ANALYSIS OF THE DETERMINANTS OF POL II PAUSING

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

vorgelegt von

MARTYNA GAJOS

Berlin, 2022

ERSTGUTACHTER: Prof. Dr. Martin Vingron
ZWEITGUTACHTER: Prof. Dr. Achim Tresch
TAG DER DISPUTATION: 27.06.2022

PUBLICATIONS

This thesis is built on a project which stemmed from discussions with Andreas Mayer, Alena van Bömmel and Martin Vingron. The results were published in Nucleic Accid Research (Gajos *et al.* [21]) and are available at https://doi.org/10.1093/nar/gkab208.

In accordance with the standard scientific protocol, throughout the chapters of this thesis I will use the personal pronoun we to indicate work done by myself, with the supervision and support from my adviser Andreas Mayer. Any results which were done in collaboration will be mentioned accordingly in the text.

ACKNOWLEDGMENTS

I am thankful to my adviser, Andreas Mayer, for providing me with the opportunity to work as a PhD candidate in his group at the Max Planck Institute for Molecular Genetics. I am grateful for his supervision and the opportunity he gave me to collaborate on multiple projects. I would like to thank my supervisor Martin Vingron for accepting me into the International Max Planck Research School for Computational Biology and Scientific Computing, advising my work and reviewing this thesis. I would also like to thank Alena van Bömmel for being a member of my thesis advisory committee and providing me with valuable feedback about my project. Finally, I am grateful to Achim Tresch for agreeing to serve as my external supervisor and reviewing the work in this thesis.

I would further like to thank my group members, who proof-read this work and supported me throughout my years as a PhD candidate. They were always ready to share their expertise. I truly enjoyed our group meetings that were filled with fruitful discussions and provided inspiration for avenues to explore in my project. I was lucky to share the office with Annkatrin Bressin, who was always willing to evaluate my ideas and who also assisted me with translating to German, and Olga Jasnovidova, who helped me navigate the world of experimental biology. I am grateful to Jelena Ulicevic for providing me not only with food, but also with food for thought during our lunch breaks. I would like to thank Siddharth Annaldasula for proof-reading my thesis and the opportunity to see him grow as a bioinformatician. I would like to acknowledge the help of Kirsten Kelleher on all PhD related issues and proof-reading chapters of this thesis.

Last but not least, I would like to thank Michael, my family and my friends for their unwavering support. I would not have made it without you.

CONTENTS

1 INTRODUCTION

Every organism that we know of depends on proteins to maintain homeostasis and react to internal and external cues. Proteins are encoded by genes, which are made up of DNA. When the cell requires a particular protein, the nucleotide sequence of the gene encoded by a portion of the long DNA molecule is first copied into RNA. In the second step, those RNA copies are used as templates to synthesize proteins. The regulation of this two-step process, called gene expression, plays a critical role in determining which proteins are present in a cell and in what amounts [1].

Gene transcription is the first regulatory step of gene expression. In eukaryotes, all protein-coding genes in the cell nucleus and a large set of non-protein encoding genes are transcribed by the key enzyme called RNA polymerase II (Pol II). For many years, gene transcription had been thought to be regulated only during the first step of transcription, called transcription initiation, when Pol II is recruited to the gene promoter. The following step of transcription, transcription elongation, had been considered as a continuous process during which subsequent RNA nucleotides are added at a steady rate. However, this view was challenged by the discovery of transcriptional pausing that interrupts the phases of productive nucleotide addition by halting the polymerase. Pol II pausing during early transcription elongation emerged as a checkpoint and a rate-limiting step in gene expression [52].

The research on transcriptional pausing has been stimulated by a rapid development of genome-wide methods capturing the position of transcribing Pol II with single-nucleotide precision and in a DNA strand-specific manner. These methods allow to generate Pol II occupancy tracks, in which transcriptional pausing creates local enrichments of signal (Figure 1.1). However, the determinants of transcriptional pausing *in vivo* remain unclear.

The main aim of this thesis was to investigate the causes of Pol II pausing. To achieve this goal, we scrutinized the characteristics and potential shortcomings of Native Elongating Transcript sequencing (NET-seq) [50], which is one of the high-resolution methods of Pol II profiling, and we improved the NET-seq analysis pipeline. We designed a tool to detect pausing sites in the high-resolution Pol II occupancy tracks and examined the distribution

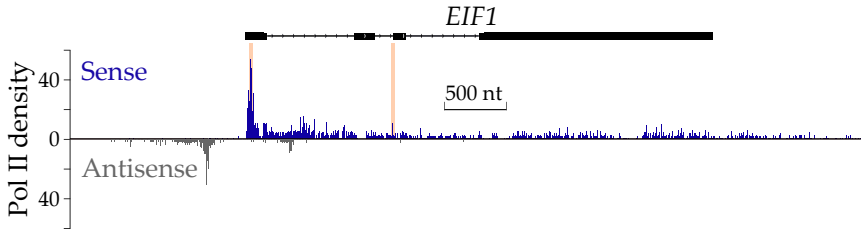

Figure 1.1: **Pol II occupancy track.** The gene track shows Pol II occupancy in both sense and antisense direction of transcription. Selected two pausing sites in promoter-proximal and gene-body region are marked in orange. The data was obtained using Native Elongating Transcript sequencing.

of the pausing sites over human genome. We then set out to identify the determinants of Pol II pausing in an unbiased manner based on the underlying DNA sequence. We created a large number of features, including factors that were previously linked to transcriptional pausing, but also factors that were not yet connected to Pol II pausing. To predict the predisposition of a genomic site to evoke Pol II pausing, we tested machine learning models such as logistic regression and two tree-based ensemble models. Examining feature importance helped us to identify the most important features in the model, namely the main determinants of Pol II pausing in different genomic regions. Finally, we addressed the question of the conservation of transcriptional pausing mechanisms, by analyzing the pausing determinants in various model organisms using publicly available NET-seq data.

1.1 THESIS STRUCTURE AND OBJECTIVES

Following this Chapter, a more detailed introduction into the biological background is given in Chapter 2. It discusses the selected aspects of the process of transcription and presents the current knowledge of the RNA polymerase pausing and the experimental methods for studying this phenomenon. Chapter 3 covers the bioinformatical prerequisites and introduces selected statistical concepts used in genomics. Chapter 4 provides an introduction to machine learning and a detailed description of the machine learning algorithms applied to classification problems. Chapter 5 focuses on the pausing site detection in the NET-seq data. It presents the potential problems encountered analyzing NET-seq data, proposes peak calling approaches for sparse data and characterizes detected pausing sites. Chapter 6 examines potential causes of transcriptional pausing in different genomic regions. It provides an in-depth description of training machine learning models distinguishing pausing and non-pausing sites and presents the interpretation of the obtained models. The conclusions of this thesis and potential open questions are discussed in Chapter 7. Chapter 8 provides short description of author's contributions to other projects conducted in Mayer's group.

2 BIOLOGICAL BACKGROUND

This chapter introduces the selected aspect of molecular biology based on the book *Molecular Biology of The Cell* [1], focusing on the process of transcription. It presents the phenomenon of RNA polymerase pausing and discusses the current knowledge of the causes of this phenomenon. Finally, we provide an overview of the experimental techniques applied to study transcription.

2.1 DNA

In all living organisms, the whole genetic information of an organism, called a genome, is stored in **DNA**. Each DNA molecule consists of two long DNA strands. The building blocks of DNA are called nucleotides and are composed of two parts: deoxyribose, a sugar that forms the backbone of the DNA strand, with a phosphate group attached to it and a base. Four nucleobases are commonly found in the DNA: adenine (A), cytosine (C), guanine (G) and thymine (T). The two DNA strands are held together by hydrogen bonds formed between complementary bases: guanine pairs with cytosine and adenine pairs with thymine. The process of binding complementary base pairs is called hybridization and its counterpart process, in which the interactions between the strands are dissociated, is referred to as melting. A thermodynamical stability of a double-strand nucleotide chain can be characterized using its melting temperature, at which a nucleotide duplex dissociates into single strands.

DNA's secondary structure is determined by the base-pairing of the two strands twisted around each other to form a double helix. Other forces affecting the DNA's geometry are the stacking interactions between the neighbouring bases in the same strand, which are stabilized by Van der Waals forces and hydrophobic interactions. The different combinations of nucleotide order in the nucleotide strands lead to various local shapes of the DNA. The local DNA geometry can be described using several parameters that characterize the distances, angles and energies between neighbouring and pairing nucleotides. There are also two grooves in the right-handed double helix, which are called major and minor grooves based on their relative size.

The right-handed helical structure formed by the complementary DNA strands is called B-DNA and is considered to be the canonical form. However, certain DNA sequence patterns can fold into secondary conformations that differ from that canonical form. There have been several types of non-B DNA identified based on the structures they can form, which in turn depend on their motif sequences. The non-B DNA forms include G-quadruplexes that incorporate stems build of guanines, Z-DNA that is a left-handed double-stranded helix, mirror repeats and A-phased repeats composed of three to nine adenines or thymines that create a curvature in the double helix.

The frequencies of individual nucleotides exhibit significant fluctuations across genomes and genomic regions. An example illustrating the biases is the GC content at promoters, which are DNA sequences located upstream of genes that promote the initiation of transcription of genes. Promoters

of lower eukaryotes are characterized by low GC content, defined as the percentage of guanines and cytosines, whereas mammalian promoters are GC-rich [66]. Additionally, in the human genome, GC richness varies between genomic regions, with a higher GC content observed in the proximity of human promoters than in the gene bodies [49]. It is because human promoters contain many regulatory elements that are often GC-rich. Moreover, promoters of expressed genes are unmethylated, which prevents mutational decay of cytosines and results in an increased GC content.

2.2 TRANSCRIPTION

Gene **transcription** is the process of copying a segment of DNA into **RNA**. Like DNA, RNA is assembled as a chain of nucleotides, but with a couple of differences. In contrast to DNA, which forms a double helix in cells, RNA exists as a single strand chain. The sugar creating the backbone of the RNA chain is ribose instead of deoxyribose. Additionally in the RNA chain, uracil (U) is incorporated instead of thymine. Uracil is a demethylated form of thymine and it can form a pair of hydrogen bonds with adenine; therefore each of the bases present in DNA has a complementary base available within the set of RNA nucleobases. This complementarity of the deoxyribonucleotides and ribonucleotides is the basis of the nucleotide addition in transcription. This process begins with unwinding a short fragment of the double helix of DNA, in a way that exposes a short stretch of unpaired nucleotides. One of the DNA strands is then used as a template for RNA synthesis. Subsequent nucleotides are added to the growing RNA chain and the order of nucleotides in RNA is determined by the order in the DNA template. If the newly added ribonucleotide is complementary to the deoxyribonucleotide in the DNA template, it will form a covalent bond with the RNA chain in an enzymatic reaction. The RNA chain produced during the transcription is called a **transcript** and its sequence is complementary to the reversed sequence of the DNA template.

2.2.1 *RNA polymerases*

Transcription is performed by enzymes called DNA-dependent RNA polymerases. A RNA polymerase composition in a cell depends on the species. In bacteria and archaea, there is only one RNA polymerase that consists of multiple subunits. There are multiple types of multi-subunit RNA polymerases observed in eukaryotes, each responsible for the synthesis of a distinct subset of RNA. The catalytic core of RNA polymerases is conserved in prokaryotes and eukaryotes. Additionally, single-subunit RNA polymerases can be found in eukaryotic chloroplasts and mitochondria.

Transcription in eukaryotic cells is performed by five nuclear polymerase complexes, from which two (Pol IV and Pol V) exist only in plants. The other polymerase complexes (I, II, and III) are present in all eukaryotes and have evolved to perform separated functions. Specifically, Pol I carries out the high-level synthesis of only a single transcript, the precursor rRNA (pre-rRNA), which is processed into 28S, 5.8S and 18S rRNAs. On the other hand, the set of transcripts synthesized by Pol II is broad. Pol II transcribes all protein-coding genes and a vast majority of non–protein-coding genes, including genes encoding small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs) and micro RNAs (miRNAs). Pol III is often associated with

the synthesis of a small set of highly expressed infrastructural RNAs, such as different tRNA species and the 5S rRNA. However, Pol III also transcribes a significant number of other ncRNAs, including enhancers, RNase P and MRP, spliceosomal U6 snRNAs, vault RNAs, Y RNAs, virus-encoded RNAs, short interspersed repeated DNA elements–encoded RNAs, 7SL and 7SK RNA.

2.2.2 *Elements of transcribing Pol II structure*

During productive transcription, Pol II is the main part of a transcription elongation complex (TEC), which is minimally composed of the RNA polymerase, double-stranded DNA template, and nascent RNA being synthesized. A close view of the structure of such minimal TEC is presented in Figure 2.1. Additionally, TEC includes several transcription elongation factors that increase transcription processivity and assist Pol II with passing the encountered obstacles.

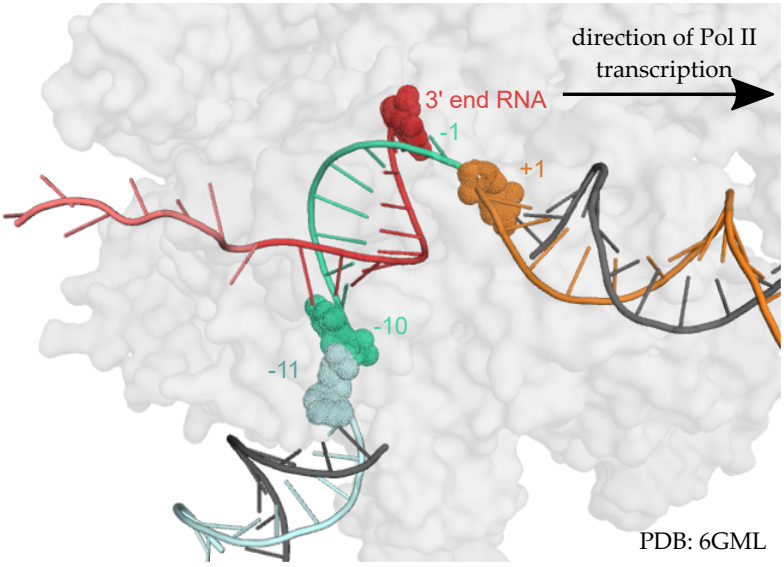

Figure 2.1: **A structure of the RNA-DNA hybrid formed by transcribing Pol II** (PDB structure ID: 6GML [71]). The nascent RNA is depicted in red and pink. The red part of the nascent RNA together with the template DNA fragment marked in green forms a 10 base pair long RNA-DNA hybrid. The template strand fragments positioned upstream and downstream of the RNA-DNA hybrid are in light blue and orange respectively. Numeration of the template strand position is done in respect to the currently transcribed position -1. The core subunits of Pol are depicted in light grey. The direction of the Pol II transcription is indicated with an arrow.

