 500 most variable genes across XX cells, with individual cells colored by sequencing time point. The black line represents the principal graph describing the pseudotime trajectory of the projected cells as computed by Monocole2 DDRTree method (left). Scatter plot representing *Xist* normalized expression levels and scaled pseudotime for each XX cell colored by sequencing time point.

as pseudotime. In this section we projected and visualized each cell through the Uniform Manifold Approximation and Projection (UMAP) [121] dimensionality reduction

procedure aiming to explore the similarity of their transcriptomes while inspecting the expression of the X chromosome together with some known *Xist* regulators. Furthermore we used Moncole DDRTree algorithm [151] to estimate pseudotime measures and to verify if it can recapitulate the ongoing cellular differentiation and XCI processes.

In a first step (Fig. 7a, left) UMAP dimensionality reduction was applied to compare the transcriptomic profiles of XX and XO cells throughout cellular differentiation. The second dimension of the UMAP embedding clearly separates the undifferentiated cells from the ones undergoing induced cellular differentiation, suggesting that the change of culture condition induced a major change in the transcriptomic profiles. On the other hand the first dimension clusters cells based on their sequencing time point. This UMAP embedding shows that XX and XO cells cluster separately at day 0 and 1 of differentiation, while the two cell lines group together at later time points. This result confirms the previously observed difference between the transcriptomic regulation of female mESCs characterized by one and two X chromosomes in undifferentiated cells and right upon differentiation, where the presence of a second X chromosome considerably affects the regulation of the pluripotency network leading to delayed exit from the ground pluripotency state [171]. On the other hand at later stages of differentiation the two populations show similar transcriptomic profiles, indeed the XX and XO cells tend to cluster together after three days of induced cellular differentiation.

When coloring every single cell (Fig. 7a, right) by the percentage of X-linked UMI counts, we can observe the expected basal difference in X-linked expression between XX and XO undifferentiated mESCs. Notably while a subset of XX cells decreases the mRNA contribution of the X chromosome, others increase its expression throughout differentiation. This difference is clarified when visualizing the the normalized expression of the *Xist* gene. Indeed the former set of cells express *Xist* which induces the initiation of the X-silencing process (XCI), while the latter fail the gene's up-regulation which leads to an increase of X-linked expression (X up-regulation). This mechanism was previously observed in differentiating male mESCs and in male pre/post-implantation embryos in vivo [16, 54, 99, 103, 106, 122, 197]. On the other hand almost all the XO cells, with few exceptions, do not express *Xist* and also undergo X up-regulation over time. Consistently with this observation the XX cells clustering more closely to their XO counterparts are the ones which reduced the global expression of the X chromosome, while the XX cells that have not completed the XCI process have a more distal cell clustering.

The UMAP method was then applied only to XX cells similarly to the previous analysis, and the UMAP embedding was combined with the gene expression predictions of the RNA velocity method [98] (Fig. 7b, left). This method fits gene-wise linear models to the spliced and unspliced not-AS gene expression matrices in order to predict future changes in mature mRNA gene expression. The RNA velocity model was fitted for 3433 autosomal and 174 X-linked genes, for which enough spliced and unspliced transcripts were available to fit a linear model and predict the future number of mRNA transcribed

molecules (Eq. 17, Eq. 18). The vector field gene expression predictions of the RNA velocity method (arrows) suggest that XX cells move along a single differentiation trajectory. This was in accordance with the expression of some marker genes (Fig. 7b, right), such as the down-regulation of naive pluripotency factors (*Nanog* and *Esrrb*) and up-regulation of *Xist* and *Dnmt3a*, a marker of primed pluripotency [130, 131, 171]. Some levels of expression heterogeneity is observed across the XX cells sequenced at the same stage of cellular differentiation. Notably this embedding highlights an intermediate group of XX cells between the day 1 and day 2 clusters, which is composed by cells sequenced after 2-4 days of induced differentiation. These cells are characterized by high expression of the pluripotency markers and low expression of *Dnmt3a* and *Xist*.

The pseudotime analysis of XX cells also revealed that undifferentiated cells clustered distantly from cells undergoing differentiation. The pseudotime of each sequenced XX cell was measured based on the expression of the 500 most variable genes over time through the Monocle2 DDRTree method (Eq. 14). A scaled measure of pseudotime was then obtained dividing the observed pseudotime estimates by the maximum value across all XX cells (Eq. 15). However, the estimated pseudotime measurements were highly correlated with the sequencing time point and failed to capture transcriptomic differences associated to *Xist* expression levels.

This analysis reveals a considerable difference in the transcriptomic profiles of XX and XO mESCs which reduces over time as the result of the ongoing cellular differentiation and X-silencing processes. Our data show that while *Xist* expressing cells decrease the X-linked global expression, the XX cells failing the gene's up-regulation increase X-linked expression over time. Notably both the *Xist* up-regulation and X-silencing processes seem to be very asynchronous throughout cellular differentiation in XX mESCs. Finally, the pseudotime measures estimated for XX cells fails to explain the difference in expression of the *Xist* gene, but rather seems to be driven by the global differences between sequencing time points.

5 Xist and X chromosome expression

This section aims to further explore the association between *Xist* up-regulation and XCI process which was described in the previous section. Specifically we are going to inspect the transcriptional regulation of the X chromosome in *Xist* positive or negative XX and XO mESCs with both not-AS and AS resolutions. Where the latter approach enables us to further characterize the *Xist* expressing XX mESCs, and to explore the two processes separately for the subpopulations silencing the B6 or the Cast alleles.

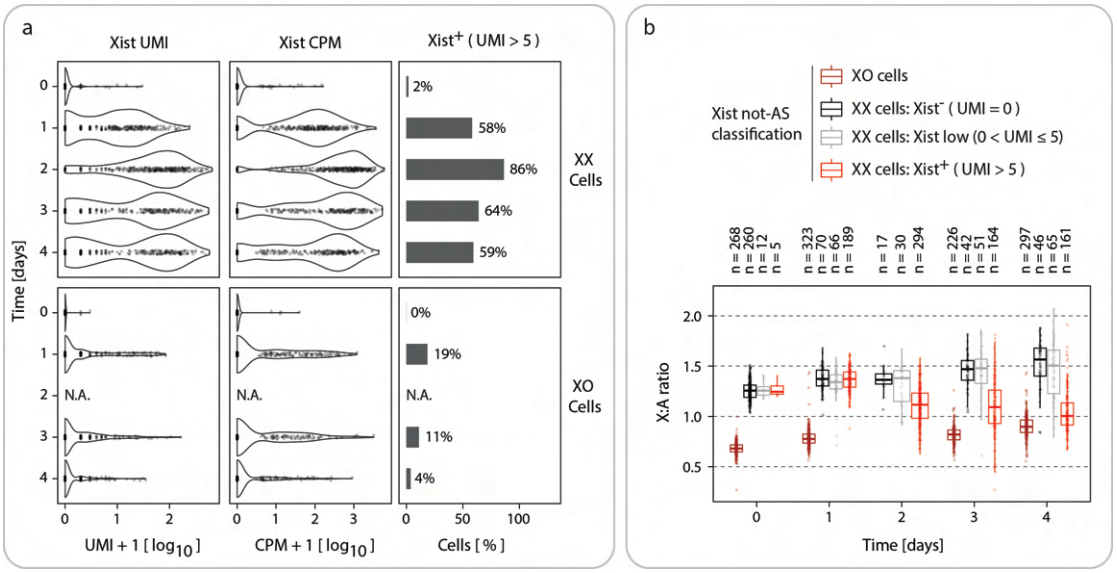


FIGURE 8: (a) Violin plots representing *Xist* not-AS UMI and CPM expression levels, and percentage of $Xist^+$ cells in XX and XO mESCs throughout sequencing time points. Every point in the violin plot represents Xist expression of each single cell. (b) Box plots of bootstrapped X:A ratios grouped by cell line and *Xist* classification, throughout developmental time.

5.1 not-AS gene expression

Following data filtering and normalization, every cell was classified based on its not-AS UMI expression of the *Xist* gene. Every XX cell which detected *Xist* with more than 5 not-AS UMI counts was classified as $Xist^+$, the ones not detecting *Xist* expression as $Xist^-$, and the remaining cells ($0 < y_{Xist,j} \leq 5$) as *Xist*-low.

The barplots (Fig. 8a) show that none of the XO and only 2% of the XX undifferentiated cells detected *Xist* expression with more than 5 not-AS UMI counts. The percentage of $Xist^+$ XX cells sharply increased upon cellular differentiation, with *Xist* reaching its peak in expression after two days of induced cellular differentiation and decreasing at later time points. Notably *Xist* expression levels were very heterogeneous across the XX cells sequenced at the same time point. Indeed *Xist* UMI expression varied from around 10 molecules for XX cells sequenced at day 1, to more than 100 molecules at later stages of cellular differentiation. A much lower fraction of XO cells were classified as $Xist^+$, and their *Xist* expression levels were considerably smaller compared to their XX counterparts. Although an unexpectedly high percentage of XO cells were $Xist^+$ at day 1, the vast majority of these cells transcribed less than 10 *Xist* molecules across all sequencing time points. The detection rate for this experiment was estimated to be around 30%, given that the number of mRNAs present in an ES cell has been estimated to be around 400,000 molecules [28], 120,000 of which were detected per cell (Fig. 4b). The actual mean copy number of the *Xist* RNA in *Xist*-expressing cells would thus increase from 79 at day 1 to 243-314 at the later time points, which is in good agreement with a previous estimate of around 300 *Xist* transcribed molecules per cell [185].

Aiming to investigate the XCI process using data with not-AS resolution, we compared the expression of the X chromosomes relative to the autosomes (Fig. 8b). A robust estimate for this measure, commonly referred to as the X:A ratio (Eq. 10), was obtained using a bootstrapping procedure similarly to a previous study [16]. This approach compares the average UMI expression of autosomal and X-linked genes while accounting for the considerably different number of genes in these two sets (8902 and 374 genes, respectively). This analysis reveals that *Xist*$^+$ XX cells initiate the XCI process, down-regulating the expression of X-linked genes relative to their autosomal counterparts, only after two days of induced cellular differentiation. As expected, *Xist*$^+$ XX cells progressively reduce the expression of the X chromosome at later time points as a result of the ongoing XCI process (Fig. 8b, Appendix 1a). Indeed the median X:A ratio of 1.26 across all XX undifferentiated cells significantly reduced to 1.02 at day 4 in *Xist*$^+$ XX cells (Mann-Whitney U two-sided test: p-value $< 2.2 \cdot 10^{-16}$). Importantly, the observed down-regulation of X-linked genes is not affected by the choice of the UMI threshold used to classify *Xist*$^+$ XX cells (Appendix 1b).

On the other hand, the median X:A ratio of *Xist*$^-$ XX cells significantly increased to 1.57 after 4 days of cellular differentiation compared to all XX undifferentiated cells (Mann-Whitney U two-sided test: p-value $< 1.2 \cdot 10^{-13}$). Similarly, XO cells increased the expression of the X chromosome over time, with the median X:A ratio increasing from 0.68 in undifferentiated cells to 0.89 after 4 days of differentiation (Mann-Whitney U two-sided test: p-value $< 2.2 \cdot 10^{-16}$). This result confirms the observation of the previous section and the significant change in X-linked expression of *Xist*$^-$ XX cells and XO cells throughout differentiation. Specifically X up-regulation is a mechanism which is thought to have evolved to compensate for the loss of the genes located on the Y chromosome as stated by Ohno's hypothesis [140], although this is still controversial [55, 103, 200]. Finally, the XX cells characterized by low levels of *Xist* expression (*Xist* low) are characterized by very heterogeneous X:A ratio values throughout differentiation. This suggests that the few *Xist* molecules transcribed on the *Xist* expressing allele might not be sufficient to initiate the XCI process while the *Xist* negative allele successfully undergoes X up-regulation over time. This combined effect might explain the similar transciptional behavior observed in *Xist* low and *Xist*$^-$ XX cells.

These results confirm the conclusions of several previous studies [20, 42, 201] stating that *Xist* up-regulation precedes and is necessary for the initiation of the X chromosome inactivation process. Indeed while the majority of XX cells express *Xist* already after one day of induced cellular differentiation, the XCI process initiates from day 2 onward. Furthermore, only *Xist* expressing XX cells undergo X inactivation, while *Xist*$^-$ XX cells and XO cells upregulate the expression of X-linked genes over time.

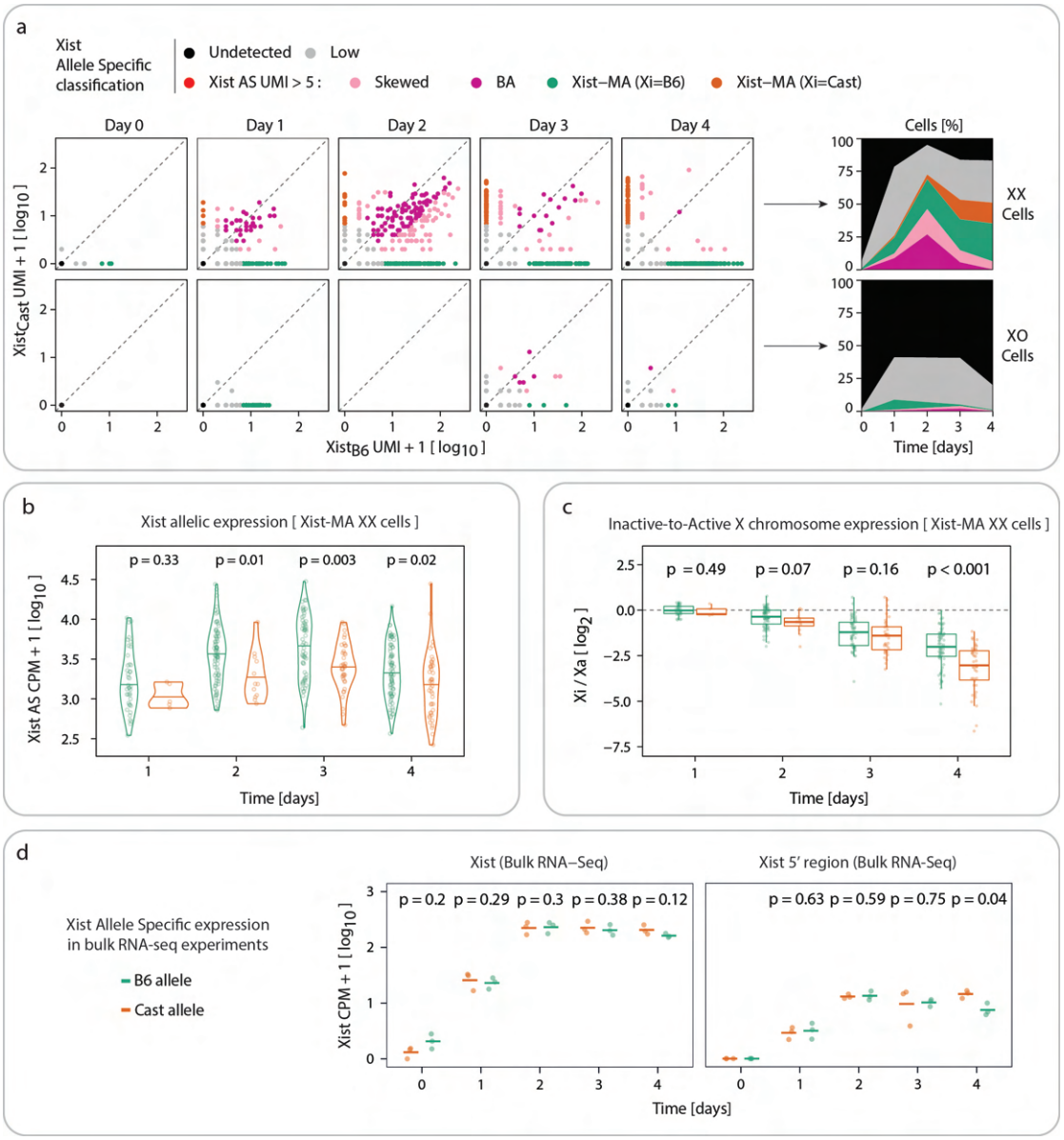


FIGURE 9: (a) *Xist* AS UMI counts associated to the B6 allele (x-axis) and Cast allele (y-axis) for each individual cell colored by *Xist* AS classification (left), and percentage of cells assigned to each *Xist* AS class (right) throughout developmental time and cell lines. (b) Violin plot comparing *Xist* allelic expression levels from the B6 (green) and Cast chromosomes (orange) in *Xist*-MA XX cells. For each time point p-values of a Mann-Whitney U two-sided test are shown. (c) Ratios between inactive and active X chromosome expression (excluding *Xist*) for each *Xist*-MA XX cell. For each time point p-values of a Mann-Whitney U two-sided test are shown. (d) Scatter plot comparing the the normalized expression of *Xist* (left: full gene body, right: 5' peak locus) of bulk RNA-Sequencing data of the TX1072 mESCs collected with 3 replicate samples per time point. For each time point the average value (horizontal bar) and the p-values of a two-sided unpaired Student's T-test are shown.

5.2 Allele-specific (AS) gene expression

Given the stochastic nature of *Xist* up-regulation from one or the other parental X chromosomes, allelic resolution is crucial to investigate XX mESCs undergoing *Xist* up-regulation and random XCI process in their endogenous context. In this subsection we

are going to further characterize *Xist* expressing XX mESCs based on the expression of the gene on the paternal and maternal alleles.

The XX cells which did not express *Xist* at the not-AS level (namely, $y_{Xist,j} = 0$) were classified as *Undetected*, while the ones with up to 5 AS UMI counts (namely, $0 < y^{B6}_{Xist,j} + y^{Cast}_{Xist,j} \leq 5$) as *Low*. Any XX cell with more than 5 *Xist* AS UMI counts was classified as: *Monoallelic* (*Xist*-MA) if all counts were assigned to a single allele, *Skewed* if more than 80% of *Xist* counts mapped to one allele, and *Biallelic* (BA) if each allele accounted for at least 20% of all *Xist* AS UMI counts (Eq. 9). The scatter plot shows the number of *Xist* UMI transcripts assigned to either parental alleles for each XX and XO cell throughout cellular differentiation, where the classification procedure described above is summarized by the area plots (Fig. 9a). As previously observed by the not-AS gene quantification, the vast majority of XO cells transcribed very few or no *Xist* molecules throughout developmental time. After one day of cellular differentiation around 9% of XO cells monoallelically expressed *Xist* from the B6 allele, while almost no *Xist*-MA cell was detected at later time points. Unexpectedly a small fraction of XO cells showed *Xist* expression from the Cast allele, which should not be detected in this cell line lacking the Castaneous X chromosome. This unexpected result was likely caused by sequencing errors at *Xist*'s SNP loci which resulted in the erroneous assignment of these transcripts to the Cast allele.

As showed by the not-AS results, the XX cells start transcribing *Xist* already after one day of induced cellular differentiation and reach the highest fraction of *Xist* positive cells (*Xist* AS UMI > 5) at day 2. Notably, *Xist* AS quantification highlighted transient biallelic *Xist* expression at the early stages of cellular differentiation. Indeed at day 1 and 2 at least half of *Xist* positive XX cells expressed the gene on both alleles, which reached its peak after two days of differentiation with 27% and 19% of all XX cells classified as Biallelic or Skewed, respectively. *Xist* biallelic gene expression was then resolved to a monoallelic fashion at later time points of differentiation. Indeed at day 4, 88% of *Xist* positive cells transcribed *Xist* on a single allele. Importantly, the observed transient biallelic *Xist* expression was not affected by the choice of the UMI threshold used to identify *Xist* positive cells (Appendix 1c). Furthermore, the same trend was highlighted by RNA-FISH at day 2,3 and 4 of cellular differentiation (Appendix 1d). The image classification of 100 cells across three biological replicates per time point indeed showed that on average 49% of XX cells exhibited two *Xist* RNA clouds at day 2, which reduced to 12% and 5% at days 3 and 4 respectively. Importantly, the transient biallelic expression of *Xist* has been recently pointed out by in vivo experiments of random XCI [128, 181].

After four days of differentiation a larger fraction of cells monoallelically expressed *Xist* on the B6 allele than on the Cast allele (29% and 16%, respectively). The observed preferential inactivation of the B6 allele is in agreement with previous findings in B6xCast

F1 hybrid cells, which associated this effect to differing X-controlling elements (Xce) [34, 149].

The B6 allele showed significantly higher *Xist* expression levels compared to the Cast allele starting from day 2 of cellular differentiation (Fig. 9b). This observation is in agreement with the prediction of the stochastic model of XCI onset, where the faster *Xist* up-regulation from one allele results in preferential inactivation of that chromosome [128]. However the higher expression of *Xist* on the B6 allele was not confirmed by bulk RNA-sequencing of TX1072 mESC triplicate samples collected in parallel to the single cell experiment (Fig. 9d). Indeed when comparing the normalized expression of *Xist* on the two alleles, the unpaired Student's T-tests show no evidence of higher expression on the B6 allele. These results are consistent both when quantifying the gene's expression levels based on all its transcripts, and also when the quantification is restricted to the transcripts aligned to its 5'-end peak. These results suggest a potential bias in the single cell protocol.

Finally, when comparing the overall X-linked gene expression levels (excluding *Xist*) on the *Xist*-expressing (Xi) and *Xist*-negative (Xa) alleles for the XX cells classified as monoallelically expressing *Xist* (Eq. 11), the Cast X chromosome appeared to be silenced significantly faster than the B6 allele (Fig. 9c).

Overall, the allele-specific *Xist* expression quantification reveals a transient biallelic *Xist* expression which is then resolved to a monoallelic state, with the preferential inactivation of the B6 chromosome, and the faster silencing of the Cast allele.

5.3 Silencing kinetics

Aiming to explore the allele-specific kinetics which characterize the *Xist* up-regulation and XCI processes, we computed for each XX mESC the fraction of B6 transcripts for the *Xist* gene (*Xist* ratio) and for the X chromosome (X chromosome ratio). Where the latter ratio was computed upon excluding the *Xist* transcripts.

