USING INTERPRETABLE MACHINE LEARNING TO UNDERSTAND GENE SILENCING DYNAMICS DURING X-CHROMOSOME INACTIVATION.

Lisa Corina Barros de Andrade e Sousa

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

Berlin, Januar 2020

Erstgutachter: **Prof. Dr. Annalisa Marsico**
Zweitgutachter: **Prof. Dr. Marcel H. Schulz**

Tag der Disputation: 18.12.2020

All we have to decide is what to do with the time that is given to us.

— J. R. R. Tolkien

PREFACE

PUBLICATION & CONTRIBUTIONS

The work discussed in this thesis grew from a collaboration project between the Labs of Edda Schulz, Edith Heard, John Lis and Annalisa Marsico. Edith Heard and John Lis had the idea to quantify gene silencing dynamics on the X chromosome with an allele-specific PRO-seq time course experiment. The cell line used for the time course experiment was provided by Edda Schulz. The experiments (PRO-seq, mRNA-seq and pyrosequencing time course) were conducted in the labs of Jon Lis (Iris Jonkers), Edith Heard (Julie Chaumeil, Christel Picard) and Edda Schulz (Ilona Dunkel). Laurene Syx pre-processed the allele-specific expression data. Discussions between Annalisa Marsico and Edda Schulz led to the idea to systematically analyse the influence of epigenetic and genomic factors on X chromosomal gene silencing dynamics with machine learning methods. Edda Schulz came up with the idea of computing silencing half-times from the time course data and Annalisa Marsico started to collect epigenetic and genomic data sets that could be used for a machine learning model.

I would like to thank Annalisa Marsico for involving me in this project and for her contribution to the model design through many fruitful discussions. Furthermore, I would like to thank Annalisa Marsico, Edda Schulz and Edith Heard for their contribution to the interpretation of the obtained results. The results were published in Genome Research (Barros de Andrade E Sousa et al., 2019) with the help of Edda Schulz and Annalisa Marsico, who largely contributed to writing the paper, and Iris Jonkers, Edith Heard and John Lis, who reviewed and edited the paper.

ACKNOWLEDGEMENTS

First of all I would like to thank my supervisor Annalisa Marsico for her scientific guidance, support and many interesting discussions throughout my time at the Max Planck Institute. I appreciate a lot that she put me in charge of this interesting project, which ended up being the main topic of my PhD thesis. Furthermore, I would like to thank her for giving me the chance to attend many interesting conferences, symposia and summer schools and for encouraging me to present my projects at international conferences. One last thing that shall not be forgotten are the amazing christmas dinners with the best italian food to which we were invited every year.

I would also like to thank Edda Schulz for her scientific support during my time as a PhD and for commenting on the thesis manuscript. I would further like to thank the IMPRS coordinators Kirsten Kelleher and Fabian Feutlinske for their help and support during my PhD and for the organisation of many interesting events. I was very happy to be part of the International Max Planck Research School for Computational Biology and Scientific Computing through which I

was able to attend many interesting courses and workshops.

I wish to thank all my colleagues of the RNA Bioinformatics group, especially my office mates Sabrina Krakau, Roman Schulte-Sasse and Stefan Budach for all the interesting discussions and laughter we shared. Special thanks to Roman Schulte-Sasse for commenting on my thesis manuscript and the ability to listen and type on a key board at the same time. Many thanks also go to Lam-Ha Ly and Virginie Stanislas, who made my lunch breaks about so much more than eating food together. Last but not least, I would like to thank Sara Hetzel for being the best first intern and for always having time for a chat. It wouldn't have been the same without you guys.

I would also like to thank my parents, who supported me in every situation and who put up with me during christmas when I was in the last weeks of writing my thesis. Last but not least, and most importantly, I would like to thank Sebastian Thieme, who stood by my side no matter what happened and who continuously supported me, especially during the time of writing my thesis. There are no words that can express my gratitude. Thanks for being amazing!

CONTENTS

1 INTRODUCTION

The genomes of females and males differ substantially with respect to their pair of sex determining chromosomes: females have two X chromosomes while males have one X and one Y chromosome. The process of X-Chromosome Inactivation (XCI) is a dosage compensation mechanism in mammalian females that ensures equal levels of X-linked gene expression between XX females and XY males by transcriptional inactivation and heterochromatinization of either the paternal or maternal X chromosome during early embryonic development. In placental mammals, the master regulator of the XCI process is the long non-coding RNA *Xist* , which is responsible for gene silencing and conversion of the entire X chromosome into silent heterochromatin. Once *Xist* is upregulated in a monoallelic fashion, the *Xist* RNA starts to spread along the future inactive X chromosome in *cis*, thereby creating an *Xist* RNA domain from which the transcription machinery is excluded. X-linked genes that are being silenced are recruited from the periphery of the X chromosome into the *Xist* RNA domain. Those early events are followed by the loss of active and recruitment of repressive chromatin marks, the spatial reorganization of the X chromosomal architecture as well as the repositioning of the X chromosome inside the nucleus. Once XCI is established in female somatic cells, the silenced state is stably propagated through the clonal cell population, leading to a mosaic of clonal groups of cells, where either the paternal or maternal X chromosome is silenced (see Figure 1.1)

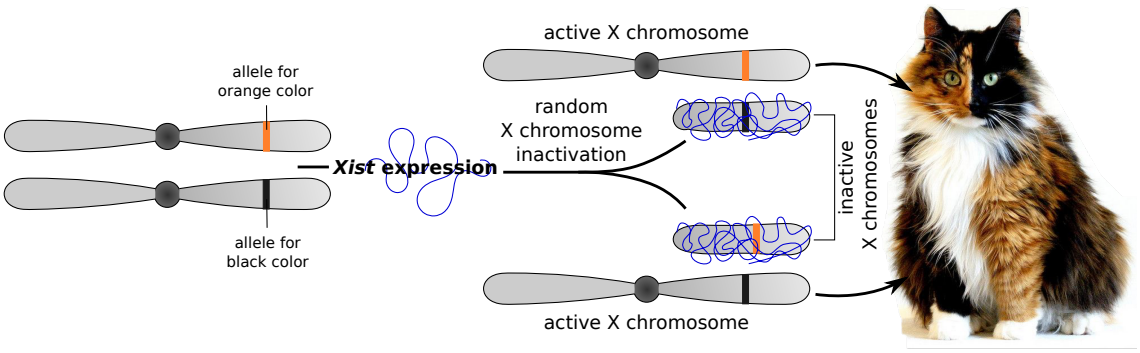

Figure 1.1: X chromosome inactivation in female calico cats. X chromosome inactivation causes the black and orange fur coloring of female calico cats. In those cats, the color coding gene is located on the X chromosome, where the allele for the orange color is located in the one and the allele for the black color on the other X chromosome. Since X chromosome inactivation is a random process, an orange and black color mosaic is produced. The colors tend to occur in patches, because the silencing state is propagated through the clonal cell population and sister cells tend to remain close together during later stages of development. In contrast, male calico cats are either solid orange or solid black, depending on the X chromosome that is inherited from the mother. The picture of the calico cat is adapted from `tah-heetch.com`.

The dynamics of *Xist*-mediated silencing are highly variable across genes, with some genes being silenced early, while others being silenced later during the silencing process. A small fraction

of X chromosomal genes is even able to escape this silencing process and remains active on both X chromosomes in female somatic cells. The underlying mechanisms that define the silencing kinetics of X-linked genes remain poorly understood. While a variety of epigenetic and genomic factors have been implicated in controlling gene-specific silencing dynamics, none of these factors alone can predict whether and to what extent a gene will be silenced upon XCI, and the associations with measured silencing efficiencies is generally weak. This in turn suggests that a gene's susceptibility to *Xist*-mediated silencing is potentially controlled by a complex combination of different silencing factors. Since most *in vivo* and *in vitro* studies analyze the influence of a silencing factor in isolation from other silencing factors, very little is known about the interplay of different factors and their relative contribution to the overall silencing dynamics. Few *in silico* studies attempted to predict gene silencing on the X chromosome by integrating DNA or chromatin features into a machine learning (ML) model in order to identify important silencing determinants. However, those studies either focus on only a specific set of silencing factors (e.g. factors previously shown to be related to *Xist*-mediated silencing) and / or investigate only a specific subset of X-linked genes, making them less generalizable to the silencing dynamics of all X-linked genes. Hence, defining the combinatorial feature patterns that underlie the differential susceptibility to XCI remains an open but important question, particularly as genes that are not fully silenced are implicated in diseases, such as autoimmune syndromes and tumorigenesis.

The main goal of this thesis was to investigate the interplay of different silencing factors in order to uncover different *Xist*-mediated silencing pathways. We expected to find different silencing pathways, because the different functional domains of *Xist* (repeat-A to -F) recruit different silencing complexes (e.g. PRC 1 and SPEN/HDAC3) and susceptibility to each silencing pathway might be determined by distinct feature patterns. To put the different pieces of the XCI puzzle together, we set out to identify silencing determinants in an unbiased and combinatorial manner based on chromosome-wide measured gene silencing dynamics. Therefore, we measured gene silencing dynamics at the level of the nascent transcriptome using allele-specific Precision nuclear Run-On sequencing (PRO-seq) and collected a large number of epigenetic and genomic as well as primary DNA sequence factors, including factors that were identified as silencing determinants before but also factors that were not yet associated to *Xist*-mediated silencing. To predict the silencing susceptibility of X chromosomal genes, we used a non-linear ML model - a Random Forest (RF) model - in order to account for the potential combinatorial nature of the silencing factors. We specifically chose the RF algorithm, because it has a reduced risk of overfitting, even when features correlate and the training set is small with high class imbalances - all properties that were present in our data set. Classical feature importance helped us to identify the most important features in the model, i.e. the main determinants of gene silencing, which included known but also unknown silencing factors. To identify the combinations of epigenetic and genomic factors, which predispose X-linked genes to be silenced efficiently or escape XCI, we had to go beyond classical features importance. We had to extract the combinatorial feature patterns from the RF model that arise from non-linear relationships within the data. We solved this problem by implementing a forest-guided clustering approach that stratifies the data points (X-linked genes) into subgroups according to different combinations of features (silencing factors). Thereby we were able to identify two silencing pathways that might be associated with the different silencing efficiencies of X chromosomal genes.

Thesis outline

Following this Chapter, a more detailed introduction into the biological background is given in Chapter 2. It gives a broad overview on gene regulation in general, including genome-wide experiments for gene regulation studies, as well as a more detailed overview on gene regulation during the process of XCI. Chapter 3 gives an introduction to the core ML concepts and a more detailed description of the main ML algorithms used throughout this thesis. Chapter 4 reviews different approaches to study XCI *in vivo* or *in vitro* and the computational approaches that attempt to predict the gene silencing status from different features.

The training process and interpretation of the our model, which is used to identify the most important gene silencing determinants and combination of those, is described in Chapter 5, 6 and 7. Chapter 5 focuses on the experimental quantification of chromosome-wide gene silencing dynamics and on the computation of gene-specific silencing half-times from allele-specific PRO-seq time course data. Chapter 6 describes the training process of our ML models based on the computed silencing half-times and the collected epigenetic and genomic as well as primary DNA sequence features. Furthermore, it describes the validation of one ML model with three different strategies. The interpretation of the trained ML models is described in Chapter 7. Here, a classical feature importance approach is used to identify the main determinants of *Xist*-mediated gene silencing. In addition, the idea of the forest-guided clustering approach is introduced and applied to the trained models to extract the complex combinatorial rules underlying gene silencing during XCI. The conclusions of this thesis and potential open questions are discussed in Chapter 8.

2 BIOLOGICAL BACKGROUND

The first part of this chapter gives a general introduction to molecular biology with a focus on the topic of gene regulation and provides an overview on the experimental techniques that can be used to explore the different areas of this topic. The second part of this chapter gives an overview on the current knowledge in the field of X chromosome inactivation.

2.1 INTRODUCTION TO GENE REGULATION

All living organisms on our planet are made of the same basic functional unit - the cell. Most living organisms are single cells but even the most complex organisms, like us humans who consist of more than 10^{13} cells, originate from one single cell. The cells in a multicellular organism are able to develop into different cell types, each performing a specialized function in the organisms. A few essential types of molecules are required for each cell to perform its function: ***DNA***, the building block of every genome; ***RNA***, a copy of a certain genomic region that either has its own functionality or serves as an intermediate between DNA and proteins; ***proteins***, the molecules that regulate the different processes in a cell. Hence, although most living organisms have very distinct phenotypes, they paradoxically share the same basic building blocks and mechanisms.

The following sections are based on the book Molecular Biology of *The Cell* (Alberts et al., 2014) and give an introduction to the basics of molecular biology, with a focus on the aspect of gene regulation. Therefore, the first section gives a short introduction to the genome, followed by an overview of transcription and its regulation in the second and third section, respectively.

2.1.1 Genome

A major breakthrough in molecular biology was the discovery of the DNA structure in the early 1950s by James D. Watson and Francis H.C. Crick (Watson et al., 1953). Each DNA molecule consists of two long complementary DNA strands. Those strands are chains of **nucleotides**, where each nucleotide is composed of a sugar (deoxyribose) backbone and a base that can be either adenine (A), cytosine (C), guanine (G) or thymine (T). The nucleotides of each DNA strand are covalently linked together by a phosphate group between the 3' carbon atom of the one nucleotide and the 5' carbon atom of the adjacent nucleotide, giving the DNA strand a certain orientation with a downstream (5' to 3') and an upstream (3' to 5') direction. Both DNA strands are twisted around each other and held together by hydrogen bonds between complementary bases where A and T bind with two hydrogen bonds and G and T bind with three hydrogen bonds, giving the DNA its characteristic three-dimensional double helix structure (Figure 2.1).

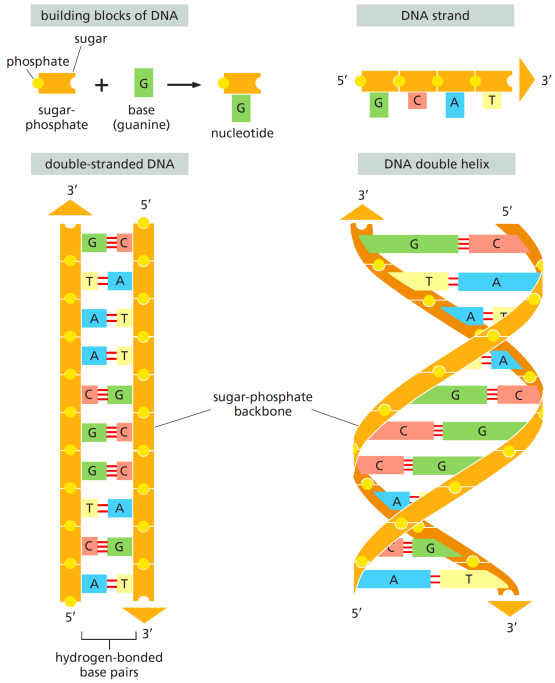

Figure 2.1: DNA and its building blocks. Each DNA strand is built from four types of nucleotides that are covalently linked by a phosphate group. Each nucleotide consist of two parts: a sugar-phosphate backbone and a base (A, C, G, and T). Two DNA strands are held together via hydrogen bonds between complementary bases, resulting in the characteristic double helix, where both DNA strands are twisted around each other. Reprinted from (Alberts et al., 2014).

The set of DNA molecules in a cell, also called the ***genome*** of a cell, contains all the genetic information necessary to build and maintain a functioning cell. This information is inherited to the next generation by a mechanisms called DNA replication that creates two identical copies of DNA from one original DNA molecule. Due to errors in the replication process or other types of DNA damage, the genome differs slightly between individuals of the same species and more drastically between different species giving rise to the huge diversity of species on our planet but also to the slight phenotypic differences between individuals that make everyone of us unique.

2.1.2 Transcription

The genome contains all the genetic information necessary to produce the different molecules that are required for proper cell functionality. Since the information is stored in very long DNA molecules, it would be inefficient if the whole DNA sequence was read each time a specific molecule is needed. The central dogma of molecular biology, a fundamental principle that applies to all cells in every living organism, states that every molecule originates from a specific region in the genome, referred to as ***gene***. Those genes are segments of short or long DNA sequences that are read independently from each other, enabling the cell to produce only the required molecules. The final product of a gene can either be a protein or a **noncoding RNA (ncRNA)** , which is a functional RNA that performs regulatory or catalytic functions in the cell but lacks protein coding capacity. The process of converting the genetic information encoded in each gene into a functioning protein is called gene expression. Gene expression can be divided into two major steps: transcription of an RNA intermediate from DNA, called messenger RNA (mRNA), and translation of the RNA intermediate into a functional protein (Figure 2.2). The transcription process is mediated by a protein complex called RNA polymerase that synthesizes

RNA molecules by reading the corresponding gene from the 5' to 3' end. While prokaryotes (cellular organisms without cell nucleus) only have one type of RNA polymerase, eukaryotes (cellular organisms with cell nucleus) have three different types of RNA polymerase. Here, the focus is on the transcription of mRNAs and ncRNAs in eukaryotes by RNA Polymerase II (RNAPII).

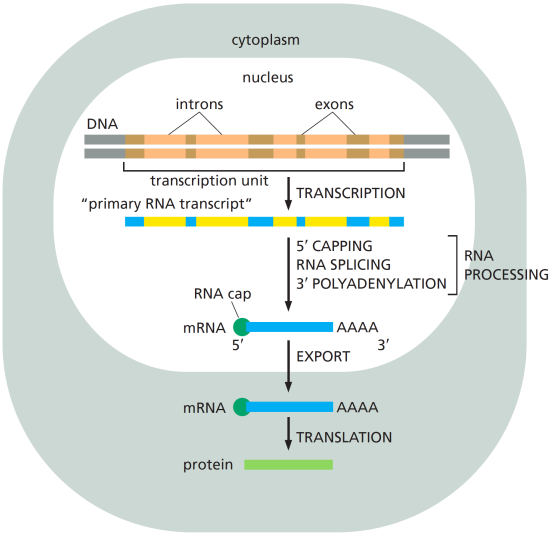

Figure 2.2: The central dogma of molecular biology. Each protein is produced in a two step process from a certain functional genomic region, called gene. The gene is first transcribed into a precursor mRNA, which is then processed into a mature mRNA and exported into the cytoplasm where it is translated into a functional protein. Reprinted from (Alberts et al., 2014).

Transcription can be seen as a three-step process with an initiation, elongation and termination step. During the initiation step specific initiation factors, called ***general transcription factors***, recruit the RNAPII to the ***transcription start site*** **(TSS)**. The TSS indicates the starting point for the RNA synthesis and is located at 5' end of the gene. The transition from initiation into elongation requires the exchange of co-factors and certain conformational changes in the RNAPII molecule, which are associated with its C-Terminal Domain (CTD). The CTD is a RNAPII specific domain that does not exist in RNAP I or RNAP III and consists of multiple heptapeptide repeats with a consensus amino acid sequence Tyrosine - Serine - Proline - Threonine - Serine - Proline - Serine. A phosphate group can be added to each of those residues. During the initiation step the CTD is unphosphorylated. Phosphorylation of Serine5 and Serine7 residues helps the RNAPII to disengage from the cluster of general transcription factors and leave the TSS to move along the DNA. While sliding along the DNA, the RNAPII unwinds small portions of the double helix and uses the gene as a template to synthesize an RNA that is a reverse complement to the gene itself. During the elongation step, the CTD serves as scaffold for several elongation factors, which help the RNAPII to move along the gene body without dissociating from the DNA before it reaches the end of the transcribed gene. Before entering the termination step, the CTD is phosphorylated at the Serine2 and the phosphate groups at Serine5 and Serine7 are gradually removed. Once the RNAPII reaches the 3' end of the gene, called the transcription termination site, it cleaves the RNA transcript and dissociates from the DNA. The transcribed RNA can either be a ncRNA or a precursor mRNA, which is later translated into a protein.

Precursor mRNAs are modified at both ends (addition of a 5' cap and a 3' poly-A tail) during the transcription process. This mechanism allows the cell to access if the precursor mRNA is intact before transforming it into a mature mRNA through a process called splicing. Splicing

is a process, where certain non-coding parts of the RNA (introns) are removed and the coding parts (exons) are merged together. The mature mRNAs are then exported to the cytoplasm and translated into proteins. ncRNAs on the other hand are not translated into proteins but are able to fold into complex three-dimensional structures, giving rise to their specific regulatory and catalytic functions. The class of ncRNAs can be divided into two major groups based on the length of the RNA: short ncRNAs (sncRNAs) with a length < 200 nucleotides and long noncoding RNAs (lncRNAs) with a length > 200 nucleotides. sncRNAs include functional RNAs, involved in the processes of transcription and translation (e.g. t-RNAs, r-RNAs and small nuclear RNAs), and regulatory RNAs, involved in regulation of gene expression (e.g. micro RNAs, small interfering RNAs and interacting RNAs). lncRNAs on the other hand are a not well defined group of large, heterogeneous ncRNAs involved in regulation of gene expression.

2.1.3 Transcriptional Regulation

Multicellular organisms are composed of many different cell types such as liver, brain or blood cells, all with very distinct phenotypes. Since all those cells are built from the same genetic information, which is read by the same basic mechanisms, the only way to achieve phenotypic differences is through regulation of gene expression. However, the genetic information encoded in the DNA is not actively altered to control gene expression. Instead, gene expression is regulated on a transcriptional and post-transcriptional level or through degradation of no longer needed molecules.

Transcriptional regulation is a process that controls the transcriptional rate of each gene, allowing the cell to produce high amounts of certain RNAs and low amounts of other RNAs according to its needs. Transcriptional regulation is primarily modulated by so-called *cis*-regulatory elements, which are genomic regions that can be bound by different transcriptional regulators. For instance, general transcription factors, which help to recruit RNAPII, assemble at a specific *cis*-regulatory element called gene ***promoter***, a genomic region that is located right upstream of the gene's TSS. Other important *cis*-regulatory elements are ***enhancers***, which are also bound by transcription regulators and can interact with gene promoters. Transcription regulators do not only comprise general transcription factors, which exclusively assemble at the gene promoter, but also ***sequence-specific transcription factors*** that bind to promoter and enhancer regions in a sequence-specific manner. Those regulators usually assemble in groups through recognition of specific DNA sequences within the *cis*-regulatory element. Once bound to the DNA they either act as activators through positive regulation or as repressors through negative regulation of transcription.

Another factor that influences gene transcription is the 3D architecture of the DNA, which affects the accessibility of *cis*-regulatory regions for transcription regulators. Since the DNA double helix is a very long molecule (in human about 2m in length), it needs to be compacted into higher order structures in order to fit into the nucleus. The needed degree of ***compaction*** is achieved by a complex of histone proteins and DNA, called chromatin, whose basic repeating unit is the ***nucleosome***. Each nucleosome is formed by 147 nucleotides of double stranded DNA that is wound around an octamer of histone proteins, consisting of two copies of each core histone

(H2A, H2B, H3 and H4). Those core histones are highly conserved across eukaryotic organisms, showing their fundamental role in organizing and compacting the DNA inside the nucleus. The chromatin itself is further folded into a series of loops and coils, which allows *cis*-regulatory elements to overcome large genomic distances and get into spatial proximity to interact with each other (e.g. enhancers and their interacting promoter can be located far away on a linear scale) . Different segments of folded DNA are further organized into spatially separated domains, so-called ***topologically associating domains (TADs)*** (Dixon et al., 2012). TADs generally help to create spatial proximity between certain *cis*-regulatory elements but also prevent deleterious interaction between other *cis*-regulatory elements by placing them in separate TADs, thereby creating spatial distance between them (Kim et al., 2011). TADs are formed with the help of ***structural proteins***, like CTCF and cohesin, that regulate the spatial folding and 3D structure of the chromatin (Seitan et al., 2013; Zuin et al., 2014). The highest-order structure of compacted DNA is called the ***chromosome*** (Figure 2.3). Each eukaryotic cell divides its DNA into several chromosomes. Humans, for instance, have 22 pairs of autosomes (homologous chromosomes, with one copy being inherited from the mother and the other one from the father) and two additional sex chromosomes, where females have a homologous pair of X chromosomes and males one X and one Y chromosome.

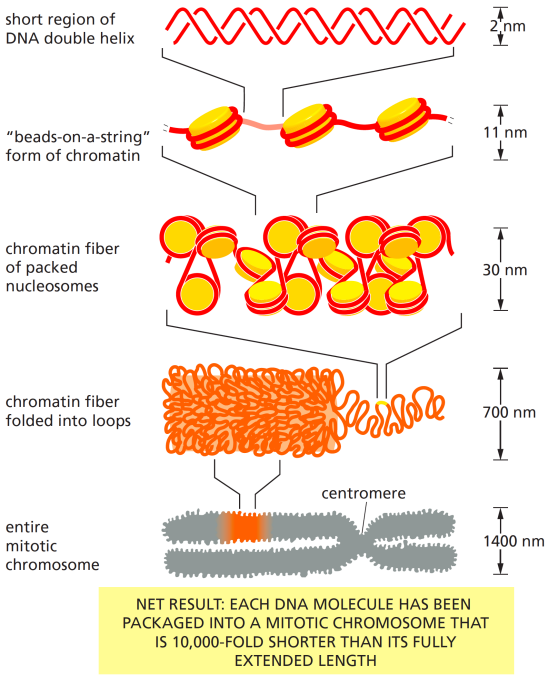

Figure 2.3: Levels of chromatin compaction. A heterodimer of histone complexes and 147 nucleotides of DNA double helix forms the basic compaction unit of the chromatin. The chromatin is further folded into loops and coils to achieve the maximal degree of compaction of the DNA in form of highly compacted chromosomes. Reprinted from (Alberts et al., 2014).

The static genetic information, encoded in the DNA, can be interpreted on a dynamic level through different levels of chromatin compaction. Packing DNA into higher-order structures is a quite dynamic process, where parts of the genome can be packed more loosely in order to make the DNA accessible or compacted to make the DNA inaccessible for transcription regulators. The level of compaction is usually described by two states: heterochromatin, the highly condensed form, and euchromatin, the less condensed form of chromatin. Euchromatin is usually located at the center of the cell nucleus and harbours many actively transcribed genes, because *cis*-regulatory elements

are made accessible for the transcription machinery. If euchromatic regions are converted into heterochromatin, the genes within that region are usually switched off. Heterochromatin contains only few, mainly inactive, genes and is located in proximity to the nuclear membrane, sometimes attached to the nuclear lamina, which is a fibrillar network, associated with the nuclear membrane and composed of lamins and lamin-associated proteins (Dechat et al., 2010). The sites where the chromatin attaches to the nuclear lamina are called **Lamina-associated Domains (LADs)** (Guelen et al., 2008).

Local conversion of the chromatin state can be achieved through covalent modifications of the DNA or histone proteins. This additional layer of information is known as the ***epigenome***. Epigenetic mechanisms include for instance the enzymatic modification of histone tails. Each core histone has an N-terminal amino acid tail, which is subject to different types of covalent modifications, such as acetylation of lysines or mono-, di- and trimethylation of lysines as well as phosphorylation of serines. The different types of ***histone modifications*** are associated with different parts of the gene and can be related to different chromatin states, dividing them into active, elongation and repression marks. H3K4me3 as well as histone acetylation are active marks that are found at accessible promoter regions. H3K27me3 and H3K9me3, on the other hand, are repressive marks found in heterochromatic regions where gene expression is inactivated (Kim et al., 2012). H3K36me3 and H3K79me2 are strongly associated with transcription elongation and are deposited along the gene body by RNAPII-bound complexes during the transcription of the target gene. Deposition of elongation marks is thought to prevent spurious transcription behind the moving RNAPII at the opened DNA strand. Covalent histone marks are frequently removed or added to the histone tails depending on the chromatin state. Those changes are induced by specific protein complexes, so called ***chromatin remodelling complexes***, that can read, catalyze or remove histone modification. Acetyl groups, for example, are added by a set of histone acetyl transferases (HATs) and removed by different histonedeacetylase complexes (HDACs). Methyl groups, on the other hand, are removed by a set of histone demethylases and added by different histone methyltransferases. For instance, the enzyme EZH2, which is part of the Polycomb Repressive Complex 2 (PRC2), catalyzes the histone modification H3K27me3 (Cao et al., 2002; Czermin et al., 2002). Another epigenetic mechanism is the covalent modification of the DNA that does not change the DNA sequence itself. ***DNA methylation*** describes the methylation of a cytosine nucleotide in a GC context. Clusters of GC dinucleotides with a minimum length of 200 nucleotide pairs and a GC content higher than 50% are called ***CpG islands*** and are typically found at gene promoters (Gardiner-Garden et al., 1987). DNA methylation of gene promoter GC dinucleotides can lead to impaired binding of transcription factors or to binding of methyl-CpG-binding domain proteins (MBDs), which recruit chromatin remodelling complexes that alter the histone modification pattern around the promoter region to form highly compacted heterochromatin. Hence, DNA methylation is predominantly found at highly compacted chromatin regions and regulates gene expression through repression of gene transcription at promoter level. In addition, DNA methylation provides a mechanism through which gene expression patterns can be stably inherited to daughter cells because GC sequences are base-paired exactly to the same sequence on the opposite strand. Hence, the parental strand can serve as a methylation template during DNA replication leading to a direct inheritance of the DNA methylation pattern to progeny cells.

Interestingly, the product of transcription, the RNA itself can function as a transcriptional regulator as well. For instance, RNAs belonging to the class of lncRNAs are able to regulate gene transcription through various mechanisms: recruitment of regulatory protein complexes to the gene promoter or enhancer; inhibition of transcription factor binding; competing transcription (due to colliding polymerase) of a lncRNA located on the opposite strand of the target gene; or re-organization of the chromatin structure through recruitment of chromatin remodelling complexes (Long et al., 2017). One of the most prominent examples for transcriptional regulation through lncRNAs is the process of X chromosome inactivation in which one copy of the two X chromosomes in female somatic cells is entirely inactivated to achieve gene dosage compensation between females and males, which only have one copy of the X chromosome. The process of X chromosome inactivation shows the power of transcriptional regulation within a cell: inactivation of more than a thousand genes on one of the two essentially identical X chromosomes that are located in the same nucleus and are exposed to the same transcription regulators. The topic of X chromosome inactivation will be further elucidated in Section 2.2.

2.1.4 Experimental Methods in Gene Regulation Studies

The previous section introduced the principles of gene transcription and its various regulatory mechanisms. The different aspects of transcription and its regulation can be explored with different experimental protocols. The basis for most protocols is a technique called high throughput sequencing, where short segments of DNA or RNA of interest are sequenced. More precisely, the DNA segment of interest is sheared into small fragments of uniform size by sonication or enzymes and then amplified by a factor of ~ 1000 to generate more signal for the following sequencing step. The sequencing step generates millions of so-called sequencing reads and computational methods are used to determine their original position in the genome in a process called mapping (Ambardar et al., 2016). The following section gives an overview on how high throughput sequencing is used to quantify gene expression and to explore the regulatory mechanisms of gene transcription.

Techniques to Measure Gene Expression

RNA-seq. RNA sequencing (RNA-seq) provides a snapshot of the transcriptome in a biological sample (e.g. a cell) at a given time point by measuring the abundance of all RNA transcripts and their isoforms in that cell (Mortazavi et al., 2008). In a first step, the RNA of interest is isolated from the biological sample and reverse transcribed into cDNA, which are then sheared into fragments of uniform size. Adapters are ligated to the 3' and 5' ends of the cDNA fragments to create a sequencing library, which is then used for high-throughput sequencing. In a last step, the sequencing reads are mapped to a reference genome. The RNA-seq experiment can be adapted to capture only certain parts of the transcriptome. For instance, the coding transcriptome can be measured by a targeted RNA-Seq protocol, called mRNA-seq, that enriches for RNA transcripts with a 3' poly-A tail, an attribute of RNAs that are translated into proteins.

PRO-seq. Precision nuclear Run-On sequencing (PRO-seq) is an experimental method that maps the location of active RNA polymerases at base-pair resolution in a genome-wide manner, providing a snapshot of active gene transcription at a given time point in the cell (Kwak et al., 2013). In

the first step of the protocol, biotin-labeled nucleotide triphosphates (biotin-NTPs) are carried out in isolated nuclei to be incorporated into the 3' end of nascent RNAs by transcriptionally engaged RNAPII. The biotin-labeled nascent RNAs are isolated, fragmented and reverse transcribed into cDNA to create a sequencing library. Next, adapters are ligated to the 3' and 5' end of the cDNA fragments before being sequenced from the 3' end via high-throughput sequencing. The sequencing reads are then mapped to a reference genome.

Pyrosequencing. Pyrosequencing is a DNA sequencing method based on the sequencing by synthesis principle that detects light emitted due to pyrophosphate release when nucleotides are incorporated by DNA polymerase (Ronaghi et al., 1998). In a first step, the isolated DNA is fragmented and amplified. Added nucleotides are then incorporated by DNA polymerase at the 3' end of the DNA fragments, thereby releasing pyrophosphate that is converted to adenosine triphosphate (ATP) by ATP sulfurylase. ATP then catalyzes the conversion of luciferin to oxyluciferin, a process that emits light with an intensity proportional to the amount of consumed ATP. The light intensity emitted by this process is determined by a light detector and can be used to infer the number and type of incorporated nucleotides.

Protocols to Study Chromatin Modifications

ChIP-seq. Chromatin Immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is an experimental method used for genome-wide profiling of DNA-binding proteins like transcription factors (Johnson et al., 2007; Robertson et al., 2007) as well as histone modifications (Barski et al., 2007). In the first step of the protocol, covalent bonds are established between DNA and proteins by crosslinking the protein to the chromatin, typically with formaldehyde treatment. Next, the chromatin is sheared into small fragments through sonication and the DNA fragments bound by a protein are co-immunoprecipitated via a protein-specific antibody to isolate the DNA fragments with the protein of interest. The cross-link between DNA and protein is then reversed and the purified DNA fragments are prepared for sequencing (DNA amplification) and sequenced from the 5' end using high-throughput sequencing. The obtained sequencing reads are then mapped to the reference genome.

Protocols to Study DNA Modifications

DNA methylation. Several experimental protocols are available to detect the proportion of methylated cytosines in the genome. A common technique to analyse DNA methylation patterns is Whole Genome Bisulfite Sequencing (WGBS) (Cokus et al., 2008). High-throughput sequencing techniques cannot distinguish between methylated and unmethylated cytosines because the DNA methylation pattern is erased during PCR amplification. To overcome this problem, an intermediate step is introduced where the genomic DNA is treated with sodium bisulfite, converting unmethylated cytosines into uracils while methylated cytosines are protected from bisulfite-induced conversion. Unmethylated cytosines appear as thymines after sequencing because the DNA polymerase reads uracils as thymines during the amplification of the bisulfite treated DNA. The methylation status of a CpG site is calculated as the proportion of reads with a

thymine where the reference sequence has a cytosine for a given CpG site (commonly referred to as beta value). A disadvantage of the WGBS is the difficulty of discriminating between the classical 5mC-modification and 5hmC or 5fC. Enrichment-based methods are able to distinguish between the different cytosine modifications though immunoprecipitation of methylated DNA, using for example DNA-methylation-specific antibodies (MeDIP). However, the resolution of enrichment-based techniques is much lower compared to WGBS.

Protocols to Study Chromatin Structure

Chromosome conformation capture. Chromosome conformation capture (3C) is an experimental technique that investigates the 3D chromatin structure through measured interactions between genomic loci (Dekker et al., 2002). The loci of those fragments can be very distant from each other on a linear genomic scale but get into spatial proximity through DNA looping. Many techniques have evolved from 3C, e.g. 4C, 5C, HiC or HiCap, but the basis for the different protocols is the same. Crosslinking agents are used to preserve the 3D structure of the chromatin within the nucleus from which the DNA is isolated and sheared into DNA fragments. DNA fragments that lie in close spatial proximity are ligated to capture regions of interacting DNA. The ligated DNA is amplified and sequenced via high-throughput sequencing. The sequencing reads are mapped to a reference genome to identify the two interacting regions and calculate their interaction frequency. The interaction frequencies between different loci identify genomic regions that frequently interact with each other and regions that are in spatial distance to each other, enabling the reconstruction of the 3D chromatin architecture. While 3C only detects interactions for a specific region of interest, HiC detects interactions for the entire genome (Lieberman-Aiden et al., 2009). HiCap is similar to HiC but focuses on interactions where one of the two interacting regions is a promoter (Sahlén et al., 2015).

2.2 INTRODUCTION TO X-CHROMOSOME INACTIVATION

All females are beautiful mosaics, we just don't have the fur to show it.

— Edith Heard

Mammalian females have two X chromosomes (XX) while males have one X and one Y chromosome (XY). The pair of sex-determining chromosomes is derived from an ancestral pair of autosomes, which lost homology during the mammalian evolution. The Y chromosome lost more than 97% whereas the X chromosome lost less than 5% of its ancestral genes. Hence, the Y chromosome is small and harbours less than 100 genes while X chromosome is large and harbours more than 1000 genes, generating a strong dosage imbalance of X-linked gene expression between XX females, containing two gene copies, and XY males, containing only one gene copy of X-linked genes. Double dosage of X-linked genes is lethal probably due to an increased dosage of genes that are responsible for cell function and differentiation (Takagi et al., 1990). Hence, mammalian females have developed a dosage compensation mechanism that generates an inactive X chromosome (Xi) and maintains an active X chromosome (Xa) during early female development. This process is called ***X-Chromosome Inactivation (XCI)*** and equalizes X-linked

gene expression by inactivating one of the two X chromosomes through transcriptional silencing and heterochromatinization.

The concept of XCI was first introduced in 1961 by Mary Lyon, who proposed that either the paternal or maternal X chromosome becomes inactivated during early embryonic development and that the inactive state is stably inherited through cell divisions (Lyon, 1961). In 1991, Brown et al. discovered the master regulator of the XCI process - the ***X inactive specific transcript (Xist)*** - a long non-coding RNA that is responsible for gene silencing and heterochromatinization (Brown et al., 1991). This discovery was a milestone in answering the questions of how the XCI process is regulated but also raised new questions concerning the mechanisms that allow *Xist* to induce chromosome-wide gene silencing and heterochromatinization.

In the past 50 years, the process of XCI has been studied extensively, typically in mouse model systems, more specifically in female mouse embryonic stem cells (mESCs). Female mESCs, that are derived from the inner cell mass of the blastocyst, harbour two active X chromosomes and recapitulate the early stages of XCI upon induced *in vitro* differentiation quite well (Rastan et al., 1985). In mouse, two waves of XCI are observed during early female development (Figure 2.4A, Figure 2.4B). The first wave of XCI is called ***imprinted XCI*** and starts shortly after fertilization during the 2- to 8-cell stage of embryonic development, where all cells selectively inactivate the paternal X chromosome. Imprinted XCI is maintained during pre-implantation embryogenesis up to the blastocyst stage. The imprint is retained in the cells of the trophectoderm and primitive endoderm, which will form the extraembryonic tissues, while cells of the inner cell mass (ICM), which that will form pluripotent epiblast cells, reactivate the paternal X chromosome. A second wave of XCI, called ***random XCI***, occurs after implantation in pluripotent epiblast cells. When epiblast cells enter differentiation, random XCI is rapidly triggered and leads to random silencing of either the paternal or maternal X chromosome. Once established, random XCI is stably maintained in somatic cells. Since the epiblast cells give rise to the embryo proper, the embryo proper will develop as a mosaic of cells with either a silent paternal or maternal X chromosome. Certain key steps in the process of XCI have been identified over past years: spreading of *Xist* RNA along the X chromosome in *cis*, exclusion of the transcription machinery, recruitment of repressive chromatin marks, spatial reorganization of the X chromosomal architecture and repositioning of the X chromosome inside the nucleus. Once established, random XCI is stably inherited through clonal cell propagation.

The importance of understanding the mechanisms behind XCI is depicted by the multitude of genetic disorders or X-linked disease that are caused by an aberrant number of X chromosomes or an impaired X inactivation process. A disrupted XCI process, for example through loss of *Xist* expression, has been implicated in tumorigenesis (Yang et al., 2018). In female cancers, like breast and ovarian cancer, *Xist* expression is lost and low *Xist* expression highly correlates with advanced tumor stages (Kobayashi et al., 2016; Zheng et al., 2018). However, in male-specific cancers, like testicular cancer, *Xist* expression is upregulated, showing the diverse roles of the master regulator *Xist* in tumorigenesis (Kawakami et al., 2003). X-linked genetic disorders are caused by an aberrant number of X chromosomes like the Turner's syndrome in females where one X chromosome (XO) is missing, or the Klinefelter syndrome in males where one or more additional X chromosomes (XXY) are present. Both karyotypes show severe phenotypic variations form the regular XX

females or XY males phenotypes, although in Klinefelter males the additional X chromosome is inactivated and in regular females only one X chromosome is active (Tüttelmann et al., 2010). This shows that even slight changes in the dosage of certain genes caused by an additional or missing inactive X chromosome can have a high impact on biological systems. Hence, it is important to understand the underlying mechanisms that drive the process of XCI, to develop new treatments for X-linked disease and genetic disorders. Unfortunately, our understanding of the XCI process is far from being complete, although various aspects of the underlying mechanisms have been described in many studies.

Since most of our knowledge about XCI and its mechanisms comes from mouse models and the analysis in this thesis are based on a mouse model experiment, this introduction on XCI focusses on XCI in mouse, more specifically on random XCI in mouse. XCI in human is similar to XCI in mouse but has some substantial differences. However, the details and differences of XCI in human won't be issued here. In the subsequent chapters,the current knowledge on the process of XCI is described, starting with the X-inactivation center, where the master regulator *Xist* is expressed from, followed by current knowledge on how *Xist* is able to spread and then silence the genes on the X chromosome. Finally, the changes in the 3D architecture of the X chromosome during XCI and how certain genes are able to escape the process of XCI are described. The main reference for those chapters is (Galupa et al., 2018), if not stated otherwise.

2.2.1 *Xist* and the X-Inactivation Center

One of the first discovered long noncoding RNAs (lncRNAs) in mammals is the X inactive specific transcript (Xist) whose gene is located on the X chromosome (Borsani et al., 1991; Brockdorff et al., 1991). The *Xist* gene produces a 17 kilobase long lncRNA, which is capped, spliced, and polyadenylated but remains localized within the nucleus, closely associated with the X chromosome it is expressed from (Brown et al., 1992). The monoallelic upregulation of *Xist* is highly correlated with the onset of the XCI process and has been shown to be essential for the process of XCI (Marahrens et al., 1997; Penny et al., 1996). Hence, *Xist* is seen as the master regulator of the XCI process. After being expressed, the *Xist* RNA spreads along the future inactive X chromosome (Xi) in *cis* and silences the genes on Xi through the recruitment of multiple proteins involved in transcriptional silencing. The *Xist* gene is conserved across species, but unique to placental mammals (Grant et al., 2012). Some of the highly conserved parts of the *Xist* gene are comprised of tandem repeated sequence elements, named ***repeats A-F***, and have been shown to bind RNA-binding proteinss (RBPs) (Figure 2.5) (Brockdorff, 2018; Brown et al., 1992; Chu et al., 2015; Wutz et al., 2002). One of the most important and best studied repeat element is the repeat-A, which is crucial for the gene silencing function of the *Xist* RNA by interacting with RBPs like SPEN and RBM15 (for further information head to Section 2.2.3) (McHugh et al., 2015; Moindrot et al., 2015; Monfort et al., 2015). The repeat-B and -C elements recruit the Polycomb Repressive Complexes (PRC) 1 and 2 through binding of the RBP hnRNPK (Brockdorff, 2017; Pintacuda et al., 2017), while repeat-E was shown to bind the RBP CIZ1, which is important for the localization of *Xist* on the Xi (Ridings-Figueroa et al., 2017; Sunwoo et al., 2017). The function of repeat-D and -F are still unknown. Overall, it seems like *Xist* is acting as a scaffold to recruit important RBPs

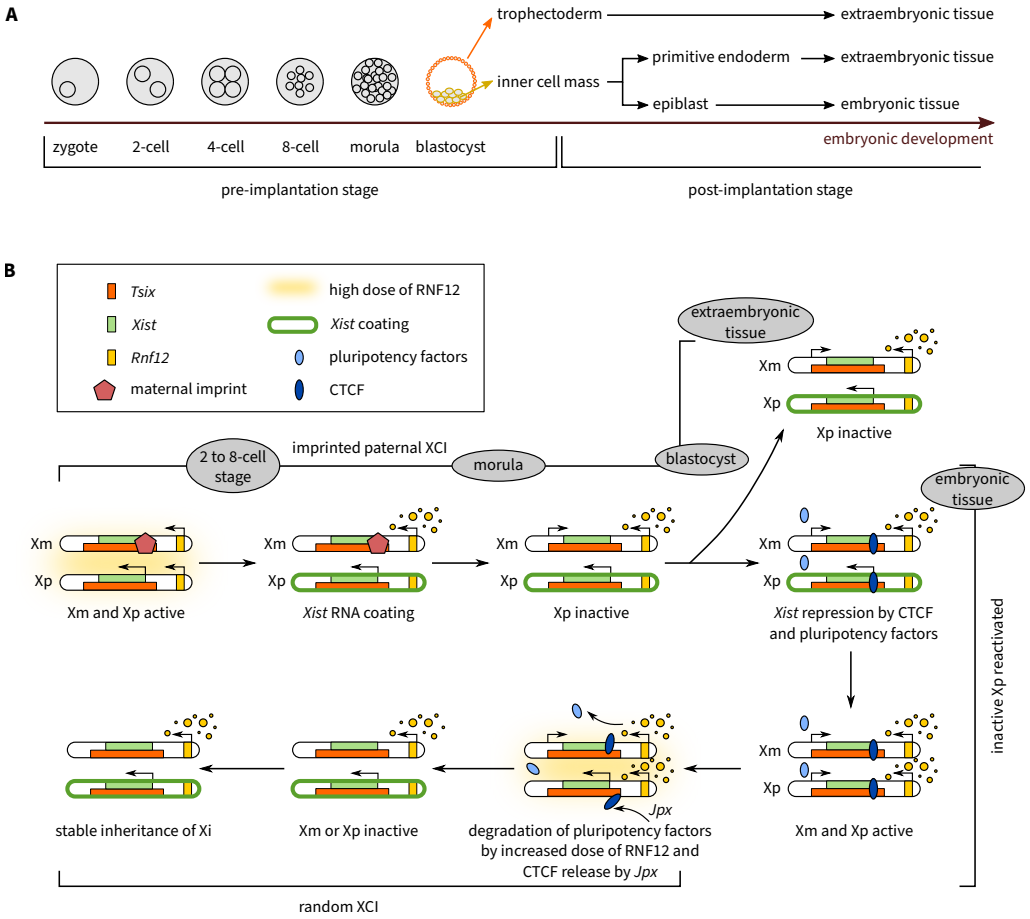

Figure 2.4: The two waves of XCI during early female development in mouse. **(A)** The pre-implantation stages of early embryonic development are characterized by a relatively synchronous doubling of cell until the 8-cell stage. After the 8-cell stage the embryo undergoes a process known as compaction to become a morula. The cells on the outside of the morula differentiate into the trophectoderm, while the cells inside the morula become the inner cell mass (ICM), resulting in the formation of the blastocyst. The epiblast and primitive endoderm cells are derived during the second differentiation event from the ICM. Trophectoderm and primitive endoderm cells form the extraembryonic tissue (e.g. placenta), while the epilast give rise to the embryo proper. **(B)** In mouse, two waves of XCI are known. During embryonic pre-implantation development of female mice, starting at the 2-cell stage, the paternal X chromosome undergoes global silencing associated with the establishment of imprinted X chromosome inactivation. The paternal imprint is propagated to the extraembryonic lineages of trophectoderm and primitive endoderm, while the epiblast cells reactivate the inactive X chromosome and a random X chromosome inactivation occurs *de novo* after implantation in the epiblast cells, the progenitor cells of the embryo proper. Inspired by (Augui et al., 2011).

during the XCI process.

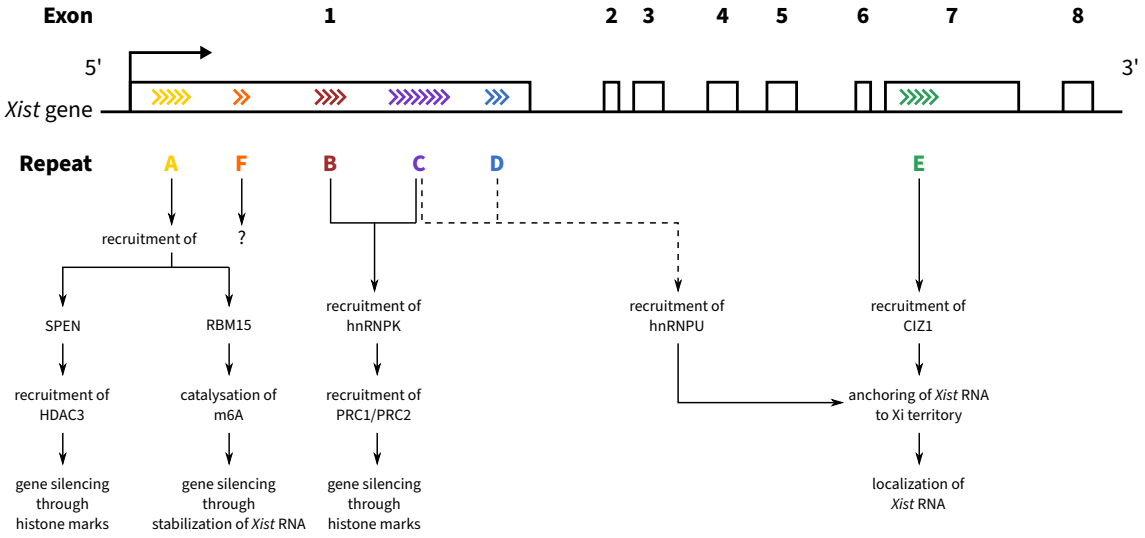

Figure 2.5: Repetitive elements of the *Xist* gene and their proposed functions in XCI. The repeat elements A-F were shown to be important for the localization and silencing function of *Xist* through the recruitment of RBPs. While Repeat-A, -B and -C are important for the silencing function, Repeat-E is involved in the localization of *Xist* on the Xi. Inspired by (Brockdorff, 2018; Gendrel et al., 2014).

The *Xist* gene is located on the X chromosome, more specifically in a region called the ***X-inactivation center (Xic)***. The Xic is defined as the minimal genetic region, containing all required *cis*-acting elements, for initiation of the XCI process if at least two chromosomes are present, which carry this defined region. This suggests that the two copies of the Xic influence each other in *trans*, while triggering XCI in *cis* (Augui et al., 2011). The Xic is organized into two topologically associating domain (TAD): the *Xist* TAD and the *Tsix* TAD (Figure 2.5-2) (Nora et al., 2012). TADs are domains with increased 3D interactions between loci within the respective domain. The two TADs separate the Xic into two domains of positive and negative gene regulation potentially facilitating the interaction between repressive and activating regulatory elements of *Xist* (Tsai et al., 2008). The *Xist* promoter is located in the *Xist* TAD but the complete *Xist* gene spans both TADs. While the *Xist* TAD harbours most of its known activators (e.g. *Rnf12* or *Jpx*), the *Tsix* TAD includes most of the known *Xist* repressors (e.g. *Tsix* or *Xite*) (Table 2.1, Table 2.2). The major known repressor of *Xist* is its antisense transcript ***Tsix*** that spans the complete *Xist* gene and represses *Xist* upregulation in *cis* by being transcribed across the *Xist* promoter (Luikenhuis et al., 2001; Stavropoulos et al., 2001). The *Tsix* TAD also harbours Xite (X-inactivation intergenic transcription element), an enhancer of the *Tsix* promoter, which lies upstream of the *Tsix* transcription start site (TSS) and is regulated by different pluripotency factors (Ogawa et al., 2003; Stavropoulos et al., 2005). Pluripotency factors are transcription factors and epigenetic regulators (e.g. OCT4, SOX2, NANOG KLF4 or REX1) that hold embryonic stem cells in the pluripotent state through repression of genes that are required for differentiation. A key activator of *Xist* is the *trans*-acting protein RNF12, which has a ubiquitin ligase activity that degrades *Xist* repressors REX1 and whose gene lies within the *Xist* TAD (Gontan et al., 2012; Jonkers et al., 2009). It was shown that RNF12 has to be transcribed from both X chromosomes to reach a critical dosage that is necessary to activate *Xist* on one of the two X chromosomes. Once that threshold is reached, RNF12 will be inactivated in *cis*, lowering RNF12 levels, which prevents

the onset of *Xist* on both X chromosomes. However, deletion of *Rnf12* on one X chromosome does not prevent XCI, suggesting that additional, probably redundant, *Xist* activation mechanisms exist (Monkhorst et al., 2008).

Table 2.1: Genes within the *Xist* TAD.

Locus	coding potential	functional during XCI
Xist	lncRNA	master regulator of XCI • coats Xi in *cis* • triggers gene silencing on Xi • triggers chromatin remodelling and structural reorganization of Xi
Jpx	lncRNA	*Xist* activator, acts in *cis* or *trans* by binding to *Xist* repressor CTCF
Ftx	lncRNA	*Xist* activator, acts in *cis*
X-pairing region (Xpr)	protein-coding	not implicated in XCI
Rnf12 also called: *Rlim*	protein-coding	*Xist* activator, acts in *trans* by targeting REX1 for degradation

Table 2.2: Genes within the *Tsix* TAD.

Locus	coding potential	functional during XCI
Nap1/2	protein-coding	not implicated in XCI
Linx	lncRNA	influences structure of Tsix TAD
Cdx4	protein-coding	not implicated in XCI
Chic1	protein-coding	unknown
Tsx	protein-coding & ncRNA	unknown
Xite	ncRNA	*Tsix* activator, acts in *cis* by being an enhancer for *Tsix* promoter
Tsix	lncRNA	*Xist* repressor, acts in *cis* by being transcribed across the *Xist* promoter

2.2.2 *Xist* localization to the inactive X chromosome

The first step in the process of XCI is the counting step where the cell determines if and how many X chromosomes should be inactivated. Still very little is known about the exact molecular

mechanisms of the counting step but it was shown that both X chromosomes become transiently colocalized before *Xist* expression is upregulated. This colocalization step is mediated by the X-pairing region, which lies upstream of the *Xist* locus within the Xic. It was proposed that the physical proximity of both Xic helps the cell to count the number of chromosomes by sensing the second Xic (Augui et al., 2011). However, a recent study showed that neither *Xic* pairing nor nuclear lamina localization influences choice-making or monoallelic *Xist* upregulation (Pollex et al., 2019). Another recent study proposed that the cooperation of a *cis*-acting repressor (e.g. Tsix) and a *trans*-acting activator (e.g. RNF12 or Jpx) is sufficient for the mono-allelic upregulation of *Xist* . In this model, a double dosage of the *trans*-acting activator is required to overcome the repression of the *cis*-acting repressor, a mechnism that would prevent *Xist* upregulation if only one X chromosome is present in the cell (Mutzel et al., 2019).

The initiation of the XCI process is marked by the monoallelic upregulation of *Xist* and down-regulation of pluripotency factors, which are thought to negatively regulate *Xist* expression by promoting *Tsix* transcription. The *trans*-acting factor RNF12 targets pluripotency factors like REX1 for degradation, leading to the downregulation of Xist repressor *Tsix* through loss of interaction between *Tsix* and its enhancer Xite (Galupa et al., 2015). Subsequently, RNF12 is quickly silenced by *Xist* RNA to prevent the expression of *Xist* on the second X chromosome. *Xist* upregulation is further facilitated by the release of CTCF, which is bound to the *Xist* promoter, through the lncRNA Jpx (Sun et al., 2013). Once *Xist* is stably upregulated from the Xic of the future Xi, its lncRNA localizes to the X chromosome in *cis*. The molecular mechanisms that ensure *Xist* localization to the correct X chromosome remain poorly understood. However, localization seems to be independent of direct RNA-DNA binding. In fact, certain matrix proteins have been identified to interact with the *Xist* RNA to embed *Xist* in the nuclear matrix of the Xi (Brockdorff, 2018). One of the identified matrix proteins that was shown to contribute to the localization of *Xist* is the RBP CIZ1. CIZ1 binds the repeat-E element on the *Xist* RNA, thereby anchoring the lncRNA to the nuclear territory of the future Xi. The function of CIZ1 seems to be dependent on another matrix protein, the RBP hnRNPU, whose knockout leads to the dispersal of *Xist* RNA although it is still associated to CIZ1 (Sunwoo et al., 2017).