The nascent RNA synthesis takes place in the catalytic core, which is situated in the centre of the Pol II, inside a cleft formed by polymerase's largest subunits. The catalytic core is physically separated from the regulatory factors, which interact with the outside surface of the polymerase. The template DNA and nascent RNA form an 8-9 nucleotide long RNA-DNA hybrid is positioned inside the cleft. A single-stranded template DNA enters the cleft from one side and exits it from the other side rejoined with the

complementary non-template strand. The nascent RNA exits the cleft from the same site as the double-stranded DNA, but using the RNA exit channel.

2.2.3 *Transcriptional cycle*

The process of the gene transcription performed by Pol II can be divided into three steps: initiation, elongation, and termination. In short, during the transcription initiation phase, Pol II is recruited to the gene's promoter by general transcription factors, forming a closed promoter complex [11]. To form the open promoter complex, the DNA strands are 'melted' to create an unwound DNA region, which enables the insertion of the template strand into the active centre of the polymerase. The RNA synthesis is launched from the **transcription start site** (TSS), which is located at the 5′ end of the gene. When a certain length of the nascent RNA is reached, initiation factors are released and a stable **transcription elongation complex** (TEC) is formed over the RNA-DNA hybrid. The polymerase then moves along the gene body, unwinding small portions of the DNA, and uses it as a template to add nucleotides to the growing nascent RNA chain. Transcript elongation is not a monotonous process. The movement of Pol II across genes is discontinuous, interrupted by transcriptional pausing that may either hinder or assist transcript elongation, or even prematurely terminate transcription. To ensure a robust synthesis of long RNA molecules, TEC has a very stable architecture. The high stability of TEC makes transcription termination a rather complicated step. First, the chromatin template at the gene's end slows down the transcription. Once the polymerase reaches the 3′ end of the gene, the RNA transcript is cleaved. Then the remaining transcript is unravelled and degraded, which induces conformational changes in the polymerase triggering its disassociation from the DNA template. The cleaved nascent RNA undergoes modifications dependant on the type of RNA synthesized and the disassociated polymerase is ready to be recruited to a gene's promoter.

2.2.4 *Pol II transcript processing*

The process of elongation is not a smooth progression over a gene body. Pol II needs to overcome obstacles such as nucleosomes or stalling due to misincorporation of nucleotides. Moreover, most of the Pol II transcripts have to be co-transcriptionally processed. All of those alterations have the potential to affect and be affected by the Pol II speed during transcription.

Pol II transcribes various classes of transcripts that undergo different co-transcriptional alterations. One class of Pol II transcripts is messenger RNA (mRNA) which serves as a template for protein synthesis. The co-transcriptional processing of mRNA consists of multiple steps such as 5′ end capping, splicing, 3′ end cleavage and maturation. RNA splicing transforms a precursor mRNA into a mature mRNA by removing non-coding regions of RNA called introns and joining coding regions called exons. This process is catalyzed by the spliceosome, which is an RNA-protein complex. Multiple variants of the final mRNA coding sequence can be obtained by rearranging the pattern of exons that are joined in a process called alternative splicing. Interestingly, interdependence between the Pol II elongation rate and the exon inclusion patterns has been noticed: low Pol II speed or internal pauses

are connected with the inclusion of alternative exons, whereas exclusion of these exons is observed for a highly elongating Pol II [40].

Another class of Pol II transcripts are microRNA (miRNA). They are small non-coding RNAs that regulate the expression of a large proportion of mRNAs by binding nascent RNA transcripts, gene promoter regions or enhancer regions and exerting further effects via epigenetic pathways. The biogenesis of miRNAs is carried out in two subsequent processing events happening in the nucleus and in the cytoplasm respectively. The nuclear processing event happens co-transcriptionally [56] and results in the production of precursor miRNAs, termed pre-miRNAs. Those are subsequently exported to the cytoplasm to undergo the final processing event. The resulting miRNA duplexes are then incorporated into regulatory complexes e.g. the RNA-induced silencing complex [54].

2.3 TRANSCRIPTIONAL PAUSING

Transcriptional pausing was initially discovered *in vitro* for bacterial RNA polymerases. In two studies performed in 1973, transcriptional pausing was observed in a form of an accumulation and disappearance of RNA transcripts of discrete intermediate lengths throughout the transcription reaction [52]. These observations lead to the realization that the movement of Pol II during the elongation phase is highly dynamic and discontinuous. Transcriptional pausing, which interrupts the processive progression of Pol II along the gene body, may assist or hinder the elongation, or even prematurely terminate transcription [59]. In this section, we discuss the most studied type of pausing in human cells, namely promoter-proximal pausing, which occurs in the early stage of elongation. We consider the *trans*-acting factors that are implicated in transcriptional pausing. Finally, we review the DNA motifs reported to be associated with transcriptional pausing in various model organisms.

Promoter-proximal pausing

Promoter-proximal pausing was first described at the heat-shock protein gene locus in *Drosophila melanogaster* [22]. Therefore, it was hypothesized that transcriptionally engaged Pol II remains paused at the promoter-proximal region, awaiting signals for its rapid release and activation following external cues. Since then, promoter-proximal pausing was shown not to be exclusive to stimulus-responsive genes [15, 42, 59]. It appears to be a widespread phenomenon at metazoan genes, with the main Pol II density peak occurring 20 to 60 nucleotides downstream of the transcription start site for most of the genes [58].

Over a dozen of *trans*-acting factors is involved in the establishment and release of promoter-proximal pausing by Pol II in mammalian cells. According to our current understanding, the two most prominent transcription factors necessary for the maintenance of Pol II promoter-proximal pausing are DSIF (DRB sensitivity-inducing factor) and NELF (negative elongation factor) [59, 75]. As shown by the structure of paused Pol II, these two transcription factors stabilize the complex in a conformation characterized by a tilted DNA–RNA hybrid [71]. Such conformation prevents the translocation of the template DNA and in turn impairs the addition of incoming nucleotides to the nascent RNA chain. However, it is still unclear whether NELF and DSIF

directly induce the tilted conformation of hybrid or it is the underlying DNA sequence that favours such a tilted conformation [59].

Factors affecting Pol II elongation

Throughout the gene body, Pol II efficiency is affected by structured chromatin. Nucleosomes, which consist of 147 base pairs of DNA wrapped around a protein core, pose an obstacle to the elongating Pol II. There are two major pausing sites at nucleosomal DNA [41, 59]. In gene-body nucleosomes, elongating Pol II accumulates directly upstream of the centre of the nucleosome, suggesting that this pausing is caused by the physical barrier created by nucleosomes. The other site is situated at the nucleosome entry site and is observed at the first nucleotide downstream of the transcription start site, termed "+1 nucleotide". It has been hypothesized that promoter-proximal pausing occurs at the entry site of the +1 nucleotide; however, closer inspection of the sequencing data in *Drosophila melanogaster* has indicated that there are two distinct promoter-proximal pausing sites. The downstream one corresponds to +1 nucleosome position, whereas the proximal pausing site is situated upstream, closer to the transcription start site [42].

Another obstacle to the elongating Pol II can be created by DNA-associated protein complexes and other polymerases. Conflicts are occurring between transcribing RNA polymerases and DNA polymerases that replicate DNA, even though processes of transcription and replication are generally separated in space and time [53]. Especially prone to such collisions are very long genes, which require an entire cell cycle or longer to be transcribed [30]. Collisions can also occur between converging polymerases and co-directionally between a travelling (upstream) and a paused (downstream) polymerase. Such collisions can lead to transcriptional stalling, and even in some cases to premature termination of transcription [59].

The nucleotide content of the DNA and DNA–RNA hybrid can also affect the processivity of Pol II. Repetitive DNA sequences rich in A–T base pairs form weaker RNA–DNA hybrids and in turn, destabilize the transcribing Pol II complex and cause pausing [65]. Likewise in *Saccharomyces cerevisiae*, GC-rich template sequences have on average fewer Pol II pauses compared with AT-rich sequences [77]. It shows that certain DNA sequences can promote pausing by destabilizing the RNA–DNA hybrid [59]. Conversely, recent analyses conducted using human cell lines has shown that genes with GC-rich sequences display robust hallmarks of Pol II pausing [69]. The proposed model predicts formation of DNA secondary structures upstream of the pausing site that stabilize Pol II in a paused state.

DNA motifs associated with transcriptional pausing

In addition to the aforementioned factors, DNA motifs, which are are short patterns of nucleotides in DNA, have been implicated in transcriptional pausing. Figure 2.2 shows a collection of motifs linked to transcriptional pausing. The first pausing motif was identified in the promoters of stalled genes in *Drosophila* embryos. This motif, termed 'pause button', is located within a 200 base-pair window centered at the transcription start site and consists of two pairs of CG dinucleotides separated by two nucleotides [31]. GC-rich motifs underlying promoter-proximal pausing sites were also uncovered in human cell lines, with pausing happening at cytosine in the DNA template

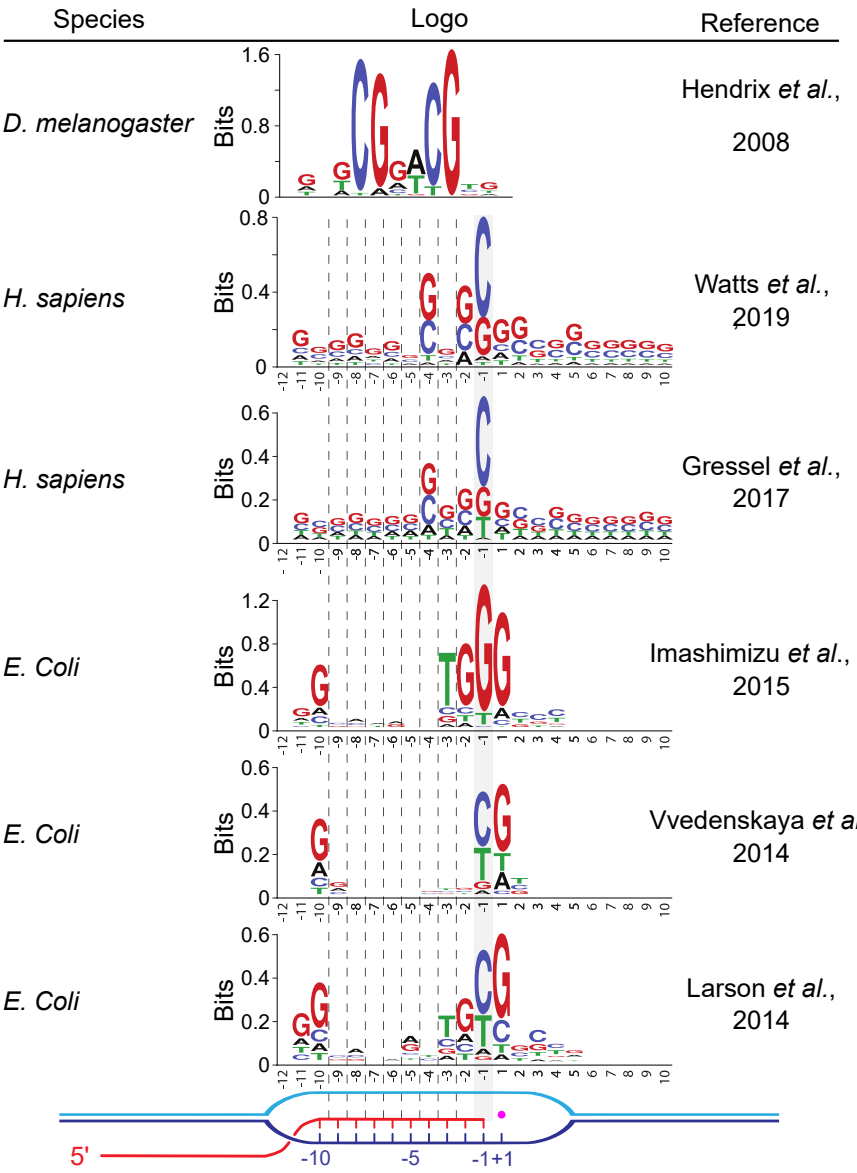

Figure 2.2: **Collection of DNA motifs implicated in transcriptional pausing.** Information content logos (middle column) together with the model organism used (first column) and the reference (last row). The logos are aligned to the schematic view of the transcription bubble (below) using dashed lines. The pink dot corresponds to a Mg^{2+} ion marking the active site of Pol II. -1 refers to the last nucleotide of the nascent RNA. +1 indicates the position in the DNA template where the next incoming NTP binds. This model is based on recent evidence from structural studies indicating that the RNA-DNA hybrid that spans the active site of the mammalian Pol II elongation complex is 9–10 bp long [4].

[26, 73]. The best characterization of DNA sequence-induced pausing was achieved for bacteria, where pausing is not limited to the promoter-proximal region, but occurs frequently throughout the gene body. Pausing sequences identified for bacteria exhibit positions with high information content at the ends of the RNA-DNA hybrid [34, 43, 72].

2.4 POL II PROFILING METHODS

The most commonly used approaches for studying the Pol II progression over the genome during transcription include Pol II profiling methods. The main idea behind all of these approaches is to get a snapshot of the distribution of polymerases over the genome in a large population of cells. The locations of slow polymerase progression can be inferred from the obtained Pol II distribution, by finding the regions with a higher relative abundance of the polymerase. Such snapshots can be obtained by stopping the transcription and then sequencing and quantifying either DNA fragments occupied by the polymerase (Pol II ChIP-seq) or the 3′ ends of the nascent transcripts (Run-On assays and NET-seq). Below, we describe relevant selected Pol II profiling methods based on Mayer *et al.* [52].

2.4.1 *Pol II Chromatin Immunoprecipitation sequencing*

Chromatin Immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is an experimental method used for genome-wide profiling of DNA-binding proteins. It can be used to determine the position of Pol II along genomes *in vivo*. Using formaldehyde, Pol II is reversibly crosslinked with the chromatin by covalent bonds. Next, the chromatin is fragmented and a Pol II specific antibody is used to co-immunoprecipitate DNA fragments bound by the polymerase. The isolation of the DNA fragments of interest is followed by the reversal of the crosslinking and the amplification of the purified DNA fragments. The relative amount of the captured DNA is assessed by high-throughput sequencing.