The violin plots (Fig. 10a) show the distribution of the *Xist* and X chromosome ratios across all XX cells, separately for each *Xist* AS class and sequencing time point. Namely, values of 0.5 reflect equal expression levels from the two alleles, while ratios around the values of 1 and 0 highlight monoallelic expression. The X chromosome ratio portrays the XCI trend which was already observed in the previous section. Indeed, $Xist^-$ cells show nearly equal expression of the two X chromosomes throughout developmental time. On the other hand, the distribution of the X chromosome ratio in $Xist^+$ cells broadens over time starting at day 2, defining two separate populations at day 4. These two represent the subset of $Xist^+$ cells which have almost completely silenced the Cast and the B6 X chromosomes, respectively identified by ratios approaching values of 1 and 0. A fraction of *Xist* low cells shows much lower X-silencing levels, while the majority of them

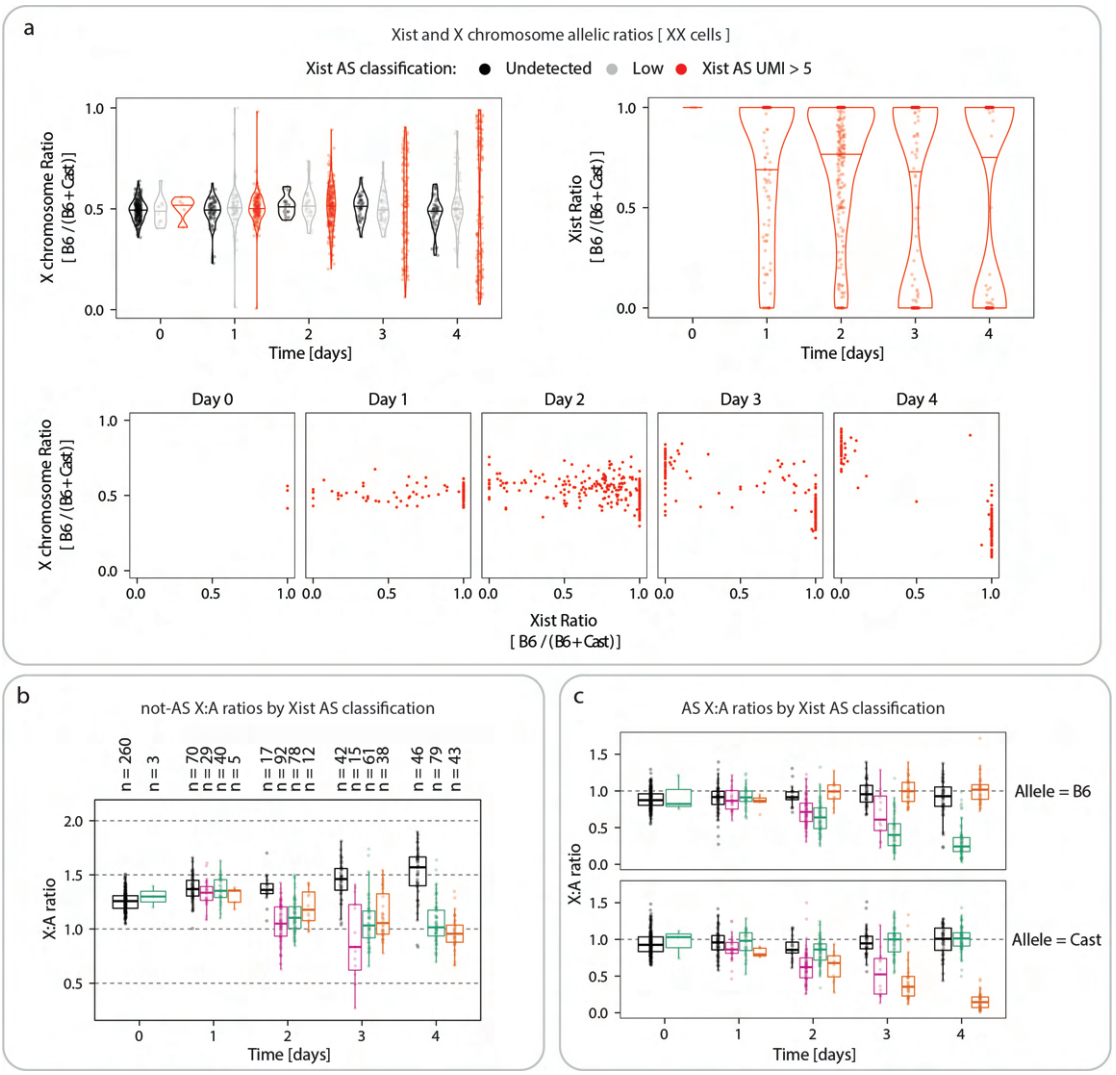


FIGURE 10: (a) Violin plots showing the distribution of the B6 allelic expression ratio for the: entire X chromosome excluding *Xist* (left) and *Xist* gene (right), coloring each XX cell by *Xist* AS classification. Scatter plot showing the observed ratios for each *Xist*$^+$ XX cells throughout developmental time (bottom). (b) Box plot showing the not-AS bootstrapped X:A expression ratios, grouping XX cells by *Xist* AS classification. (c) Box plot showing the AS bootstrapped X:A expression ratios, grouping XX cells by *Xist* AS classification

maintain both X chromosomes equally active throughout time. As previously observed (Fig. 9a), the distribution of the *Xist* ratio highlights that some *Xist*$^+$ cells undergo transient biallelic gene expression up to day 2 of differentiation, while a subset of XX cells monoallelically express *Xist* already after a single day of differentiation. The *Xist* and X chromosome allelic ratios observed for each *Xist*$^+$ cell can be represented through a scatter plot. This plot reveals that *Xist* monoallelic expression precedes the silencing of the *Xist*-expressing X chromosome, which reflects *Xist*-induced chromosome-wide gene silencing. Moreover, transient biallelic *Xist* expression is resolved to a monoallelic state around day 2 of cellular differentiation, and *Xist* biallelic cells show similar expression levels of the two X chromosomes.

Notably the extent of biallelic X-silencing can not be inspected using the above allele-specific ratios, however it can be investigated through the not-AS and allele-specific X:A ratios (Eq. 10). The not-AS X:A ratios (Fig. 10b) show that X-linked gene silencing of *Xist* biallelic cells was even more pronounced than the one observed for *Xist* monoallelic cells (p=0.45/0.01/0.08 at day 1/2/3, Mann-Whitney U two-sided test). Moreover, the AS X:A ratios show that *Xist* biallelic XX cells reduce X-linked gene expression on both alleles (Fig. 10c).

In summary, the allele-specific analyses quantitatively assess the relationship between *Xist* expression and global gene silencing. These results suggest that *Xist* expression is necessary for X-linked gene silencing, which started around two days after *Xist* was initially upregulated. Upon *Xist* monoallelic expression, the differentiating XX cells silence the *Xist*-expressing X chromosome. On the other hand *Xist* biallelic expression induces the silencing of both X chromosomes with similar kinetics. *Xist* biallelic XX cells then resolve *Xist* expression to a monoallelic state, and undergo *Xist*-induced X-silencing in cis. Notably the AS X:A ratios of Xist Undetected cells do not show any significant up-regulation over time, previously observed at not-AS resolution. The discrepancy between these two might arise from the reduced number of X-linked genes which could be quantified with allelic resolution.

5.4 RNA-velocity predicted X-linked expression

In order to apply the concept of RNA velocity, for every XX cell and annotated gene the number of spliced and unspliced transcripts was quantified as the amount of uniquely aligned reads with a unique UMI barcode sequence overlapping the gene's exonic and intronic regions, respectively.

The scatter plots (Fig. 11a) represent the total number of X-linked spliced and unspliced molecules which could be assigned to either parental alleles for each cell throughout differentiation (159 X-linked genes). Where every XX cell, which is represented by a point in the plot, is colored according to its *Xist* AS classification. Notably, the expected XCI trend was observed in both spliced and unspliced transcripts' quantification. Indeed *Xist* monoallelic cells decreased the number of X-linked molecules transcribed by the *Xist*-expressing allele over time, while *Xist* negative and biallelic cells showed similar expression levels of the two X chromosomes. Observing the expected XCI trend in both spliced and unspliced quantifications motivated the application of the RNA velocity method to predict the future not-AS spliced X-linked expression of each XX cell (Eq. 17).

The number of predicted not-AS X-linked spliced molecules estimated by the RNA velocity method were then visualized within a lower dimensional space which separates the cells silencing the B6 and the Cast alleles (Fig. 11b). Such cellular embedding was

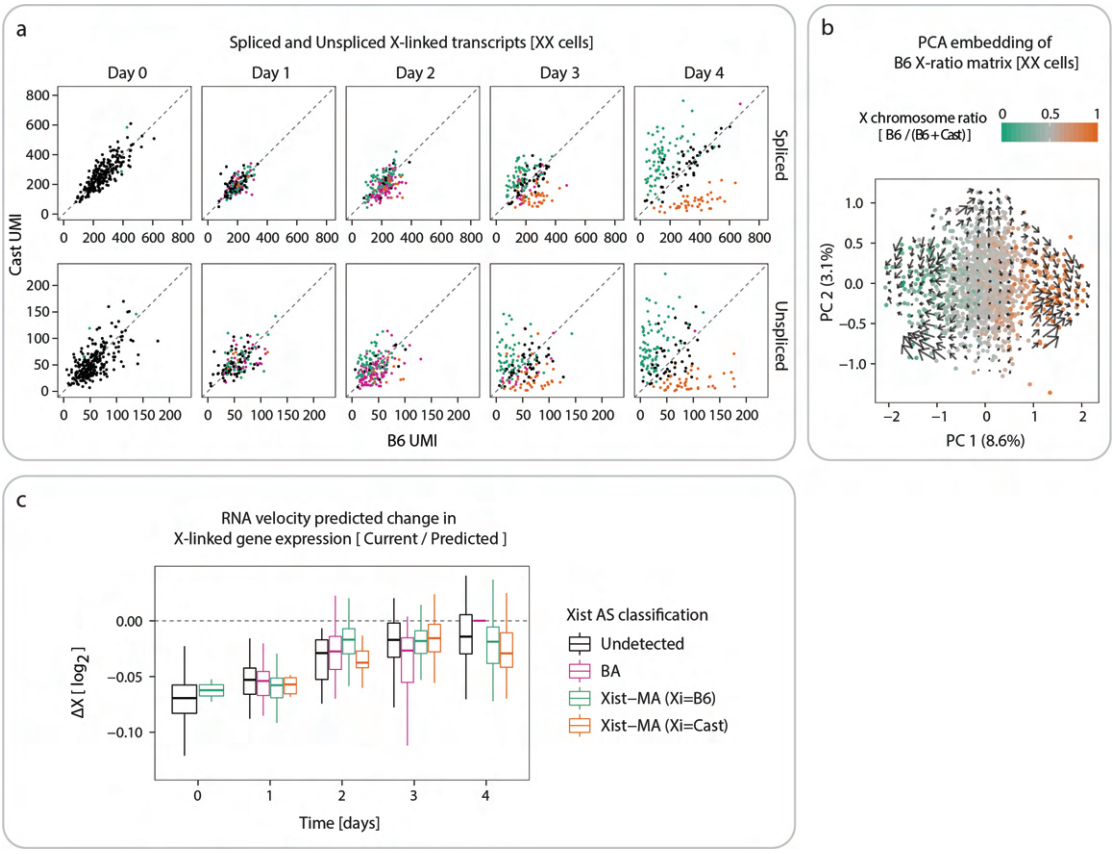


FIGURE 11: (a) Scatter plots showing spliced (top) and unspliced (bottom) reads mapping to the X chromosome on the B6 and Cast alleles. XX cells are colored by *Xist* AS classification. (b) XX cells embedding on the first two principal components of the allelic expression ratio matrix of X-linked genes. Arrows indicate the predicted change in spliced transcripts as estimated by RNA velocity method. (c) Box plot of the log-ratio between RNA-velocity estimates of current and predicted normalized X-chromosomal gene expression (ΔX), grouping XX cells by *Xist* AS classification

computed applying the Principal Component Analysis (PCA) method to the matrix of X-linked B6 ratios. Every XX cell sequenced throughout developmental time was then projected onto the first two principal components and colored by its X chromosome ratio value. Where the position of each cell in this embedding is defined by its X-linked B6 ratio, while its future state (arrow) was predicted based on the vector of not-AS velocities estimated for the same X-linked genes (Eq. 17, Eq. 20, Appendix 1e). The first principal component, which accounts for 8.6% of all the variability in the X-linked B6 ratio matrix, clearly separates the XX cells silencing the B6 (left) or Cast (right) X chromosomes from the cells which have not yet initiated the XCI process (center). Notably, the latter show almost no predicted change in their X-linked B6 ratios while the former are predicted to proceed towards the completion of the XCI process.

The RNA-velocity predicted states (arrows) can be explained by the fact that the gene-wise X-linked ratios between spliced and unspliced transcripts are in steady state for the cells that equally express both X chromosomes (PC1 = 0), since these cells have not yet initiated to down-regulate their X-linked unspliced transcripts. Therefore both

their predicted X-linked gene expression and X-linked B6 ratio levels will be similar to the observed ones. On the other hand, the velocities estimated for the cells which have already initiated the XCI process account for the observed under-representation of X-linked unspliced transcripts, which results in predicting a future down-regulation of X-linked transcripts. Specifically, the cells characterized by more extreme X chromosome ratios are likely down-regulating the unspliced transcripts of multiple genes, which results in fewer predicted X-linked transcripts and more extreme predicted ratios compared to the cells with observed X chromosome ratios closer to 0.5. This explains why the XX cells showing higher extent of X-silencing are also predicted to proceed faster towards the state corresponding to the completion of the XCI process.

The RNA velocity gene-wise linear models were then used to estimate the change in X-linked gene expression for each XX cell as the ratio between the measured and future predicted X chromosome gene expression (Eq. 21), which is denoted as ΔX (Fig. 11c). Where this ratio is expected to increase upon initiation of the XCI process, and decrease upon X up-regulation. Accordingly, both *Xist* monoallelic and biallelic cells result in slightly higher ΔX values compared to *Xist* negative cells after two days of induced cellular differentiation, when XCI is first initiated.

These results show that the expected AS decrease in X-linked reads could also be observed in both spliced and unspliced gene expression quantifications, which motivates the use of the RNA velocity method to predict the future expression of X-linked reads. Accordingly, the predicted ΔX values show higher values in *Xist* Monoallelic and Biallelic cells compared to *Xist* Undetected cells at day 2, which corresponds to the time point when X-silencing is first observed. This suggests the potential use of this measure to identify putative regulators of *Xist* and XCI.

6 Identification of putative Xist and XCI regulators

The aim of this section is to identify a set of genes which regulates *Xist* expression or plays a role in the initiation of the XCI process. The expression levels of such genes are expected to be associated to variations in *Xist* or overall X-linked expression levels. In these analyses, *Xist* expression was quantified by *Xist* not-AS CPM values (Fig. 8a) while the changes in X chromosome expression levels were estimated by the RNA-velocity predicted change in X-linked gene expression, which is referred to as ΔX (Fig. 11c).

Based on the observed *Xist* CPM or ΔX values across all XX cells throughout differentiation, a set of putative regulators was identified using two different approaches relying on differential expression and correlation analyses. On one hand, putative regulators were identified through MAST differential expression (DE) analysis method comparing the gene-wise expression levels of XX cells characterized by high and low *Xist* CPM or

ΔX values. On the other hand, the Spearman's correlation coefficient between every gene's CPM expression values and *Xist* CPM or ΔX values was computed, and a putative set of regulators was identified as the subset of genes whose correlation coefficients were significantly different from zero.

Although these analyses were performed separately for each sequencing time point, the identification of such regulators was restricted to days 1 and 2 of differentiation, since these represents the earliest stages of *Xist* up-regulation and XCI, respectively (Fig. 8).

6.1 Identify regulators based on Xist expression

In order to perform MAST DE analysis, the XX cells were first clustered into 7 groups applying the K-means clustering algorithm (K = 7) to *Xist* not-AS CPM expression levels observed across all time points. The cells assigned to the top three K-means classes were classified as *Xist*-high cells, the ones assigned to the bottom three groups as *Xist*-low, while the ones assigned to the intermediate *Xist* expression class (K-4) were excluded from the DE analyses (Fig. 12a). The optimal value for K was chosen to minimize the within-cluster sum of squares value, while ensuring that at least 50 cells were assigned to the *Xist*-high and *Xist*-low groups at each time point throughout cellular differentiation.

MAST DE analysis method was then performed separately for each time point with at least 10 XX cells assigned to the *Xist*-high and *Xist*-low groups, testing for each gene the significance of the difference in expression between these two groups (H_0 : $\log_2 FC = 0$). Since none of the undifferentiated XX cells was assigned to the *Xist*-high class, DE analysis was only performed for cells undergoing differentiation. The number of autosomal and X-linked genes deemed to be differentially expressed between the *Xist*-high and *Xist*-low cells at each time point of differentiation (namely, with a Benjamini-Hochberg adjusted p-value: FDR ≤ 0.05) are summarized by the barplots (Fig. 12b). The red and blue bars represent the number of differentially expressed genes (DEGs) showing significantly higher expression in the *Xist*-high and *Xist*-low groups, respectively. After one day of induced cellular differentiation only two autosomal genes were deemed as differentially expressed between the two groups of cells, while the number of DEGs sharply increased over time presumably due to global X-dosage effects such as modulation of the differentiation-promoting MAPK signalling pathway and DNA hypomethylation [171, 180, 205]. As expected, the number of X-linked genes which significantly down-regulated their expression in the *Xist*-high group increased throughout differentiation as a result of the ongoing XCI process of *Xist* expressing cells. Since the up-regulation of *Xist* leads to the down-regulation of X-linked genes, putative regulators of *Xist* and the XCI process were identified restricting the analysis to autosomal DEGs and up-regulated X-linked DEGs identified at the earliest stages of cellular differentiation. On the other hand, significantly down-regulated X-linked genes

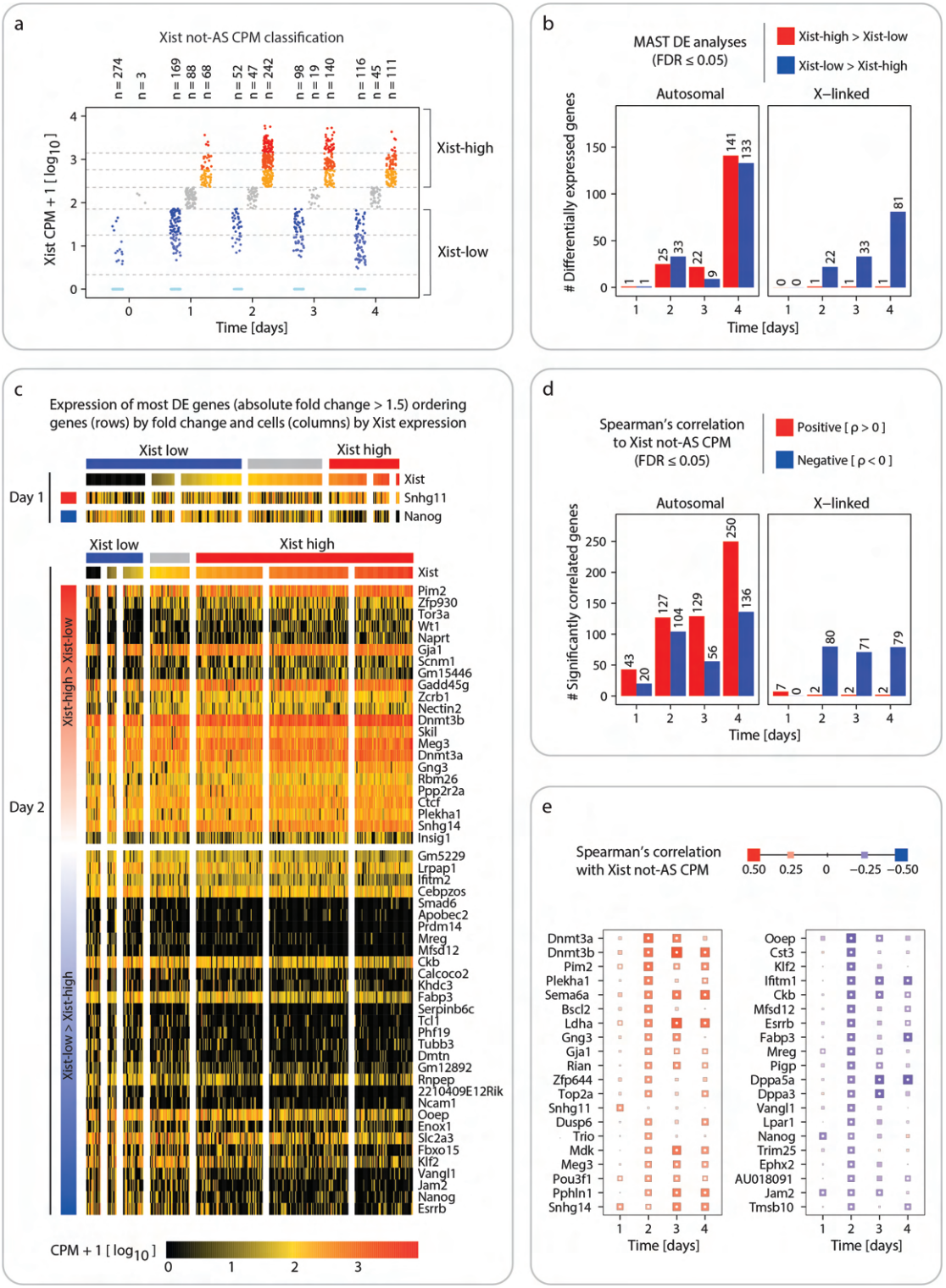


FIGURE 12: (a) *Xist* not-AS CPM K-means (K=7) clustering at each sequencing time point. (b) Number of significantly differentially expressed genes (DEG: FDR ≤ 0.05) between *Xist*-high and *Xist*-low XX cells throughout cellular differentiation. (c) Expression of DEGs (rows) with absolute fold change above 1.5 at day 1 (top) and day 2 (bottom) in single cells (columns). X-linked genes with *Xist*-low>*Xist*-high are excluded. (d) Number of genes with CPM expression significantly correlated (FDR $\leq$ 0.05) to *Xist* CPM expression. (e) Top 20 genes showing highest significantly positive (left) or negative (right) correlations to *Xist* CPM expression at day 1 or 2 (excluding pseudogenes), ordered by decreasing absolute correlation coefficient. Size and color indicate the correlation coefficient as indicated. White dots represent significant correlations (FDR ≤ 0.05).

and pseudogenes were not deemed as putative regulators since their association to *Xist* might be confounded with the overall XCI process.

The heatmaps show the cell-wise CPM expression of every autosomal or up-regulated X-linked DEG with an absolute fold change greater than 1.5, which were identified for XX cells after 1 and 2 days of differentiation (Fig. 12c). Where cells (columns) are grouped by *Xist* K-means classification, while the genes (rows) are ordered by decreasing fold change. After one day of induced cellular differentiation, the known *Xist* regulator *Nanog* ($\log_2$FC = -2.15) was significantly down-regulated in the *Xist*-high group, while the lncRNA *Snhg11* ($\log_2$FC = 1.81) was deemed as a putative *Xist* activator. After 2 days of induced cellular differentiation, we identified a number of previously reported pluripotency factors implicated in *Xist* repression (such as *Nanog*: $\log_2$FC = -2.23; *Klf2*: $\log_2$FC = -1.93; *Prdm14*: $\log_2$FC = -1.22) together with other pluripotency-associated factors (such as *Esrrb*: $\log_2$FC = -2.86; *Fbxo15*: $\log_2$FC = -1.91; *Tcl1*: $\log_2$FC = -1.48) [120, 130, 143] which were significantly down-regulated in the *Xist*-high group, hence were deemed to be putative *Xist* repressors. On the other hand, a number of genes involved in transcriptional regulation and signalling were up-regulated in the *Xist*-high group, hence were deemed to be putative *Xist* activators. These included some transcription factors (such as *Wt1*: $\log_2$FC = 1.74; *Gm15446*: $\log_2$FC = 1.57), DNA methyltransferases and splicing factors (such as *Zcrb1*: $\log_2$FC = 1.49; *Dnmt3b*: $\log_2$FC = 1.15; *Dnmt3a*: $\log_2$FC = 0.98) and genes modulating the MAPK and TGF-β/Smad signalling pathways (such as *Gadd45g*: $\log_2$FC = 1.51; *Skil*: $\log_2$FC = 1.11), which were reported in several previous studies [61, 78, 90, 170, 196]. Interestingly, the X-linked gene *Pim2* was found to be significantly up-regulated in the *Xist*-high group starting at day 2 of differentiation ($\log_2$FC = 2.41/1.32/2.06 at day 2/3/4 respectively). This gene encodes an oncogenic kinase which cooperates with the *Myc* transcription factor[123] which was never reported to play a role in the XCI process, making it an interesting candidate for future studies. For each sequencing time point the gene-wise average CPM expression level and $\log_2$FCs between the two groups are shown through scatter plots (Appendix 2a) with significantly up- and down-regulated genes (FDR $\leq$ 0.05) colored in red and blue, respectively.

Similarly to a previous work [95], putative *Xist* regulators were identified for each time point throughout cellular differentiation testing if the Spearman's correlation coefficient (namely, ρ) between *Xist* CPM expression and the normalized expression of every other detected gene was significantly different from zero, $H_0 : \rho = 0$. In concordance with the previous DE analyses, the number of significantly correlated genes (namely, with a Benjamini-Hochberg adjusted p-value FDR $\leq$ 0.05) increased over time and most of the significant X-linked genes were negatively correlated to *Xist* expression as a result of the ongoing *Xist*-mediated XCI process (Fig. 12d). The identification of *Xist* regulators was again limited to XX cells sequenced after one and two days of induced cellular differentiation, excluding negatively correlated X-linked genes and pseudogenes. The results of

these analyses are summarized by the dot plot (Fig. 12e), which highlights the correlation coefficients over time for the 20 genes resulting in the highest significantly positive or negative correlation coefficients to *Xist* CPM expression at day 1 or day 2. The DE and correlation analyses were highly concordant. Indeed, at day 1 of differentiation the correlation analysis identified *Snhg11* (ρ= 0.26) as a putative *Xist* activator and *Nanog* (ρ= -0.27) as its repressor. In addition to these, the correlation analysis also highlighted the cell stem marker gene *Jam2* (ρ= -0.25) as a negative *Xist* regulator, and the lncRNA *Snhg14* (ρ= 0.245) as a putative *Xist* activator at day 1 [169]. Also the results for day 2 of cellular differentiation were highly concordant to the DE analysis results for the same time point. Indeed also this analysis identified *Dnmt3a* (ρ= 0.36), *Dnmt3b* (ρ= 0.34) and *Pim2* (ρ= 0.34) as *Xist* putative activators, in addition to the transcription factor *Zfp644* (ρ= 0.26) and signalling genes *Gng3* (ρ= 0.28), *Dusp6* (ρ= 0.26) and *Trio* (ρ= 0.25). Furthermore, *Esrrb* (ρ= -0.30), *Klf2* (ρ= -0.32) and *Nanog* (ρ= -0.25) were again deemed as *Xist* putative repressors, in addition to other pluripotency factors such as *Dppa5a* (ρ= -0.28), *Dppa3* (ρ= -0.28), the Wnt signalling pathway regulator *Vangl1* (ρ= -0.28) and the E3 ubiquitin ligase *Trim25* (ρ= -0.26).