It is hypothesized that *Xist* RNA is able to spread across the future Xi in *cis* via so-called ***early* Xist *entry sites***, since X-linked genes with promoters close to early *Xist* entry sites were shown to be silenced early during the XCI process (Borensztein et al., 2017; Engreitz et al., 2013). Early *Xist* entry sites are sites that frequently interact with the *Xist* locus in 3D space and therefore are in spatial proximity to the *Xist* locus. Hence, the *Xist* RNA is able to rapidly reach the early *Xist* entry sites through proximity transfer because those distal genomic regions come in close contact to the *Xist* locus by chromosome folding. From there, *Xist* has been proposed to first propagate locally into gene-dense and then into intergenic regions (Simon et al., 2013). This spreading mechanism allows *Xist* RNA to initiate gene silencing across the entire X chromosome.

2.2.3 *Xist* -mediated repression of the inactive X chromosome

Once *Xist* RNA has coated the entire X chromosome in *cis*, it forms a ***transcriptionally silent compartment (TSC)*** that is depleted of RNA Polymerase II and associated transcription factors.

X-linked genes that become silenced are recruited from the periphery of the Xi territory into the TSC. The TSC mainly consists of a specific class of repetitive DNA elements, called LINEs (long interspersed nuclear elements), which are silenced during the early stages of XCI, prior to X-linked genes but independent of the A-repeat, which in turn is crucial for X-linked gene silencing. Genes that are located in LINE-rich regions are silenced more efficiently than genes located in LINE-poor regions, which tend to remain outside the TSC. A specific subset of young LINE-1 elements is expressed during later stages of XCI upon *Xist* -induced heterochromatin formation and remains outside the TSC. Such active LINEs potentially serve as way stations that facilitate local propagation of *Xist* and silencing of X-linked genes in escape-prone regions outside the TSC (Chow et al., 2010). The repeat hypothesis is supported by the fact that the X chromosome is enriched for LINEs compared to autosomes and that autosomes, which carry an *Xist* transgene, are silenced with lower efficiency relative to the X chromosome (Balaton et al., 2016).

The silencing of X-linked genes seems to depend on the A-repeat since recruitment of X-linked genes into the TSC is impaired in A-repeat mutants. Several RBPs have been identified to interact with the A-repeat, for instance, the transcriptional repressor SPEN, which recruits the NCoR-HDAC3 corepressor complex, leading to the loss of euchromatic histone marks by deacetylation of histone tails (Balaton et al., 2018). Another RBP that has been shown to bind to the A-repeat is RBM15. RBM15 interacts with WTAP, a core subunit of the m6A RNA methyltransferase complex that catalyses methylation of the N6-adenosine (m6A) residues of RNAs (Mira-Bontenbal et al., 2016; Moindrot et al., 2015; Patil et al., 2016). Different regions of the *Xist* RNA, including the A-repeat, are targeted by m6A modifications, possibly contributing to the stability and recognition of *Xist* RNA by RBPs (Rocha et al., 2017).

Another RBP that has been implicated in *Xist* -mediated gene silencing is hnRNPK, which does not bind to the A-repeat but instead to the B- and at lower levels also to the C-repeat. hnRNPK is thought to initiate Polycomb recruitment through interaction with PCGF3 and PCGF5, both proteins of the core subunit of non-canonical PRC1 (ncPRC1) (Figure 2.6) (Almeida et al., 2017; Pintacuda et al., 2017). After being recruited, ncPRC1 catalyzes mono-ubiquitylation of lysine 119 in histone H2A (H2AK119ub1), a repressive chromatin mark that directly contributes to gene silencing but also enables indirect recruitment of PRC2 through binding of JARID2 (Cooper et al., 2016; Rocha et al., 2014). PRC2 catalyses histone H3 lysine 27 trimethylation (H3K27me3), which is also a repressive chromatin mark. H3K27me3 promotes gene silencing as well but is also responsible for the recruitment of canonical PRC1 (Cao et al., 2002; Wang et al., 2004). Polycomb Group Proteins seem to be important to stabilize gene silencing but not necessary for the initiation of the silencing process mediated by *Xist* (Bousard et al., 2019). Hence, they represent the first repressive epigenetic layer in the process of XCI before a stable silent state is established on the Xi through accumulation of additional repressive histone marks (e.g. H3K9me2/3), incorporation of the H2A histone variant macroH2A and DNA methylation of CpG islands.

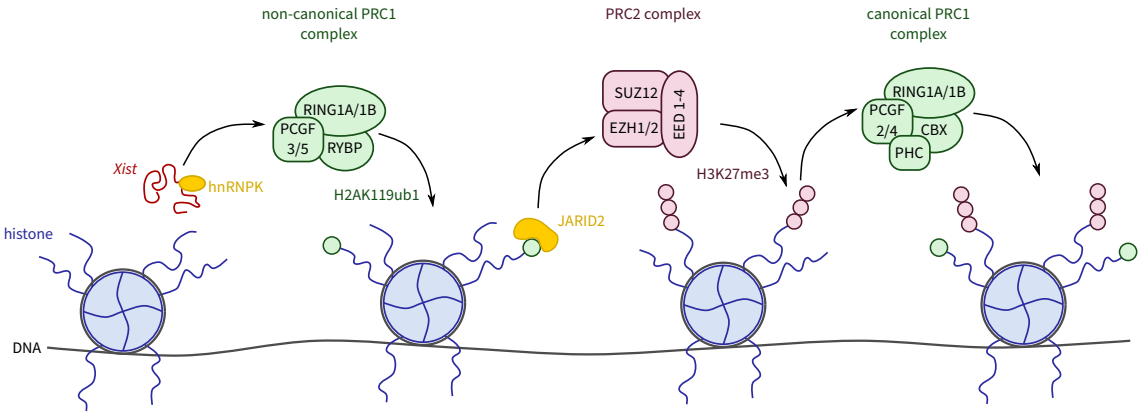

Figure 2.6: Current model for PRC1/PRC2 recruitment during early XCI. The RBP hnRNPK binds to the repeat-B and -C elements of *Xist* and helps to recruit non-canonical PRC1, which in turn deposits the histone mark H2AK119ub1. JARID2 is able to interact with PRC1 mark H2AK119ub1, leading to the recruitment of PRC2 to the Xi. PRC2 in turn deposits the repressive mark H3K27me3 through which canonical PRC1 is recruited. Inspired by (Galupa et al., 2018).

2.2.4 Structural reorganization of the inactive X chromosome

The upregulation of *Xist* on the future Xi initiates a series of events that lead to the silencing of X-linked genes but also triggers the spatial reorganisation and the localization of the Xi to the nuclear periphery.

Despite condensation due to spatial reorganization, the Xi does not have the typical heterochromatic structure with a high degree of compaction. Instead it only has a 1.2-fold higher compaction than the Xa (Naughton et al., 2010). More dramatic changes seem to occur in the chromatin structure of the Xi. The amount of 3D interactions on a chromosome is generally high, even within inactive chromosomal regions, because it leads to the compartmentalization of the chromosome into TADs. Allele-specific HiC studies, however, could show that the Xi lacks any complex three-dimensional structure because 3D interactions of silenced genes with other loci on the Xi are completely missing. Hence, the Xi is mostly devoid of TADs but has a conserved bipartite structure that is formed by CTCF-mediated superloops and separates the Xi into two large superdomains rather than individual smaller TADs as observed on other chromosomes. The anchor point between those two superdomains is the macrosatellite DXZ4, a repeat array that is able to bind CTCF, and has been shown to be essential for the bipartite structure of the Xi. *Xist* seems to play a crucial role in the formation of the two superdomains as well, because deletion of *Xist* from the Xi leads to the recovery of TAD structures on the Xi by increased cohesin binding. This in turn shows that *Xist* is continuously required to maintain the silent status of the Xi by controlling the spatial organization of the Xi.

The dramatic spatial reorganization of the Xi during the XCI process results in the characteristic condensed structure of the Xi, also known as ***Barr body***. Barr and Bertram first described the Xi as a nucleolar satellite because the Xi is commonly found at the periphery of the nucleolus or in proximity to the nuclear lamina after inactivation. The interaction between *Xist* lncRNA and the lamin B receptor, a component of the nuclear lamina, enables the recruitment of Xi to the nuclear

periphery. The attachment of the Xi to the nuclear lamina seems to be an important step in the XCI process because *Xist* RNA that is deficient for lamin B receptor binding fails to induce *Xist* -mediated gene silencing (Balaton et al., 2018; Monfort et al., 2017). Anchoring of the Xi to the nucleolus is mediated by the X-linked lncRNA Firre and the microsatellite DXZ4, which is also responsible for superdomain formation (Jégu et al., 2017). Perinucleolar association seems to be important for the stable maintenance of repressive marks, because the loss of this association leads to the erosion of H3K27me3 histone mark.

2.2.5 Escapees

Xist -mediated silencing of Xi genes occurs with different kinetics during the XCI process. Certain groups of genes are silenced early other later during the XCI process while some genes, so-called ***escapees***, are even able to escape XCI. Escapees are an exception in the XCI process because they are expressed from both, the Xa and Xi, leading to the assumption that they might have a female-specific function. Although escapees are expressed from both X chromosomes, the genes on the Xi are expressed at lower levels, typically only around 33%, compared to their counterparts on the Xa (Keniry et al., 2018). Hence, in most studies escapees are defined by having an expression level of at least 10% of the Xa gene expression (Balaton et al., 2018). In mice, 3 to 7% of X-linked genes escape XCI on average with some genes escaping in all cell types (constitutive escapees) and others escaping in a cell-type specific manner (facultative escapees) (Balaton et al., 2016). Facultative genes often have cell-type specific functions while constitutive escapees are genes enriched among Y-linked homologs or genes of the Xic.

The underlying mechanisms that define the silencing kinetics of X-linked genes are not yet fully understood. It has been proposed that early silenced genes lie in close genomic proximity to the Xic or to early *Xist* entry sites. Marks et al. could show that gene silencing dynamics correlate with the genomic distance from the gene to the Xic (Marks et al., 2015). The more distantly a gene is located from the Xic, the later it is silenced during XCI. Furthermore, promoter regions of early silenced genes seems to be depleted for active chromatin marks such as H3K4me3 but enriched for repressive chromatin marks like H3K27me3 and H3K9me3. Escapee promoters on the other hand are enriched in active but depleted in repressive chromatin marks as well as PRC2 components like EZH2, which are responsible for the deposition of repressive chromatin marks (Table 2.3). In addition, the histone variant macroH2A, which shows a fourfold enrichment on the Xi compared to the Xa, is locally depleted at escapee promoters (Balaton et al., 2016; Carrel et al., 2017; Pinheiro et al., 2017; Sahakyan et al., 2018).

Table 2.3: Enrichment and depletion of epigenetic marks at promoters of X-linked genes. Adapted from (Balaton et al., 2016).

Epigenetic Mark	Function	Silenced Gene	Escapee
H3K4me2	active	depleted	enriched
H3K4me3	active	depleted	enriched
H3K9ac	active		enriched
H3K9me1	active		enriched
H3K27ac	active	depleted	enriched
RNA Polymerase II	active	depleted	enriched
H3K9me3	repressive	enriched	
H3K27me3	repressive	enriched	depleted
H4K20me3	repressive	enriched	
macroH2A	repressive	enriched	depleted
Xist RNA	repressive	enriched	depleted

Gene silencing dynamics seem to further be affected by the nuclear position and folding of the gene loci. As mentioned before, the Xi is mainly devoid of TADs because silenced genes on the Xi lack 3D interactions with other loci. Escapees, however, show increased 3D interactions with each other, leading to the formation of mini TAD-like structures that potentially help escapees to loop out of the transcriptionally silent compartment (TSC) formed by *Xist* RNA (Jégu et al., 2017; Splinter et al., 2011). Binding of YY1 or CTCF near promoters of escapees potentially plays a role in the formation of those mini TAD-like structures and the interaction between escapee promoters and their *cis*-regulatory elements (Chen et al., 2016; Filippova et al., 2005). Escapees remain at the periphery of the TSC even at later stages of differentiation, most likely due to better accessibility of the transcription machinery that is excluded from the TSC. Supporting this idea, RNA Polymerase II is more frequently found at escapees than at silenced genes. Despite all the findings described above, still very little is known about how escapees establish and maintain their active state in the heterochromatic environment on the Xi.

3 MACHINE LEARNING BACKGROUND

"In the 1990's "data mining" was an exciting and popular new concept. Around 2010, people instead started to speak of "big data." Today, the popular term is "data science". However, during all this time, the concept remained the same: use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems in science, commerce, healthcare, government, the humanities, and many other fields of human endeavor."

— (Leskovec et al., 2014)

Recent technological advances in molecular science have made it possible to analyse biological systems in a high throughput fashion. The possibility to perform high throughput experiments at low cost led to a dramatic increase of generated biological data, heralding the era of "Big Data" in biology - a pervasive buzzword for the huge amount of (biological) data that is generated every day. The speed of data growth can be observed for instance at the European Bioinformatics Institute (EMBL-EBI), which maintains big data bases like Ensembl, where the total disk capacity for storing biological data grew from around 60 petabytes at the end of 2015 to just over 160 petabytes by the end of 2018 and a continuing exponential growth is expected (Cook et al., 2019).

The generated biological data ranges from microscopic imaging to protein structures or genomic sequences and consortia of leading research institutes were able to assemble the generated data into huge biological data sets, like the Encyclopedia of DNA Elements (ENCODE) project that has collected more than 9000 genomic datasets from ChIP-seq, RNA-seq, Hi-C and other experiments (ENCODE Project Consortium 2004). The availability of such huge data sets gives us the opportunity to answer challenging biological questions, where the underlying mechanisms are complex and depend on the interplay of many different regulatory factors. Often, data sets of different sources and types need to be integrated and analysed in a data-driven manner to understand how regulatory systems work together and to uncover the patterns that explain those complex mechanisms. However, efficient analysis of such large and complex data sets is merely impossible by visual investigation or traditional statistical methods (e.g. pairwise correlations). Instead, machine learning (ML) algorithms offer the opportunity to systematically extract information from the integrated data sets and gain new biological insights by generating data-driven hypotheses that can be validated later on with biological experiments.

But what exactly is the magic behind those ML algorithms? The subject of ML can be seen as a joint subfield of statistics and computer science that deals with the development of computational algorithms, which are able to identify complex patterns in large data sets and make predictions based on a given input data set. Hence, ML algorithms can help to extract knowledge and gain insights from structured and unstructured data, considering thousands of observations and their relationship between each other at once. ML has many applications in different fields within science and industry, ranging from healthcare to intelligent process automation and is nowadays

used in every area of our lives: shopping platforms recommend products we might be interested in; online maps propose alternatives routes to not get caught up in a traffic jam and Google's ML based program AlphaGo beats the reigning champion of the ancient board game Go. (Silver et al., 2016)

Generally, ML algorithms can be categorized into two broad classes: ***supervised*** and ***unsupervised ML*** algorithms (Figure 3.1). An input data set usually consists of a set of observations for which we have a set of input features and optionally a set of known output measurements that can be either continuous or categorical. We call a data set ***labelled*** if we have known output measurements and ***unlabelled*** if known output measurements are missing. Supervised ML algorithms aim to find underlying pattern in labelled data by learning a function on given input features that approximates the known output measurement. Supervised ML applications in bioinformatics are for example the prediction of the gene expression status from different chromatin marks (Karlić et al., 2010). Unsupervised ML algorithms, on the other hand, discover underlying patterns in unlabelled data by extracting structures from the data itself. A typical application is the discovery of recurring biological patterns from different epigenetic or genomic data sets e.g. the systematic annotation of gene regulatory elements from ChIP-seq data by ChromHMM (Ernst et al., 2012).

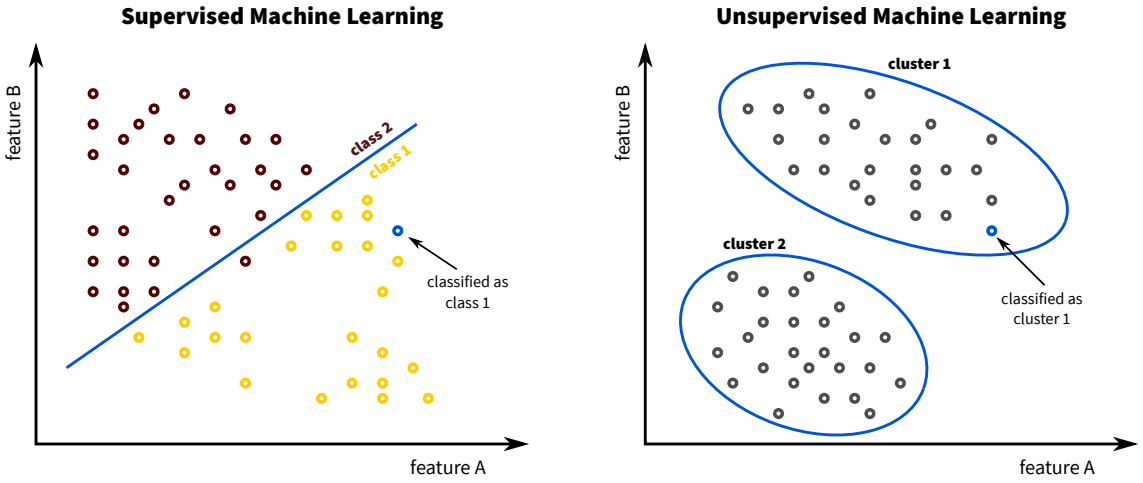

Figure 3.1: Supervised vs unsupervised machine learning. The goal of supervised learning is the identification of a function that explains the output measurements of each observation by the input features. Usually we either have a regression (continuous measurement values) or a classification (categorical measurements) problem. In the latter case we try to find a decision boundary that best separates the classes (left panel). The goal of unsupervised learning is to find structure in unlabelled data. A common technique to find groups of similar observations is clustering (right panel).

The choice for a particular ML algorithm, whether supervised or unsupervised, strongly depends on the biological question to be answered and the data available for answering that question. Every ML algorithm has its advantages and disadvantages and the choice of an appropriate ML algorithm is influenced by different properties of the collected data set. For instance, properties like multicollinearity, where similar features with correlating values are included in the data set, curse of dimensionality, where the data set has fewer observations than features, and confounding factors, where measurements are affected by batch effects caused by different laboratory condi-

tions, all require different strategies. Hence, every complex biological question requires specific data preparation and a tailored ML solution.

This chapter gives an introduction to the core ML concepts and a more detailed description of the main ML algorithms used throughout this thesis. In Section 3.1 the basic mathematical notations are defined. Section 3.2 gives an introduction to the main concepts of the supervised ML methods used in this thesis, while Section 3.3 gives an introduction to the unsupervised ML algorithms used in this thesis.

3.1 MATHEMATICAL NOTATIONS

In this section, the statistical framework and basic notations for this thesis are defined. All notations are based on the second chapter of the Statistical Learning Book of Hastie et. al. (Hastie et al., 2009). Let $Z = (x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)$ be a data set of n independent identically distributed (i.i.d.) observations. Then for every observation i we have a vector of p input features $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ with associated known output measurement y_i. Input features, which are typically called ***predictor*** or ***independent variable***, can be continuous or categorical and are represented as a $n \times p$ matrix $X = (X_1, \dots, X_j, \dots, X_p)$, where $X_j = (x_{j1}, x_{j2}, \dots, x_{jn})$, with n observations and p features. The vector of known output measurements $Y = (y_1, y_2, \dots, y_n)$ is typically called the ***response*** or ***dependent variable*** and can be either quantitative (continuous response) or qualitative (categorical response). For X and Y we assume a relation of $Y = f(X) + \epsilon$ where f is a fixed but unknown function on X and ϵ is a random error term with mean zero, independent of X.

3.2 SUPERVISED MACHINE LEARNING

The goal of any supervised machine learning (ML) algorithm is to find complex structures in data sets with hundreds of observations that can be used to predict the response of new unseen observations (predictions) and understand the dependence of the response on different predictors (interpretation). To relate the response to the predictors, the algorithm tries to find a function $\hat{f}(X)$ of the predictor variables that approximates the response variable accurately such that $Y \approx \hat{f}(X)$ for any observation (x_i, y_i). If the function approximates a continuous response, we call it a ***regression problem***, whereas if the function maps the predictor variables to a categorical response, we call it a ***classification problem***.

Supervised ML algorithms can be further divided into ***parametric*** and ***non-parametric*** approaches. Parametric methods simplify the approximation function to a known functional form, e.g. assuming that $\hat{f}$ is linear in X, which creates a linear model of the form:

$$\hat{f}(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

This assumption greatly simplifies the modelling process because now we just have to estimate the $p + 1$ coefficients $\beta_0, \dots, \beta_p$ of the predefined function from the given data set instead of fitting an arbitrary p-dimensional function to the predictor variables. Prominent examples of parametric

models are linear as well as generalized linear models. Non-parametric approaches, on the other hand, do not make any explicit assumptions on the functional form of the approximation function. Instead, the functional form is learned from the given data set itself, making it possible to fit the approximation function $\hat{f}$ to a wider range of functional forms. Common examples of non-parametric methods are k-nearest neighbors, decision trees and Support Vector Machines with radial basis function Kernels. In general, parametric models are easier to interpret but constrained by the functional form that needs to be specified beforehand, which might not match the true underlying function f. In contrast, non-parametric approaches are more flexible because they do not make any assumptions on the functional form of the approximation function but they are more prone to overfitting and also harder to interpret (Hastie et al., 2009).

Model ***overfitting*** occurs when the ML model is too complex and therefore, is not generalizable to new unseen data. Complex models usually predict the response of the observations they are trained on very well (small learning error) but perform poorly on new unseen observations (large generalization error). Model complexity is commonly defined by the number of fitting parameters in the model (e.g. coefficients $\beta_0, \dots, \beta_p$ in a linear regression model), meaning that the model gets more complex as we add more parameters. But how exactly do we identify model overfitting? One way of identifying overfitting is to calculate the generalization error on an independent data set that was not used to optimize the model parameters and train the final model. Therefore, the given set of observations $Z = (x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)$ is divided into three non overlapping and independent data sets: ***training set*** $Z_{training}$, ***validation set*** $Z_{validation}$ and ***test set*** Z_{test}. The training set is the largest data set and is used to fit the approximation function of the model, while the validation set is used to calculate the model performance on different sets of hyperparameters. Hyperparameters are higher-level structural properties of the model which need to be fixed before fitting the approximation function because they cannot be learned from the training data.

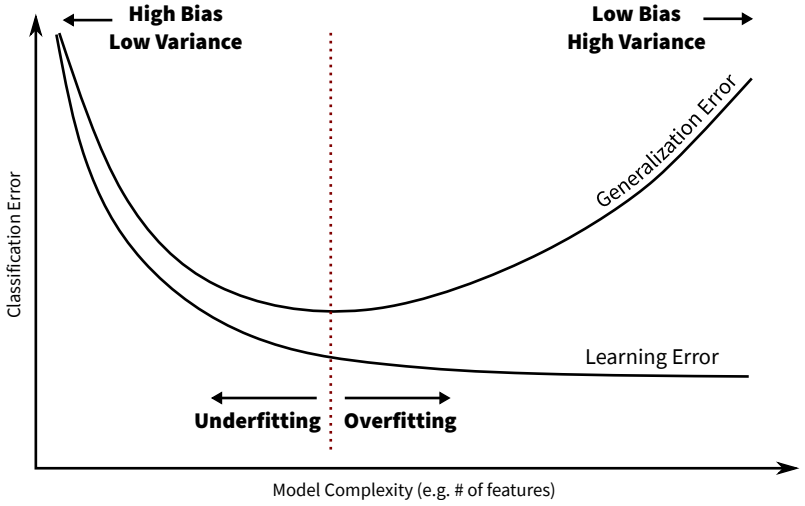

Figure 3.2: Relationship between Learning and Generalization Error. Model overfitting can be identified through comparison of learning and generalization error. The learning error tends to monotonically decrease with increasing model complexity, while the generalization error tends to increase if the model gets too complex.

The ***learning error***, which is the error made on the validation set, is used to select the best set of hyperparameters on which the final model $\hat{f}_{final}$ is trained. The test set is then used to access the performance of the final model by predicting the outcome of the observations in Z_{test} with the final model: $\hat{Y}_{test} = \hat{f}_{final}(X_{test})$ and comparing it to the ground truth Y_{test}. The error made on the test set is called the ***generalization error*** and describes how good the model performs on unseen observations. By comparing the learning error with the generalization error, we can identify model overfitting and then chose a model complexity that leads to the lowest generalization error to avoid overfitting (Figure 3.2).

In situations where the available input data set is small, it is possible to divide the data into only two subsets, training and test set, and perform model training and hyperparameter tuning on the same subset, the training set. Therefore, methods like *k*-***fold cross-validation*** or ***bootstrapping***, that use different splits of the training set for every set of hyperparameters, can be applied to get the final model (Izenman, 2008). In *k*-fold cross-validation, the training set is divided into K distinct subgroups, called folds. The model is trained on $K - 1$ folds while the k^{th} fold is used as validation set to compute the learning error. This process is repeated until each of the K folds was used once as validation set and the learning error is averaged across all folds (Geisser, 1975). Bootstrapping on the other hand, is a data resampling method that draws a bootstrap sample of size N with replacement from a training set of size N. Hence, some observations are represented multiple times in the bootstrap sample while others are left out. Following the 632+ bootstrap rule, typically $^1/_3$ of the observations are left out of the bootstrap sample (Efron et al., 1997). In total, B bootstrap samples are drawn and a model is fitted to each of the bootstrap samples. The observations that were left out of the bootstrap sample - Out of Bag data - are then used to compute the learning error, which is then averaged over the B bootstraps (Efron et al., 1994).

The focus of this thesis lies on classification problems, more specifically on binary classification problems, where the response variable Y is divided into two classes $Y \in 0, 1$. Here, the classification task is to train a classifier $\hat{f} : X \longrightarrow 0, 1$, which maps the predictor variables of an observation to one of the two classes, e.g predict the disease status (yes/no) of a patient based on gene expression data. There exist a variety of methods that perform binary classification tasks, like logistic regression, Support Vector Machines, Neural Networks or Random Forests. In this thesis, the focus is on the parametric logistic regression method and the non-parametric Random Forest algorithm. The following sections explain both methods and are based on Izenman, 2008 and Hastie et al., 2009.

3.2.1 Linear Models

Linear models are a large class of supervised machine learning (ML) algorithms that assume a linear relationship between predictor variable X and response variable Y and can be used to solve both, regression (continuous response) and classification (categorical response) problems.

One of the simplest linear models is ***linear regression***, where the linear relationship between predictor variables X and a continuous response Y is modelled with a weighted linear combination of the predictor variables:

$$Y = f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

where X_j is the j^{th} predictor variable and $\beta_0, \dots, \beta_p$ are a set of unknown model coefficients that have to be estimated from the training set $Z_{training}$. The most common method to estimate $\beta_0, \dots, \beta_p$ is the least squares approach, which selects the set of coefficients that minimizes the residual sum of squares:

$$\hat{\beta} = \arg\min_{\beta} RSS(\beta)$$

where the residual sum of squares (RSS) is defined as:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

The estimated model coefficient $\hat{\beta}_j$ represents the mean change in Y for one unit of change in X_j. We can then use the linear model with the estimated model coefficients $\hat{\beta}_0, \dots, \hat{\beta}_p$ to make predictions for new unseen observations:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j$$

The performance of a linear regression model can be assessed with different metrics. One possible performance metric is the R^2 statistic, which calculates the proportion of variability in Y that can be explained by the linear regression model:

$$R^2 = \frac{TSS - RSS}{TSS}$$

where the total sum of squares (TSS) is defined as:

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

and measures the amount of variability in the response Y, whereas the RSS measures the amount of variability that is left unexplained by the linear regression model. Hence, an R^2 value close to one indicates that a large fraction of the variability in Y is explained by the model, while an R^2 value close to zero indicates that a large fraction of the variability in Y is left unexplained by the model.

Linear models can also be applied to classification problems, where the probability that the response variable Y belongs to a certain class K is modelled:

$$f(X) = Pr(Y = K|X)$$

The simplest classification setting is a binary classification, where Y is divided into two classes ($K = 2$). In this case, $f(X)$ gives the posterior probability of Y belonging to the positive class:

$$f(X) = Pr(Y = 1|X)$$

and should take values between zero and one. A transformation function $T(X) : X \longrightarrow [0, 1]$ can be used to model the relationship between predictor variables and probabilities, i.e. the logistic function:

$$p(X) = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

which transforms a real number to values between $[0, 1]$. By log-transforming the logistic function, we get a linear regression model of the log-odds, also called a ***logistic regression*** model:

$$log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

which allows us to predict the posterior probability of $Y = 1$. The idea is to find a set of coefficients $\beta_0, \dots, \beta_p$ that maximizes the posterior probability for $Y = 1$ and minimizes the posterior probability for $Y = 0$. A commonly used method to solve this problem is the maximum likelihood estimator:

$$\hat{\beta} = \arg\max_{\beta} L(\beta_0, \dots, \beta_p)$$

where the likelihood L is the product of both probabilities:

$$L(\beta_0, \dots, \beta_p) = \prod_{i:\, y_i=1} p(x_i) \times \prod_{i:\, y_i=0} (i - p(x_i))$$

and $p(X)$ is related to the model coefficients via the logistic function $p(X) = \frac{e^{\beta X}}{1+e^{\beta X}}$. The coefficients of a logistic regression model indicate the change in log-odds for a one unit increase in X. A positive coefficient β indicates that an increase in X is associated with an increased posterior probability, while a negative coefficient β indicates a decreased posterior probability. We can use the logistic regression model with the estimated model coefficients $\hat{\beta}_0, \dots, \hat{\beta}_p$ to make predictions for new unseen observations:

$$\hat{p}(X) = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

which can be used to assess the performance of a linear regression model. Therefore, we assign each new observation to one of the two classes, based on its predicted probability to belong to the positive class, and compare the assigned class to the true observed class y_i. Different performance metrics make use of this comparison to evaluate the logistic regression model. One of the simplest performance measures is the ***misclassification error***:

$$error = \frac{FP + FN}{n}$$

where FP includes all observations, which were wrongly predicted as $Y = 1$ and FN includes all observations, which were wrongly predicted as $Y = 0$.

If a linear model (regression or classification model) was fit to a large number of predictor variables with a high predictor to observations ratio ($n << p$ situations), one has to be careful when interpreting the estimated model coefficients $\hat{\beta}_0, \dots, \hat{\beta}_p$. In a $n << p$ situations, the linear model is underdetermined and hence, the system of linear equations has many feasible solutions, where small changes in the training set may lead to highly varying results, even when the model bias stays constant. Generating sparsity in the input feature space can help to overcome this problem,

because it reduces the number of predictor variables used to fit the model. The interpretation of model coefficients associated to correlated predictor variables is highly problematic as well, because for a set of equally important but highly correlated predictor variables, often only one predictor has a high coefficient, i.e. is interpreted as an important predictor, while the others have low coefficients, i.e. are wrongly interpreted as unimportant predictors. The joint selection of correlated features into the model can help to overcome this problem. Both, the sparsity and grouping problem, can be addressed with ***regularization*** methods that penalize large model coefficients, leading to a sparse input feature space, and encourage a grouping effect, leading to the joint selection of correlated features. In the following, the application of different regularization techniques to a logistic regression model are explained. However, those regularization techniques can also be applied to other machine learning models.

A common regularization method that addresses the sparsity problem is the ***Lasso*** regularization. Lasso constrains the maximum likelihood estimator with an additional penalty term that shrinks the coefficients towards zero:

$$\hat{\beta}_{\lambda}^{L} = \arg\max_{\beta} L(\beta_0, \dots, \beta_p) - \lambda \sum_{j=1}^{p} |\beta_j|$$

where $\lambda \sum_{j=1}^{p} |\beta_j|$ is the shrinkage penalty, using the L_1 norm as a constraint, and $\lambda \geq 0$ is a tuning parameter, controlling the impact of both terms in the equation and thereby regulating the trade-off between the goodness of fit and size of coefficients. The constraint region of the L_1 norm shrinks the coefficients towards zero and forces some of them to be equal to zero if λ is large enough (Figure 3.3). Hence, the Lasso regularization can be used to perform feature selection but is not able to address the grouping problem that arises from correlated features.

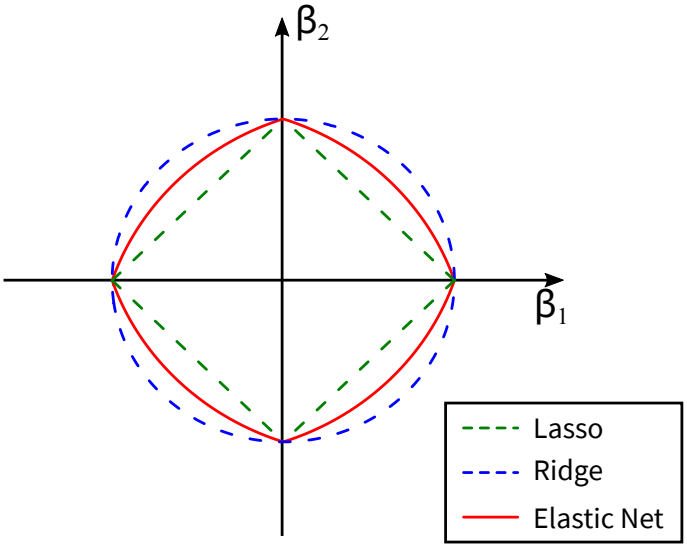

Figure 3.3: Regularization of Linear Models. Regularization is achieved by adding an additional shrinkage penalty term to the objective function. The constraint region of Lasso (green dotted line) has a diamond shape with corners, which helps to generate a sparse model by shrinking some coefficients to exactly zero. The constraint region of Ridge (blue dotted line) has a circular shape with convex edges, which encourage a grouping effect. Elastic Net (red line) combines both penalty terms, leading to a constraint region that has convex edges but also singularities at vertices, which generates a sparse model and encourages the grouping of correlated variables.

Ridge regularization addresses the grouping problem by using the quadratic L_2 norm as a shrinkage penalty:

$$\hat{\beta}_{\lambda}^{R} = \arg\max_{\beta} L(\beta_0, \dots, \beta_p) - \lambda \sum_{j=1}^{p} \beta_j^2$$

The constraint region of Ridge regression encourages a grouping effect by which groups of correlated features tend to have similar model coefficients, i.e. similar importance in the model (Figure 3.3). However, Ridge regularization will keep all features in the final model, because it is not able to shrink the coefficients to zero and hence, does not provide a sparse model.

A solution to this problem is ***Elastic Net***, which combines the shrinkage penalties from Lasso and Ridge (Figure 3.3) to perform feature selection in combination with a grouped selection of correlated predictors:

$$\hat{\beta}_{\lambda}^{EN} = \arg\max_{\beta} L(\beta_0, \dots, \beta_p) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

where λ_1 and λ_2 control the impact of both shrinkage penalties. If $\lambda_1 \approx \lambda_2$ we have a balanced impact of feature selection and feature grouping, while $\lambda_1 >> \lambda_2$ or $\lambda_1 << \lambda_2$ emphasizes either the selection or the grouping effect.

Regularization methods help to better interpret the model coefficients of a linear model by shrinking unimportant predictors to zero and/or giving correlated predictors equal importance. However, in cases of $n << p$, where we have many more predictor variables than observations, linear models still suffer from the curse of dimensionality, where parameter estimates become highly unstable. Nevertheless, regularized linear models often provide an adequate and interpretable description of how the response variable is affected by the predictor variables. However, the interpretability comes at the cost of assuming a linear dependence between predictor and response variables, but for many biological problems, a linear relationship between response and predictor variables is not given. In such cases it can be beneficial to use non-parametric ML models that make no assumptions on the relationship between response and predictor variable. A large class of non-parametric models are tree-based models that will be described in the next section.

3.2.2 Random Forest

Random Forest is a commonly used non-parametric machine learning (ML) algorithm, which is based on the concepts of decision trees and ensemble learning.

Decision Trees

Decision trees are non-parametric tree-based learning methods that recursively split the predictor space X into smaller subsets such that the resulting subgroups of Y are as homogeneous as possible. A decision tree can be represented by a tree graph with one root node that contains the whole input data set, many internal nodes that represent the splitting points on the predictor variables X and terminal nodes that are not further split and contain a subgroups of Y belonging

to a certain class (Figure 3.4). The first step in constructing a decision tree is to split the root node based on the predictor variable X into two terminal nodes to improve the homogeneity of the response variable Y in each terminal node compared to the root node. This splitting step is repeated with the two terminal nodes, which now become internal nodes, to successively improve the homogeneity of the response variable in each terminal node until we reach a predefined stopping criterion. A natural stopping criterion is the node purity, where the tree is grown until the terminal nodes are homogeneous, hence all members in a terminal node belong to the same class. This however, often leads to model overfitting, because the model starts being too complex and tends to learn the noise in the data as well. An appropriate stopping criterion serves the purpose of finding a balance between too complex models, which overfit the data, and too simple models, which underfit the data, both leading to a high generalization error. A suitable stopping criterion can be a minimum member size of each terminal node, where the terminal node is then marked with the most frequent class of its members, or a maximum depth for the tree, where the longest path from the root node to a terminal node should not exceed a certain threshold. By applying such methods we "prune" the decision tree.

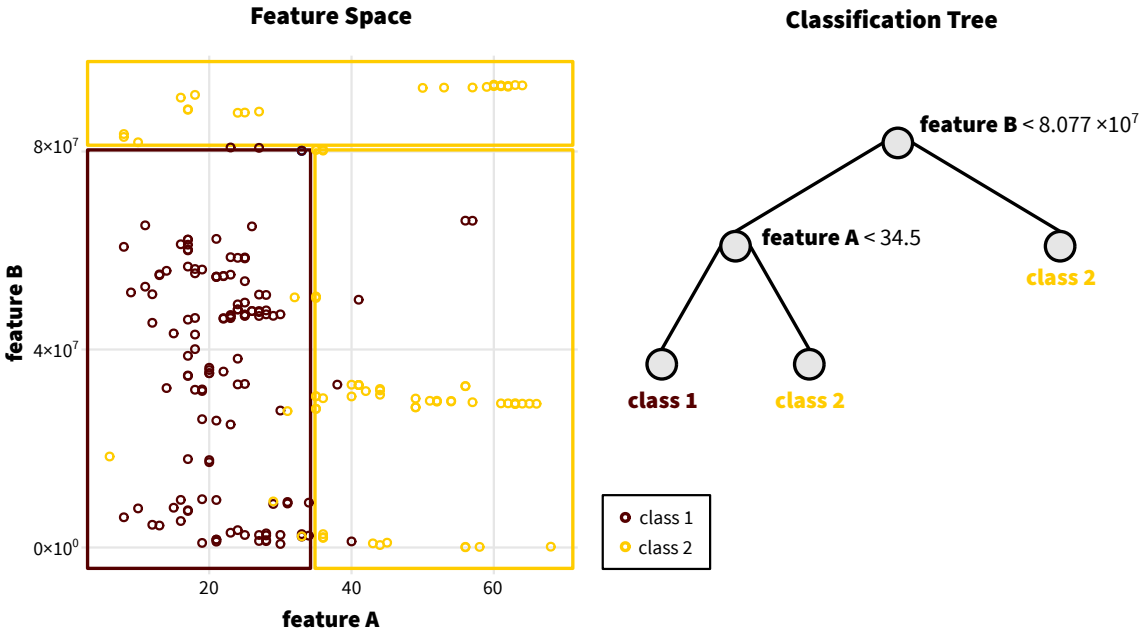

Figure 3.4: Example of a classification tree. Classification of a response variable into two classes (red and yellow) in the two-dimensional input feature space with a classification tree. The classification tree makes two splits on feature A and feature B and has three terminal nodes, representing the subgroups of the response variable. The splits divide the feature space into rectangles, where most points belong to either one of the two classes and the region is labelled with the majority class.

Decision trees can be used for regression or classification tasks. Here, the focus is on decision trees used for classification that are build with a popular tree-based method called CART (Classification and Regression Trees) which was introduced in 1984 by Breiman et al. (Breiman et al., 1984). The construction of a globally optimal decision tree has been proven to be an NP complete problem (Laurent et al., 1976). Hence, decision tree algorithms like CART are based on heuristics that build the decision tree in a top-down, greedy approach resulting in locally optimal decision trees

because at each step we make the best split for that particular step instead of choosing a less optimal split that would lead to a better tree in some future steps. The CART algorithm constructs a decision tree, by dividing the predictor space X into l distinct and non-overlapping regions R_l (defined by the split points s of the internal nodes) such that the homogeneity of Y in each region is maximized, which is equivalent to minimizing the node impurity in that region. Therefore, for each node m we select a predictor X_j with splitting point $s \in S_j$, where S_j is the set of all possible split points for X_j, such that we get the maximal decrease in node impurity for the two daughter nodes $m1$ and $m2$. The node impurity for a classification problem with K classes can be measured by different impurity indices. The most popular impurity index is the ***Gini index***. Let the proportion of members in node m belonging to class $k \in K$ be defined as:

$$p_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

where I is the identity function, R_m is the region of node m containing in total N_m observations and $\sum_{k=1}^{K} p_{mk} = 1$. Then the Gini index for node m is defined by:

$$G_m = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) = 1 - \sum_{k=1}^{K} p_{mk}^2$$

Small Gini index values indicate a more pure node than higher Gini index values, where $0 \leq G_m \leq 1$. To compute the optimal split for a node m, the algorithm computes for every unique value s in the space of each predictor X_j the reduction in impurity of the two daughter nodes $m1$ and $m2$ compared to the parent node m:

$$\Delta G_m(X_j, s) = G_m - \left(\frac{|N_{m1}|}{|N_m|} \times G_{m1}(X_j, s) + \frac{|N_{m2}|}{|N_m|} \times G_{m2}(X_j, s) \right)$$

for all $X_j \in X$ and $s \in S_j$. The best split s_{opt} for node m is the one that has the highest reduction in impurity, hence $\Delta G_m(X_j, s)$ with the largest value. This split point is then used to split the region R_m of node m into two sub-regions: $R_{m1} = \{X | X_j < s_{opt}\}$ and $R_{m2} = \{X | X_j \geq s_{opt}\}$.

The constructed decision tree can then be used to predict the class of a new unseen observation: start at the root node, drop the new observation down the left or right daughter node, depending on its value of the predictor variable that was used at that split, repeat until a terminal node is reached. For each new observation that falls into terminal node l we will make the same prediction, which is the majority class of the response values in region R_l.

The partitioning of the predictor space X into smaller subsets allows us to model nonlinear, complex relationships between predictor and response variables. However, if we have a linear dependency of the response variable on the predictor variables, a linear model, e.g. a logistic regression model, will most likely outperform a decision tree model because the linear model can approximate the linear dependency by a linear function whereas the decision tree has to use a step function. In addition, decision trees are very sensitive to changes in the input data and are prone to overfitting when constructing trees that are too complex. To avoid such problems one can build a model based on an ensemble of decision trees, trained on bootstrapped input data.

Ensemble Learning

The predictive performance of weak ML models like decision trees can be improved by a technique called ensemble learning, which combines a group of weak predictor models, to form a strong ensemble learner. The idea behind ensemble learning is to improve the predictive performance by reducing the variance term of the generalization error. The generalization error of a predictor model can be decomposed into two elements, a bias term and a variance term (Figure 3.5). While the bias measures how well the function learned on the predictor variables approximates the true underlying function, the variance describes how much this bias varies across different training sets. Interestingly, the variance of the average of B i.i.d. predictor models with variance σ^2, is $\frac{\sigma^2}{B}$ while the bias of the average is the same as that of an individual predictor model. Hence, by building an ensemble learner from a set of B individual predictor models, we can theoretically reduce the variance by a factor equal to the number of models in the ensemble. Thereby, we are able to reduce the generalization error and improve the predictive performance of the ML model. Common methods for ensemble learning are ***Bagging*** (Bootstrap Aggregation), Boosting or Randomization. Here, the focus is on Bagging, a method introduced in 1996 by Breiman et al. (Breiman, 1996) that can be used to aggregate multiple decision trees to form a strong ensemble model.

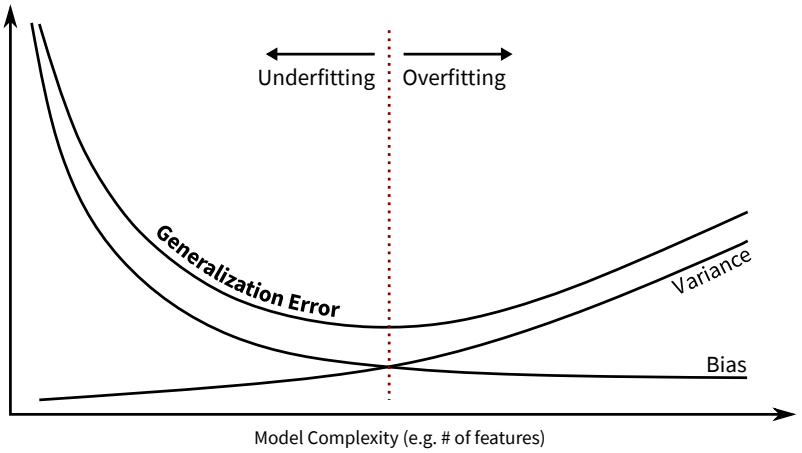

Figure 3.5: Bias-Variance decomposition of the Generalization Error. The generalization error can be decomposed into a bias and a variance term. The bias monotonically decreases, while the variance monotonically increases with increasing model complexity. To avoid overfitting of the model due to model complexity, which in turn increases the generalization error, we have to find the optimal trade-off between bias and variance.

In Bagging, trees are fully grown, hence have a low bias, but predictions are averaged over multiple trees which reduces variance. To build a bagged model, B bootstrap samples Z_b are drawn from the training set $Z_{training}$ and on each bootstrap sample a decision tree $\hat{f}_b(Z_b)$ is trained (Figure 3.6). To obtain the predicted class for a new observation i, the majority class across all trained decision trees in the bagged model is calculated:

$$\hat{f}^B_{bag}(x_i) = majority\ vote\{\hat{f}_b(x_i)\}^B_1$$

where $\hat{f}_b(x_i)$ is the class prediction of the b^{th} decision tree in the ensemble based on the predictor variables x_i of observation i. The learning error of the bagged model can be calculated during the training phase through the ***Out of Bag (OOB) error***. The OOB error is an unbiased estimate of the learning error because it is calculated on the OOB data, which is the data that was not used to train the bagged model. To calculate the OOB error, the majority vote of the predictions for a training observation i over all decision trees $\hat{f}_b$, in which observation i was part of the OOB data of $\hat{f}_b$, is calculated. The fraction of OOB observations that were classified incorrectly is then the OOB error. It has been shown that OOB error estimates are nearly identical to k-fold cross validation estimates (Hastie et al., 2009).

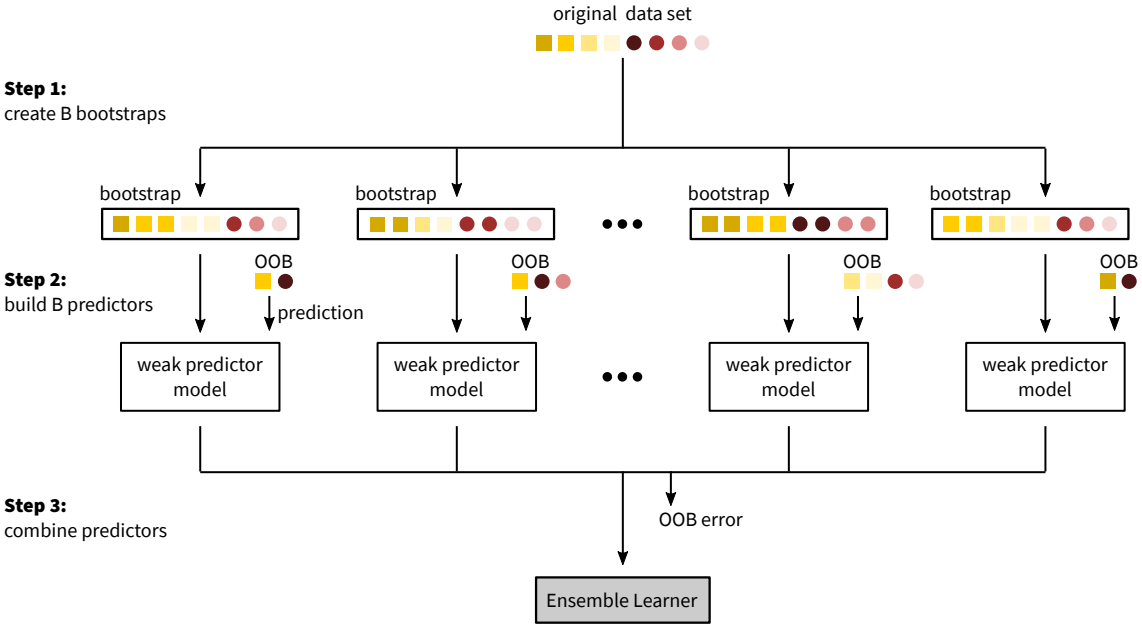

Figure 3.6: Construction of a bagged model. A bagged model is a collection of weak prediction models (e.g. decision trees), each build on a bootstrapped data set of the original input data, which are then combined to form a strong ensemble learner.

Notably, the assumption of variance reduction in Bagging only holds for i.i.d. predictor models. If the predictor models are simply identically distributed but not necessarily independent, e.g. the common case that some decision trees in the ensemble are correlated amongst each other, the variance of the average is not $\frac{\sigma^2}{B}$ but instead $\rho\sigma^2 + \frac{1-\rho}{B} \times \sigma^2$. By increasing B, the number of predictor models in the ensemble, the second term becomes negligible small, however, the first term remains and therefore limits the benefits of Bagging by the amount of correlation ρ between the predictor models. To decrease the overall amount of correlation in the ensemble, the predictor models have to be decorrelated. The solution to this problem is implemented in an algorithm called Random Forest.

Random Forest Algorithm

The Random Forest (RF) algorithm was introduced in 2001 by Breiman et al. (Breiman, 2001) and extends the Bagging algorithm by building an ensemble of decorrelated decision trees. Decision trees become correlated if only few features are strong predictors of the response variable, leading to the majority of decision trees having a similar structure (the strong predictor is used as first split in many trees) and therefore highly correlated predictions. To reduce the correlation between decision trees, RF performs ***random feature selection*** at each node prior to the selection of the optimal split s_{opt}. Hence, the reduction in node impurity is only computed on a random subset of predictor variables, which reduces the chance that strong predictors are always used as first splits.

Definition (Random Forest). A Random Forest is an ensemble classifier consisting of a collection of B tree-structured classifiers $\{\hat{f}(x, \Theta_b)|b = 1, \dots, B\}$ where the Θ_b is the number of randomly selected features and each tree casts a unit vote for the most popular class at input x.

In short, RF creates an ensemble of decision trees by fitting unpruned decision trees to a set of B different bootstrap samples, while selecting at each split a random subset of $p' < p$ input features as candidates for splitting. The class of a new unseen observation x is then predicted as the majority class across predictions for x made with all trees in the ensemble. By averaging the predictions over a large ensemble of high variance but low correlation and low bias decision trees, RF is able to improve the variance reduction of Bagging and efficiently reduce both components - bias and variance - of the generalization error. Figure 3.7 and Algorithm 1 summarize the training and prediction steps for a RF classification model.

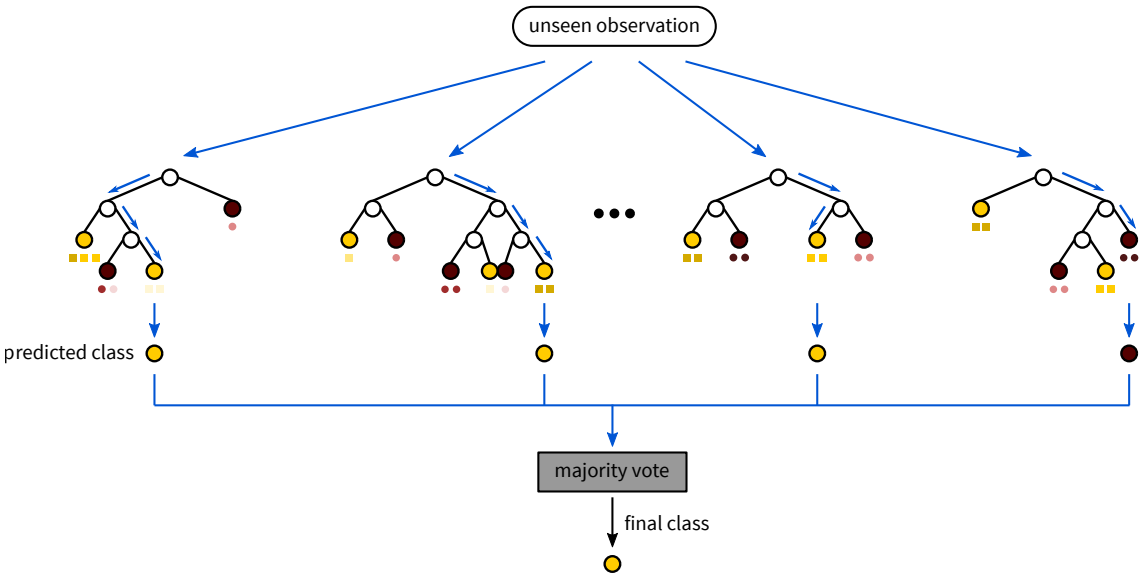

Figure 3.7: Classification with a Random Forest model. To classify a new unseen observation with a RF model, the class of the unseen observation is predicted with each decision tree in the model. The final class is then determined by a majority vote over the predicted classes.

Algorithm 1: Random Forest for Classification

Input: data set $X = (X_1, X_2, \dots, X_p)$,
silencing class $Y = (y_1, y_2, \dots, y_n)$

1) Train the RF model

for *b=1 to B* **do**
  draw a bootstrap sample Z_b from the training set $Z_{training}$;
  train a decision tree model T_b on the bootstrapped data by:
  repeat
    for each terminal node starting at the root node:
      randomly select $p' < p$ of the original predictor variables;
      compute the best split s_{opt} based on the p' predictor variables;
      split node m into two daughter nodes $m1$ and $m2$ with
      $R_{m1} : X|X_j < s_{opt}$ and $R_{m2} : X|X_j \geq s_{opt}$;
  until *stopping criterion reached*;
end
return *ensemble of trees* $\{T_b\}_1^B$

2) Make predictions

for *new observation x* **do**
  let $\hat{C}_b(x)$ be the class prediction of the b^{th} RF tree;
  then $\hat{C}_{rf}^B(x) = majority\ \ vote\{\hat{C}_b(x)\}_1^B$
end

As in Bagging, we can use the OOB error rate as an unbiased estimate for the learning error because it accesses the model performance on the OOB data set that was not used to train the model. In addition, Breiman (Breiman, 2001) could show that the generalization error (GE) converges to an upper bound when the number of decision trees in the RF model is large enough:

$$GE \leq \frac{\bar{\rho}(1 - s^2}{s^2}$$

where $\bar{\rho}$ is the mean correlation between decision trees and s is the strength of the decision trees $\{\hat{f}(x, \Theta_b)|b = 1, \dots, B\}$ in the model (e.g. a measure of accuracy of a decision tree in the model). Hence, if the number of decision trees in the RF model is chosen large enough, overfitting of the RF model can be avoided.

Interpretation of Random Forest Models

Complex supervised ML models are often considered to be "Black Boxes" because it can be hard to understand why certain predictions have been made by the model. It means that although the model correctly predicts the outcome of an observation, we cannot explain the logic behind those

predictions. But why aren't we just satisfied with an accurately predicting model?

"The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks."

— (Doshi-Velez et al., 2017)

Particularly in biology, it is more and more important to not just accurately predict the outcome of a biological system with an ML model but to also be able to uncover the mechanisms behind those biological systems that led to a certain outcome. To uncover the underlying mechanisms of a biological system, we have to work on the interpretability of our ML models, which are able to learn the underlying patterns in our data. Interpretability means, for example, to understand which features play the most important role in predicting the outcome of an observation or which combination of features lead to a certain outcome. The advantage of parametric and simple non-parametric models like logistic regression or decision trees, respectively, is that their interpretability is straightforward: the importance of a predictor variable j directly corresponds to its coefficient β_j in the linear model or to its position in the decision tree. However, for more complex models like RFs the interpretation is not as straightforward as for decision trees, because in RFs we have an ensemble of differently structured decision trees that are hard to analyse separately.

The visualization of single decision trees, implemented in different tools, only makes sense for RF models with small numbers of decision trees (Hänsch et al., 2015; Yang et al., 2012). Such tools visualize each decision tree separately as a 3D tree, where its structure and complexity is represented by its height and shape, e.g. a complex decision tree is represented by a tree with many branches. All trees are placed on local hills on a plateau, where the height of the hill indicates the performance of the respective decision tree, while the performance of the RF model is represented by the global height of the plateau. Correlations among decision trees are visualized by the distance and proximity of the trees on the plateau, i.e. if a tree is far away from another tree on the plateau the correlation between the decision trees is low. A major drawback of those visualization tools is that they become hard to interpret for larger collections of decision trees, because there are too many instances to plot and analyse.