ChIP-seq is a widely used method and therefore several limitations of this approach were recognized. As the crosslinking binds both the template and non-template strand of the DNA, ChIP-seq lacks DNA strand specificity. Moreover, the spatial resolution of classical ChIP assays is limited, due to the relatively large size of the DNA fragments obtained during fragmentation. Furthermore, ChIP-seq reveals the location of all polymerases, not only the transcriptionally active ones, which might be not desired in some studies. Additionally, the strength of the obtained profiles depends strongly on the specificity and quality of the antibody used.

2.4.2 *Native Elongating Transcript sequencing methods*

Native Elongating Transcript sequencing (NET-seq) methods allow us to determine *in vivo* the precise position of all transcribing polymerase complexes. The main idea of this approach is to capture not only the position of the polymerase itself, but the 3′ end of the nascent RNA that marks the position of the active centre of a polymerase and is a hallmark of ongoing transcription. The nascent transcripts are captured without relying on crosslinking the polymerase, but by exploiting the high stability of the DNA-RNA-polymerase complexes.

NET-seq library preparation is initiated by stopping the Pol II transcription using the transcription inhibitor called α-amanitin. Then, the cells are fractionated in a way that allows retrieving nuclei. The chromatin is isolated from the nucleus together with the transcribing polymerases and nascent RNA molecules attached, exploiting the high stability of the DNA-RNA-polymerase complexes. The RNA is purified from those complexes and is

further processed in a way that protects its 3′ ends including the last added nucleotide. Next, a linker is ligated to the 3′ end of the RNA molecule. The linker comprises a random molecular barcode and sequences necessary for obtaining complementarity to the primers for the reverse transcriptase, PCR amplification, and sequencing. The random molecular barcode consists of 6 or 10 (depending on the library) random nucleotides that enable the removal of PCR duplicates and the recognition of the reverse transcriptase artifact in the bioinformatical analyses. The next steps prepare the library for the high-throughput sequencing. To allow for Illumina sequencing and balance the PCR amplification, the RNA molecules are fragmented and only fragments of the desired length are selected. Then, cDNA is synthesized using reverse transcriptase. The cDNA molecules are later amplified. In the last step of the experimental part of the protocol, the read insert is sequenced from the 3′ end together with the random molecular barcode. In this way, the identity and abundance of the 3′ ends of purified transcripts are revealed and further bioinformatical processing produces the Pol II occupancy profiles that reflect the position of the polymerase with a single nucleotide resolution and in a strand-specific manner. The main steps of the procedure are also presented in Figure 2.3.

In addition to the 3′ ends of nascent RNAs, NET-seq captures non-nascent RNA species and RNA processing intermediates that enter the library during the purification step. Those RNA molecules are selected probably thanks to the strong bonds that they can form either the nascent RNA or the chromatin [47]. These background RNA sequences can be computationally identified and removed, but processing the data in this manner decreases the fraction of reads that are informative about the Pol II position. This caveat can be overcome with a variant of NET-seq, termed HiS-NET-seq, which involves short metabolic labelling using modified nucleotides such as 4-thiouridine before the inhibition of the transcription. Thanks to the labelling, the newly produced RNAs can be enriched by selecting only the molecules that include the modified nucleotides.

Another problem that complicates the NET-seq analyses is the presence of artificial reads, whose 3′ ends do not reflect the 3′ end of nascent RNAs. The artificial reads can be generated during the cDNA synthesis step for those RNA molecules that harbour sequences partially complementary to the reverse transcriptase primer. However, the presence of reads originating from the RT mispriming can be limited using a variant of NET-seq termed 'nested NET-seq' [21], in which the design of the linker prevents PCR amplification of those reads.

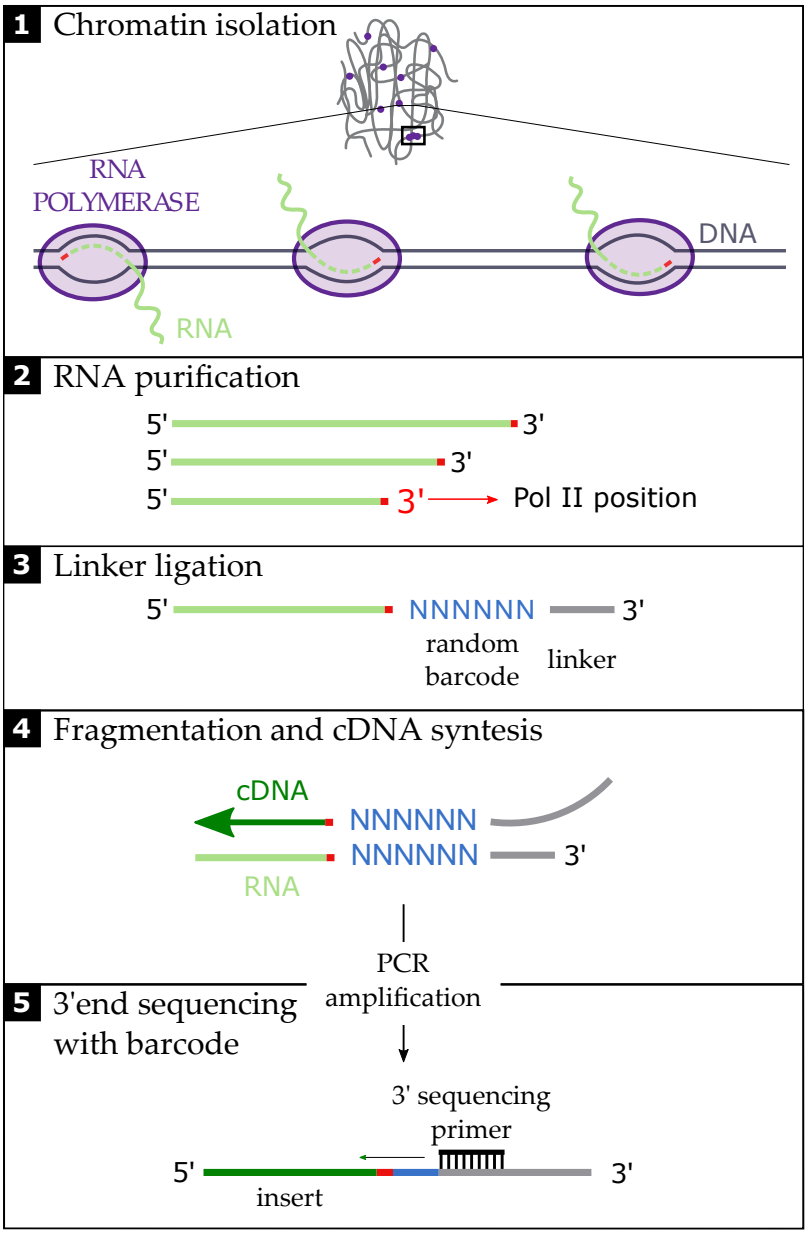

Figure 2.3: **Schematic overview of the NET-seq protocol.** Numbers in the corners indicate the subsequent steps. In the first step, the nuclear chromatin is isolated together with the transcribing polymerases and nascent RNA molecules attached to it. The second step comprises the purification of the RNA. Next, a linker including a random molecular barcode is ligated to the RNA molecule. In the fourth step, RNA is fragmented and cDNA is synthesized. The molecules are later amplified to allow for the high throughput sequencing. In the last step of the experimental part of the protocol, the read insert is sequenced from the 3′ end together with the barcode. This figure is adapted from Mayer *et al.* [51].

3 BIOINFORMATICAL BACKGROUND

This chapter covers the bioinformatical prerequisites used in the thesis. It presents selected statistical distributions used in genomics, focusing on modeling the expected number or expected maximum number of reads per position in NET-seq data. It introduces the concept of hypothesis testing and the special case of multiple hypothesis testing. Finally, we discuss various bioinformatical concepts helping us understand the results such as sequence motifs.

3.1 NET-SEQ POL II OCCUPANCY TRACK

Most high-throughput sequencing assays produce genomic tracks originating from multiple cells. A genomic track is a series of data units positioned on a line representing a reference genome in a line-based coordinate system [27]. A track element is a basic informational unit of data with associated genomic coordinates that may or may not be explicitly specified. In the case of a Pol II ChIP-seq, a typical genomic track represents the number of reads spanning each nucleotide position.

Consider a polymerase occupancy track obtained just for a single cell. In one cell, at most one active centre of a polymerase can be observed at a single nucleotide in a haploid genome. If we obtained a polymerase occupancy track from just a single cell, only those nucleotide positions at which active centres of the polymerases were situated when the transcription was stopped would be marked with 1s. Such a genomic track could provide information about e.g. the frequency of the transcription initiation (by examining the distances between subsequent polymerases transcribing the same gene) or the scale of the divergent transcription (by assessing the number of transcription start sites with polymerases transcribing in the opposite directions). However, it does not inform us about the time that polymerase spends at a given genomic position nor allow us to recognize the positions at which polymerase spends significantly more time. To approximate the time spent by a polymerase at a single-nucleotide position, we need information coming from multiple cells.

The NET-seq track can be seen as a sum of polymerase occupancy tracks coming from single cells. Therefore, we need to keep in mind that our observations are made for a population of cells, which is possibly heterogeneous. The signal intensity at a position corresponds to the number of cells in which we observed the polymerase at this position. Having included a large enough number of cells, the expected number of polymerases at a locus is the same for every position assuming a constant transcription speed. In case a signal intensity is higher for a position in a locus, it means that polymerase is encountered more frequently there and we assume that it spends more time at this position. At the same time, if the average number of polymerases between two loci differs, the possible explanations include differences in the initiation rate and the average transcriptional speed between those loci.

3.2 STATISTICAL DISTRIBUTIONS USED IN GENOMICS

A **random variable** is a variable that, when measured during an experiment, takes a numeric value from a set of possible values. The probability model, such as a **probability distribution**, defines relative likelihoods of the various possible values. The probability model depends on the process (e.g. the way we conduct the experiment) that generates the random variable.

Here, we motivate the use of selected discrete distributions for modeling the counts per position forming a genomic track. The number of reads x_i being assigned to the position i depends on the proportion p_i of the DNA fragments in the library that originate from the position i. If the total number of DNA fragments in the library is much larger than the number of sequenced reads N, which is true for the sequencing experiments, the number of reads x_i being assigned to the position i comes from a binomial distribution. The number of trials N and the probability of success p_i of the binomial distributions are given by the number of sequenced reads N and the proportion p_i of the DNA fragments in the library that originate from the position i. Typically, the number of sequenced reads N is known, but it is often our goal to estimate the probability of success p_i. For a large number of sequenced reads N and small probability p_i, which is the case for NGS experiments, Poisson distribution is a close approximation of the binomial distribution.

3.2.1 *Poisson distribution*

The **Poisson distribution** is used to describe the distribution of rare events in a large population. For the Poisson random variable $X \sim Poisson(\lambda)$, the probability mass function of the Poisson distribution is defined using a single parameter λ:

$$Pr(X = x|\lambda) = \frac{\lambda^x e^{-x}}{x!}, \qquad x \in \mathbb{N}.$$

The possible values of a Poisson random variable X are all the non-negative integers. Additionally, the parameter λ specifies both the expected number of counts per interval and its variance.

In genomics, the Poisson distribution is frequently used to model the expected number of reads in a genomic range and to detect the ranges with unexpectedly high read counts.

3.3 HYPOTHESIS TESTING

A **hypothesis** is a proposed explanation for a phenomenon or, in statistical terms, a statement that is supposed to be true (see Motulsky [57]). One of the most commonly used methods for making decisions or judgments about the value of a particular observation is to perform a hypothesis test. A hypothesis test involves two hypotheses: the null hypothesis and the alternative hypothesis. The **null hypothesis** (H_0) is a statement to be tested, whereas the **alternative hypothesis** (H_A) is a statement that is considered to be an alternative to the null hypothesis. The hypothesis test aims to see if the null hypothesis should be rejected in favour of the alternative hypothesis. For example, we can test the hypothesis that the observed Pol II peak is not stronger than expected given the local Pol II occupancy.

The p**-value** is the probability of obtaining test results at least as extreme as the observed results, under the assumption that the null hypothesis is correct. A connected term is the **significance level of the test** α that describes the probability of rejecting the null hypothesis, in case the null hypothesis is true. If the obtained p-value is lower than the statistical significance level chosen before testing the hypothesis, it suggests that the observed data is inconsistent with the null hypothesis and that the null hypothesis may be rejected. Such a result is then said to be **statistically significant**. However, it is important to keep in mind that even if a result is statistically significant it does not automatically mean that the finding is scientifically or clinically significant. The observed effect might still be too small to be interesting or worthy of further investigation.

There are two types of errors that we can make testing a hypothesis: we can reject the null hypothesis when it is actually true (**type I error**) or we can accept a false null hypothesis (**type II error**). The type I error is controlled by the chosen significance level α, as the lower, the significance level is the lower is the probability of incorrectly rejecting the true null hypothesis. However, a low significance level increases the chance that a significant difference will not be found, even if the null hypothesis is false.

We can distinguish two types of statistical hypothesis tests: parametric and non-parametric ones. In **parametric tests**, we make assumptions about the type and parameters of the probability distribution from which the sample is drawn. In contrast, **non-parametric hypothesis testing** does not require these assumptions, meaning that the data can be collected from a sample that does not follow a specific distribution. In general, parametric tests have higher statistical power than non-parametric tests, meaning they are more likely to correctly reject the null hypothesis when the alternative hypothesis is true. However, violating the assumptions about the distribution might lead to incorrect or misleading results of the analysis. Therefore, non-parametric tests have broader applicability, due to their distribution-free nature.

A subcategory of non-parametric tests, which gained popularity with increasing computer power, is **resampling**. It can be used for simulating an empirical distribution of an estimator. For example, we can estimate the expected height of a peak given the local read density in NET-seq data without explicitly using the Gumbel distribution by reshuffling the read assignment to the position and extracting the maximum number of reads per position multiple times. In this way, we obtain a simulated distribution of maxima that can be further used for estimating the expected value of the maximum or its confidence intervals.