6.2 Identify regulators based on predicted variation in X chromosome expression

A similar analysis was performed based on the ratio between the current and the RNA-velocity predicted overall X-linked expression, namely ΔX. Since this ratio varies greatly across developmental time, the subset of XX cells characterized by high and low ΔX values were identified applying the K-means clustering algorithm (K = 3) separately for each sequencing time point (Fig. 13a). The cells assigned to the top K-means class were classified as ΔX-high cells (red), the ones assigned to the bottom group as ΔX-low (blue), while the remaining cells (grey) were excluded from the following DE analyses.

MAST DE analysis method was then performed separately for each time point throughout cellular differentiation, comparing the gene-wise expression difference between the ΔX-high and ΔX-low XX cells ($H_0 : \log_2 FC = 0$). Similarly to the previous section, the number of autosomal and X-linked genes deemed to be differentially expressed between the two groups at each time point throughout differentiation (namely, with a Benjamini-Hochberg adjusted p-value: FDR ≤ 0.05) are summarized by the barplots (Fig. 13b). After one day of induced cellular differentiation only two X-linked genes were deemed as differentially expressed between the two groups of cells, while the number of DEGs sharply increased at day 2 following the initiation of the XCI process. Similarly to the previous analysis, putative regulators were identified restricting the analysis to the earliest time points of differentiation while excluding pseudogenes and down-regulated X-linked genes, whose difference in expression level might be the result and not the cause of the XCI process.

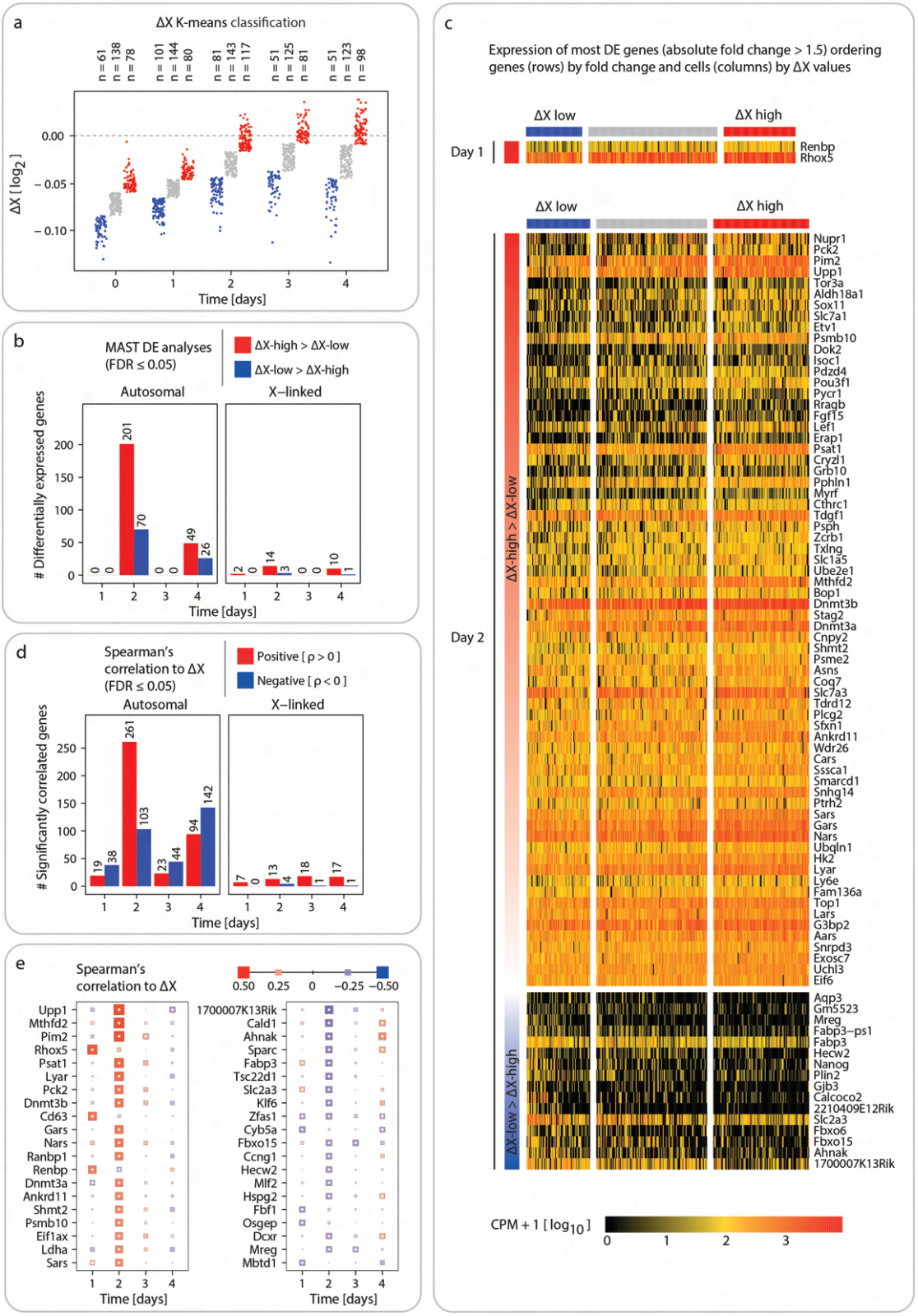


FIGURE 13: (a) RNA-velocity predicted change in X-linked gene expression (ΔX) K-means (K=3) classification at each time point. (b) Number of significantly differentially expressed genes (DEG: FDR ≤ 0.05) between ΔX-high and ΔX-low XX cells throughout cellular differentiation. (c) Expression of genes (rows) with FDR ≤ 0.01 and absolute fold change above 1.5 at day 1 (top) and day 2 (bottom) in single cells (columns). X-linked genes with ΔX-low>ΔX-high are excluded. (d) Number of genes with CPM expression significantly correlated (FDR ≤ 0.05) to ΔX value. (e) Top 20 genes showing highest significantly positive (left) or negative (right) correlations to ΔX value at day 1 or 2 ordered by decreasing absolute correlation coefficient, excluding pseudogenes. Size and color indicate the correlation coefficient as indicated. White dots represent significant correlations.

The heatmaps (Fig. 13c) show the cell-wise CPM expression of every autosomal or up-regulated X-linked gene with BH-adjusted pvalue (FDR) smaller than 0.01 and an absolute fold change greater than 1.5. The cells (columns) are grouped by ΔX K-means classification, while the genes (rows) are ordered by decreasing fold change. After one day of induced cellular differentiation the X-linked transcription factor *Rhox5* ($\log_2$FC = 1.53) whose overexpression was previously shown to prevent the exit from mESCs' pluripotent state [16, 41, 63], and the protein coding gene *Renbp* ($\log_2$FC = 1.65) were significantly up-regulated by the ΔX-high cells. Similarly to the previous analysis, at day 2 of cellular differentiation the ΔX-high cells significantly down-regulated a number of known plutipotency-associated genes (such as *Fbxo15*: $\log_2$FC = -1.94; *Esrrb*: $\log_2$FC = -1.66; *Nanog*: $\log_2$FC = -1.31; *Prdm14*: $\log_2$FC = -0.91), and up-regulated several genes involved in transcriptional regulation and signalling (such as *Sox11*: $\log_2$FC = 2.05; *Lef1*: $\log_2$FC = 1.78; *Sox4*: $\log_2$FC = 1.52; *Zcrb1*: $\log_2$FC = 1.43; *Dnmt3b*: $\log_2$FC = 1.26; *Dnmt3a*: $\log_2$FC = 1.19). Notably, the X-linked gene *Pim2* ($\log_2$FC = 2.46) was again deemed as significantly up-regulated in the ΔX-high cells. For each sequencing time point, the gene-wise average CPM expression level and $\log_2$FCs between the two groups are shown through scatter plots (Appendix 2b) with significantly up- and down-regulated genes (FDR $\leq$ 0.05) colored in red and blue, respectively.

A set of putative *Xist* regulators was then identified testing if the Spearman's correlation coefficient (namely, ρ) between the ΔX values and the normalized expression of every detected gene was significantly different from zero, $H_0 : \rho = 0$ (Fig. 13d). The results of these analyses are summarized by the dot plot (Fig. 13e), which highlights the correlation coefficients over time for the 20 genes with the highest significantly positive or negative correlations to ΔX values at day 1 and day 2. The results of the correlation analyses were in agreement with the previous DE analyses. At day 1 of differentiation, also the correlation analysis highlighted the X-linked *Rhox5* (ρ= 0.41) and *Renbp* (ρ= 0.33) as *Xist* putative regulators, in addition to other genes. At day 2 of differentiation, *Pim2* (ρ= 0.42) was again identified as one of the top *Xist* activators together with DNA methyltransferases *Dnmt3b* (ρ= 0.35) and *Dnmt3a* (ρ= 0.32), and other genes involved in transcriptional regulation (*Lef1*: ρ= 0.26; *Sox4*: ρ= 0.21; *Sox11*: ρ= 0.21; *Zcrb1*: ρ= 0.19). Furthermore, also the pluripotency factors *Fbxo15* (ρ= -0.24), *Prdm14* (ρ= -0.2), *Nanog* (ρ= -0.2) and *Esrrb* (ρ= -0.18) were deemed as *Xist* putative repressors.

6.3 Putative Xist and XCI regulators

The results of the above 8 analyses (*Xist*/ΔX, day1/day2, DE/Correlation) were integrated, focussing on the autosomal and up-regulated X-linked genes deemed as significant (FDR $\leq$ 0.05) activators or repressors in at least 3 analyses (Fig. 14a and 14b, respectively).

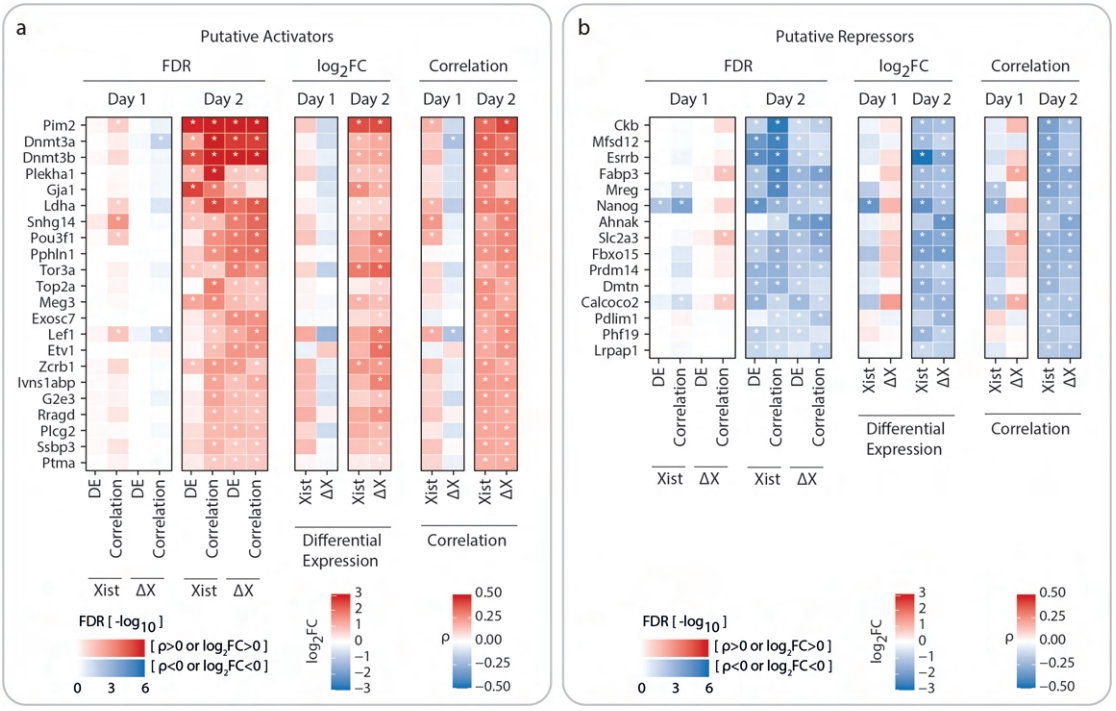


FIGURE 14: Putative activators (a) and repressors (b) of early XCI identified through correlation and differential expression (DE) analyses based on *Xist* expression and early gene silencing (ΔX) at day 1 and 2 of differentiation. For each gene deemed as significant (FDR ≤ 0.05) in at least 3 of the 8 analyses (day1/2, *Xist*/ΔX, DE/Correlation): the Benjamini-Hochberg corrected p-values (FDR, left), the $\log_2$-transformed fold changes ($\log_2$FC, middle) and Spearman's correlation coefficients (ρ, right) are shown. Pseudogenes and X-linked genes with negative correlation coefficients ρ or $\log_2$FCs are not shown. Asterisks indicate statistically significant tests.

After one day of induced cellular differentiation, the analyses identified a handful of genes as putative regulators of XCI initiation. Among these, the known *Xist* repressor *Nanog* [130] was significantly differentially expressed between the *Xist*-high and *Xist*-low groups, and its expression was negatively correlated to *Xist* expression. On the other hand, *Nanog* expression levels did not significantly differ between the two ΔX groups, nor significantly correlated to the ΔX values.

At day 2 of differentiation, the analyses identified a much larger number of genes as putative regulators of XCI initiation. Several pluripotency factors such as *Nanog*, *Prdm14*, *Esrrb* and *Fbxo15* [10, 130, 145, 188, 202] were deemed as putative *Xist* repressors by both *Xist* and ΔX analyses. On the other hand, a number of genes involved in transcriptional regulation or signalling were deemed as putative *Xist* activators. Among the genes promoting the initiation of XCI: the transcription factor *Pou3f1*, which is associated with early ESC differentiation, the DNA methyltransferases *Dnmt3a* and *Dnmt3b*, and the splicing factor *Zcrb* [10, 78, 194]. Notably the X-linked kinase *Pim2*, which cooperates with the *Myc* transcription factor [86, 123], was deemed as a putative activator by several analyses and exhibited the highest significance and fold change between

the two *Xist* and ΔX groups after two days of differentiation, making it an interesting candidate for further studies.

Interestingly, this analysis only identified a subset of the previously proposed regulators of *Xist* and XCI process. A comprehensive analysis of genes that have been implicated in *Xist* regulation before showed that *Nanog*, *Klf2*, *Klf4* and *Prdm14* were correlated with *Xist* and early XCI, while other pluripotency factors such as *Oct4* (*Pou5f1*) and *Sox2* were not (Appendix 2c, 2d). On the other hand, the above analyses unexpectedly deemed the known *Xist* repressor *Ctcf* as a putative activator [57], and the *Xist* activator *Rnf12* (*Rlim*) as a putative repressor [76, 92].

Taken together these results show that the down-regulation of naive pluripotency factors, in particular *Nanog*, combined with the up-regulation of early differentiation factors, such as *Pou3f1* and *Dnmt3a* and *Dnmt3b*, seem to play a role in the initiation of *Xist* expression and the initiation of the XCI process. It has to be noticed that these analyses identify genes whose expression is significantly associated to a change in *Xist* or X-chromosome expression levels. These analyses however do not clarify if these genes directly regulate the changes in expression levels, or rather if they are involved in a more complex regulatory network leading to differential expression. One way to investigate their role in the XCI process would be to design an experiment where their expression gets significantly reduced or inhibited, for example through shRNA guides or CRISPR-Cas9 technologies respectively, and analyse the transcriptomic profiles of mESCs at the earliest stages of differentiation. Observing perturbed *Xist* expression levels or failure/delay in the XCI process with such experimental settings would provide more insights in the role of these genes and of their co-regulators.

7 Allele-specific silencing dynamics

The analysis of *Xist* allelic expression revealed that while this gene was expressed at higher levels and by a larger fraction of XX mESCs on the B6 allele, the silencing of X-linked genes proceeded faster on the Cast X chromosome (Fig. 9b,c). The detection bias of the *Xist* gene might however be a technical artifact caused by a mapping bias towards the reference genome (B6), indeed no difference was detected in parallel bulk RNA-sequencing experiments (Fig. 9d).

In order to verify if the allelic gene counts were consistently higher on the B6 allele throughout the entire genome, we computed for each autosomal and X-linked annotated gene the ratio between the total number of B6 and Cast UMI counts grouping XX cells by time point and *Xist* AS classification (Appendix 3a). While the X-linked genes in *Xist*-MA cells showed the expected preferential detection of the *Xist*-negative allele over time as a result of the ongoing XCI process, the median B6-to-Cast ratio of autosomal and X-linked genes in *Xist* negative cells ranged between values of 0.99 and 1.05 throughout

cellular differentiation. This result suggested the absence of a strong genome-wide allelic mapping bias both in autosomal and X-linked genes.

The aim of the following analysis is to identify X-linked genes showing differential silencing kinetics between the two alleles, and to classify the silencing speed of each gene, while accounting for the faster silencing of the Cast allele compared to the B6 X chromosome.

7.1 XCI progress and linear model fit

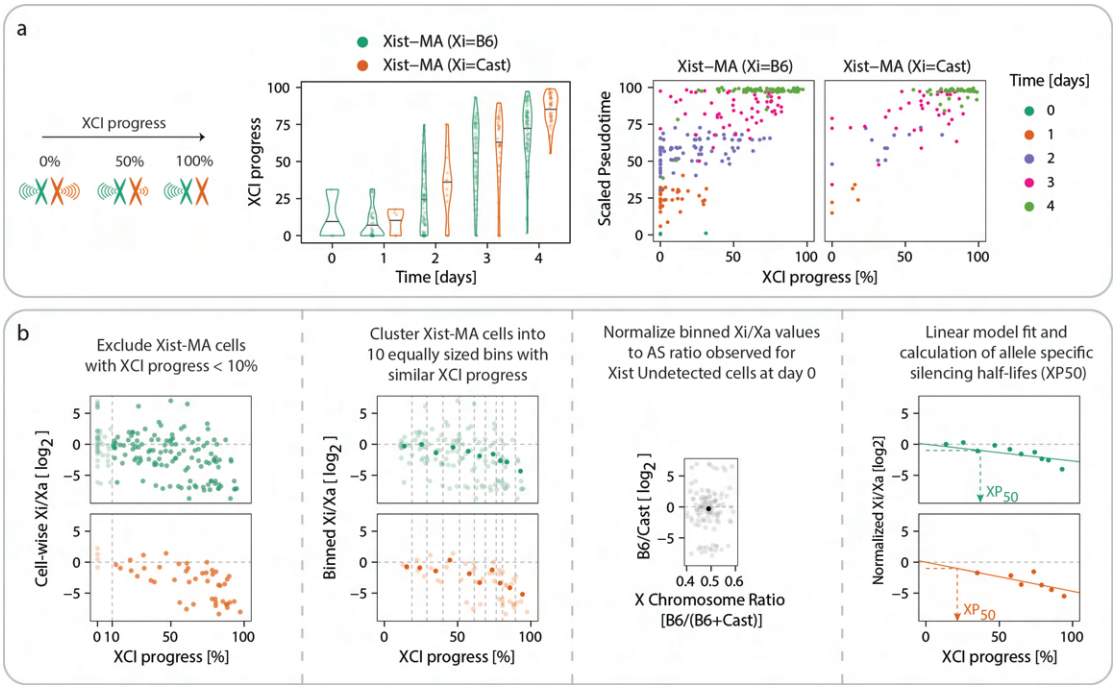


FIGURE 15: (a) Schematic representation of XCI progress (XP), defined as the percentage of X-silencing on Xi relative to the *Xist*-negative allele (left); violin plot of observed XP values in *Xist*-MA cells (center); comparison of XCI progress with scaled pseudotime, coloring *Xist*-MA cells by time point (right). (b) Step-by-step procedure for the identification of differentially silenced X-linked genes, shown for an example gene (Eif1ax). Transparent dots indicate individual cells and solid dots the binned values for cells with similar XP values. The solid line shows the log-linear fit, used to estimate the XP_{50} value (dashed arrows).

In this analysis we modeled the extent of silencing of every X-linked gene with respect to the silencing of the entire *Xist*-expressing chromosome, and compared the gene-wise trends observed in the two *Xist*-MA populations aiming to identify the genes with different silencing dynamics between the two alleles.

For every *Xist*-MA XX cell we defined a measure (Fig. 15a) representing the extent of chromosome-wide silencing, referred to as the "XCI progress" (XP, Eq. 12). This is computed as the percentage of silencing of the *Xist*-expressing X chromosome (Xi) relative to the *Xist*-negative allele (Xa). The XP values (Fig. 15a) were highly variable across the *Xist*-MA cells collected at the same time point or associated with similar

pseudotime values, reflecting the asynchronous nature of the XCI process in XX mESCs. Similarly, we quantified the silencing of each X-linked gene (Xi:Xa, Eq. 13) as the ratio between the UMI counts observed on the inactive and active alleles.

In order to identify differentially silenced X-linked genes (Fig. 15b), the analysis was restricted to the *Xist*-MA cells which had already initiated the XCI process, namely excluding cells with an XP value smaller than 10%. Aiming to robustly estimate the XP and Xi:Xa values, all *Xist*-MA cells sequenced throughout cellular differentiation were clustered into groups with similar extent of chromosome-wide silencing. This was achieved dividing the range of XP values observed across all *Xist*-MA cells into B equally sized bins ($B = 10$), and calculating for each bin and *Xist*-MA population the binned XP and Xi:Xa values by aggregating the AS UMI counts of every single cell assigned to the same bin (Eq. 28).

For each gene, the presence of basal expression skewing between the two alleles was accounted for by normalizing the binned Xi:Xa values to the B6:Cast ratio computed across *Xist*-negative undifferentiated cells, whose difference in expression between the two alleles is not caused by silencing nor by differentiation (Eq. 29). For each *Xist*-MA populations and X-linked gene, a zero-intercept linear model was then fitted to the $\log_2$-transformed normalized Xi:Xa values, regressing out the binned XP values (Eq. 30).

In order to quantify the silencing kinetics of each gene in each *Xist*-MA population, the estimated allele-specific slopes were used to compute the allelic XP_{50} values (Eq. 31). This measure represents the expected percentage of chromosome-wide silencing corresponding to a two-fold decrease in the gene expression on Xi relative to Xa.

Finally differentially silenced genes were identified through an ANOVA F test, which compares the fit of the allele-specific linear model to a simpler model fitting a single slope to both *Xist*-MA populations (Eq. 32).

7.2 Identification of differentially silenced genes

The allele-specific linear model fit was restricted to X-linked genes with at least 5 bins characterized by a minimum of 5 cells and at least 25 allele-specific counts. Only 35 X-linked genes passed this filtering step and could be fitted a linear model on both *Xist*-MA populations, while 39 additional genes could only be fitted for the population of cells which monoallelically expressed *Xist* on the B6 allele. The higher number of genes fitted for the latter subpopulation is due to the larger number of cells which monoallelically express *Xist* and undergo silencing on the B6 allele.