Hence, alternative measures for the interpretation of bigger RF models had to be developed. In 2004, Adele Cutler and Leo Breiman introduced the RAFT tool, which focuses on the visualization of RF models using four different interpretation measures: variable importance, proximities, prototypes and interactions (Cutler et al., 2004). Below, those four interpretation measures are explained in more detail.

Variable Importance. The variable importance captures the contribution of a predictor variable to the prediction of the response to access its importance for the classification problem. Variable importance measures can be used for different purposes: 1) variable selection, find the minimum number of predictor variables that are sufficient to correctly classify the response variable and 2) interpretation, find the most important predictor variables that highly contribute to the classification of the response variable. Generally, more important predictor variables are assumed to be closer to the root node, where they partition big parts of the observations. Hence, naive variable importance approaches quantify the frequency by which the predictor variable was used as split

point in the RF model or calculate for each predictor variable the average depth (relative distance to the root node) of the first split point across all decision trees in the model.

More elaborated approaches are the mean decrease in impurity (MDI) and the mean decrease in accuracy (MDA), which can also be applied to other ML methods. The MDI for predictor variable X_j measures the weighted reduction of impurity (using the Gini index) for all nodes m where X_j was used as splitting variable, averaged over all decision trees in the RF. The drawback of the MDI measure is that it is biased towards continuous predictor variables and variables with many categories because predictor variables that have more potential splitting points are more likely to produce a good split point by chance (Strobl et al., 2007). MDA, on the other hand, measures the difference in prediction accuracy of the model before and after permuting the predictor variable X_j. The random permutation of the values of X_j breaks up its original association to the response variable Y. The logic behind MDA is that if the original predictor X_j was linked to the response Y, the prediction accuracy of the model will drop substantially if the link between X_j and Y is broken through permutation. However, if X_j is unrelated to Y the random permutation of X_j should not affect the prediction accuracy of the model. To compute the MDA for a predictor variable X_j we calculate the prediction accuracy on the OOB observations for each decision tree in the model. Next, we randomly permute the values of X_j in the OOB observations and once more calculate the prediction accuracy of the OOB observations for each decision tree in the model. The difference between the prediction accuracy of the RF model with the original and the permuted predictor is then averaged over all decision trees in the model:

$$MDA(X_j) = \frac{1}{B}\sum_{b=1}^{B} MDA^b(X_j)$$

where $MDA^b(X_j) = L(y_b, \hat{y}_b) - L(y_b, \hat{y}_b^{\pi})$ with $L()$ being the prediction accuracy, $\hat{y}_b$ being the predicted values for the OOB observations in decision tree b on the original X_j and $\hat{y}_b^{\pi}$ after permuting X_j. MDA values close to or below zero indicate that the predictor variable does not contribute to or is even detrimental for the classification of the response variable, while positive MDA values are indicative for a relationship between predictor and response variables. The higher the decrease in prediction accuracy for a predictor variable, the stronger the relationship between predictor and response variables and therefore the more important the predictor variable for the classification problem. MDA was shown to be a more reliable indicator than MDI because it shows no preference for continuous predictor variables or variables with many categories (Strobl et al., 2007). In addition, MDA values can be computed separately for each class. There also exist extensions of the MDA measure where joint effects of predictor variables are captured by jointly permuting the values of the predictor variables across the OOB observations (Bureau et al., 2005).

Variable Importance has also been shown to be superior to univariate screening methods like Fisher's exact test because such methods consider each variable in isolation while variable importance can capture the interactions between predictor variables in the context of a decision tree. This in turn, enables the detection of important predictors that have little predictive power on their own but important interaction effects with other predictor variables (Lunetta et al., 2004). Nevertheless, one has to keep in mind that MDA is a measure of predictive - not causal - importance of the predictor variable.

Proximities. Proximities indicate which observations are close together in the eye of the RF model. To calculate the proximity for a pair of observations we count how often those two observations end up in the same terminal node across all decision trees and divide the count by the number of decision trees in the RF model. Proximities range between zero (observations never end up together in a terminal node) and one (observations always end up together in a terminal node) and are represented by a $n \times n$ matrix where each entry is the proximity between two observations. The proximity measure shows the connectivity between observations and can be used to identify structures in the data of a RF model, e.g. find outliers and misclassified observations or find groups of observations that are classified in the same way by the same set of rules.

Prototypes. A prototype is an artificial observation for class k that provides a condensed view on how the predictor variables relate to the classification. For each class we can calculate a small number of prototypes that are representative for that class. To find a prototype for class k, we find the n-nearest neighbors for each observation among all observations, where neighbors are defined via proximities. Next, we select the observation that has the largest number of class k observations among its neighbors. The prototype is calculated for each predictor variable separately by defining the median (continuous predictor variables) or most frequent value (categorical predictor variable) among those neighbors as the prototype of class k for a specific predictor variable. Depending on the homogeneity of each class, one or more prototypes should be calculated. While a small, homogeneous classes can be summarized by one prototype, more diverse classes will require more prototypes. The prototype represents a group of observations that are classified in a similar way by the RF model (due to the definition of neighbors via proximities) and indicate which features have on average high or low values for this group of observations.

Interactions. Interactions explore the relationship between predictor variables within the RF model. The interaction measure describes the influence of predictor variable X_i on predictor variable X_j in the RF model (e.g. caused by correlation of variable X_i with variable X_j). Two predictor variables X_i and X_j are defined to interact with each other if a split on X_i makes a nearby split on X_j systematically more or less possible. If a split on X_i has no influence on a nearby split on X_j, both variables are considered to be independent from each other, i.e. they do not interact. Variable interactions can be derived from the reduction in Gini impurity or a joint permutation test, based on the MDA (Kelly et al., 2012).

In the RAFT tool, variable importance and prototypes are visualized as profile plots with the predictor variables on the x-axis and the variable importance / prototype on the y-axis, while proximities and interactions are visualized as heatmaps with observations and predictor variables on both axis, respectively (Figure 3.8). Another tool called Random Forest INspEctor (ReFINE) visualizes proximities, interactions and prototypes in a slightly different way (e.g. prototypes are visualized in a scatterplot for all pairs of predictor variables) and adds a visualization for each decision tree in the model (Kuznetsova, 2014).

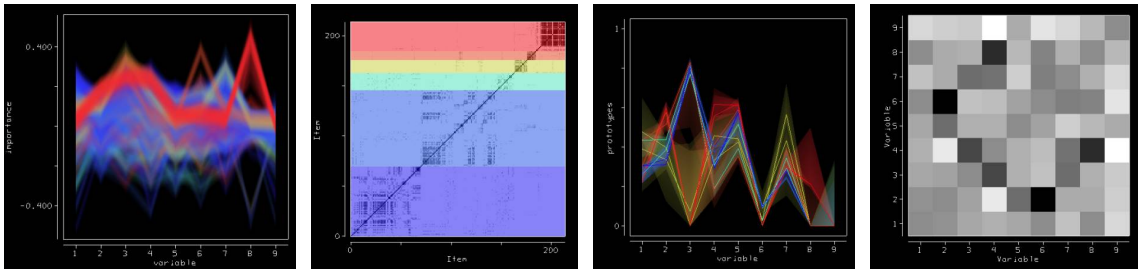

Figure 3.8: RAFT tool visualizations. **(A)** Visualization of variable importance, where each line represents the importance of a predictor variable for a specific observation in the training set (colors indicate class membership). **(B)** Proximities are visualized in a heatmap, where each entry represents the proximity for a pair of observations (colors indicate class membership, darker colors represent higher values). **(C)** Visualization of prototypes, where each line represents the prototype of a predictor variable for a specific observation in the training set (colors indicate class membership). **(D)** Interactions are visualized in a heatmap, where each entry represents interaction between a pair of observations (the darker the color, the stronger the interaction). Reprinted from (Cutler et al., 2004)

A different way to interpret the results of a RF model is to visualize the relationship between predictor and response variables through so-called partial dependence plots. Partial dependence plots decompose the high dimensional prediction function into a sequence of partial functions:

$$f(X) = f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)$$

which can be each represented in a two-dimensional plot (Figure 3.9), representing the relationship of one predictor variable to the response (implemented for instance in the package R package `edarf` (M. Jones et al., 2016)). Interactions between predictor variables, which were captured by the RF model, can be visualized in 3D plots of the two interacting predictor variables and the response variable. However, with increasing number of predictor variables it becomes hard to identify the pairs of interacting predictor variables or to capture latent interaction effects. The Forest Floor package extends the idea of partial dependence plots to identify latent interactions between predictor variables and applies a color gradient to each 2D partial dependence plot to identify and highlight interactions between predictor variables (Welling et al., 2016).

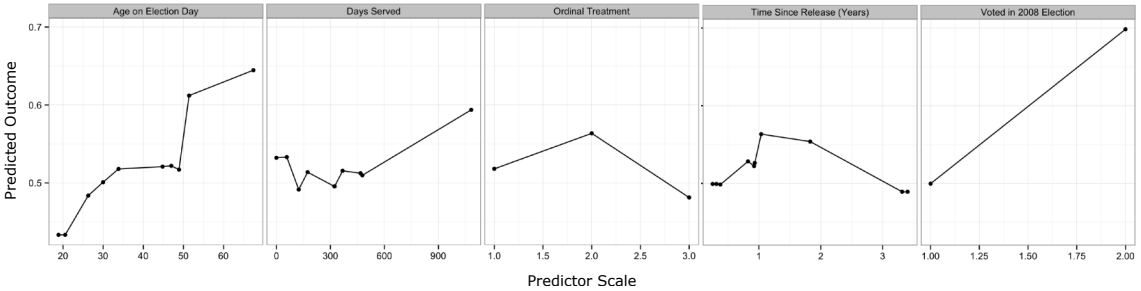

Figure 3.9: Partial dependence plots for RF visualizations. Example of a partial dependence plot for five different predictor variables, representing the relationship between predictor and response variable (x-axis: predictor variable; y-axis: predictions from a RF model). The plot for the Age on Election Day predictor variable, shows for instance that low values of the predictor variable are related to the negative class, while high values are related to the positive class. Reprinted from (M. Jones et al., 2016).

Advantages of using Random Forest Models

In molecular biology, we often have the case that different biological pathways with different interacting factors lead to the same outcome, e.g. a set of genes belonging to the positive class is transcriptionally activated by pathway A while the remaining genes of the positive class are transcriptionally activated by pathway B. Those combinatorial patterns are hard to detect by simple ML classifiers, like logistic regression models, which assume a linear relationship between predictor and response variable. Non-parametric ML learning algorithm, like RFs, on the other hand are able to approximate the relationship between predictor and response variables with arbitrary functional forms that allow to capture complex higher-order interaction effects and non-linear relationships between predictor and response variables. This in turn helps to uncover the combinatorial interaction patterns of biological systems. Due to the random subsampling in the splitting step of a RF model, they can be applied to problems where we have multicollinearity among predictor variables or more predictor variables than observations (curse of dimensionality). This is why RFs are also well suited for biological problems with large sets of correlating predictor variables but limited amount of samples like in large-scale association studies for complex genetic disease where each Single Nucleotide Polymorphism represeAlvareznts one predictor variable, but often only few biological samples are available (Bureau et al., 2005; Díaz-Uriarte et al., 2006; Lunetta et al., 2004). Additionally, RF has a variety of available interpretation measures that help us to not only make accurate predictions with the learned model but also to understand the logic behind those predictions. The underlying mechanisms that lead to a certain outcome can be detected by measures like feature importance or proximities, which point to the most informative predictor variables and help to find clusters of outcomes that are predicted by the same rules.

Better Practice for the Application of Random Forest Models

A RF model has two hyperparameters that have to be tuned during the training process: the number of decision trees in the model and the number of predictor variables that get randomly chosen at each split. As mentioned before, the generalization error of a RF model converges to an upper bound if the number of trees in the forest is large enough. Hence, the number of decision trees should be chosen as large as possible, limited by the available compute time, to improve the predictive power and avoid overfitting of the model. The number of randomly chosen predictor variables controls the amount of correlation between decision trees in the RF model. If we choose a value equal to the number of input features, the RF model reduces to Bagging on unpruned decision trees. As mentioned above, the generalization error of a RF model depends on the strength of each individual decision tree (bias) and the correlation between those decision trees (variance). By reducing the number of randomly selected features, we reduce the variance of the model but at the same time we increase the bias of each individual tree because we might not find the optimal predictor variable for each split. Hence, the number of randomly selected features is a tradeoff between bias and variance in the model and we can use the OOB error to find the best tradeoff for our model.

Another important issue is the training of a RF model on an unbalanced data set. Unbalanced data sets usually have a large proportion of observations belonging to one class but only a small

fraction belonging to the other class. Training a RF model on such a data set will result in large variations between prediction errors for the different classes because the algorithm will minimize the overall error rate but not the error rates for each class. Hence, if the algorithm classifies most observations according to the bigger class, the overall error rate will be small but the error rate for the small class might be very high, because most of its observations will be misclassified. To overcome this problem, two modifications to the RF algorithm exist: balanced RFs, where the bigger class is undersampled and weighted RFs, where higher weights are assigned to the smaller class when calculating the error rate.

3.3 UNSUPERVISED MACHINE LEARNING

Unsupervised machine learning (ML) methods analyse the underlying structures of the input data X, to get a better understanding of the relationship between input features or between observations, without having an associated outcome Y that can guide the analysis. The most prominent unsupervised ML approaches are ***clustering*** and ***dimensionality reduction*** which both try to simplify the input data with different strategies. Dimensionality reduction methods, like principal component analysis or non-negative matrix factorization, transform the input data from a high-dimensional into a low-dimensional representation through a small set of latent variables that are inferred from the predictor variables and explain a good fraction of the variance in the input data. Clustering on the other hand describes a group of methods which try to partition the input data into previously unknown homogeneous subgroups or so-called 'clusters'. The input data set can be clustered based on the p input features, to find subgroups among observations, or based on the n observations, to discover subgroups among features. Depending on the context, we can also cluster the input data on both, observations and features simultaneously. Hence, in the further description the term 'instances' as is used as a proxy for either observations, input features or both.

Clustering methods can be divided into ***hierarchical*** and ***non-hierarchical*** approaches. Hierarchical clustering methods split the input data set recursively into two subclusters based on pairwise dissimilarity measures until each cluster contains a single instance. In this thesis the focus is on non-hierarchical clustering methods which seek to partition the input data into a pre-specified number of distinct and non-overlapping clusters such that the instances within each subgroup are more similar to each other than to instances of other subgroups (Hastie et al., 2009; Izenman, 2008). In the final clustering, each instance belongs to at least one of the clusters but no instance belongs to more than one cluster.

To partition the input data into clusters, we first need to define what it means for two instances to be similar or dissimilar. The applied dissimilarity measure is computed from the input data and usually varies between different clustering methods, leading to quite different clustering results. Let x_{ij} be an instance of the input matrix with $i = 1, \dots, n$ observation on $j = 1, \dots, p$ input features, then the pairwise dissimilarities between observations i and i' is defined by:

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} d_j(x_{ij}, x_{i'j})$$

where $d_j(x_{ij}, x_{i'j})$ is the dissimilarity between the values of the j^{th} input feature. Popular dissimilarity measures for $d_j(x_{ij}, x_{i'j})$ are the squared Euclidean distance: $d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$ or Manhattan City-Block distance: $d_j(x_{ij}, x_{i'j}) = |x_{ij} - x_{i'j}|$. It should be noted that it is recommended to standardize the input data when the input features are on different scales and therefore the variability between features is quite high. One of the most popular non-hierarchical clustering methods is the *K-**means*** algorithm (MacQueen, 1967), where the squared Euclidean distance is chosen as dissimilarity measure, requiring quantitative input data. K stands for the total number of desired cluster and needs to be specified beforehand. After choosing K, each instance will be assigned to exactly one of the K clusters $C_1, \ldots, C_K$ such that the within-cluster variation across all clusters is minimized:

$$minimize_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

where the within-cluster variation is defined by the average squared Euclidean distance:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

and describes how much the instances in each cluster differ from each other, normalized by the total number of instances in the k^{th} cluster $|C_k|$. To solve this minimization problem the K-means algorithm (Figure 3.10, Algorithm 2) iteratively computes the centroids of each cluster (e.g. mean over input features for the observations of the k^{th} cluster) and then assigns each instance to the closest cluster (defined by the squared Euclidean distance).

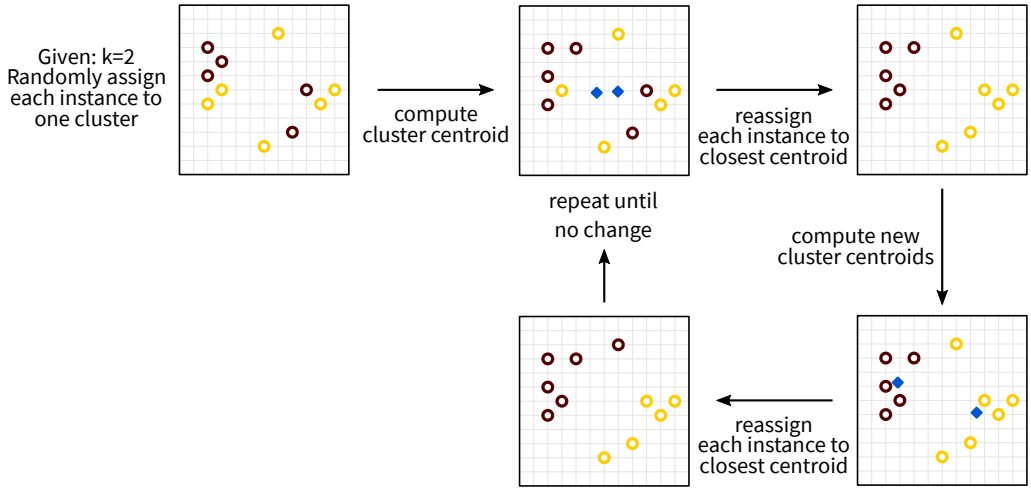

Figure 3.10: Steps in a K-means clustering algorithm. K-means clustering partitions unlabelled data into a predefined number (K) of clusters based on the computation of K cluster centroids and the assignment of each instance to the closest centroid.

Algorithm 2: K-means clustering

Input: data set $X = (X_1, X_2, \ldots, X_p)$,
number of clusters K

1) Randomly assign each instance to one of the K clusters and use as initial cluster assignment

2) Reassign instances

```
repeat
    for each cluster k:
        a) compute cluster centroid
        b) assign each instance to the closest centroid
until cluster assignment does not change;
```

The solution to the minimization problem obtained by the K-means algorithm will be a local optimum, depending on the initial random assignment of instances to the clusters. Hence, it is recommended to run the algorithm with different initial assignments and then choose the best solution based on the within-cluster variation of each clustering. In addition, the K-means algorithm lacks robustness against outliers because the squared Euclidean distance inflates the large distances generated by the outlier.

A generalization of the K-means clustering methods to arbitrary defined dissimilarity measures $D(x_i, x_{i'})$ is the K-***medoids*** method which is not restricted to the squared Euclidean distance (Vinod, 1969). Hence, k-medoids is robust against outliers and the input data is not required to be quantitative. In some settings the input data might be represented directly by a proximity matrix, a symmetric $n \times n$ similarity matrix with non-negative entries and ones on the diagonal. The K-medoids clustering method replaces centroids by medoids, which are a representative instance for each cluster, i.e. an actual observation. Hence, we do not need to explicitly compute cluster centroids from the original data. This in turn allows the application to input data that is only represented by a proximity matrix because we can transform the proximity matrix (P) into a distance matrix (D) with $D = 1 - P$. The PAM (Partitioning Around Medoids) algorithm is an implementation of the K-medoids clustering (Figure 3.11, Algorithm 3), where the medoids of each cluster are assigned initially at random and then replaced by another instance if the exchange reduces the dissimilarity between instances within the cluster (Kaufman et al., 1990).

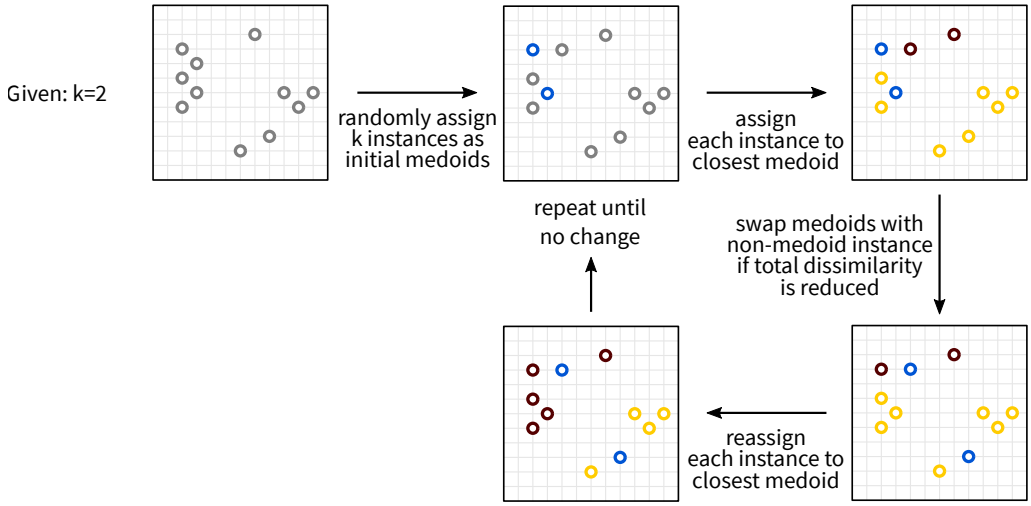

Figure 3.11: Steps of Partitioning Around Medoids clustering. PAM clustering partitions unlabelled data into a predefined number (K) of clusters based on the selection of K cluster medoids and the assignment of each instance to the closest medoid.

Algorithm 3: Partitioning Around Medoids clustering

Input: distance matrix $D(x_i, x_{i'})$,
number of clusters K

1) Randomly assign K instances as initial medoids

2) Reassign instances

1 **repeat**
2 **for each** cluster k:
3 a) assign each non-medoid instance to the nearest medoid
4 b) swap medoid with non-medoid instance if the total dissimilarity to all other instances in the cluster is reduced
5 **until** *cluster assignment does not change*;

Since K-medoids clustering can be applied on proximity matrices, we can use this clustering method to further interpret the results of a RF model. Clustering the observation of a RF model based on their proximities can help to find groups of observations that are classified by a similar set of rules and to determine the features that define the different groups of observations. A complete description of this approach and an application to *Xist* -mediated gene silencing during X-chromosome inactivation is given in Chapter 7.

4 STATE OF THE ART IN STUDYING AND MODELLING THE PROCESS OF XCI

The dynamics of X-Chromosome Inactivation (XCI) as well as the different properties of the inactive X chromosome (Xi) have been widely studied with different *in vivo* and *in vitro* mouse models. The prevailing *in vitro* system to study random XCI are female mouse embryonic stem cells (mESCs), which have two active X chromosomes and recapitulate random XCI when differentiation is triggered. mESCs are derived from the inner cell mass (ICM) of preimplantation embryos and are considered to be in ground/naïve state, which is characterised by the ability to form any cell type of the embryo proper upon differentiation (Vallot et al., 2016). Female trophoblast stem cells on the other hand, can be used to study imprinted XCI *in vitro*. Trophoblast stem cells are derived from the trophectoderm or extraembryonic ectoderm, where imprinted XCI with an inactivated paternal X chromosome is retained (Roberts et al., 2011). *In vivo* studies of random XCI usually analyse different mouse tissues, where differentiation and thus random XCI has already occurred, while *in vivo* studies of imprinted XCI follow early embryonic development from the 2-cell until blastocyst stage, where imprinted XCI is established.

4.1 *IN VIVO* AND *IN VITRO* MOUSE MODELS ARE USED TO IDENTIFY ESCAPEES

Several genome-wide studies have analysed gene silencing kinetics during random XCI and have identified genes that escape random XCI (Andergassen et al., 2017; Berletch et al., 2015; Borensztein et al., 2017; Calabrese et al., 2012; Marks et al., 2015; Splinter et al., 2011; Wu et al., 2014; Yang et al., 2010). Depending on the study design either *in vivo* or *in vitro* mouse models are used to identify escaping genes. *In vivo* studies use different types of mouse tissue with concluded random XCI (e.g. Brain, Ovary, Spleen) to analyse the tissue-dependency of escapees. In such studies, escapees can be identified with RNA fluorescent in situ hybridization (FISH) experiments, where the absence or presence of an RNA is measured through fluorescent probes that bind to candidate genes via sequence complementarity (Hosoi et al., 2018). *In vitro* studies differentiate mESCs into specific cell types to identify escapees or combine differentiating mESC models with time course experiments to follow gene silencing kinetics during the process of random XCI. During random XCI either the paternal or maternal X chromosome becomes silenced. Hence, to follow the silencing dynamics on the X chromosome or to determine the silencing status of X chromosomal genes, the silenced X chromosome has to be identified. Some studies use a *Hprt* mutation test to separate cells with a *Hprt* mutation on Xa, which become resistant to the toxic Purinbase 6-Thioguanin, from the remaining cells via flow sorting (Albertini, 2001), while other studies completely skew XCI towards a pre-specified X chromosome, i.e. through insertion of a stop signal in the *Xist* antisense transcript *Tsix* . Once the Xi is identified, allele-specific gene expression analysis can be used to determine the amount of expression coming from Xi and

Xa. Highly polymorphic hybrid mouse lines are used to measure allele-specific gene expression. Such hybrid mouse lines are derived from an intercrossing of two genetically distant mouse strains (typically *Mus musculus domesticus* (C57BL/6J) is crossbred with either *Mus musculus castaneus* (CAST/Ei) or *Mus spretus*), each with a distinct Single Nucleotide Polymorphism (SNP) pattern. Allele-specific RNA-seq exploits the strain-specific SNPs to measure allele-specific gene expression by counting the allele-specific reads that contain strain-specific SNPs (Babak et al., 2008; Lu et al., 2017). Allele-specific gene expression is then used to map gene expression back to the Xi and Xa in order to follow the silencing dynamics on the Xi or to identify silenced or escaping genes on the Xi (note: only genes containing strain-specific SNPs can be analysed with allele-specific RNA-seq). The majority of studies define silenced or escaping genes as genes, which have an allele-specific Xi expression $< 10\%$ or $\geq 10\%$ of the active X chromosome expression levels, respectively.

Few studies focus on imprinted XCI, where the paternal X chromosome is always inactivated and hence, the Xi is already known beforehand (Andergassen et al., 2017; Borensztein et al., 2017; Calabrese et al., 2012). Those studies identify escapees in extraembryonic tissue or follow early embryonic development with allele-specific RNA sequencing. Details for each study are listed in Table 4.1.

Table 4.1: Genome-wide XCI studies that identify escapees. The column # is used to reference each study in the text. The column study references the corresponding publication. The column *XCI* specifies if random or imprinted XCI was studied and if the study was conducted *in vivo* or *in vitro*. The column *Experiment* specifies what kind of experimental technique was used to identify escapees. The last column *Mouse model* specifies a) which mouse strains were crossbred, b) which cell types or tissues were analysed and c) which chromosome was silenced and the corresponding technique to select the inactive X chromosome.

#	Study	XCI	Experiment	Mouse model
1	(Yang et al., 2010)	random (*in vitro*)	allele-specific RNA-seq	a) *M. spretus* × C57BL/6 b) embryonic kidney cells c) Xi: C57BL/6 ⟶ *Hprt* mutation on *M. spretus*
2	(Splinter et al., 2011)	random (*in vitro*)	identification of new escapees via 4C interactions with known escapees and validation of new escapees via allele-specific RT-PCR on genes with SNPs	a) C57BL/6 x CAST/Ei b) mESCs differentiated into NPCs c) Xi: C57BL/6 or CAST/Ei ⟶ clone extraction
3	(Calabrese et al., 2012)	imprinted (*in vitro*)	allele-specific RNA-seq	a) C57BL/6 x CAST/Ei b) trophoblast stem cells c) Xi: paternal

4	(Wu et al., 2014)	random (*in vivo*)	allele-specific RNA-seq	a) C57BL/6 x CAST/Ei b) brain tissue c) Xi: C57BL/6 ⟶ *Hprt* mutation on CAST/Ei Xi: CAST/Ei ⟶ *Hpr*t mutation on C57BL/6
5	(Berletch et al., 2015)	random (*in vivo*)	allele-specific RNA-seq	a) *M. spretus* × C57BL/6 b) spleen, ovary and brain tissue c) Xi: *M. spretus* ⟶ *Xist* mutation on C57BL/6
6	(Marks et al., 2015)	random (*in vitro*)	allele-specific RNA-seq	a) *M. musculus* 129 × CAST/Ei b) mESCs differentiated into NPCs with time course data: 0, 2, 3, 4, 8 days after differentiation c) Xi: *M. musculus* 129 ⟶ clone extraction or stop signal in *Xist* antisense transcript *Tsix* on *M. musculus* 129 Xi: CAST/Ei → clone extraction
7	(Borensztein et al., 2017)	imprinted (*in vivo*)	allele-specific RNA-seq	a) C57BL/6 x CAST/Ei b) time course data: 2, 4, 8, 16, 32 cell stage and blastocyst c) Xi: paternal
8	(Andergassen et al., 2017)	imprinted random (*in vivo*)	allele-specific RNA-seq	a) FVB/NJ × CAST/Ei random XCI: b) 5 embryonic, 2 neonatal and 9 adult stage tissues c) Xi: FVB/NJ ⟶ bias in *Xist* expression imprinted XCI: b) 3 extraembryonic tissues c) Xi: paternal

4.2 *XIST* TRANSGENES HELP TO IDENTIFY SILENCING DETERMINANTS

Transgenic mESCs with an inducible *Xist* transgene allow for tightly controlled expression of *Xist* RNA. An inducible transgene can be integrated at different X chromosomal or autosomal locations. Such transgenic models can be used to access the silencing efficiency of *Xist* in autosomes, measured by the total number of silenced genes on the autosomes, in order to understand the dependency of *Xist* on certain genomic conditions (Russell, 1963). For instance, Loda et al. induced ectopic *Xist* expression from different X-linked and autosomal (chromosome 12) loci in mESCs to investigate the mechanisms of *Xist*-mediated gene silencing (Loda et al., 2017). They could show that an *Xist* transgene can recapitulate *Xist* function in an autosomal context. However, the silencing efficiency was lower on chromosome 12 compared to chromosome X, indicating that the silencing ability of *Xist* is independent of its genomic location, but the silencing efficiency of *Xist* is position dependent. Analysis of the silencing efficiency on chromosome 12 revealed for instance, that efficiently silenced genes, which are located far from the *Xist* integration site, are enriched for LINE elements around their transcription start site (TSS).

4.3 *XIST* MUTANTS HELP TO UNDERSTAND THE FUNCTION OF *XIST* REPEAT ELEMENTS

Several chromosome-wide studies used *Xist* mutants to analyse the importance of the different conserved *Xist* repeats-A to -F for the silencing function of *Xist* . Typically, *Xist* mutants have a mutated *Xist* allele that produces an *Xist* RNA, which is lacking one of its repeat elements (Wutz et al., 2002). *Xist* mutant studies can be used to analyse the impact of a specific repeat element on imprinted XCI, where the paternal *Xist* allele is mutated. Sakata et. al. studied the importance of the repeat-A element for the silencing function of *Xist* RNA during imprinted XCI in trophoblasts (Sakata et al., 2017). The mutated *Xist* allele was expressed from its endogenous locus and produced an *Xist* RNA that is lacking the repeat-A sequence. The mutated *Xist* RNA was able to coat the X chromosome in *cis* but failed to silence the majority of X-linked genes. Surprisingly, it still retained the silencing ability for a subset of X chromosomal genes, suggesting that not all X-linked genes are silenced via the *Xist* repeat-A element. Other mutant studies analysed the contribution of *Xist* repeats to random XCI, using an inducible mESCs system where the maternal or paternal *Xist* allele is mutated. Bousard et al. created several *Xist* mutants in mESC, lacking the repeats-A,-B,-C and -F (Bousard et al., 2019). The mutated *Xist* allele was mono-allelically upregulated from its endogenous *Xist* locus with a tetracycline-inducible *Xist* promoter (*Xist*-tetOP), which can be activated by doxycycline. For both repeat-A and repeat-BC mutants (*Xist* RNA lacking both, repeat-B and -C), the authors measured the extent to which transcriptional silencing could still be induced in both mutants via RNA-seq. Their analysis confirmed that the repeat-A element is important for the initiation of gene silencing and the repeat-B and -C elements are important for the interaction with Polycomb Group Proteins. In addition, they could show that repeat-B and -C elements are not necessary to initiate X-linked gene silencing but are rather important for the stabilization of the repressive state. Similar results were reported by a study from Nesterova

et al., where the repeat-A and repeat-BC mutants were created in mESCs carrying an inducible endogenous *Xist* allele (Nesterova et al., 2019). The study showed that repeat-A and its interacting partner SPEN is responsible for silencing the majority of X-linked genes, except a small fraction of weakly expressed genes, while a deletion of the repeat-BC element leads to a complete loss of *Xist*-dependent H2AK119ub and H3K27me3. Colognori et al. further investigated the role of the repeat-B element in female mouse embryonic fibroblasts, where the repeat-B element has been deleted from endogenous *Xist* (Colognori et al., 2019). They could show that *Xist* and polycomb complexes depend on each other and that the repeat-B element plays an important role in *Xist* spreading and the formation of the *Xist* cloud.

4.4 *IN SILICO* STUDIES TRY TO UNCOVER GENOMIC PROPERTIES THAT INFLUENCE *XIST*-MEDIATED GENE SILENCING

The previous sections described how silenced and escaping genes can be identified with different mouse models and how *Xist* transgenes and *Xist* mutants can be used to identify specific silencing factors that influence silencing dynamics on the X chromosome. However, the relative contribution of each silencing factor and their possibly combinatorial nature is hard to access with studies that focus on only a couple of specific factors or only on a set of certain X chromosomal genes. To overcome this problem, few *in silico* studies used different machine learning approaches to systematically analyse the combination and contribution of sequence features and chromatin states to *Xist*-mediated gene silencing on the X chromosome (Carrel et al., 2006; Nesterova et al., 2019; Wang et al., 2006).

Early studies used statistical approaches to systematically investigate the relationship between sequence features and X chromosome inactivation. Furthermore, machine learning models were used to predict the silencing status of X chromosomal genes from sequence features (Carrel et al., 2006; Wang et al., 2006). Both studies derived the silencing status of X chromosomal genes from human fibroblast-derived somatic cell hybrids containing one inactivated X chromosome. Genes were defined as escapee if they were expressed in at least 75% (Wang et al., 2006) or 8 out of 9 (Carrel et al., 2006) hybrids. Carrel et al. identified enriched oligomers in the neighborhood of silenced and escaping genes: 12-mers that have a 5-fold enrichment around the TSS of silenced vs escaping genes. The counts of overrepresented oligomers (in 50kb, 100kb and 250kb windows) around the TSSs of silenced and escaping genes were used as input features for a linear discriminant analysis (LDA) classifier to predict the silencing status of X chromosomal genes (silenced genes vs escapees). The classifier achieved a leave-one-out CV performance of > 90% accuracy (silenced: 93% and escapee: 90%) and test set performance of > 80% accuracy (silenced: 100% and escapee: 81%) for the 100kb window (best performance). However, an interpretation of the classification results was not possible, because the dimension of the input feature space (248 overrepresented oligomers) was reduced by principal components analysis to fit the number of genes in the training set (71 genes: 31 silenced and 40 escapees). Wang et al. systematically analysed the neighborhood of silenced and escaping genes by comparing the distributions of DNA

sequence features, including CpG islands, repeat counts as well as 3-mer and 5-mer sequence counts, in different windows (2kb up to 100kb) surrounding each gene TSS. Their analysis revealed that the most informative window sizes are larger windows (50kb and 100kb), which show the greatest difference between silenced and escaping genes for sequence features. Furthermore, they could show that long long interspersed nuclear elements (LINEs) and mammalian-wide interspersed repeat elements (MIRs) are significantly enriched around TSSs of silenced genes, while ALU repetitive elements as well as short motifs containing ACG/CGT are enriched around TSSs of escapees. In addition, a linear Support Vector Machine (SVM) classifier was trained on 110 X chromosomal genes with the described DNA sequence features (using only the 50kb and 100kb windows) as input features to predict the silencing status of X chromosomal genes. The SVM model correctly predicted 81% of the genes (silenced: 85% and escapee: 76%), measured with leave-one-out CV, and LINE-1 / LINE-2 elements were among the important classification features. However, both studies built their machine learning models only on a subset of X chromosomal genes. Those genes lie on the X-added region (XAR), located on the shorter arm of the human X chromosome, which contains about equal numbers of genes that are silenced and that escape X chromosome inactivation. Both models show considerably lower performance when predictions are made for all X chromosomal genes: the LDA classifier achieved an accuracy of 56% for silenced and 80% for escapees; the SVM classifier achieved an accuracy of 92% for silenced but only 17% for escapees. Furthermore, model predictions of both models were not validated experimentally.

A more recent study, used a machine learning model to identify key determinants of gene silencing efficiency during XCI (Nesterova et al., 2019). Gene silencing efficiencies of X chromosomal genes were measured with allele-specific Chromatin RNA-seq before (day 0) and after induction of *Xist* (day 1) in an inducible mESC system under differentiating conditions (129 × Cast mESC line (XX) with an *Xist*-tetOP):

$$SE = \left[\frac{Xi}{Xi + Xa}\right]_{day1} - \left[\frac{Xi}{Xi + Xa}\right]_{day0}$$

where Xi and Xa indicate expression from the inactive and active allele, respectively. Genes with a silencing efficiency $SE < -0.2$ were defined as highly silenced, while genes with $-0.05 < SE < -0.2$ were defined as lowly silenced. ChromHMM was used to assign one of 12 mESC chromatin states (e.g. heterochromatin, active promoter or gene body) to each gene promoter based on a 4kb window around the gene TSS. A Random Forest model was used to predict the silencing efficiency of X chromosomal genes (highly vs lowly silenced) based on the chromatin state of the gene promoter, the gene expression level at day 0 as well as the linear and 3D distance of the gene promoter to the *Xist* locus. Feature importance identified the linear and 3D distance to the *Xist* locus as well as active promoter and repressed chromatin states as the most informative features. However, the results are hard to interpret because chromatin states summarize several epigenetic marks in one state and do not provide a detailed view on the importance of specific epigenetic marks. In addition, the model performance was not striking with an AUC of 0.71.

The models described above show that the usage of machine learning models is a good approach to systematically analyse the impact of epigenetic or DNA sequence features on *Xist*-mediated gene silencing. However, all models show some drawbacks (e.g. lack of generalizability to all X-linked genes or difficulty to interpret the model results), which can be avoided by building an integrative machine learning model that includes as many epigenetic but also genomic features as

available and is trained on all, not just a subset of well-suited, X chromosomal genes. Such a model could help to understand the interplay between known silencing factors and potentially uncover new players in the XCI process - an important first step towards a system-level understanding of the XCI process.

5 QUANTIFICATION OF *XIST*-MEDIATED GENE SILENCING DYNAMICS

Gene silencing dynamics during the process of XCI are usually measured in differentiating mouse embryonic stem cells (mESCs) (see Section 4.1 for an overview on different mouse model systems). Such mouse models limit the temporal resolution of population measurements due to the asynchronous nature of the XCI process in differentiating mESCs. Here, an inducible mouse model system was used that allowed to overcome the asynchronous nature of XCI by inducing *Xist* expression through a doxycycline-inducible promoter from its endogenous locus. The inducible system was combined with an allele-specific Precision nuclear Run-On sequencing (PRO-seq) time course experiment to quantify chromosome-wide gene silencing dynamics with high temporal resolution. In comparison to mRNA-seq, which is commonly used to quantify gene silencing dynamics (see Section 4.1), PRO-seq measures gene expression at the level of nascent transcriptome, which allows a direct readout of gene silencing dynamics.

5.1 EXPERIMENTAL DATA

Changes in X chromosomal gene expression were measured after ectopic *Xist* induction in a ***time course experiment*** with dense temporal resolution. The female TX1072 mESC line was used for this experiment. This cell line is derived from a cross between the two mouse strains C57BL/6 (*B6*) x CAST/EiJ (*Cast*) and carries a ***doxycycline-inducible promoter*** in front of the *Xist* gene on the *B6* X chromosome that can be activated by doxycycline (dox) (Schulz et al., 2014). *Xist* up-regulation from the endogenous locus on the *B6* X chromosome was induced through dox treatment and ***allele-specific gene expression*** was measured before and at different time points after dox treatment with two different high-throughput techniques: PRO-seq and mRNA-seq (Figure 5.1).

PRO-seq experiment. *Xist* expression was induced by dox treatment in undifferentiated female TX1072 mESCs and the nascent transcriptome was measured by allele-specific PRO-seq at different time points up to 24 hours (h) of dox treatment (Section 2.1.4 for further details on the PRO-seq protocol). Samples were collected before (0h) and at time points 0.5, 1, 2, 4, 8, 12 and 24h after dox treatment. The gene expression on the B6 X chromosome decreases gradually over time, with most genes being silenced after 12h - 24h (Figure 5.2A). Samples before and after 24h of dox treatment were collected in duplicate to be able to assess reproducibility. The obtained allele-specific sequencing data is highly reproducible, since replicates generated for the first and last time point of the experiment (0h, 24h) are strongly correlated (Pearson correlation > 0.94; (Figure 5.2B). The PRO-seq procedure and allele-specific mapping of PRO-seq reads is described in more detail in the Appendix (Section A.1) as it was done by Laurene Syx (Heard Lab).

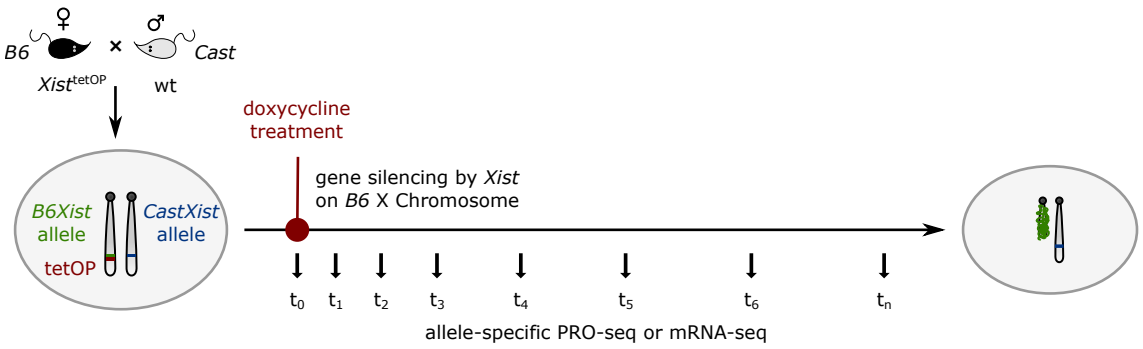

Figure 5.1: Measuring gene silencing dynamics. Schematic overview of the experimental setup used for PRO-seq and mRNA-seq experiments. The used hybrid female mESC line (*B6* × *CAST*) carried a doxycycline-inducible promoter (tetOP) in front of the endogenous *Xist* gene on the *B6* allele. *Xist* expression was induced through dox treatment and silencing kinetics were measured through allele-specific PRO-seq or mRNA-seq at different time points.

mRNA-seq experiment. Two additional data sets were generated to analyse the effects of ectopic XCI. Allele-specific mRNA-seq was performed in undifferentiated and differentiating female TX1072 mESCs before and after dox treatment (Section 2.1.4 for further details on the mRNA-seq protocol). For undifferentiated mESCs, allele-specific mRNA-seq data was collected before (0h) and at time points 2, 4, 8, 12 and 24h after dox treatment with two replicates for each time point. The majority of genes on the B6 X chromosome is silenced after 12h - 24h, as observed for the PRO-seq experiment (Figure 5.2A). For differentiating mESCs, allele-specific mRNA-seq data was collected before (0h) and at time points 8, 16, 24 and 48h after dox treatment with two replicates for each time point (Figure 5.2A). The mRNA-seq procedure and allele-specific mapping of mRNA-seq reads is described in more detail in the Appendix (Section A.1) as it was done by Laurene Syx (Heard Lab).

Pyrosequencing experiment. A pyrosequencing experiment was performed for validation of a few candidate genes (silenced genes and escapees) that were not captured in the PRO-seq experiment (Section 2.1.4 for further details on the pyrosequencing protocol). Samples were collected before dox treatment (0h) and at time points 4, 8, 12, 24h after dox treatment with 3 replicates for each time point. The pyrosequencing procedure is described in more detail in the Appendix (Section A.1).

All data described above was generated by our collaborators in the Labs of Edith Heard (Laurene Syx, Julie Chaumeil, Christel Picard, Chong-Jian Chen), John Lis (Iris Jonkers) and Edda Schulz (Ilona Dunkel). Edda Schulz provided the TX1072 mESCs line. Iris Jonkers performed the PRO-seq experiment and Chong-Jian Chen as well as Laurene Syx processed the allele-specific PRO-seq data. Julie Chaumeil and Christel Picard performed both mRNA-seq experiments and Laurene Syx processed the allele-specific mRNA-seq data. Ilona Dunkel performed the pyrosequencing experiment. All generated raw and processed sequencing data was submitted to the NCBI Gene Expression Omnibus (GEO; `http://www.ncbi.nlm.nih.gov/geo/`) under accession number GSE121144. The first quantitative analysis (Figure 5.2) as well as all subsequent analysis steps were done by Lisa Barros de Andrade e Sousa (Marsico Lab).

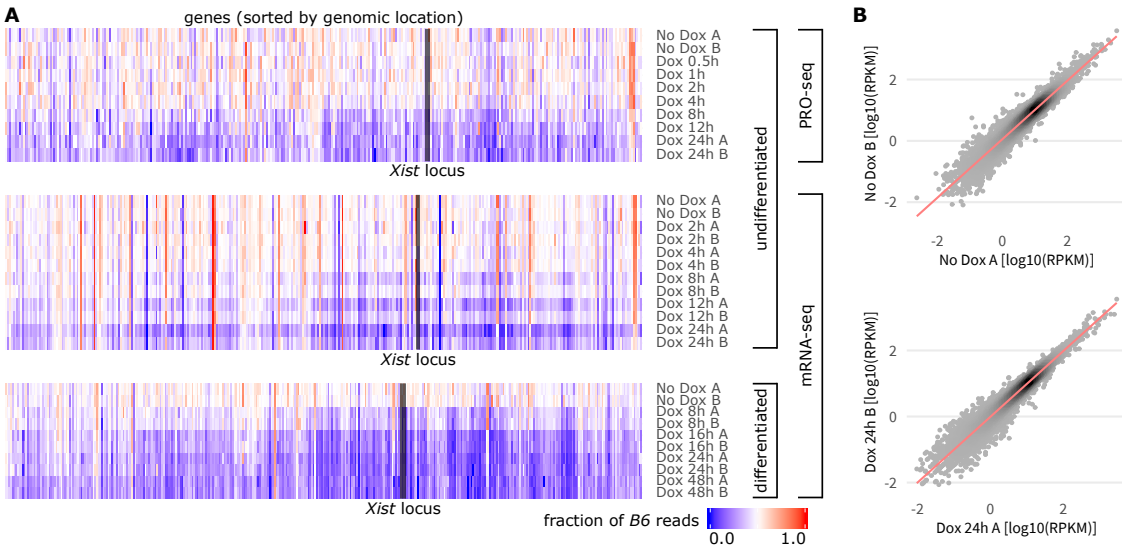

Figure 5.2: Allele-specific data obtained from time course experiments. **(A)** Comparison of PRO-seq (undifferentiated mESC, *upper panel*) and mRNA-seq data (undifferentiated mESCs, *middle panel*; differentiated mESCs, *lower panel*). Fraction of B6 reads are shown for all genes covered in all three data sets, ordered by genomic position. **(B)** Scatterplots of the log10 RPKM of all covered autosomal genes of no doxycycline sample A and B (*upper panel*) and doxycycline 24 hours sample A and B (*lower panel*). The data was highly reproducible, since replicates generated for the first and last time point of the experiment are strongly correlated.

5.2 SILENCING HALF-TIMES AS MEASURE OF SILENCING DYNAMICS

To quantify the changes in silencing dynamics of X chromosomal genes due to *Xist* activation, we estimated gene-specific ***silencing half-times*** that indicate the time point when transcription from the *B6* X chromosome is reduced by 50% compared to the uninduced control.

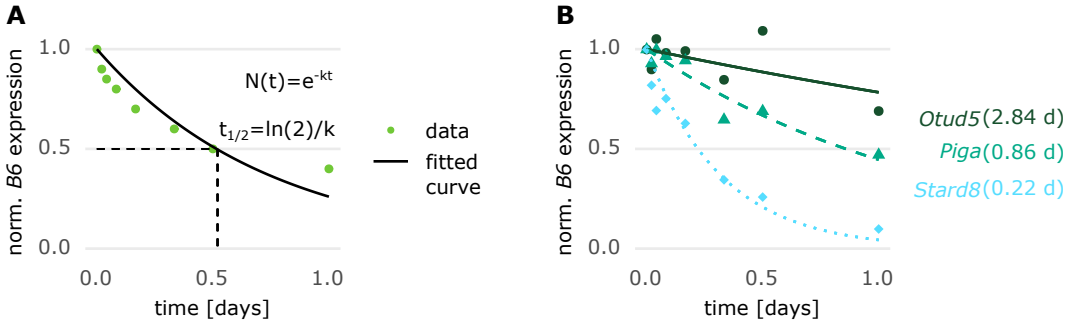

Figure 5.3: Computation of silencing half-times. **(A)** Schematic overview on how gene silencing half-times were estimated from allele-specific time course data through fitting of an exponential decay function. **(B)** Three examples of fitted gene silencing half-times (in parentheses) from the PRO-seq data set: one early silenced gene (*Stard8*), one late silenced gene (*Piga*) and one potential escapee (*Otud5*).

To calculate gene-specific silencing half-times, we normalized the allele-specific PRO-seq and mRNA-seq counts for sequencing depth, corrected for basal skewing (different transcriptional activity at the two active X chromosomes in the absence of dox) and fitted an exponential decay function to the time course data of each gene from which we estimated the gene-specific silencing half-time (Figure 5.3A), as described in detail in the next paragraph.

Out of 2610 genes annotated on chromosome X, 1630 genes had at least one SNP and could be used for allele-specific mapping. In order to confidently compute silencing dynamics using read coverage from all time points, we discarded genes without a minimum read coverage of > 10 reads for all time points (Table 5.1).

Table 5.1: Filtering steps in computation of gene half-times.

filtering step	# of genes after filtering		
	PRO-seq	mRNA-seq undiff.	mRNA-seq undiff.
genes annotated on Chromosome X		2610	
genes containing at least one SNP		1630	
minimum read coverage per timestep > 10	341	374	401
basal skewing between 0.2 and 0.8	330	353	379
sqrtRSS < 1.5	296	346	379
active TSS identified	280	320	349

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

$$f_{B6}^{t} = \frac{reads_{B6}^{t}}{reads_{B6}^{t} + reads_{CAST}^{t}}$$

For the PRO-seq data set, f_{B6}^{t} was averaged across replicates (0, 24h), resulting in a total of eight time points ($t = 0, 0.5, 1, 2, 4, 8, 12, 24$h). For the undifferentiated mRNA-seq data set we discarded replicate B due to insufficient read coverage and only used replicate A, resulting in a total of six time points ($t = 0, 2, 4, 8, 12, 24$h). For the differentiated mRNA-seq data set we averaged f_{B6}^{t} across replicates for each time point, resulting in a total of five time points ($t = 0, 8, 16, 24, 48$h). Genes with a strong ***basal skewing***, showing different transcriptional activity at the two active X chromosomes in the absence of dox, ($f_{B6}^{0} < 0.2$ or $f_{B6}^{0} > 0.8$) were removed from the data set (Table 5.1). The allelic ratio for the normalized counts was calculated as follows:

$$ratio^{t} = \frac{read_{B6}^{t}}{read_{CAST}^{t}} = \frac{f_{B6}^{t}}{1 - f_{B6}^{t}}$$

and ***normalized to the uninduced control*** ($t = 0$) to correct for basal skewing:

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

The expectation is that gene transcription gradually reduces over time due to *Xist*-mediated silencing. Exponential decay functions are typically used to model transcript decay over time (Lugowski et al., 2018; Rabani et al., 2011; Wada et al., 2017). The transcript abundance at a given time point t ($N(t)$) is defined by two parameters, the abundance at $t = 0$ (N_0) and the decay rate (k). Hence, the amount of remaining transcripts at time point t is described by the equation:

$$N(t) = N_0 \times e^{-kt}$$

In our case, k represents the silencing rate of a specific gene, the transcript abundance $N(t)$ is given by the normalized counts $norm^t$ and $N_0 = 1$ because the data is normalized to the uninduced control. The equation can be fit to the time course data using nonlinear least squares, which is implemented in the `nls` function of the `stats` R package. Once the ***exponential decay function*** is fit, the half-time of each gene can be calculated as:

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

A maximum value of $k = 5$ was defined (corresponding to a half-time of 3.5 days) as higher half-times cannot be reliably estimated from our data, due to the limited range of time points from 0 to 24h / 48h. The goodness of fit was evaluated via the square root of the sum of squared residuals *sqrtRSS*, defined as:

$$sqrtRSS = \sqrt{\sum_t (norm^t - N(t))^2}$$

Genes with sum of squared residuals $sqrtRSS > 1.5$ are indicative of a bad exponential fit and were discarded (Table 5.1).

The resulting silencing half-times range from 0 to 3.5 days. Genes with low silencing half-times are silenced very early during XCI process, whereas high silencing half-times are indicative of potential escapees (Figure 5.3B).

5.3 IDENTIFICATION OF ACTIVE TRANSCRIPTION START SITES

Since we wanted to associate silencing dynamics with promoter features in the downstream analysis (Chapter 6), we further filtered out genes, for which no ***active transcription start site (TSS)*** could be identified. To identify the active gene TSS that is used in embryonic stem cells, we annotated ***regulatory regions (RR)*** based on the PRO-seq data before dox treatment (0h) with the dREG tool (Core et al., 2014; Danko et al., 2015). RRs are defined as regions which harbor bidirectional transcription from the PRO-seq signal. Both replicates were analysed separately and de-novo RRs with a quality score of 0.8 or higher were selected. Those RRs are indicative of active TSSs and were used to assign each gene to its active TSS (Figure 5.4A with example in Figure 5.4B).

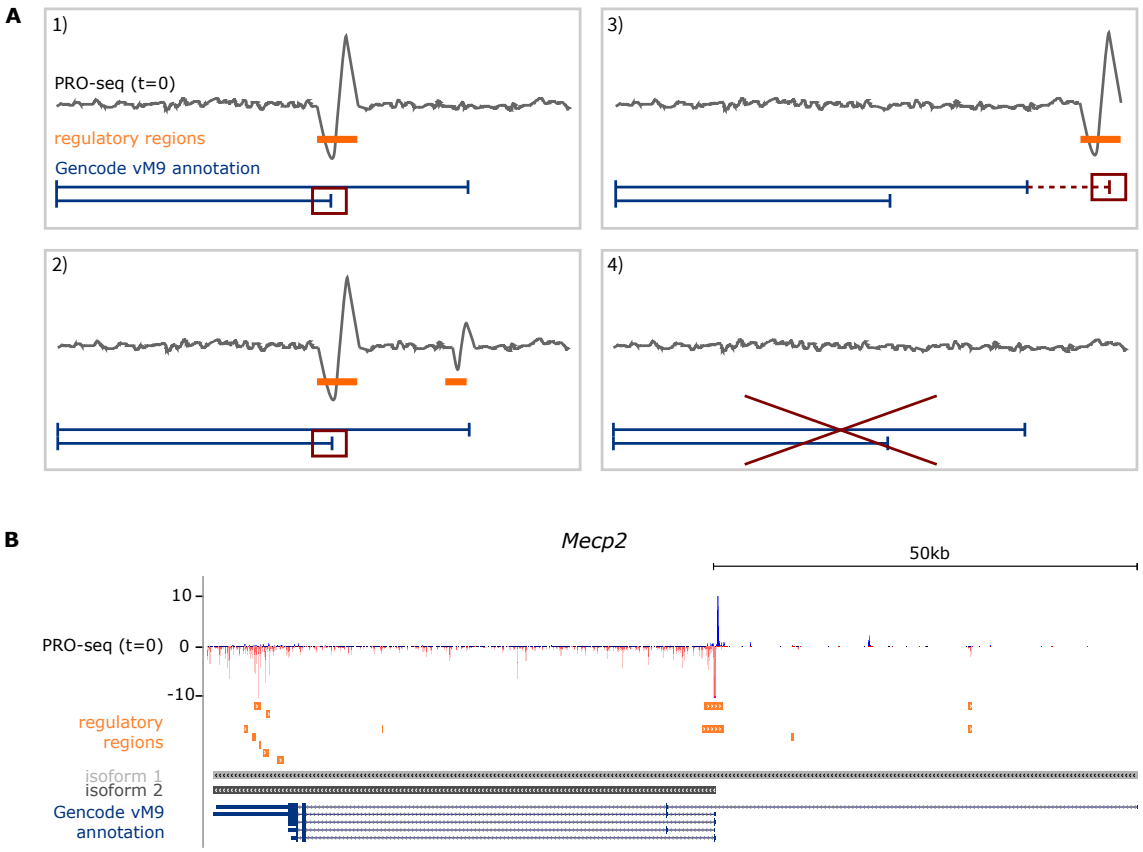

Figure 5.4: Assignment of X-linked gene to its active TSS. (A) GENCODE M9 gene annotation (blue) and annotated regulatory regions (orange) from the PRO-seq data (0h), identified with the dREG tool (Core et al., 2014; Danko et al., 2015) are used to assign each gene to its corresponding active TSS (red box, panel 1-3). Genes, for which no regulatory region can be identified within +/ – 1000bp around the annotated gene TSS, were discarded (panel 4). **(B)** As an example, the *Mecp2* gene on the (–) strand of chromosome X is shown. Its assigned active TSS is the one corresponding to isoform 2, as it overlaps a regulatory region defined by a bi-directional peak in the PRO-seq track.