3.3.1 *Multiple testing correction*

With every test made, we face a risk of committing a type I error with the probability equal to the significance level α. With the increasing number of tests, the probability of rejecting at least one null hypothesis by chance increases, and the probability of committing the type I error in at least one test can be calculated using the chosen value of α. In other words, it is impossible to interpret small p-values without knowing how many comparisons were made, as some p values are likely to be small just by chance. Therefore, to minimize the chance of spuriously rejecting a true null hypothesis, we have to apply a multiple testing correction method.

There are multiple methods designed to lower the probability of falsely rejecting true null hypotheses, by controlling either the family-wide error or the false discovery rate. The **false discovery rate** (FDR) is the expected fraction of tests declared statistically significant in which the null hypothesis is actually true, or in other words, the probability that a null hypothesis is true given that the null hypothesis has been rejected [24]. To infer the frequency of true null hypotheses, we use the distribution of p-values. The distribution of p-values consists of two components: a uniformly distributed p-values between 0 and 1 for true null hypotheses, and a right-skewed distribution of p-values for false null hypotheses. The main goal of false discovery rate control is to set significance levels for a collection of tests in such a way that the proportion of true null hypotheses among tests declared significant is lower than a specified threshold. Benjamini and Hochberg, who developed the original false discovery rate control method [3], proposed controlling the false discovery rate control for a study with n tests with maximum false discovery rate d by declaring the k tests with the smallest p-values significant, where k denotes the largest index i for which $p_i \leq d\frac{i}{n}$ for p-values sorted in ascending order.

3.4 SEQUENCE MOTIFS

Sequence motifs are short patterns of nucleotides in DNA or RNA. They are often a hallmark of a genomic location that is important for a biological function. They can indicate sequence-specific binding sites for proteins, mRNA processing sites (including splicing, editing, cleavage and polyadenylation) and others [19]. The motifs are usually constructed by creating a collection of sequence fragments of the same length from positions at which the biological phenomenon of interest occurs. For example, in the case of transcriptional pausing, we can collect all the genomic sites at which polymerase pauses and then extract 50 nucleotide long sequences around these sites.

Sequence motifs are frequently used to describe the genomic locations where transcription factors interact with DNA in a sequence-specific manner. These genomic locations are called as Transcription Factor Binding Sites and the preferentially bound sequences can be represented using **Transcription Factor Binding Motifs** (**TFBM**) [37]. Similarly, **RNA-binding protein motifs** provide a useful framework to describe the propensity of RNA-binding proteins to interact with RNA [23].

One of the possible representations of sequence motifs is the **consensus sequence**, for which every position shows the most frequent nucleotide at that site in the set of sequences. A more informative description of a motif can be provided through a **Position Frequency Matrix** (PFM), where we record how often each base occurs in known sites, rather than only keeping track of the most common base at each position. Such a PFM can be visualized using the **sequence logo** with the height of each nucleotide proportional to its relative frequency. The standard logo plots tend to visually highlight letters that are enriched, meaning where a nucleotide appears more frequently than expected. The height $h_i(n)$ of the nucleotide n at the position i of the logo can be calculated as:

$$h_i(n) = f_i(n) log_2 \frac{f_i(n)}{b(n)}, \tag{3.1}$$

where $f_i(n)$ is the frequency of the nucleotide n at the position i and $b(n)$ is the background frequency of the nucleotide n [19]. The background frequency is an important parameter used in the enrichment score calculation, where it describes the expected frequency of encountering a nucleotide. A uniform background frequency equal to $\frac{1}{4}$ for a four-letter nucleotide alphabet is commonly used. However, the background frequencies can be adjusted to better reflect the prior knowledge about the region of origin of the sequences. For example, as the GC content varies not only between species but also between different genomic regions in the same organism [60], it is crucial to define the background composition adequately to the research question. Assuming improper background distribution might mask potentially interesting patterns.

In cases where identifying depletions is also interesting, Enrichment Depletion Logos (EDLogos) can prove useful [18]. For EDLogo, the height $h_i(n)$ of the nucleotide n at the position i of the logo is an absolute value $|r_i(n)|$ of the ratio $r_i(n)$, which can be calculated as:

$$r_i(n) = log_2 \frac{f_i(n)}{b(n)}, \tag{3.2}$$

where $f_i(n)$ is the frequency of the nucleotide n at the position i and $b(n)$ is the background frequency of the nucleotide n. The nucleotide n is then plotted at the position i above the x axis if $r_i(n)$ is positive, or below the x axis if $r_i(n)$ is negative.

4

MACHINE LEARNING BACKGROUND

This chapter introduces the supervised learning algorithms used for classification problems. It presents the methods to evaluate both a model's performance and its ability to generalize the predictions to the unseen data. Finally, we discuss how a machine learning user can understand and interpret the prediction made by a machine learning model, using random forest as an example.

4.1 TERMS AND NOTATION

Machine learning was first defined in 1959 by Samuel [62] as a "field of study that gives computers the ability to learn without being explicitly programmed". Machine learning models are used to extract meaningful information from existing data to improve performance in a given task. We present the basic term and concepts used in this field of study based on Chapter 1 of Hastie *et al.* [29] unless stated otherwise. The process in which machine learning models infer knowledge from the collected data is called **training**. Since machine learning techniques are applied to a wide spectrum of problems, there are multiple ways of categorizing machine learning algorithms (see Chapter 5 of Goodfellow *et al.* [25]). One of the most commonly encountered division between machine learning approaches recognizes supervised and unsupervised learning. Both learning approaches experience an **input**, which is a data set consisting of many **examples**, alternatively called **data points**. An example is a collection of features and is typically represented as a vector $x \in \mathbb{R}_n$, where each entry x_i of the vector is another feature. A **feature** is a quantitative or qualitative measurement describing an object or an event that we want the machine learning system to process. An input is usually represented as a matrix X, where each row corresponds to an example and each column to a feature. Additionally in supervised learning, every example is connected to a **label**. A vector of labels corresponding to all of the examples is called an **output** and is usually denoted using a vector $y \in \mathbb{R}^n$, where each entry y^j of the vector is a label. The goal of **supervised learning** is to use the inputs to predict the values of the output. In the **unsupervised learning**, only the features are observed and the values of the outcome are not available. In the case of an unsupervised learning problem, the task is to describe the internal structure of the input data. This shows that the choice of the machine learning approach can often be motivated by the data availability and the task we are interested in solving.

4.2 SUPERVISED LEARNING

The goal of supervised machine learning is to learn from labelled data. Depending on the character of the outputs, we can distinguish the following types of supervised learning problems: regression and classification (see Chapter 2 of Hastie *et al.* [29]). In a **regression** problem, the output is a quantitative measurement. To solve this problem, the machine learning algorithm needs to find a function $f : \mathbb{R}_n \longrightarrow \mathbb{R}$. In the **classification**

problem, the output is a qualitative measurement. In this problem, we expect machine learning algorithm to decide to which of k categories an example belongs. To solve this problem, the algorithm needs to find a function $f : \mathbb{R}_n \longrightarrow \{1, \ldots, k\}$.

The **binary classification** is a special case of the classification problem, where we want to discriminate between two categories and the numerical value of the label can take two possible values $y^j \in \{0, 1\}$. The examples and their labels are commonly referred to as a positive example, when the numerical value of the label equals 1, and a negative example otherwise.

4.2.1 *Logistic regression*

Logistic regression can be seen as an adaptation of the linear regression to a classification problem. In difference to linear regression, which predicts quantitative outputs, the outcomes of the binary classification problem can take only values 0 or 1. We can address this problem by predicting the probability of the example belonging to one of the classes e.g. the positive class. The probability of an example belonging to the positive class determines its probability of belonging to the negative class, as the sum of those two probabilities needs to add up to 1. In this way, we transformed the binary classification problem into a regression problem. However, linear regression might return any real number as a prediction, whereas to predict probability we need to ensure that the obtained value lies between 0 and 1. One way to solve this problem is to use the logistic sigmoid function to squash the output of the linear function into the interval from 0 to 1 and interpret that value as a probability.

The logistic function $\sigma(x)$ is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{4.1}$$

In the linear regression model, the relationship between predicted outcome $\hat{y}$ and features of the example is modelled with a linear equation:

$$\hat{y}^{(i)} = \beta^T x^{(i)} + \beta_0, \tag{4.2}$$

where β is a vector of parameters and β_0 is an intercept. Squashing the linear equation using the logistic function, gives us the probability that the example $x^{(i)}$ belongs to the positive class with the label $y^{(i)}$:

$$P(y^{(i)} = 1 | X = x^{(i)}) = \frac{1}{1 + e^{-(\beta^T x^{(i)} + \beta_0)}}. \tag{4.3}$$

The probability that the example $x^{(i)}$ belongs to the negative class with the label $y^{(i)}$ can be then expressed as:

$$P(y^{(i)} = 0 | X = x^{(i)}) = 1 - P(y^{(i)} = 1 | X = x^{(i)}). \tag{4.4}$$

The above mentioned conditional probabilities are often denoted as $p_1(x; \theta)$ and $p_0(x; \theta)$ for the positive and the negative class accordingly, to emphasize the dependence on the entire parameter set $\theta = \{\beta_0, \beta^T\}$. To make the final assignment of the classes, we set a probability boundary threshold, e.g. 0.5 for a balanced binary classification.

The model parameters are fitted using the **maximum likelihood**. It is easier to use negative log-likelihood function, which can be expressed as:

$$l(\theta) = -\sum_{i=1}^{N}\{y_i \log p_1(x_i;\theta) + (1-y_i)\log(1-p_1(x_i;\theta)\}, \tag{4.5}$$

where N is the number of training examples. To fit the model, we need to find the parameters that minimize the negative log-likelihood. We do it by setting the derivatives of the negative log-likelihood function to 0 and solving the obtained equations.

We often want to limit the number of non-zero parameters β to be able to easily recognize the important predictors and increase model interpretability. We can obtain the desired effect by adding a term to the negative log-likelihood function that penalizes parameter vector θ if its norm is large, thus shrinking the coefficients β. One of the commonly used shrinkage method is the elastic-net. The additional penalty term in the elastic-net can be expressed as:

$$\frac{1}{C}\sum_{j=1}^{p}\{\alpha\beta_j^2 + (1-\alpha)|\beta_j|\}, \tag{4.6}$$

where C is the inverse of the regularization strength and the parameter α takes values between 0 and 1. Parameter α regulates the characteristic of regularization. When α equals 0, the first term in the sum of the Equation 4.6 disappears and the Lasso regularization is performed, and when α equals 1, only the first term is retained and the ridge regularization is applied.

4.2.2 *Decision trees*

A **decision tree** (see Izenman [35]) learns a sequence of if-else questions about individual features in order to infer the labels of the examples. Those questions are being asked at nodes, and branches represent the possible outcomes. A decision tree consist of a **root node**, where the first decision is made, **parent nodes**, which are split into **daughter nodes**, and **leaves**, which do not split anymore and contain labels. There may be multiple leaves with the same label.

A CART (classification and regression tree) is constructed by either splitting or not splitting each node of the tree into two daughter nodes at each of the if-else questions. In this way, the feature space is divided into non-overlapping (hyper-)rectangles, which are then assigned a label.

4.2.2.1 *Classification trees*

In the case of classification trees, the leaves take values of the class labels. In difference to the logistic regression, that uses a single linear boundary to divide the classes, classification trees use multiple (hyper-)planes to divide the feature space. Such an approach can yield superior classification performance, in the case of the classes being separated by a clearly non-linear boundary. However, greater complexity of a decision tree might allow it to learn training examples 'by heart' and lead to a poor performance on new examples.

Growing a classification tree involves: (1) choosing the feature and criterion to use for splitting at each node, (2) determining when a node should become a leaf and (3) finding an optimal label assignment at each leaf.

The last problem concerning finding an optimal label assignment for a leaf has a relatively simple solution. The winner-takes-all approach is applied. A leaf takes the label of the most represented class, in terms of number of examples, from all the examples that ended up in the leaf following the sequence of if-else questions.

The main idea guiding the process of finding the optimal split feature and criterion for each node is that class membership should become more homogeneous or pure in each of the branches stemming from the node. A **node impurity** is lowest when all examples split to the node belong to the same class and the highest when both of the classes are equally represented. There are two commonly used functions assessing homogeneity of a node: the **cross-entropy** and the **Gini index**. The Gini index is used by default for the classification tree growing, but in practice there is not much difference between these two types of node impurity functions. In the binary class case, the Gini index $i(\eta)$ of a node η is defined as:

$$i(\eta) = 2p(1-p), \tag{4.7}$$

where p is the fraction of examples belonging to the positive class at node η. The goodness of a split is then determined by the decrease in impurity defined in terms of Gini indexes. We calculate it as the difference between the Gini index of the parent node and the weighted average of Gini of indexes of the daughter nodes, where the weights are proportional to the fraction of examples of the parent node that are directed to the daughter node. Mathematically, the goodness of split ξ can be expressed as:

$$\Delta i(\eta, \xi) = i(\eta) - (r \cdot i(\eta_r) + (1-r) \cdot i(\eta_l)) \tag{4.8}$$

r is the fraction of examples of the node η that are directed to the right daughter node η_r and the left daughter node is denoted as η_l. The best split is the one that has the largest value of the goodness of split. We can see that the maximum goodness of split for a node η takes equals the Gini index $i(\eta)$ in case each when each of the daughter nodes is pure and their Gini indexes equal zero.

The tree-growing procedure starts at the root node, which consists of all of the learning examples. The values of the goodness of split of all possible splits for a single variable are calculated for the root node in order to find the best split for that variable. The best split of a node is defined as the one that has the largest value of goodness of split over all single-variable best splits at that node. We use the best split of the root node to separate the examples into the daughter nodes. We repeat the procedure of finding the best split for each daughter node using only the examples that are split into the given daughter node. This sequential splitting process of building a tree layer-by-layer is called **recursive partitioning**. We call a tree **saturated** when all of the nodes are pure and they cannot be split any further. In a high-dimensional classification problem, the tree can easily get overwhelmingly large, especially if it is allowed to grow until saturation.

This brings us to the last aspect that needs to be determined, namely deciding when to stop growing the tree. One possible approach is to grow a 'large' tree and then prune off branches (from the bottom up) until the obtained subtree is of the 'right' size. This approach is commonly referred to as **pruning**. As there are multiple possible ways of pruning a tree, finding the 'right' tree is the crucial part of the process. The main measure guiding the process is the misclassification rate that should be minimal for the optimal

tree. For the binary classification problem, the **resubstitution estimate** of the misclassification rate in a node is defined as the fraction of the examples belonging to the less represented class, in terms of number of examples, from all the examples that ended up in the node following the sequence of if-else questions. However, approaches alternative to the resubstitution estimate are used in practice for estimating the misclassification rate, due to its poor properties e.g. nondifferentiability.