The allelic XP_{50} values of the 35 X-linked genes which were analyzed on both *Xist*-MA populations can be visualized through a scatter plot (Fig. 16a, left) that compares the

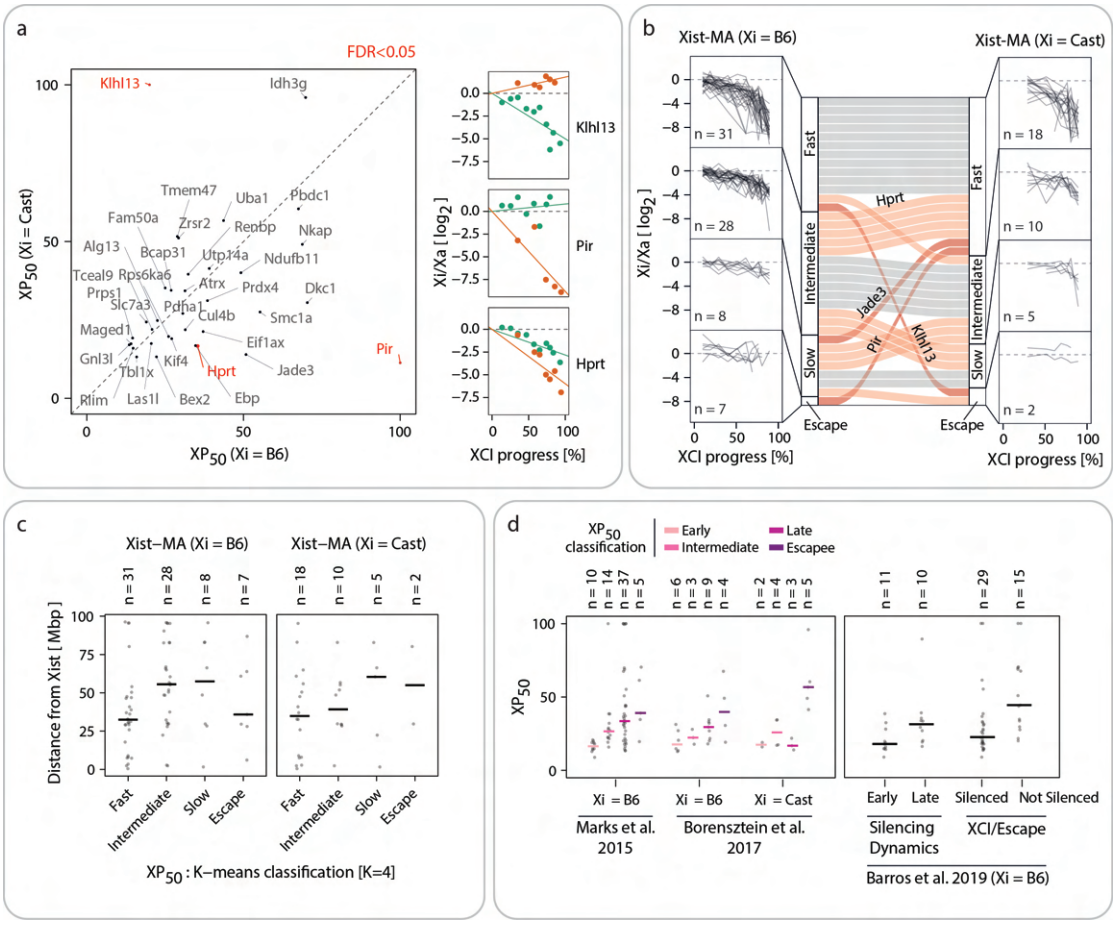


FIGURE 16: (a) Comparison of XP_{50} values estimated for the B6 and Cast chromosomes. Genes with significantly different silencing dynamics (ANOVA F test: BH-corrected p-value ≤ 0.05) are colored in red and shown on the right panels. (b) Allele-specific K-means clustering (K=4) of genes according to their XP_{50} values (side panels). The bars compare the allelic classifications for the 35 genes for which the XP_{50} value could be estimated on both chromosomes (center). (c) Genomic distance from the *Xist* gene for genes grouped according to their XP_{50} values as in (b): on the B6 (left, 74 genes) and Cast X chromosomes (right, 35 genes), respectively. Dots represent individual genes and the horizontal bars show the median value. (d) Comparison of the estimated XP_{50} values with previously determined silencing classes. Dots represent individual genes and the horizontal bars show the median value.

estimated gene-wise silencing dynamics of the two alleles. Upon normalization for the allelic global silencing dynamics, as expected the majority of genes resulted in XP_{50} values located on the proximity of the diagonal line, meaning that these X-linked genes were silenced with similar kinetics on the two alleles. On the other hand, the ANOVA F test identified three genes (*Klhl13*, *Pir* and *Hprt*) as significantly differentially silenced between the two alleles (Fig. 16a, right). Specifically *Klhl13* escaped silencing and was up-regulated on the Cast allele while being silenced on the B6 allele. On the other hand both *Pir* and *Hprt* were silenced significantly faster on the Cast allele. Notably, the results of this analysis were robust upon variations of the XP threshold and the number of equally sized bins (Appendix 3b), consistently identifying these three genes as being significantly differentially silenced between the two alleles.

The K-means clustering algorithm (K=4) was then applied to the estimated XP_{50} values separately for the B6 and Cast alleles (74 and 35 X-linked genes, respectively) in order to classify the allelic silencing kinetics of each gene as: fast, intermediate, slow, or escaping silencing (Fig. 16b). The majority of the X-linked genes which could be analyzed on both *Xist*-MA populations were assigned to the same or to the neighboring allelic silencing classes (19 grey and 13 orange bars, respectively), while the significantly differentially silenced genes *Klhl13* and *Pir* together with *Jade3* (ANOVA F test: FDR = 0.19) were assigned to the fast silencing class on one allele and to the slow or escaping group on the other (Fig. 16b, central panel). Furthermore, the average genomic distance from *Xist* in mega base pairs (Mbp) was computed for each X-linked gene, and the genes were grouped according to their K-means allelic silencing class (Fig. 16c). In accordance with previous studies [16, 60, 119], the X-linked genes located in proximity of the *Xist* locus were silenced faster compared to distal ones, with the exception of genes escaping the XCI process.

These results were then validated comparing the estimated allelic XP_{50} values with some previous studies which classified the allele-specific silencing dynamics of X-linked genes in female mice (Fig. 16d). The estimated XP_{50} values relative to genes silenced on the B6 allele were in good agreement with a previous analysis which measured the RNA expression of (129 x Cast) female mESCs by bulk RNA-Sequencing, where the silencing of the 129 X chromosome (closely related to the B6 mouse strain) was induced by a transcriptional stop in the Tsix of the 129 allele [119]. The allele-specific XP_{50} values were also concordant with the results of a previous study which used scRNA-sequencing to characterize the transcriptomic profile of pre-implantation (B6 x Cast) female mouse embryos in order to analyze the silencing dynamics of genes on both X chromosomes [16], although this comparison was restricted to a limited number of X-linked genes (Fig. 16d, left). Furthermore, the estimated allelic XP_{50} values were in good agreement with a study which measured the nascent transcriptome of (B6 x Cast) female mESCs by allele-specific PRO-seq, where *Xist* up-regulation and silencing on the B6 allele were induced by doxycycline treatment (Fig. 16d, left) [60]. Indeed, the estimated XP_{50} values for genes silencing the B6 allele which were classified as "Silencing Dynamics: Early" were significantly lower than the ones classified as "Silencing Dynamics: Late" (one-sided Wilcoxon rank sum test, p-value = 0.008), while the XP_{50} values of genes classified as "XCI/Escape: Silenced" were significantly lower than the ones classified as "XCI/Escape: Not Silenced" (one-sided Wilcoxon rank sum test, p-value = 0.0007). Notably, while all the 11 genes classified as "Silencing Dynamics: Early" were assigned to the fast or intermediate silencing class based on the K-means classification of their B6 XP_{50} values, discordant results were obtained for two X-linked genes (*Klhl13* and *Mmgt1*) which were assigned to the fast silencing class in our analysis while being classified as "XCI/Escape: Not Silenced" by the analysis of PRO-seq data.

This analysis shows that, when accounting for the overall faster silencing of the Cast X

chromosome, the majority of genes seems to have similar silencing kinetics on the two alleles. Nonetheless, the silencing dynamics of a number of genes appears to be altered by genetic variations between the B6 and Cast X chromosomes.

8 Experimental validation on non-random XCI cell line

The results of the differential silencing analysis were also validated through an orthogonal experimental approach, where mESCs underwent non-random XCI of either X chromosomes upon induced cellular differentiation.

8.1 Generation of ΔXic cell lines

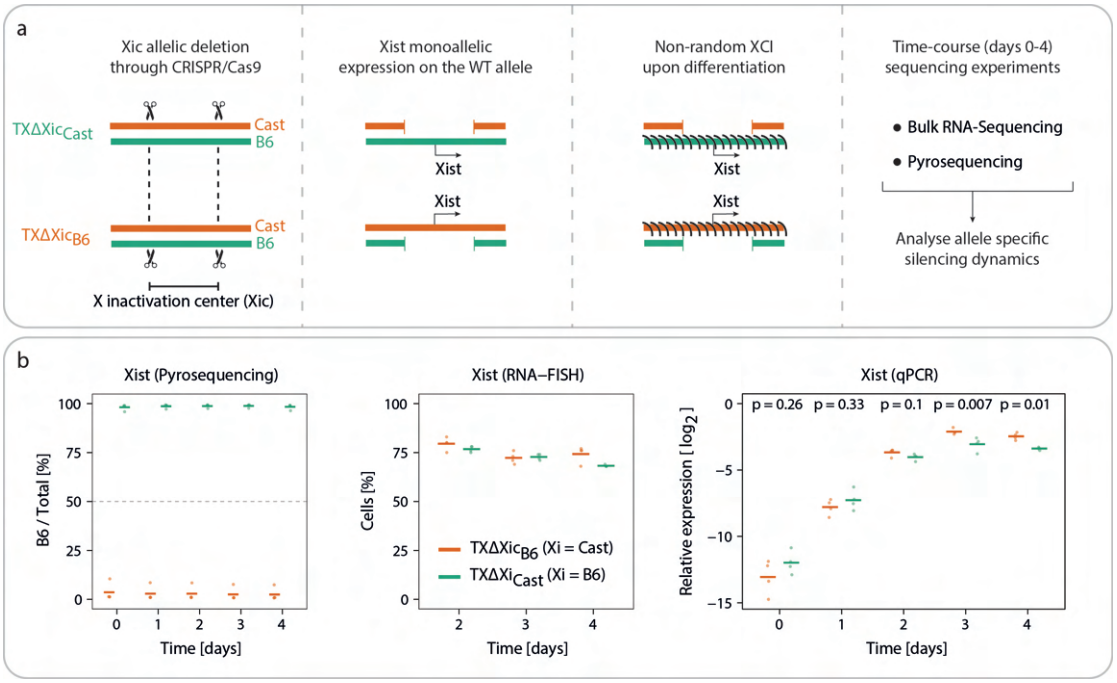


FIGURE 17: (a) Schematic representation of the generation of ΔXic cell lines, which undergo non-random XCI upon differentiation. (b) Comparison of *Xist* expression patterns in differentiating TXΔXic$_{B6}$ and TXΔXic$_{Cast}$ mESCs: allelic expression was quantified by Pyrosequencing (left), the percentage of *Xist*-positive cells was estimated by RNA-FISH (center), and relative expression was assessed by qPCR (right). For the latter plot, at each time point p-values of a two-sided unpaired Student's T-test are shown.

To assess the silencing dynamics of the B6 and Cast X chromosomes independently, two cell lines were generated deleting the X-inactivation center (Xic) of TX1072 mESCs on the B6 (TXΔXic$_{B6}$) or on the Cast (TXΔXic$_{Cast}$) alleles (Fig. 17a, Appendix 4a,b). Upon induced cellular differentiation, the presence of a single copy of the *Xist* gene on the wild-type (WT) X chromosome leads to monoallelic *Xist* expression, which initiates the non-random XCI of the WT allele. Therefore, TXΔXic$_{B6}$ and TXΔXic$_{Cast}$

mESCs respectively silenced the Cast and B6 X chromosomes. This experimental setting enabled the study of allele-specific silencing dynamics through bulk assays, such as Pyrosequencing and bulk RNA-sequencing experiments.

As expected both cell lines showed monoallelic expression of the *Xist* gene on the WT X chromosome (Fig. 17b, left), with a similar percentage of *Xist* expressing cells (Fig. 17b, center), showing an increase in *Xist* expression throughout cellular differentiation (Fig. 17b, right). Notably, the qPCR quantifications of *Xist* relative expression levels over time showed similar allele-specific expression levels at the earliest time points, and significantly higher expression of *Xist* on the Cast allele after 4 days of induced cellular differentiation. This strengthens the hypothesis that the significantly higher expression of *Xist* on the B6 allele which was previously observed in *Xist* monoallelic cells could be the result of a technical artifact (Fig. 9b,d).

8.2 Pyrosequencing

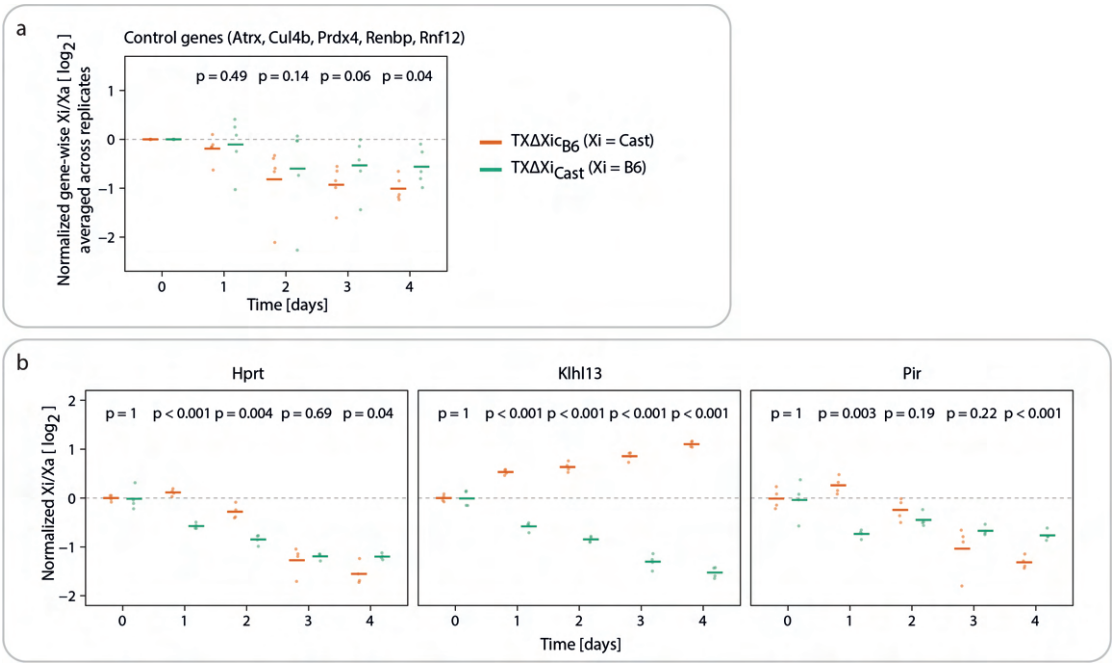


FIGURE 18: (a) Xi:Xa ratios, for 5 genes with similar XP_{50} values on both alleles, averaged across four biological replicates and normalized to the average ratio observed in undifferentiated cells (day 0). Every dot represents the average value of a single gene, and the horizontal bar their average value. For each time point p-values of a two-sided paired Student's T-test are shown. (b) Xi:Xa expression ratios in each replicate sample, for genes previously deemed as differentially silenced between the two alleles, normalized to the average ratio observed in undifferentiated cells (day 0). Every dot represents the ratio observed in each replicate sample, and the horizontal bar their average value. For each time point p-values of a two-sided unpaired Student's T-test are shown.

The allelic gene expression levels of $TX\Delta Xic_{B6}$ and $TX\Delta Xic_{Cast}$ cells throughout cellular differentiation were first measured through Pyrosequencing, which performs quantitative

sequencing over individual SNPs on cDNA. The allelic expression of the three genes previously deemed as being differentiatially silenced between the two mouse strains (*Hprt*, *Klhl13* and *Pir*) and five X-linked control genes which resulted in similar XP_{50} values (*Atrx*, *Cul4b*, *Prdx4*, *Renbp*, *Rnf12*) was measured at each time point on four biological replicates (Appendix 4c).

The allelic expression (Eq. 33) observed across the five control genes again highlights that, at the latest time points of cellular differentiation, the Cast allele is silenced faster compared to the B6 X chromosome (Fig. 18a).

Similarly to the previous analysis on single cell RNA-seq data, the Xi:Xa ratios observed on each biological replicate for the genes deemed as differentially silenced were first normalized to the average value observed in undifferentiated cells (day 0), and then compared between the two cell lines at each time point of differentiation. The results agree with the previous findings, showing that *Klhl13* escapes silencing on the Cast allele while being silenced on the B6 allele, and that *Hprt* and *Pir* are silenced significantly slower on the B6 allele after 4 days of cellular differentiation.

8.3 Bulk RNA-sequencing

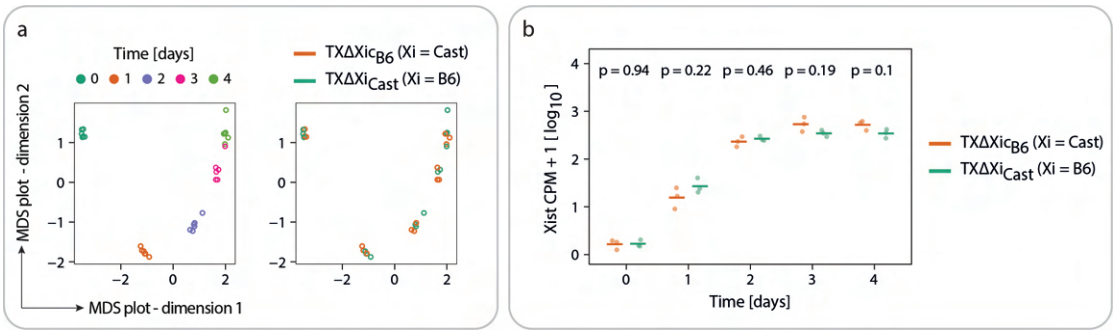


FIGURE 19: (a) Multi-Dimensional Scaling (MDS) plot, where the distance between each pair of samples is computed as the root-mean-square of the 500 genes with the largest $\log_2$ fold changes between the two samples. Every sequenced sample is colored by time point (left) and cell line (right). (b) *Xist* CPM expression in the TXΔXic$_{B6}$ and TXΔXic$_{Cast}$ cell lines throughout cellular differentiation. The pvalues on top derive by the two-sided unpaired Student's T-tests comparing the values observed in the two cell lines at each time point.

The genome wide gene expression levels of TXΔXic$_{B6}$ and TXΔXic$_{Cast}$ mESCs throughout induced cellular differentiation were then measured through bulk RNA-Sequencing experiments on three biological replicates per time point and cell line.

The difference in the transcriptomic profile of each sequenced sample can be explored through a multi-dimensional scaling (MDS) plot, which projects every sample on a two-dimensional subspace where the distance between each pair of samples is computed as the root-mean-square of the 500 genes with the largest $\log_2$ fold changes between the two samples (Fig. 19a). As expected, the MDS dimensionality reduction method clearly separates the undifferentiated samples from the ones sequenced throughout cellular differentiation, and highlights that the samples cluster together by sequencing time point.

In agreement with the previous qPCR quantification of *Xist* relative expression (Fig. 17b, right) the observed CPM expression levels were not deemed as significantly different between the two cell lines, although *Xist* expression was slightly higher on the Cast allele after 3 and 4 days of induced cellular differentiation (Fig. 19b).

8.4 Allelic XCI and validation of differentially silenced genes

The analysis of the Pyrosequencing and single-cell RNA-seq data suggested that the Cast X chromosome was silenced faster than the B6 allele over time (Fig. 9c, Fig. 18a). The same chromosome-wide difference in silencing kinetics is also observed through the analysis of bulk RNA-sequencing data (Fig. 20a) when comparing the Xi:Xa ratios of each X-linked gene (left) and of the entire chromosome X (right) between the two cell lines (Eq. 34).

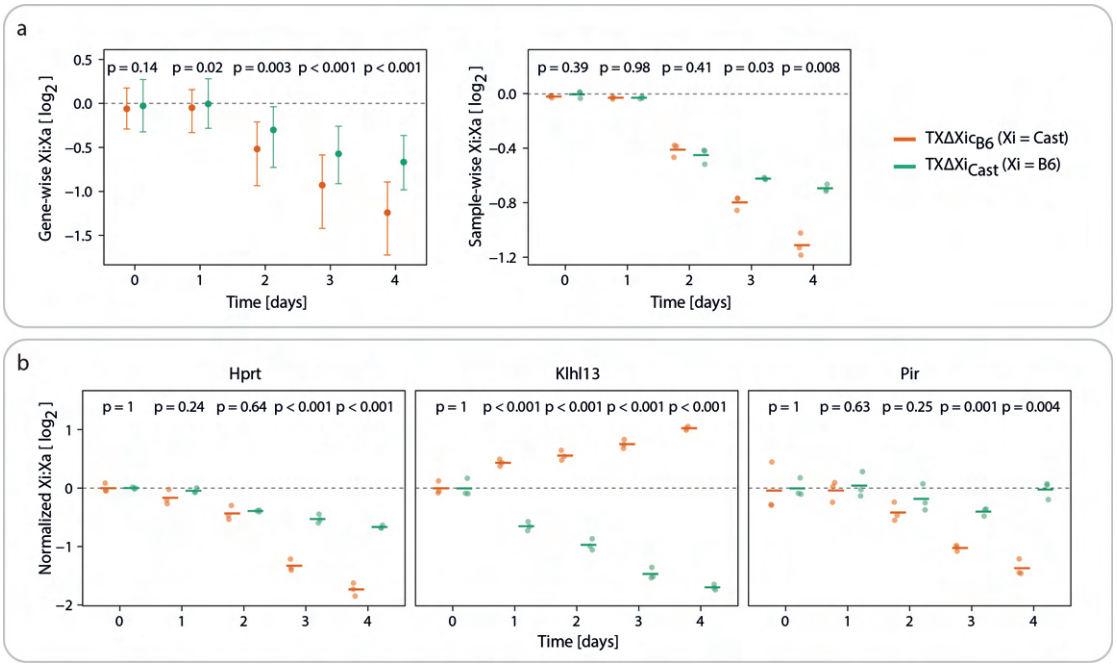


FIGURE 20: (a) Xi:Xa expression ratios for X-linked genes outside the CRISPR-deleted region (660 genes) computed: for each X-linked gene (median and first/third quartiles), summing up the expression of three biological replicates (left); and for each replicate sample (individual replicates and average), summing up the expression of X-linked genes (right). For each time point p-values of a two-sided paired Student's T-test (left) and p-values of a two-sided unpaired Student's T-test (right) are shown. (b) Xi:Xa expression ratios in each replicate sample, for genes previously deemed as differentially silenced between the two alleles, normalized to the average ratio observed in undifferentiated cells (day 0). Every dot represents the ratio observed in each replicate sample, and the horizontal bar their average value. For each time point p-values of a two-sided unpaired Student's T-test are shown.

The Xi:Xa ratios (Eq. 35) of every gene which was previously deemed as differentially silenced between the two X chromosomes (Fig. 16a) were normalized to the average baseline ratio observed across the three undifferentiated replicates (day 0), and the normalized ratios of the two cell lines were compared at each time point of cellular differentiation (Fig. 20b). The results of this analysis agreed with the differential silencing analysis of single cells. Indeed, *Klhl13* escaped silencing and was up-regulated on the Cast allele while being silenced on the B6 allele. On the other hand, both *Hprt* and *Pir* showed significantly lower Xi:Xa values on the Cast allele at late stages of differentiation, which was in agreement with their faster silencing on the Cast X chromosome.

8.5 X chromosome silencing map

While the single-cell RNA-seq data enabled the analysis of a restricted number of X-linked genes with allele-specific resolution, the higher sequencing depth and full length read coverage of bulk RNA-sequencing data could be exploited to explore the silencing

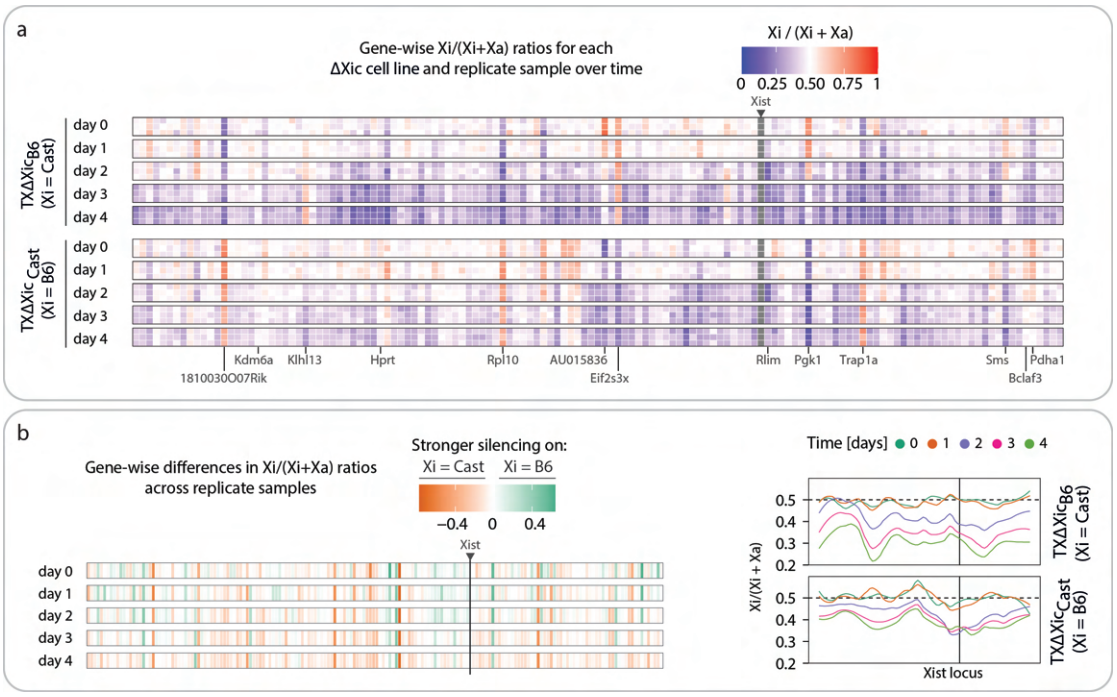


FIGURE 21: (a) Gene-wise expression ratios map (n = 136) of the mutated allele (Xi/Xt) across X-linked genes for each ΔXic cell line, time point and replicate, ordered by genomic position. The genes are ordered by their genomic position, and the Xist locus is highlighted by a black triangle. (b) Difference between the allelic ratios computed lumping counts observed across replicate samples on the two cell lines (left); LOESS fit (span = 1/5) of the allelic ratios observed for each time point and cell line (right)

dynamics of a much larger set of genes. This analysis aims to define a refined silencing map of the X chromosome for cells silencing the B6 ($TX\Delta Xic_{Cast}$) or the Cast ($TX\Delta Xic_{B6}$) alleles throughout differentiation.