Regulatory regions with overlapping genomic ranges between replicates were merged into one region. Most of the identified RRs overlapped known gene promoters. If a RR was found within +/ – 100bp of an annotated gene TSS, the TSS was chosen as active TSS for that gene (Figure 5.4A, panel 1). If multiple gene TSSs were found to overlap RRs, the TSS overlapping the RR with the strongest signal (i.e. highest score) was chosen for that gene (Figure 5.4A, panel 2). If no RR was found within +/ – 100bp of an annotated gene TSS, the genomic search space was extended to +/ – 1000bp. If a RR could be found within +/ – 1000bp of an annotated gene TSS, a novel alternative TSS, coincident with the middle point of the RR, was defined for that gene (Figure 5.4A, panel 3). If no RR could be found also within the enlarged region, the gene was discarded (Figure 5.4A, panel 4). For the PRO-seq data set, we filtered out 16 genes for which the active TSS could not be defined confidently, leading to a final data set of 280 genes with assigned active TSS on mouse genome mm10 for which we could estimate silencing half-times (see Supplemental Table S2 in Barros de Andrade E Sousa et al., 2019). For both mRNA-seq data sets, we filtered out 26 and 30 genes without active TSS (via the regulatory regions defined with the PRO-seq data set) for

undifferentiated and differentiated mESCs, respectively. The remaining genes were mapped to the mouse genome mm9 with the `liftOver` tool from UCSC Genome Browser (Kuhn et al., 2007)). For 233 genes, half-time could be estimated from all 3 data sets. The estimated half-times of all three data sets ranged from several hours up to several days (Figure 5.5).

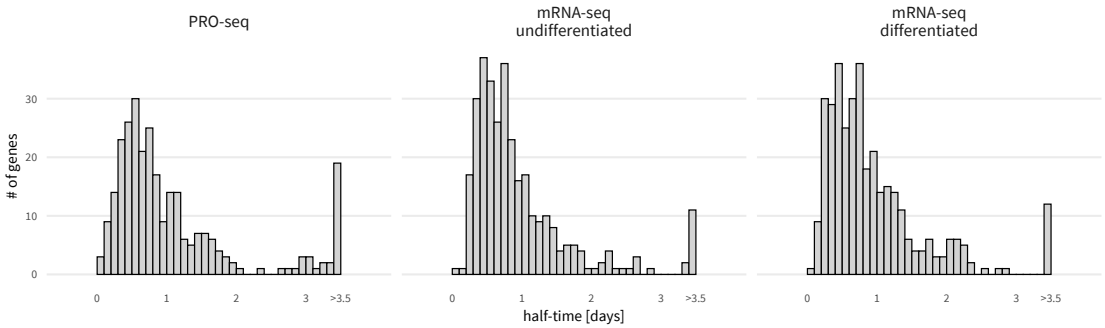

Figure 5.5: Distribution of silencing half-times. Distribution of computed half-times for X-linked genes of PRO-seq (undifferentiated mESCs) and mRNA-seq (undifferentiated and differentiated mESCs) experiments.

5.4 COMPARISON OF *IN VITRO* AND *IN VIVO* SILENCING DYNAMICS

Xist started to be upregulated from the *B6* X chromosome about 1h after dox treatment and reached a plateau after 4h (Figure 5.6A, Figure A2). Global gene expression on the *B6* X chromosome was gradually reduced over time, starting at 4h of dox treatment, while autosomal gene expression stayed constant over time, showing that only X-linked genes are affected by *Xist*-mediated gene silencing (Figure 5.6B).

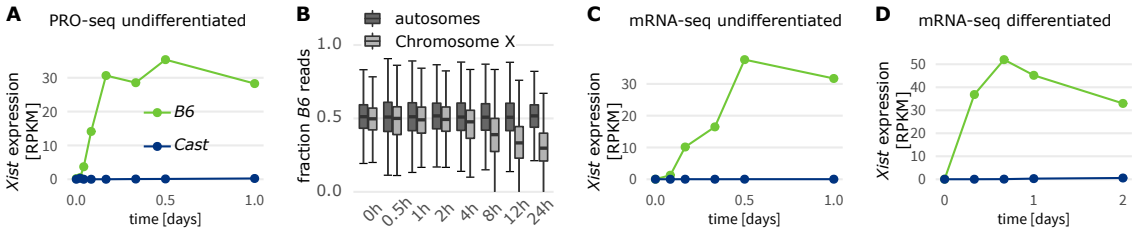

Figure 5.6: *Xist* expression over time. **(A)** *Xist* expression from the *B6* and *Cast* chromosomes for PRO-seq experiment over 24 hours time course. **(B)** Distribution of the fraction of *B6* reads for autosomal and X-linked genes over time in undifferentiated mESCs (PRO-seq). **(C-D)** *Xist* expression from the *B6* and *Cast* chromosomes in undifferentiated mESCs and differentiated mESCs measured by mRNA-seq over 24 and 48 hours time course, respectively.

To ensure that the relative silencing dynamics across genes, when XCI is induced in undifferentiated mESCs, are comparable to those in the cellular context where XCI occurs endogenously,

we generated two additional data sets, where ***allele-specific mRNA-seq*** was performed at different time points of a 24h and 48h dox treatment in ***undifferentiated and differentiating mESCs***, respectively (Section 5.1, Figure 5.6C-D). The computed half-times were comparable between these two data sets (Figure 5.7A, Pearson correlation coefficient: $r = 0.75$), suggesting that the differentiation process only has a minor impact on relative gene silencing dynamics. When comparing half-times estimated from the two different data types (mRNA-seq vs PRO-seq) correlation was generally a bit lower, independent of the cellular context (Figure 5.7B, Figure 5.7C, Pearson correlation coefficient: $r = 0.52$ and $r = 0.51$), which would be expected given that PRO-seq measures the direct transcription dynamics, whereas mRNA-seq kinetics are influenced by transcription, RNA-processing and degradation.

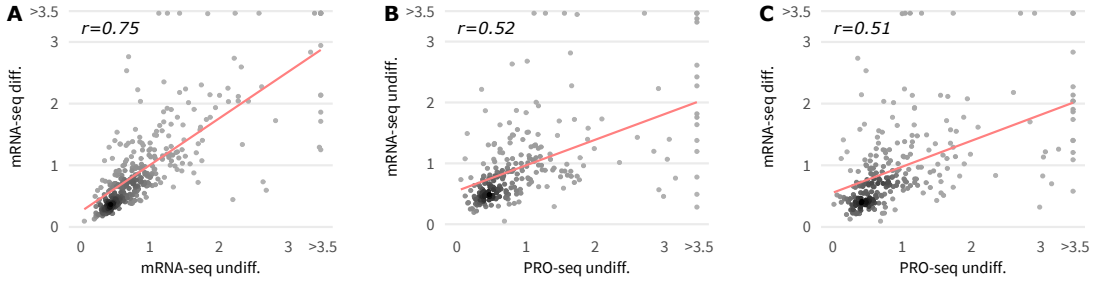

Figure 5.7: Comparison of PRO-seq-based silencing half-times to mRNA-seq data sets. (A-C) Comparison of estimated half-times (in days) between PRO-seq and mRNA-seq data sets with fitted regression lines (red). Pearson correlation coefficients are indicated.

In addition, we compared our estimated half-times to previous studies, which identified genes that escape random XCI *in vitro* and *in vivo* (Table 4.1 study #1, #2, #4, #5, #6). Genes that were identified as escapee in at least one study have significantly higher half-times than genes that were not identified as escapee (Wilcoxon Rank Sum Test: $p = 9.06 \times 10^{-6}$, Figure 5.8A), indicating that our estimated half-times recapitulate the findings of previous studies. Interestingly, ***constitutive escapees*** (genes identified as escapee in at least three samples/tissues) have higher half-times than ***facultative escapees*** (ANOVA Tukey Test: $p = 0.05$, Figure 5.8B), suggesting that facultative escapees are subject to XCI in our data set because they are specific to other cell types. To ensure that *in vitro* studies are comparable to *in vivo* studies, we compared the escapees identified in studies #1, #2 and #6, which were conducted *in vitro*, to those of studies #4 and #5, which were conducted *in vivo*. Most facultative escapees (> 90%) are only identified by either *in vitro* or *in vivo* studies while all constitutive escapees are identified by both types of studies, suggesting that the identification of escapees is not affected by the study type itself but rather by the cell-type specificity of certain escapees.

Furthermore, we compared our computed half-times to the silencing classes defined in study #6 (Marks et al., 2015), which used a dox-independent strategy to make XCI non-random. The different silencing classes in this study (early, intermediate, late, escapee) were defined based on time course data measured during differentiation of mESCs into Neural Progenitor Cells (NPCs). The defined silencing classes are in good agreement with the half-times estimated from the PRO-seq data (Figure 5.8C), which suggests that dox-induced XCI recapitulates endogenous gene

silencing dynamics. We also compared our *Xist*-induced gene silencing half-times in mESCs to the dynamics of imprinted XCI measured in pre-implantation mouse embryos of study #7 (Borensztein et al., 2017). The gene classification in that study (early: 16-cell stage; intermediate: 32-cell stage; late: blastocyst stage or escapee: still expressed in blastocyst stage) was once more in good agreement with the silencing half-times estimated from the PRO-seq data (Figure 5.8D).

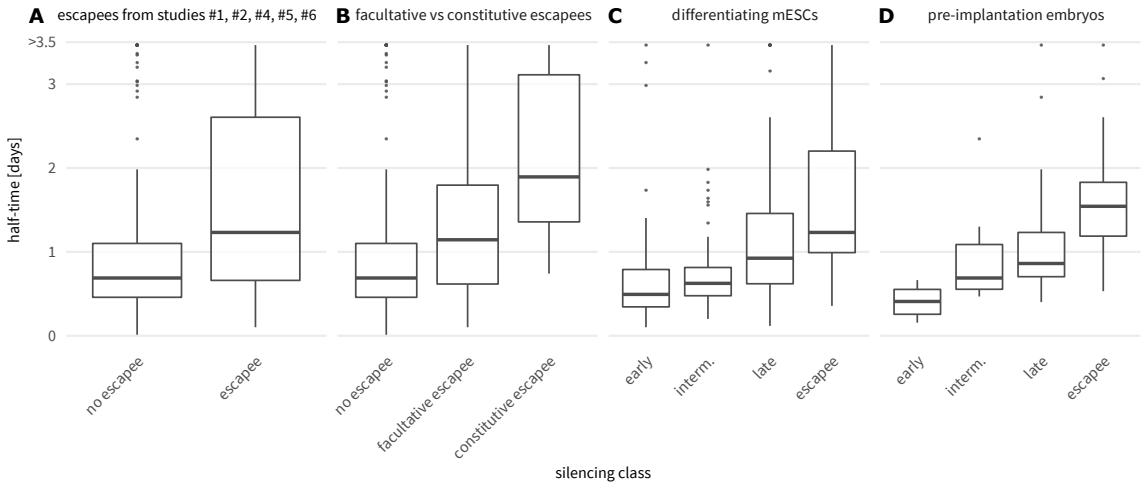

Figure 5.8: Comparison of PRO-seq-based silencing half-times to silencing classes of previous studies. **(A)** Distribution of computed silencing half-times for genes previously identified as non escapees and escapees (Table 4.1 studies #1, #2, #4, #5, #6) **(B)** Distribution of half-times for genes defined as constitutive escapees (identified in at least 3 tissues/samples) or facultative escapees, which are potentially cell-type specific escapees. **(C)** Distribution of half-times within silencing classes defined previously in differentiating mESCs (Table 4.1 study #6). **(D)** Distribution of half-times within silencing classes defined previously in pre-implantation mouse embryos (Table 4.1 study #7).

In conclusion, estimated half-times were comparable between undifferentiated and differentiated mESCs, suggesting that random XCI is mainly dependent on the silencing function of *Xist* and less dependent on the accompanying differentiation process. Furthermore, a comparison to previous studies on random XCI, suggests that our dox-inducible *in vitro* mouse model recapitulates endogenous gene silencing dynamics and is comparable to *in vivo* mouse models. However, in contrast to previous studies, we are able to follow silencing dynamics on the Xi with a high temporal resolution (resolution in hours vs resolution in days (Marks et al., 2015)), which gives us the possibility to not only distinguish between silenced genes and escapees but to also differentiate between genes that are silenced very early in the silencing process and genes that are silenced with slower kinetics. The computed silencing dynamics will be used in the next Chapter as input for a machine learning model to better understand the dependency of *Xist*-mediated silencing on epigenetic, genomic and DNA sequence factors.

6 MODELLING *XIST*-MEDIATED GENE SILENCING DYNAMICS

Chapter 6 explains and introduces the machine learning (ML) model used to predict gene silencing dynamics on the X chromosome. A schematic overview of the different steps required to build such an ML model is given in Figure 6.1. First, different epigenetic and genomic data sets were pre-processed and integrated into one feature matrix (Figure 6.1A, Section 6.1). Second, an ML model was trained on the integrated feature matrix (Section 6.2.1 and Section 6.2.2) and its predictions were validated with different experimental approaches (Figure 6.1B, Section 6.2.3). The last step, described in Chapter 7, focuses on the interpretation of the trained machine learning (Figure 6.1C).

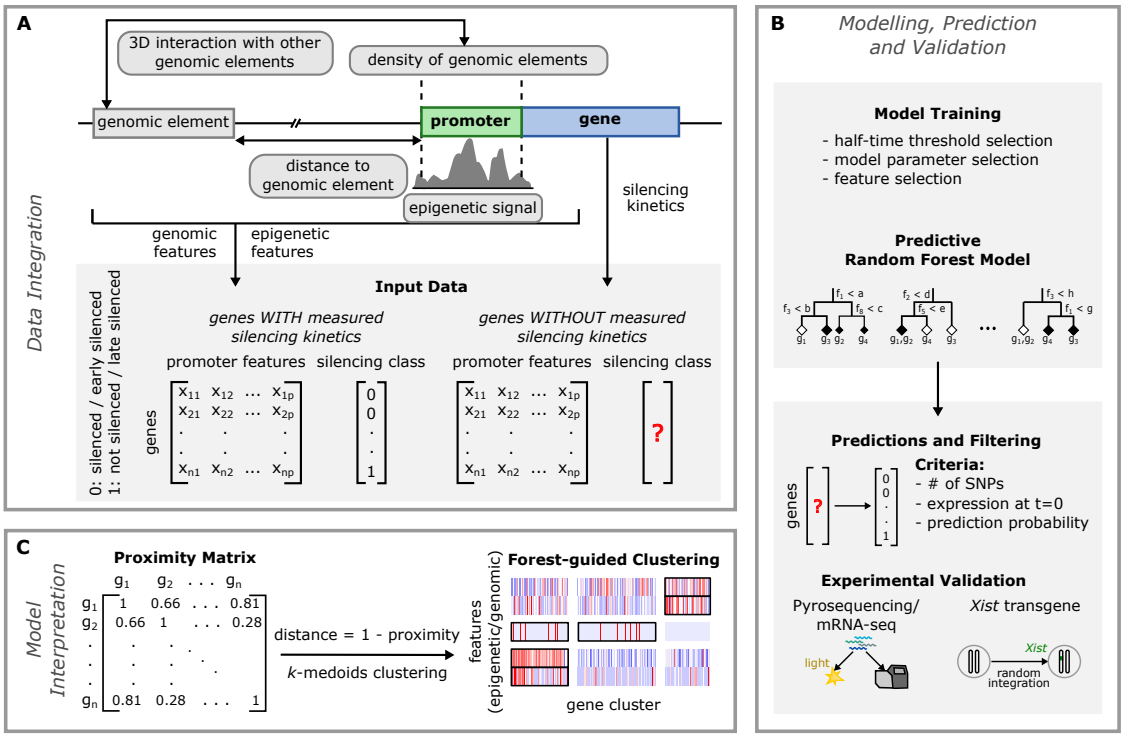

Figure 6.1: Schematic overview of our modeling approach. **(A)** Epigenetic and genomic as well DNA sequence feature input data for the model are collected, and feature matrices are computed for all X-linked genes with estimated half-times (labeled) and without estimated half-times (unlabeled). **(B)** After model training, the XCI/escape model is then used to predict the silencing class of all unlabeled X-linked genes given the same set of input features. The predictions are validated by comparing them to measured half-times from undifferentiated mRNA-seq data, with pyrosequencing experiments (few selected genes) and with measured silencing dynamics of genes in six transgenic mESCs clones. **(C)** A forest-guided clustering approach was developed for model interpretation. A proximity matrix between genes is computed from the trained model and converted into a distance matrix. Clusters of genes and their most significant associated features are displayed as a heatmap.

The code for the modelling pipelines and the different feature interpretation approaches is available on GitHub: `https://github.com/marsicoLab/xist_mediated_gene_silencing`.

6.1 FEATURE ENGINEERING

To analyse the impact of ***epigenetic and genomic factors*** as well as primary ***DNA sequence elements*** on the silencing status of X chromosomal genes, we collected a large amount of different data sets from various sources (Figure 6.1A). Epigenetic features were computed from high-throughput data sets (ChIP-seq and Bisulfite-seq) measuring genome-wide signals of chromatin modifications, chromatin modifiers, Transcription Factor (TF) binding and components of the transcriptional machinery. Since these data sets were generated in undifferentiated mESCs, they correspond to the chromatin state before *Xist* induction. In addition to the epigenetic features, we defined several genomic and structural features, such as gene density, the frequency of 3D chromatin interactions with different genomic elements and the linear distance to other genomic features, such as the distance to the *Xist* locus or the next full-length LINE element. DNA sequence features were represented by the distribution of k-mer sequences in a defined genomic window. A complete list of the features used in the model is given in (Table 6.1). Since we wanted to analyse the predictive power of epigenetic and genomic features from two different *cis*-regulatory elements - promoters and enhancers - we either used the active gene transcription start site (TSS) (see Section 5.3 for identification of active TSSs) or the center of the enhancer region as reference point for the computation of the different model features described below.

Table 6.1: Overview on different features used for modeling.

Epigenetic features	transcriptional regulators	CDK9, E2F1, HCFC1, MAX, MED1, MED12, NIPBL, RNAPII (unphosphorylated, S2p, S5p, S7p), SIN3A, TAF1, TAF3, TBP, MYC, ESRRB, KLF4, MAFK, NANOG, MYCN, OCT4, SOX2, TCF3, TCFCP2L1, YY1, ZNF384
	histone modifications	activation: H3K27ac, H3K9ac, H3K4me1, H3K4me3, H3K36me3, H3K79me2
		repression: H2AK119ub1, H3K27me3)
	DNA modifications	DNA methylation (BS-seq), 5fC (MeDip), 5hmC (MeDip
	chromatin remodelling complexes	RING1B (PRC1), CBX7 (PRC1) RYBP (PRC1), KMT6/EZH2 (PRC2), SUZ12 (PRC2), KMT2B/MLL2, KDM1A/LSD1, KDM2A, KDM2B, HDAC1, HDAC2, HDAC3, TET1
	structural proteins	CTCF, SMC1, SMC3
	others	H2A.Z, OGT, BRG1, CBX3
Genomic features	genomic elements	distance to the *Xist* locus, TAD borders, LADs, full-length LINEs
		overlap with *Xist* entry sites, LADs, CpG islands
		density of genes (1 Mb) and full-length LINE (700 kb)
		CpG content
	3D structure	number Hi-C all, Hi-C promoter, HiCap all, HiCap promoter, HiCap enhancer
		strength Hi-C all, Hi-C promoter, Hi-C *Xist*
DNA sequence features	density of 3-mers (100 kb), density of 5-mers (100 kb)	

6.1.1 Epigenetic Features

ChIP-seq experiments are the standard tool for genome-wide profiling of protein-DNA interactions, including TF binding, RNAPII occupancy and epigenetic modifications (see Section 2.1.4 for further description of ChIP-seq experiments). We downloaded a collection of 138 publicly available ***ChIP-seq libraries on undifferentiated mESCs*** with matching control libraries from the Gene Expression Omnibus (GEO) database (Edgar et al., 2002) (Table A1). ChIP-seq and control reads (downloaded as sra or fastq files) were aligned to the mouse genome mm9 with `Bowtie2` (with number of mismatches = 1) (Langmead et al., 2012). Obtained sam files were converted into bam files using `samtools`, only keeping alignments with a MAPQ score > 10 (Li et al., 2009).

Replicates were pooled for further analysis to obtain better coverage. All ChIP libraries containing less than three million uniquely mapped reads were removed from the collection, as the read coverage would be too sparse to infer robust ChIP signals.

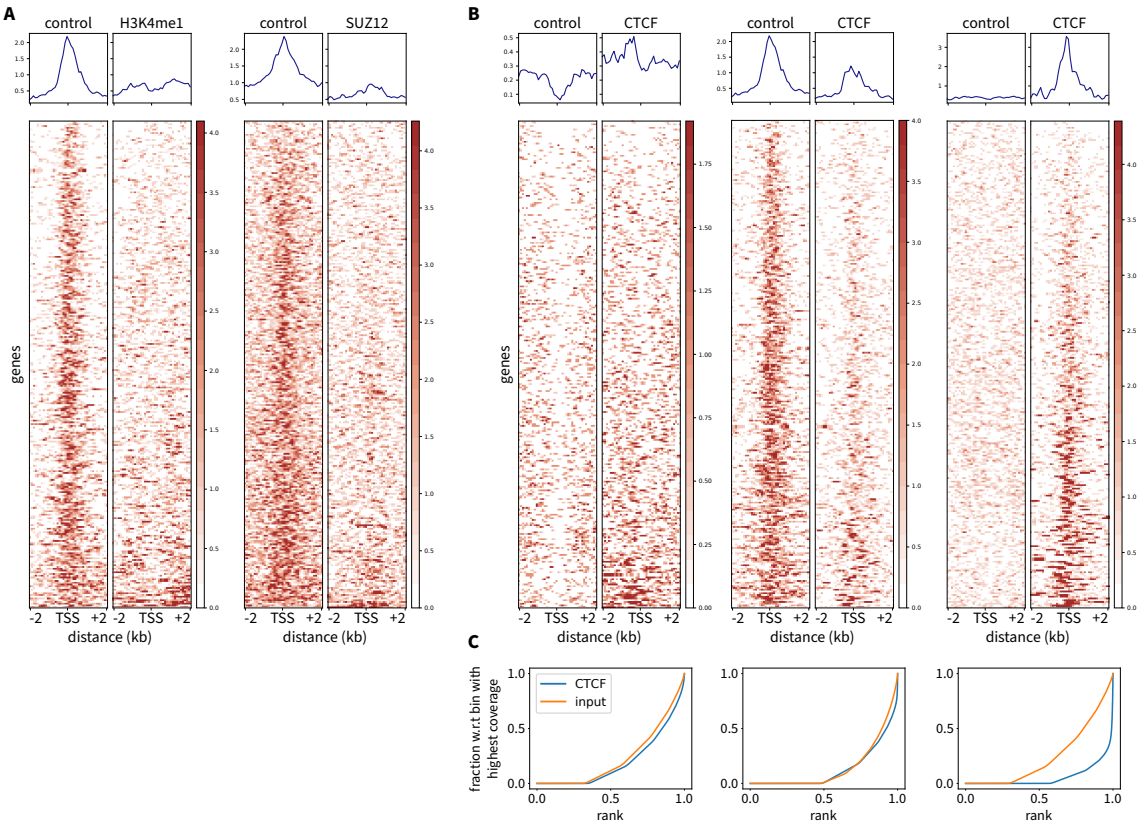

Figure 6.2: ChIP-seq library filtering with deepTools heatmap. **(A)** `DeepTools heatmaps` are visualized for two ChIP-seq experiments, H3K4me1 (GEO: GSE29184, left panel) and SUZ12 (GEO: GSE66830, right panel) and their respective controls. Shown is the ChIP-seq signal at the +/ − 2 kb region around the TSS of each gene (280 X chromosomal genes with computed half-times). For a ChIP-seq dataset on H3K4me1 or SUZ12 we expect an average signal enrichment (i.e. a peak) at the gene promoter for the experiment but not for the control. However, both experiments show a higher enrichment for the control signal in this region than the experiment itself (where very little signal is present). This evidence makes the quality of both data sets doubtful and therefore those libraries, as well as other data sets showing similar characteristics, were excluded from further analysis to avoid biases in the modelling process. **(B-C)** For each epigenetic feature only one GEO library is selected for further analysis based on `deepTools heatmaps` (B) and `fingerprint` plots (C). An example is shown for CTCF (from left to right: GSE25777, GSE29184, GSE28247), where the GSE28247 dataset is selected out of three libraries because: 1) the fingerprint plot shows that the cumulative distribution of the reads from the control experiment is closer to the diagonal, indicative of a uniform read distribution, compared to the other two libraries, 2) the fingerprint plot shows that the read distribution of the ChIP experiment has a steep rise towards the end of the plot, which is indicative of a peaked read distribution at CTCF binding sites, compared to the other two libraries and 3) the heatmap (top plots) clearly shows signal enrichment for CTCF at the −/ + 2 kb region around gene TSSs compared to control, indicative of a good signal to noise ratio, which is not the case for the GSE29184 and the GSE25777 libraries.

The obtained reads from a ChIP-seq experiment are usually a mixture of the actual ChIP signal reads and background noise reads, which originate from different sources of experimental biases. Hence, we had to ensure that the quality of each ChIP-seq data set was high enough to separate the ChIP signal from the background noise. The `deepTools` package (Ramírez et al., 2014) was used for ***quality control of the ChIP-seq data***: `fingerprint` plots, created with the `plotFingerprint` function and heatmap summary plots, created with `bamCoverage`, `computeMatrix` and `plotHeatmap` functions, were created for each ChIP and corresponding control library. `Fingerprint` plots (e.g. Figure 6.2C) produce a profile of cumulative read coverage from bins of specified size (bin size = 500) across the genome and allow to assess the antibody enrichment of the ChIP-seq experiment, i.e. is there sufficient antibody enrichment to separate the ChIP signal from the background noise. Heatmap plots (e.g. Figure 6.2B) produce an enrichment profile around a predefined region (e.g. promoters) from bins of specified size (bin size = 100) and allow to access the signal to noise ratio, i.e. does the observed signal correspond to the expected signal (based on prior knowledge). ChIP libraries were filtered by manual inspection based on the antibody enrichment and signal to noise ratio (Figure 6.2A). For some features more than one ChIP-seq library was downloaded, when experiments from different labs were available in GEO, and the most high-quality data set for each feature was chosen based on the signal to noise ratio (`deepTools heatmap`, Figure 6.2B) and the antibody enrichment (`deepTools fingerprint`, Figure 6.2C). After applying the aforementioned filtering steps we were left with 58 high-quality ChIP-seq libraries that were used for further analysis (Table A1).

After completion of all filtering steps (Table 6.2), we defined regions of enrichment based on the `deepTools heatmap`, which was used to show the average distribution of ChIP-seq signal around the promoter. Regions of enrichment can be short and localized in case of some TF binding or long and diffused in case of histone modifications. Hence, for each feature the width of the enrichment region was chosen according to the feature type and the observed enrichment in the `deepTools heatmap` plot at the reference point (promoter or enhancer). For instance, the enrichment region for promoter features was mainly located around the active gene TSS and only for few features, such as elongation marks (H3K36me3, RNAPII S2P, and H3K79me2), the signal was averaged over the entire gene body region (Table A1).

Table 6.2: Filtering steps in ChIP library preprocessing.

filtering step	# of ChIP libraries after filtering
number of downloaded data sets	138
remove ChIP libraries with < 3 Mio. reads	125
manual filtering with heatmap plots	84
selecting the best ChIP library for each feature	58

To remove experimental source of biases from the actual ChIP signal, we had to ***normalize the ChIP signal to a control sample***, i.e. a sequenced sample from either Input DNA (DNA isolated from cells that were treated similarly to the ChIP-seq sample) or IgG (enrichment of a non specific antibody). Several different normalization methods for ChIP-seq experiments were developed

over the past decades, one of the simplest methods being for example the log2-ratio of ChIP over control signal. Here, we use a more elaborated normalization approach, implemented in the R package `normR` (Helmuth et al., 2016; Kinkley et al., 2016). `normR` jointly models ChIP and control reads over the whole genome with a binomial m-component mixture model where one component models the background noise and the remaining m-1 components models the signal. In our case only a two-component model is used: one component to account for the background and one component to account for the ChIP signal. The fitted background component allows to inspect the enrichment in a certain genomic region and is used to compare ChIP read counts for that region to the expected read counts under the fitted background component (Figure A4). This model can then be used to calculate a normalized enrichment for each region, where the fold change of ChIP vs control read counts of each region is regularized (windows with zero counts get zero enrichment) and standardized (to values between zero and one, where zero means no enrichment and one means 100% enrichment), making read counts comparable between different ChIP experiments. We used `normR` to normalize each ChIP library to the corresponding control library in order to remove the background signal. For each of the remaining 58 ChIP-seq libraries ((after quality filtering steps, see Table 6.2)we used the normalized read counts in the specified regions as epigenetics features for the machine learning model.

In addition to the aforementioned ChIP-seq features, we computed the level of 5fC and 5hmC DNA methylation within 1000 bp around each gene promoter using published MeDIP data (for description see Section 2.1.4) in mESC (Pastor et al., 2011; Raiber et al., 2012). The MeDip data was processed in the same way as the ChIP-seq data using `normR` to normalize the MeDIP signal to the control signal. We furthermore computed the level of 5mC DNA methylation within 1000 bp around each gene promoter using published ***Whole Genome Bisulfite Sequencing (WGBS)*** data (for description see Section 2.1.4) in mESC (Stadler et al., 2011). For each C in a CG context, the total number of reads and the number of methylated reads is given, from which the percentage of methylation (# methylated reads / # total reads) can be computed. We computed the average methylation level over all CG sites within a 1000 bp region around each reference point as proxy for the promoter and enhancer methylation level (subsequently referred to as DNA methylation (WGBS)).

6.1.2 Genomic Features

In addition to the 59 epigenetic features (58 ChIP-seq, 1 WGBS feature), we collected 18 genomic and structural features, which are listed below.

Xist* distance**. We defined the linear distance of each reference point to the TSS of the *Xist* gene as a genomic feature: ***distance to Xist. The gene annotation for *Xist* was taken from GENCODE v. M9 on mm10 and lifted over to mm9.

***Xist* early sites**. Engreitz et al. defined the genomic coordinates of few early site (between 100 KB and 1 MB in size) on the X Chromosome, which have been identified as regions coated by *Xist* at an early stage of XCI, i.e. sites where *Xist* transfers itself from its transcription locus in order to initiate spreading across the X Chromosome (Engreitz et al., 2013). We define the overlap of

each X-linked gene / enhancer with these early sites as a genomic feature: ***overlap with* Xist *entry sites***. This feature is a dichotomic feature where a value of '1' indicates an overlap with an early entry site, while '0' indicates no overlap.

Genes. We defined the number of annotated genes within a 1Mb region around each reference point as a genomic feature: ***gene density***. Gene annotation was taken from GENCODE v. M9 on mm10 and lifted over to mm9.

LINE elements. Long interspersed nuclear elements are a type of repetitive DNA that is derived from transposons. We defined the distance of each reference point to the closest full-length LINE as a genomic feature: ***distance to LINE***. Furthermore, we defined the number of full-length LINEs within the 700 kb region around each reference point as a genomic feature: ***LINE density***. The genomic annotation of full-length LINEs in mESCs was taken from Penzkofer et al. (Penzkofer et al., 2017) and includes 1594 LINEs on Chromosome X. We downloaded the following data set of Penzkofer et al. from the L1Base (v2): Mouse Full-Length LINE-1 Elements [FLnI-L1] (Ens84.38) (14076 Entries, Last Update: 2016-09-27). LINE annotation was downloaded on mm10 and lifted over to mm9.

TADs. Topologically associated domains are chromatin units with a high frequency of long-range DNA interactions between loci within the same unit but with low interaction frequency between loci of adjacent units (Section 2.1.3). We defined the distance of each reference point to the border of the closest TAD as a genomic feature: ***distance to TAD border***. TADs are annotated from Hi-C data on mESC, which is taken from Dixon et al. (Dixon et al., 2012).

LADs. Lamina Associated Domains are genomic regions that interact with the nuclear lamina (Section 2.1.3). We defined the distance of each reference point to the closest LAD boundary as a genomic feature: ***distance to LAD***. Furthermore, we defined the occurrence of a LAD within a 1000 bp region around the reference point as a genomic feature: ***overlap with LADs***. This feature is dichotomic: a value of '1' indicates an overlap with a LAD, '0' indicates no overlap. The genomic annotation of LADs in mESCs was taken from Peric-Hupkes et al. (Peric-Hupkes et al., 2010).

3D interactions. Long-range 3D interaction between genomic loci can be measured with chromosome conformation capture (3C). While 3C only detects interactions for a specific region of interest, Hi-C experiments detect long-range DNA interactions across the entire genome. HiCap, on the other hand, combines Hi-C with sequence capture of promoter regions to identify promoter-anchored 3D chromatin interactions at high-resolution (Section 2.1.4). We obtained the number of 3D interactions and the average strength of 3D interactions (sum(Hi-C interactions strength) / number of interactions) of each gene TSS with other genomic elements from mESC Hi-C data (Schoenfelder et al., 2015). We defined separate features for all interactions (***number Hi-C all*** and ***strength Hi-C all***), interactions with other promoters only (***number Hi-C promoter and strength Hi-C promoter***) or with the *Xist* locus (***strength Hi-C* Xist**). In addition, we computed three features from HiCap data on mESCs (Sahlén et al., 2015): ***number HiCap all***, which corresponds to the total number of interactions of each gene's promoter / enhancer with other elements, such as other promoters or enhancer regions, averaged over two replicates;

number HiCap promoter, which corresponds to the number of interactions of each gene's promoter / enhancer with other promoters only; ***number HiCap enhancers***, which corresponds to the number of interactions of each gene's promoter / enhancer with enhancer elements only.

CpGs. CpG dinucleotides are cytosine nucleotides that are followed by a guanine nucleotide in the DNA sequence. The CpG content of a specific region is the amount of CpG dinucleotides within that region. We defined the normalized CpG content within the 1000 bp region around each reference point as a genomic feature: ***CpG content***. We computed the normalized CpG content as the ratio of observed over expected CG dinucleotides (Marsico et al., 2013):

$$\frac{\#GpGs/L}{((\#G + \#C)/2L)^2}$$

where L is the length of the considered region. CpG islands are regions of DNA (> 200 bp) in which the GC content exceeds 50%. We defined the occurrence of a CpG island within a 1000 bp region around each reference point as a genomic feature: ***overlap CpG island***. This feature is dichotomic: a value of '1' indicates an overlap with a CpG island, '0' indicates no overlap. CpG island annotation was taken from the UCSC genome browser (mm9) (Kuhn et al., 2007).

6.1.3 DNA Sequence Features

We choose to use 3-mer and 5-mer sequences to analyse the impact of the primary DNA sequence on the XCI process, because Wang et al. showed that the distributions of certain 3-mers and 5-mers are consistently different between silenced and escaping genes (Wang et al., 2006). We retrieved the distribution of all possible ***3-mer*** (64) and ***5-mer*** (1024) sequences within a 100 kb window surrounding each gene TSS (i.e. 50 kb upstream and 50 kb downstream of each gene TSS). We choose a large genomic window of 100 kb, because Wang et al. showed that the greatest sequence feature difference between silenced and escaping genes can be found in larger window sizes of 50 to 100 kb (Wang et al., 2006). We used `bedTools` to extract the DNA sequence from mouse genome mm9 within a 100 kb windows around each active gene TSS (see Section 5.3 for identification of active TSSs) and then computed the occurrences of each k-mer within the respective windows with the R package `kmer`. The k-mer counts for $k = 3$ and $k = 5$ were merged into one feature matrix, resulting in a feature matrix with 1088 (64 3-mers, 1024 5-mers) DNA sequence features for the 100 kb window.

In summary, we retrieved the distribution of 1088 primary DNA sequence features around each gene promoter. In addition, we computed the enrichment of 59 epigenetic features and defined 18 genomic and structural features for each gene promoter and enhancer from various sources of publicly available data sets. Those features will be used in the next sections as predictor variables for different machine learning models to predict the silencing status of X chromosomal genes from promoter- or enhancer-associated features.

6.2 PREDICTING GENE SILENCING DYNAMICS FROM PROMOTER-ASSOCIATED FEATURES

Promoters are *cis*-regulatory elements, which are located right upstream of the gene TSS and play a major role in transcriptional regulation of each gene in the genome. Promoters are usually bound by a variety of general and specific transcription factors and the adjacent nucleosomes typically show a specific modification pattern, which were shown to be predictive of the gene expression status (Karlić et al., 2010). Since promoters play such an important regulatory role, we analysed if the primary DNA sequence or a pre-marking of the ***promoter*** with certain epigenetic and genomic factors is predictive of gene silencing dynamics mediated by *Xist* .

6.2.1 A Linear Model Fails to Predict Gene Silencing Dynamics

Linear models are one of the simplest machine learning models that are easy to interpret and often outperform fancier nonlinear models, especially when the size of the training set is small. Since the assembled data for our modelling task includes only 280 observation (genes with measured silencing kinetics), we started with a simple linear regression model to model the continuous values of the gene silencing dynamics from promoter-associated features. Here, we only used the ***epigenetic and genomic feature matrix***, because the ***DNA sequence feature matrix*** contained many more predictor variables (1088) than observations (280), which made it unsuitable for linear regression models that suffer from the curse of dimensionality problem in $n << p$ situations.

We used the enrichment of the 59 epigenetic features at promoter level as well as the 18 computed genomic features for all 280 genes with measured half-times as input matrix for the ***linear regression model***. The input matrix was standardized by applying z-score transformation to every predictor variable, except the dichotomous predictors. The computed silencing half-times (Section 5.2) served as response variable in the linear regression model. We trained the linear model (with the `lm` function of `stats` R package) on 80% of the genes in our data set and left 20% of the genes as an independent test set to access the model performance. Unfortunately, the linear regression model had low predictive power with an average R^2 value of 0.2 across 200 independently trained linear regression models. This could be due to different sources of biases and imprecisions in the estimated half-times. For instance, we set a maximum value of 3.5 days for all silencing half-times (higher half-times could not be reliably estimated due to the limited time point range of 24 hours), whereby very late silenced genes and escapees get the same half-times. In addition, the exponential decay function that we used to fit the time course data, did not perfectly fit the data of all genes, introducing bias into the computed half-times.

To overcome those problems, we divided our gene set into different silencing classes to predict the silencing status of each gene instead of predicting the actual silencing half-time. We trained two logistic regression models to distinguish 1) ***silenced*** from ***not silenced*** genes, referred to as ***XCI/escape model*** and 2) ***early silenced*** from ***late silenced*** genes, referred to as ***silencing dynamics model***. The XCI/escape model is used to identify those factors that are important for

silencing in general, and the silencing dynamics model is used to find those factors that influence the kinetics of gene silencing. Therefore, the continuous half-time values were assigned to discrete classes according to fixed thresholds. Those thresholds were chosen from a range of different possible thresholds for each class (Table 6.3) and a grid search was performed to find the threshold combination that would minimize the misclassification error (as defined in Section 3.2.1). The ranges of possible thresholds were inferred from the distribution of silencing half-times computed from the PRO-seq experiment (Figure 5.5). Genes that did not fall into one of the two model classes were excluded from the analysis.

Table 6.3: Ranges of half-times for choosing class thresholds.

silencing class	half-time ranges
silenced genes	$t_{(1/2)} < [0.9, \dots, 1.4]$
not silenced genes	$t_{(1/2)} > [1.4, \dots, 2]$
early silenced genes	$t_{(1/2)} < [0.5, \dots, 0.7]$
late silenced genes	$[0.7, \dots, 1] < t_{(1/2)} < [1, \dots, 1.4]$

For the grid search, we trained a ***logistic regression model*** (with the `glm` function of the `glmnet` R package) for each threshold combination on a training set and assessed the model performance on an independent test set. We had a strong class imbalance for most threshold combinations, i.e. the class of silenced genes was on average three times larger than the class of not silenced genes, which is expected since most genes are silenced during XCI and only few genes are able to escape the process of XCI. Such a class imbalance can lead to high prediction errors for the smaller class, because the classifier mainly learns the properties of the bigger class and achieves a good average error rate by assigning most observations to the bigger class. A common method to overcome this problem is to downsample the bigger class in order to achieve a better class error rate for the smaller class. Hence, we downsampled the bigger class (e.g. silenced class) and used a balanced data set (containing the same number of genes ($n = size\ smaller\ class$) for both classes) as training set. We held out 20% of the genes for a test set and set the number of training genes to: $n_{training} = n \times 0.8$ for each class, where n is the size of the balanced data set. All remaining genes $n_{test} = n \times 0.2$ were used as independent test set. However, we required a minimum size for the test set of $n_{test} > 10$ and excluded threshold combinations where $n_{training} < 25$ for one of the two classes. For each threshold combination, we then calculated the misclassification error on the test set genes. We averaged the misclassification error over 200 trained models for the same threshold combination to obtain stable results. The misclassification errors for all threshold combinations are listed in Table A2 and Table A3. We identified several threshold combinations with similar minimum misclassification errors of around 44% and 45% for the XCI/escape and silencing dynamics model, respectively. The high misclassification error for both models shows that simply replacing the linear regression by a linear classification model does not help to improve the predictive power of the model. This could be due to the large number of predictor variables with variable degree of correlation among each other (Figure A5), of which some potentially introduce noise into the model.

To account for correlation among features and to remove noisy features from the model, we regularized the logistic regression model with ***Elastic Net*** to achieve both, feature selection and grouping of correlated features (see Section 3.2.1). We implemented the regularized logistic regression model with the `glm` function of the `glmnet` R package. We set the parameter α, which regulates the impact of both shrinkage penalties, to $\alpha = 0.5$ to have a balanced impact of feature selection and feature grouping. We trained a regularized logistic regression model for each threshold combination in the same way we trained the simple logistic regression model, adding an additional cross-validation step to tune the shrinkage parameters λ. Both models had several threshold combinations with similar minimum misclassification error of around 35% (see Table A2 and Table A3). Hence, using the Elastic Net regularization improved the model performance by ~ 10% for both models, probably by removing features that introduce noise into the model. Nevertheless, the predictive power of the regularized logistic regression model was still low, indicating that no single linear combination of features or rules could be defined to discriminate, for example, silenced from not silenced genes.

6.2.2 A Random Forest Model Can Predict Gene Silencing Dynamics

Previous studies have identified the linear distance to the *Xist* locus as a main determinant of gene silencing on the X chromosome (Marks et al., 2015; Nesterova et al., 2019). Yet, many genes do not follow this trend, as they are close to the Xist locus but escape XCI or are located in the distal regions of the X chromosome but are silenced early Figure A3. This indicates that the susceptibility of genes to *Xist* -mediated silencing is likely to be controlled by a complex combination of different features. In addition, it was shown that a subset of X chromosomal genes can be silenced independently of the *Xist* repeat-A element, while the majority of X chromosomal genes requires the repeat-A element for silencing (see Section 4.3 for details on *Xist* mutant experiments (Sakata et al., 2017)). To account for the potential combinatorial nature of silencing pathways, we switched from a linear to a non-linear classification model - a ***Random Forest (RF) model*** - to predict the silencing susceptibility of X chromosomal genes (Figure 6.1B). RFs are non-parametric classifiers which make use of multiple decision trees to learn non-linear classification tasks. The use of multiple trees makes the method robust to outliers and noise, and reduces the risk of overfitting, also with a small number of training examples (in $n << p$ situations), strong class imbalance and correlated features - properties which are all present in our data sets.

We implemented two RF models with the `randomForest` R package to distinguish silenced from not silenced genes (XCI/escape model) and early from late silenced genes (silencing dynamics model) based on the epigenetic and genomic feature matrix (EGm) but also based on the DNA sequence feature matrix (Sm). Therefore, we used the enrichment of the 59 epigenetic features at promoter level as well as the 18 computed genomic features as input matrix for the epigenetic and genomic feature model. For the DNA sequence feature model, we used the distribution of 1088 3-mer and 5-mer sequence features within a 100 kb window around each gene TSS as input matrix. In contrast to the input matrix of the linear models, we did not apply any standardization, e.g. z-score transformation, to the input matrix of the RF model, because RFs are invariant to scaling of input features and can handle continuous and categorical predictor variables. The

continuous half-time values were assigned to discrete classes, as described for the logistic regression model, and a grid search was performed to find the threshold combination that would minimize the OOB error of the RF model (see Section 3.2.2 for definition of OOB error). For the grid search, we trained a collection of 500 RF models for each threshold combination and calculated the average OOB error. We trained each RF model with a large number of decision trees (parameter `ntree = 1000`) to avoid overfitting of the RF model. The number of training observations for each decision tree (parameter `sampsize`) was chosen as described for the logistic regression model to train each decision tree on a balanced subset of the data, thereby avoiding a classification in favour of the larger class. The `mtry` parameter, which defines the number of randomly selected features at each node and represents the trade-off between bias and variance of the RF model, was optimized during training such that the OOB error of the RF model is minimized. The XCI/escape model based on the EGm had several threshold combinations with similar minimum OOB error rates of around 27%, while the XCI/escape model trained on the Sm had only two threshold combinations with similar minimum OOB error rates of around 31% (see Table A2 and Table A3). Hence, the RF model trained on the EGm performed slightly better than the RF model trained on the Sm. For both silencing dynamics models, trained on EGm and Sm, we found four threshold combinations with similar minimum OOB error rates of around 30% and 27%, respectively (see Table A2 and Table A3). In comparison to the XCI/escape model, the silencing dynamics model performed better on the Sm. A comparison of the linear classification model based on the EGm to the corresponding RF classification model, showed that the model performance improved by 15-17%, indicating that the process of XCI is of non-linear nature.

Regularizing the logistic regression model to perform feature selection led to a high performance increase due to the removal of weaker or redundant features, which potentially introduce noise into the model. Since we train one RF model on 77 epigenetic and genomic features and another RF model on 1088 sequence features, we checked whether the performance of the RF models could be improved through ***feature selection*** as well. RF provides an internal measure of variable importance (related to the relevance of each feature in the classification) that can be used for feature selection as proposed by Díaz-Uriarte and Alvarez (Díaz-Uriarte et al., 2006) and Genuer et al. (Genuer et al., 2008). They propose recursive feature elimination, where 20% of variables that have the smallest variable importance are recursively eliminated or sequential feature introduction, where the k most important variables are sequentially introduced into the model (Díaz-Uriarte et al., 2006; Genuer et al., 2008). The set of features leading to the smallest error rate is then selected for the final model. Here, we performed feature selection with a modified version of the sequential feature introduction approach. Instead of using the global feature importance measure, we used a class-wise feature importance measure to include important features for both classes. In order to identify the class-wise most important features, we computed the mean decrease in accuracy (MDA, see Section 3.2.2) of each feature. The MDA is computed for every feature in the model and for each class separately, as some features might contribute more to the prediction of one class than the other. Next, we ranked the features according to their MDA for each class separately and then computed the OOB error rate on the top x features from both classes (including only features with MDA > 0). x was optimized to obtain the feature set with minimal OOB error rate, which was then used to retrain the RF model (Algorithm 4).

Algorithm 4: Random Forest Feature Selection

Input: data set $X = (X_1, X_2, \dots, X_p)$,
silencing class $Y = (y_1, y_2, \dots, y_n)$,
feature importance from the RF model trained on all features MDA_i with $i = 1, .., p$

1) Compute OOB error for each set of top x features

`fi0` = all p features X_i sorted by their MDA_i for class 0;
`fi1` = all p features X_i sorted by their MDA_i for class 1;
`fi0, fi1` = remove feature X_i when $MDA_i < 0$ for `fi0, fi1`;
`p'` = number of features X_i with $MDA_i > 0$;
for *x=1 to p'* **do**
    `top_x_features = fi0[1 : x] ∪ fi1[1 : x]`;
    train RF model `RF(top_x_features,Y)` and save OOB error
return *table with OOB error rates for p' sets of top x features*

2) Choose optimal number of top x features

`optimal_x` = x which lead to minimal OOB error rate in step 1
retrain RF on `top_x_features`

To test whether feature selection based on the ***top x features*** performs significantly better than feature selection based on a Random subset of x input features, we computed an ***empirical p-value*** for each threshold combination. The empirical p-value was calculated via a bootstrap test by randomly sampling x features from the whole set of input features, where x is the number of top features that led to the best OOB error rate. We repeated the sampling 500 times and computed each time the ***expected OOB error rate (EER)*** for that run. The empirical p-value is then defined as:

$$p - value = sum(\#of\ times\ EER < CER)/500$$

where CER is the ***computed OOB error rate*** for the top x features that led to the best OOB error rate. The CER is considered to be significant if the empirical p-value is smaller than 0.05. For all threshold combinations of the XCI/escape model based on the EGm, we obtained empirical p-values smaller than 0.05, indicating that the set of top x features generally performs better than a random subset of x input features.

The model performance of the RF model trained on the EGm improved by 5-8%, resulting in an OOB error rate of 22% for both models. In comparison, a previously trained RF model on chromatin states only obtained an accuracy of 71% (see Section 4.4) (Nesterova et al., 2019). The best performing thresholds combinations of the RF models with feature selection were similar to those without feature selection. The RF models trained on the Sm benefited the most from feature selection, most likely due to the high number of input features. Feature selection improved the model performance by 13% and 11%, leading to an improved error rate of 18% and 16% for

the XCI/escape and silencing dynamics model, respectively. Yet, the best performing thresholds combinations of the RF models with feature selection were similar to those without feature selection. The OOB error rate for the silenced / early silenced class was on average only 3% lower than the OOB error rate for the not silenced / late silenced class. This is a great improvement compared to previous machine learning models trained on DNA sequence features, where the chromosome-wide classification was always in favour of one class, while the other class had a much lower accuracy (LDA classifier: 56% accuracy for silenced genes; SVM classifier: 17% accuracy for escapees; see Section 4.4) (Carrel et al., 2006; Wang et al., 2006).

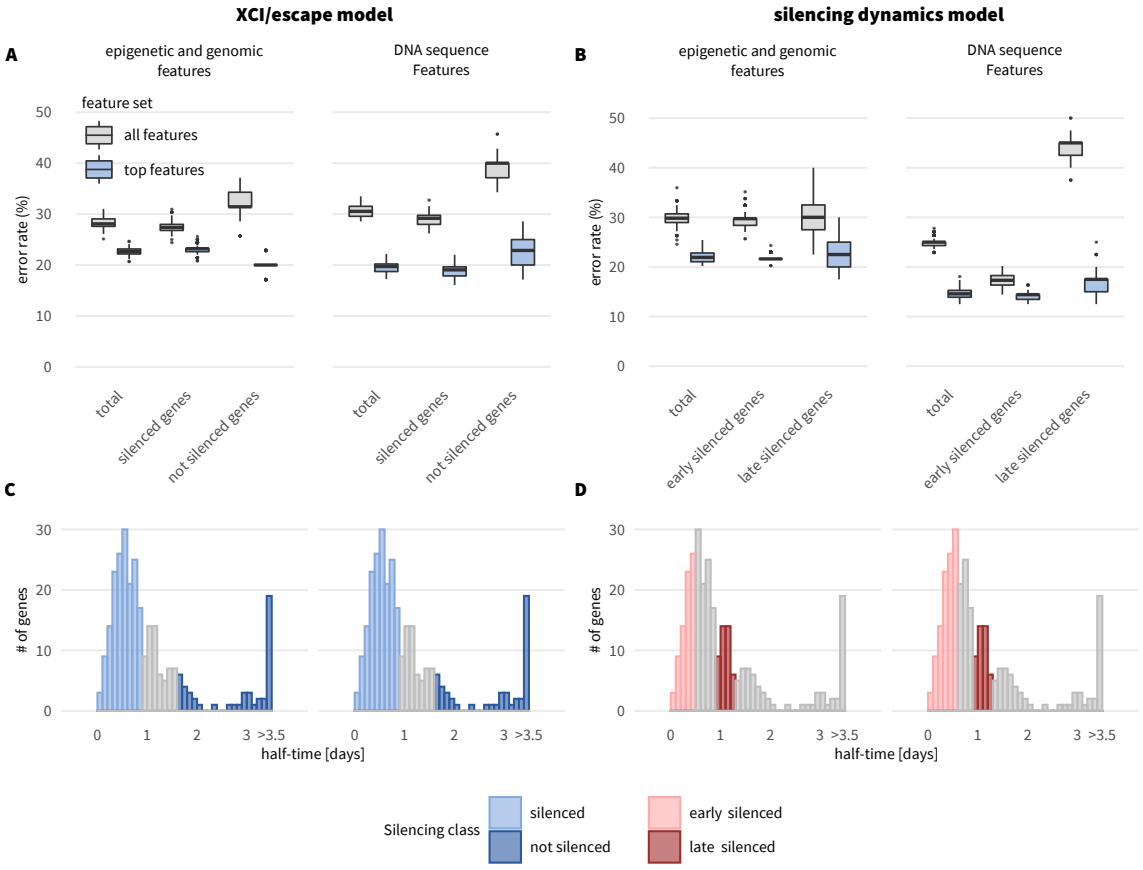

Figure 6.3: Model performance. **(A-B)** RF model performance measured with OOB error rate for the best performing thresholds of (A) the XCI/escape model and (B) the silencing dynamics model trained on the epigenetic and genomic or DNA sequence feature matrix. Each box in the plot represents the distribution of error rates over 500 trained RF models. Error rates are reported for both classes combined (total) and for the prediction of each individual class (silenced and not silenced class for the XCI/escape model, early and late silenced class for the silencing dynamics model). In addition, error rates are reported for models trained on the complete set of features (all 77 epigenetic and genomic features or all 1088 DNA sequence features) and models retrained on the top x features for feature selection (XCI/escape EGm: top 10 features; XCI/escape Sm: top 16 features; silencing dynamics EGm: top 8 features; silencing dynamics Sm: top 6 features). **(C-D)** Distribution of 280 X chromosomal genes with estimated half-times. The halftime ranges used to define the model classes for the best performing threshold combination are indicated for the both models, (C) XCI/escape and (D) silencing dynamics, and both feature matrices, epigenetic and genomic or DNA sequence feature matrix.

For further analysis (model validation and model interpretation), we selected the ***best performing threshold combination*** for each model, which is the threshold combination with smallest total OOB error rate that fulfills the following criterion: the absolute difference between the class errors should not exceed 3% (the average difference between class error rates): $|OOB_{class0} - OOB_{class1}| < 3$. For both XCI/escape models the best performing threshold combination was $t_{1/2} < 0.9$ for the silenced class (168 genes) and $t_{1/2} > 1.6$ for the not silenced class (50 genes). The model trained on the EGm and Sm had an OOB error rate of 22.44% and 19.01%, respectively, after feature selection was performed (EGm: top 10 features; Sm: top 16 features; Figure 6.3A). The best performing threshold combination of the silencing dynamics models had an OOB error rate of 21.56% and 15.57% for the EGm and Sm, respectively, after feature selection (EGm: top 8 features; Sm: top 6 features; Figure 6.3B). The early silenced class had a threshold of $t_{1/2} < 0.5$ and $t_{1/2} < 0.6$ (74 and 104 genes) for the EGm and the Sm, respectively, while the late silenced class had a threshold of $0.9 < t_{1/2} < 1.3$ (40 genes) for both, EGm and the Sm (Figure 6.3C-D).