Additionally, the tree growing can be terminated if a certain stop criterion is met. The most commonly used ones include defining the maximum depth of the tree and the minimum size of a leaf, allowing a split to happen at a node only if the daughter nodes are larger than a certain critical size.

4.2.2.2 *Regression trees*

In the case of regression trees, a leaf takes a constant value as a label. Growing a regression tree involves the same steps as growing a classification tree. However, the principles leading all of the steps are different. The process of finding the optimal split for each node is guided by the least squares criterion. The main idea is that the optimal split should minimize the sum of the distances between the examples and the mean value of all of the examples in the corresponding daughter nodes. The next step, namely deciding when to stop growing a regression tree, is resolved by growing an overly large tree and pruning it or by setting a stop criterion such as the maximum depth of the tree. Finally, a leaf takes the average of outputs of all examples in that leaf as its label.

4.2.3 *Random forest*

Ensemble learning is a machine learning paradigm where a prediction model is built by combining the strengths of a simpler base model, called **base learners**. Ensemble methods include: bagging, boosting and stacking. The key idea of **bagging** is to average many noisy but unbiased models, and in this way reduce their variance (see Chapter 15 of Hastie *et al.* [29]). To obtain the prediction in a classification problem, a committee of base learners each cast a vote for the predicted class and classifies based on the majority vote.

The major problem with decision trees is their instability: a small change in the data might lead to a very different series of splits and therefore to a different label assignment. At the same time, if grown with sufficient depth, decision trees have relatively low bias, which makes them perfect candidates for bagging. **Random forest** is a modification of a bagging technique that builds a large collection of randomized trees. The randomization is obtained by constructing each individual tree based on a different randomly sampled subset of the training observations, and determining each split within a tree from a randomly sampled subset of features. The scheme of classification with a random forest is presented in Figure 4.1.

4.2.4 *Boosting trees*

Boosting is another of the ensemble learning techniques. Similarly to bagging, boosting combines the outputs of many base models to produce a powerful committee (see Chapter 10 of Hastie *et al.* [29]). However, the resemblance to

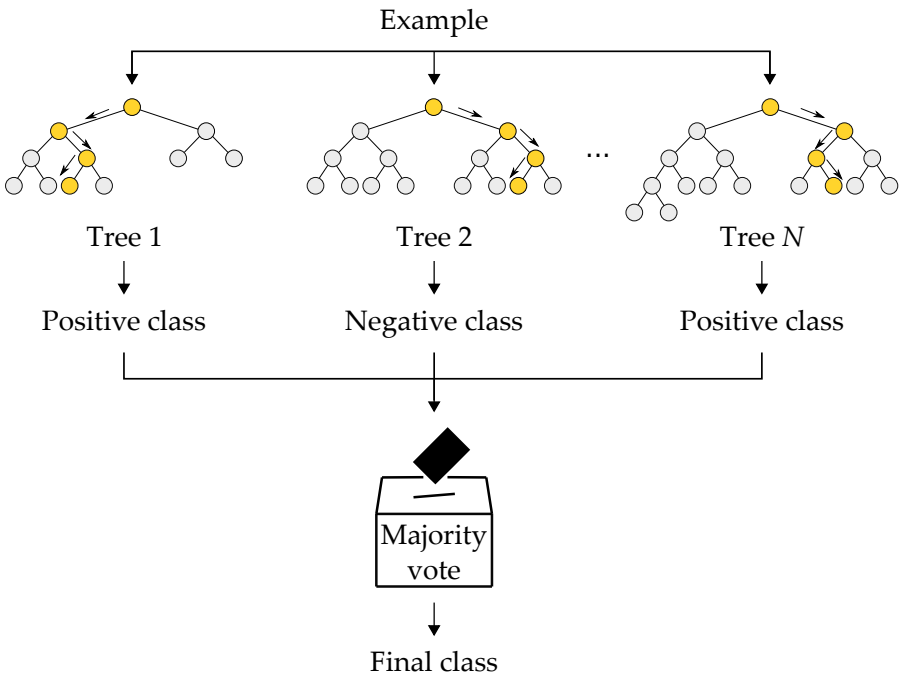

Figure 4.1: **Classification with the random forest.** Each tree is presented with a new example and predicts its class label. The final class assignment is performed based on the majority of votes of individual classification trees.

bagging is only superficial, as both teaching a single base model and voting is performed differently. Boosting produces a sequence of weak classifiers that are trained using repeatedly modified versions of the original data. The final prediction is usually produced combining the predictions from all classifiers through a weighted majority vote, where more accurate base learners have higher influence on the final prediction.

Gradient Boosting is a popular boosting algorithm, in which each predictor in the ensemble corrects its predecessor's error. Each predictor is trained using the original input and the negative gradient of a loss function, also called pseudo-residuals, as labels. Counterintuitively, the gradient boost classification algorithm uses regression trees of a limited maximum depth. A leaf of a regression tree might return any real number $\hat{y}$ as a prediction, whereas the goal of the classification is predicting a class label. Similarly to the logistic regression, the logistic function 4.1 is used to transform real-valued leaf labels $\hat{y}$ to the predicted probability $\hat{p}$ of the example belonging to the positive class:

$$\hat{p} = \frac{1}{1 + e^{-\hat{y}}}. \tag{4.9}$$

The Equation 4.9 can be used to derive the relation between the predicted value $\hat{y}$ and the predicted probability $\hat{p}$:

$$\hat{y} = \frac{\hat{p}}{1 - \hat{p}}. \tag{4.10}$$

The above Equation 4.10 is also commonly referred to as a logit transformation. As we can see, the predicted value $\hat{y}$ corresponds to the log odds

of the example belonging to the positive class. As leaf labels correspond to the log odds, a leaf of the regression tree in gradient boosting does not take the average of outputs of all examples in that leaf as its label. Instead, the negative log-likelihood function (Equation 4.5) is used to find the optimal label of a leaf, given the examples that are directed to the leaf.

The first tree in the chain consists of just a single leaf and gives the same initial prediction $\hat{y}_0$ for all of the examples. The initial prediction is calculated as log of the odds of encountering a positive example in the whole learning set. The pseudo-residuals Δ_i used to train the i-th tree are calculated as the difference between the observed probabilities p, equal 1 and 0 for the positive and negative examples respectively, and the probability predicted by the $(i-1)$-th tree.

An important parameter used in training gradient boosted trees is shrinkage. The prediction of each tree in the ensemble is shrunk after it is multiplied by a learning rate r, which takes a value between 0 and 1. As there is a trade-off between the learning rate and the number of estimators, decreasing the learning rate should be compensated by increasing the number of estimators, or otherwise it might prevent the ensemble from reaching the desired performance. The final prediction of the log odds for an example is the sum of the shrunk predictions of all trees in the chain. The class probability is then retrieved using Equation 4.9. The scheme of classification with gradient boosted trees is presented in Figure 4.2.

4.3 MODEL EVALUATION

4.3.1 *Performance measures*

There are several measures to assess the performance of a model. In case of the binary classifications, the most commonly used are: accuracy, precision, recall, true positive rate, and false positive rate. Those parameters are defined using four terms that describe possible results of a single binary classification. Knowing the labels, for every example we can describe the prediction at a specified probability threshold as a true positive, a true negative, a false positive or a false negative. The application of those terms to a binary classification can be visualized using the **confusion matrix** shown in Table 4.1.

Table 4.1: Confusion matrix presenting possible results of a classification.

Predicted label	**True label**	
	Positive	**Negative**
Positive	true positive (TP)	false positive (FP)
Negative	false negative (FN)	true negative (TN)

Accuracy describes the proportions of true classifications ($TP + TN$) among all classifications:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.11}$$

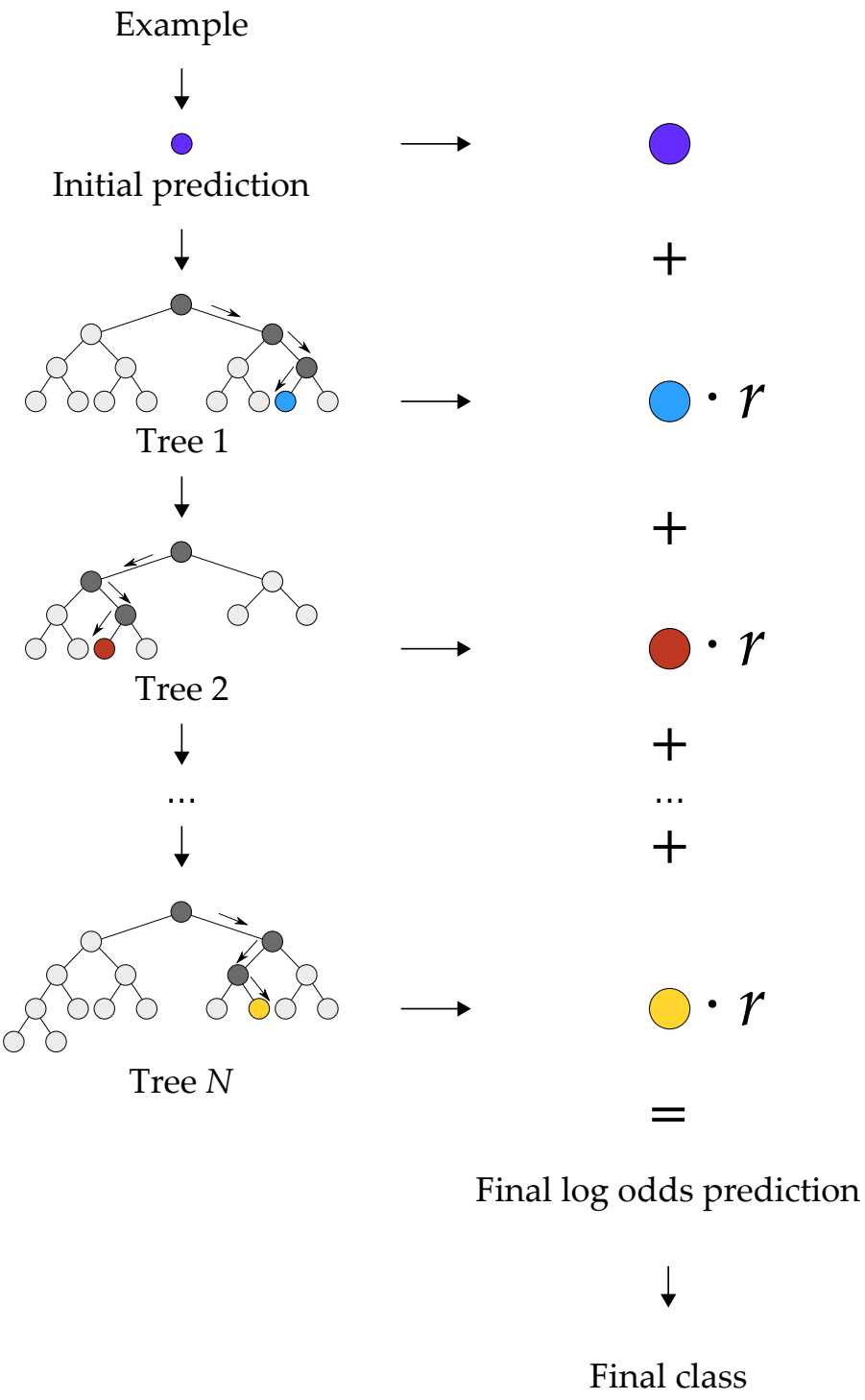

Figure 4.2: **Classification with the gradient boosted trees.** The first tree in the chain consists only of a single leaf and includes the initial prediction that is the same for every example. Each tree is presented with an example and a pseudo-residual obtained from the previous tree. The final class assignment is performed based on the majority of the sum of the shrunk predictions of all trees in the chain.

Accuracy measures reliability of the model. However, in case of an unbalanced data set, for which cardinalities of the classes are uneven, accuracy might give misleading results, as even a model always predicting the class with more examples will yield a high accuracy.

Precision, also called positive predictive value, measures the ratio of truly positive classifications to all positive predictions ($TP + FP$):

$$precision = \frac{TP}{TP + FP}. \tag{4.12}$$

It is a complement of the False Discovery Rate (FDR).

Recall, alternatively called **true positive rate (TPR)**, refers to the fraction of positive examples ($TP + FN$) that is detected by the model:

$$recall = \frac{TP}{TP + FN}. \tag{4.13}$$

Recall shows how sensitive the model is in identifying the positive examples.

False positive rate (FPR) refers to the proportion of the number of negative examples wrongly categorized as positive and the total number of actual negative events ($TN + FP$):

$$specifity = \frac{FP}{TN + FP}. \tag{4.14}$$

Receiver operator characteristics (ROC) and precision recall (PR) curves are diagnostic plots that can serve multiple purposes such as comparing performance of different classifiers or finding an optimal probability threshold for imbalance classification tasks. The ROC curve connects pairs consisting of the false positive rate (on the x-axis) and the true positive rate (on the y-axis) for a list of increasing probability thresholds. A diagonal line on the plot from the bottom-left to top-right indicates the scores for a random classifier and a point in the top left of the plot indicates a model with perfect skill. The area under the ROC Curve, so-called ROC AUC, provides yet another measure assessing the performance of a model. It takes a single value between 0.5 (corresponding to a random guess) and 1 (indicating perfect performance).

In case of an imbalance data set, ROC curve, similarly to the accuracy, might give misleading results. The **precision-recall (PR) curve** provides a more informative alternative in this case. The PR curve connects pairs consisting of the recall (on the x-axis) and the precision (on the y-axis) for a list of increasing probability thresholds. A point in the top right of the plot indicates a model with perfect skill and a constant line $y = a$, where a is the fraction of examples of the larger class, indicates the scores for a random classifier. Exemplary ROC and precision-recall curves are shown in Figure 4.3.

4.3.2 *Generalization error*

The main challenge and objective of training a machine learning model is obtaining a model that is able to **generalize**; that is, to make good predictions when applied to a data set independent of the one used to fit the model. The resubstitution estimate, introduced in Section 4.2.2, uses the same data that was used to derive the predictor, therefore the result is an overly optimistic view of prediction accuracy (see Chapter 1 of Izenman [35]).