This analysis was restricted to the subset of X-linked genes with at least 50 AS counts (Xi + Xa) detected for each cell line, sequencing time point and replicate sample (136 genes). For each cell line, replicate sample and X-linked gene, we computed the fraction of counts assigned to WT allele (Xi) which undergoes XCI upon differentiation (Fig. 21a). The comparison between the two cell lines revealed a subset of genes which were solely expressed by the B6 (*1810030O07Rik*, *Rpl10*, *Trap1a*) or by the Cast (*Eif2s3x*) alleles. For the latter set, this analysis showed three genes (*Pgk1*, *Sms* and *AU015836*) which seem to be almost uniquely expressed by the Cast allele at day 0 and 1, and then get reactivated only on the $TX\Delta Xic_{B6}$ allele. This map also highlights the known escapee *Kdm6a* [13], which has almost equal biallelic expression levels throughout sequencing time.

In order to compare the silencing kinetics of the alleles over time, we robustly estimated the Xi/(Xi+Xa) gene-wise ratios at each time point lumping the gene counts derived from each replicate sample, and computed the difference across biological replicates between the values observed on the two cell lines at each time point (Fig. 21b, left).

The overall gene-silencing trend across the X chromosome was then summarized for each time point by the LOESS fit of the observed fractions in the two cell lines (Fig. 21b, right). This analysis revealed that the difference in silencing kinetics between the two mouse strains was a chromosome-wide effect. Indeed the vast majority of X-linked genes under investigation were silenced faster on the Cast allele. Interestingly it can be noticed that, while TXΔXic$_{B6}$ mESCs silenced the WT allele progressively over time, the mESCs silencing the B6 allele seemed to reach a silencing plateau after two days of induced cellular differentiation.

Overall this analysis revealed that the silencing of X-linked genes proceeds faster throughout the entire Cast allele over time, while the B6 allele seems to reach a silencing peak after two days of differentiation. Moreover the silencing map describes a subset of genes uniquely expressed on one or the other allele.

8.6 Differential silencing analysis

Both the Pyrosequencing and bulk RNA-Sequencing data analyses comparing the allelic expression of putatively differentially silenced genes between the two cell lines over time (Fig. 18b, 20b) confirmed the results of the differential silencing analysis on single cell RNA-seq data (Fig. 16a). However a strong limitation of the analyses on the ΔXic lines is that they did not take into account the difference in global silencing kinetics between the two alleles (Fig. 21a,b). For this reason, it is difficult to conclude whether the difference observed for Hprt and Pir is due to a gene-specific effect, or rather if it is a consequence of the overall faster silencing of the Cast X chromosome. This issue was taken into account performing a differential silencing analysis similar to the one previously shown for single cell RNA-seq data (Fig. 22).

The following analysis (Fig. 22a) was restricted to the set of X-linked genes with at least 500 AS gene counts across all replicate samples for each time point throughout differentiation (177 genes). For each time point and cell line, a robust measure of silencing of the WT X chromosome (XP) and gene-wise silencing (Xi/Xa) was estimated lumping the allele-specific gene counts observed across replicate samples. For each gene and cell line, the observed silencing ratios were normalized to the value of undifferentiated cells (day 0) and a log-linear model was fitted to the normalized Xi/Xa ratios relative to the XP values observed over time. The allelic silencing dynamic of each X-linked gene was then measured through the XP_{50} statistic, and these values were used to assign every gene to an allele-specific silencing class through K-means clustering algorithm (Fig. 21a, left): fast (k=1), intermediate (k=2), slow (k=3) or escaping XCI (k=4). The density plots of the allele-specific XP_{50} values revealed a larger number of genes escaping the silencing process on B6 compared to the Cast allele (58 and 37 genes, respectively). Notably a set of genes which are known to escape the XCI process (namely: *Ddx3x*, *Pbdc1*, *Kdm5c* and *Kdm6a*) [13] were assigned to the escape class on both cell lines. Similarly

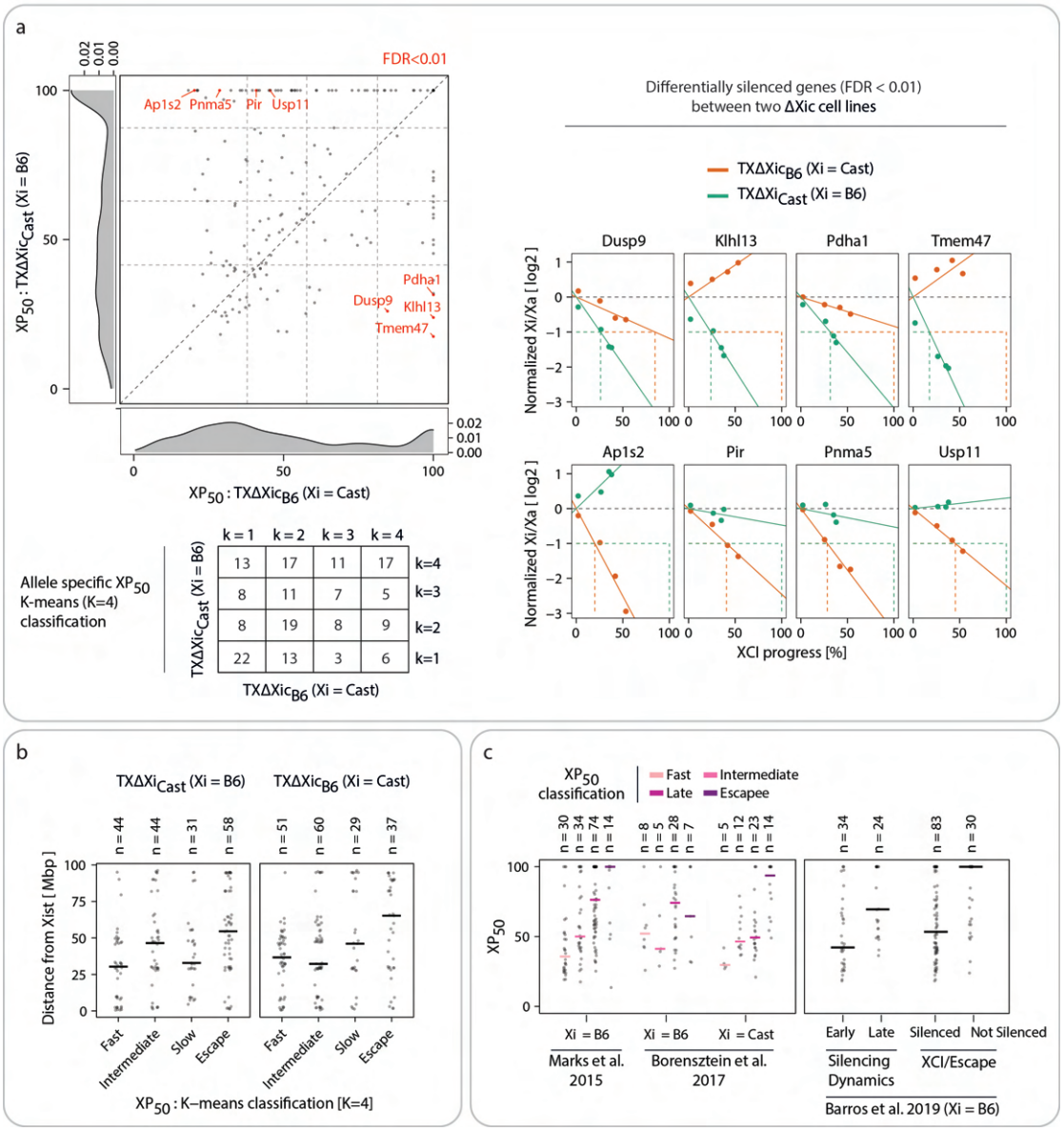


FIGURE 22: (a) Comparison of XP_{50} values estimated for the $TX\Delta Xic_{B6}$ and $TX\Delta Xic_{Cast}$ cell lines. Genes with significantly different silencing dynamics (ANOVA F test: BH-corrected p-value ≤ 0.05) are colored in red and shown on the right panels. (b) Allele-specific K-means clustering (K=4) of genes according to their XP_{50} values (side panels). The bars compare the allelic classifications for the 35 genes for which the XP_{50} value could be estimated on both chromosomes (center). (c) Genomic distance from the *Xist* gene for genes grouped according to their XP_{50} values as in (b): on the B6 (left, 74 genes) and Cast X chromosomes (right, 35 genes), respectively. Dots represent individual genes and the horizontal bars show the median value. (d) Comparison of the estimated XP_{50} values with previously determined silencing classes. Dots represent individual genes and the horizontal bars show the median value.

to the single-cell RNA-seq data analysis, putative differentially silenced genes were then identified through an ANOVA F test which compared the fit of the allele-specific linear model to a simpler model fitting the Xi/Xa ratios of the two cell lines with a single slope. Overall, this analysis identified 65 differentially silenced X-linked genes with FDR$\leq$0.05,

and 8 genes with FDR≤0.01 (Fig. 22a, right). This analysis validated the results of the previous single cell RNA-seq data analysis, confirming that *Klhl13* (FDR = 0.003), *Pir* (FDR = 0.009) and *Hprt* (FDR = 0.025) are differentially silenced between the two alleles. Furthermore, this analysis revealed the presence of a much larger number of genes with different silencing kinetics between the two alleles compared to the set of genes which were pointed out by the single cell RNA-seq data analysis (Fig. 16a/b).

Similarly to the single cell assay (Fig. 16c), the average genomic distance from *Xist* in mega base pairs (Mbp) was computed for each X-linked gene, and genes were grouped according to their K-means allelic silencing class (Fig. 22b). This comparison revealed that the distance from *Xist* of fast silenced genes was significantly smaller than the one observed for intermediate/slow silenced genes in the TXΔXic$_{Cast}$ cell line, while it was not in the TXΔXic$_{B6}$ cell line (two-sided unpaired Student's T-test, Bonferroni adjusted p-values: 0.02 and 0.13, respectively).

Furthermore, the allele-specific estimates of the gene-wise XP_{50} values largely agreed with some previous studies which classified the allele-specific silencing dynamics of X-linked genes in female mice [16, 119] (Fig. 22c). Moreover, the estimated XP_{50} values for genes silencing the B6 which were classified [60] as "Silencing Dynamics: Early" were significantly lower than the ones classified as "Silencing Dynamics: Late" (one-sided Wilcoxon rank sum test, p-value = 0.0004), while the XP_{50} values of genes classified as "XCI/Escape: Silenced" were significantly lower than the ones classified as "XCI/Escape: Not Silenced" (one-sided Wilcoxon rank sum test, p-value = 0.0003).

Overall the analyses of bulk RNA-sequencing and Pyrosequencing data from the TXΔXic cell lines confirmed the overall faster silencing of the Cast allele compared to the B6 allele. Where this difference derives by a chromosome-wide faster silencing of genes on the Cast allele, and not by a subset of differentially silenced genes. Moreover the above analyses confirmed that Klhl13 escapes silencing on the Cast allele, and that Hprt and Pir are silenced significantly faster on the Cast allele even when accounting for the overall difference in global silencing kinetics between the two alleles. Finally, the analysis of bulk RNA-sequencing data identified a larger set of differentially silenced genes compared to the previous single cell RNA-sequencing data analysis, and the estimated allelic gene-wise silencing kinetics were in good agreement with previous studies.

4 Conclusions

In the present work we studied the endogenous onset of Xist up-regulation and rXCI in mESCs undergoing cellular differentiation with allelic and strand specific transcriptomic resolutions. This was achieved performing single cell RNA-Sequencing of TX1072 female mESCs in undifferentiated state (2i&Lif) and throughout four days of cellular differentiation (induced upon 2i&Lif removal). Specifically, the design of this experiment enabled us to explore the transcriptomic profile of each cell at the not-AS and AS levels.

The read alignment procedure resulted a median of 400000 reads per cell, 30% of which uniquely aligned to exonic regions and carried a unique UMI barcode, while around 4% were exonic and overlapped at least one high-confidence SNP between the B6 and Cast genomes. These two subsets of reads were used to quantify the not-AS and AS transcriptomic profiles of each cell across approximately 9000 and 5000 murine annotated genes, respectively.

The analysis of not-AS gene counts revealed that XX cells up-regulated the lncRNA gene *Xist* after a single day of cellular differentiation, while *Xist*-mediated gene silencing initiated after one additional day. Both events are however very asynchronous throughout cellular differentiation, indeed approximately 20% of XX cells at day 4 did not express *Xist* nor initiated the rXCI process. Notably $Xist^{+}$ XX cells decreased the X:A ratio throughout differentiation as a result of *Xist*-mediated XCI process, while $Xist^{-}$ XX and XO mESCs underwent X-linked up-regulation over time.

The analysis of AS gene expression levels enabled us to better understand the mechanisms which characterize both *Xist* up-regulation and random XCI processes. As a result of the Xce effect, a higher fraction of TX1072 XX cells preferentially silenced the B6 allele, although the chromosome-wide gene silencing proceeded faster on the Cast allele. We observed transient biallelic *Xist* up-regulation in around 50% of cells at day 2 which was coupled by the partial silencing of both X chromosomes. Notably the silencing extent of these cells was even more pronounced than in their monoallelic counterparts. However the biallelic *Xist* expression and X-silencing was later resolved to a monoallelic state by four days of differentiation. The transient presence of *Xist* biallelic cells was also confirmed by orthogonal RNA-FISH experiments performed on the same cell line.

Differential expression analysis between *Xist* highly and lowly expressing cells, together with gene-wise correlation analyses with respect to *Xist* CPM expression levels were performed in order to identify putative *Xist* regulators. Moreover, we used the RNA-velocity method to model the spliced and unspliced transcripts of X-linked genes which estimated the future change in expression of the genes on the X chromosome (ΔX). Similarly to the previous analysis, we performed differential expression between ΔX high and low cells and correlation analyses. All these analyses were performed separately for XX cells sequenced at day 1 and 2, which correspond to the time points when we first observed a subset of cells initiating the *Xist* up-regulation and XCI processes. The

analyses based on *Xist* expression and ΔX provided highly concordant results identifying a set of known and novel regulators of *Xist* expression and *Xist*-mediated gene silencing.

We then accounted for the difference in X-silencing kinetics between the two alleles and fitted gene-wise log-linear models separately for the two *Xist*-MA populations in order to estimate and classify the silencing speed of each X-linked gene, and to further identify X-linked genes showing differential silencing kinetics between the two alleles. This analysis revealed that, when correcting for strain-specific global silencing effects, most genes showed similar silencing dynamics while only a handful of genes showed silencing differences between the B6 and Cast alleles. Furthermore the strain-specific classifications of the silencing speed of X-linked genes were highly concordant with previous studies, and the genes in proximity of the Xist locus resulted to be silenced significantly faster than distal ones, with the exception of genes escaping the XCI process.

Finally, these results were validated by orthogonal experiments (Pyro-Sequencing, RNA-FISH, qPCR and bulk RNA-Sequencing) performed on female B6xCast mESCs cell lines undergoing non-random XCI, achieved by the heterozygous deletion of the Xic locus on either alleles. The analyses of Pyro-sequencing and bulk RNA-Sequencing data confirmed the silencing trends observed for the genes deemed as differentially silenced by the scRNA-Seq data, and the faster silencing of the Cast allele. Furthermore, the analysis of bulk RNA-Seq data revealed the presence of a larger set of differentially silenced genes between the B6 and Cast X chromosomes, and further highlighted that the differences in silencing kinetics between the two alleles affected most of the genes along the X chromosome.

5 Discussion

In this study we profiled the transcriptomic regulation of mESCs during *Xist* up-regulation and random XCI at the early stages of cellular differentiation. This was achieved by the transcriptomic profiling of XX hybrid cell lines throughout four days of cellular differentiation. The presence of a high number of polymorphisms between the two mouse strains combined with the use of 3'-end scRNA-Sequencing technologies enabled us to explore the transcriptomic profiles of each mESC with allelic and strand specific resolution.

1 Xist gene expression quantification and regulation

The efficient C1-HT protocol library preparation combined with deep sequencing resulted in a median of 120,000 mRNA molecules per cell, which was significantly higher than other UMI-based methods [204]. The efficiency of this scRNA-Seq experiment enabled the quantification of the allelic expression of more than 5000 murine genes, including *Xist* and 158 other genes on the X chromosome. The relatively high number of X-linked genes whose transcripts overlapped high confidence SNPs was crucial to investigate the endogenous rXCI process of differentiating mESCs.

1.1 Xist transcripts' alignment

Differently from the vast majority of murine annotated genes, *Xist* (mm10, chrX: 103,453,491-103,482,714) did not show the expected 3'-end read alignment bias. Indeed the vast majority of its reads uniquely aligned to the 5'-end of its transcript, and were almost completely absent at its 3'-end locus.

Xist 5'-end read coverage was also observed in a previous study relying on the CEL-Seq2 3'-end biased sequencing protocol [83]. This suggests that the mild denaturation conditions of single cell protocols might not be sufficient to render *Xist* 3'-end polyA tails accessible for reverse transcription. This observation is in agreement with the problematic amplification of *Xist* single cell mRNA libraries both in vitro ESCs and in vivo blastocysts observed in previous studies [15, 171].

The inspection of *Xist* 5' locus revealed the presence of a 25bp long polyA sequence upstream of *Xist* reads' alignment region, which was likely targeted for reverse transcription. Nonetheless the quantification of *Xist* expression based on these transcripts showed the expected trend in gene expression. Indeed while the transcripts were almost completely absent in XO mESCs and undifferentiated XX mESCs, differentiating XX mESCs showed an increasing number of transcripts throughout developmental time.

This led us to the conclusion that the internal poly-A sequence was preferentially targeted for reverse transcription while *Xist* 3'-end was inaccessible, and that *Xist* 5'-end aligned reads could still provide a good quantification of the gene's expression levels.

The unexpected alignment of Xist reads highlights the importance of relying on strand specific sequencing protocols whenever quantifying the expression of genes characterized by anti-sense transcripts. This is especially relevant in studies involving *Xist*-mediated gene silencing where *Xist* 5'-end biased reads could be erroneously used to quantify *Tsix* expression, one of *Xist*'s major repressors [36, 38, 103]. Furthermore this rises questions on how many other genes might not be detected by single cell sequencing protocols due to the incapacity of the polyT sequencing primer to successfully ligate the polyA tail of the target mRNAs molecules and perform reverse transcription.

1.2 *Xist* allele-specific gene expression

The overlap between *Xist* transcripts and high confidence SNPs between the B6 and Cast mouse strains enabled the quantification of its expression with allelic resolution.

The transcriptomic profiling of *Xist* AS expression in XX mESCs confirmed the well known Xce effect which characterizes hybrid murine cell lines. Indeed we observed a much larger fraction of $Xist^+$ cells monoallelically expressing the gene on the B6 allele compared to Cast, leading to a higher fraction of cells which transcriptionally silence the B6 X chromosome [29, 31–33].

Comparing the *Xist* gene expression levels between the two monoallelic populations we observed a significantly higher expression of the B6 allele compared to Cast in scRNA-Sequencing data. Although the two alleles seemed to be equally expressed across all the autosomal genes whose transcripts overlapped strain specific SNPs, the fact that many other X-linked genes showed higher expression on the B6 allele suggests that this effect might be a technical artifact caused by the preferential mapping towards the reference genome. This was indeed confirmed by the analysis of *Xist* allelic expression levels in TX1072 and TXΔXic XX mESCs through bulk assays such as qPCR and bulk RNA-Sequencing, which revealed no significance difference between the B6 and Cast alleles.

Interestingly, the scRNA-Sequencing data revealed transient biallelic expression of *Xist* starting already at day 1 and reaching a peak at day 2 of differentiation with more than half of $Xist^+$ cells showing expression from both X chromosomes. Notably this effect was observed regardless on the UMI threshold used to define *Xist*-expressing cells. The fraction of cells expressing *Xist* on both alleles then decreased after the second day of differentiation, and the gene's expression was resolved to a monoallelic state at later time points. This trend was also confirmed by RNA-FISH experiments which showed that around 50% of differentiating mESCs have two *Xist* RNA clouds at day 2 of differentiation, and that this percentage considerably reduces at later time points coupled

with the increasing percentage of cells characterized by a single *Xist* RNA cloud. The transient biallelic *Xist* expression pattern observed in our work confirms the observations of previous experiments both in vitro differentiating mESCs and in vivo mouse embryos [80, 128, 181], and supports the model of stochastic *Xist* expression stating that the two alleles might independently undergo *Xist* up-regulation [124, 127, 128].

2 Transcriptional regulation of Xist and X chromosome

2.1 X chromosome regulation throughout differentiation

Similarly to a previous study we used the X:A ratio as a measure to explore the XCI process by comparing the expression of the X chromosomes and autosomes throughout cellular differentiation [16].

The not-AS X:A ratios revealed that *Xist*$^-$ XX and XO cells significantly increase the ratio throughout cellular differentiation, as a result of the X chromosome up-regulation (XCU) process. This process is thought to have evolved to compensate the loss of genes on the Y chromosomes [140], and was previously observed in differentiating male mESCs and in male pre/post-implantation embryos [16, 99, 106, 122, 197]. Surprisingly the AS X:A ratios of *Xist*$^-$ XX cells did not reveal any significant up-regulation over time. The discordance between the not-AS and AS ratios could be caused by the much smaller number of genes with allelic quantification which might not be enough to capture the extent of up-regulation observed at the not-AS level.

A recent work analysed Smart-Seq3 unstranded scRNA-Sequencing data to study B6xCast F1 hybrid mESCs differentiating into EpiSCs [103]. The authors concluded that the XCU process could only be observed on the active allele of XX cells which completed the XCI process (XaXi, such that: X chromosome ratio $\in$ [0, 0.1] or [0.9, 1]), while no up-regulation was observed for XX cells with two active X chromosomes (XaXa, such that: X chromosome ratio $\in$ [0.4, 0.6]). The discrepancy between this result and ours could be explained by the authors' definition of the XaXa subpopulation. Indeed the XaXa group would include not only the vast majority of *Xist* Undetected cell, but also all the *Xist* biallelic cells, together with a large fraction of *Xist* monoallelic cells. Such heterogeneous classification might mitigate the XCU effect across the XaXa population, which we solely observed in differentiating *Xist* Undetected XX and XO cells.

Xist$^+$ XX cells significantly decrease their X:A ratio starting one day after *Xist* up-regulation. The analysis of the X:A ratios separately for *Xist* monoallelic and biallelic cells revealed that the latter initiate the silencing of both X chromosomes prior to re-solving their *Xist* expression to a monoallelic state and silencing a single allele. This observation is in line with the model which postulates that both *Xist* up-regulation and X-silencing might take place independently on both alleles, and that *Xist* biallelic

expression would be reverted to a monoallelic state upon the complete silencing of an essential X-linked *Xist* activator which serves as a negative feedback to ensure the silencing of a single X chromosome [124, 128]. Notably the biallelic silencing observed in mice resembles the dampening effect observed in early human embryos, although mice seem to resolve the transient biallelic expression to a monoallelic state much faster compared to humans [147, 168].

The analysis of X:A and Xi/Xa ratios observed for the *Xist* monoallelic cells revealed that the X chromosome undergoes a significantly more efficient silencing on the Cast allele compared to its B6 counterpart. Notably the analysis of bulk RNA-Sequencing data not only confirmed this observation, but also showed that this effect can be observed chromosome-wide. Moreover this analysis revealed that while the Cast allele gradually decreases its expression after each day of cellular differentiation, the B6 X chromosomes seems to reach a silencing plateau after two days of cellular differentiation. These results suggest that polymorphisms between the two mouse strains, independently on *Xist* expression levels, modulate the silencing kinetics throughout the entire genome.

2.2 Identification of Xist putative regulators

Given the very heterogeneous and asynchronous expression patterns of *Xist* in differentiating XX mESCs, which was also reported in previous studies [36, 38], we identified putative *Xist* regulators either by performing DE analyses between *Xist* highly and lowly expressing cells or through correlation analyses with respect to *Xist* normalized gene expression levels, similarly to a previous work [95].