Based on the PRO-seq derived silencing half-times and the trained RF models, we classified all genes according to whether they are subject to XCI or escape (silenced / not silenced) and whether they are silenced with slow or fast kinetics (early / late). Genes with intermediate half-times between the classes were excluded from the analysis (see gap between groups in Figure 6.3C). The resulting classes largely agree with those previously defined in differentiating mESCs, using a dox-independent strategy to make random XCI non random, and in pre-implantation embryos, where classes are defined based on the different cell stages during imprinted XCI (Figure 6.4 as example for EGm models).

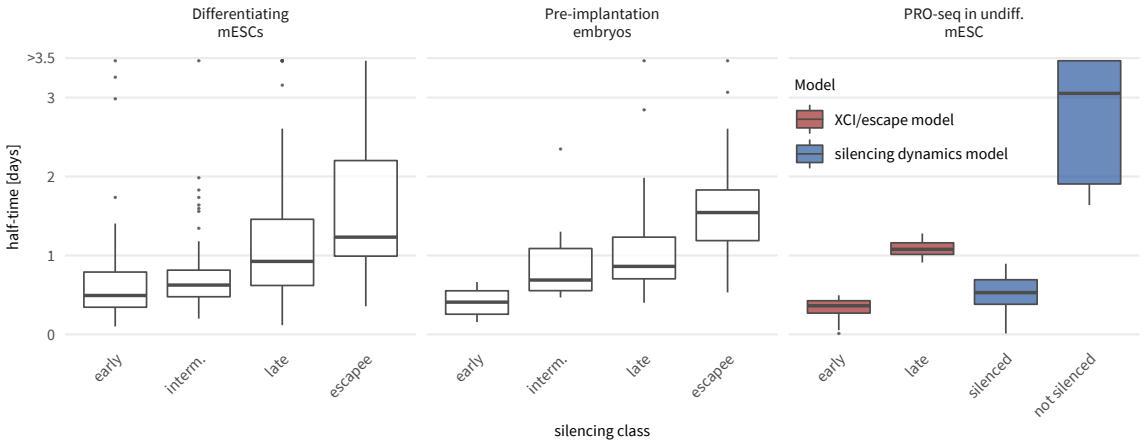

Figure 6.4: Comparison of silencing classes. Distribution of half-times within silencing classes defined previously in differentiating mESCs (Marks et al., 2015), in pre-implantation mouse embryos (Borensztein et al., 2017), and the classes used for RF model trained on epigenetic and genomic features: XCI/escape model (blue) and silencing dynamics model (red).

Using a RF model in combination with feature selection, instead of a regularized logistic regression model, improved the model performance of the XCI/escape and silencing dynamics model by 13% each. This indicates that no single linear combination of features but instead a combination of different silencing pathways is required to silence the genes on the X chromosome.

6.2.3 Validating Predictions from the Promoter-associated Random Forest Model

The trained promoter-associated RF models, distinguishing silenced from not silenced (XCI/escape model) and early from late silenced genes (silencing dynamics model), performed well on the 77 epigenetics and genomic features as well as 1088 DNA sequence features with OOB error rates between 16-22%. To confirm the model performances in an independent way, we further ***validated the XCI/escape model*** trained on the epigenetic and genomic feature matrix with three different validation approaches: 1) ***experimental validation*** of selected candidate genes, 2) comparison of predicted gene classes to half-times computed from ***mRNA-seq time course*** data (in undifferentiated mESCs) and 3) prediction of silencing kinetics in ***transgenic clones*** (Figure 6.5).

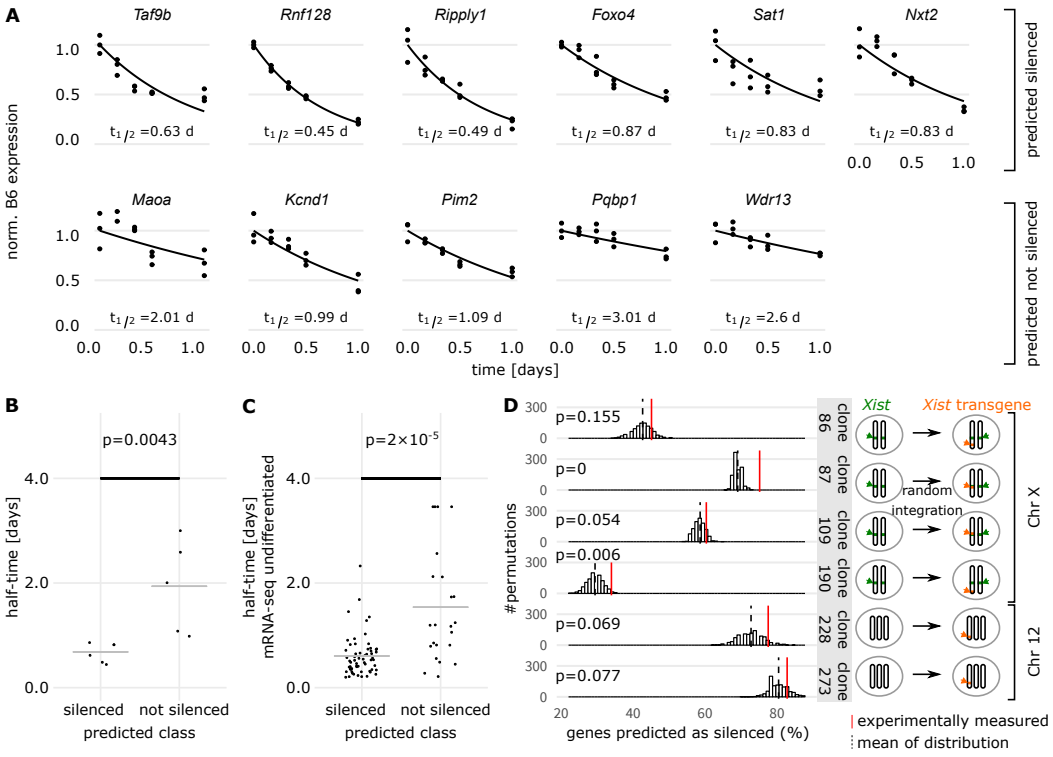

Figure 6.5: Experimental validation of model predictions. **(A)** Half-times of six candidate genes predicted as silenced (top) and five candidate genes predicted as not silenced (bottom) were validated experimentally. Allele-specific expression analysis was measured with pyrosequencing at different time points during 24 h of doxycycline treatment in TX1072 cells with three replicates per time point. Individual data points (dots), the fitted exponential decay function (line), and the estimated silencing half-times are shown. **(B)** Scatter plot of the silencing half-times estimated in (A). **(C)** Scatter plot of undifferentiated mRNA-seq half-times for genes predicted as silenced and not silenced by our XCI/escape model. The gray lines in B and C indicate the mean, and the p-value (Wilcoxon rank-sum test) indicates a significant difference between the mean of the two distributions. **(D)** Fraction of genes correctly predicted as silenced by the XCI/escape model (red lines) for six cell lines in which an inducible *Xist* transgene was integrated in different chromosomal locations (orange, cartoon on the right). The background distributions of silenced predictions was used to estimate empirical p-values (histogram, black dashed line represents the mean).

Each validation approach is described in more detail below. The results from all three validation steps confirm that our machine learning model can successfully predict the silencing status of X chromosomal genes based solely on promoter-associated features.

Validation with Pyrosequencing

We selected candidate genes for experimental validation among those X-linked genes that were not included in the training set of the XCI/escape model due to missing half-times, either because of insufficient read coverage from the PRO-seq data or because of a poor fit to the exponential decay function (see Section 5.2). Given our trained XCI/escape model, we predicted the silencing class for the respective genes based on their computed epigenetic and genomic features and then selected a few candidate genes for experimental validation according to the following criteria: 1) sufficient expression for experimental detection at time point 0 (PRO-seq RPKM > 1, based on non-allele specific mapping), 2) at least one SNP in exonic regions for allele-specific signal detection and 3) probability of a gene to be predicted in a certain class (silenced or not silenced) higher than 80%, averaged over 500 trained RF models.

For experimental validation of the selected candidate genes, an independent doxycycline (dox) induced time course experiment was performed and ***pyrosequencing*** was used to assess the allele-specific expression of 6 genes predicted to be silenced (Figure 6.5A, top) and 5 genes predicted to be not silenced (Figure 6.5A, bottom, for experimental details see Section 5.1). For each gene the signal fraction coming from the B6 X chromosome was reported before dox treatment and at four time points after dox treatment (t = 4, 8, 12, 24). We normalized the signal fraction fraction of each time point to the uninduced control and fitted the normalized signals to an exponential decay function to obtain silencing half-times for each gene (as described for the PRO-seq and mRNA-seq data in Section 5.2). The half-times of the silenced genes ranged from 0.45 to 0.87 days, and those of the not silenced genes ranged from 0.99 to 3.01 days. The half-time difference between silenced and not silenced genes was highly significant (Wilcoxon Rank Sum Test: p = 0.0043, Figure 6.5B). In addition, the half-times of all silenced genes fell in the silenced category ($t_{1/2}$ < 0.9), while the half-time of 3 out of 5 not silenced genes fell in the not silenced category ($t_{1/2}$ > 1.6).

Validation with mRNA-seq Data

To further validate the XCI/escape model, we compared the model predictions for all X-linked genes that were not included in the training set of the XCI/escape model (as described above) to the silencing half-times that were estimated from the ***mRNA-seq time course*** experiment in undifferentiated mESCs (see Section 5.2). Genes predicted as not silenced exhibited much longer silencing half-times than genes predicted as silenced (Wilcoxon Rank Sum Test: $p = 1.710^{-5}$, Figure 6.5C). In addition, we identified three genes (B630019k06Rik, Porcn, Ssr4) that have not previously been reported as escapees, but are predicted to be not silenced by the XCI/escape model and are also measured as not silenced in the mRNA-seq experiment. Those genes are potentially novel escapees, which were not identified in previous studies.

Validation with Different *Xist* Transgenes

Xist *transgenes*, which are activated in a doxycycline dependent manner, can be integrated in different X chromosomal or autosomal locations and are used to access the silencing efficiency of *Xist* in a different genomic context (see Section 4.2 for details on transgenic clones). We assessed the capability of our model to predict gene silencing efficiencies in such transgenic models, to verify the generalizability of our model. To this end, we made use of a study that performed allele-specific mRNA-seq before and after 5 days of *Xist* induction in transgenic mESC clones (GEO: GSE92894), that carry such a transgene in different positions on chromosome X or on chromosome 12 (Loda et al., 2017).

For our validation, we only used transgenic clones, where the genomic location of the *Xist* transgene is precisely reported in the paper and the experiment has been performed in undifferentiating mESCs (listed in Table 6.4). In total, we analysed 2 clones on chromosome 12 and 4 clones on chromosome X, where the *Xist* transgene is located in different positions with respect to its endogenous locus.

Table 6.4: Summary of clones used for validating the result from the XCI/escape model.

Clone	Chr	Allele	Karyotype	Integration site (mm9)
86	ChrX	*Cast*	diploid	chrX:130936613-131094303
87	ChrX	*Cast*	diploid	chrX:100655712-100678556
109	ChrX	*Cast*	diploid	chrX:100678562-100679597
190	ChrX	*Cast*	diploid	chrX:166414854-166443668
228	Chr12	*Cast*	diploid (duplicated chr12)	chr12:110315558-110351738
273	Chr12	*Cast*	diploid (duplicated chr12)	chr12:99721510-99727910

Loda et al. defined the allele-specific expression ratio (AER) for each gene i in clone j as:

$$AER_{ij} = \frac{reads_{Cast}}{reads_{Cast} + reads_{129}}$$

To define gene silencing kinetics in each transgenic clone *j*, we computed the normalized allelic expression ratio (AER^{norm}) as a measure of silencing kinetics for each gene *i* in each clone *j* (as we did for the PRO-seq data in Section 5.2):

$$AER_{ij}^{norm} = \frac{AR_{ij}^{5\ days}}{1 - AR_{ij}^{5\ days}} \times \frac{1 - AR_{ij}^{uninduced}}{AR_{ij}^{uninduced}}$$

Gene *i* was defined as silenced in clone *j* if $AER_{ij}^{norm} < 0.9$. Different cutoffs for defining silenced genes were tested and led to very similar results. Here, we used a cutoff of 0.9, a reasonable value for all clones.

To check whether our model correctly predicts gene silencing efficiencies in the different transgenic clones, we computed the epigenetic and genomic features for all genes of each transgenic

clone *j*. Therefore, we had to adapt the distance to *Xist* feature to the genomic location of each *Xist* transgene. We then applied our XCI/escape model trained on the PRO-seq data to predict the silencing class of all genes in each clone *j*. From the predictions, we defined the ratio of ***correctly predicted silenced (CPS) ratio*** genes as the fraction of genes predicted to be silenced by our model within all genes that were silenced in clone *j* (Figure 6.5D, red lines). These values varied considerably between clones depending on the size of the chromosomes and the location of the transgene. To test the significance of the obtained CPS ratio for each clone, we computed an ***empirical p-value*** with a bootstrap test by randomly sampling x_j genes, where x_j equals the number of silenced genes in clone *j*, from the background set of all genes detected in that clone. We repeated the sampling 1000 times and computed each time the ratio of ***expected predicted silenced genes (EPS) ratio*** for that run. The empirical p-value was then defined as:

$$p - value = sum(\#\ of\ times\ EPS\ ratio < CPS\ ratio)/1000$$

The CPS ratio is considered significant if the empirical p-value for clone *j* is smaller than 0.1. A significant empirical p-value indicates that our XCI/escape model is able to predict a proportion of silenced genes in clone *j* that is significantly higher than the random expectation. For 5 out of 6 clones the CPS ratio was significantly higher than expected for a random sample (Figure 6.5D, cp. red line to background distribution). Although potentially limited by the efficiency of the transgenes (Figure A6), this analysis shows that our model can, to some extent, be generalized even to other chromosomes.

The results of our three validation steps confirmed that our machine learning model can predict X chromosomal gene silencing based solely on epigenetic and genomic features. However, to understand the epigenetic and genomic mechanisms that govern *Xist*-mediated silencing it is also important to derive the biological meaning behind our trained RF model.

6.3 MODELLING GENE SILENCING DYNAMICS FROM ENHANCER-ASSOCIATED FEATURES

Transcriptional regulation is not only mediated by gene promoters, but also by other important *cis*-regulatory elements like enhancers. In contrast to promoters, which are located right upstream of the gene TSS, enhancers can be located far away from the regulated gene but are brought into spatial proximity to the gene promoter via 3D chromatin interactions. To get a more complete picture of the factors that influence *Xist* -mediated gene silencing, we analysed different ***enhancer*** features and their association with gene silencing dynamics. To address this question, we downloaded enhancer annotations in mESCs from a high-resolution, genome-wide mapping of promoter-enhancer and enhancer-enhancer interactions determined with the HiCap technique (Sahlén et al., 2015). HiCap enables the identification of promoter anchored 3D chromatin interactions by combining Hi-C with sequence capture of annotated promoters. Genomic regions that are enriched in either H3K27ac or DNA hypersensitive sites and are connected to promoters with a HiCap interaction that is supported by at least three reads in both replicates, were defined as enhancers by Sahlén et al. (Sahlén et al., 2015) and used for our analysis. The set of putative HiCap enhancers comprises 654 unique genomic regions on the X chromosome, where each putative

enhancer can interact with more than one gene promoter and each gene promoter can have more than one interacting enhancer. We were able to map 365 enhancers to 110 X chromosomal gene promoters (with computed half-times), with an average of 3.3 interacting enhancers per gene promoter (maximum: 19).

To analyse the association of enhancer features with different gene silencing dynamics, we computed the enrichment of our 59 epigenetic and 18 genomic features (Table 6.1) within the defined enhancer region, similarly to what we did with the promoter regions. If the length of an enhancer region was below 1000 bp we extended the region to +/ – 500 bp around the center of the enhancer. Next, we assigned each enhancer the silencing class of its associated gene (i.e. silenced or not silenced) and then used the assigned silencing class and the computed enhancer-associated features as input for a RF model. Unfortunately, the resulting RF model was biased, because some genes, which are linked to many enhancers (up to 19 enhancers per gene) were overrepresented in the training set, while others, which are linked to only one enhancers, were underrepresented in the model. Hence, the model results were not reliable. To overcome this problem, we had the idea to train an RF model on a selected subset of enhancers by choosing only one enhancer per gene, e.g. the enhancer with the strongest 3D interaction for each gene. Unfortunately, when choosing only one enhancer per gene, the training set for the RF model was not large enough to meet our minimum size requirements (defined in logistic regression model of Section 6.2.1) for any available threshold combination. Hence, we were not able to model gene silencing dynamics from enhancer-associated features. To still get some insights into the impact of enhancer features in the process of XCI, we performed a simple association analysis, described in Section 7.3.

7 IDENTIFYING THE MAIN DETERMINANTS OF *XIST*-MEDIATED GENE SILENCING DYNAMICS

"Interpretability is the degree to which a human can understand the cause of a decision."

— (Miller, 2019)

In Chapter 6 we showed that gene silencing dynamics on the X chromosome can be predicted with high accuracy from promoter-associated features, including epigenetic and genomic (called in the following EGm, see Section 6.1.1 and Section 6.1.2) as well as DNA sequence features (called in the following Sm, see Section 6.1.3). We built two binary classification models for each set of input features, to predict whether a gene is silenced or not (XCI/escape model), and whether it is silenced early or late (silencing dynamics model) using the RF classification described in Section 6.2.2. In this chapter we focus on the biological interpretation of the RF models trained in Chapter 6, starting with different commonly used interpretation techniques (introduced in the background Section 3.2.2), followed by a new approach to determine combinatorial feature patterns from RF models (see Figure 6.1).

7.1 RANDOM FOREST INTERPRETATION MEASURES IDENTIFY RELATIONSHIPS BETWEEN FEATURES AND MODEL CLASSES

In modelling, often only a small fraction of the collected input features is associated to the outcome. To better understand which features are associated with gene silencing in general (XCI/escape model) and which features are associated with the kinetics of gene silencing (silencing dynamics model), we identified, in a first step, the most ***important features*** among the large set of input features in a variable importance analysis (Section 3.2.2). In RF models, important features are often used as split points in the first node levels (close to the root node), where they partition big parts of the observations. Hence, to get a first impression on the importance of each feature, we obtained the frequency by which each feature was used as a split point (***split point frequencies***) in the first three node levels of all decision trees in the RF model. The obtained split point frequencies indicate that the linear distance of the gene TSS to the *Xist* locus and gene density around the gene TSS play an important role for silencing in general but also for the kinetics of gene silencing (Figure A7). Furthermore, the Polycomb Repressive Complexes (PRC) 1 mark RING1B, the histone mark H3K4me1, as well as the chromatin remodelling complexes HDAC2 and TET1 seem to be important for the classification of genes into silenced and not silenced, while the pluripotency factor SOX2, the CpG content around the gene TSS as well as the distance to the next TAD border and LINE element seem to be important for the classification of genes into

early and late silenced. The split point frequencies for the DNA sequence features do not give such a clear picture, probably because many k-mers are quite similar to each other and are used alternately as a split point in the first node levels (Figure A8). Nonetheless, an AT rich sequence context seems to play a more important role for silencing in general, while a GC rich context seems to play a more important role for the dynamics of gene silencing.

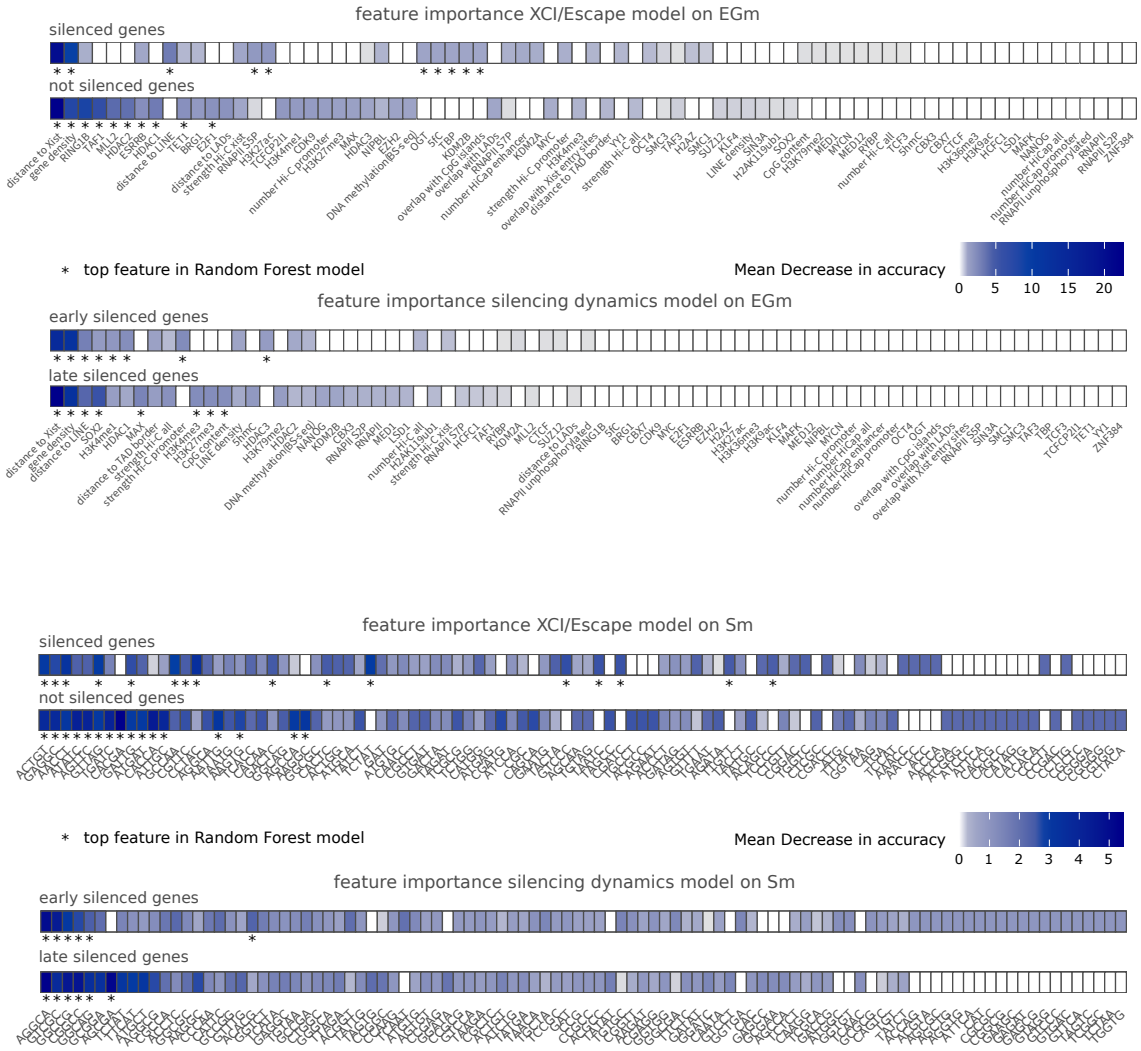

Figure 7.1: Feature importance for XCI/escape and silencing dynamics model. For each model (trained on epigenetic and genomic feature matrix (EGm) or DNA sequence feature matrix (Sm)), features are ranked class wise according to their importance for the classification, quantified by the mean decrease in accuracy (MDA). (*) The top features of each class that are used to build the final model (XCI/escape: 10 (EGm) and 16 (Sm) features; silencing dynamics: 8 (EGm) and 6 (Sm) features; described in Section 6.2.2).

The investigation of the first node levels of a RF model provides a good overview on the features that significantly contribute to the classification. At the same time it does not consider features that have important interaction effects with other features in the model but low predictive power. To get a more detailed view on the important model features, RF provides internal measures of feature importance, based on the OOB data. Here, we chose the ***mean decrease in accuracy***

(MDA, introduced in Section 3.2.2) as feature importance criterion, because it does not only identify features with high predictive power but also features with low predictive power but important interaction effects with other features in the model. We calculated the MDA for every feature in the model and for each class separately, as some features might contribute more to the prediction of one class than the other. Furthermore, we averaged the feature importance (MDA) over a collection of five hundred Random Forest models to obtain stable results and considered only features with MDA > 0. All features identified as important with the split point frequency analysis (above, Figure A7, Figure A8) were also among the most important features in the MDA analysis (Figure 7.1, Figure A9).

The linear distance of the gene TSS to the *Xist* locus as well as the gene density were again the most important features for both models. In contrast to the split frequency analysis, the MDA analysis showed that the distance to the next LINE element, is not only an important feature for the classes of early and late silenced genes but also for the class of silenced genes. While the PRC1 component RING1B and histone deacetylase HDAC2 were again among the important features for silencing in general, the PRC2 deposited histone mark H3K27me3 and the histone deacetylase HDAC3 seem to play an important role for the silencing kinetics. The histone deacetylase HDAC1 is an important feature for both models. The MDA analysis also identified several additional important features for both models. Among the top features specific for the XCI/escape model we found features associated with active transcription, such as H3K27ac and RNAPII S5P and the general transcription factor TAF1. In contrast, for the silencing dynamics model several features related to 3D chromatin organisation (e.g. distance to LADs, 3D interactions with other genomic loci) seem to be important. Interestingly, the top two epigenetic and genomic features for each class have much higher MDA values (MDA > 8%) compared to the remaining top features (MDA 2 – 5%), while the top DNA sequence features all have similar, moderate MDA values (MDA < 6%). This suggests that few epigenetic and genomic features have a big impact in the classification of silencing classes, while no specific k-mer but rather a specific AT or GC rich sequence context seems to be important for the classification of silencing classes.

The variable importance analysis was able to identify the most relevant features in the classification process but did not provide insights into the RF decision making process and the relationship between features and outcome, i.e. is the feature positively or negatively related to a silencing class? To investigate the relationship between features and outcome, different measures and tools were developed (see Section 3.2.2). Here, we focussed on the internal RF ***prototype*** measure, which is calculated by the `randomForest` R package, and a ***partial dependence*** analysis, which is implemented in the `edarf` R package.For both analysis we only used the top x features that were identified in the RF feature selection process and yield to the model with the highest predictive power (XCI/escape: 10 (EGm) and 16 (Sm) features; silencing dynamics: 8 (EGm) and 6 (Sm) features; described in Section 6.2.2). Prototypes are calculated for each class and feature and show the general trend (e.g. enrichment or depletion) of each feature for each class (for further description see Section 3.2.2). For instance, Figure 7.2A shows that the 5-mer TTGTT frequently occurs in the 100 kb region around gene promoters (count values as defined in Section 6.1.3) of both silencing classes, while the 5-mer AATTG occurs more frequently around gene promoters of silenced genes. Partial dependence plots show the relationship between features and model predictions, represented by a linear function on the feature (for further description see Section 3.2.2). For

instance, the third panel of Figure 7.3B shows that a gene will be predicted with high probability as not silenced (probability of not silenced class > 0.7) if it is located far away from *Xist* locus. Combining the results from both analysis, we can get insights into the relationship between features and silencing classes. The analysis revealed that silenced genes are placed in an AT rich sequence context (Figure 7.2) and have higher levels of sequence-specific transcription factors (e.g. ESRRB) and chromatin remodelling complexes, such as MLL2, TET1, RRC1 component RING1B and histone deacetylases HDAC1 and HDAC2 (Figure 7.3). In agreement with the previous feature importance analysis, not silenced genes are located further away from the *Xist* locus and are found in gene dense regions.

Figure 7.2: Prototypes and partial dependence plots for XCI/escape model trained on DNA sequence features. Prototypes and partial dependence plots are shown for the top 16 features of XCI/escape model trained on DNA sequence features. **(A)** Prototypes computed for each feature and class. A prototype is a representative for the respective class that helps to relate the features to the classes, i.e. how frequent is a specific k-mer in a certain class. The y-axis of a prototype plot shows the actual feature value, i.e the k-mer counts in a 100 kb window surrounding the gene TSS as defined in Section 6.1.3, while the displays the top x features. **(B)** Partial dependence plot for each feature. A partial dependence plot shows the relation of each feature to the prediction probability for both classes, i.e. which class is predicted if we observe a high count of a certain k-mer. The x-axis corresponds to the actual feature value, i.e. the k-mer counts in a 100 kb window surrounding the gene TSS (as defined in Section 6.1.3) and the y-axis corresponds to the prediction probability of either class for a specific feature value.

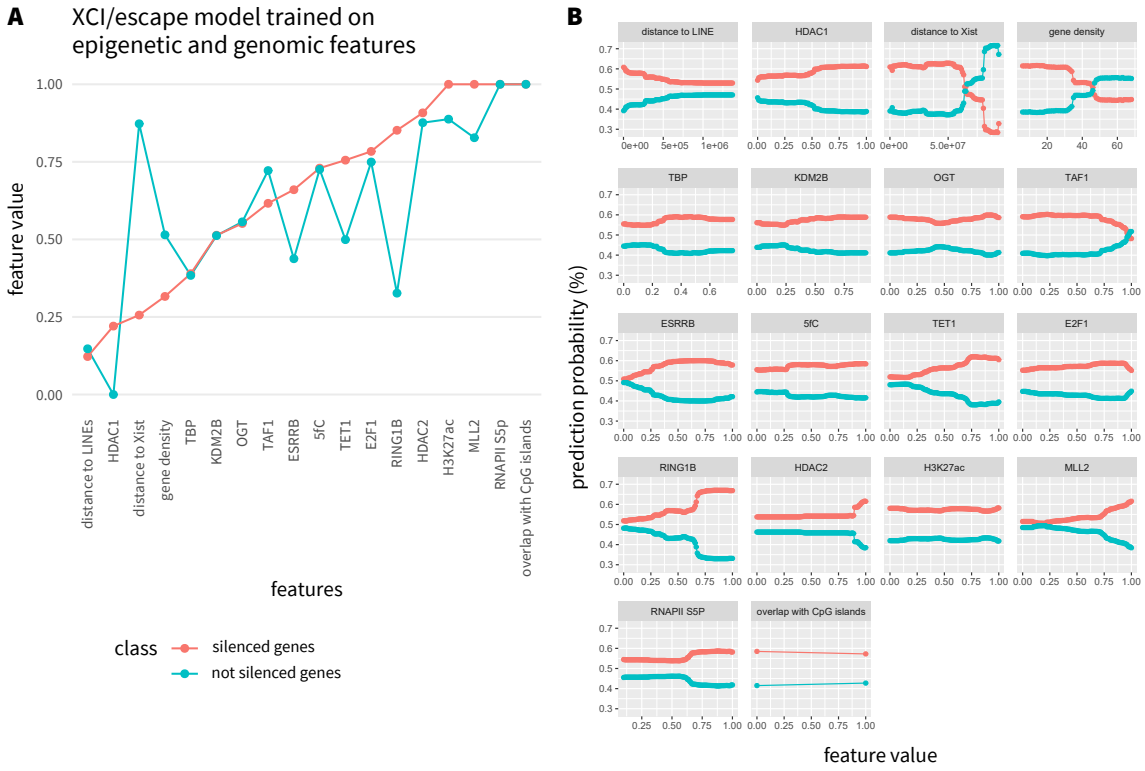

Figure 7.3: Prototypes and partial dependence plots for XCI/escape model trained on epigenetic and genomic features. Prototypes and partial dependence plots are shown for the top 10 features of XCI/escape model trained on epigenetic and genomic features. **(A)** Prototypes computed for each feature and class. A prototype is a representative for the respective class that helps to relate the features to the classes, i.e. is an epigenetic feature enriched or depleted in a certain class. The y-axis corresponds to the actual feature value, i.e. the enrichment of epigenetic features or scaled distance / density measures (e.g. distance to *Xist* is scaled by the maximum distance to the *Xist* locus in the data set to put the feature value on the same scale as the epigenetic features). **(B)** Partial dependence plot for each feature. A partial dependence plot shows the relation of each feature to the prediction probability for both classes, i.e. which class is predicted if a certain epigenetic feature is enriched. The x-axis corresponds to the enrichment of epigenetic features (as defined in Section 6.1.1) or distance / density measures (as defined in Section 6.1.2) and the y-axis corresponds to the prediction probability of either class for a specific feature value.

A similar pattern can be observed for late silenced genes, which are also found in gene dense regions, far away from the *Xist* locus (but not as far as not silenced genes) but also far away from full-length LINE elements (Figure 7.4). Late silenced genes are placed in a GC rich sequence context (Figure 7.5). In accordance, early silenced genes have a lower CpG content than late silenced genes and are, in addition, either depleted in SOX2 or have high levels of SOX2.

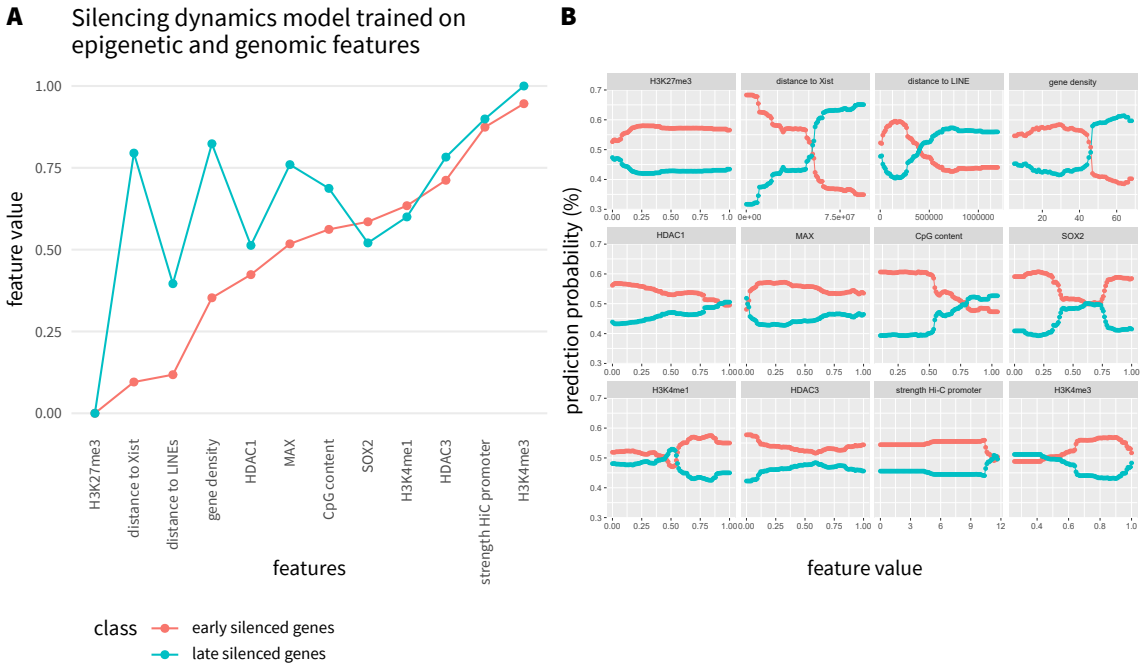

Figure 7.4: Prototypes and partial dependence plots for silencing dynamics model trained on epigenetic and genomic features. Prototypes and partial dependence plots are shown for the top 8 features of silencing dynamics model trained on epigenetic and genomic features. **(A)** Prototypes computed for each feature and class. A prototype is a representative for the respective class that helps to relate the features to the classes, i.e. is an epigenetic feature enriched or depleted in a certain class. The y-axis corresponds to the actual feature value, i.e. the enrichment of epigenetic features or scaled distance / density measures (e.g. distance to *Xist* is scaled by the maximum distance to the *Xist* locus in the data set to put the feature value on the same scale as epigenetic features). **(B)** Partial dependence plot for each feature. A partial dependence plot shows the relation of each feature to the prediction probability for both classes, i.e. which class is predicted if a certain epigenetic feature is enriched. The x-axis corresponds to the enrichment of epigenetic features (as defined in Section 6.1.1) or distance / density measures (as defined in Section 6.1.2) and the y-axis corresponds to the prediction probability of either class for a specific feature value.

Unfortunately, no clear trend is captured for the remaining features, which could be due to an enrichment or depletion in only a subset of (early) silenced genes. This in turn demonstrates the major drawback of those interpretation techniques. They only provide a linear view on the relationship between features and silencing class although RFs are used to model non-linear relationships. In chapter 6, we switched from a linear to a RF model due to the non-linear nature of our classification problem. Hence, the goal of the next section is to extract and visualize the combinatorial rules that might be captured by the RF model.

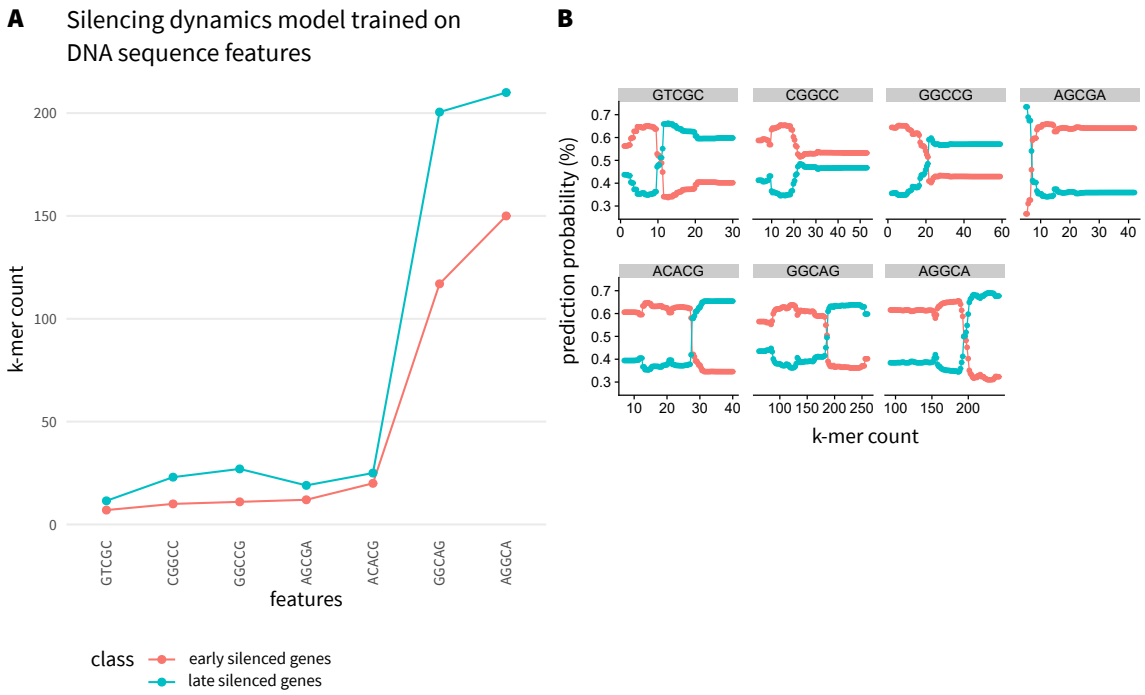

Figure 7.5: Prototypes and partial dependence plots for silencing dynamics model trained on DNA sequence features. Prototypes and partial dependence plots are shown for the top 6 features of silencing dynamics model trained on DNA sequence features. **(A)** Prototypes computed for each feature and class. A prototype is a representative for the respective class that helps to relate the features to the classes, i.e. how frequent is a specific k-mer in a certain class. The y-axis of a prototype plot shows the actual feature value, i.e. the k-mer counts in a 100 kb window surrounding the gene TSS as defined in Section 6.1.3, while the displays the top x features. **(B)** Partial dependence plot for each feature. A partial dependence plot shows the relation of each feature to the prediction probability for both classes, i.e. which class is predicted if we observe a high count of a certain k-mer. The x-axis corresponds to the actual feature value, i.e. the k-mer counts in a 100 kb window surrounding the gene TSS (as defined in Section 6.1.3) and the y-axis corresponds to the prediction probability of either class for a specific feature value.

7.2 FOREST-GUIDED CLUSTERING IDENTIFIES COMBINATORIAL FEATURE PATTERNS FROM RANDOM FOREST MODELS

The variable importance analysis above pinpoints the individual contribution of each feature to the classification problem, but cannot identify the role of correlated features and of feature combinations associated with different silencing pathways, which ultimately determine the silencing class of each gene. To stratify the genes into subgroups according to different combinations of classification rules, we implemented a ***forest-guided clustering*** approach.

7.2.1 Forest-Guided Clustering

We used the proximity measure (see Section 3.2.1) between genes to cluster genes that are regulated by the same set of rules with the k-medoids clustering algorithm (see Section 3.2.2).

Clustering on the Proximity Matrix

The ***proximity*** measure between two genes i and j represents the frequency with which those genes occur in the same terminal nodes of a tree in the RF, intuitively defining how close those genes are in the RF model. The proximity matrix is a $N \times N$ symmetric matrix (with N =total number of genes) and can be calculated from a trained RF model (using the R package `randomForest`). Each entry in the proximity matrix lies in the interval $[0, 1]$. The values $1 - proximity[i, j]$ are squared distances in an Euclidean space (Breiman et al., 2003) and can therefore be used as distance measures:

$$distance[i, j] = 1 - proximity[i, j]$$

for a k-medoids clustering (using the pam function of the `cluster` R-package). The proximity matrix values and the class predictions used for clustering are averaged over 500 RF models.

Determining the optimal number of clusters

Similarly to k-means clustering, k-medoids clustering requires setting the number of clusters k in advance. We developed a ***scoring system*** to choose the optimal number of clusters k, which minimizes the model bias while restricting the model complexity. The model bias measures how well the trained model (with a certain value of k) approximates the expected model, while the variance is related to the model complexity, since complex models usually have a high variance and poor generalization capability.

We define the ***model bias*** by the $mixture\_index_k$ that penalizes values of k yielding a clustering with a high degree of mixture, i.e. clusters containing genes from both silencing classes. For the definition of the $mixture\_index_k$, we introduce a mixture measure for each cluster i that is defined as:

$$mixture\_index_i = 4\left(\frac{x_{i0}}{n_i} \times \frac{x_{i1}}{n_i}\right)$$

where n_i is the number of genes in cluster i and x_{ij} with either $j = 0$ or $j = 1$, is the number of genes in cluster i belonging to the silencing class j. The maximum mixture value for each cluster i is 0.25 in case of a completely mixed cluster where 50% of genes belong to one class and 50% to the other class. We multiply the mixture value by a scaling factor of 4 to obtain a number between 0 and 1. A small adjustment to this formula is needed in case of class imbalances. The smaller class needs to be scaled to the size of the larger class in a way that both classes have comparable influence on the mixture value. Hence, the number of genes belonging to the smaller class x_{sj} in cluster i are scaled by:

$$scaled x_{sj} = x_{sj} + \frac{x_{sj}}{n_{small}} \times (n_{large} - n_{small})$$

where n_{small} is the total number of genes belonging to the smaller class and n_{large} is the total number of genes belonging to the larger class. The $mixture\_index_k$ for a given number of clusters k represents the average degree of mixture per cluster across all k clusters:

$$mixture\_index_k = \frac{\sum_{i=1}^{k} mixture\_index_i}{k}$$

The smaller the value of the $mixture\_index_k$, the better the separation of both class into separate clusters.

On the other hand, we restrict the ***model variance*** by discarding too complex models (i.e. partitions with many clusters, high values of k) and thereby avoiding overfitting. Therefore, we analyse the stability of the forest-guided clustering for each value of k. We assess the stability of each cluster i in the clustering with the average Jaccard Similarity between the original cluster A and three hundred bootstrapped clusters B_b:

$$JS_i(A|B) = \frac{\sum_{b=1}^{300} \frac{|A \cap B_b|}{|A \cup B_b|}}{300}$$

using the function `clusterboot` of the R package `fpc`. Therefore, the genes are resampled via bootstrapping and each bootstrap is clustered with the k-medoids clustering procedure described above. Jaccard similarities values, which are smaller than or equal to 0.5 are an indication of a dissolved cluster, while values higher than 0.6 are usually indicative of stable patterns in the data (Hennig, 2008). We define a clustering to be stable if each cluster i in the partition has an average Jaccard Similarity $JS_i(A|B) > 0.6$. Only stable clusterings, i.e. clustering with low variance, are considered as clustering candidates for selecting an optimal value of k based on the minimal bias. Hence, the optimal number of clusters k is the one yielding the minimum $mixture\_index_k$, while having a stable clustering.

An Analysis of Variance (ANOVA) test was performed to find features with significant differences across clusters. The results of the k-medoids clustering are visualized as heatmaps, displaying the top 10 features, which have a significant variation across clusters according to the p-value of the ANOVA test, and a few other features with significant differences that provide interesting biological insights (Figure 7.6A, Figure 7.8A).

7.2.2 Identifying Feature Combinations Associated with Different Silencing Pathways

Genes of the same silencing class are largely expected to cluster together according to a certain combination of features. Given the non-linear nature of our classification problem, we also expect, to some extent, genes from the same silencing class to be grouped in different clusters according to different combinations of features. Considering the scoring system described above, the optimal number of clusters for both XCI/escape models was $k = 3$, with stable clusters ($JS_i > 0.9$ and $JS_i > 0.6$ for EGm and Sm, respectively, Figure A10). The minimal value of $mixture\_index_k$ for the silencing dynamics model was achieved for $k = 3$ ($JS_i > 0.8$, Figure A10) and $k = 2$ ($JS_i > 0.8$, Figure A10) for the EGm and Sm, respectively.

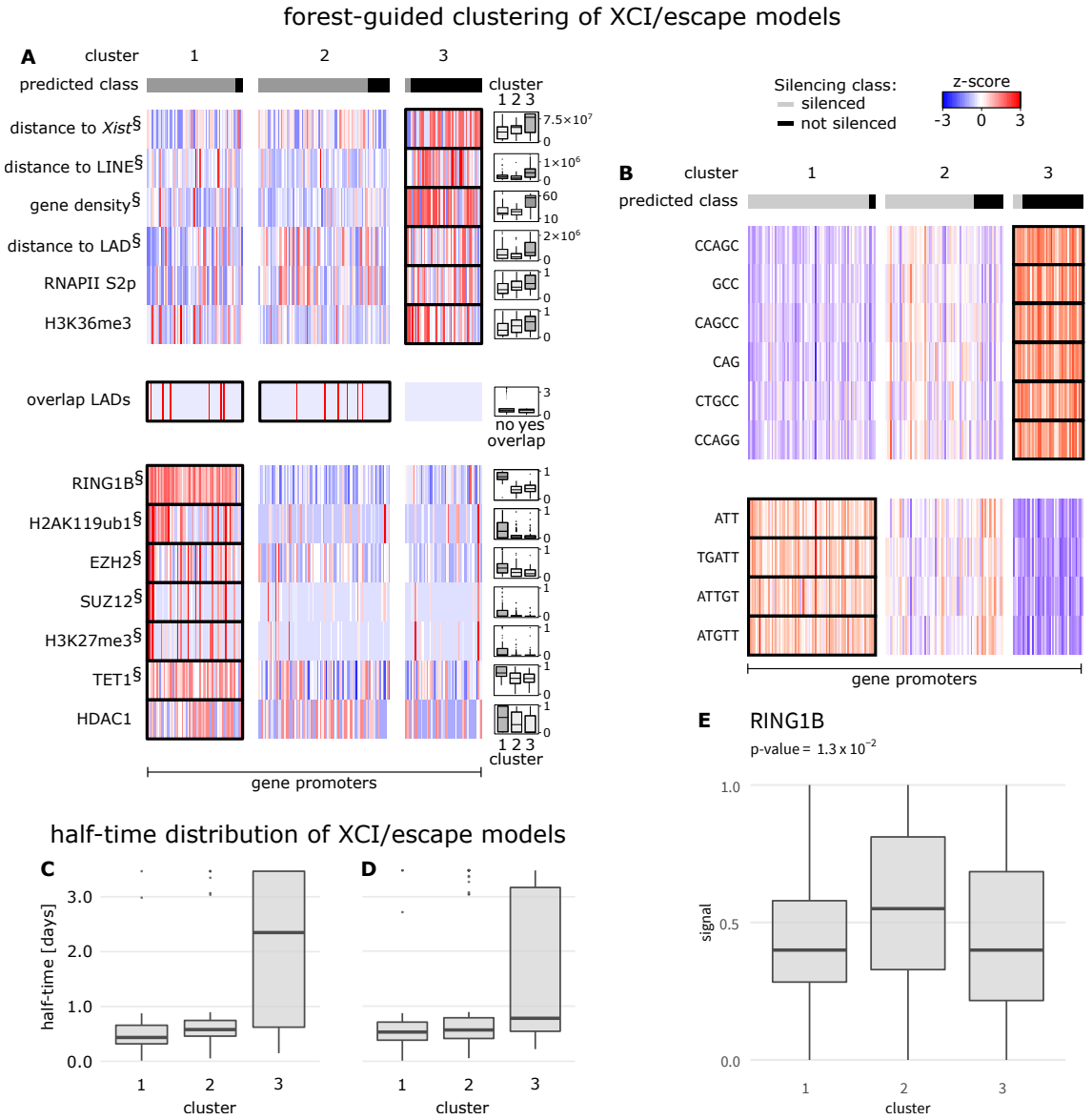

Figure 7.6: Forest-guided clustering for the XCI/escape model. Results from the forest-guided clustering of the XCI/escape model, trained on **(A)** epigenetic and genomic as well as **(B)** DNA sequence features, visualized as a heatmap. Columns indicate the genes grouped by cluster; rows correspond to features with significant differences among clusters (ANOVA test). (§) The top 10 most significant features from the ANOVA test. Distributions of epigenetic and genomic features in each cluster are shown in the box plots next to the heatmap (exception: overlap LADs, where the number of genes in each category is shown). **(C)** Silencing half-time distribution in each cluster of XCI/escape model trained on epigenetic and genomic features. **(D)** Silencing half-time distribution in each cluster of XCI/escape model trained on DNA sequence features. **(E)** Distribution of RING1B in each cluster of the XCI/escape model trained on DNA sequence features, indicates that RING1B is significantly enriched in cluster 2 (p-value of Wilcoxon Rank Sum Test).

The results are visualized in a heatmap showing the genes (columns), grouped by cluster, and a subset of features (rows) selected based on whether they have significant differences across

clusters (Figure 7.6, Figure 7.8). For the XCI/escape model trained on epigenetic and genomic features, genes in clusters 1 and cluster 2 are mainly predicted as silenced and genes in cluster 3 as not silenced (Figure 7.6A, Figure 7.6C).

The escaping genes seem to cluster close to the centromeric region of the mouse X chromosome, which is located far from the *Xist* locus (distance to *Xist* > 70 Mb). Furthermore, genes tend to escape when they are far from LINE elements (distance to LINEs> 400 kb) and LADs (distance to LADs> 800 kb), are located in gene dense regions and are enriched for transcription elongation marks such as RNAPII S2P and H3K36me3. Previous studies pointed out that some genes escape in all cell types (constitutive escapees), while others escape in a cell type specific manner (facultative escapees, see Section 2.2.5, Section 5.4), which poses the interesting question whether constitutive and facultative escapees are enriched for certain epigenetic or genomic features.

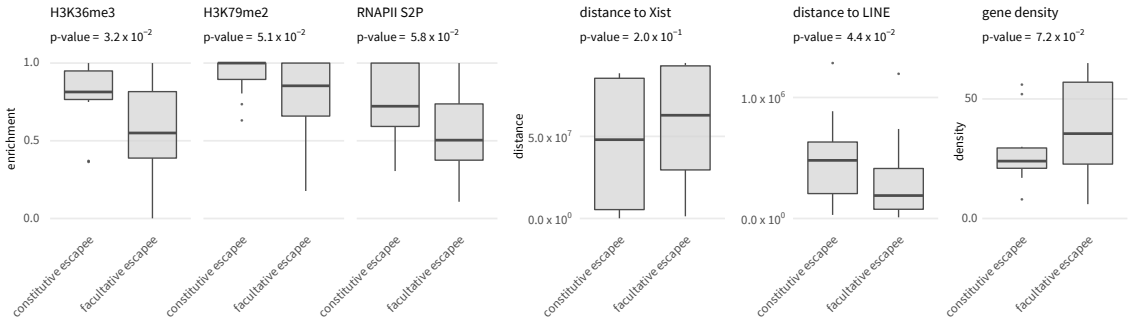

Figure 7.7: Enriched features at constitutive and facultative escapees. Each boxplot shows the differences for epigenetic and genomic features between constitutive and facultative escapees. Only those features where we observe significant differences between constitutive (genes identified as escapee in at least three samples/tissues, listed in Table 4.1) and facultative escapees (p-value of Wilcoxon Rank Sum Test) are displayed.

Figure 7.7 shows that constitutive escapees have higher levels of elongation marks (H3K36m3, RNAPII S2p) compared to facultative escapees, indicating that constitutive escapees might be expressed at higher levels than facultative escapees before XCI is initiated. Facultative escapees on the other hand, seem to be located in gene denser regions and closer to LINE elements. Interestingly, the distance to the *Xist* locus seems to play a less important role for constitutive escapees than for facultative escapees, probably because constitutive escapees are also enriched among genes of the X-inactivation centre (Xic) that help to maintain a stable *Xist* expression, i.e. the constitutive escapee Ftx, which is located in the Xic and is an activator of *Xist* . The class of silenced genes is divided into two clusters: cluster 1, which is marked by a repressive chromatin state (PRC1/2, HDAC1) and bound by TET1, and cluster 2, which is depleted for those marks (Figure A11, Figure A12). A similar pattern can be observed for the forest-guided clustering of the DNA sequence feature model (Figure 7.6B, Figure 7.6D). Again, the class of silenced genes is divided into two clusters: cluster1, which shows a strong enrichment in AT-rich k-mers, and cluster 2, which shows no specific k-mer preferences. However, cluster 2 is significantly enriched for PRC1 mark RING1B (Figure 7.6E), indicating that there might exist two different silencing pathways: one sequence-specific and one Polycomb-specific silencing pathway.

7.2.3 Identifying Combinatorial Rules Associated to the Kinetics of Gene Silencing

In the next step, we investigated the factors that would distinguish early from late silenced genes based on epigenetic and genomic features. The forest-guided clustering approach produced two early silenced clusters (cluster 1 and cluster 2) with lower half-times and one late silenced cluster (cluster 3) with higher half-times (Figure 7.8A, Figure 7.8C).

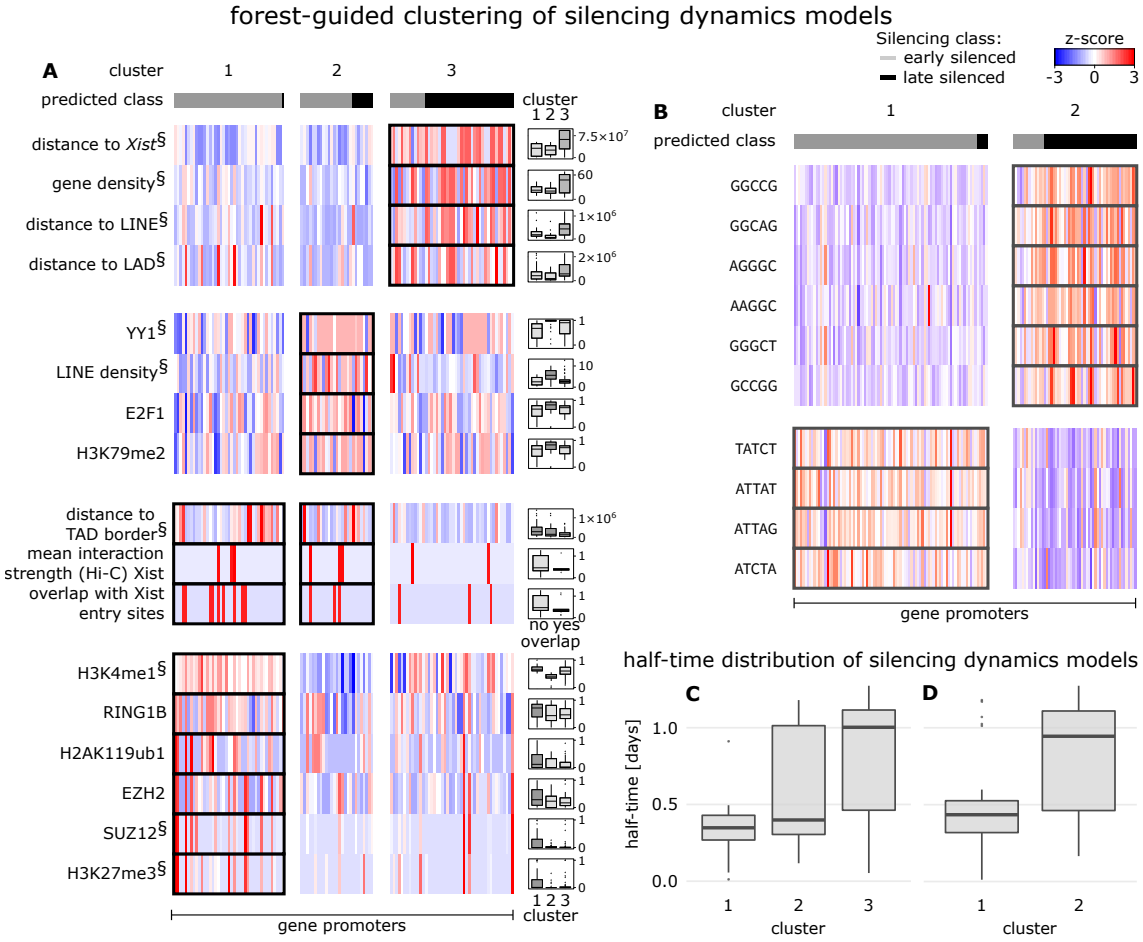

Figure 7.8: Forest-guided clustering for the silencing dynamics model. Results from the forest-guided clustering of the silencing dynamics model, trained on **(A)** epigenetic and genomic as well as **(B)** DNA sequence features, visualized as a heatmap. Columns indicate the genes grouped by cluster; rows correspond to features with significant differences among clusters (ANOVA test). (§) The top 10 most significant features from the ANOVA test. Distributions of epigenetic and genomic features in each cluster are shown in the box plots next to the heatmap (exception: overlap with *Xist* entry sites and mean interaction strength (Hi-C) *Xist* , where the number of genes in each category is shown). **(C)** Silencing half-time distribution in each cluster of silencing dynamics model trained on epigenetic and genomic features. **(D)** Silencing half-time distribution in each cluster of silencing dynamics model trained on DNA sequence features.