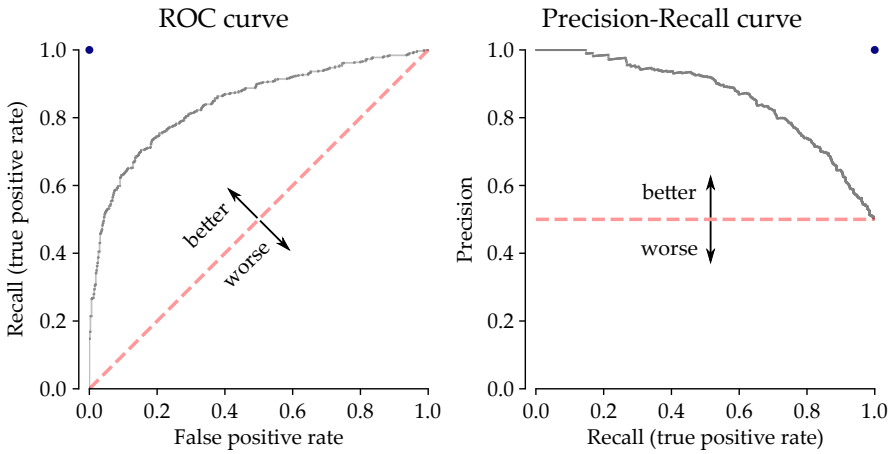

Figure 4.3: **ROC and PR curve.** Examplary ROC (left) and precision-recall (right) curves in grey. Curves of random classifiers marked with red, dashed lines. The dark blue dots show the scores of the perfect classification.

To get a more realistic estimate of accuracy, if the data set is large enough, it is a common practice to separate the data into three non-overlapping and independent data sets: a learning set, a validation set, and a test set. A **learning** or **training set**, as the name suggests, is used to train the model. A **validation set** is a data set used for model selection and assessment of competing models. A **test set** is a data set to be used for assessing the performance of the final model. The predictor is built using only the examples from the learning set (see Chapter 5 of Goodfellow *et al.* [25]). Then, the fitted model can be used to predict the output values of the learning set and compute an error measure called the **learning** or **training error**. In case of classification, we calculate the **learning error** as the fraction of all misclassified examples of the learning data set and we reduce this error. What separates machine learning from optimization is that we want the generalization error, also called the test error, to be low as well. The **generalization error** is defined as the expected value of the error on a new input. We typically estimate the generalization error of a machine learning model by measuring its **test error**, which in the classification problem is defined as the fraction of all misclassified examples of the training data set.

The question that might arise here is how can we affect performance on the test set when we can observe only the training set? The main assumption is that the training, validation, and test subsets of the data are independent and each generated by the same underlying distribution. If we sample from the underlying distribution repeatedly to generate the training set and the test set, the only difference between the two conditions is the name we assign to the data set we sample. Therefore the expected training set error is exactly the same as the expected test set error, as both expectations are formed using the same data set sampling process. As in reality, model parameters are chosen in a way that reduces the training set error, the expected test error is greater than or equal to the expected value of the training error.

A well performing machine learning model should be able to make both the training error and the difference between the training and test errors small. We can tell that the trained model is not performing well, if we observe one of these two behaviours: underfitting or overfitting. **Underfitting** describes

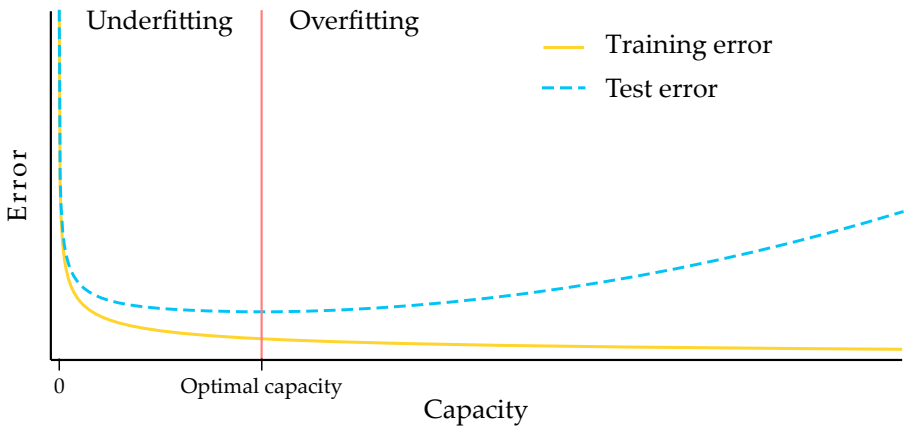

Figure 4.4: **A typical relationship between the errors and model capacity.** Training error decreases with increasing model capacity, whereas test error decreases till the optimal model is reached and increases for higher capacities. Models with too low capacities tend to underfit, leading to high training and test errors. Models with too high capacities tend to overfit, leading to a large difference between training and test errors.

the situation when the model is not able to obtain sufficiently low error rate even on the training set. **Overfitting** can be diagnosed when we observe sufficiently low error value on the training set, but the model fails to obtain the desired performance on the test set, making the gap between the training and test error large. Both of the problems have a basis in the **model capacity**, which reflects model ability to fit a wide variety of functions. Whereas models with low capacity may struggle to fit the training set leading to underfitting, models with high capacity tend to overfit by memorizing properties of the training set that do not help them to make predictions on the test set. The relation between the training and generalization errors and the model capacity is depicted in Figure 4.4. In general, a machine learning algorithm performs best when its capacity is appropriate for the true complexity of the task they need to perform.

4.3.3 *Hyperparameter optimization*

One way of searching for the model with an optimal capacity is trying various learning algorithms that specify different families of functions the model can choose from in order to reach the training objective. The capacity of a machine learning model can also be affected by adapting the algorithm's hyperparameters. **Hyperparameters** are settings that control the algorithm's behavior. In difference to model parameters, their values are not found by the learning algorithm itself, but they are set prior to the learning. An example of a hyperparameter is the number of trees in the random forest or the maximum number of features considered for splitting a node of a decision tree.

Settings controlling the capacity of the model cannot be learnt on the training set. Such hyperparameters would always choose the maximum possible model capacity, later resulting in overfitting, if learnt on the training

set. Other types of hyperparameters are not learnt because the setting is difficult to optimize. To find the hyperparameters and avoid overfitting, we need an independent data set to measure the prediction errors. A **validation set** consists of examples that the training algorithm does not observe and that do not belong to the test set. Using the validation set to tune the hyperparameters allows selecting the optimal model configuration.

4.3.4 *Cross-validation*

Dividing the data set into a fixed training set and a fixed test set can be problematic if the resulting sets are small. A small training set might not provide enough examples for the model to extract meaningful information and a small test set adds statistical uncertainty around the estimated test error, making it difficult to compare performance of different algorithms on the given task.

There are alternative ways of estimating the mean test error for small data sets. The main idea behind them is to repeat the training and testing on different, randomly chosen subsets of the original data set. The most commonly used procedure is the *k*-**fold cross-validation**, in which a partition of the data set is formed by splitting it into k non-overlapping subsets. There are k rounds of the training and testing computations. On trial i, the i-th subset of the data is set aside as the test set, and the rest of the data is used as the training set. Finally, the test error is estimated by taking the average test error across k trials. The k-fold cross-validation allows us not only to train and test the model on a small data set, but it also allows us to estimate the confidence interval of the measures of the performance and shows us how sensitive the model is to the selection of the subset of the training data set. Those benefits come at the price of increased computational cost.

4.4 INTERPRETABILITY

Machine learning **interpretability** is defined as the degree to which a machine learning user can understand and interpret the prediction made by a machine learning model. There are a couple of reasons why we might want to make machine learning models interpretable. The interpretability of the model is highly desired in scientific applications, where machine learning is often used not only to make future predictions but also to extract knowledge from complex data. We can use an interpretable model as a source of knowledge about the modeled problem, as during the process of learning, the model receives insight into the problem that allows it to give correct predictions. Additionally, being able to explain the predictions of the machine learning model facilitates spotting potential sources of errors and biases that the model has learnt. An overview of making supervised machine learning interpretable has recently been provided by Molnar [55] and all the definitions and concepts in this section are based on this book unless stated otherwise.

One way to achieve interpretability is to use only a subset of algorithms that creates interpretable models, such as the logistic regression or the decision tree. However, often the inherent interpretability of the model arises from its simplicity and comes at a cost of lower prediction accuracy than the one achieved for more complex algorithms. Another solution is to use model-agnostic interpretation measures. The great advantage of model-

agnostic interpretation methods is the flexibility obtained by separating the explanation from the model. The same interpretation method can be used for any type of model, facilitating model comparison in terms of interpretability.

4.4.1 *Permutation feature importance*

The **permutation feature importance** was first introduced by Breiman [7] for random forests and then developed into a model-agnostic version by Fisher *et al.* [20]. The idea behind this approach is straightforward: if a feature is important for making the prediction, removing the information about this feature will increase the model's prediction error. The importance of a feature is measured by calculating the decrease in the model's performance after permuting the values of the feature. Shuffling values of an unimportant feature leaves the predictions unchanged, because the model ignored the feature for the prediction. However, if the model heavily relied on a feature for prediction, shuffling its values increases the model error. Another approach for removing the feature's information is deleting the feature altogether, retraining the model and then comparing the model errors. Nevertheless, retraining with a reduced data set creates a different model than the one we want to examine, therefore limiting its usability for estimating the feature importance. Additionally, retraining a model can be computationally costly, and just permuting a feature can save a lot of time.

To get the ranked list of the most important features according to the permutation importance, first, the error measure is calculated for the original model. Then, we permute the values of one of the features and we calculate the difference between the error measured for the data set with the permuted feature and the error of the original model. We repeat the second step for all the features of the input. The higher the increase of the error is upon the permutation of a feature, the more important the feature is. An additional decision involves choosing if the process should be performed on the training or test data set. Both choices give us insights into different aspects. Using the training set informs us which features the model relies on for making predictions, whereas values of the importance obtained on the test set tells us which features contribute to the performance of the model on unseen data.

There are some possible problems encountered using the permutation feature importance. As the algorithm depends on randomly shuffling the feature values, repeating the process might give different results. A solution to stabilize the measure is averaging the importance measures over repetitions. Other problems might arise in case of a data set with correlated features. First of all, adding a correlated feature can decrease the importance of the associated feature by splitting the importance between both features, as in this case the association between the permuted feature and the outcome is not fully broken due to the presence of the correlated, unpermuted feature. Another problem is that permuting a feature might produce unrealistic data instances when two or more features are correlated. In this case, we measure the importance based on examples that might not occur in reality, which might create irrelevant estimates of the feature importance.

5

DETECTION OF POL II PAUSING SITES IN NET-SEQ DATA

In this chapter, we examine the NET-seq Pol II occupancy tracks, focusing on the positions with highest signal and previously reported artefacts. We propose refinements to the NET-seq data processing that improve the recognition of the problematic positions and limit their influence on the pausing site detection. We present a resampling-based peak caller suitable for sparse occupancy tracks, such as NET-seq, and we compare it to the testing approach based on the Poisson model. Finally, we investigate the peaks detected in NET-seq, focusing on the sequences underlying transcriptional pausing sites found in various genomic regions and organisms.

5.1 MOTIVATION

A preparation of every NGS library includes many molecular manipulations that may introduce some form of biases, in turn resulting in a skewed representation of the original molecules. Such a skew representation can affect accurate quantification, lead to false results, or mask potentially interesting patterns [6]. Therefore, understanding the shortcomings of the experimental method of choice and foreseeing the potential biases is a very important part of every bioinformatical analysis.

Even though the NET-seq has been introduced over a decade ago [13], most of the papers focus on the experimental protocol and biological conclusions reached using the method [13, 50, 51] and the complete guideline listing all necessary bioinformatical steps is missing. Such a guideline would facilitate the analyses, accelerate the detection of possible arifacts in the library, providing robust insight into the biological phenomena of interest faster.

In order to enhance our understanding of transcriptional pausing, we need to be able to recognized pausing sites in the polymerase occupancy tracks. Local enrichments of the polymerase occupancy can be detected and assessed using one of many peak calling algorithms and tools. However, most of the approaches were developed for ChIP-seq genomic tracks that are characterized by much lower spatial resolution than NET-seq data. As a result, a thorough parameter tuning is necessary before applying commonly used peak callers to detect and asses the statistical significance of a peak in sparse, zero-inflated NET-seq data. Additionally, some of the peak calling softwares require input in a form that is characteristic for ChIP-seq experiments such as paired-end sequenced data (whereas NET-seq experiment produces single-end data) or a control data set. The simplest solution is developing a peak caller that is designed for identifying pausing sites from the high-resolution genome-wide methods, that detects the peaks with a single-nucleotide resolution and performs the statistical assessments taking into account the sparsity of the data.

A rigorous processing of NET-seq data followed by a pause detection using a properly calibrated peak caller can enhance our understanding of transcriptional pausing. First, a high-resolution genomic map of Pol II

pausing sites can be used to determine the regions where the polymerase is more likely to pause and answer the questions whether transcriptional pausing in human cells is limited to promoter-proximal locations. Second, a set of high-confidence pausing sites is necessary to find the key drivers of Pol II pausing and potential motifs determining DNA sequence-induced pausing. Lastly, the positional information about transcriptional pausing can be integrated with the genome annotation and other genomic data in order to correlate the pausing with other molecular events and gain insight into the role of transcriptional pausing.

5.2 METHODS

5.2.1 *NET-seq library processing*

After obtaining the NET-seq reads, bioinformatical analyses are required to generate the desired genomic tracks. In this section we describe the NET-seq specific issues that we address in the preprocessing and the changes implemented to the previously published approach [51]. The used tools and the applied settings are listed in Supplemental Table S1.

First steps of the NET-seq preprocessing consist of removing adapters and controlling the read quality. As for most of the NGS libraries, NET-seq preparation includes a PCR amplifications step. To ensure an accurate quantification of the number of polymerases encountered at every genomic position, we remove PCR duplicates based on the Unique Molecular Identifiers (UMIs), also referred to as barcodes. This step is particularly important for NET-seq, as PCR amplification rates are high. We store the information about the number of PCR copies of every read and use it to assess the library quality. Resulting reads consist of two parts: a read insert (corresponding to the fragment of a nascent RNA) and a barcode.