At day 1 when we first observed *Xist* up-regulation, only *Nanog* was deemed as a significant negative regulator of *Xist* expression levels in both analyses, while other known *Xist* regulators were not. This result suggests that the down-regulation of *Nanog*, more than other known pluripotency factors, might play a crucial role in the initiation of *Xist* up-regulation [10, 130]. On the other hand, at day 2 the down-regulation of a number of pluripotency factors previously implicated in *Xist* regulation [10, 130, 145, 188, 202], including *Nanog*, *Esrrb* and others, was coupled with the up-regulation of early differentiation factors such as *Pou3f1*, the de-novo methyltransferases *Dnmt3a* and *Dnmt3b*, the polycomb-like protein *Phf19* and the splicing factor *Zcrb* which all together played a role on early silencing of the X chromosome, as previously reported by other studies [10, 78, 194].

Interestingly these analyses highlighted a number of genes as putative *Xist* regulators which were not previously reported. Specifically both the correlation and DE analyses highlighted the X-linked kinase *Pim2*, an oncogene cooperating with the *Myc* transcription factor which is involved in cell survival and proliferation, as the gene showing the highest positive association to *Xist* expression [86, 123]. Furthermore the correlation

analyses at day 1 and both analyses at day 2 deemed the autosomal lncRNA *Snhg14*, a poorly studied gene only expressed from the paternally inherited chromosome [4], and the protein coding *Ldha*, a gene which was reported to interact with *Rac1* to promote glycolysis and cancer [107], as putative early activators. Given their significant association to *Xist* expression level at the time point when we first observed *Xist* up-regulation, it would be interesting to further investigate the regulatory roles of these genes.

Nonetheless the DE and correlation analyses have also revealed some unexpected associations to *Xist* expression and initiation of silencing. Indeed *Ctcf*, which encodes a protein preventing *Xist* up-regulation [186], was deemed as a positive regulator of *Xist* expression by correlation analyses both at day 1 and 2 of differentiation. Furthermore *Rlim*, acting as a trans-activator of Xist transcription [7, 76], appeared to be strongly down-regulated by $Xist^+$ cells upon the initiation of silencing at day 2. *Rlim* might however represent the essential X-linked *Xist* activator leading to the transition from biallelic to monoallelic *Xist* expression, which was discussed in the previous section [124, 128]. Its down-regulation observed after two days of differentiation might indeed be the result of its fast silencing in $Xist^+$ cells.

2.3 Strain-specific silencing kinetics

In this work we developed a strategy to account for global differences in silencing kinetics between the two mouse strains, aiming to classify the strain-specific silencing dynamic of each gene and to identify differentially silenced ones.

This analysis revealed that, when accounting from strain specific differences, most of the genes under investigation showed similar silencing dynamics on the two X chromosomes. Our gene-wise classification of silencing speed based on genes' silencing half-lifes was highly concordant with the results obtained in previous studies [16, 60, 119]. Similarly to these studies we also observed that genes located in the proximity of the *Xist* locus were silenced faster compared to distal ones, with the exception of genes escaping the silencing process.

Moreover this analysis led to the identification of three differentially silenced genes between the two mouse strains, namely: *Klhl13*, *Pir* and *Hprt*. The analysis of Pyrosequencing and bulk RNA-Sequencing data on the TXΔXic cell lines confirmed their significantly different silencing dynamics between the two alleles. These data revealed that *Klhl13* was indeed escaping silencing and getting significantly up-regulated throughout differentiation on the Cast chromosome, while being silenced on the B6 allele. Recent studies showed that *Klhl13* inhibits X-linked differentiation, and that its silencing might enable the differentiation of cells with two active X chromosomes [72, 171, 180]. Therefore its escape might have evolved to postpone and account for the faster silencing of the Cast allele. On the other hand, both *Pir* and *Hprt* were confirmed to be differentially

silenced on the two alleles, however these data could not clarify if these differences were caused by the overall faster silencing of the Cast allele or by polymorphisms between the two mouse strains. Finally, the analysis of bulk RNA-Seq data from the TXΔXic cell lines enabled the characterization of the allele-specific silencing dynamics of many more X-linked genes compared to the scRNA-Sequencing assay. Indeed it revealed the presence of a much larger set of differentially silenced genes between the two mouse strains, known escapees and genes solely expressed by a single allele.

This approach could be extended to explore the XCI status in other contexts such as in primary human cells. Although humans carry much fewer heterozygous SNPs compared to the hybrid murine cell line which was investigated throughout this study, scRNA-Sequencing technologies could be combined with variant calling methods to de novo identify SNPs loci aiming to explore the roles of escapees and allele specific XCI in disease susceptibility between the sexes [105, 192].

3 Outlook

3.1 Limitations

This work provides a detailed picture of the transcriptomic regulation of mESCs throughout *Xist* up-regulation and random XCI, however it presents some limitations.

The major limitation and advantage of this study is represented by the protocol used for the scRNA-Sequencing experiment. On one hand this protocol enables the collection of strand-specific reads, the use of UMI barcodes and the visual inspection of cells isolated on each IFC. On the other, strand specificity comes at the cost of a considerably lower read coverage throughout the mouse genome. As previously mentioned throughout this work, the use of a 3'-end biased protocol restricted the number of genes which could be explored with allelic resolution compared to full length assays. This was also clear when comparing the differential expression and silencing analyses performed using the bulk and single cell assays.

Furthermore the reduced transcript coverage of 3'-end biased protocols limits the allele specific analysis to rely on a handful of SNP loci whenever assigning a SNP-aligned read to either alleles. This might result in erroneous AS quantifications whenever sequencing errors occur on SNP loci, which was observed for a handful of XO cells showing biallelic or Cast monoallelic *Xist* expression. Nonetheless the use of strand specific reads enabled us to achieve an unbiased quantification of sense-antisense genes such as *Xist* and *Tsix*, which was crucial for all downstream analyses.

A further technical limitation of this study was the absence of XO cells at day 2 of differentiation, which was the time point where we first observed XCI in XX cells.

Nonetheless we could still draw meaningful observations from the XO cell lines such as their X-linked up-regulation throughout cellular differentiation, although this cell line was not extensively explored as this project mostly focused on XX mESCs.

Finally, an additional limitation of this study is that since mESCs undergo *Xist* up-regulation and rXCI in a more heterogeneous and less synchronized manner compared to in vivo embryos, the results of the present work might differ from the ones observed in vivo differentiating mouse embryos.

3.2 Future studies

The differential expression and correlation analyses of *Xist* expression levels and of the RNA-velocity predicted change in X-linked expression (ΔX) revealed a number of novel putative regulators.

These analyses however reflect mere associations of their expression levels to changes in expression of *Xist* or ΔX, hence it can not be concluded if these genes are direct regulators or rather if they are part of a more complex regulatory network leading to a significant change in expression. In order to elucidate their regulatory role, future studies could perturb these genes' expression though small interfering RNA (siRNA), short hairpin RNA (shRNA) or CRISPR-Cas9 screens which would reveal how their reduced or absent transcription affects the expression of other genes, together with the *Xist* up-regulation and XCI processes.

Moreover it would be interesting to repeat the experiments and analyses described throughout this work on a hybrid cell line characterized by paternally inherited B6 and maternally inherited Cast alleles, aiming to verify if and how the paternal iXCI affects the downstream gene-silencing process during rXCI. Finally the Xce effect and differential gene silencing could be further explored combining the scRNA-Sequencing with scATAC-sequencing and scHiC experiments, aiming to better understand how strain-specific polymorphisms affect chromatin accessibility and chromatin tri-dimensional architectures.

6 Appendix

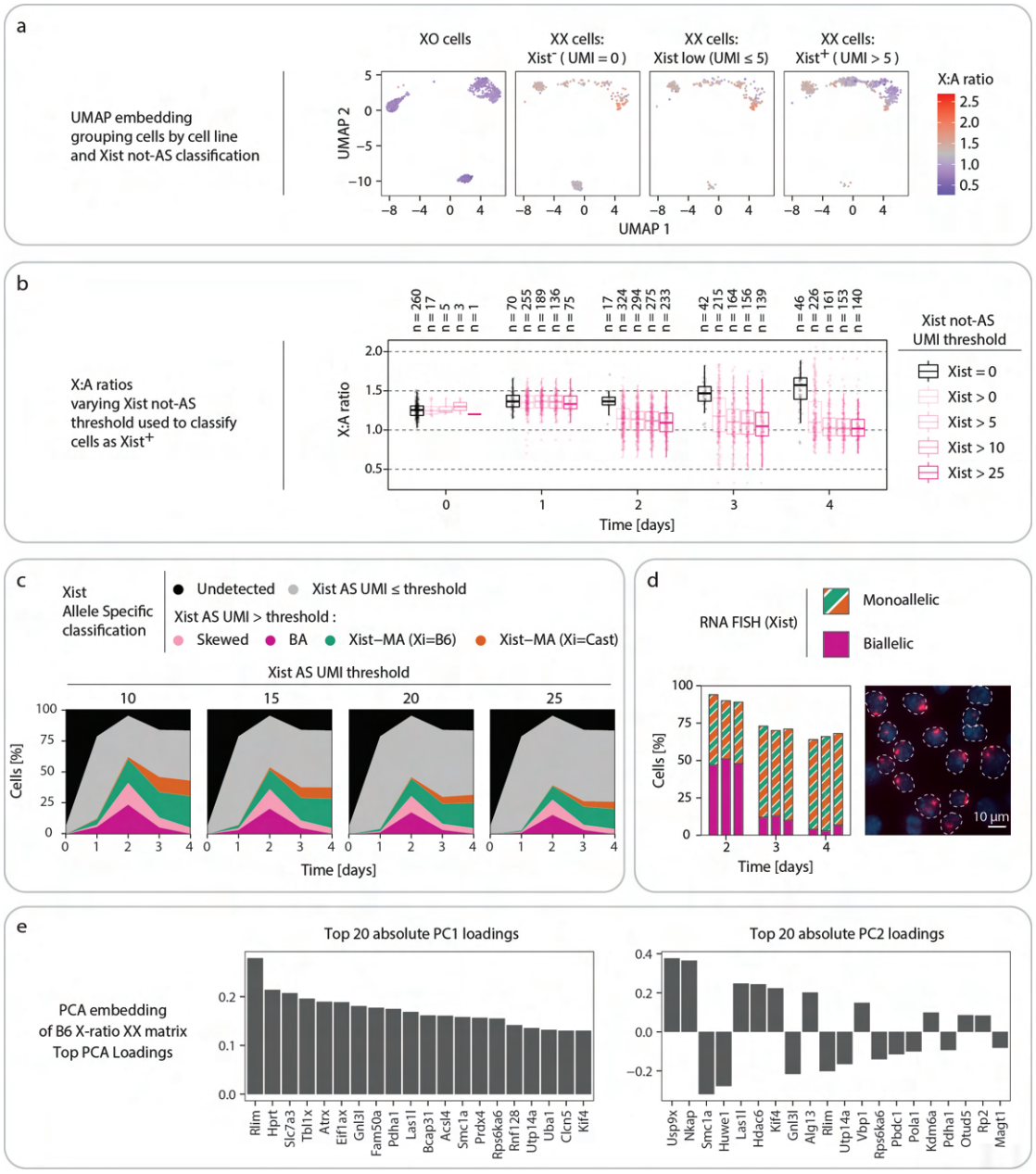


FIGURE 1: (a) UMAP embedding of single cells grouped by cell line and *Xist* classification and colored by bootstrapped X:A ratios values. (b) Box plots of bootstrapped X:A ratios obtained when varying the *Xist* not-AS used to classify cells as *Xist*$^+$, throughout developmental time. (c) Percentage of XX cells assigned to each *Xist* AS class obtained when varying the *Xist* AS UMI threshold used to classify *Xist*-expressing cells, throughout developmental time. (d) RNA FISH of *Xist* in TX1072 mESCs at days 2-4 of cellular differentiation. (left) Bar graph showing the quantification of three biological replicates over time (100 cells were counted for each replicate). (right) An example image at day 2 of differentiation, where dotted lines indicate the outline of cell nuclei stained with Dapi (blue) and Xist (red). (e) Bar plots representing the top 20 absolute loadings for the first two principal components of the embedding defined for the centered B6 X-ratio matrix computed on XX mESCs.

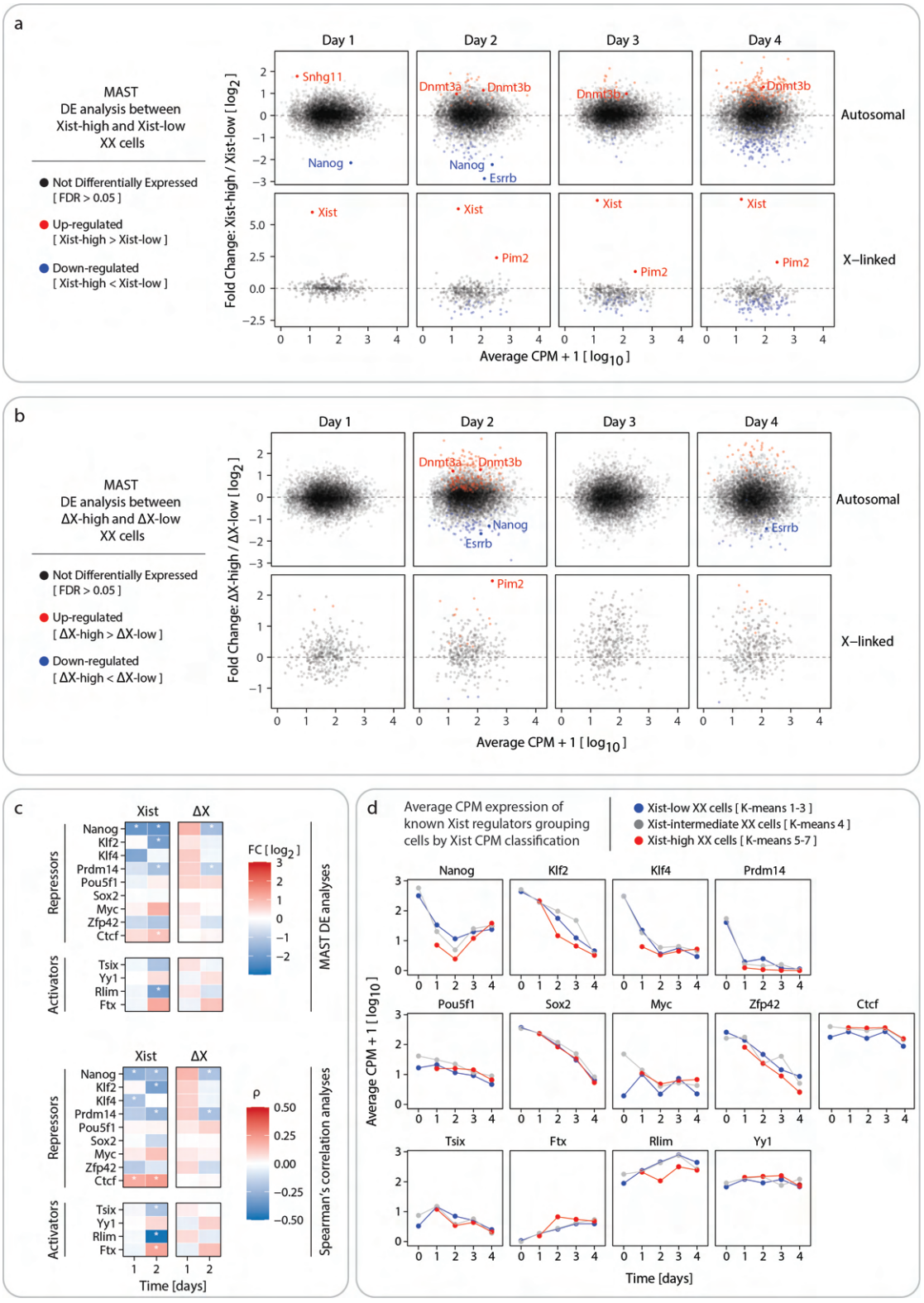


FIGURE 2: (a) For each sequencing time point, the scatter plots show the results of MAST DE analysis between *Xist*-high and *Xist*-low XX cells where significantly up-/down-regulated genes (FDR *le* 0.05) are colored in red/blue while the not significant genes are colored in black. Where every point represents the average normalized expression and observed $\log_2$FCs (*Xist*-high vs *Xist*-low cells) for each single gene being tested. (b) Same scatter plot as described in (a), obtained when performing MAST DE analysis between the ΔX-high and ΔX-low XX cells. (c) Heatmaps showing the results of MAST DE and Correlation analyses obtained when comparing *Xist*-high and *Xist*-low XX cells (left) or ΔX-high and ΔX-low XX cells (right) for a set of genes which have been previously implicated in *Xist* regulation. The boxes labeled with a white star represent significant (FDR *le* 0.05) results. (d) Line plots showing the average normalized expression of the same genes showed in (c) throughout developmental time, computed separately for XX cells assigned to different *Xist* not-AS expression classes.

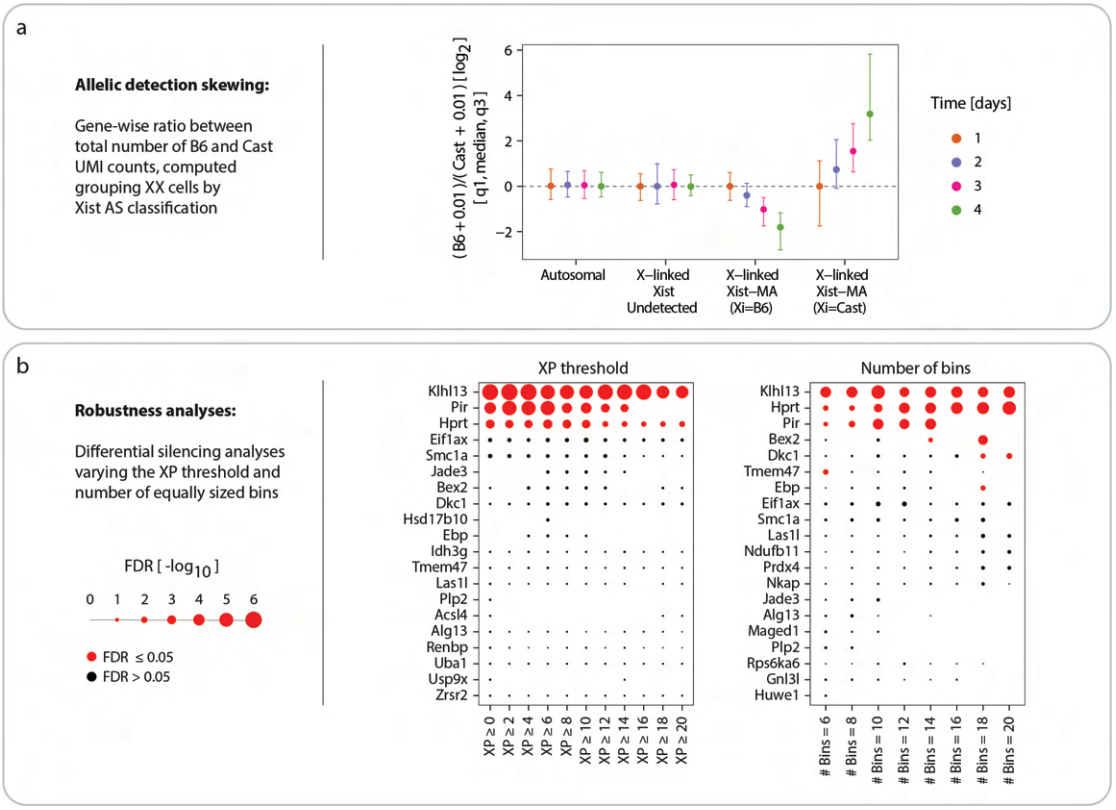


FIGURE 3: (a) Plot showing the gene-wise ratios e (median and first/third quartiles) computed lumping up all B6 and Cast AS UMI counts across all XX cells, separately for autosomal and X-linked genes (divided by *Xist* Undetected and MA cells), throughout developmental time. (b) Dot plot showing the significance (dot size) of X-linked genes when performing the differential silencing analysis, obtained when varying the XP thresholds and the number of bins.

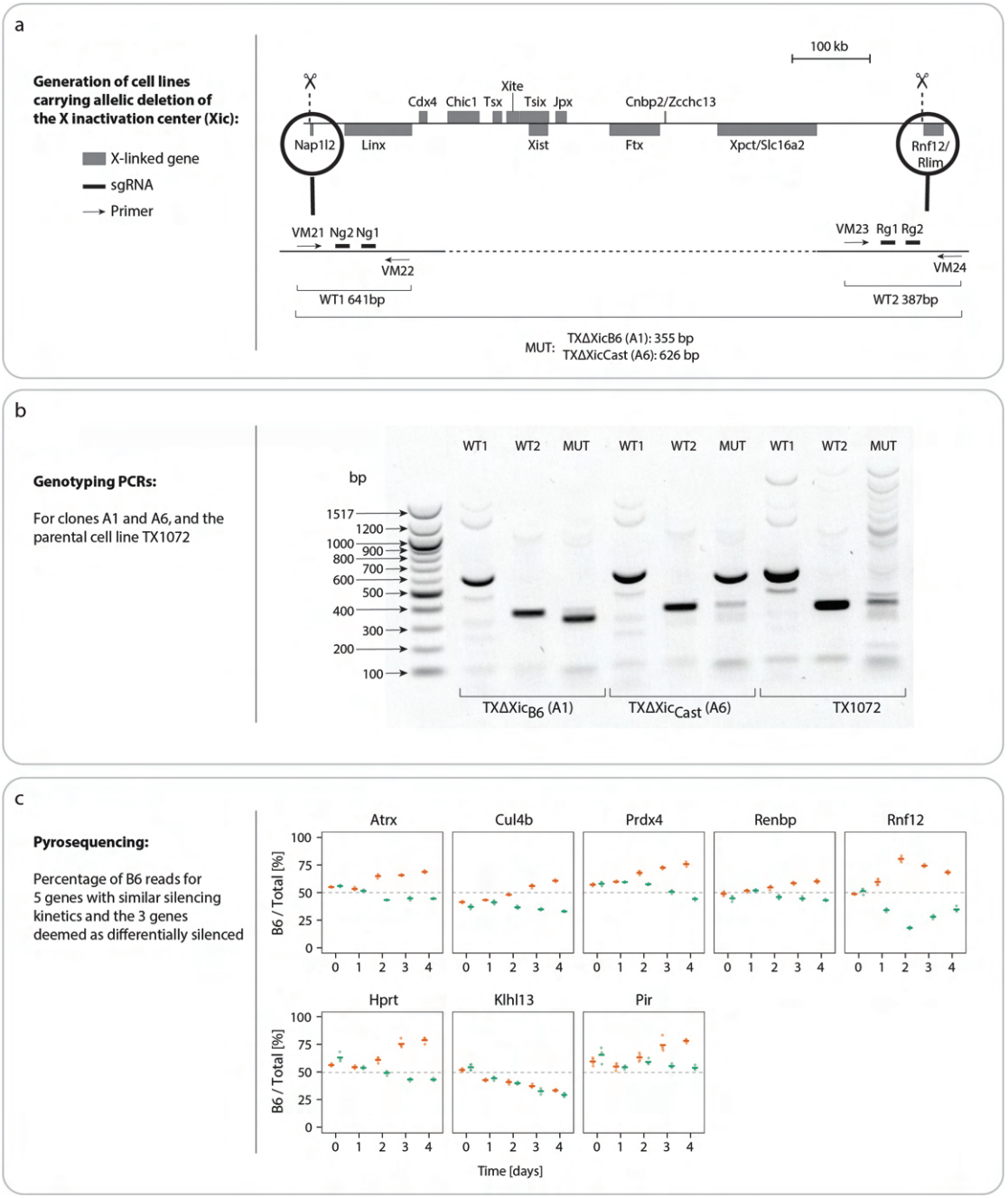


FIGURE 4: (a) Schematic representation of the X inactivation center (top). Genes on the plus strand are shown above the line and genes on the minus strand below. Scissors demark the deleted region. The position of the sgRNA used to generate the deletion (bars) and of the primers (arrows) used for genotyping are shown together with the expected sizes for the PCR products. Namely, TXΔXic$_{B6}$ carries the deletion on the B6 chromosome (chrX:103,182,701-103,955,531, mm10), while TXΔXic$_{Cast}$ on the Cast allele (chrX:103,182,257-103,955,698, mm10). (b) Genotyping PCRs for clones A1 and A6 and the parental cell line TX1072. The experiment was performed twice with similar results and the identity of the PCR bands was confirmed by Sanger sequencing. (c) Scatter plots showing the percentage of Pyro-sequencing reads assigned to the B6 allele for the TXΔXic$_{B6}$ (orange) and TXΔXic$_{Cast}$ (green) cell lines, for each replicate sample (dot) with the horizontal bar representing the average value. Data are shown for five genes with similar silencing kinetics (top) and for the three genes deemed as differentially silenced between the two TXΔXic cell lines.