Again, one early silenced cluster (cluster 1) is pre-marked by Polycomb repressed chromatin (H2AK119ub1, RING1B, EZH2, SUZ12, H3K27me3) and also H3K4me1. The other early silenced cluster (cluster 2) is mainly characterized by a preferential location of genes in LINE-dense regions and an enrichment of transcription elongation related features (E2F1 subunit and H3K79me2) as well as transcription factor YY1 (Figure A13, Figure A14). Genes in both early silenced clusters tend to be far away from TAD borders, to overlap with *Xist* entry sites and to exhibit strong 3D contacts with the *Xist* locus. The late silenced genes in cluster 3 are mainly characterized by genomic features. They are located in gene dense regions, distant from the *Xist* locus (though not as far as not silenced genes) and far from LINE elements and LADs. In addition, late silenced genes show an enrichment of GC-rich k-mers, while early silenced genes tend to be enriched in AT-rich k-mers (Figure 7.8B, Figure 7.8D).

7.2.4 Contribution of Different *Xist* Repeats to Gene Silencing Pathways

The results of the forest-guided clustering showed that there seem to be two distinct silencing pathways, one influenced by the sequence context and another one controlled by Polycomb Group Protein complexes. Proteins of the Polycomb complex PRC1 were shown to interact with the repeat-B and -C elements on the *Xist* RNA in previous studies (see Section 2.2.1). To investigate how the different **Xist *repeat elements*** are associated with the two silencing pathways, we analyzed data from two previous studies: one study that analysed repeat-A mutants in trophoblasts *in vivo* (Sakata et al., 2017) and another study that analyzed different repeat mutants in mESCs (Bousard et al., 2019) (see Section 4.3).

Sakata et al. measured gene silencing kinetics with allele-specific RNA-seq and provided measurements of the percentage of paternal reads, the fraction of reads expressed from the paternal X chromosome after silencing has occurred. The percentage of paternal reads were then used to define whether a gene is affected by the absence of the repeat-A element or not: gene i is dependent on the repeat-A if the percentage of paternal reads of gene i is greater than 10% (according to Sakata et al.), otherwise gene i is silenced independently of the repeat-A (Sakata et al., 2017). We used the information about the repeat dependency to analyse if any of the clusters of the epigenetic and genomic XCI/escape model is enriched in repeat dependent or independent genes, which would indicate an association of the repeat-A element with one of the two silencing pathways.

Bousard et al. analysed the impact of an repeat-A or repeat-BC depleted *Xist* RNA on gene silencing with an dox-inducible mESC model and measured gene expression via RNA-seq before and after 2 days of dox induction (Bousard et al., 2019). We used the provided data to compute the difference in fold-change between repeat-A or repeat-BC with the wild type form:

$$\Delta fc = log_2 \left(\frac{Dox\ 2\ Days}{no\ Dox} \right)_{mutant} - log_2 \left(\frac{Dox\ 2\ Days}{no\ Dox} \right)_{WT}$$

If $\Delta fc > t$, we considered the silencing to be reduced in the mutant lacking the corresponding repeat element. As the effect on gene silencing is much milder for the repeat-BC than for the

repeat-A mutant, the threshold t was set differently for the two mutants: a higher (more strict) threshold $t = 1$ for the repeat-A and a lower threshold $t = 0.5$ for the repeat-BC, corresponding to the 30% and 60% lower quantile of the fc distribution, respectively. Based on these thresholds we divided our gene set into repeat dependent (silencing affected by the removal of the repeat element, $\Delta fc > t$) and repeat independent genes (silencing not affected by the removal of the repeat element, $\Delta fc < t$).

Figure 7.9: Relation between clustering results and repeat dependency. The proportion of genes shown to undergo silencing in repeat-A and repeat-BC mutants (i.e. *Xist* carrying a deletion of either repeat-A or both repeat-B and -C) is shown for each cluster of the XCI/escape model. In detail, repeat dependent genes refers to genes that have an impaired silencing in the mutants, while repeat independent genes refers to those genes that could still undergo silencing in the repeat mutants, and not measured refers to those genes in our dataset which were not covered by Sakata et al. (Sakata et al., 2017) or Bousard et al. (Bousard et al., 2019). Cluster 2 has a significant enrichment of repeat-A dependent genes compared to cluster 1, whereas repeat-BC dependent genes show an enrichment in cluster 1 compared to cluster 2.

A Fisher's exact test was performed to test for repeat-A or repeat-BC dependent genes in the clusters of the epigenetic and genomic XCI/escape model (Figure 7.9). Cluster 1 is enriched for repeat-BC dependent genes, whose silencing efficiency is impaired in a repeat-BC mutant, compared to cluster 2 (odd ratio = 1.6, Fisher's exact test: $p = 0.19$), while cluster 2 is enriched for repeat-A dependent genes, whose silencing is affected by the removal of the repeat-A element from the *Xist* RNA (mESC data: odd ratio = 2.3, Fisher's exact test: $p = 0.09$ and trophoblast data: odd ratio = 2.76, Fisher's exact test: $p = 0.003$). These findings are consistent with the recent observation that PRC1/2 recruitment requires *Xist* repeat-B and -C and the idea that genes that require Polycomb for silencing are already pre-marked by PRC1/2 and the associated histone modifications (cluster 1). Genes, which are placed in an AT-rich environment, on the other hand, seem to require the repeat-A element for silencing, which activates the SPEN/HDAC3 silencing pathway (cluster 2). Interestingly, matrix-associated/attached regions (MARs), which determine chromatin structure and accessibility, were shown to have an AT-rich sequence context. The special AT-rich binding protein 1 (SATB1) is an MARs binding protein, which serves as docking stations for histone modifying enzymes and is involved in the compaction of chromatin structure, was implicated in the initiation of gene repression by *Xist* (Agrelo et al., 2009). Another interesting observation is that genes, which rely on the repeat-A element for silencing, are located further away from the *Xist* locus than genes that are silenced independently of the A-repeat element. This in turn would be in accordance with the AT-rich sequence context, because MARs and SATB1

localize to the periphery of the chromosome.

The above observations further support the idea that there exist at least two different silencing pathways by which X chromosomal genes can be silenced.

7.3 CONTRIBUTION OF ENHANCER FEATURES TO GENE SILENCING

In Section 6.2.3, we set out to build a RF model on enhancer-associated features. Unfortunately, the enhancer training set was too small to build a reliable RF model. To still get some insights into silencing associated enhancer features, we used a Wilcoxon Rank Sum Test to identify enhancer features that show significant differences between silenced vs not silenced genes (silencing thresholds as defined in Section 6.2.2 for the RF XCI/escape model).

Most of the 110 X chromosomal gene promoters analysed in Section 6.2.3 could be linked to more than one enhancer region. Therefore, we inspected differences between silenced and not silenced genes for features at 1) all enhancers connected to a gene, 2) only the strongest enhancer (i.e. with the best read support), 3) only the closest enhancer to each gene. Following the results of the Wilcoxon Rank Sum Test (Figure 7.10), not silenced genes are preferentially associated to enhancers with high levels of H3K27ac, as well as features related to active transcription (e.g. RNAPII signal) and strong 3D interactions with other genomic regions, all hallmarks of strong enhancer activity. Additionally, we observe a significant pre-marking of CTCF signal at enhancers of not silenced genes compared to silenced genes. On the other hand, enhancers of silenced genes have a smaller genomic distance to the *Xist* locus, LINE elements and LADs, similarly to promoters of silenced genes.

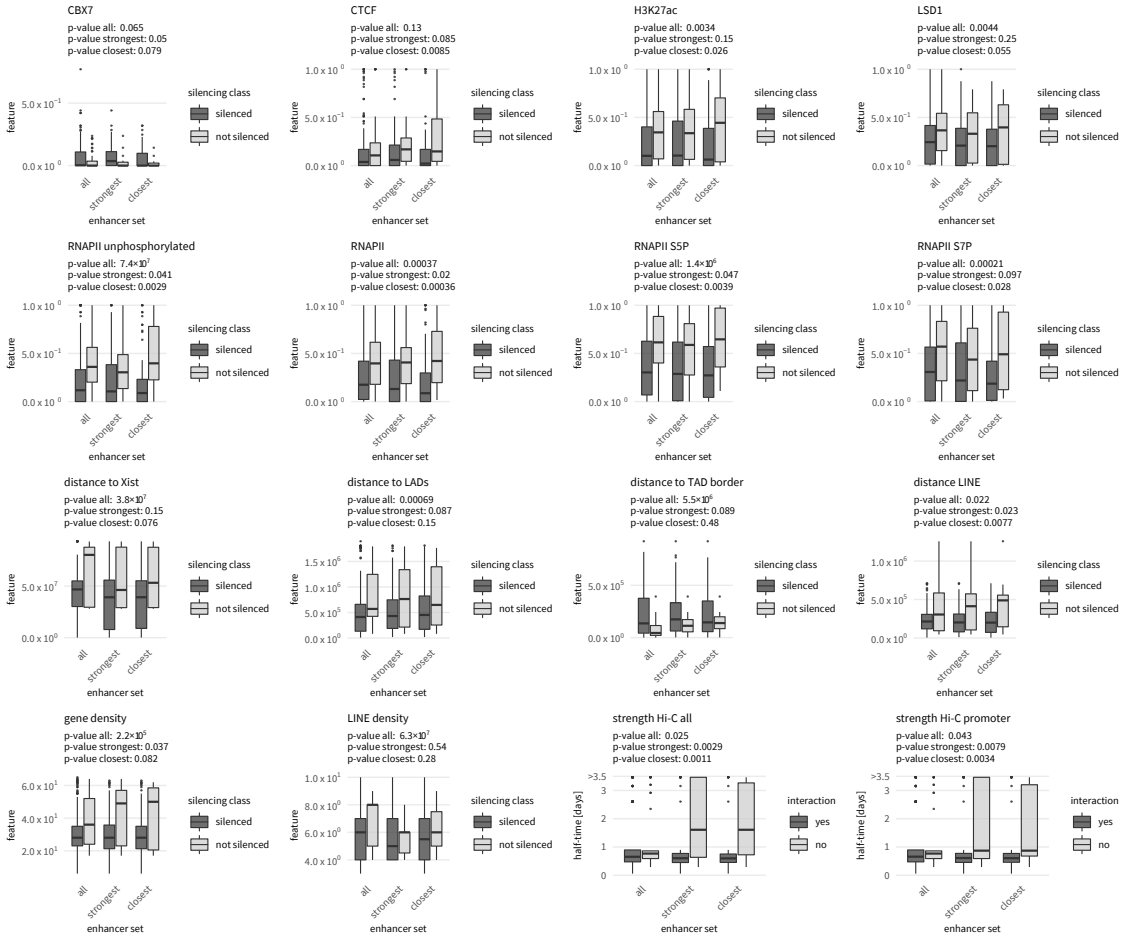

Figure 7.10: Enriched features at enhancers of genes with measured half-times. For each gene we defined putative enhancers via HiCap 3D chromatin promoter-enhancer interactions from Sahlén et al. (Sahlén et al., 2015). Each boxplot shows differences between silenced and not silenced genes for epigenetic and genomic features at 1) all enhancers connected to the gene 2) only the strongest enhancer and 3) only the closest enhancer to each gene. Only those features where we observe significant differences between the class of silenced versus not silenced genes (p-value of Wilcoxon Rank Sum Test) are displayed.

8 DISCUSSION AND CONCLUSION

The process of X-Chromosome Inactivation (XCI) and its master regulator *Xist* has been studied extensively over the past 50 years with various *in vivo*, *in vitro* and *in silico* approaches. Although many key factors in the XCI process have been discovered, we still are missing an overall picture on the interplay of silencing factors and the resulting silencing pathways, because previous studies limited their analysis to only a few genes or certain factors that were shown or hypothesized to have an impact on the XCI process. In the present thesis, those two main issues are addressed to improve the understanding of XCI.

To understand *Xist*-mediated gene silencing on a chromosome-wide level, we measured silencing kinetics on the silenced X chromosome with high temporal resolution through allele-specific expression analysis in undifferentiated mouse embryonic stem cells (mESCs). We assessed silencing dynamics via allele-specific PRO-seq, which measures nascent transcription and therefore, allows for a higher temporal resolution (resolution in hours) compared to mRNA-seq (resolution in days) that has been used in previous studies (Borensztein et al., 2017; Marks et al., 2015). The high temporal resolution enabled a more direct quantification of silencing kinetics, because instantaneous changes in transcription by transcriptionally engaged RNAPII can be captured. Moreover, the use of an inducible system allowed us to uncouple XCI from differentiation and to avoid the use of mutations to ensure non-random XCI.

Previous studies that analysed the XCI process *in vivo* or *in vitro* , focused on only a few promoter-associated factors, often selected based on prior knowledge, whose association with gene silencing was usually shown by hypothesis testing, not taking interactions with other silencing factors into account (Kelsey et al., 2015; Loda et al., 2017; Marks et al., 2015). Unlike those studies, we set out to identify silencing determinants in an unbiased manner, without being influenced by prior knowledge, through the integration of a large number of publicly available epigenetic, genomic and DNA sequence data sets from mESCs. We used the integrated data sets to train a machine learning model that is able to systematically analyse the contribution and combination of features that predispose X chromosomal genes to *Xist*-mediated silencing. Based on computed silencing half-times, derived from the PRO-seq time course experiment, we trained two separate classification models - the XCI/escape and the silencing dynamics model - to identify those factors that are important for silencing in general and those that influence the kinetics of gene silencing. The low predictive power of a linear model was an indicator for the potential combinatorial nature of silencing pathways. This was one of the main reasons why we opted for a RF model, a non-parametric machine learning method, which makes use of multiple decision trees to learn a non-linear classification task. Furthermore, RF models have a reduced risk of overfitting when the number of trees is chosen large enough, even in cases with strong class imbalances, correlated features and small training sets - all properties we exhibit in our data set. Our trained RF models outperformed previous machine learning models that were build on DNA sequence features or

chromatin states (Carrel et al., 2006; Nesterova et al., 2019; Wang et al., 2006). Our RF model, trained on epigenetic and genomic features (22% error rate), generally performed much better than the RF model of Nesterova et al. that was trained on chromatin states (29% error rate). In addition, the used chromatin states represent a mixture of certain epigenetic factors and therefore, only give an overview on the type of features, i.e. repressive or active features, that might contribute to the silencing process but not on specific epigenetic features that might be important for silencing. Two previous models trained on DNA sequence features, using linear discriminant analysis (Carrel et al., 2006) and support vector machines (Wang et al., 2006), reported a slightly better performance than our RF model trained on DNA sequence features. However, those models were only trained on a pre-selected subset of X-chromosomal genes coming from a specific region on the X chromosome. When applied to all genes on chromosome X those models showed a very poor performance for one of the model classes (LDA: 46% error rate for silenced genes; SVM: 83% error rate for escapees), indicating that the trained models are not generalizable to all X chromosomal genes. In contrast, our RF models were trained on all X chromosomal genes, for which we could compute silencing half-times, and achieved equally good error rates for both silencing classes in the model. This in turn indicates that our models are generalizable to genes on the whole X chromosome and are able to learn important properties of both classes from the data. Model predictions for X chromosomal genes without computed silencing half-times were used to further verify the generalizability of our model through 1) experimental testing of candidate genes, 2) validation of model predictions with an independent mRNA-seq data set and 3) prediction of the silencing susceptibility to *Xist* transgenes located on an autosome. The results of all three validation approaches confirmed that our RF model can predict *Xist*-mediated gene silencing based solely on epigenetic and genomic features.

As described above, we trained two separate RF classification models to analyse the impact of silencing factors on the silencing dynamics but also on silencing in general by discretizing our computed silencing half-times into two silencing classes for each model (silenced vs not silenced and early vs late silenced). Since the computed silencing half-times are continuous values, it would have been desirable to use a regression setting to avoid setting a threshold for each silencing class. However, one has to keep in mind that the fitted silencing half-times can be noisy, because the silencing kinetics do not always follow an exponential decay function and hence, a discretization of the fitted silencing half-times made the RF model more robust. In addition, we had to apply a maximum value for the computed silencing-halftimes. This cutoff was necessary because the PRO-seq experiment had a limited range of time points between 0 and 1 day and therefore, half-times above the maximum cutoff would not have been reliable estimates. The consequence was a missing graduation for genes with higher half-times, which made the use of a classification setting more reasonable. Another problem of the limited range of time points was the resulting potential mixture of very late silenced genes and escapees. Some late silenced genes are only fully silenced after more than one day of *Xist* expression (Marks et al., 2015), which means that the class of not silenced genes potentially includes a few very late silenced genes as well. One possibility to tackle both problems would be an extension of the experiment up to several days, in order to avoid setting a half-time cutoff and to get a better half-time resolution for late silenced genes as well as escapees.

To uncover the combinatorial rules that control *Xist*-mediated silencing dynamics we went one step beyond classical variable importance analysis and introduced a forest-guided visualization

scheme. A variable importance analysis provides a ranking of features by importance for the classification task, giving us a linear view on the nonlinear RF model. The forest-guided visualization scheme, on the other hand, takes the nonlinearity of a RF model into account by visualizing the combinatorial rules that characterize certain groups of equally regulated genes (gene clusters), thereby providing insights into the RF decision making process.

The determinants of silencing for groups of clustered genes, recovered by the forest-guided clustering method, recapitulated previous observations but also shed light on novel players or combination of features potentially controlling a gene's susceptibility to *Xist*-mediated silencing (see Figure 8.1). The linear distance and 3D interactions with the *Xist* locus are thought to be the prime determinants of early *Xist* spreading (Engreitz et al., 2013). Our model found the same features to be highly predictive of gene silencing dynamics, an association that had previously been described (Borensztein et al., 2017; Marks et al., 2015) and suggests that efficient *Xist* coating is required for early silencing. However, previous studies had shown that *Xist* RNA initially tends to spread to gene-dense and LINE-poor regions (Engreitz et al., 2013; Simon et al., 2013), but in our analysis, gene density was associated with reduced silencing and LINE elements were found in proximity of early silenced genes. A similar association has been reported previously (Chow et al., 2010; Loda et al., 2017) and suggests that *Xist* coating is not the only determinant of silencing.

The *Xist* RNA recruits several protein complexes that mediate gene silencing, such as SPEN, which binds directly to the repeat-A element, and Polycomb Repressive Complexes (PRC), which are indirectly recruited by the repeat-B element (Brockdorff, 2017; Chu et al., 2015; Monfort et al., 2015; Wutz et al., 2002). Our model identified groups of genes associated with each of these silencing pathways. Repeat-B / PRC associated genes are already enriched for PRC1 and PRC2 prior to the onset of XCI, suggesting that polycomb pre-marking might promote and even accelerate gene silencing and/or reinforce *Xist* spreading, as suggested by a recent study (Colognori et al., 2019). A similar enrichment of PRC components has previously been found at genes susceptible to ectopic silencing by *Xist* transgenes (Kelsey et al., 2015; Loda et al., 2017). Repeat-B associated genes are also enriched for demethylase TET1, which was identified as a stable partner of OGT in the nucleus of ESCs (Vella et al., 2013). Notably, the *Ogt* gene is located on the X chromosome itself and was identified as escapee in our as well as previous studies (Andergassen et al., 2017; Calabrese et al., 2012; Marks et al., 2015; Splinter et al., 2011). In the nucleus, OGT is responsible for modifications of important transcriptional regulators like transcription factors or the C-terminal domain of RNAPII, which leads to the inhibition of RNAPII activation and elongation (Comer et al., 2001). Interestingly, we found OGT among the top features of silenced genes in our variable importance analysis, which suggests that OGT, which is an escapee itself, might contribute to X-linked gene silencing through inhibition of RNAPII activity within the *Xist* cloud. Although we did not find a clear epigenetic signature at repeat-A associated silenced genes in the XCI/escape model, we did observe a strong enrichment for AT-rich k-mers (in accordance with the results of (Wang et al., 2006)) around the promoters of such genes, suggesting that the sequence context plays a more important role for the repeat-A silencing pathway than the epigenetic environment. The AT-rich sequence context and the greater genomic distance to the *Xist* locus of repeat-A associated silenced genes points towards the placement of those genes in matrix-associated/attached regions (MARs), which are bound by the special AT-rich binding protein 1 (SATB1) (Bode et al., 1992). SATB1 itself does not recognize a specific primary DNA sequence but rather recognizes an AT-rich sequence

context, which would explain why no specific k-mer showed a high feature importance in our RF interpretation analysis (Belle et al., 1998). In the study of Agrelo et al. SATB1 was implicated in the initiation of *Xist* -mediated gene silencing and *Xist* was shown to localize along SATB1-organized chromatin (Agrelo et al., 2009). However, a later study reported that SATB1 is dispensable for X chromosome inactivation in fibroblasts, which could be due to the cell-type specific expression of SATB1 (Dickinson et al., 1992; Nechanitzky et al., 2012). This shows that the role of SATB1 in the process of XCI has to be further elucidated. Early silenced genes in the silencing dynamics model that are not enriched for PRC, are located in particularly LINE-dense regions, suggesting that LINE elements contribute to silencing of polycomb-independent genes as well.

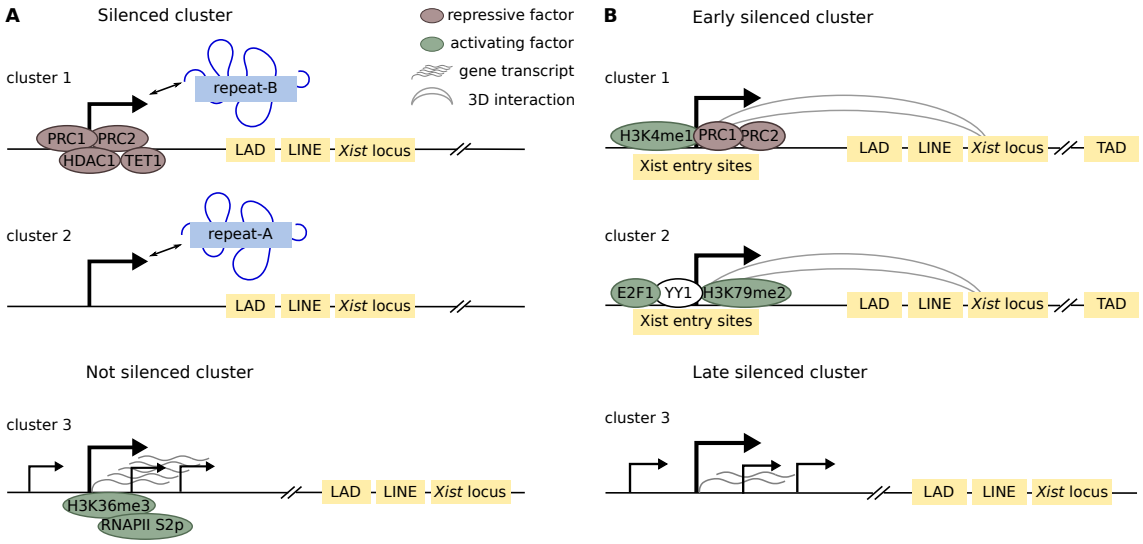

Figure 8.1: Schematic view of epigenetic and genomic mechanisms that predispose X chromosomal genes to *Xist*-mediated silencing. The potential silencing mechanisms are derived from the forest-guided clustering of the RF models trained on the set of epigenetic and genomic data. **(A)** XCI/escape model. Schematic view of the feature combinations promoting gene silencing (clusters 1 and 2) or escape (cluster 3). **(B)** Silencing dynamics model. Schematic view of the features associated with early (cluster 1 and 2) and late gene silencing (cluster 3).

Previous studies looking at post-XCI cells have proposed a role of CTCF in XCI (Berletch et al., 2015; Filippova et al., 2005) and have found a moderate enrichment of CTCF prior to XCI at promoters of escapees compared to promoter of silenced genes (Loda et al., 2017). While we did not find CTCF as one of the discriminating promoter features in our XCI/escape model, we observed a significant enrichment of CTCF signal at enhancers of not silenced X-linked genes, suggesting a potential role of CTCF in gene silencing mediated by chromatin looping between enhancers and promoters. Finally, our analysis identified several structural features that appear to modulate the dynamics of silencing. A high connectivity of some genes, i.e. how much the gene is involved in 3D interactions with other genomic elements, is associated with faster silencing, maybe because *Xist* RNA can spread more easily to these genes through proximity transfer. Moreover, early silencing preferentially occurs at genes that are close to a LADs, which generally contain repressed genes (Steensel et al., 2017), while genes placed in a GC-rich sequence context

and close to TAD boundaries tend to be silenced late.

To investigate the influence of the experimental technique (PRO-seq vs mRNA-seq), used to measure gene silencing dynamics, on the obtained conclusions from the model interpretation analysis, we additionally trained an XCI/escape model on silencing half-times computed from the undifferentiated mRNA-seq data set (see Appendix (Section A.2)). A feature importance and forest-guided clustering analysis of this model gave similar results to the model interpretation analysis for the XCI/escape model trained on undifferentiated PRO-seq data, indicating that the experimental technique has no mayor impact on the drawn conclusions.

The Random Forest approach allowed us to quantify the relative contribution of several features that have been previously associated with XCI (e.g. linear distance to *Xist* , LINE elements, enrichment for PRC1 and PRC2) and suggested new features, which can be tested in more detail in future studies, like TET1 / OGT and some pluripotency factors, such as ESRRB and SOX2, which have recently been implicated in reactivation of the X chromosome during reprogramming (Janiszewski et al., 2019). Additional features could be included in the model to further improve our ability to predict silencing susceptibility and a detailed experimental investigation of the different silencing pathways elicited by *Xist* would facilitate the interpretation of the features that predict silencing dynamics as well as escape from XCI. We also showed that the sequence context seems to have an influence on gene silencing, since (early) silenced genes that are not enriched in PRC components are placed in an AT-rich sequence context. In future work, it would be interesting to look at the sequence context in more detail, to investigate the enrichment of certain oligomers or sequence motifs that potentially contribute the *Xist* spreading and the repeat-A associated silencing function.

The forest-guided clustering approach helped us to derive the two potential silencing pathways, one controlled by the repeat-A element and a specific sequence context, the other one controlled by the repeat-B element and a specific epigenetic context. Nevertheless, there are few features, which are among the top feature in the variable importance analysis that are not captured by the forest-guided clustering. A potential explanation for this observation is the existence of alternative feature combinations, which are only captured by a minority of decision trees in the RF but lead to equally good performances. In our szenario, for instance, the feature distance to *Xist* is predominantly used as a first split point in the decision trees and therefore, heavily influences the remaining split points. In some cases, however, this feature is not among the randomly chosen features for the calculation of the first split point and therefore, another feature is chosen, which potentially leads to a different tree structure with other features used as split points. However, if this alternative feature combination is only captured by a minority of decision trees (e.g. where distance to *Xist* is not available as first split point), it only has a minor influence on the computed proximity matrix and therefore, is not captured by the forest-guided clustering. One potential idea to solve this problem, is to aggregate differently structured decision trees into clusters of decision trees before calculating the proximity matrix. The proximity matrix could then be calculated separately for each group of similarly structured decision trees, which would then be visualized via the forest-guided clustering to uncover all alternative feature combinations. Decision trees could be aggregated via classification results (i.e. cluster decision trees whose classification results for all data points correlate) or via feature co-occurrences (i.e. cluster decision trees that are built on the same set of features). Nevertheless, the forest-guided clustering approach enabled a

systems-level view of the DNA sequence context and epigenetic landscape of the X chromosome in mESCs, which in turn allowed to get new insights into the complex mechanisms behind the different *Xist*-mediated silencing pathways.

ABBREVIATIONS

A adenine

AER allelic expression ratios

ANOVA Analysis of Variance

ATP adenosine triphosphate

biotin-NTP biotin-labeled nucleotide triphosphate

AUC Area Under the Curve

bp base pair

C cytosine

CART Classification and Regression Trees

CER computed OOB error rate

ChIP-seq Chromatin Immunoprecipitation followed by high-throughput sequencing

CTD C-Terminal Domain

CPS correctly predicted silenced

dox doxycycline

EER expected OOB error rate

EGm epigenetic and genomic feature matrix

ENCODE Encyclopedia of DNA Elements

EMBL-EBI European Bioinformatics Institute

EPS expected predicted silenced genes

FP False Positive

FN False Negative

G guanine

GE generalization error

GEO Gene Expression Omnibus

FISH fluorescent in situ hybridization

HAT histone acetyl transferase

HDAC histonedeacetylase complexe

h hour

ICM inner cell mass

i.i.d. independent identically distributed

JS Jaccard Similarity

LAD Lamina-associated Domain

LDA linear discriminant analysis

LINE long interspersed nuclear element

lncRNA long noncoding RNA

MARs matrix-associated/attached regions

MBD methyl-CpG-binding domain protein

MDA mean decrease in accuracy

MDI mean decrease in impurity

mESC mouse embryonic stem cell

MeDIP Methylated DNA immunoprecipitation

MIR mammalian-wide interspersed repeat element

ML machine learning

mRNA-seq messenger RNA-sequencing

NPC Neural Progenitor Cell

PAM Partitioning Around Medoids

PRC Polycomb Repressive Complexes

PRO-seq Precision nuclear Run-On sequencing

OOB Out of Bag

RBP RNA-binding proteins

ReFINE Random Forest INspEctor

RF Random Forest

RNAPII RNA Polymerase II

RNA-seq RNA sequencing

RR regulatory regions

RSS residual sum of squares

RPKM Reads Per Kilobase of transcript, per Million mapped reads

Sm DNA sequence feature matrix

SNP Single Nucleotide Polymorphism

tetOP tetracycline operator (tetO) controlled promoter

TF Transcription Factor

TSC transcriptionally silent compartment

TSS transcription start site

sncRNA short ncRNA

SVM Support Vector Machine

T thymine

TAD topologically associating domain

TSC transcriptionally silent compartment

WGBS Whole Genome Bisulfite Sequencing

Xa active X chromosome

XAR X-added region

XCI X-Chromosome Inactivation

Xi inactive X chromosome

Xic X-inactivation center

Xist X inactive specific transcript

LIST OF FIGURES

LIST OF TABLES

A APPENDIX

A.1 EXPERIMENTAL PROCEDURES AND DATA PROCESSING

The following data was generated in the Lab of Edith Heard and John Lis (see Section 5.1 for contribution of each Lab). All raw and processed sequencing data generated were submitted to the NCBI Gene Expression Omnibus (GEO; `http://www.ncbi.nlm.nih.gov/geo/`) under accession number GSE121144.

ES cell culture. The female TX1072 cell line is a F1 hybrid mouse ESC line derived from a cross between the 57BL/6 (*B6*) and CAST/EiJ (*Cast*) mouse strains that carries a doxycycline responsive promoter in front of the *Xist* gene on the *B6* chromosome and an rtTA insertion in the *Rosa26* locus (Schulz et al., 2014). Cells were grown on gelatin-coated flasks in serum-containing ES cell medium (DMEM (Sigma), 15% 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). Cells were seeded at a density of 10^5 cells/cm^2 coated with gelatin two days before the experiment. *Xist* was induced by supplementing the medium with 1 μg/ml Doxycycline. Samples were collected before doxycycline treatment (0 h) and with dense temporal sampling at time points 0.5, 1, 2, 4, 8, 12 and 24 h (PRO-seq), 2, 4, 8, 12 and 24 h (mRNA-seq) and 4, 8, 12, 24 h after treatment (Pyro-sequencing). Samples without doxycycline and 24 h doxycycline were collected in duplicate to be able to assess reproducibility. To induce differentiation cells were cultured in DMEM, supplemented with 15% FBS and 0.1mM β-mercaptoethanol, and collected at 0, 8, 16, 24 and 48 h for mRNA-seq.

PRO-seq. For each timepoint ~ 10^7 nuclei were isolated by washing the cells twice with ice-cold PBS, and once with 15 ml swelling buffer (10 mM Tris-Cl, pH 7.4, 300 mM Sucrose, 3 mM $CaCl_2$, 2 mM $MgAc_2$, 5 mM DTT). Then, 15 ml cell lysis buffer (10 mM Tris-Cl, pH 7.4, 300 mM Sucrose, 3 mM $CaCl_2$, 2 mM $MgAc_2$, 0.5% NP-40, 1 mM PMSF, EDTA-free protease inhibitors (1 tablet for 50 ml buffer; Roche), 5 mM DTT) is added and cells are scraped off the plate into a 50 ml tube and spun at 900 g and $4^\circ C$ in a swing bucket centrifuge for 5 minutes. Supernatant is removed and the cell pellet is resuspended in 5 ml cell lysis buffer, transferred to a 7 ml dounce homogenizer and dounced 50 times on ice. Dounced cells are moved to 15 ml tube and spun at 1200 g and $4^\circ C$ in a swing bucket centrifuge for 5 minutes. Supernatant is removed and the nuclei are counted, snap frozen and stored in glycerol storage buffer (50 mM Tris-Cl, pH 8.3, 40% glycerol, 0.1 mM EDTA, 5 mM $MgAc_2$, 1 mM PMSF, EDTA-free protease inhibitors (1 tablet for 50 ml buffer; Roche), 5 mM DTT). Run-on and library preparation was performed as previously described (Mahat et al., 2016) using the single biotin-CTP nucleotide run-on protocol to prolong run-on and increase sequence length. In short, run-on was performed with 10^7 nuclei in 100 ml glycerol storage buffer and 100 ml

pre-heated nuclear run-on mix, to get a final concentration in the run-on of 5 mM Tris-HCl, pH 8, 2.5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl, 0.025 mM biotin-11-CTP, 0.25 mM CTP, 0.125 mM ATP, UTP and GTP, 0.5% sarkosyl and RNase inhibitor. Run-on was done for 5 minutes at 37 $\circ$ *C* and stopped by adding 500 ml TRIzol LS. RNA isolation, base hydrolysis, biotinylated-RNA enrichment steps, enzymatic modifications of RNA, adapter ligations, reverse transcription, amplification and library size selection were done as described previously (Mahat et al., 2016). Libraries were sequenced on the HiSeq 2000 Illumina sequencer (single-end, 100bp). For each library at least 50 Mio reads were generated. Adapter sequences were trimmed with `cutadapt` (v1.8.2). Nucleotides with poor 3' base quality (BAPQ < 20) were trimmed and reads of < 20 bp were discarded. After quality control between 30 to 50 million reads remained. Ribosomal reads were first removed by alignment to the rRNA reference (GenBank identifiers:18S, NR_003278.3; 28S, NR_003279.1; 5S, D14832.1; and 5.8S, KO1367.1) using `Bowtie1` (v1.0.0) and allowing 2 mismatches in the seed (options: `-m 1 -l 20 -n 2`) (Langmead et al., 2009). Then, non-ribosomal reads were mapped to both parental genomes. To do this, the VCF file (mgp.v5.merged.snps_all.dbSNP142.vcf) reporting all SNP sites from 36 mouse strains, based on mm10, was downloaded from the Sanger database. `SNPsplit` (v0.3.0) was used to reconstruct the Cast genome from the mm10 reference (Krueger et al., 2016). Only random best alignments with fewer than two mismatches (options: `-M 1 -v 2 -l 20`) were kept for downstream analyses. We applied an allele-specific RNA-seq strategy as described in Borensztein et al. (Borensztein et al., 2017). Briefly, mapping files of both parental genomes were merged for each sample and `SAMtools mpileup` (v1.1) was then used to extract the base-pair information at each genomic position (Li et al., 2009). Read counts mapping to the paternal and maternal genomes, respectively, were summed up across all SNPs present in the same gene. To avoid allele specific bias, we checked the genotypes using a ChIP-seq input from the same cell line. Therefore, only SNPs covered by at least 10 reads in this input sample and having an allelic ratio range between 0.25 and 0.75 were kept for downstream analysis (17, 035, 327 SNPs in total). RPKM values were calculated using gene count table, generated with GENCODE annotation (M9) and `HTSeq` (v0.6.1) (Anders et al., 2015).

mRNA-seq. Cells were lysed by direct addition of 1 ml TRIzol (Invitrogen), 200μl of Chloroform was added and after 15 min centrifugation (12000xg, 4$^\circ C$) the aqueous phase was mixed with 700μl 70% ethanol and applied to a Silica column (Qiagen RNAeasy Mini kit). RNA was then purified according to the manufacturers recommendations, including on-column DNAse digestion. Concentration and purity were checked on a Nanodrop. In case of a low 260/230 ratio, extra ethanol precipitation was performed. RNA profiles were then checked by Bioanalyzer (Agilent RNA 6000 Nano kit) and 1ug of RNA from each condition was used for mRNA-seq. Single Index kit was used, and 12 cycles of PCR were set up. Final libraries were quantified with *Qubit dsDNA HS Assay Kit*, and qualified with *LabChIP® GX system* (PerkinElmer). Then 2 equimolar pools of 16 libraries each were prepared at 10nM. The exact molarity of the pools were assess by qPCR using the *KAPA Library Quantification Kit Illumina* on CFX96 system (Biorad). Then each pool was sequenced on 1 flowcell of HiSeq 2000 system (paired-end, 100bp reads) in PE100, in order to target ~ 100M cluster per sample. The first ten bases from all reads were removed, due to their low quality, using `FASTX toolkit` (v0.0.13). Reads were then mapped to both parental genomes with `TopHat2` (v2.1.0). Only random best alignments with less than two mismatches were kept for downstream analyses. We applied the same allele-specific RNA-seq strategy used for PRO-seq

data analysis.

Pyrosequencing. For pyrosequencing, RNA was extracted using the Direct-zol RNA MiniPrep kit (Zymo Research) and DNase digest was performed using Turbo DNA free kit (Ambion). 1ug RNA was reverse transcribed into cDNA using Superscript III Reverse Transcriptase (Invitrogen). An amplicon containing a SNP is amplified by PCR from cDNA using GoTaq Flexi G2 (Promega) with 2.5 mM $MgCl_2$ or HotStarTaq (Qiagen) for 40 cycles. The PCR product was sequenced using the Pyromark Q24 system (Qiagen). ***Supplemental Table S6*** in (Barros de Andrade E Sousa et al., 2019) contains the forward and reverse Primers used for the validation of 11 candidate genes with Pyrosequencing: six predicted as silenced, 5 predicted as not silenced by the XCI/escape model.

A.2 XCI/ESCAPE MODEL ON UNDIFFERENTIATED MRNA-SEQ DATA

To investigate the differences between the undifferentiated PRO-seq and mRNA-seq data we trained an XCI/escape model on the gene half-times computed from the undifferentiated mRNA-seq data in the same way as we did for the PRO-seq data set, and compared the results with those obtained from the PRO-seq-based XCI/escape model. The accuracy of the RNA-seq model is comparable to the accuracy of the PRO-seq model, and many of the important top features used for classification largely agree between the two models. This is expected given that the Pearson correlation coefficient between the computed half-times from PRO-seq and the undifferentiated mRNA-seq experiment is 0.5.

Distance to *Xist* , gene density, distance to LINEs or TAD boundaries are among the top features which are conserved between the PRO-seq and the mRNA-seq model Figure A1A. We also explored whether the silencing rules retrieved from the forest-guided clustering on the PRO-seq model still hold for the mRNA-seq model. Similarly to the PRO-seq clustering, we observe a partition of genes into three clusters: 2 silenced clusters and 1 not silenced cluster. Figure A1B shows enriched features in the mRNA-seq clusters which were significant in the PRO-seq clustering. However, the distinction between PRC1/2-enriched cluster 1 versus cluster 2 in the mRNA-seq clustering is not as prominent as in the PRO-seq model, indicating that PRO-seq is most probably more sensitive to detect different silencing pathways than mRNA-seq.

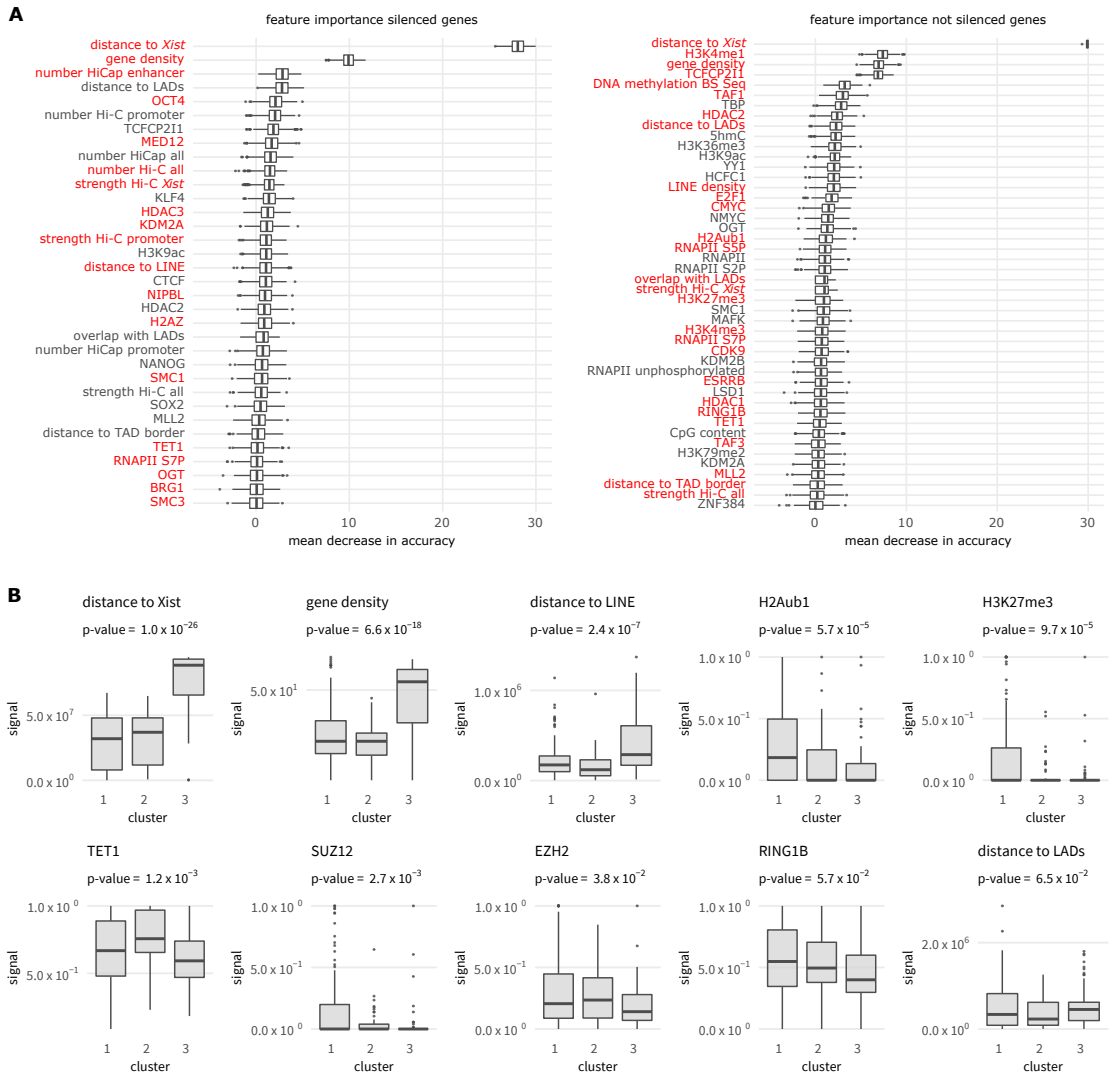

Figure A1: XCI/escape model training on half-times computed from undifferentiated mRNA-seq data in mESCs **(A)** Feature importance of the XCI/escape model on the undifferentiated mRNA-seq data set. Features are ranked based on mean decrease in accuracy (MDA) and only features with a MDA > 0 are shown. Features at the top are more important than features at the bottom. Features marked in red correspond to discriminating features (MDA > 0) also detected in the PRO-seq model. **(B)** Boxplots showing the enrichment of features across the three clusters of the mRNA-seq model. Here we show the feature enrichment in the clusters of the mRNA-seq model of the top 10 most significant features in the ANOVA test of the PRO-seq model.

A.3 SUPPLEMENTAL FIGURES AND TABLES

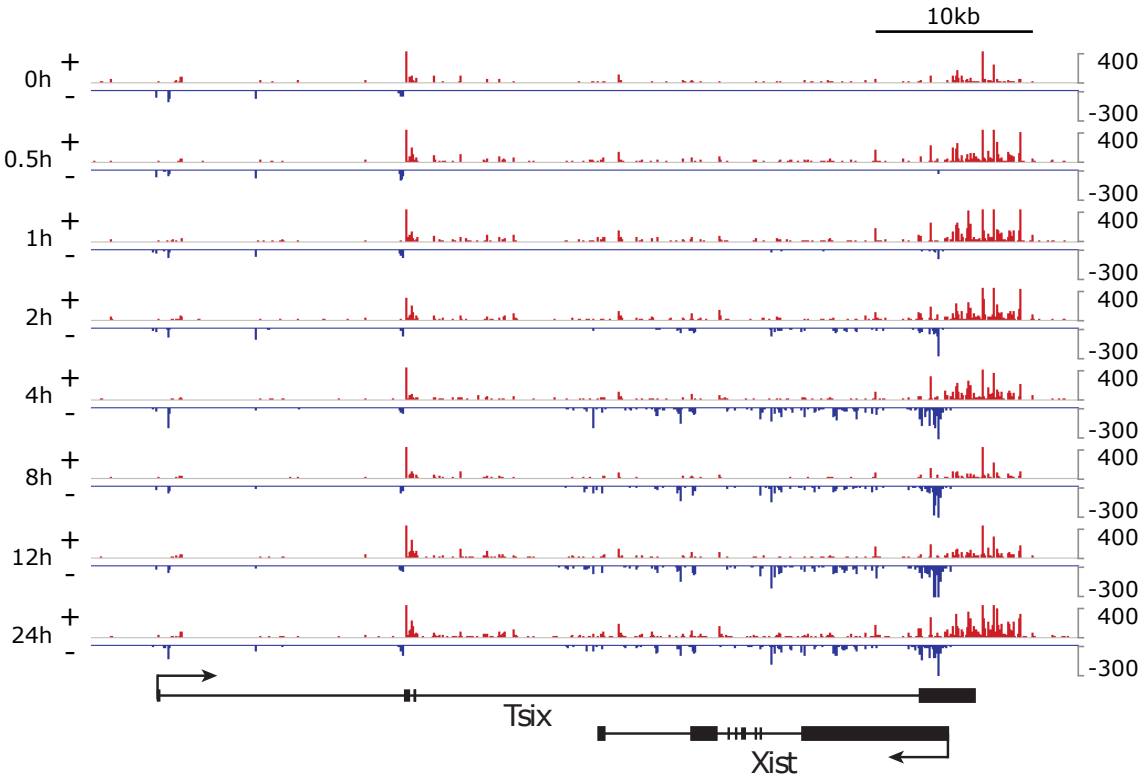

Figure A2: *Tsix* / *Xist* locus. Strand-specific read density at the *Tsix* / *Xist* locus. Plus-strand is shown in red, minus strand in blue; the y-axis indicates reads per million.

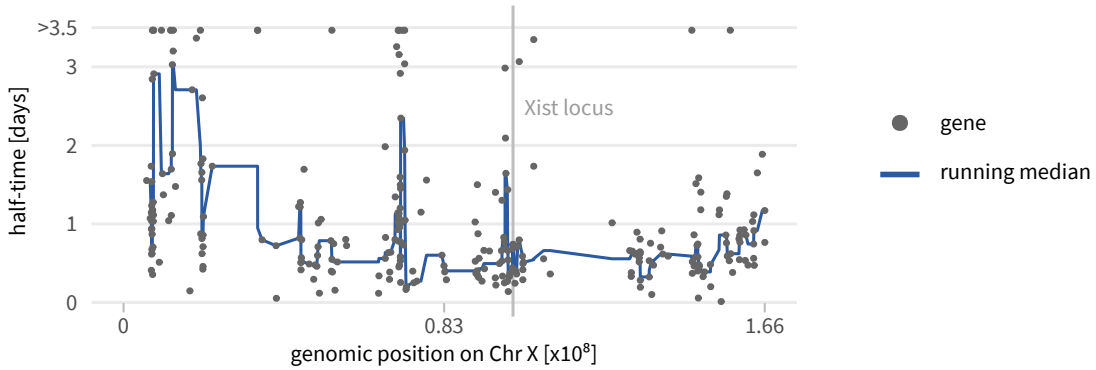

Figure A3: Gene half-times vs genomic position of genes. Estimated half-times (black circles) for all genes in the PRO-seq data set along the X chromosome. A fitted smooth curve of the half-times is displayed as a blue line, and the Xist locus is marked with a gray line.

Table A1: Metadata for ChIP-seq features.

ChIP-seq feature	GEO Accession Number	filtering reason	enrichment region start (wrt TSS)	enrichment region end (wrt TSS)
5fC	GSE40148		−500	500
5hmC	GSE28682		−500	500
5mC	GSE28682	removed due to bad deepTools heatmap		
BRG1	GSE14344		−500	500
CBX3	GSE44242		−250	250
CBX7	GSE42466		−500	500
CDK9	GSE44286		−500	500
CMYC	GSE11431		−500	500
COREST	GSE27841	removed due to bad deepTools heatmap		
CTCF	GSE25777	removed due to bad deepTools heatmap		
CTCF	GSE28247		−500	500
CTCF	GSE29184	removed due to bad deepTools heatmap		
E2F1	GSE11431		−750	750
ESRRB	GSE11431		−500	500
EZH2	GSE46536	removed due to bad deepTools heatmap		
EZH2	GSE49431		−500	500
EZH2	GSE55697	removed because better experiment available for same feature		
EZH2	GSE66830	removed due to bad deepTools heatmap		
H2Aub1	GSE34518		−500	500
H2AZ	GSE36114	removed due to bad deepTools heatmap		
H2AZ	GSE39237	removed because better experiment available for same feature		
H2AZ	GSE53208		−500	500
H3K27ac	GSE31039		−750	750
H3K27ac	GSE36114	removed due to bad deepTools heatmap		
H3K27me3	GSE12241	removed due to bad deepTools heatmap		

H3K27me3	GSE36114	removed due to bad deepTools heatmap		
H3K27me3	GSE41589	removed because better experiment available for same feature		
H3K27me3	GSE47949		−1000	1000
H3K27me3	GSE55697	removed because better experiment available for same feature		
H3K36me2	GSE41589	removed due to bad deepTools heatmap		
H3K36me3	GSE11724	removed because better experiment available for same feature		
H3K36me3	GSE12241	removed due to bad deepTools heatmap		
H3K36me3	GSE31039	removed due to bad deepTools heatmap		
H3K36me3	GSE34518	removed because better experiment available for same feature		
H3K36me3	GSE36114	removed due to bad deepTools heatmap		
H3K36me3	GSE41589		0	gene end
H3K4me1	GSE11172	removed because better experiment available for same feature		
H3K4me1	GSE29184	removed due to bad deepTools heatmap		
H3K4me1	GSE31039	removed due to bad deepTools heatmap		
H3K4me1	GSE32218	removed because better experiment available for same feature		
H3K4me1	GSE36114	removed due to bad deepTools heatmap		
H3K4me1	GSE47949		−1000	1000
H3K4me2	GSE11172	removed because coverage was < 3 Mio.		
H3K4me2	GSE36114	removed due to bad deepTools heatmap		
H3K4me3	GSE11724	removed because better experiment available for same feature		

H3K4me3	GSE12241	removed because better experiment available for same feature		
H3K4me3	GSE29184	removed because better experiment available for same feature		
H3K4me3 GSE31039	removed because better experiment available for same feature			
H3K4me3	GSE32218		−1000	1000
H3K4me3	GSE36114	removed due to bad deepTools heatmap		
H3K79me2	GSE11724		0	3000 or gene end
H3K9ac	GSE31039		−1000	1000
H3K9me3	GSE12241	removed due to bad deepTools heatmap		
H3K9me3	GSE18371	removed due to bad deepTools heatmap		
H3K9me3	GSE31039	removed due to bad deepTools heatmap		
H3K9me3	GSE32218	removed due to bad deepTools heatmap		
H3K9me3	GSE47894	removed due to bad deepTools heatmap		
H4K20me3	GSE12241	removed due to bad deepTools heatmap		
HCFC1	GSE36030		−500	500
HDAC1	GSE27841		−250	250
HDAC2	GSE27841		−250	250
HDAC3	GSE116480		−500	500
KAP1	GSE41903	removed because coverage was < 3 Mio.		
KDM2A	GSE40860		−500	500
KDM2B	GSE37930	removed because better experiment available for same feature		
KDM2B	GSE40860		−750	750
KLF4	GSE11431		−250	250
LAMINB	GSE28247	removed due to bad deepTools heatmap		

LSD1	GSE18515	removed because better experiment available for same feature		
LSD1	GSE27841		−750	750
MAFK	GSE36030		−500	500
MAX	GSE48175		−500	500
MBD1A	GSE39610	removed because coverage was < 3 Mio.		
MBD1B	GSE39610	removed because coverage was < 3 Mio.		
MBD2A	GSE39610	removed because coverage was < 3 Mio.		
MBD2T	GSE39610	removed because coverage was < 3 Mio.		
MBD3A	GSE39610	removed because coverage was < 3 Mio.		
MBD4	GSE39610	removed because coverage was < 3 Mio.		
MECP2	GSE39610	removed because coverage was < 3 Mio.		
MED1	GSE22562		−500	500
MED12	GSE22562		−500	500
MI2B	GSE27841	removed due to bad deepTools heatmap		
MLL2	GSE48172		−500	500
NANOG	GSE11431	removed because coverage was < 3 Mio.		
NANOG	GSE11724	removed because better experiment available for same feature		
NANOG	GSE44286		−500	500
NIPBL	GSE22562		−500	500
NMYC	GSE11431		−500	500
OCT4	GSE11431	removed because better experiment available for same feature		
OCT4	GSE11724	removed because better experiment available for same feature		
OCT4	GSE44286		−500	500
OGT	GSE39154		−500	500
P300	GSE11431	removed due to bad deepTools heatmap		

P300	GSE28247	removed due to bad deepTools heatmap		
P300	GSE29184	removed due to bad deepTools heatmap		
PHF19	GSE41589	removed due to bad deepTools heatmap		
PHF19	GSE41609	removed due to bad deepTools heatmap		
RAD21	GSE25777	removed due to bad deepTools heatmap		
REST	GSE27841	removed due to bad deepTools heatmap		
RING1B	GSE34518		−500	500
RING1B	GSE42466	removed because better experiment available for same feature		
RING1B	GSE55697	removed because better experiment available for same feature		
RNAPII	GSE12241	removed because coverage was < 3 Mio.		
RNAPII	GSE28247		−500	500
RNAPII	GSE29184	removed because better experiment available for same feature		
RNAPII_8WG16	GSE34518		−500	500
RNAPII_S2P	GSE34518		0	gene end
RNAPII_S5P	GSE34518		−500	500
RNAPII_S7P	GSE34518		−500	500
RYBP	GSE42466		−500	500
SETDB1	GSE18371	removed due to bad deepTools heatmap		
SIN3A	GSE24841		−500	500
SIN3A	GSE24841	removed because better experiment available for same feature		
SMAD1	GSE11431	removed because coverage was < 3 Mio.		
SMC1	GSE22562		−500	500
SMC3	GSE22562		−500	500

SOX2	GSE11431 removed because coverage was < 3 Mio.			
SOX2	GSE11724	removed because better experiment available for same feature		
SOX2	GSE44286		−500	500
STAT3	GSE11431	removed due to bad deepTools heatmap		
SUZ12	GSE11431	removed due to bad deepTools heatmap		
SUZ12	GSE11724	removed due to bad deepTools heatmap		
SUZ12	GSE42466	removed because better experiment available for same feature		
SUZ12	GSE44286	removed because better experiment available for same feature		
SUZ12	GSE49431		−750	750
SUZ12	GSE55697	removed because better experiment available for same feature		
SUZ12	GSE66830	removed due to bad deepTools heatmap		
TAF1	GSE30959		−500	500
TAF1	GSE36114	removed due to bad deepTools heatmap		
TAF3	GSE30959		−750	750
TBP	GSE30959		−500	500
TCF3	GSE11724		−500	500
TCFCP2I1	GSE11431		−500	500
TET1_C	GSE24841		−500	500
TET1_N	GSE24841	removed because better experiment available for same feature		
YY1	GSE68195		−500	500
ZC3H11A	GSE36030	removed due to bad deepTools heatmap		
ZNF384	GSE36030		−500	500

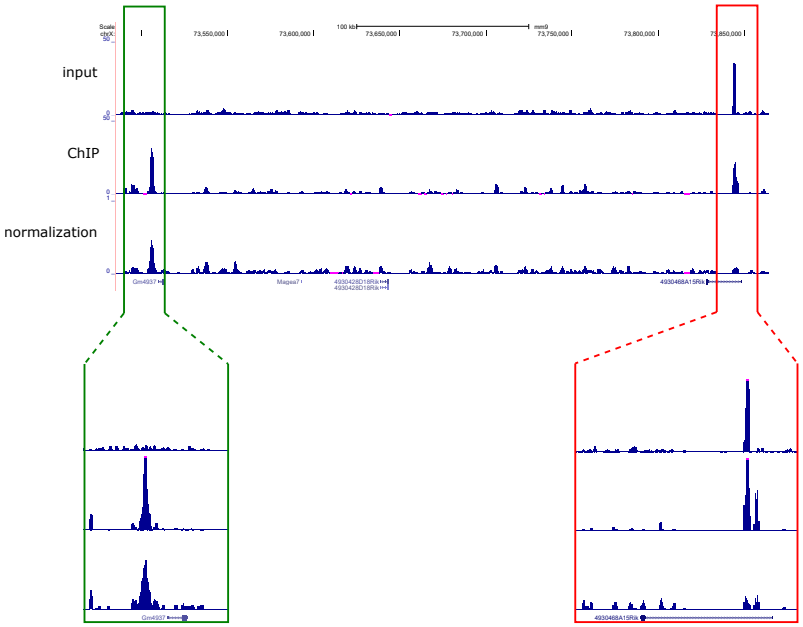

Figure A4: Example of ChIP-seq signal normalization with normR. Two genomic loci on Chromosome X are shown. The green box highlights a region with no or little uniform signal in the control but a sharp peak in the ChIP library. The normalized track correctly shows that the signal corresponding to the sharp peak is still maintained after normalization. In contrast, the red box highlights a region with a peak signal in both the control and the ChIP library. The normalized track correctly shows that the peak in this region is rescaled after normalization to the control signal.