Next, we create a dictionary linking a read insert with all the barcodes attached to it. That allows us to speed up the read mapping, because each read insert is mapped only once. To obtain the number of reads mapping to a region, we multiply each instance of a read insert mapping by the number of different barcodes attached to the read insert. In this way, we remove PCR-duplicates as we include each unique pair of a mapped insert and a barcode only once. The 3′ ends of read inserts are used for creating the Pol II occupancy tracks.

Mapping of the read inserts is a crucial step and there are several problems that need to be addressed in case of NET-seq. First, as read inserts correspond to nascent RNA fragments, we expect to encountered both spliced and unspliced moleculers. Therefore, we decided to use the read aligner STAR, which is designed for RNA-seq data and allows mapping against the genome while taking into account information about possible splice junctions. Moreover, many read mappers trim ends of the reads to optimize the alignments. Since 3′ ends of read inserts correspond to the locations of polymerases, we disable this feature in STAR to ensure that the start positions of the alignments correspond to the start positions of the original nascent fragments. Another mapping related question is how to deal with reads mapping to multiple locations. Such multi-mapping reads often originate from repetitive regions, which are scattered around the genome and often situated in the intronic regions of genes. For NET-seq peak detection such reads constitute a major challenge, because they can create pile-ups at multiple locations,

that are later not distinguishable from the Pol II pausing positions. The easiest and most stringent way is to discard all multi-mapping reads and only perform the analyses using only uniquely mapping reads. It is a commonly used approach and therefore in this thesis we discard multi-mapping reads.

Next issue that needs to be addressed is the RT-artefact. Reverse transcription (RT) is a step in the NET-seq library preparation in which a cDNA molecules are produced based on selected nascent RNAs. The RT-primer can anneal in a nonspecific way to the RNA template, which in turn gives rise to reads with spurious and incorrect 3′ ends. Such RT-artefact is particularly dangerous in assays requiring single-nucleotide precision like NET-seq, as it can cause misinterpretation of data [64]. Moreover, with the increased PCR amplification, the number of errors within a barcode rises. The errors creates read originating from the same cDNA molecule with a number of different barcodes, which allows such PCR duplicates escape detection during the deduplication step and later produces pile-ups of signal indistinguishable from real peaks.

Barcodes are not only a solution to recognize PCR duplicated reads, they also facilitate detection of the RT-prone positions. In NET-seq libraries, a molecular barcode consist of N (usually 6 or 8) random nucleotides attached directly to the 3′ end of the purified RNA. Since reads originating from RT-mispriming do not have the barcodes, the part of the sequence extracted as a barcode should be identical to the N nucleotides downstream of the mapped 3′ end of the read insert (referred to later as a downstream sequence). This property allows us to recognize all reads originating from RT-mispriming that do not harbour any amplification or sequencing errors. However, in case of short fragments characterized by increased amplification PCR duplication, not all barcodes are identical to the downstream sequence. Therefore, we introduced a recognition of RT-prone positions based on the similarity between the barcode and the downstream sequence that allows for d mismatches. A position is called RT-prone and masked if the fraction f of reads with barcodes similar to the downstream sequence is high. The fraction f is calculated using all reads (including PCR duplicates) mapped to the position position, and a barcode is called similar if the number of mismatches between the barcode and the downstream sequence is lower than a predefined number of mismatches d.

One last problem that needs to be solved bioinformatically arises from the RNA purification step. As NET-seq uses nuclear chromatine purification to enrich for the nascent RNA, it selects not only nascent RNA produced by Pol II, but also all other chromatin-associated RNA species [50]. Those include nascent RNA produced by other human polymerases (Pol I and Pol III), transient product of nascent RNA processing (e.g. splicing or miRNA maturation), and RNA species executing their functions in proximity of chromatin and nascent RNA such as snRNAs (e.g. splicing), snoRNAs or miRNAs (e.g. mRNA silecing). To avoid misinterpretation of the data, the signal in the above mentioned regions are masked. The complete list of all masked regions is included in the subsection below, called *Creating masking regions for NET-seq*. Further justification for masking the listed region accompanied with analyses is presented in the Result section of this Chapter. Finally, as mentioned before, only the 3′ ends of not masked, PCR deduplicated read inserts are used for creating the Pol II occupancy tracks.

5.2.2 *Creating masking regions for NET-seq*

We created masking files corresponding to regions known to be transcribed by other RNA polymerases than Pol II and known NET-seq contaminants. The regions were determined using the following annotations: GENCODE (v37), miRBase v22.1 and the UCSC's RepeatMasker. To find regions transcribed by mitochondrial polymerase, we extracted genes from GENCODE located on the mitochondrial chromosome. The Pol I transcribed genes were extracted from RepeatMasker using terms: rRNA, rRNA_pseudogene, LSU-rRNA_Hsa, and SSU-rRNA_Hsa, whereas Pol III transcribed genes were extracted from GENCODE annotation of tRNAs and from RepeatMasker using terms: 5S, 7SK, HY1, HY3, HY4, HY5, Y_RNA, tRNA, U6, BC200, and vaultRNA. For finding chromatin-associated RNA species, we looked for snRNA and snoRNA in GENCODE, and for the terms: U1, U2, U3, U4, U5, U6, U7, U8, U13, U14, and U17 in RepeatMasker. Masking splicing intermediates was executed by extracting 3′ ends of exons (excluding the terminal ones) and 3′ ends of introns using GENCODE annotation. Additionally, positions at the 3′ end of transcripts were masked to exclude the signal originating from full length, cleaved transcripts. As our analyses were limited to protein-coding and long non-coding genes, the end positions of the maturated miRNAs were not explicitly masked, as we have masked the whole pre-miRNA regions using miRBase.

5.2.3 *Pausing site detection*

The algorithm was designed for detecting signal intensity peaks from a single nucleotide resolution Pol II occupancy data corresponding to potential transcriptional pausing sites. However, after adjusting the parameters, it can be used to detect peaks in any sparse data with a single nucleotide precision. The algorithm consists of two main steps: peak detection and peak evaluation. In the first step, all local maxima are detected. A local maximum is a nucleotide position for which the neighbouring positions have lower signal intensity. Formally speaking, we are interested in position i for which we observe $x_{i-1} < x_i$ and $x_i > x_{i+1}$, where x_i is the signal x intensity at the position i. In the second step, a statistical test is applied for every detected peak to test if the peak has a significantly higher value than either the expected value (using the Poisson distribution) or the expected value of the maximum (using the empirical distribution) given the local Pol II density. To control for multiple testing, all obtained p-values are corrected using the Benjamini-Hochberg procedure. For downstream analyses, we considered pausing sites with a corrected p-value smaller than 0.05 as significant. We also define high-confidence pausing sites as pausing sites that are detected in all biological replicates at the same nucleotide position. Only high-confidence pausing sites are used for downstream analysis.

5.2.3.1 *Peak evaluation using Poisson distribution*

The approach presented here is a modification of the MACS algorithm [76], that allows for the application to sparse data with a single nucleotide resolution. The expected number of reads from a genomic region has a Poisson distribution with dynamic parameter λ_i that varies along the genome. The parameter λ_i is the average number of reads per position within the

window of a given length L centred at the position i. For libraries resulting in sparse data tracks, meaning data tracks with a lot of positions with no signal, only the positions with the signal are included. Therefore, the parameter λ_i for the position i is calculated as a ratio of the number of reads M and the number of positions l with non-zero signal. For NET-seq, we use a window of length L equal 200 nucleotides, a size that is an equivalent of four footprints of a Pol II.

5.2.3.2 *Peak evaluation using non-parametrical testing*

A nonparametric resampling approach is applied for every detected peak to test if the peak has a significantly higher value than the expected given the local Pol II density. Local Pol II density for the position i is described by the number of reads M and the number of positions l with non-zero signal within the window of a given length L (here 200 nucleotides) centred at the position i. A value of local maximum is simulated by redistributing M reads over l positions and extracting the maximum number of reads per position from the newly obtained read distribution. The redistribution is conducted following the null hypothesis that Pol II does not accumulate at any position and a read has an equal probability of being assigned to each of the positions in the local window of length l. Such resampling generates a pool of N (here 10000) simulated values of local maxima. Next, the p-value is estimated using the fraction of simulation experiments in which the simulated value of local maximum is greater or equal to the observed local maximum.

5.2.4 *Assigning pausing sites to genomic regions*

Based on their location, pausing sites were classified into one of four major categories: promoter-proximal, gene-body, antisense or intergenic. For defining the regions, we used GENCODE annotations (v37). A pausing site was classified as 'promoter-proximal' if located within 300 nucleotides downstream of the transcription start site (TSS). Pauses between +301 and the 3'-most polyadenylation (pA) site of a gene were classified as 'gene-body' pauses. If the pausing site was situated on the opposite strand of a gene in a region between 1000 nucleotides upstream of the most upstream TSS and the most downstream pA site, the pausing site was classified as 'antisense'. In case of an overlap between the gene-body region of one gene and the antisense region of another gene, the pausing location remains undetermined. All pausing sites located outside the listed regions (promoter-proximal, gene-body, antisense) were classified as intergenic pauses. Gene-body pausing sites were further specified into two subclasses. Gene-body pausing sites were 'exonic' if they overlap with annotated exons, otherwise they were labelled 'intronic'. A schematic view of the genomic region classification is featured in Figure 5.1.

5.3 EXPERIMENTAL DATA

For examining the sources of NET-seq signal, we analysed two different variants of NET-seq: standard NET-seq and HiS-NET-seq. For the standard NET-seq, results for two lines are presented: HeLa S3 (two biological replicates) and K562 (two technical replicates). For the HiS-NET-seq, we analysed two biological replicates produced using K562 cell lines. Additionally, two

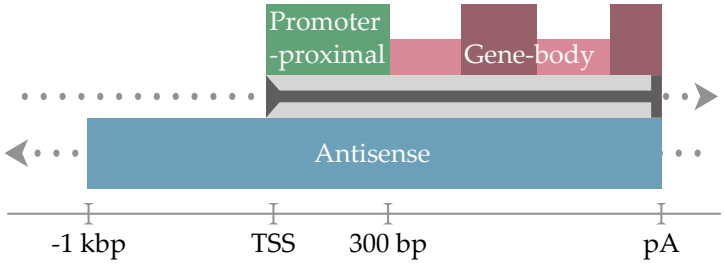

Figure 5.1: **Scheme of different genomic regions of interest.** The direction of transcription is indicated by arrowheads.

technical replicates of a standard NET-seq library prepared using HEK293T cells were examined to assess the robustness of the peak calling algorithm. The summary of the sample information is included in Table 5.1.

Table 5.1: Summary of sample information.

Sample name	Cell line	NET-seq type	Number of sequenced reads
standard HeLa Rep1	HeLa S3	standard	175,303,176
standard HeLa Rep2	HeLa S3	standard	768,583,061
standard K562	K562	standard	42,167,250
standard K562 reseq.	K562	standard	320,432,287
labeled K562 Rep1	K562	HiS-NET-seq	76,774,699
labeled K562 Rep2	K562	HiS-NET-seq	67,578,055
standard HEK293T Rep1	HEK293T	standard	109,078,738
standard HEK293T Rep2	HEK293T	standard	105,211,781

Due to their high sequencing depth, the standard NET-seq libraries obtained for HeLa S3 cells served as a base for characterization of the transcriptional pausing in human cell lines. Additionally, we re-analysed available NET-seq data for *Escherichia coli* [43] (GEO accession: GSM1367304), *Saccharomyces cerevisiae* [28] (GEO accessions: GSM1673641 and GSM1673642) and *Arabidopsis thaliana* [38] (GEO accessions: GSM3814845 and GSM3814846).

5.4 RESULTS

5.4.1 *Examining potential artifact positions*

We pre-processed the chosen NET-seq libraries as described in the Method section of this Chapter without masking any regions. Before detecting Pol II pausing sites, we examined the genomic regions that have previously been reported to generate NET-seq signal that does not reflect Pol II occupancy. Those regions can be divided into three categories:

- loci transcribed by other human polymerases,
- positions corresponding to 3′ ends of transient product of nascent RNA processing,
- positions corresponding to 3′ ends of the chromatin associated and nascent transcript associated species.

There are several reason to inspect the above mentioned regions. First of all, NET-seq signal accumulation detected in those locations can easily be mistaken by the Pol II pausing sites, and in turn lead to false biological conclusions about the phenomena. Second, the amount of signal in the above listed can serve as a quality check of the NET-seq library, and the relationship between the experimental parameters and abundance of those undesired RNA species can guide the improvements of the experimental procedures. Below, we discuss the categories of the NET-seq contaminants and examine the signal obtained in the corresponding regions.

Other human RNA polymerases

We expect to see the NET-seq signal over genes that are known to be transcribed by other human RNA polymerases, namely Pol I and Pol III, as the purification step in NET-seq does not include an additional selection of Pol II transcribed loci. Additionally, if the purification of the nuclei is not conducted properly, we can observe NET-seq signal over the mitochondrial chromosome, in the loci transcribed by the mitochondrial polymerase. To quantify the contribution of the nascent RNAs produced by RNA polymerases other than Pol II, we created a list of transcribed regions in human genome that are not synthesised by Pol II as described in the Method section. Then, we used deepTools package [61] to create heatmap summary plots of NET-seq signal in those regions.

Figure 5.2 shows the average NET-seq signal for rRNA genes transcribed by Pol I. In all libraries, we can see an accumulation of NET-seq signal at the 3′ end of rRNA genes. The signal coming from the Pol I transcription over gene bodies is less pronounced and it is visible only in the libraries prepared using HeLa S3 cell line. The majority of the regions with the highest signal intensity encodes 5S rRNA.