Bibliography

[1] Robin L Adrianse, Kaleb Smith, Tonibelle Gatbonton-Schwager, Smitha P Sripathy, Uyen Lao, Eric J Foss, Ruben G Boers, Joachim B Boers, Joost Gribnau, and Antonio Bedalov. Perturbed maintenance of transcriptional repression on the inactive x-chromosome in the mouse brain after xist deletion. *Epigenetics & chromatin*, 11(1):1–13, 2018.

[2] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data2. *Genome biology*, 11(10):R106, 2010.

[3] Montserrat C Anguera, Weiyuan Ma, Danielle Clift, Satoshi Namekawa, Raymond J Kelleher III, and Jeannie T Lee. Tsx produces a long noncoding rna and has general functions in the germline, stem cells, and brain. *PLoS genetics*, 7(9): e1002248, 2011.

[4] Shadi Ariyanfar and Deborah J Good. Analysis of snhg14: A long non-coding rna hosting snord116, whose loss contributes to prader–willi syndrome etiology. *Genes*, 14(1):97, 2023.

[5] Sandrine Augui, Elphège P Nora, and Edith Heard. Regulation of x-chromosome inactivation by the x-inactivation centre. *Nature Reviews Genetics*, 12(6):429–442, 2011.

[6] Bradley P Balaton and Carolyn J Brown. Escape artists of the x chromosome. *Trends in Genetics*, 32(6):348–359, 2016.

[7] Tahsin Stefan Barakat, Nilhan Gunhanlar, Cristina Gontan Pardo, Eskeatnaf Mulugeta Achame, Mehrnaz Ghazvini, Ruben Boers, Annegien Kenter, Eveline Rentmeester, J Anton Grootegoed, and Joost Gribnau. Rnf12 activates xist and is essential for x chromosome inactivation. *PLoS genetics*, 7(1):e1002001, 2011.

[8] Tahsin Stefan Barakat, Friedemann Loos, Selma van Staveren, Elvira Myronova, Mehrnaz Ghazvini, J Anton Grootegoed, and Joost Gribnau. The trans-activator rnf12 and cis-acting elements effectuate x chromosome inactivation independent of x-pairing. *Molecular cell*, 53(6):965–978, 2014.

[9] Murray L Barr and Ewart G Bertram. A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature*, 163(4148):676–677, 1949.

[10] Antonio Barral, Isabel Rollan, Hector Sanchez-Iranzo, Wajid Jawaid, Claudio Badia-Careaga, Sergio Menchero, Manuel J Gomez, Carlos Torroja, Fatima Sanchez-Cabo, Berthold Göttgens, et al. Nanog regulates pou3f1 expression at the exit from pluripotency during gastrulation. *Biology open*, 8(11), 2019.

[11] PR Baverstock, M Adams, RW Polkinghorne, and M Gelder. A sex-linked enzyme in birds—z-chromosome conservation but no dosage compensation. *Nature*, 296 (5859):763–766, 1982.

[12] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[13] Joel B Berletch, Wenxiu Ma, Fan Yang, Jay Shendure, William S Noble, Christine M Disteche, and Xinxian Deng. Escape from x inactivation varies in mouse tissues. *PLoS Genet*, 11(3):e1005079, 2015.

[14] Giancarlo Bonora, Vijay Ramani, Ritambhara Singh, He Fang, Dana L Jackson, Sanjay Srivatsan, Ruolan Qiu, Choli Lee, Cole Trapnell, Jay Shendure, et al. Single-cell landscape of nuclear configuration and gene expression during stem cell differentiation and x inactivation. *Genome Biology*, 22(1):1–36, 2021.

[15] Maud Borensztein, Ikuhiro Okamoto, Laurène Syx, Guillaume Guilbaud, Christel Picard, Katia Ancelin, Rafael Galupa, Patricia Diabangouaya, Nicolas Servant, Emmanuel Barillot, et al. Contribution of epigenetic landscapes and transcription factors to x-chromosome reactivation in the inner cell mass. *Nature communications*, 8(1):1297, 2017.

[16] Maud Borensztein, Laurène Syx, Katia Ancelin, Patricia Diabangouaya, Christel Picard, Tao Liu, Jun-Bin Liang, Ivaylo Vassilev, Rafael Galupa, Nicolas Servant, et al. Xist-dependent imprinted x inactivation and the early developmental consequences of its failure. *Nature structural & molecular biology*, 24(3):226, 2017.

[17] Aurélie Bousard, Ana Cláudia Raposo, Jan Jakub Żylicz, Christel Picard, Vanessa Borges Pires, Yanyan Qi, Cláudia Gil, Laurène Syx, Howard Y Chang, Edith Heard, et al. The role of xist-mediated polycomb recruitment in the initiation of x-chromosome inactivation. *EMBO reports*, 20(10):e48019, 2019.

[18] FA Brook and RL Gardner. The origin and efficient derivation of embryonic stem cells in the mouse. *Proceedings of the National Academy of Sciences*, 94(11):5709–5712, 1997.

[19] Carolyn J Brown and Huntington F Willard. The human x-inactivation centre is not required for maintenance of x-chromosome inactivation. *Nature*, 368(6467): 154–156, 1994.

[20] Carolyn J Brown, Andrea Ballabio, James L Rupert, Ronald G Lafreniere, Markus Grompe, Rossana Tonlorenzi, and Huntington F Willard. A gene from the region of the human x inactivation centre is expressed exclusively from the inactive x chromosome. *Nature*, 349(6304):38–44, 1991.

[21] Carolyn J Brown, Ronald G Lafreniere, Vicki E Powers, Gianfranco Sebastio, Andrea Ballabio, Anjana L Pettigrew, David H Ledbetter, Elaine Levy, Ian W Craig, and Huntington F Willard. Localization of the x inactivation centre on the human x chromosome in xq13. *Nature*, 349:82–84, 1991.

[22] Carolyn J Brown, Brian D Hendrich, Jim L Rupert, Ronald G Lafreniere, Yigong Xing, Jeanne Lawrence, and Huntington F Willard. The human xist gene: analysis of a 17 kb inactive x-specific rna that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71(3):527–542, 1992.

[23] Spencer W Brown. Heterochromatin. *Science*, 151(3709):417–425, 1966.

[24] Davide Cacchiarelli, Xiaojie Qiu, Sanjay Srivatsan, Anna Manfredi, Michael Ziller, Eliah Overbey, Antonio Grimaldi, Jonna Grimsby, Prapti Pokharel, Kenneth J Livak, et al. Aligning single-cell developmental and reprogramming trajectories identifies molecular determinants of myogenic reprogramming outcome. *Cell systems*, 7(3):258–268, 2018.

[25] J Mauro Calabrese, Wei Sun, Lingyun Song, Joshua W Mugford, Lucy Williams, Della Yee, Joshua Starmer, Piotr Mieczkowski, Gregory E Crawford, and Terry Magnuson. Site-specific silencing of regulatory elements as a mechanism of x inactivation. *Cell*, 151(5):951–963, 2012.

[26] John D Calaway, Alan B Lenarcic, John P Didion, Jeremy R Wang, Jeremy B Searle, Leonard McMillan, William Valdar, and Fernando Pardo-Manuel de Villena. Genetic architecture of skewed x inactivation in the laboratory mouse. *PLoS Genetics*, 9(10):e1003853, 2013.

[27] Sarah Carmona, Benjamin Lin, Tristan Chou, Katti Arroyo, and Sha Sun. Lncrna jpx induces xist expression in mice using both trans and cis mechanisms. *PLoS genetics*, 14(5):e1007378, 2018.

[28] Mark G Carter, Alexei A Sharov, Vincent VanBuren, Dawood B Dudekula, Condie E Carmack, Charlie Nelson, and Minoru SH Ko. Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. *Genome biology*, 6(7):R61, 2005.

[29] BM Cattanach and JH Isaacson. Controlling elements in the mouse x chromosome. *Genetics*, 57(2):331, 1967.

[30] BM Cattanach and D Papworth. Controlling elements in the mouse: V. linkage tests with x-linked genes. *Genetics Research*, 38(1):57–70, 1981.

[31] BM Cattanach and JN Perez. Parental influence on x-autosome translocation-induced variegation in the mouse. *Genetics Research*, 15(1):43–53, 1970.

[32] BM Cattanach and CE Williams. Evidence of non-random x chromosome activity in the mouse. *Genetics Research*, 19(3):229–240, 1972.

[33] BM Cattanach, CE Pollard, and JN Perez. Controlling elements in the mouse x-chromosome: I. interaction with the x-linked genes. *Genetics Research*, 14(3): 223–235, 1969.

[34] Lisa Helbling Chadwick, Lisa M Pertz, Karl W Broman, Marisa S Bartolomei, and Huntington F Willard. Genetic control of x chromosome inactivation in mice: definition of the xce candidate interval. *Genetics*, 173(4):2103–2110, 2006.

[35] Julie Chaumeil, Patricia Le Baccon, Anton Wutz, and Edith Heard. A novel role for xist rna in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes & development*, 20(16):2223–2237, 2006.

[36] Geng Chen, John Paul Schell, Julio Aguila Benitez, Sophie Petropoulos, Marlene Yilmaz, Björn Reinius, Zhanna Alekseenko, Leming Shi, Eva Hedlund, Fredrik Lanner, et al. Single-cell analyses of x chromosome inactivation dynamics and pluripotency during differentiation. *Genome research*, 26(10):1342–1354, 2016.

[37] Wenan Chen, Yan Li, John Easton, David Finkelstein, Gang Wu, and Xiang Chen. Umi-count modeling and differential expression analysis for single-cell rna sequencing. *Genome biology*, 19(1):70, 2018.

[38] Shangli Cheng, Yu Pei, Liqun He, Guangdun Peng, Björn Reinius, Patrick PL Tam, Naihe Jing, and Qiaolin Deng. Single-cell rna-seq reveals cellular heterogeneity of pluripotency transition and x chromosome dynamics during early mouse development. *Cell reports*, 26(10):2593–2607, 2019.

[39] Ci Chu, Qiangfeng Cliff Zhang, Simão Teixeira Da Rocha, Ryan A Flynn, Maheetha Bharadwaj, J Mauro Calabrese, Terry Magnuson, Edith Heard, and Howard Y Chang. Systematic discovery of xist rna binding proteins. *Cell*, 161 (2):404–416, 2015.

[40] Corinne Chureau, Sophie Chantalat, Antonio Romito, Angélique Galvani, Laurent Duret, Philip Avner, and Claire Rougeulle. Ftx is a non-coding rna which affects xist expression and chromatin structure within the x-inactivation center region. *Human molecular genetics*, 20(4):705–718, 2011.

[41] Paolo Cinelli, Elisa A Casanova, Syndi Uhlig, Priska Lochmatter, Takahiko Matsuda, Takashi Yokota, Thomas Rülicke, Birgit Ledermann, and Kurt Bürki. Expression profiling in transgenic fvb/n embryonic stem cells overexpressing stat3. *BMC developmental biology*, 8(1):57, 2008.

[42] Christine Moulton Clemson, John A McNeil, Huntington F Willard, and Jeanne Bentley Lawrence. Xist rna paints the inactive x chromosome at interphase: evidence for a novel rna involved in nuclear/chromosome structure. *The Journal of cell biology*, 132(3):259–275, 1996.

[43] Philippe Clerc and Philip Avner. Role of the region 3 to xist exon 6 in the counting process of x-chromosome inactivation. *Nature genetics*, 19(3):249–253, 1998.

[44] Dena E Cohen, Lance S Davidow, Jennifer A Erwin, Na Xu, David Warshawsky, and Jeannie T Lee. The dxpas34 repeat regulates random and imprinted x inactivation. *Developmental cell*, 12(1):57–71, 2007.

[45] David Colognori, Hongjae Sunwoo, Andrea J Kriz, Chen-Yu Wang, and Jeannie T Lee. Xist deletional analysis reveals an interdependency between xist rna and polycomb complexes for spreading along the inactive x. *Molecular cell*, 74(1): 101–117, 2019.

[46] Thomas Conrad and Asifa Akhtar. Dosage compensation in drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nature Reviews Genetics*, 13(2):123–134, 2012.

[47] Sarah Cooper, Anne Grijzenhout, Elizabeth Underwood, Katia Ancelin, Tianyi Zhang, Tatyana B Nesterova, Burcu Anil-Kirmizitas, Andrew Bassett, Susanne M Kooistra, Karl Agger, et al. Jarid2 binds mono-ubiquitylated h2a lysine 119 to mediate crosstalk between polycomb complexes prc1 and prc2. *Nature communications*, 7(1):13661, 2016.

[48] Györgyi Csankovszki, Barbara Panning, Brian Bates, John R Pehrson, and Rudolf Jaenisch. Conditional deletion of xist disrupts histone macroh2a localization but not maintenance of x inactivation. *Nature genetics*, 22(4):323–324, 1999.

[49] David B Cunningham, Dominique Segretain, Danielle Arnaud, Ute C Rogner, and Philip Avner. The mousetsxgene is expressed in sertoli cells of the adult testis and transiently in premeiotic germ cells during puberty. *Developmental biology*, 204 (2):345–360, 1998.

[50] Simão Teixeira da Rocha, Valentina Boeva, Martin Escamilla-Del-Arenal, Katia Ancelin, Camille Granier, Neuza Reis Matias, Serena Sanulli, Jen Chow, Edda Schulz, Christel Picard, et al. Jarid2 is implicated in the initial xist-induced targeting of prc2 to the inactive x chromosome. *Molecular cell*, 53(2):301–316, 2014.

[51] Alessandra Dal Molin, Giacomo Baruzzo, and Barbara Di Camillo. Single-cell rna-sequencing: assessment of differential expression analysis methods. *Frontiers in genetics*, 8:62, 2017.

[52] E Debrand, C Chureau, D Arnaud, P Avner, and E Heard. Functional analysis of the dxpas34 locus, a 3 regulator of xist expression. *Molecular and cellular biology*, 19(12):8513–8525, 1999.

[53] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

[54] Christine M Disteche. Dosage compensation of the sex chromosomes. *Annual review of genetics*, 46:537–560, 2012.

[55] Christine M Disteche. Dosage compensation of the sex chromosomes and autosomes. In *Seminars in cell & developmental biology*, volume 56, pages 9–18. Elsevier, 2016.

[56] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[57] Mary E Donohoe, Susana S Silva, Stefan F Pinter, Na Xu, and Jeannie T Lee. The pluripotency factor oct4 interacts with ctcf and also controls x-chromosome pairing and counting. *Nature*, 460(7251):128–132, 2009.

[58] François Dossin and Edith Heard. The molecular and nuclear dynamics of x-chromosome inactivation. *Cold Spring Harbor Perspectives in Biology*, 14(4): a040196, 2022.

[59] François Dossin, Inês Pinheiro, Jan J Żylicz, Julia Roensch, Samuel Collombet, Agnès Le Saux, Tomasz Chelmicki, Mikaël Attia, Varun Kapoor, Ye Zhan, et al. Spen integrates transcriptional and epigenetic control of x-inactivation. *Nature*, 578(7795):455–460, 2020.

[60] Lisa Barros de Andrade e Sousa, Iris Jonkers, Laurène Syx, Ilona Dunkel, Julie Chaumeil, Christel Picard, Benjamin Foret, Chong-Jian Chen, John T Lis, Edith Heard, et al. Kinetics of xist-induced gene silencing can be predicted from combinations of epigenetic and genomic features. *Genome research*, 29(7):1087–1099, 2019.

[61] Gabriela Ecco, Marco Cassano, Annamaria Kauzlaric, Julien Duc, Andrea Coluccio, Sandra Offner, Michaël Imbeault, Helen M Rowe, Priscilla Turelli, and Didier Trono. Transposable elements and their krab-zfp controllers regulate gene expression in adult tissues. *Developmental cell*, 36(6):611–623, 2016.

[62] Jesse M Engreitz, Amy Pandya-Jones, Patrick McDonel, Alexander Shishkin, Klara Sirokman, Christine Surka, Sabah Kadri, Jeffrey Xing, Alon Goren, Eric S Lander, et al. The xist lncrna exploits three-dimensional genome architecture to spread across the x chromosome. *Science*, 341(6147):1237973, 2013.

[63] Yong Fan, Mona F Melhem, and J Richard Chaillet. Forced expression of the homeobox-containing gene pem blocks differentiation of embryonic stem cells. *Developmental biology*, 210(2):481–496, 1999.

[64] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):278, 2015.

[65] Giulia Furlan, Nancy Gutierrez Hernandez, Christophe Huret, Rafael Galupa, Joke Gerarda van Bemmel, Antonio Romito, Edith Heard, Céline Morey, and Claire Rougeulle. The ftx noncoding locus controls x chromosome inactivation independently of its rna products. *Molecular cell*, 70(3):462–472, 2018.

[66] Rafael Galupa. Lppnx lncrna: The new kid on the block or an old friend in x-inactivation choice? *Proceedings of the National Academy of Sciences*, 120(7): e2218989120, 2023.

[67] Rafael Galupa and Edith Heard. X-chromosome inactivation: a crossroads between chromosome architecture and gene regulation. *Annual review of genetics*, 52:535–566, 2018.

[68] Rafael Galupa, Elphège Pierre Nora, Rebecca Worsley-Hunt, Christel Picard, Chris Gard, Joke Gerarda van Bemmel, Nicolas Servant, Yinxiu Zhan, Fatima El Marjou, Colin Johanneau, et al. A conserved noncoding locus regulates random monoallelic xist expression across a topological boundary. *Molecular cell*, 77 (2):352–367, 2020.

[69] Michal R Gdula, Tatyana B Nesterova, Greta Pintacuda, Jonathan Godwin, Ye Zhan, Hakan Ozadam, Michael McClellan, Daniella Moralli, Felix Krueger, Catherine M Green, et al. The non-canonical smc protein smchd1 antagonises tad formation and compartmentalisation on the inactive x chromosome. *Nature communications*, 10(1):30, 2019.

[70] Marnie E Gelbart and Mitzi I Kuroda. Drosophila dosage compensation: a complex voyage to the x chromosome. 2009.

[71] Anne-Valerie Gendrel, Mikael Attia, Chong-Jian Chen, Patricia Diabangouaya, Nicolas Servant, Emmanuel Barillot, and Edith Heard. Developmental dynamics and disease potential of random monoallelic gene expression. *Developmental cell*, 28(4):366–380, 2014.

[72] Oriana Genolet, Anna A Monaco, Ilona Dunkel, Michael Boettcher, and Edda G Schulz. Identification of x-chromosomal genes that drive sex differences in embryonic stem cells through a hierarchical crispr screening approach. *Genome Biology*, 22(1):1–41, 2021.

[73] Luca Giorgetti, Rafael Galupa, Elphège P Nora, Tristan Piolot, France Lam, Job Dekker, Guido Tiana, and Edith Heard. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157(4): 950–963, 2014.

[74] Luca Giorgetti, Bryan R Lajoie, Ava C Carter, Mikael Attia, Ye Zhan, Jin Xu, Chong Jian Chen, Noam Kaplan, Howard Y Chang, Edith Heard, et al. Structural organization of the inactive x chromosome in the mouse. *Nature*, 535(7613):575–579, 2016.

[75] Rutger AF Gjaltema, Till Schwämmle, Pauline Kautz, Michael Robson, Robert Schöpflin, Liat Ravid Lustig, Lennart Brandenburg, Ilona Dunkel, Carolina Vechiatto, Evgenia Ntini, et al. Distal and proximal cis-regulatory elements sense x chromosome dosage and developmental state at the xist locus. *Molecular Cell*, 82 (1):190–208, 2022.

[76] Cristina Gontan, Eskeatnaf Mulugeta Achame, Jeroen Demmers, Tahsin Stefan Barakat, Eveline Rentmeester, Wilfred van IJcken, J Anton Grootegoed, and Joost Gribnau. Rnf12 initiates x-chromosome inactivation by targeting rex1 for degradation. *Nature*, 485(7398):386–390, 2012.

[77] Jennifer A Marshall Graves. Mammals that break the rules: genetics of marsupials and monotremes. *Annual review of genetics*, 30(1):233–260, 1996.

[78] Maxim VC Greenberg and Deborah Bourc'his. The diverse roles of dna methylation in mammalian development and disease. *Nature reviews Molecular cell biology*, pages 1–18, 2019.

[79] Vaijayanti Gupta, Michael Parisi, David Sturgill, Rachel Nuttall, Michael Doctolero, Olga K Dudko, James D Malley, P Scott Eastman, and Brian Oliver. Global analysis of x-chromosome dosage compensation. *Journal of biology*, 5(1): 1–22, 2006.

[80] Aurélia Guyochin, Sylvain Maenner, Erin Tsi-Jia Chu, Asma Hentati, Mikael Attia, Philip Avner, and Philippe Clerc. Live cell imaging of the nascent inactive x chromosome during the early differentiation process of naive es cells towards epiblast stem cells. *Plos one*, 9(12):e116109, 2014.

[81] Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845, 2016.

[82] R Scott Hansen, Theresa K Canfield, Alan D Fjeld, and Stanley M Gartler. Role of late replication timing in the silencing of x-linked genes. *Human molecular genetics*, 5(9):1345–1353, 1996.

[83] Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, et al. Cel-seq2: sensitive highly-multiplexed single-cell rna-seq. *Genome biology*, 17(1):77, 2016.

[84] Edith Heard and Christine M Disteche. Dosage compensation in mammals: fine-tuning the expression of the x chromosome. *Genes & development*, 20(14):1848–1867, 2006.

[85] Emil Heitz. *Das heterochromatin der moose*. Bornträger, 1928.

[86] Masahiro Hiasa, Jumpei Teramachi, Asuka Oda, Ryota Amachi, Takeshi Harada, Shingen Nakamura, Hirokazu Miki, Shiro Fujii, Kumiko Kagawa, Keiichiro Watanabe, et al. Pim-2 kinase is an important target of treatment for tumor progression and bone loss in myeloma. *Leukemia*, 29(1):207–217, 2015.

[87] Andreas Hierholzer, Corinne Chureau, Alessandra Liverziani, Nerea Blanes Ruiz, Bruce M Cattanach, Alexander N Young, Manish Kumar, Andrea Cerase, and Phil Avner. A long noncoding rna influences the choice of the x chromosome to be inactivated. *Proceedings of the National Academy of Sciences*, 119(28):e2118182119, 2022.

[88] Azusa Inoue, Lan Jiang, Falong Lu, and Yi Zhang. Genomic imprinting of xist by maternal h3k27me3. *Genes & development*, 31(19):1927–1932, 2017.

[89] Yuichiro Itoh, Esther Melamed, Xia Yang, Kathy Kampf, Susanna Wang, Nadir Yehya, Atila Van Nas, Kirstin Replogle, Mark R Band, David F Clayton, et al. Dosage compensation is less effective in birds than in mammals. *Journal of biology*, 6(1):1–15, 2007.

[90] Nadine S Jahchan and Kunxin Luo. Snon in mammalian development, function and diseases. *Current opinion in pharmacology*, 10(6):670–675, 2010.

[91] Yesu Jeon and Jeannie T Lee. Yy1 tethers xist rna to the inactive x nucleation center. *Cell*, 146(1):119–133, 2011.

[92] Iris Jonkers, Tahsin Stefan Barakat, Eskeatnaf Mulugeta Achame, Kim Monkhorst, Annegien Kenter, Eveline Rentmeester, Frank Grosveld, J Anton Grootegoed, and Joost Gribnau. Rnf12 is an x-encoded dose-dependent activator of x chromosome inactivation. *Cell*, 139(5):999–1011, 2009.

[93] Philippe Julien, David Brawand, Magali Soumillon, Anamaria Necsulea, Angélica Liechti, Frédéric Schütz, Tasman Daish, Frank Grützner, and Henrik Kaessmann. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS biology*, 10(5):e1001328, 2012.

[94] Edda Koina, Julie Chaumeil, Ian K Greaves, David J Tremethick, and Jennifer A Marshall Graves. Specific patterns of histone marks accompany x chromosome inactivation in a marsupial. *Chromosome Research*, 17:115–126, 2009.

[95] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Jason CH Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N Natarajan, Alex C Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, et al. Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell stem cell*, 17(4):471–485, 2015.

[96] Felix Krueger and Simon R Andrews. Snpsplit: Allele-specific splitting of alignments between genomes with known snp genotypes. *F1000Research*, 5, 2016.

[97] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. Rna velocity of single cells. *Nature*, 560 (7719):494, 2018.

[98] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. Rna velocity of single cells. *Nature*, 560 (7719):494, 2018.

[99] Anton JM Larsson, Christos Coucoravas, Rickard Sandberg, and Björn Reinius. X-chromosome upregulation is driven by increased burst frequency. *Nature structural & molecular biology*, 26(10):963–969, 2019.