Table A2: Performance of different machine learning methods for XCI / escape model.

silencing threshold	# of silenced genes / # of not silenced genes / size training set	**Logistic Regression** (error rate **total**)	**Elastic Net Regression** (error rate **total** / silenced / not silenced)	**Random Forest** (epigenetic & genomic features model)			**Random Forest** (DNA sequence feature model on 100kb window)	
				all features (error rate total / silenced / not silenced)	top features (error rate total / silenced / not silenced)	permutation test top features (p-value)	all features (error rate total / silenced / not silenced)	top features (error rate total / silenced / not silenced)
sil.: $t_{1/2}$ < 0.9 not sil.: $t_{1/2}$ > 1.4	168 / 64 / 51	**44.33**	**34.81** / 34.59 / 36.77	**29.18** / 28.03 / 32.2	**25.49** / 25.08 / 26.57	0.016	**36.35** / 35.34 / 38.97	**24.3** / 24.07 / 24.91
sil.: $t_{1/2}$ < 0.9 not sil.: $t_{1/2}$ > 1.5	168 / 57 / 46	**46.06**	**35.05** / 34.91 / 36.59	**26.96** / 25.8 / 30.35	**23.03** / 21.69 / 26.97	0.002	**37.63** / 36.43 / 41.16	**24.88** / 25.07 / 24.32
sil.: $t_{1/2}$ < 0.9 not sil.: $t_{1/2}$ > 1.6	168 / 50 / 40	**48.07**	**35.94** / 35.7 / 39.05	**27.95** / 26.85 / 31.63	**22.44** / 22.95 / 19.97	0.024	**30.55** / 28.82 / 38.86	**19.01** / 18.89 / 21.8
sil.: $t_{1/2}$ < 0.9 not sil.: $t_{1/2}$ > 1.7	168 / 44 / 34	**49.66**	**37.59** / 37.53 / 38.4	**26.53** / 25.58 / 30.14	**21.11** / 20.31 / 24.1	0.020	**35.91** / 34.23 / 42.32	**26.31** / 26.13 / 27
sil.: $t_{1/2}$ < 0.9 not sil.: $t_{1/2}$ > 1.8	168 / 40 / 30	**49.68**	**38.6** / 38.5 / 40	**28.87** / 27.93/ 32.8	**22.69** / 22.69 / 22.7	0.008	**35.69** / 33.7 / 44.05	**25.81** / 24.83 / 29.9
sil.: $t_{1/2}$ < 0.9 not sil.: $t_{1/2}$ > 1.9	168 / 37 / 27	**48.89**	**37.41** / 37.4 / 37.6	**30.02** / 29.31 / 33.28	**24.12** / 24.59 / 22	0.014	**31.84** / 29.81 / 41.08	**21.84** / 20.97 / 25.78
sil.: $t_{1/2}$ < 0.9 not sil.: $t_{1/2}$ > 2.0	168 / 35 / 25	**49.01**	**38** / 37.82 / 40.6	**28.28** / 27.55 / 31.79	**21.31** / 20.17 / 25.13	0.028	**38.1** / 37.14 / 41.32	**23.41** / 22.48 / 26.56
sil.: $t_{1/2}$ < 1.0 not sil.: $t_{1/2}$ > 1.4	177 / 64 / 51	**43.73**	**34.74** / 34.36 / 38.46	**29.79** / 29.39 / 30.89	**26.47** / 26.18 / 27.27	0.027	**39.05** / 37.74 / 42.69	**28.12** / 28.77 / 26.31
sil.: $t_{1/2}$ < 1.0 not sil.: $t_{1/2}$ > 1.5	177 / 57 / 46	**46.05**	**35.24** / 35.08 / 37.23	**27.75** / 26.58 / 31.38	**24.49** / 23.59 / 27.3	0.008	**39.33** / 38.05 / 43.33	**25.69** / 25.5 / 26.28
sil.: $t_{1/2}$ < 1.0 not sil.: $t_{1/2}$ > 1.6	177 / 50 / 40	**48.46**	**36.76** / 36.81 / 36.1	**28.06** / 27.04 / 31.68	**21.43** / 20.42 / 24.99	0.012	**39.92** / 38.73 / 44.12	**23.27** / 22.52 / 25.92
sil.: $t_{1/2}$ < 1.0 not sil.: $t_{1/2}$ > 1.7	177 / 44 / 34	**49.95**	**37.79** / 37.72 / 38.75	**27.12** / 26.24 / 30.64	**22.47** / 22.39 / 22.79	0.010	**38.89** / 37.68 / 43.73	**23.87** / 23.2 / 26.59

sil.: $t_{1/2}$ < 1.0 not sil.: $t_{1/2}$ > 1.8	177 / 40 / 30	**49.12**	**39.21** / 39.18 / 39.65	**29.26** / 28.22 / 33.85	**24.12** / 23.93 / 24.98	0.008	**38.26** / 36.87 / 44.4	**23.68** / 22.44 / 29.15
sil.: $t_{1/2}$ < 1.0 not sil.: $t_{1/2}$ > 1.9	177 / 37 / 27	**48.57**	**38.53** / 38.43 / 40.05	**31.37** / 30.86 / 33.83	**22.48** / 22.4 / 22.83	0.000	**33.56** / 31.91 / 41.46	**24.05** / 23.14 / 28.38
sil.: $t_{1/2}$ < 1.0 not sil.: $t_{1/2}$ > 2.0	177 / 35 / 25	**47.95**	**39.6** / 39.49 / 41.15	**29.63** / 29.08 / 32.42	**23.49** / 23.93 / 21.26	0.034	**32.99** / 31.31 / 41.49	**21.13** / 20.48 / 24.4
sil.: $t_{1/2}$ < 1.1 not sil.: $t_{1/2}$ > 1.4	191 / 64 / 51	**43.46**	**37.28** / 37.12 / 39	**32.04** / 31.33 / 34.17	**27.92** / 26.98 / 30.74	0.002	**39.42** / 38.05 / 43.5	**27.65** / 28.68 / 24.59
sil.: $t_{1/2}$ < 1.1 not sil.: $t_{1/2}$ > 1.5	191 / 57 / 46	**45**	**37.23** / 37.08 / 39.18	**31.29** / 30.87 / 32.7	**29.72** / 28.59 / 33.48	0.016	**39.88** / 38.58 / 44.25	**26.45** / 25.14 / 30.84
sil.: $t_{1/2}$ < 1.1 not sil.: $t_{1/2}$ > 1.6	191 / 50 / 40	**48.67**	**37.96** / 37.97 / 37.75	**29.62** / 28.67 / 33.27	**22.53** / 21.81 / 25.26	0.008	**40.46** / 39.52 / 44.04	**24.38** / 23.67 / 27.08
sil.: $t_{1/2}$ < 1.1 not sil.: $t_{1/2}$ > 1.7	191 / 44 / 34	**49.5**	**39.69** / 39.59 / 41.3	**28.57** / 27.55 / 33.02	**23.78** / 24.14 / 22.22	0.000	**39** / 38.08 / 42.95	**26.16** / 25.89 / 27.36
sil.: $t_{1/2}$ < 1.1 not sil.: $t_{1/2}$ > 1.8	191 / 40 / 30	**49.23**	**41.54** / 41.6 / 40.55	**32.06** / 31.29 / 35.77	**25.77** / 25.35 / 27.8	0.028	**37.72** / 36.35 / 44.3	**22.08** / 21.68 / 24
sil.: $t_{1/2}$ < 1.1 not sil.: $t_{1/2}$ > 1.9	191 / 37 / 27	**49.29**	**40.21** / 40.22 / 40.1	**32.93** / 32.54 / 34.95	**24.3** / 24.7 / 22.21	0.002	**33.93** / 32.53 / 41.14	**21.28** / 20.6 / 24.81
sil.: $t_{1/2}$ < 1.1 not sil.: $t_{1/2}$ > 2.0	191 / 35 / 25	**49.79**	**41.05** / 40.91 / 43.35	**33.5** / 33.8 / 31.88	**24.17** / 24.31 / 23.39	0.004	**33.54** / 32.45 / 39.49	**21.03** / 20.95 / 21.43
sil.: $t_{1/2}$ < 1.2 not sil.: $t_{1/2}$ > 1.4	205 / 64 / 51	**43.82**	**38.88** / 38.7 / 41.08	**33.55** / 32.86 / 35.76	**28.14** / 27.67 / 29.66	0.006	**39.53** / 38.45 / 43	**27.81** / 27.75 / 28
sil.: $t_{1/2}$ < 1.2 not sil.: $t_{1/2}$ > 1.5	205 / 57 / 46	**45.91**	**39** / 39 / 39.09	**31.31** / 30.28 / 35.01	**29.01** / 28.28 / 31.63	0.044	**40.01** / 39.02 / 43.58	**23.03** / 22.04 / 26.6
sil.: $t_{1/2}$ < 1.2 not sil.: $t_{1/2}$ > 1.6	205 / 50 / 40	hb48.37	hb39.6 / 39.69 / 38.1	**29.7** / 28.7 / 33.81	**23.86** / 23.25 / 26.34	0.000	**34.89** / 33.9 / 40.69	**19.08** / 19.01 / 19.49
sil.: $t_{1/2}$ < 1.2 not sil.: $t_{1/2}$ > 1.7	205 / 44 / 34	**48.85**	**41.18** / 41.36 / 38.15	**30.44** / 29.81 / 33.37	**25.04** / 24.32 / 28.43	0.002	**39.98** / 38.99 / 44.64	**22.02** / 21.65 / 23.77
sil.: $t_{1/2}$ < 1.2 not sil.: $t_{1/2}$ > 1.8	205 / 40 / 30	**49.29**	**42.67** / 42.73 / 41.55	**33.7** / 33.68 / 33.78	**25.99** / 25.71 / 27.4	0.000	**37.96** / 36.77 / 44.05	**21.76** / 20.84 / 26.5

sil.: $t_{1/2}$ < 1.2 not sil.: $t_{1/2}$ > 1.9	205 / 37 / 27	**49.68**	**41.58** / 41.77 / 38.2	**34.1** / 33.83 / 35.59	**24.75** / 24.84 / 24.26	0.000	**34.06** / 32.72 / 41.51	**18.34** / 17.82 / 21.19
sil.: $t_{1/2}$ < 1.2 not sil.: $t_{1/2}$ > 2.0	205 / 35 / 25	**50.06**	**42.64** / 42.75 / 40.65	**35.48** / 35.42 / 35.86	**26.53** / 27.25 / 22.33	0.016	**40.19** / 39.46 / 43.16	**22.99** / 22.82 / 23.68
sil.: $t_{1/2}$ < 1.3 not sil.: $t_{1/2}$ > 1.4	211 / 64 / 51	**45.2**	**39.29** / 39.14 / 41.08	**33.79** / 33.15 / 35.89	**29.53** / 29.63 / 29.19	0.004	**40.39** / 39.09 / 44.69	**24.63** / 24.3 / 25.72
sil.: $t_{1/2}$ < 1.3 not sil.: $t_{1/2}$ > 1.5	211 / 57 / 46	**46.2**	**38.76** / 38.59 / 41.27	**31.06** / 30.01 / 34.98	**28.56** / 27.46 / 32.66	0.028	**40.51** / 39.38 / 44.7	**25.83** / 24.98 / 28.98
sil.: $t_{1/2}$ < 1.3 not sil.: $t_{1/2}$ > 1.6	211 / 50 / 40	**48.57**	**40.41** / 40.41 / 40.4	**29.76** / 28.87 / 33.51	**23.37** / 22.55 / 26.82	0.000	**39.13** / 38.22 / 43	**23.36** / 22.25 / 28
sil.: $t_{1/2}$ < 1.3 not sil.: $t_{1/2}$ > 1.7	211 / 44 / 34	**49.19**	**41.25** / 41.27 / 41	**30.56** / 30.08 / 32.85	**25.08** / 25 / 25.5	0.002	**42.06** / 41.19 / 46.23	**22.38** / 21.96 / 24.41
sil.: $t_{1/2}$ < 1.3 not sil.: $t_{1/2}$ > 1.8	211 / 40 / 30	**49.33**	**42.63** / 42.72 / 40.9	**33.65** / 33.44 / 34.77	**27.82** / 28.32 / 25.1	0.038	**37.71** / 36.45 / 44.3	**23.86** / 23.01 / 28.3
sil.: $t_{1/2}$ < 1.3 not sil.: $t_{1/2}$ > 1.9	211 / 37 / 27	**49.54**	**41.68** / 41.65 / 42.2	**34.24** / 34.02 / 35.51	**24.81** / 24.4 / 27.16	0.000	**35.02** / 33.75 / 42.22	**17.57** / 16.87 / 21.51
sil.: $t_{1/2}$ < 1.3 not sil.: $t_{1/2}$ > 2.0	211 / 35 / 25	**49.59**	**41.54** / 41.47 / 42.85	**36.19** / 36.81 / 32.46	**25.55** / 26.33 / 20.83	0.018	**35.5** / 34.59 / 40.97	**21.93** / 21.22 / 26.17
sil.: $t_{1/2}$ < 1.4 not sil.: $t_{1/2}$ > 1.4	216 / 64 / 51	**44.62**	**38.65** / 38.42 / 41.54	**33.36** / 32.52 / 36.22	**27.58** / 27.05 / 29.37	0.002	**40.15** / 38.9 / 44.37	**25.63** / 25.5 / 26.06
sil.: $t_{1/2}$ < 1.4 not sil.: $t_{1/2}$ > 1.5	216 / 57 / 46	**46.97**	**38.27** / 38.14 / 40.36	**30.70** / 29.85 / 33.93	**27.48** / 25.48 / 35.05	0.006	**39.33** / 37.94 / 44.6	**26.41** / 25.89 / 28.39
sil.: $t_{1/2}$ < 1.4 not sil.: $t_{1/2}$ > 1.6	216 / 50 / 40	**48.62**	**40.75** / 40.88 / 38.45	**29.68** / 28.71 / 33.88	**23.3** / 22.71 / 25.86	0.000	**38.5** / 37.66 / 42.16	**21.64** / 20.62 / 26.04
sil.: $t_{1/2}$ < 1.4 not sil.: $t_{1/2}$ > 1.7	216 / 44 / 34	**49.84**	**40.52** / 40.51 / 40.7	**30.16** / 29.55 / 33.2	**24.13** / 23.51 / 27.18	0.000	**39.32** / 38.34 / 44.09	**23.6** / 22.69 / 28.09
sil.: $t_{1/2}$ < 1.4 not sil.: $t_{1/2}$ > 1.8	216 / 40 / 30	**49.05**	**42.16** / 42.16 / 42.1	**33.66** / 33.38 / 35.17	**28.09** / 27.97 / 28.73	0.026	**36.7** / 35.26 / 44.45	**20.42** / 19.67 / 24.5
sil.: $t_{1/2}$ < 1.4 not sil.: $t_{1/2}$ > 1.9	216 / 37 / 27	**49.05**	**41.66** / 41.75 / 39.9	**33.96** / 33.79 / 34.96	**24.52** / 24.31 / 25.73	0.011	**34.67** / 33.37 / 42.27	**24.07** / 23.48 / 27.51

sil.: $t_{1/2}$ < 1.4 not sil.: $t_{1/2}$ > 2.0	216 / 35 / 25	**48.93**	**42.84** / 43.05 / 38.65	**36.01** / 35.5 / 39.15	**27.17** / 27.88 / 22.82	0.003	**34.61** / 33.6 / 40.86	**20.65** / 20.19 / 23.49		

Table A3: Performance of different machine learning methods for silencing dynamics model.

silencing threshold (early silenced, late silenced)	# of silenced genes / # of not silenced genes / size training set	**Logistic Regression** (error rate **total**)	**Elastic Net Regression** (error rate **total** / silenced / not silenced)	**Random Forest** (epigenetic & genomic features model) (error rate **total** / silenced / not silenced)		**Random Forest** (DNA sequence feature model on 100kb window) (error rate total / silenced / not silenced)	
				all features (error rate **total** / silenced / not silenced)	top features (error rate **total** / silenced / not silenced)	all features (error rate **total** / silenced / not silenced)	top features (error rate **total** / silenced / not silenced)
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.7 < $t_{1/2}$ < 1.0	74 / 48 / 38	**49.85**	**44.67** / 45.17 / 42.9	**39.05** / 40.41 / 36.96	**23.27** / 22.15 / 23.99	40.14 / 39.7 / 40.83	18.37 / 18.03 / 18.92
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.7 < $t_{1/2}$ < 1.1	74 / 62 / 50	**47.76**	**43.42** / 43.27 / 43.71	42.27 / 44.37 / 39.76	26.11 / 25.38 / 26.98	41.1 / 37.81 / 45.03	20.63 / 18.24 / 23.48
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.7 < $t_{1/2}$ < 1.2	74 / 75 / 59	**46.73**	**39.02** / 37.93 / 40.03	40.1 / 37.95 / 42.23	28.41 / 30.15 / 26.68	42.39 / 39.27 / 45.47	24.87 / 23.54 / 26.19
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.7 < $t_{1/2}$ < 1.3	74 / 79 / 59	**44.69**	**39.2** / 37.97 / 40.12	36.73 / 34.06 / 39.24	27.73 / 28.81 / 26.73	41.11 / 38.22 / 43.82	26.48 / 26 / 26.94
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.7 < $t_{1/2}$ < 1.4	74 / 84 / 59	**45.74**	**40.05** / 39.37 / 40.46	36.15 / 33.67 / 38.33	26.71 / 28.08 / 25.5	39.3 / 36.76 / 41.55	25.6 / 26.24 / 25.02

early sil.: $t_{1/2}$ < 0.5 late sil.: 0.8 < $t_{1/2}$ < 1.1	74 / 40 / 30	**50.5**	**44.16** / 43.93 / 45.15		35.37 / 34.85 / 36.34		25.06 / 26.38 / 22.64		43.35 / 36.89 / 55.3		22.65 / 21.62 / 24.55
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.8 < $t_{1/2}$ < 1.2	74 / 53 / 42	**47.03**	**39.12** / 38.56 / 40.73		41.53 / 39.48 / 44.4		25.95 / 22.84 / 30.29		42.69 / 35.3 / 53.02		23.37 / 22.49 / 24.6
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.8 < $t_{1/2}$ < 1.3	74 / 57 / 46	**47.69**	**37.62** / 37.48 / 37.95		**30.97** / 27.42 / 35.6		**21.4** / 19.71 / 23.6		41.66 / 35.08 / 50.21		22.78 / 21.81 / 24.04
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.8 < $t_{1/2}$ < 1.4	74 / 62 / 50	**46.21**	**37.76** / 37.42 / 38.46		39.27 / 36.94 / 42.06		**22.77** / 21.58 / 24.19		38.09 / 33.54 / 43.52		23.43 / 23.57 / 23.26
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.9 < $t_{1/2}$ < 1.2	74 / 36 / 26	**49.61**	**35.86** / 35.4 / 38.1		32.13 / 31.34 / 33.74		**21.99** / 21.37 / 23.28		28.56 / 19.95 / 46.28		19.27 / 18.6 / 20.66
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.9 < $t_{1/2}$ < 1.3	74 / 40 / 30	**49.03**	**34.13** / 32.52 / 41.2		**29.33** / 29.5 / 29		**21.56** / 20.43 / 23.24		28.7 / 20.59 / 43.7		19.11 / 18.43 / 20.35
early sil.: $t_{1/2}$ < 0.5 late sil.: 0.9 < $t_{1/2}$ < 1.4	74 / 45 / 35	**48.21**	**36.15** / 34.91 / 41		**29.73** / 28.46 / 31.83		23.82 / 22.32 / 26.27		31.43 / 25.97 / 40.4		19.62 / 18.54 / 21.38
early sil.: $t_{1/2}$ < 0.5 late sil.: 1.0 < $t_{1/2}$ < 1.4	74 / 36 / 26	**49.76**	**38.72** / 38.73 / 38.7		33.59 / 32.35 / 36.14		24.12 / 22.08 / 28.31		31.06 / 25.19 / 43.11		17.25 / 15.59 / 20.67
early sil.: $t_{1/2}$ < 0.6 late sil.: 0.7 < $t_{1/2}$ < 1.0	104 / 48 / 38	**50.57**	**47.01** / 47.52 / 43.6		43.36 / 43.21 / 43.68		27.84 / 28.98 / 25.38		44.59 / 43 / 48.04		19.96 / 20.38 / 19.04

early sil.: $t_{1/2}$ < 0.6 late sil.: 0.7 < $t_{1/2}$ < 1.1	104 / 62 / 50	**48.86**	**44.89** / 44.92 / 44.79	40.15 / 38.37 / 43.13	27.34 / 28.01 / 26.2	44.67 / 40.75 / 51.26	19.57 / 20.21 / 18.48
early sil.: $t_{1/2}$ < 0.6 late sil.: 0.7 < $t_{1/2}$ < 1.2	104 / 75 / 60	**45.64**	**39.37** / 38.77 / 41.13	43.43 / 41.52 / 46.07	30.14 / 27.98 / 33.14	41.46 / 35.12 / 50.27	23.06 / 21.19 / 25.65
early sil.: $t_{1/2}$ < 0.6 late sil.: 0.7 < $t_{1/2}$ < 1.3	104 / 79 / 63	**45.44**	**39.19** / 39.37 / 38.75	43.33 / 40.75 / 46.71	29.44 / 27.04 / 32.61	39.61 / 32.65 / 48.76	19.52 / 17.1 / / 22.71
early sil.: $t_{1/2}$ < 0.6 late sil.: 0.7 < $t_{1/2}$ < 1.4	104 / 84 / 67	**45.32**	**40.61** / 40.42 / 41.03	42.76 / 40.67 / 45.34	32.03 / 29.94 / 34.6	40.71 / 34.46 / 48.45	25.51 / 23.89 / 27.52
early sil.: $t_{1/2}$ < 0.6 late sil.: 0.8 < $t_{1/2}$ < 1.1	104 / 40 / 30	**49.78**	**44.45** / 44.1 / 47	36.36 / 35.7 / 38.08	27.88 / 26.73 / 30.84	41.13 / 35.5 / 55.75	30.65 / 29.17 / 34.5
early sil.: $t_{1/2}$ < 0.6 late sil.: 0.8 < $t_{1/2}$ < 1.2	104 / 53 / 42	**48.63**	**38.51** / 38.06 / 41.05	42.68 / 40.64 / 46.69	28.44 / 24.14 / 36.87	38.15 / 29.1 / / 55.92	23.81 / 24.08 / 23.28
early sil.: $t_{1/2}$ < 0.6 late sil.: 0.8 < $t_{1/2}$ < 1.3	104 / 57 / 46	**46.44**	**37.99** / 37.91 / 38.36	**29.18** / 26.81 / 33.5	23.3 / 21.65 / 26.32	37.7 / 29.27 / / 53.09	22.36 / 21.23 / 24.42
early sil.: $t_{1/2}$ < 0.6 late sil.: 0.8 < $t_{1/2}$ < 1.4	104 / 62 / 50	**46.08**	**38.02** / 37.69 / 39.46	41.58 / 39.37 / 45.3	28.66 / 26.06 / 33.01	39.38 / 30.75 / 53.87	24.88 / 23.12 / 27.84
early sil.: $t_{1/2}$ < 0.6 late sil.: 0.9 < $t_{1/2}$ < 1.2	104 / 36 / 26	**50.07**	**35.61** / 35.03 / 40.2	31.74 / 30.12 / 36.44	24.34 / 23.82 / 25.86	**27.91** / 21.94 / 45.17	**15.73** / 15.94 / 15.11

early sil.: $t_{1/2}$ < 0.6 late sil.: 0.9 < $t_{1/2}$ < 1.3	104 / 40 / 30	**49.32**	**34.08** / 33.41 / 39.05	32.54 / 31.05 / 36.44	24.84 / 23.83 / 27.47	**26.44** / 19.73 / 43.9	**15.57** / 15.38 / 16.05
early sil.: $t_{1/2}$ < 0.6 late sil.: 0.9 < $t_{1/2}$ < 1.4	104 / 45 / 35	**47.65**	**36.59** / 36 / 40.65	33.2 / 31.69 / 36.67	26.42 / 25.1 / / 29.48	**27.28** / 19.71 / 44.75	17.62 / 16.23 / 20.84
early sil.: $t_{1/2}$ < 0.6 late sil.: 1.0 < $t_{1/2}$ < 1.4	104 / 36 / 26	**50.41**	**37.98** / 37.85 / 38.95	37.93 / 35.69 / 44.4	27.28 / 24.76 / 34.57	30.97 / 26.38 / 44.22	18.63 / 17.83 / 20.94
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.7 < $t_{1/2}$ < 1.0	125 / 48 / 38	**50.36**	**47.63** / 48.15 / 43.15	46.31 / 45.93 / 47.3	32.47 / 31.51 / 34.96	48.47 / 48.86 / 47.46	21.63 / 22.93 / 18.25
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.7 < $t_{1/2}$ < 1.1	125 / 62 / 50	**49.67**	**44.29** / 44.24 / 44.62	43.76 / 43.36 / 44.56	32.09 / 31.58 / 33.11	50.02 / 48.96 / 52.16	25.63 / 28.66 / 19.55
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.7 < $t_{1/2}$ < 1.2	125 / 75 / 60	**47.23**	**40.09** / 39.95 / 40.73	45.96 / 43.58 / 49.93	35.63 / 34.15 / 38.09	46.88 / 43.52 / 52.48	23 / 25.18 / 19.36
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.7 < $t_{1/2}$ < 1.3	125 / 79 / 63	**45.33**	**38.99** / 38.67 / 40.25	38.63 / 35.8 / 43.1	26.1 / 24.23 / 29.06	45.25 / 41.95 / 50.46	25.74 / 25.01 / 26.91
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.7 < $t_{1/2}$ < 1.4	125 / 84 / 67	**45.47**	**39.47** / 39.09 / 40.76	45.55 / 43.9 / 48	35 / 34.14 / 36.28	45.87 / 43.5 / 49.38	25.51 / 24.96 / 26.33
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.8 < $t_{1/2}$ < 1.1	125 / 40 / 30	**49.89**	**46.08** / 45.93 / 47.5	40.11 / 40.6 / 38.57	28.36 / 27.94 / 29.68	44.39 / 40.83 / 55.5	33.26 / 34.48 / 29.45

early sil.: $t_{1/2}$ < 0.7 late sil.: 0.8 < $t_{1/2}$ < 1.2	125 / 53 / 42	**49.39**	**39.37** 39.14 41.09	/ /	35.48 32.43 42.67	/ /	27.56 26.38 30.35	/ /	42.99 37.58 55.74	/ /	27.64 25.84 31.89	/ /
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.8 < $t_{1/2}$ < 1.3	125 / 57 / 46	**47.53**	**38.16** 37.66 41.77	/ /	44.79 42.46 49.91	/ /	28.67 27.02 32.3	/ /	41.16 35.12 54.39	/ /	26.54 25.42 28.98	/ /
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.8 < $t_{1/2}$ < 1.4	125 / 62 / 50	**46.24**	**38.36** 37.89 41.25	/ /	38.93 36.68 43.45	/ /	30.62 29.05 33.76	/ /	43.34 37.78 54.55	/ /	26.68 24.91 30.26	/ /
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.9 < $t_{1/2}$ < 1.2	125 / 36 / 26	**50.18**	**35.72** 35.04 42.5	/ /	31.16 30.16 34.66	/ /	24.29 24.04 25.18	/ /	28.78 23.68 46.5	/ /	23.04 22.5 24.94	/ /
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.9 < $t_{1/2}$ < 1.3	125 / 40 / 30	**50.67**	**34.65** 33.95 41.3	/ /	32.76 31.54 36.58	/ /	23.41 22.32 26.82	/ /	**27.04** 21.52 44.3	/ /	19.81 19.46 20.9	/ /
early sil.: $t_{1/2}$ < 0.7 late sil.: 0.9 < $t_{1/2}$ < 1.4	125 / 45 / 35	**48.75**	**35.88** 35.21 41.9	/ /	34.53 33.12 38.47	/ /	25.44 24.62 27.72	/ /	30.11 24.19 46.53	/ /	21.39 20.42 24.09	/ /
early sil.: $t_{1/2}$ < 0.7 late sil.: 1.0 < $t_{1/2}$ < 1.4	125 / 36 / 26	**50.58**	**38.17** 37.93 40.5	/ /	40.93 39.95 44.36	/ /	28.97 25.11 42.37	/ /	32.1 28.83 43.44	/ /	22.91 21.65 27.28	/ /

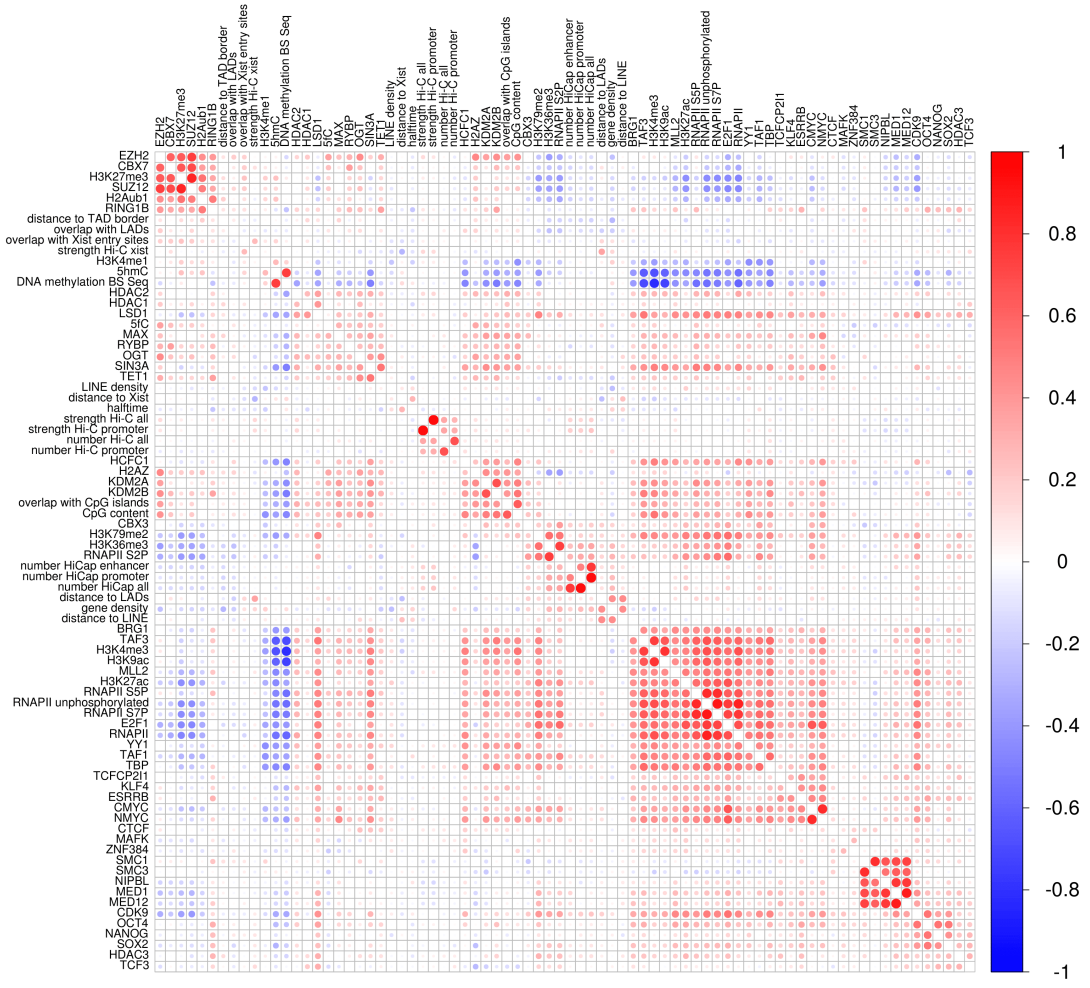

Figure A5: Feature correlation matrix. It shows the Pearson correlation coefficient for every pair of features used in the model and it is computed based on all 280 genes with estimated half-times from the PRO-seq data. Red indicates high positive correlation and blue a high negative correlation. One can observe blocks of correlate features. For example, the active marks (RNAPII, H3K4me3, H3K27ac and others) are highly correlated amongst each other while repressive features, such as PRC1 and PRC2 components and H3K27me3 form another positively correlated block but are negatively correlated with many active mark features.

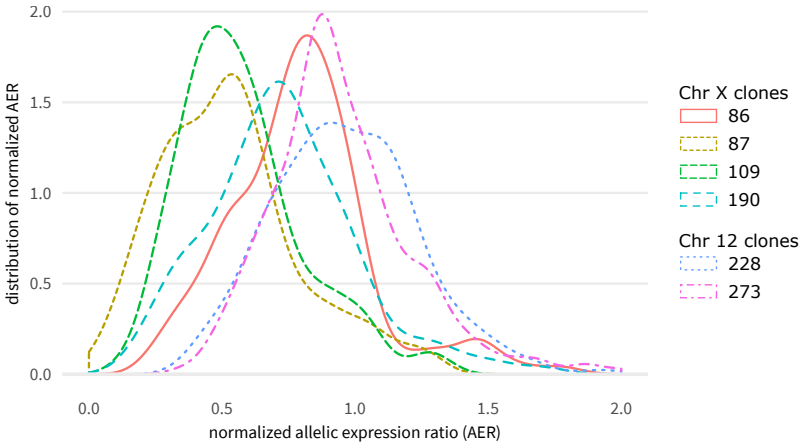

Figure A6: Distribution of normalized allelic expression ratios (AER) for each clone. Shown is the distribution of normalized AER for all genes in each of the six clones with ectopic *Xist* expression (four on Chromosome X and two autosomal locations on Chromosome 12). A normalized AER below one indicates that the gene is silenced after 2 days of doxycycline induction in the respective clone. The figure shows that overall gene silencing is less efficient on the clones on Chromosome 12 compared to the clones on Chromosome X.

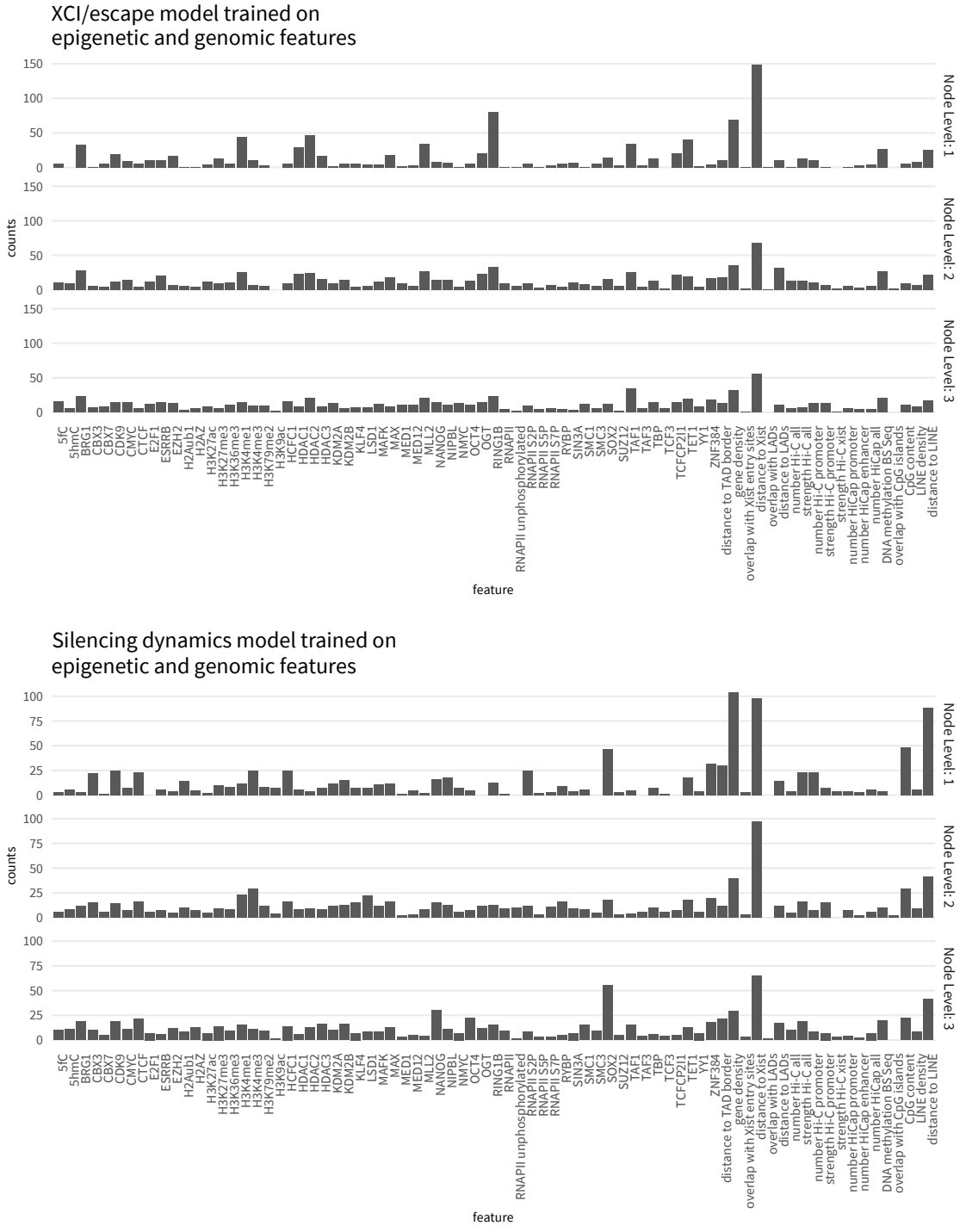

Figure A7: Split point frequencies of the RF models trained on epigenetic and genomic features. For each epigenetic or genomic feature, the bar indicates how often that feature was used as a split point in the respective node level counted over all decision trees in the RF model. The first three node levels are displayed, where the first node level represents the first split point of a decision tree.

Figure A8: Split point frequencies of the RF models trained on DNA sequence features. For each DNA sequence feature, the bar indicates how often that feature was used as a split point in the respective node level counted over all decision trees in the RF model. The first three node levels are displayed, where the first node level represents the first split point of a decision tree.

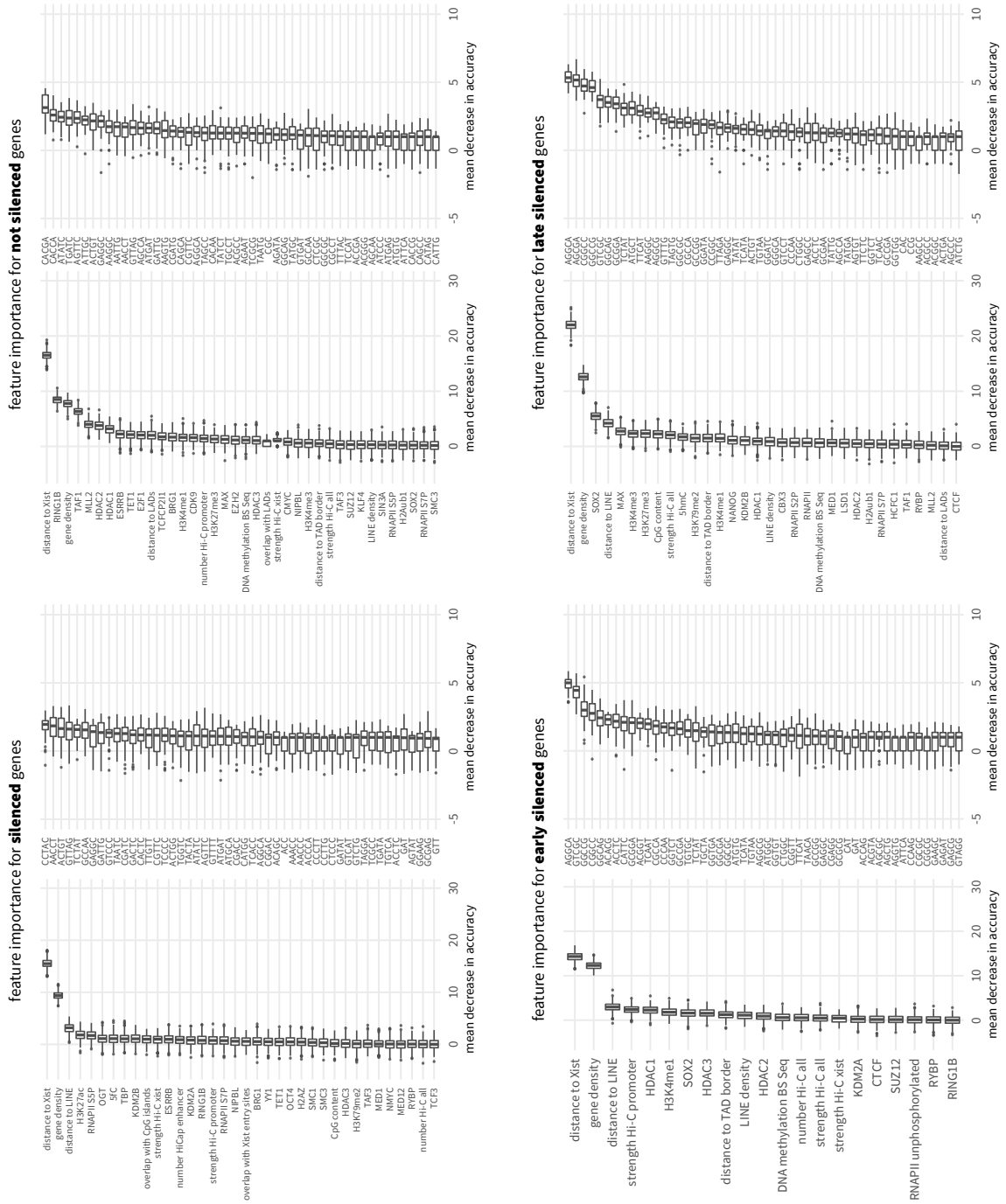

Figure A9: Feature importance for the XCI/escape and silencing dynamics model. The importance of features for Random Forest classification is measured by the mean decrease in accuracy (MDA), which is defined as the average decrease in model accuracy after permuting the values in each feature. The feature with the highest MDA (e.g. distance to *Xist*) is the most important feature for the classification. Each box in the plot corresponds to a model feature and represents the distribution of that feature's MDA over 500 Random Forest models. For simplicity, only features with MDA higher than 0 are shown.

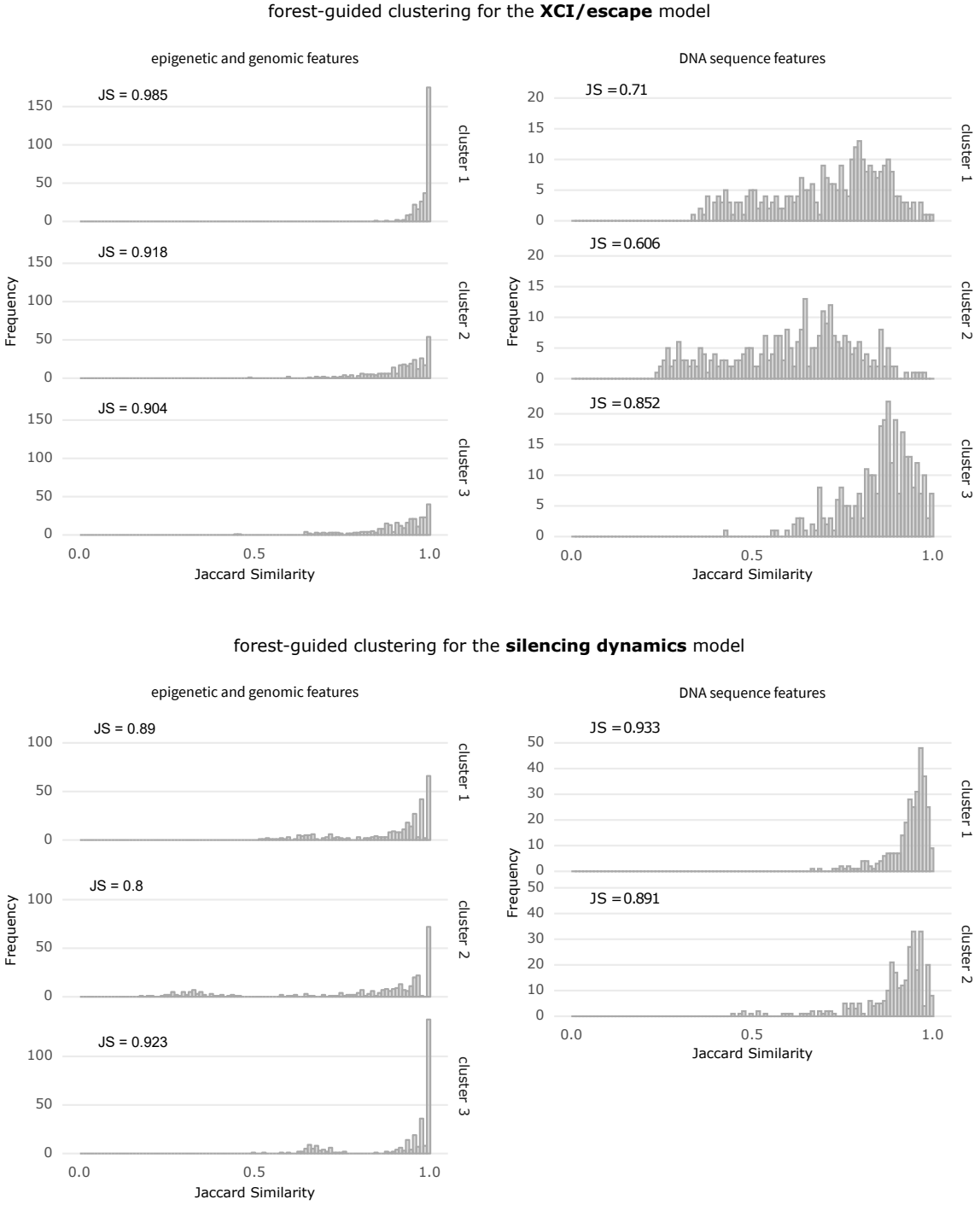

Figure A10: Cluster stability analysis for selection of optimal number of k clusters. The cluster stability analysis shows the distribution of Jaccard Similarity (JS) for each cluster over 300 bootstrap runs. Average JS values over 300 runs are reported for each cluster.

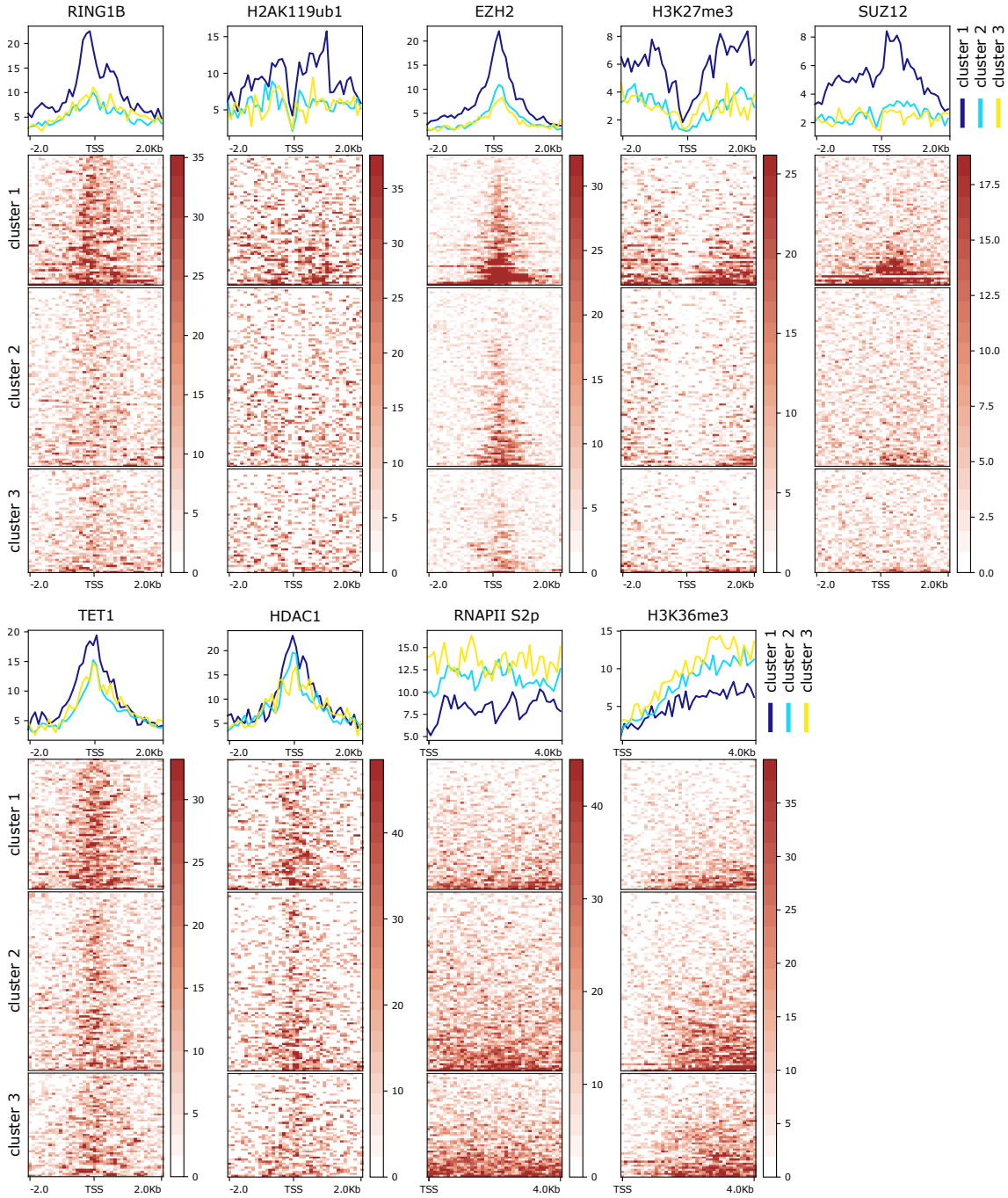

Figure A11: Enriched features from the XCI/escape model clustering. The normalized signal of epigenetic marks and other factors computed in the +/ – 2000 bp genomic region around each gene promoter is shown in the heatmaps for each of the three clusters separately. Average profile plots for the same factors are also shown above the heatmaps to highlight overall differences between clusters. Shown here are only those features which, according to the p-value of an ANOVA test, were the top most significantly different among clusters in the XCI/escape model.

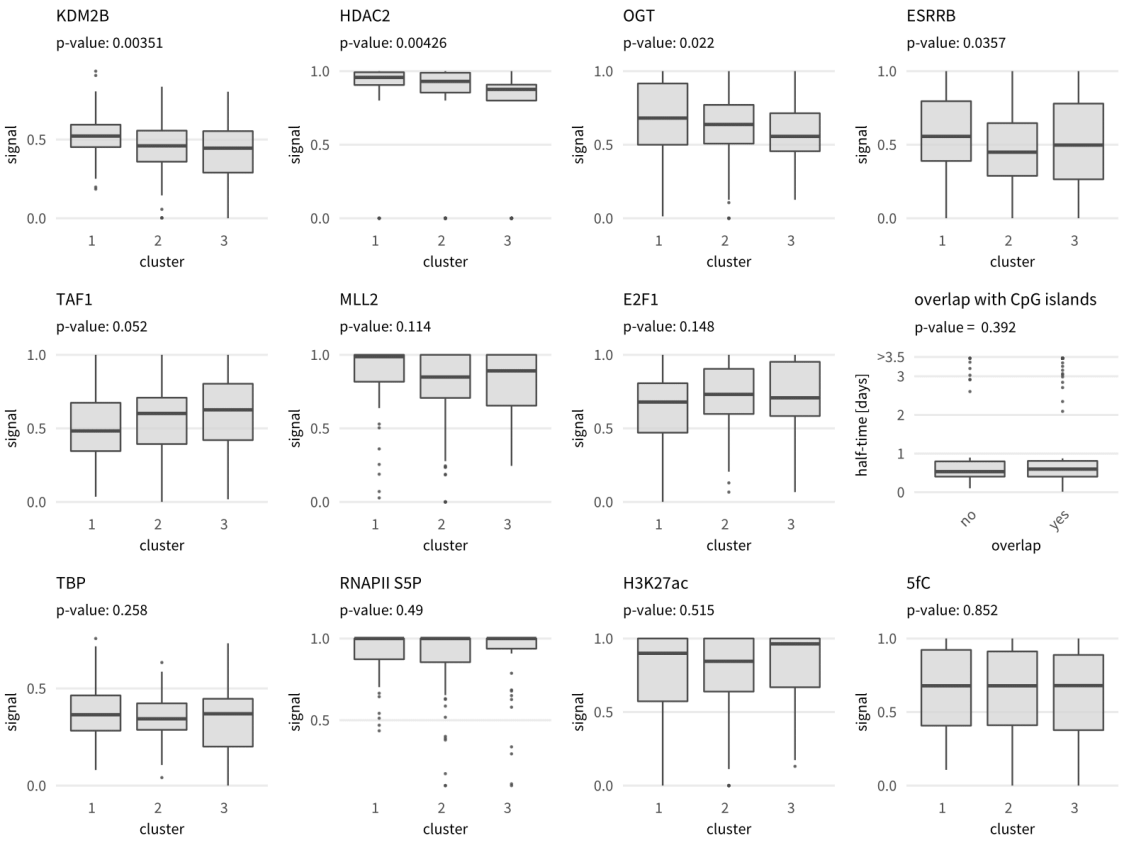

Figure A12: Top features from the XCI/escape Random Forest. The feature distributions at each gene are shown in the boxplots for each of the three clusters separately to highlight overall differences between clusters (p-value of ANOVA test indicates the significance of the differences between clusters). Shown here are epigenetic and genomic features that are among the top 10 features in the XCI/escape Random Forest model but are not among the top significant ones from the clustering.