[100] Jeannie Lee, Lance S Davidow, and David Warshawsky. Tsix, a gene antisense to xist at the x-inactivation centre. *Nature genetics*, 21(4):400–404, 1999.

[101] Jeannie T Lee and Rudolf Jaenisch. Long-range cis effects of ectopic x-inactivation centres on a mouse autosome. *Nature*, 386(6622):275–279, 1997.

[102] Jeannie T Lee and Naifang Lu. Targeted mutagenesis of tsix leads to nonrandom x inactivation. *Cell*, 99(1):47–57, 1999.

[103] Antonio Lentini, Huaitao Cheng, JC Noble, Natali Papanicolaou, Christos Coucoravas, Nathanael Andrews, Qiaolin Deng, Martin Enge, and Björn Reinius. Elastic dosage compensation by x-chromosome upregulation. *Nature Communications*, 13 (1):1854, 2022.

[104] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[105] Claude Libert, Lien Dejager, and Iris Pinheiro. The x chromosome in immune functions: when a chromosome makes the difference. *Nature Reviews Immunology*, 10(8):594–604, 2010.

[106] Hong Lin, Vibhor Gupta, Matthew D VerMilyea, Francesco Falciani, Jeannie T Lee, Laura P O'Neill, and Bryan M Turner. Dosage compensation in the mouse balances up-regulation and silencing of x-linked genes. *PLoS Biol*, 5(12):e326, 2007.

[107] Juan Liu, Cen Zhang, Tianliang Zhang, Chun-Yuan Chang, Jianming Wang, Ludvinna Bazile, Lanjing Zhang, Bruce G Haffty, Wenwei Hu, and Zhaohui Feng. Metabolic enzyme ldha activates rac1 gtpase as a noncanonical mechanism to promote cancer. *Nature Metabolism*, pages 1–17, 2022.

[108] Agnese Loda, Samuel Collombet, and Edith Heard. Gene regulation in time and space during x-chromosome inactivation. *Nature Reviews Molecular Cell Biology*, 23(4):231–249, 2022.

[109] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12): 550, 2014.

[110] Aaron TL Lun, Yunshun Chen, and Gordon K Smyth. It's de-licious: a recipe for differential expression analyses of rna-seq experiments using quasi-likelihood methods in edger. 2015.

[111] Mary F Lyon. Gene action in the x-chromosome of the mouse (mus musculus l.). *nature*, 190(4773):372–373, 1961.

[112] Mary F Lyon. Possible mechanisms of x chromosome inactivation. *Nature New Biology*, 232(34):229–232, 1971.

[113] Mary F Lyon et al. A further mutation of the mottled type in the house mouse. *Journal of Heredity*, 51:116–121, 1960.

[114] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[115] Shantha K Mahadevaiah, Helene Royo, John L VandeBerg, John R McCarrey, Sarah Mackay, and James MA Turner. Key features of the x inactivation process are conserved between marsupials and eutherians. *Current Biology*, 19(17):1478–1484, 2009.

[116] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[117] Yolanda Markaki, Johnny Gan Chong, Yuying Wang, Elsie C Jacobson, Christy Luong, Shawn YX Tan, Joanna W Jachowicz, Mackenzie Strehle, Davide

Maestrini, Abhik K Banerjee, et al. Xist nucleates local protein gradients to propagate silencing across the x chromosome. *Cell*, 184(25):6174–6192, 2021.

[118] Hendrik Marks, Tüzer Kalkan, Roberta Menafra, Sergey Denissov, Kenneth Jones, Helmut Hofemeister, Jennifer Nichols, Andrea Kranz, A Francis Stewart, Austin Smith, et al. The transcriptional and epigenomic foundations of ground state pluripotency. *Cell*, 149(3):590–604, 2012.

[119] Hendrik Marks, Hindrik HD Kerstens, Tahsin Stefan Barakat, Erik Splinter, René AM Dirks, Guido van Mierlo, Onkar Joshi, Shuang-Yin Wang, Tomas Babak, Cornelis A Albers, et al. Dynamics of gene silencing during x inactivation using allele-specific rna-seq. *Genome biology*, 16(1):1–20, 2015.

[120] Graziano Martello and Austin Smith. The nature of embryonic stem cells. *Annual review of cell and developmental biology*, 30, 2014.

[121] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[122] Hisham Mohammed, Irene Hernando-Herraez, Aurora Savino, Antonio Scialdone, Iain Macaulay, Carla Mulas, Tamir Chandra, Thierry Voet, Wendy Dean, Jennifer Nichols, et al. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell reports*, 20(5):1215–1228, 2017.

[123] Patrizia Mondello, Salvatore Cuzzocrea, and Michael Mian. Pim kinases in hematological malignancies: where are we now and where are we going? *Journal of hematology & oncology*, 7(1):95, 2014.

[124] Kim Monkhorst, Iris Jonkers, Eveline Rentmeester, Frank Grosveld, and Joost Gribnau. X inactivation counting and choice is a stochastic process: evidence for involvement of an x-linked activator. *Cell*, 132(3):410–421, 2008.

[125] Céline Morey, Danielle Arnaud, Philip Avner, and Philippe Clerc. Tsix-mediated repression of xist accumulation is not sufficient for normal random x inactivation. *Human molecular genetics*, 10(13):1403–1411, 2001.

[126] Hermann Joseph Muller. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1): 2–9, 1964.

[127] Verena Mutzel and Edda G Schulz. Dosage sensing, threshold responses, and epigenetic memory: a systems biology perspective on random x-chromosome inactivation. *Bioessays*, 42(4):1900163, 2020.

[128] Verena Mutzel, Ikuhiro Okamoto, Ilona Dunkel, Mitinori Saitou, Luca Giorgetti, Edith Heard, and Edda G Schulz. A symmetric toggle switch explains the onset

of random x inactivation in different mammals. *Nature structural & molecular biology*, 26(5):350–360, 2019.

[129] Pablo Navarro, Sylvain Pichard, Constance Ciaudo, Philip Avner, and Claire Rougeulle. Tsix transcription across the xist gene alters chromatin conformation without affecting xist transcription: implications for x-chromosome inactivation. *Genes & development*, 19(12):1474–1484, 2005.

[130] Pablo Navarro, Ian Chambers, Violetta Karwacki-Neisius, Corinne Chureau, Céline Morey, Claire Rougeulle, and Philip Avner. Molecular coupling of xist regulation and pluripotency. *Science*, 321(5896):1693–1695, 2008.

[131] Pablo Navarro, Andrew Oldfield, Julie Legoupi, Nicola Festuccia, Agnes Dubois, Mikael Attia, Jon Schoorlemmer, Claire Rougeulle, Ian Chambers, and Philip Avner. Molecular coupling of tsix regulation and pluripotency. *Nature*, 468(7322): 457–460, 2010.

[132] Tatyana B Nesterova, Sergey Ya Slobodyanyuk, Eugene A Elisaphenko, Alexander I Shevchenko, Colette Johnston, Marina E Pavlova, Igor B Rogozin, Nikolay N Kolesnikov, Neil Brockdorff, and Suren M Zakian. Characterization of the genomic xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome research*, 11(5):833–849, 2001.

[133] Tatyana B Nesterova, Claire E Senner, Janina Schneider, Tilly Alcayna-Stevens, Anna Tattermusch, Myriam Hemberger, and Neil Brockdorff. Pluripotency factor binding and tsix expression act synergistically to repress xist in undifferentiated embryonic stem cells. *Epigenetics & chromatin*, 4(1):1–10, 2011.

[134] Tatyana B Nesterova, Guifeng Wei, Heather Coker, Greta Pintacuda, Joseph S Bowness, Tianyi Zhang, Mafalda Almeida, Bianca Bloechl, Benoit Moindrot, Emma J Carter, et al. Systematic allelic analysis defines the interplay of key pathways in x chromosome inactivation. *Nature communications*, 10(1):3129, 2019.

[135] Di Kim Nguyen and Christine M Disteche. Dosage compensation of the active x chromosome in mammals. *Nature genetics*, 38(1):47–53, 2006.

[136] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L Van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, 2012.

[137] Yuya Ogawa and Jeannie T Lee. Xite, x-inactivation intergenic transcription elements that regulate the probability of choice. *Molecular cell*, 11(3):731–743, 2003.

[138] Tatsuya Ohhata, Yuko Hoki, Hiroyuki Sasaki, and Takashi Sado. Crucial role of antisense transcription across the xist promoter in tsix-mediated xist chromatin modification. 2008.

[139] S Ohno, WD Kaplan, and R Kinosita. The centromeric and nucleolus-associated heterochromatin of rattus norvegicus. *Experimental cell research*, 16(2):348–357, 1959.

[140] Susumu Ohno. *Sex chromosomes and sex-linked genes*, volume 1. Springer Science & Business Media, 2013.

[141] Susumu Ohno and TS Hauschka. Allocycly of the x-chromosome in tumors and normal tissues. *Cancer Research*, 20(4):541–545, 1960.

[142] Ikuhiro Okamoto, Catherine Patrat, Dominique Thépot, Nathalie Peynot, Patricia Fauque, Nathalie Daniel, Patricia Diabangouaya, Jean-Philippe Wolf, Jean-Paul Renard, Véronique Duranthon, et al. Eutherian mammals use diverse strategies to initiate x-chromosome inactivation during development. *Nature*, 472(7343): 370–374, 2011.

[143] Keisuke Okita, Tomoko Ichisaka, and Shinya Yamanaka. Generation of germline-competent induced pluripotent stem cells. *nature*, 448(7151):313–317, 2007.

[144] Amy Pandya-Jones, Yolanda Markaki, Jacques Serizay, Tsotne Chitiashvili, Walter R Mancia Leon, Andrey Damianov, Constantinos Chronis, Bernadett Papp, Chun-Kan Chen, Robin McKee, et al. A protein assembly mediates xist localization and gene silencing. *Nature*, 587(7832):145–151, 2020.

[145] Bernhard Payer, Michael Rosenberg, Masashi Yamaji, Yukihiro Yabuta, Michiyo Koyanagi-Aoi, Katsuhiko Hayashi, Shinya Yamanaka, Mitinori Saitou, and Jeannie T Lee. Tsix rna and the germline factor, prdm14, link x reactivation and stem cell reprogramming. *Molecular cell*, 52(6):805–818, 2013.

[146] Graeme D Penny, Graham F Kay, Steven A Sheardown, Sohaila Rastan, and Neil Brockdorff. Requirement for xist in x chromosome inactivation. *Nature*, 379(6561): 131–137, 1996.

[147] Sophie Petropoulos, Daniel Edsgärd, Björn Reinius, Qiaolin Deng, Sarita Pauliina Panula, Simone Codeluppi, Alvaro Plaza Reyes, Sten Linnarsson, Rickard Sandberg, and Fredrik Lanner. Single-cell rna-seq reveals lineage and x chromosome dynamics in human preimplantation embryos. *Cell*, 165(4):1012–1026, 2016.

[148] Greta Pintacuda, Guifeng Wei, Chloë Roustan, Burcu Anil Kirmizitas, Nicolae Solcan, Andrea Cerase, Alfredo Castello, Shabaz Mohammed, Benoît Moindrot, Tatyana B Nesterova, et al. hnrnpk recruits pcgf3/5-prc1 to the xist rna b-repeat to establish polycomb-mediated chromosomal silencing. *Molecular cell*, 68(5):955–969, 2017.

[149] Robert M Plenge, Ivona Percec, Joseph H Nadeau, and Huntington F Willard. Expression-based assay of an x-linked gene to examine effects of the x-controlling element (xce) locus. *Mammalian Genome*, 11(5):405, 2000.

[150] Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. Single-cell mrna quantification and differential analysis with census. *Nature methods*, 14(3):309, 2017.

[151] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, 14(10):979, 2017.

[152] Sohaila Rastan. Non-random x-chromosome inactivation in mouse x-autosome translocation embryos—location of the inactivation centre. 1983.

[153] Sohaila Rastan and Elizabeth J Robertson. X-chromosome deletions in embryo-derived (ek) cell lines associated with lack of x-chromosome inactivation. 1985.

[154] Liat Ravid Lustig, Abhishek Sampath Kumar, Till Schwämmle, Ilona Dunkel, Gemma Noviello, Elodie Limberg, Raha Weigert, Guido Pacini, René Buschow, Afrah Ghauri, et al. Gata transcription factors drive initial xist upregulation after fertilization through direct activation of long-range enhancers. *Nature Cell Biology*, pages 1–12, 2023.

[155] Björn Reinius, Chengxi Shi, Liu Hengshuo, Kuljeet Singh Sandhu, Katarzyna J Radomska, Glenn D Rosen, Lu Lu, Klas Kullander, Robert W Williams, and Elena Jazin. Female-biased expression of long non-coding rnas in domains that escape x-inactivation in mouse. *BMC genomics*, 11(1):1–16, 2010.

[156] Björn Reinius, Jeff E Mold, Daniel Ramsköld, Qiaolin Deng, Per Johnsson, Jakob Michaëlsson, Jonas Frisén, and Rickard Sandberg. Analysis of allelic expression patterns in clonal somatic cells by single-cell rna–seq. *Nature genetics*, 48(11): 1430–1435, 2016.

[157] Rebeca Ridings-Figueroa, Emma R Stewart, Tatyana B Nesterova, Heather Coker, Greta Pintacuda, Jonathan Godwin, Rose Wilson, Aidan Haslam, Fred Lilley, Renate Ruigrok, et al. The nuclear matrix protein ciz1 facilitates localization of xist rna to the inactive x-chromosome territory. *Genes & development*, 31(9): 876–888, 2017.

[158] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[159] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25, 2010.

[160] Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

[161] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[162] Lisa Rodermund, Heather Coker, Roel Oldenkamp, Guifeng Wei, Joseph Bowness, Bramman Rajkumar, Tatyana Nesterova, David Miguel Susano Pinto, Lothar Schermelleh, and Neil Brockdorff. Time-resolved structured illumination microscopy reveals key principles of xist rna spreading. *Science*, 372(6547):eabe7500, 2021.

[163] Clara Roidor, Laurene Syx, Emmanuelle Beyne, Dina Zielinski, Aurelie Teissandier, Caroline Lee, Marius Walter, Nicolas Servant, Karim Chebli, Deborah Bourc'his, et al. Spatio-temporal x-linked gene reactivation and site-specific retention of epigenetic silencing in the in vivo germline. *bioRxiv*, pages 2023–04, 2023.

[164] Olga Rosspopoff, Christophe Huret, Amanda J Collier, Miguel Casanova, Peter J Rugg-Gunn, Jean-François Ouimette, and Claire Rougeulle. Mechanistic diversification of xist regulatory network in mammals. *bioRxiv*, page 689430, 2019.

[165] Takashi Sado, Martin H Fenner, Seong-Seng Tan, Patrick Tam, Toshihiro Shioda, and En Li. X inactivation in the mouse embryo deficient for dnmt1: distinct effect of hypomethylation on imprinted and random x inactivation. *Developmental biology*, 225(2):294–303, 2000.

[166] Takashi Sado, Yuko Hoki, and Hiroyuki Sasaki. Tsix silences xist through modification of chromatin structure. *Developmental cell*, 9(1):159–165, 2005.

[167] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5): 547, 2019.

[168] Anna Sahakyan, Rachel Kim, Constantinos Chronis, Shan Sabri, Giancarlo Bonora, Thorold W Theunissen, Edward Kuoy, Justin Langerman, Amander T Clark, Rudolf Jaenisch, et al. Human naive pluripotent stem cells model x chromosome dampening and x inactivation. *Cell stem cell*, 20(1):87–101, 2017.

[169] Takehisa Sakaguchi, Masazumi Nishimoto, Satoru Miyagi, Atsushi Iwama, Yohei Morita, Naoki Iwamori, Hiromitsu Nakauchi, Hiroshi Kiyonari, Masami Muramatsu, and Akihiko Okuda. Putative "stemness" gene jam-b is not required for maintenance of stem cell state in embryonic, neural, or hematopoietic stem cells. *Molecular and cellular biology*, 26(17):6557–6570, 2006.

[170] Jesús M Salvador, Joshua D Brown-Clay, and Albert J Fornace. Gadd45 in stress signaling, cell cycle control, and apoptosis. In *Gadd45 Stress Sensor Genes*, pages 1–19. Springer, 2013.

[171] Edda G Schulz, Johannes Meisig, Tomonori Nakamura, Ikuhiro Okamoto, Anja Sieber, Christel Picard, Maud Borensztein, Mitinori Saitou, Nils Blüthgen, and Edith Heard. The two active x chromosomes in female escs block exit from the pluripotent state by modulating the esc signaling network. *Cell stem cell*, 14(2): 203–216, 2014.

[172] Till Schwämmle and Edda G Schulz. Regulatory principles and mechanisms governing the onset of random x-chromosome inactivation. *Current Opinion in Genetics & Development*, 81:102063, 2023.

[173] JongDae Shin, Michael Bossenz, Young Chung, Hong Ma, Meg Byron, Naoko Taniguchi-Ishigaki, Xiaochun Zhu, Baowei Jiao, Lisa L Hall, Michael R Green, et al. Maternal rnf12/rlim is required for imprinted x-chromosome inactivation in mice. *Nature*, 467(7318):977–981, 2010.

[174] JongDae Shin, Mary C Wallingford, Judith Gallant, Chelsea Marcho, Baowei Jiao, Meg Byron, Michael Bossenz, Jeanne B Lawrence, Stephen N Jones, Jesse Mager, et al. Rlim is dispensable for x-chromosome inactivation in the mouse embryonic epiblast. *Nature*, 511(7507):86–89, 2014.

[175] Marie-Christine Simmler, Bruce M Cattanach, Carol Rasberry, Claire Rougeulle, and Phil Avner. Mapping the murine xce locus with (ca) n repeats. *Mammalian Genome*, 4:523–530, 1993.

[176] Matthew D Simon, Stefan F Pinter, Rui Fang, Kavitha Sarma, Michael Rutenberg-Schoenberg, Sarah K Bowman, Barry A Kesner, Verena K Maier, Robert E Kingston, and Jeannie T Lee. High-resolution xist binding maps reveal two-step spreading during x-chromosome inactivation. *Nature*, 504(7480):465–469, 2013.

[177] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, Article 3, 2004.

[178] Miki Soma, Yoshitaka Fujihara, Masaru Okabe, Fumitoshi Ishino, and Shin Kobayashi. Ftx is dispensable for imprinted x-chromosome inactivation in preimplantation mouse embryos. *Scientific reports*, 4(1):1–6, 2014.

[179] Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, 15(4):255, 2018.

[180] Juan Song, Adrian Janiszewski, Natalie De Geest, Lotte Vanheer, Irene Talon, Mouna El Bakkali, Taeho Oh, and Vincent Pasque. X-chromosome dosage modulates multiple molecular and cellular properties of mouse pluripotent stem cells

independently of global dna methylation levels. *Stem cell reports*, 12(2):333–350, 2019.

[181] Elsa J Sousa, Hannah T Stuart, Lawrence E Bates, Mohammadmersad Ghorbani, Jennifer Nichols, Sabine Dietmann, and Jose CR Silva. Exit from naive pluripotency induces a transient x chromosome inactivation-like state in males. *Cell stem cell*, 22(6):919–928, 2018.

[182] Wolfram Stacklies, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. pcamethods—a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, 2007.

[183] Nicholas Stavropoulos, Rebecca K Rowntree, and Jeannie T Lee. Identification of developmentally specific enhancers for tsix in the regulation of x chromosome inactivation. *Molecular and cellular biology*, 25(7):2757–2769, 2005.

[184] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):1–16, 2018.

[185] Bryan K Sun, Aimée M Deaton, and Jeannie T Lee. A transient heterochromatic state in xist preempts x inactivation choice without rna stabilization. *Molecular cell*, 21(5):617–628, 2006.

[186] Sha Sun, Brian C Del Rosario, Attila Szanto, Yuya Ogawa, Yesu Jeon, and Jeannie T Lee. Jpx rna activates xist by evicting ctcf. *Cell*, 153(7):1537–1551, 2013.

[187] Hongjae Sunwoo, David Colognori, John E Froberg, Yesu Jeon, and Jeannie T Lee. Repeat e anchors xist rna to the inactive x chromosomal compartment through cdkn1a-interacting protein (ciz1). *Proceedings of the National Academy of Sciences*, 114(40):10654–10659, 2017.

[188] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126 (4):663–676, 2006.

[189] Fuchou Tang, Catalin Barbacioru, Ellen Nordman, Bin Li, Nanlan Xu, Vladimir I Bashkirov, Kaiqin Lao, and M Azim Surani. Rna-seq analysis to capture the transcriptome landscape of a single cell. *Nature protocols*, 5(3):516–535, 2010.

[190] Di Tian, Sha Sun, and Jeannie T Lee. The long noncoding rna, jpx, is a molecular switch for x chromosome inactivation. *Cell*, 143(3):390–403, 2010.

[191] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381, 2014.

[192] Taru Tukiainen, Alexandra-Chloé Villani, Angela Yen, Manuel A Rivas, Jamie L Marshall, Rahul Satija, Matt Aguirre, Laura Gauthier, Mark Fleharty, Andrew Kirby, et al. Landscape of x chromosome inactivation across human tissues. *Nature*, 550(7675):244–248, 2017.

[193] Céline Vallot, Catherine Patrat, Amanda J Collier, Christophe Huret, Miguel Casanova, Tharvesh M Liyakat Ali, Matteo Tosolini, Nelly Frydman, Edith Heard, Peter J Rugg-Gunn, et al. Xact noncoding rna competes with xist in the control of x chromosome activity during human early development. *Cell stem cell*, 20(1): 102–111, 2017.

[194] Guido van Mierlo, Gert Jan C Veenstra, Michiel Vermeulen, and Hendrik Marks. The complexity of prc2 subcomplexes. *Trends in cell biology*, 29(8):660–671, 2019.

[195] Sébastien Vigneau, Sandrine Augui, Pablo Navarro, Philip Avner, and Philippe Clerc. An essential role for the dxpas34 tandem repeat and tsix transcription in the counting process of x chromosome inactivation. *Proceedings of the National Academy of Sciences*, 103(19):7390–7395, 2006.

[196] Kay-Dietrich Wagner, Nicole Wagner, and Andreas Schedl. The complex life of wt1. *Journal of cell science*, 116(9):1653–1658, 2003.

[197] Feng Wang, JongDae Shin, Jeremy M Shea, Jun Yu, Ana Bošković, Meg Byron, Xiaochun Zhu, Alex K Shalek, Aviv Regev, Jeanne B Lawrence, et al. Regulation of x-linked gene expression during early mouse development by rlim. *Elife*, 5: e19127, 2016.

[198] Anton Wutz and Rudolf Jaenisch. A shift from reversible to irreversible x inactivation is triggered during es cell differentiation. *Molecular cell*, 5(4):695–705, 2000.

[199] Anton Wutz, Theodore P Rasmussen, and Rudolf Jaenisch. Chromosomal silencing and localization are mediated by different domains of xist rna. *Nature genetics*, 30(2):167–174, 2002.

[200] Yuanyan Xiong, Xiaoshu Chen, Zhidong Chen, Xunzhang Wang, Suhua Shi, Xueqin Wang, Jianzhi Zhang, and Xionglei He. Rna sequencing shows no dosage compensation of the active x-chromosome. *Nature genetics*, 42(12):1043–1047, 2010.

[201] Ziny C Yen, Irmtraud M Meyer, Sanja Karalic, and Carolyn J Brown. A cross-species comparison of x-chromosome inactivation in eutheria. *Genomics*, 90(4): 453–463, 2007.

[202] Xiaofei Zhang, Juan Zhang, Tao Wang, Miguel A Esteban, and Duanqing Pei. Esrrb activates oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. *Journal of Biological Chemistry*, 283(51):35825–35833, 2008.

[203] Qing Zhou, Taifu Wang, Lizhi Leng, Wei Zheng, Jinrong Huang, Fang Fang, Ling Yang, Fang Chen, Ge Lin, Wen-Jing Wang, et al. Single-cell rna-seq reveals distinct dynamic behavior of sex chromosomes during early human embryogenesis. *Molecular Reproduction and Development*, 86(7):871–882, 2019.

[204] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643, 2017.

[205] Ilona Zvetkova, Anwyn Apedaile, Bernard Ramsahoye, Jacqueline E Mermoud, Lucy A Crompton, Rosalind John, Robert Feil, and Neil Brockdorff. Global hypomethylation of the genome in xx embryonic stem cells. *Nature genetics*, 37(11): 1274–1279, 2005.

[206] Jan Jakub Żylicz, Aurélie Bousard, Kristina Žumer, Francois Dossin, Eusra Mohammad, Simão Teixeira da Rocha, Björn Schwalb, Laurène Syx, Florent Dingli, Damarys Loew, et al. The implication of early chromatin changes in x chromosome inactivation. *Cell*, 176(1-2):182–197, 2019.