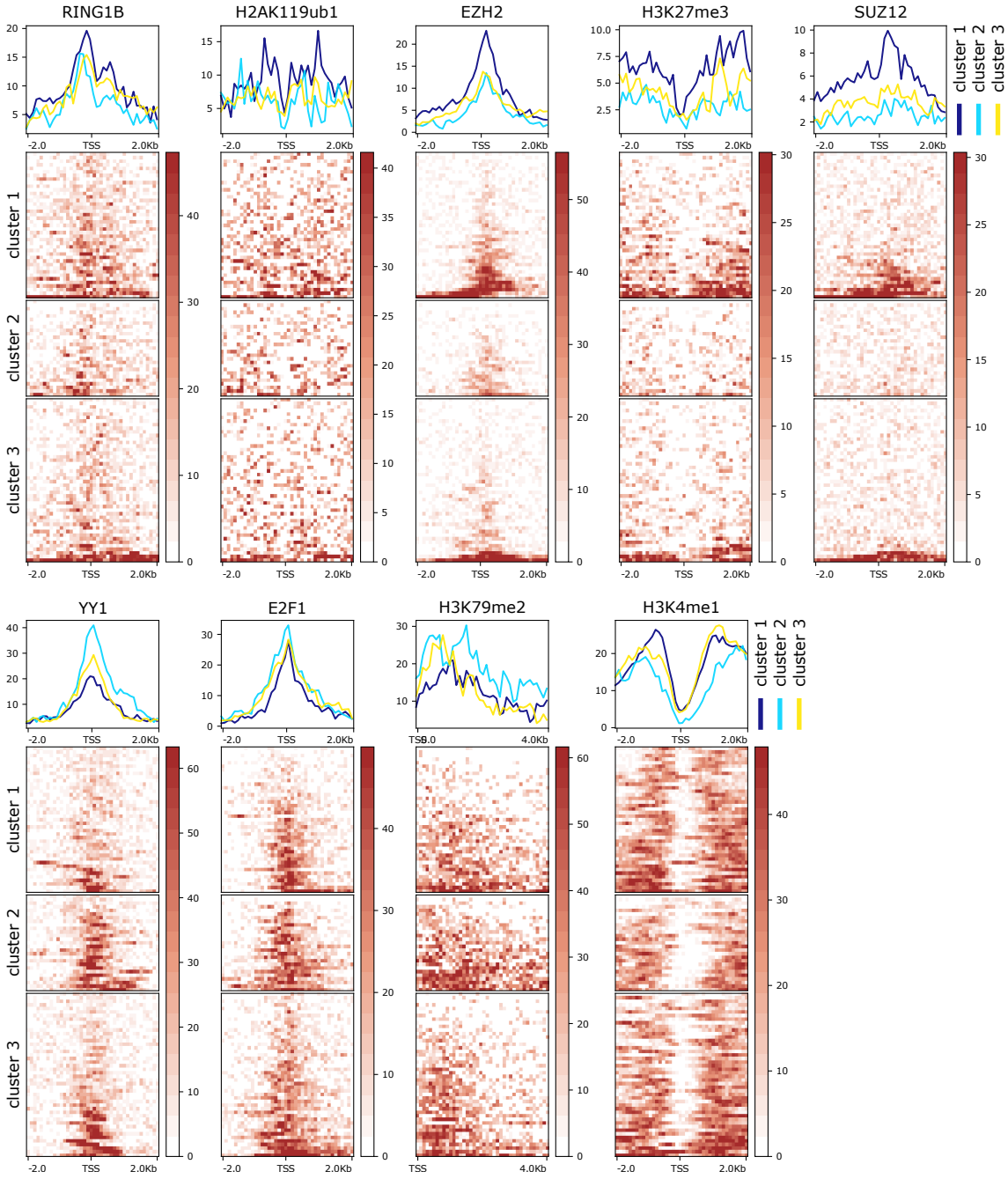

Figure A13: Enriched features from the silencing dynamics model clustering. The normalized signal of epigenetic marks and other factors, computed in the +/ – 2000 bp genomic region around each gene promoter is shown in the heatmaps for each of the three clusters separately. Average profile plots for the same factors are also shown above the heatmaps to highlight overall differences between clusters. Shown here are only those features which, according to the p-value of an ANOVA test, were the top most significantly different among clusters in the silencing dynamics model.

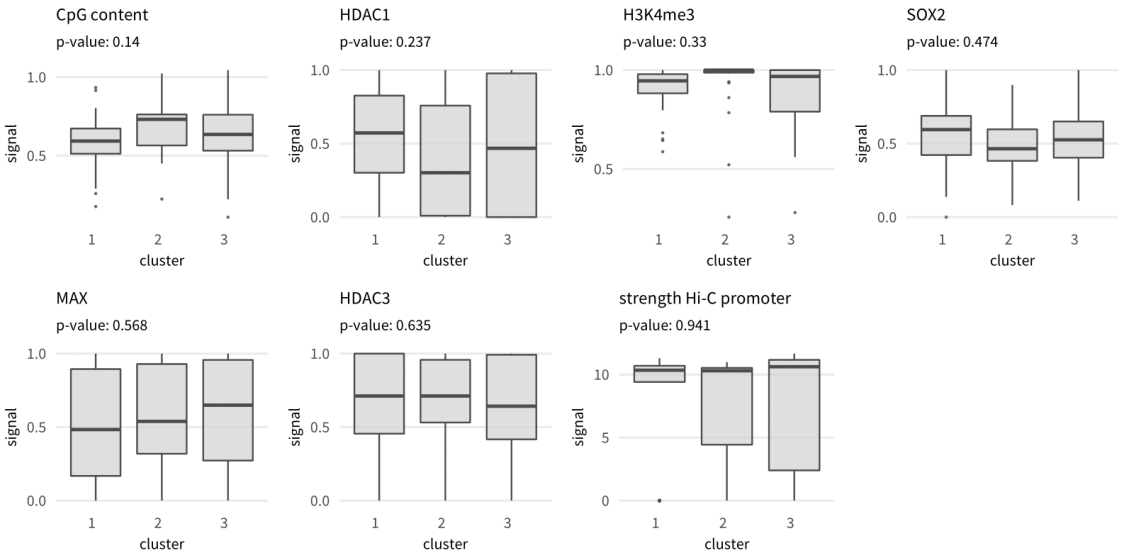

Figure A14: Top features from the silencing dynamics Random Forest. The feature distributions at each gene are shown in the boxplots for each of the three clusters separately to highlight overall differences between clusters (p-value of ANOVA test indicates the significance of the differences between clusters). Shown here are epigenetic and genomic features that are among the top 8 features in the silencing dynamics Random Forest model but are not among the top significant ones from the clustering.

BIBLIOGRAPHY

Agrelo, Ruben et al. (Apr. 2009). "SATB1 defines the developmental context for gene silencing by Xist in lymphoma and embryonic cells." In: *Developmental Cell* 16.4, pp. 507–516. DOI: 10.1016/j.devcel.2009.03.006.

Albertini, R (Oct. 2001). "HPRT mutations in humans: biomarkers for mechanistic studies." In: *Mutation Research/Reviews in Mutation Research* 489.1, pp. 1–16. ISSN: 13835742. DOI: 10.1016/S1383-5742(01)00064-3.

Alberts, Bruce, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter (Dec. 2014). *Molecular Biology Of The Cell*. 6th ed. New York, NY: Garland Science, p. 1464. ISBN: 0815344643.

Almeida, Mafalda et al. (June 2017). "PCGF3/5-PRC1 initiates Polycomb recruitment in X chromosome inactivation." In: *Science* 356.6342, pp. 1081–1084. DOI: 10.1126/science.aal2512.

Ambardar, Sheetal, Rikita Gupta, Deepika Trakroo, Rup Lal, and Jyoti Vakhlu (2016). "High throughput sequencing: an overview of sequencing chemistry." In: *Indian journal of microbiology* 56.4, pp. 394–404.

Andergassen, Daniel et al. (Aug. 2017). "Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression." In: *eLife* 6. DOI: 10.7554/{eLife}.25125.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber (Jan. 2015). "HTSeq —a Python framework to work with high-throughput sequencing data." In: *Bioinformatics* 31.2, pp. 166–169. DOI: 10.1093/bioinformatics/btu638.

Augui, Sandrine, Elphège P Nora, and Edith Heard (June 2011). "Regulation of X-chromosome inactivation by the X-inactivation centre." In: *Nature Reviews. Genetics* 12.6, pp. 429–442. DOI: 10.1038/nrg2987.

Babak, Tomas, Brian Deveale, Christopher Armour, Christopher Raymond, Michele A Cleary, Derek van der Kooy, Jason M Johnson, and Lee P Lim (Nov. 2008). "Global survey of genomic imprinting by transcriptome sequencing." In: *Current Biology* 18.22, pp. 1735–1741. ISSN: 0960-9822. DOI: 10.1016/j.cub.2008.09.044.

Balaton, Bradley P and Carolyn J Brown (Apr. 2016). "Escape artists of the X chromosome." In: *Trends in Genetics* 32.6, pp. 348–359. DOI: 10.1016/j.tig.2016.03.007.

Balaton, Bradley P, Thomas Dixon-McDougall, Samantha B Peeters, and Carolyn J Brown (Aug. 2018). "The eXceptional nature of the X chromosome." In: *Human Molecular Genetics* 27.R2, R242–R249. DOI: 10.1093/hmg/ddy148.

Barros de Andrade E Sousa, Lisa et al. (June 2019). "Kinetics of Xist-induced gene silencing can be predicted from combinations of epigenetic and genomic features." In: *Genome Research* 29.7, pp. 1087–1099. DOI: 10.1101/gr.245027.118.

Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao (May 2007). "High-resolution profiling of histone methylations in the human genome." In: *Cell* 129.4, pp. 823–837. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.05.009.

Belle, I de, S Cai, and T Kohwi-Shigematsu (Apr. 1998). "The genomic sequences bound to special AT-rich sequence-binding protein 1 (SATB1) in vivo in Jurkat T cells are tightly associated with

the nuclear matrix at the bases of the chromatin loops." In: *The Journal of Cell Biology* 141.2, pp. 335–348. DOI: 10.1083/jcb.141.2.335.

Berletch, Joel B, Wenxiu Ma, Fan Yang, Jay Shendure, William S Noble, Christine M Disteche, and Xinxian Deng (Mar. 2015). "Escape from X inactivation varies in mouse tissues." In: *PLoS Genetics* 11.3, e1005079. DOI: 10.1371/journal.pgen.1005079.

Bode, J, Y Kohwi, L Dickinson, T Joh, D Klehr, C Mielke, and T Kohwi-Shigematsu (Jan. 1992). "Biological significance of unwinding capability of nuclear matrix-associating DNAs." In: *Science* 255.5041, pp. 195–197. DOI: 10.1126/science.1553545.

Borensztein, Maud et al. (Jan. 2017). "Xist-dependent imprinted X inactivation and the early developmental consequences of its failure." In: *Nature Structural & Molecular Biology* 24.3, pp. 226–233. DOI: 10.1038/nsmb.3365.

Borsani, G et al. (May 1991). "Characterization of a murine gene expressed from the inactive X chromosome." In: *Nature* 351.6324, pp. 325–329. DOI: 10.1038/351325a0.

Bousard, Aurélie et al. (Aug. 2019). "The role of Xist-mediated Polycomb recruitment in the initiation of X-chromosome inactivation." In: *EMBO Reports*, e48019. DOI: 10.15252/embr.201948019.

Breiman, L, JH Friedman, RA Olshen, and CJ Stone (1984). "Classification and regression trees." In: *Wadsworth and Brooks, Monterey, CA*.

Breiman, Leo (Aug. 1996). "Bagging predictors." In: *Machine learning* 24.2, pp. 123–140. ISSN: 0885-6125. DOI: 10.1007/{BF00058655}.

– (2001). "Random Forests." In: *Machine Learning*.

Breiman, Leo and Adele Cutler (2003). "Manual for Setting Up, Using, and Understanding Random Forest V4.0." In: *stat.berkeley*.

Brockdorff, N, A Ashworth, G F Kay, P Cooper, S Smith, V M McCabe, D P Norris, G D Penny, D Patel, and S Rastan (May 1991). "Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome." In: *Nature* 351.6324, pp. 329–331. DOI: 10.1038/351329a0.

Brockdorff, Neil (Nov. 2017). "Polycomb complexes in X chromosome inactivation." In: *Philos. Trans. R. Soc. Lond. B. Biol. Sci* 372.1733. DOI: 10.1098/rstb.2017.0021.

– (Oct. 2018). "Local Tandem Repeat Expansion in Xist RNA as a Model for the Functionalisation of ncRNA." In: *Non-coding RNA* 4.4. DOI: 10.3390/ncrna4040028.

Brown, C J, A Ballabio, J L Rupert, R G Lafreniere, M Grompe, R Tonlorenzi, and H F Willard (Jan. 1991). "A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome." In: *Nature* 349.6304, pp. 38–44. DOI: 10.1038/349038a0.

Brown, C J, B D Hendrich, J L Rupert, R G Lafrenière, Y Xing, J Lawrence, and H F Willard (Oct. 1992). "The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus." In: *Cell* 71.3, pp. 527–542. DOI: 10.1016/0092-8674(92)90520-m.

Bureau, Alexandre, Josée Dupuis, Kathleen Falls, Kathryn L Lunetta, Brooke Hayward, Tim P Keith, and Paul Van Eerdewegh (Feb. 2005). "Identifying SNPs predictive of phenotype using random forests." In: *Genetic Epidemiology* 28.2, pp. 171–182. DOI: 10.1002/gepi.20041.

Calabrese, J Mauro, Wei Sun, Lingyun Song, Joshua W Mugford, Lucy Williams, Della Yee, Joshua Starmer, Piotr Mieczkowski, Gregory E Crawford, and Terry Magnuson (Nov. 2012). "Site-specific silencing of regulatory elements as a mechanism of X inactivation." In: *Cell* 151.5, pp. 951–963. DOI: 10.1016/j.cell.2012.10.037.

Cao, Ru, Liangjun Wang, Hengbin Wang, Li Xia, Hediye Erdjument-Bromage, Paul Tempst, Richard S Jones, and Yi Zhang (Nov. 2002). "Role of histone H3 lysine 27 methylation in Polycomb-group silencing." In: *Science* 298.5595, pp. 1039–1043. ISSN: 1095-9203. DOI: 10.1126/science.1076997.

Carrel, Laura and Carolyn J Brown (Nov. 2017). "When the Lyon(ized chromosome) roars: ongoing expression from an inactive X chromosome." In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372.1733. DOI: 10.1098/rstb.2016.0355.

Carrel, Laura, Chungoo Park, Svitlana Tyekucheva, John Dunn, Francesca Chiaromonte, and Kateryna D Makova (Sept. 2006). "Genomic environment predicts expression patterns on the human inactive X chromosome." In: *PLoS Genetics* 2.9, e151. DOI: 10.1371/journal.pgen.0020151.

Chen, Chih-Yu et al. (Nov. 2016). "YY1 binding association with sex-biased transcription revealed through X-linked transcript levels and allelic binding analyses." In: *Sci Rep* 6, p. 37324. DOI: 10.1038/srep37324.

Chow, Jennifer C et al. (June 2010). "LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation." In: *Cell* 141.6, pp. 956–969. ISSN: 1097-4172. DOI: 10.1016/j.cell.2010.04.042.

Chu, Ci, Qiangfeng Cliff Zhang, Simão Teixeira da Rocha, Ryan A Flynn, Maheetha Bharadwaj, J Mauro Calabrese, Terry Magnuson, Edith Heard, and Howard Y Chang (Apr. 2015). "Systematic discovery of Xist RNA binding proteins." In: *Cell* 161.2, pp. 404–416. DOI: 10.1016/j.cell.2015.03.025.

Cokus, Shawn J, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini, and Steven E Jacobsen (Mar. 2008). "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." In: *Nature* 452.7184, pp. 215–219. ISSN: 1476-4687. DOI: 10.1038/nature06745.

Colognori, David, Hongjae Sunwoo, Andrea J Kriz, Chen-Yu Wang, and Jeannie T Lee (Apr. 2019). "Xist Deletional Analysis Reveals an Interdependency between Xist RNA and Polycomb Complexes for Spreading along the Inactive X." In: *Mol Cell* 74.1, 101–117.e10. ISSN: 10972765. DOI: 10.1016/j.molcel.2019.01.015.

Comer, F I and G W Hart (July 2001). "Reciprocity between O-GlcNAc and O-phosphate on the carboxyl terminal domain of RNA polymerase II." In: *Biochemistry* 40.26, pp. 7845–7852. DOI: 10.1021/bi0027480.

Cook, Charles E, Rodrigo Lopez, Oana Stroe, Guy Cochrane, Cath Brooksbank, Ewan Birney, and Rolf Apweiler (Jan. 2019). "The European Bioinformatics Institute in 2018: tools, infrastructure and training." In: *Nucleic Acids Research* 47.D1, pp. D15–D22. DOI: 10.1093/nar/gky1124.

Cooper, Sarah et al. (Nov. 2016). "Jarid2 binds mono-ubiquitylated H2A lysine 119 to mediate crosstalk between Polycomb complexes PRC1 and PRC2." In: *Nat Commun* 7, p. 13661. DOI: 10.1038/ncomms13661.

Core, Leighton J, André L Martins, Charles G Danko, Colin T Waters, Adam Siepel, and John T Lis (Dec. 2014). "Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers." In: *Nature Genetics* 46.12, pp. 1311–1320. DOI: 10.1038/ng.3142.

Cutler, Adele and Leo Breiman (June 2004). *RAFT (RAndom Forest Tool)*. WEBSITE.

Czermin, Birgit, Raffaella Melfi, Donna McCabe, Volker Seitz, Axel Imhof, and Vincenzo Pirrotta (Oct. 2002). "Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase

activity that marks chromosomal Polycomb sites." In: *Cell* 111.2, pp. 185–196. ISSN: 0092-8674. DOI: 10.1016/s0092-8674(02)00975-3.

Danko, Charles G, Stephanie L Hyland, Leighton J Core, Andre L Martins, Colin T Waters, Hyung Won Lee, Vivian G Cheung, W Lee Kraus, John T Lis, and Adam Siepel (May 2015). "Identification of active transcriptional regulatory elements from GRO-seq data." In: *Nature Methods* 12.5, pp. 433–438. DOI: 10.1038/nmeth.3329.

Dechat, Thomas, Stephen A Adam, Pekka Taimen, Takeshi Shimi, and Robert D Goldman (Nov. 2010). "Nuclear lamins." In: *Cold Spring Harb Perspect Biol* 2.11, a000547. DOI: 10.1101/cshperspect.a000547.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner (Feb. 2002). "Capturing chromosome conformation." In: *Science* 295.5558, pp. 1306–1311. ISSN: 1095-9203. DOI: 10.1126/science.1067799.

Dickinson, L A, T Joh, Y Kohwi, and T Kohwi-Shigematsu (Aug. 1992). "A tissue-specific MAR/SAR DNA-binding protein with unusual binding site recognition." In: *Cell* 70.4, pp. 631–645. DOI: 10.1016/0092-8674(92)90432-c.

Dixon, Jesse R, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren (Apr. 2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." In: *Nature* 485.7398, pp. 376–380. DOI: 10.1038/nature11082.

Doshi-Velez, Finale and Been Kim (Feb. 2017). "Towards A Rigorous Science of Interpretable Machine Learning." In: *arXiv*.

Díaz-Uriarte, Ramón and Sara Alvarez de Andrés (Jan. 2006). "Gene selection and classification of microarray data using random forest." In: *BMC Bioinformatics* 7, p. 3. DOI: 10.1186/1471-2105-7-3.

Edgar, Ron, Michael Domrachev, and Alex E Lash (Jan. 2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." In: *Nucleic Acids Research* 30.1, pp. 207–210. DOI: 10.1093/nar/30.1.207.

Efron, Bradley and Robert J. Tibshirani (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton: Chapman and Hall/CRC. ISBN: 978-0-412-04231-7. DOI: 10.1007/978-1-4899-4541-9.

Efron, Bradley and Robert Tibshirani (June 1997). "Improvements on Cross-Validation: The .632+ Bootstrap Method." In: *Journal of the American Statistical Association* 92.438, pp. 548–560. ISSN: 0162-1459. DOI: 10.1080/01621459.1997.10474007.

Engreitz, Jesse M et al. (Aug. 2013). "The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome." In: *Science* 341.6147, p. 1237973. DOI: 10.1126/science.1237973.

Ernst, Jason and Manolis Kellis (Mar. 2012). "ChromHMM: automating chromatin-state discovery and characterization." In: *Nature Methods* 9.3, pp. 215–216. DOI: 10.1038/nmeth.1906.

Filippova, Galina N, Mimi K Cheng, James M Moore, Jean-Pierre Truong, Ying J Hu, Di Kim Nguyen, Karen D Tsuchiya, and Christine M Disteche (Jan. 2005). "Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development." In: *Dev Cell* 8.1, pp. 31–42. ISSN: 1534-5807. DOI: 10.1016/j.devcel.2004.10.018.

Galupa, Rafael and Edith Heard (Apr. 2015). "X-chromosome inactivation: new insights into cis and trans regulation." In: *Current Opinion in Genetics & Development* 31, pp. 57–66. DOI: 10.1016/j.gde.2015.04.002.

– (Nov. 2018). "X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation." In: *Annual Review of Genetics* 52, pp. 535–566. DOI: 10.1146/annurev-genet-120116-024611.

Gardiner-Garden, M and M Frommer (July 1987). "CpG islands in vertebrate genomes." In: *J Mol Biol* 196.2, pp. 261–282. DOI: 10.1016/0022-2836(87)90689-9.

Geisser, Seymour (June 1975). "The Predictive Sample Reuse Method with Applications." In: *Journal of the American Statistical Association* 70.350, pp. 320–328. ISSN: 0162-1459. DOI: 10.1080/01621459.1975.10479865.

Gendrel, Anne-Valerie and Edith Heard (June 2014). "Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation." In: *Annual Review of Cell and Developmental Biology* 30, pp. 561–580. DOI: 10.1146/annurev-cellbio-101512-122415.

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau (Nov. 2008). "Random Forests: some methodological insights." In: *arXiv*.

Gontan, Cristina, Eskeatnaf Mulugeta Achame, Jeroen Demmers, Tahsin Stefan Barakat, Eveline Rentmeester, Wilfred van IJcken, J Anton Grootegoed, and Joost Gribnau (Apr. 2012). "RNF12 initiates X-chromosome inactivation by targeting REX1 for degradation." In: *Nature* 485.7398, pp. 386–390. DOI: 10.1038/nature11070.

Grant, Jennifer et al. (July 2012). "Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation." In: *Nature* 487.7406, pp. 254–258. DOI: 10.1038/nature11171.

Guelen, Lars et al. (June 2008). "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." In: *Nature* 453.7197, pp. 948–951. ISSN: 1476-4687. DOI: 10.1038/nature06947.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.

Helmuth, Johannes et al. (Oct. 2016). "normR: Regime enrichment calling for ChIP-seq data." In: *BioRxiv*. DOI: 10.1101/082263.

Hennig, Christian (July 2008). "Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods." In: *J Multivar Anal* 99.6, pp. 1154–1176. ISSN: 0047259X. DOI: 10.1016/j.jmva.2007.07.002.

Hosoi, Yusuke, Miki Soma, Hirosuke Shiura, Takashi Sado, Hidetoshi Hasuwa, Kuniya Abe, Takashi Kohda, Fumitoshi Ishino, and Shin Kobayashi (Sept. 2018). "Female mice lacking Ftx lncRNA exhibit impaired X-chromosome inactivation and a microphthalmia-like phenotype." In: *Nat Commun* 9.1, p. 3829. DOI: 10.1038/s41467-018-06327-6.

Hänsch, Ronny and Olaf Hellwich (Mar. 2015). "Performance Assessment and Interpretation of Random Forests by Three-dimensional Visualizations." In: *Proceedings of the 6th International Conference on Information Visualization Theory and Applications*. SCITEPRESS - Science, and Technology Publications, pp. 149–156. ISBN: 978-989-758-088-8. DOI: 10.5220/0005310901490156.

Izenman, Alan J. (2008). *Modern Multivariate Statistical Techniques*. Ed. by G. Casella, S. Fienberg, and I. Olkin. Springer texts in statistics. New York, NY: Springer New York. ISBN: 978-0-387-78188-4. DOI: 10.1007/978-0-387-78189-1.

Janiszewski, Adrian, Irene Talon, Juan Song, Natalie De Geest, San Kit To, Greet Bervoets, Jean-Christophe Marine, Florian Rambow, and Vincent Pasque (Feb. 2019). "Dynamic Erasure of Random X-Chromosome Inactivation during iPSC Reprogramming." In: *BioRxiv*. DOI: 10.1101/545558.

Johnson, David S, Ali Mortazavi, Richard M Myers, and Barbara Wold (June 2007). "Genome-wide mapping of in vivo protein-DNA interactions." In: *Science* 316.5830, pp. 1497–1502. ISSN: 1095-9203. DOI: 10.1126/science.1141319.

Jonkers, Iris, Tahsin Stefan Barakat, Eskeatnaf Mulugeta Achame, Kim Monkhorst, Annegien Kenter, Eveline Rentmeester, Frank Grosveld, J Anton Grootegoed, and Joost Gribnau (Nov. 2009). "RNF12 is an X-Encoded dose-dependent activator of X chromosome inactivation." In: *Cell* 139.5, pp. 999–1011. DOI: 10.1016/j.cell.2009.10.034.

Jégu, Teddy, Eric Aeby, and Jeannie T Lee (May 2017). "The X chromosome in space." In: *Nature Reviews. Genetics* 18.6, pp. 377–389. DOI: 10.1038/nrg.2017.17.

Karlić, Rosa, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahovicek, and Martin Vingron (Feb. 2010). "Histone modification levels are predictive for gene expression." In: *Proc Natl Acad Sci USA* 107.7, pp. 2926–2931. ISSN: 1091-6490. DOI: 10.1073/pnas.0909344107.

Kaufman, LR and P Rousseeuw (1990). "PJ (1990) Finding groups in data: An introduction to cluster analysis." In: *Hoboken NJ John Wiley & Sons Inc* 725.

Kawakami, Takahiro, Keisei Okamoto, Hiroyuki Sugihara, Takanori Hattori, Anthony E Reeve, Osamu Ogawa, and Yusaku Okada (Apr. 2003). "The roles of supernumerical X chromosomes and XIST expression in testicular germ cell tumors." In: *J Urol* 169.4, pp. 1546–1552. DOI: 10.1097/01.ju.0000044927.23323.5a.

Kelly, Cassidy and Kazunori Okada (May 2012). "Variable interaction measures with random forest classifiers." In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 154–157. ISBN: 978-1-4577-1858-8. DOI: 10.1109/{ISBI}.2012.6235507.

Kelsey, Angela D, Christine Yang, Danny Leung, Jakub Minks, Thomas Dixon-McDougall, Sarah E L Baldry, Aaron B Bogutz, Louis Lefebvre, and Carolyn J Brown (Oct. 2015). "Impact of flanking chromosomal sequences on localization and silencing by the human non-coding RNA XIST." In: *Genome Biol* 16, p. 208. DOI: 10.1186/s13059-015-0774-2.

Keniry, Andrew and Marnie E Blewitt (June 2018). "Studying X chromosome inactivation in the single-cell genomic era." In: *Biochemical Society Transactions* 46.3, pp. 577–586. DOI: 10.1042/{BST20170346}.

Kim, Joomyeong and Hana Kim (2012). "Recruitment and biological consequences of histone modification of H3K27me3 and H3K9me3." In: *ILAR journal* 53.3-4, pp. 232–239.

Kim, Yoon Jung, Katharine R Cecchini, and Tae Hoon Kim (May 2011). "Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus." In: *Proc Natl Acad Sci USA* 108.18, pp. 7391–7396. DOI: 10.1073/pnas.1018279108.

Kinkley, Sarah, Johannes Helmuth, Julia K Polansky, Ilona Dunkel, Gilles Gasparoni, Sebastian Fröhler, Wei Chen, Jörn Walter, Alf Hamann, and Ho-Ryun Chung (Aug. 2016). "reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4(+) memory T cells." In: *Nature Communications* 7, p. 12514. DOI: 10.1038/ncomms12514.

Kobayashi, Reiko, Ryu Miyagawa, Hideomi Yamashita, Teppei Morikawa, Kae Okuma, Masashi Fukayama, Kuni Ohtomo, and Keiichi Nakagawa (Nov. 2016). "Increased expression of long non-coding RNA XIST predicts favorable prognosis of cervical squamous cell carcinoma subsequent to definitive chemoradiation therapy." In: *Oncol Lett* 12.5, pp. 3066–3074. DOI: 10.3892/ol.2016.5054.

Krueger, Felix and Simon R Andrews (June 2016). "SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes." In: *F1000Research* 5, p. 1479. DOI: 10.12688/f1000research.9037.2.

Kuhn, R M et al. (Jan. 2007). "The UCSC genome browser database: update 2007." In: *Nucleic Acids Research* 35.Database issue, pp. D668–73. DOI: 10.1093/nar/gkl928.

Kuznetsova, Natalia (Aug. 2014). "Random Forest Visualization." In: *Master Thesis*.

Kwak, Hojoong, Nicholas J Fuda, Leighton J Core, and John T Lis (Feb. 2013). "Precise maps of RNA polymerase reveal how promoters direct initiation and pausing." In: *Science* 339.6122, pp. 950–953. DOI: 10.1126/science.1229386.

Langmead, Ben and Steven L Salzberg (Mar. 2012). "Fast gapped-read alignment with Bowtie 2." In: *Nature Methods* 9.4, pp. 357–359. DOI: 10.1038/nmeth.1923.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg (Mar. 2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." In: *Genome Biology* 10.3, R25. DOI: 10.1186/gb-2009-10-3-r25.

Laurent, H and RL Rivest (1976). "Constructing optimal binary 12ision trees is NP-complete." In: *Information processing letters* 5.1, pp. 15–17.

Leskovec, A, Anand Rajaraman, and Jeffrey David Ullman (2014). *Mining of Massive Datasets*. Cambridge: Cambridge University Press. ISBN: 9781139058452. DOI: 10.1017/{CBO9781139058452}.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup (Aug. 2009). "The Sequence Alignment/Map format and SAMtools." In: *Bioinformatics* 25.16, pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

Lieberman-Aiden, Erez et al. (Oct. 2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." In: *Science* 326.5950, pp. 289–293. ISSN: 1095-9203. DOI: 10.1126/science.1181369.

Loda, Agnese et al. (Sept. 2017). "Genetic and epigenetic features direct differential efficiency of Xist-mediated silencing at X-chromosomal and autosomal locations." In: *Nature Communications* 8.1, p. 690. DOI: 10.1038/s41467-017-00528-1.

Long, Yicheng, Xueyin Wang, Daniel T Youmans, and Thomas R Cech (Sept. 2017). "How do lncRNAs regulate transcription?" In: *Sci Adv* 3.9, eaao2110. DOI: 10.1126/sciadv.aao2110.

Lu, Zhipeng, Ava C Carter, and Howard Y Chang (Nov. 2017). "Mechanistic insights in X-chromosome inactivation." In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372.1733. DOI: 10.1098/rstb.2016.0356.

Lugowski, Andrew, Beth Nicholson, and Olivia S Rissland (Mar. 2018). "Determining mRNA half-lives on a transcriptome-wide scale." In: *Methods* 137, pp. 90–98. DOI: 10.1016/j.ymeth.2017.12.006.

Luikenhuis, S, A Wutz, and R Jaenisch (Dec. 2001). "Antisense transcription through the Xist locus mediates Tsix function in embryonic stem cells." In: *Mol Cell Biol* 21.24, pp. 8512–8520. DOI: 10.1128/{MCB}.21.24.8512-8520.2001.

Lunetta, Kathryn L, L Brooke Hayward, Jonathan Segal, and Paul Van Eerdewegh (Dec. 2004). "Screening large-scale association study data: exploiting interactions using random forests." In: *BMC Genetics* 5, p. 32. ISSN: 1471-2156. DOI: 10.1186/1471-2156-5-32.

Lyon, M F (Apr. 1961). "Gene action in the X-chromosome of the mouse (Mus musculus L.)." In: *Nature* 190, pp. 372–373. DOI: 10.1038/190372a0.

M. Jones, Zachary and Fridolin J. Linder (Oct. 2016). "edarf: Exploratory Data Analysis using Random Forests." In: *The Journal of Open Source Software* 1.6, p. 92. ISSN: 2475-9066. DOI: 10.21105/joss.00092.

MacQueen, J (1967). "Some methods for classification and analysis of multivariate observations." In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1.14, p. 281.

Mahat, Dig Bijay, Hojoong Kwak, Gregory T Booth, Iris H Jonkers, Charles G Danko, Ravi K Patel, Colin T Waters, Katie Munson, Leighton J Core, and John T Lis (July 2016). "Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq)." In: *Nature Protocols* 11.8, pp. 1455–1476. DOI: 10.1038/nprot.2016.086.

Marahrens, Y, B Panning, J Dausman, W Strauss, and R Jaenisch (Jan. 1997). "Xist-deficient mice are defective in dosage compensation but not spermatogenesis." In: *Genes Dev* 11.2, pp. 156–166. DOI: 10.1101/gad.11.2.156.

Marks, Hendrik et al. (Aug. 2015). "Dynamics of gene silencing during X inactivation using allele-specific RNA-seq." In: *Genome Biology* 16, p. 149. DOI: 10.1186/s13059-015-0698-x.

Marsico, Annalisa, Matthew R Huska, Julia Lasserre, Haiyang Hu, Dubravka Vucicevic, Anne Musahl, Ulf Orom, and Martin Vingron (Aug. 2013). "PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs." In: *Genome Biol* 14.8, R84. DOI: 10.1186/gb-2013-14-8-r84.

McHugh, Colleen A et al. (May 2015). "The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3." In: *Nature* 521.7551, pp. 232–236. DOI: 10.1038/nature14443.

Miller, Tim (Feb. 2019). "Explanation in artificial intelligence: Insights from the social sciences." In: *Artificial intelligence* 267, pp. 1–38. ISSN: 00043702. DOI: 10.1016/j.artint.2018.07.007.

Mira-Bontenbal, Hegias and Joost Gribnau (Apr. 2016). "New Xist-Interacting Proteins in X-Chromosome Inactivation." In: *Current Biology* 26.8, R338–42. DOI: 10.1016/j.cub.2016.03.022.

Moindrot, Benoit, Andrea Cerase, Heather Coker, Osamu Masui, Anne Grijzenhout, Greta Pintacuda, Lothar Schermelleh, Tatyana B Nesterova, and Neil Brockdorff (July 2015). "A Pooled shRNA Screen Identifies Rbm15, Spen, and Wtap as Factors Required for Xist RNA-Mediated Silencing." In: *Cell Rep* 12.4, pp. 562–572. DOI: 10.1016/j.celrep.2015.06.053.

Monfort, Asun, Giulio Di Minin, Andreas Postlmayr, Remo Freimann, Fabiana Arieti, Stéphane Thore, and Anton Wutz (July 2015). "Identification of Spen as a Crucial Factor for Xist Function through Forward Genetic Screening in Haploid Embryonic Stem Cells." In: *Cell Rep* 12.4, pp. 554–561. DOI: 10.1016/j.celrep.2015.06.067.

Monfort, Asun and Anton Wutz (Nov. 2017). "Progress in understanding the molecular mechanism of Xist RNA function through genetics." In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372.1733. DOI: 10.1098/rstb.2016.0368.

Monkhorst, Kim, Iris Jonkers, Eveline Rentmeester, Frank Grosveld, and Joost Gribnau (Feb. 2008). "X inactivation counting and choice is a stochastic process: evidence for involvement of an X-linked activator." In: *Cell* 132.3, pp. 410–421. DOI: 10.1016/j.cell.2007.12.036.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold (July 2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7105. DOI: 10.1038/nmeth.1226.

Mutzel, Verena, Ikuhiro Okamoto, Ilona Dunkel, Mitinori Saitou, Luca Giorgetti, Edith Heard, and Edda G Schulz (Apr. 2019). "A symmetric toggle switch explains the onset of random X

inactivation in different mammals." In: *Nat Struct Mol Biol* 26.5, pp. 350–360. DOI: 10.1038/s41594-019-0214-1.

Naughton, Catherine, Duncan Sproul, Charlotte Hamilton, and Nick Gilbert (Nov. 2010). "Analysis of active and inactive X chromosome architecture reveals the independent organization of 30 nm and large-scale chromatin structures." In: *Mol Cell* 40.3, pp. 397–409. DOI: 10.1016/j.molcel.2010.10.013.

Nechanitzky, Robert, Amparo Dávila, Fabio Savarese, Stefanie Fietze, and Rudolf Grosschedl (Oct. 2012). "Satb1 and Satb2 are dispensable for X chromosome inactivation in mice." In: *Developmental Cell* 23.4, pp. 866–871. DOI: 10.1016/j.devcel.2012.09.018.

Nesterova, Tatyana B et al. (July 2019). "Systematic allelic analysis defines the interplay of key pathways in X chromosome inactivation." In: *Nature Communications* 10.1, p. 3129. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11171-3.

Nora, Elphège P et al. (Apr. 2012). "Spatial partitioning of the regulatory landscape of the X-inactivation centre." In: *Nature* 485.7398, pp. 381–385. DOI: 10.1038/nature11049.

Ogawa, Yuya and Jeannie T Lee (Mar. 2003). "Xite, X-inactivation intergenic transcription elements that regulate the probability of choice." In: *Mol Cell* 11.3, pp. 731–743. DOI: 10.1016/s1097-2765(03)00063-7.

Pastor, William A, Utz J Pape, Yun Huang, Hope R Henderson, Ryan Lister, Myunggon Ko, Erin M McLoughlin, Yevgeny Brudno, Sahasransu Mahapatra, Philipp Kapranov, et al. (2011). "Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells." In: *Nature* 473.7347, pp. 394–397.

Patil, Deepak P, Chun-Kan Chen, Brian F Pickering, Amy Chow, Constanza Jackson, Mitchell Guttman, and Samie R Jaffrey (Sept. 2016). "m(6)A RNA methylation promotes XIST-mediated transcriptional repression." In: *Nature* 537.7620, pp. 369–373. ISSN: 0028-0836. DOI: 10.1038/nature19342.

Penny, G D, G F Kay, S A Sheardown, S Rastan, and N Brockdorff (Jan. 1996). "Requirement for Xist in X chromosome inactivation." In: *Nature* 379.6561, pp. 131–137. DOI: 10.1038/379131a0.

Penzkofer, Tobias, Marten Jäger, Marek Figlerowicz, Richard Badge, Stefan Mundlos, Peter N Robinson, and Tomasz Zemojtel (Jan. 2017). "L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes." In: *Nucleic Acids Res* 45.D1, pp. D68–D73. ISSN: 0305-1048. DOI: 10.1093/nar/gkw925.

Peric-Hupkes, Daan et al. (May 2010). "Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation." In: *Mol Cell* 38.4, pp. 603–613. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2010.03.016.

Pinheiro, Ines and Edith Heard (Mar. 2017). "X chromosome inactivation: new players in the initiation of gene silencing." In: *F1000Research* 6. DOI: 10.12688/f1000research.10707.1.

Pintacuda, Greta et al. (Dec. 2017). "hnRNPK Recruits PCGF3/5-PRC1 to the Xist RNA B-Repeat to Establish Polycomb-Mediated Chromosomal Silencing." In: *Mol Cell* 68.5, 955–969.e10. DOI: 10.1016/j.molcel.2017.11.013.

Pollex, Tim and Edith Heard (Jan. 2019). "Nuclear positioning and pairing of X-chromosome inactivation centers are not primary determinants during initiation of random X-inactivation." In: *Nat Genet* 51.2, pp. 285–295. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0305-7.

Rabani, Michal et al. (May 2011). "Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells." In: *Nature Biotechnology* 29.5, pp. 436–442. ISSN: 1546-1696. DOI: 10.1038/nbt.1861.

Raiber, Eun-Ang, Dario Beraldi, Gabriella Ficz, Heather E Burgess, Miguel R Branco, Pierre Murat, David Oxley, Michael J Booth, Wolf Reik, and Shankar Balasubramanian (2012). "Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase." In: *Genome biology* 13.8, pp. 1–11.

Ramírez, Fidel, Friederike Dündar, Sarah Diehl, Björn A Grüning, and Thomas Manke (July 2014). "deepTools: a flexible platform for exploring deep-sequencing data." In: *Nucleic Acids Research* 42.Web Server issue, W187–91. DOI: 10.1093/nar/gku365.

Rastan, S and E J Robertson (Dec. 1985). "X-chromosome deletions in embryo-derived (EK) cell lines associated with lack of X-chromosome inactivation." In: *Journal of embryology and experimental morphology* 90, pp. 379–388.

Ridings-Figueroa, Rebeca et al. (May 2017). "The nuclear matrix protein CIZ1 facilitates localization of Xist RNA to the inactive X-chromosome territory." In: *Genes Dev* 31.9, pp. 876–888. DOI: 10.1101/gad.295907.117.

Roberts, R Michael and Susan J Fisher (Mar. 2011). "Trophoblast stem cells." In: *Biology of Reproduction* 84.3, pp. 412–421. DOI: 10.1095/biolreprod.110.088724.

Robertson, Gordon et al. (Aug. 2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." In: *Nature Methods* 4.8, pp. 651–657. ISSN: 1548-7091. DOI: 10.1038/nmeth1068.

Rocha, Simão T da and Edith Heard (Mar. 2017). "Novel players in X inactivation: insights into Xist-mediated gene silencing and chromosome conformation." In: *Nature Structural & Molecular Biology* 24.3, pp. 197–204. DOI: 10.1038/nsmb.3370.

Rocha, Simão Teixeira da et al. (Jan. 2014). "Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome." In: *Mol Cell* 53.2, pp. 301–316. DOI: 10.1016/j.molcel.2014.01.002.

Ronaghi, M, M Uhlén, and P Nyrén (July 1998). "A sequencing method based on real-time pyrophosphate." In: *Science* 281.5375, pp. 363, 365. DOI: 10.1126/science.281.5375.363.

Russell, L B (May 1963). "Mammalian X-chromosome action: inactivation limited in spread and region of origin." In: *Science* 140.3570, pp. 976–978.

Sahakyan, Anna, Yihao Yang, and Kathrin Plath (June 2018). "The Role of Xist in X-Chromosome Dosage Compensation." In: *Trends in Cell Biology* 28.12, pp. 999–1013. DOI: 10.1016/j.tcb.2018.05.005.

Sahlén, Pelin, Ilgar Abdullayev, Daniel Ramsköld, Liudmila Matskova, Nemanja Rilakovic, Britta Lötstedt, Thomas J Albert, Joakim Lundeberg, and Rickard Sandberg (Aug. 2015). "Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution." In: *Genome Biology* 16, p. 156. DOI: 10.1186/s13059-015-0727-9.

Sakata, Yuka, Koji Nagao, Yuko Hoki, Hiroyuki Sasaki, Chikashi Obuse, and Takashi Sado (Aug. 2017). "Defects in dosage compensation impact global gene regulation in the mouse trophoblast." In: *Development* 144.15, pp. 2784–2797. DOI: 10.1242/dev.149138.

Schoenfelder, Stefan et al. (Apr. 2015). "The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements." In: *Genome Res* 25.4, pp. 582–597. DOI: 10.1101/gr.185272.114.

Schulz, Edda G, Johannes Meisig, Tomonori Nakamura, Ikuhiro Okamoto, Anja Sieber, Christel Picard, Maud Borensztein, Mitinori Saitou, Nils Blüthgen, and Edith Heard (Feb. 2014). "The two active X chromosomes in female ESCs block exit from the pluripotent state by modulating

the ESC signaling network." In: *Cell Stem Cell* 14.2, pp. 203–216. DOI: 10.1016/j.stem.2013.11.022.

Seitan, Vlad C et al. (Dec. 2013). "Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments." In: *Genome Res* 23.12, pp. 2066–2077. DOI: 10.1101/gr.161620.113.

Silver, David et al. (Jan. 2016). "Mastering the game of Go with deep neural networks and tree search." In: *Nature* 529.7587, pp. 484–489. DOI: 10.1038/nature16961.

Simon, Noah, Jerome Friedman, and Trevor Hastie (Nov. 2013). "A Blockwise Descent Algorithm for Group-penalized Multiresponse and Multinomial Regression." In: *arXiv*.

Splinter, Erik et al. (July 2011). "The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA." In: *Genes & Development* 25.13, pp. 1371–1383. DOI: 10.1101/gad.633311.

Stadler, Michael B et al. (Dec. 2011). "DNA-binding factors shape the mouse methylome at distal regulatory regions." In: *Nature* 480.7378, pp. 490–495. ISSN: 0028-08361476-4687. DOI: 10.1038/nature10716.

Stavropoulos, N, N Lu, and J T Lee (Aug. 2001). "A functional role for Tsix transcription in blocking Xist RNA accumulation but not in X-chromosome choice." In: *Proc Natl Acad Sci USA* 98.18, pp. 10232–10237. ISSN: 0027-8424. DOI: 10.1073/pnas.171243598.

Stavropoulos, Nicholas, Rebecca K Rowntree, and Jeannie T Lee (Apr. 2005). "Identification of developmentally specific enhancers for Tsix in the regulation of X chromosome inactivation." In: *Mol Cell Biol* 25.7, pp. 2757–2769. DOI: 10.1128/{MCB}.25.7.2757-2769.2005.

Steensel, Bas van and Andrew S Belmont (May 2017). "Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression." In: *Cell* 169.5, pp. 780–791. ISSN: 00928674. DOI: 10.1016/j.cell.2017.04.022.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn (Jan. 2007). "Bias in random forest variable importance measures: illustrations, sources and a solution." In: *BMC Bioinformatics* 8, p. 25. DOI: 10.1186/1471-2105-8-25.

Sun, Sha, Brian C Del Rosario, Attila Szanto, Yuya Ogawa, Yesu Jeon, and Jeannie T Lee (June 2013). "Jpx RNA activates Xist by evicting CTCF." In: *Cell* 153.7, pp. 1537–1551. DOI: 10.1016/j.cell.2013.05.028.

Sunwoo, Hongjae, David Colognori, John E Froberg, Yesu Jeon, and Jeannie T Lee (Oct. 2017). "Repeat E anchors Xist RNA to the inactive X chromosomal compartment through CDKN1A-interacting protein (CIZ1)." In: *Proc Natl Acad Sci USA* 114.40, pp. 10654–10659. DOI: 10.1073/pnas.1711206114.

Takagi, N and K Abe (May 1990). "Detrimental effects of two active X chromosomes on early mouse development." In: *Development* 109.1, pp. 189–201.

Tsai, Chia-Lun, Rebecca K Rowntree, Dena E Cohen, and Jeannie T Lee (July 2008). "Higher order chromatin structure at the X-inactivation center via looping DNA." In: *Developmental Biology* 319.2, pp. 416–425. DOI: 10.1016/j.ydbio.2008.04.010.

Tüttelmann, F and J Gromoll (June 2010). "Novel genetic aspects of Klinefelter's syndrome." In: *Molecular Human Reproduction* 16.6, pp. 386–395. DOI: 10.1093/molehr/gaq019.

Vallot, Céline, Jean-François Ouimette, and Claire Rougeulle (July 2016). "Establishment of X chromosome inactivation and epigenomic features of the inactive X depend on cellular contexts." In: *Bioessays: News and Reviews in Molecular, Cellular and Developmental Biology* 38.9, pp. 869–880. DOI: 10.1002/bies.201600121.

Vella, Pietro et al. (Feb. 2013). "Tet proteins connect the O-linked N-acetylglucosamine transferase Ogt to chromatin in embryonic stem cells." In: *Mol Cell* 49.4, pp. 645–656. DOI: 10.1016/j.molcel.2012.12.019.

Vinod, Hrishikesh D. (June 1969). "Integer programming and the theory of grouping." In: *Journal of the American Statistical Association* 64.326, pp. 506–519. ISSN: 0162-1459. DOI: 10.1080/01621459.1969.10500990.

Wada, Takeo and Attila Becskei (Dec. 2017). "Impact of Methods on the Measurement of mRNA Turnover." In: *International Journal of Molecular Sciences* 18.12. DOI: 10.3390/ijms18122723.

Wang, Liangjun, J Lesley Brown, Ru Cao, Yi Zhang, Judith A Kassis, and Richard S Jones (June 2004). "Hierarchical recruitment of polycomb group silencing complexes." In: *Mol Cell* 14.5, pp. 637–646. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2004.05.009.

Wang, Zhong, Huntington F Willard, Sayan Mukherjee, and Terrence S Furey (Sept. 2006). "Evidence of influence of genomic DNA sequence on human X chromosome inactivation." In: *PLoS Computational Biology* 2.9, e113. DOI: 10.1371/journal.pcbi.0020113.

Watson, J D and F H Crick (Apr. 1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." In: *Nature* 171.4356, pp. 737–738. DOI: 10.1038/171737a0.

Welling, Soeren H., Hanne H. F. Refsgaard, Per B. Brockhoff, and Line H. Clemmensen (May 2016). "Forest Floor Visualizations of Random Forests." In: *arXiv*.

Wu, Hao, Junjie Luo, Huimin Yu, Amir Rattner, Alisa Mo, Yanshu Wang, Philip M Smallwood, Bracha Erlanger, Sarah J Wheelan, and Jeremy Nathans (Jan. 2014). "Cellular resolution maps of X chromosome inactivation: implications for neural development, function, and disease." In: *Neuron* 81.1, pp. 103–119. DOI: 10.1016/j.neuron.2013.10.051.

Wutz, Anton, Theodore P Rasmussen, and Rudolf Jaenisch (Feb. 2002). "Chromosomal silencing and localization are mediated by different domains of Xist RNA." In: *Nat Genet* 30.2, pp. 167–174. ISSN: 1061-4036. DOI: 10.1038/ng820.

Yang, Fan, Tomas Babak, Jay Shendure, and Christine M Disteche (May 2010). "Global survey of escape from X inactivation by RNA-sequencing in mouse." In: *Genome Research* 20.5, pp. 614–622. DOI: 10.1101/gr.103200.109.

Yang, Min, Hexin Xu, Dingju Zhu, and Huijuan Chen (2012). "Visualizing the random forest by 3D techniques." In: *Internet of Things*. Ed. by Yongheng Wang and Xiaoming Zhang. Vol. 312. Communications in computer and information science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 639–645. ISBN: 978-3-642-32426-0. DOI: 10.1007/978-3-642-32427-7_91.

Yang, Zhi, Xiaodi Jiang, Xiaofeng Jiang, and Haiying Zhao (Aug. 2018). "X-inactive-specific transcript: A long noncoding RNA with complex roles in human cancers." In: *Gene* 679, pp. 28–35. DOI: 10.1016/j.gene.2018.08.071.

Zheng, Ruinian, Shunhuan Lin, Ling Guan, Huiling Yuan, Kejun Liu, Chun Liu, Weibiao Ye, Yuting Liao, Jun Jia, and Ruopeng Zhang (Apr. 2018). "Long non-coding RNA XIST inhibited breast cancer cell growth, migration, and invasion via miR-155/CDX1 axis." In: *Biochem Biophys Res Commun* 498.4, pp. 1002–1008. DOI: 10.1016/j.bbrc.2018.03.104.

Zuin, Jessica et al. (Jan. 2014). "Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells." In: *Proc Natl Acad Sci USA* 111.3, pp. 996–1001. DOI: 10.1073/pnas.1317788111.

ABSTRACT

To equalize gene dosage between sexes, the long non-coding RNA *Xist* mediates chromosome-wide gene silencing of one X Chromosome in female mammals - a process known as X chromosome inactivation (XCI). The efficiency of gene silencing is highly variable across genes, with some genes even escaping XCI in somatic cells. A gene's susceptibility to *Xist*-mediated silencing appears to be determined by a complex interplay of epigenetic and genomic features. However, the underlying rules remain poorly understood. To advance the understanding of *Xist*-mediated silencing pathways, chromosome-wide gene silencing dynamics at the level of nascent transcriptome were quantified using allele-specific Precision nuclear Run-On sequencing. We have developed a Random Forest machine learning model that is able to predict the measured silencing dynamics based on a large set of epigenetic and genomic features and tested its predictive power experimentally. We introduced a forest-guided clustering approach to uncover the combinatorial rules that control *Xist*-mediated gene silencing. Results suggest that the genomic distance to the *Xist* locus, followed by gene density and distance to LINE elements are the prime determinants of silencing velocity. Moreover, a series of features associated with active transcriptional elongation and chromatin 3D structure are enriched at efficiently silenced genes. Generally, silenced genes seem to be separated into two distinct groups, associated with different silencing pathways: one group that requires an AT-rich sequence context and the *Xist* repeat-A for silencing, which is known to activate the SPEN pathway, and another group where genes are pre-marked by polycomb complexes and tend to rely on the repeat-B in *Xist* for silencing, known to recruit polycomb complexes during XCI. Our machine learning approach can thus uncover the complex combinatorial rules underlying gene silencing during X chromosome inactivation.

ZUSAMMENFASSUNG

Eines der beiden X chromosome in weiblichen Säugetieren muss inaktiviert, um die Dosierung von X-Chromosomalen Genen zwischen den Geschlechtern auszugegleichen. Dieser Prozess wird X Chromosom Inaktivierung (XCI) genannt und wird maßgeblich von der langen nicht-kodierenden RNA *Xist* gesteuert. Die Inaktivierung von unterschiedlichen Genen erfolgt unterschiedlich schnell. Manche Gene sind sogar in der Lage der Inaktivierung zu entgehen und sind somit weiterhin in somatischen Zellen aktiv. Die Dynamiken mit denen Gene inaktiviert werden, werden durch ein komplexes Zusammenspiel von epigenetischen und genomischen Faktoren bestimmt. Dieses Zusammenspiel wurde bis jetzt jedoch noch nicht hinreichend untersucht um aussagekräftige Rückschlüsse zu ziehen. Für ein besseres Verständnis dieses Zusammenspiels, wurde mit Hilfe allel spezifischer Precision nuclear Run-On Sequenzierung die Inaktivierungsdynamik Chromosomen weit gemessen. Diese Messungen, wie auch eine Vielzahl von epigenetischen und genomischen Faktoren, haben uns in die Lage versetzt, mit Hilfe eines Random Forest Modells, Chromosomen weite Inaktivierungsdynamiken vorherzusagen, welche durch zusätzliche Experimente validiert werden konnten. Um zu analysieren welche Faktoren in diesem Prozess zusammenspielen, haben wir einen Random Forest-gestützten Clustering Ansatz implementiert. Die Ergebnisse legen nahe, dass der genomische Abstand zum *Xist* Genlocus, sowie die Gendichte und der Abstand zu LINE Elementen, die Hauptfaktoren für die Inaktivierungsgeschwindigkeit sind. Darüber hinaus wird eine Reihe von Faktoren, wie zum Beispiel die aktive Transkription oder die 3D Struktur des Chromatins, mit schneller Inaktivierung in Verbindung gebracht. Im Allgemeinen lassen sich inaktivierte Gene in zwei unterschiedliche Gruppen unterteilen, die mit unterschiedlichen Inaktivierungspfaden in Verbindung gebracht werden können. Die eine Gruppe benötigt einen AT-reichen Sequenz Kontext und das *Xist* Repeat-A Element, das welches den SPEN-Pfad aktiviert, während die andere Gruppe eine Anreicherung an Polycomb-Komplexen benötigt und auf das *Xist* Repeat-B Element zurückgreift, welches Polycomb-Komplexe während des XCI Prozesses rekrutiert. Diese Ergebnisse zeigen, dass unser Ansatz, basierend auf maschinellem Lernen, die komplexen kombinatorischen Regeln identifizieren kann, die der Inaktivierung von Genen während des XCI Prozesses zugrunde liegen.

SELBSTÄNDIGKEITSERKLÄRUNG

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Lisa Barros de Andrade e Sousa

Berlin, Januar 2020

Lisa Corina Barros de Andrade e Sousa