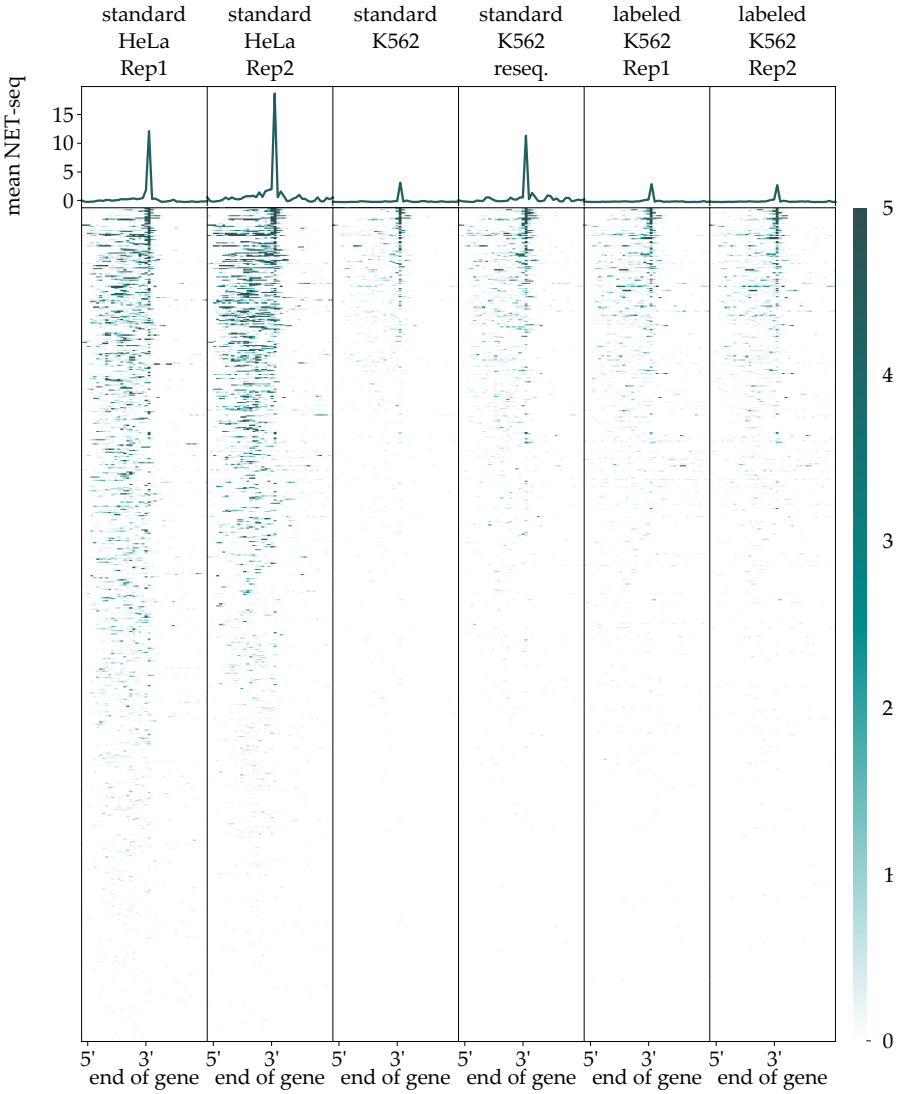

Figure 5.2: **NET-seq signal at rRNA genes visualized with deepTools.** Shown is the NET-seq signal at the region between 10 base pairs upstream of the 5′ end and 100 base pairs downstream of the 3′ end of the rRNA genes. DeepTools heatmaps are visualized for six NET-seq experiments conducted in two different cell lines (HeLa S3 and K562) and using two different NET-seq variants (standard and with labeling). Color-coded is the average NET-seq signal in 5 base pair long bins. Profiles presented above the heatmaps show the average NET-seq signal. The plots include only those genomic region, for which NET-seq signal (at least one read) was detected in at least one library (806 out of 1065 annotated rRNA genes).

Then, we visualized the average NET-seq signal for tRNA genes transcribed by Pol III in Figure 5.3. Similarly to the rRNA genes, we can see an accumulation of NET-seq signal at the 3′ end of tRNA genes in all libraries. The heatmaps show not only signal coming from the full-length tRNAs, but we can also observe the NET-seq signal over the whole length of tRNA genes coming from the transcribing Pol III.

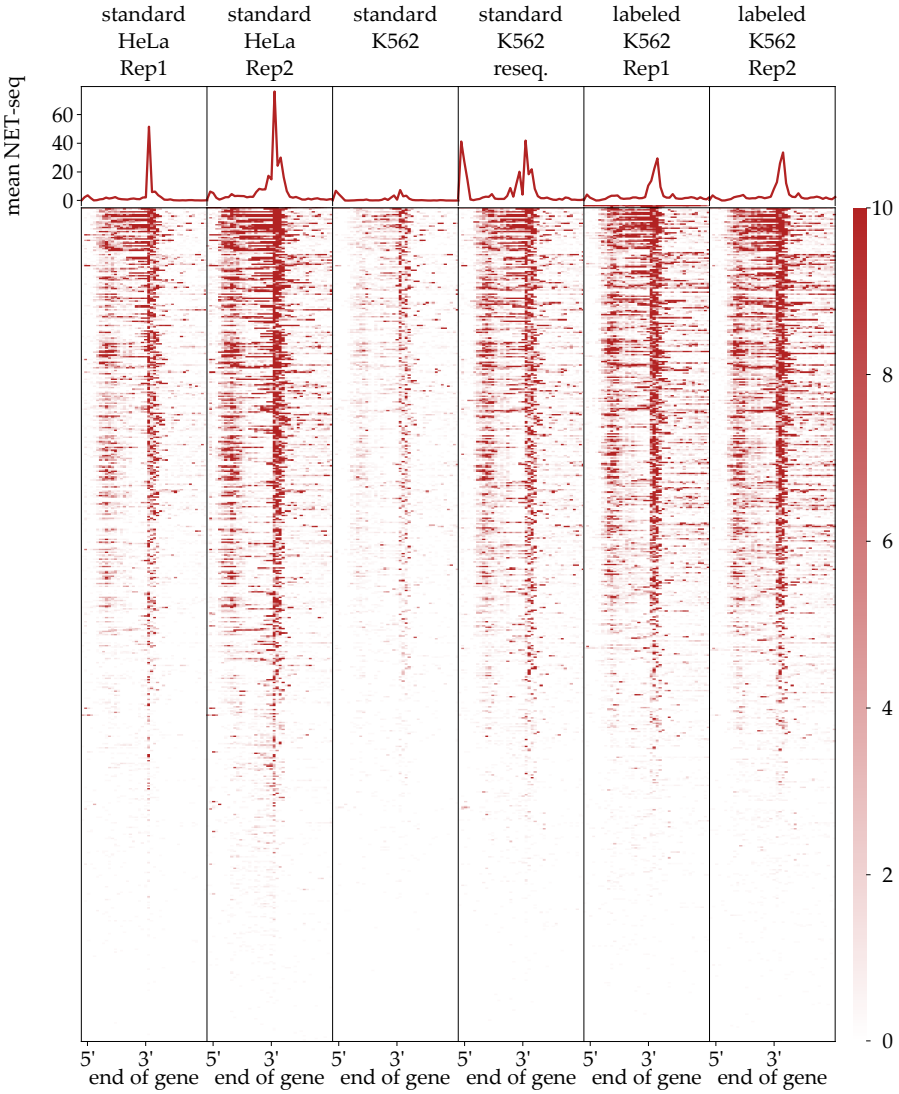

Figure 5.3: **NET-seq signal at tRNA genes visualized with deepTools.** Shown is the NET-seq signal at the region between 10 base pairs upstream of the 5′ end and 100 base pairs downstream of the 3′ end of the tRNA genes. DeepTools heatmaps are visualized for six NET-seq experiments conducted in two different cell lines (HeLa S3 and K562) and using two different NET-seq variants (standard and with labeling). Color-coded is the average NET-seq signal in 5 base pair long bins. Profiles presented above the heatmaps show the average NET-seq signal. The plots include only those genomic region, for which NET-seq signal (at least one read) was detected in at least one library (590 out of 649 annotated tRNA genes).

The full-length products of the transcription performed by the mitochondrial polymerase can also be seen in all of the analysed libraries, as Figure 5.4 shows. We can observe a very strong NET-seq at several mitochondrial genes, especially in the standard NET-seq library prepared using K562 cell line and sequenced at a high depth.

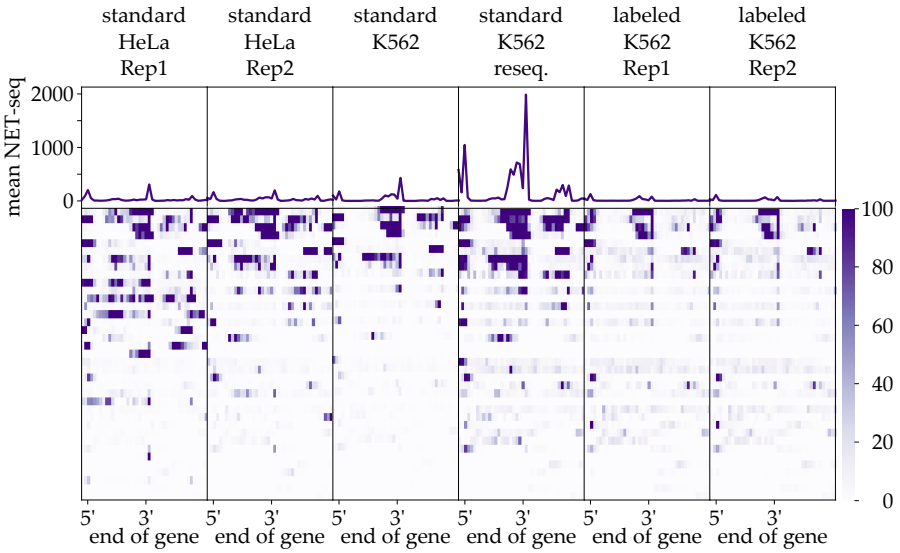

Figure 5.4: **NET-seq signal at mitochondrial genes visualized with deepTools.** Shown is the NET-seq signal at the region between 10 base pairs upstream of the 5′ end and 100 base pairs downstream of the 3′ end of the mitochondrial genes. DeepTools heatmaps are visualized for six NET-seq experiments conducted in two different cell lines (HeLa S3 and K562) and using two different NET-seq variants (standard and with labeling). Color-coded is the average NET-seq signal in 5 base pair long bins. Profiles presented above the heatmaps show the average NET-seq signal. The plots include only those genomic region, for which NET-seq signal (at least one read) was detected in at least one library (all of the annotated mitochondrial genes).

Transient product of the nascent RNA processing

In addition to nascent transcripts, NET-seq captures intermediates of the nascent RNA processing, due to their 3′ hydroxyl termini that allow them to enter the library. Those intermediates include transient product originating from the co-transcriptional RNA processes such as splicing and miRNA maturation. We assessed the contribution of the transient processing intermediates to the total NET-seq signal using heatmap summary plots generated with the deepTools package [61].

Figure 5.5 shows the average NET-seq signal for miRNA genes. We can see an accumulation of the NET-seq signal at the 3′ end of the mature miRNA. Additional pile-ups can be observe upstream and at the 5′ end and downstream of the 3′ end of miRNA. The peaks upstream of the 5′ and downstream of the 3′ end of the mature miRNA correspond to the ends of the pre-miRNA.

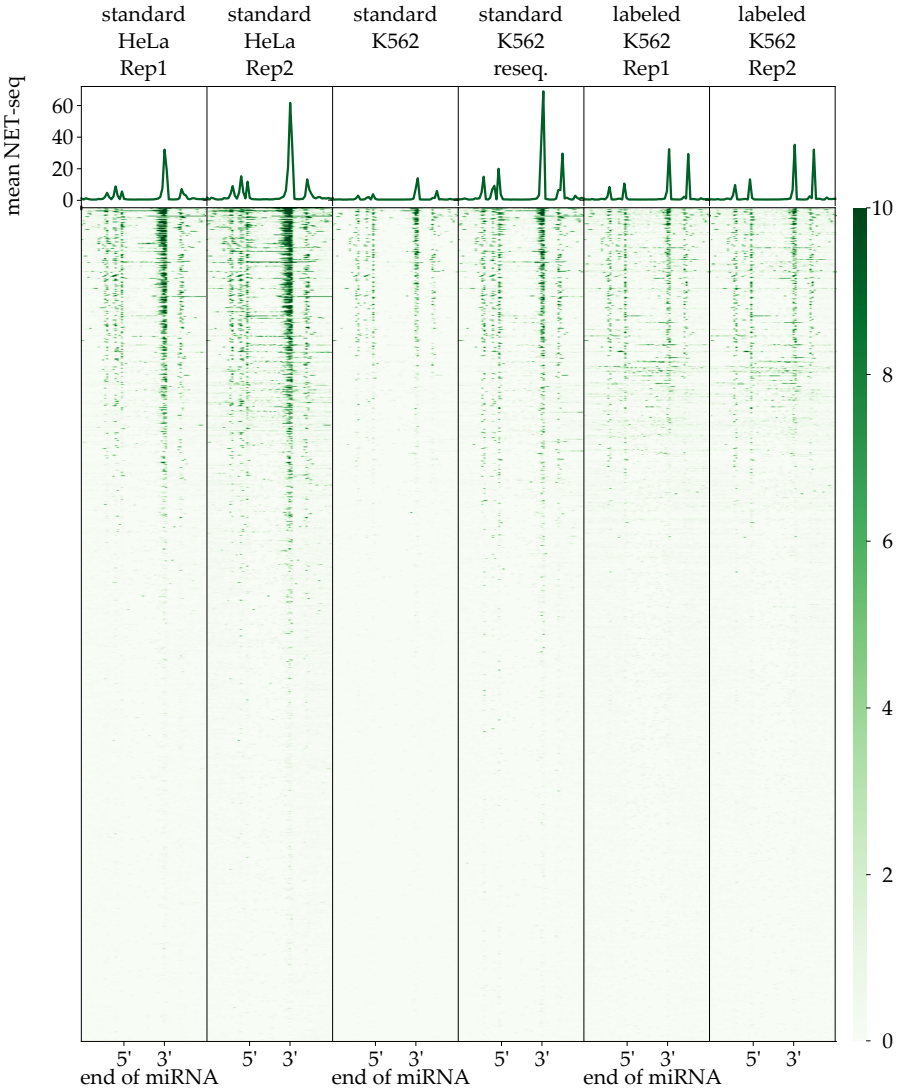

Figure 5.5: **NET-seq signal at miRNA genes visualized with deepTools.** Shown is the NET-seq signal at the region between 100 base pairs upstream of the 5′ end and 100 base pairs downstream of the 3′ end of the miRNA genes. DeepTools heatmaps are visualized for six NET-seq experiments conducted in two different cell lines (HeLa S3 and K562) and using two different NET-seq variants (standard and with labeling). Color-coded is the average NET-seq signal in 5 base pair long bins. Profiles presented above the heatmaps show the average NET-seq signal. The plots include only those genomic region, for which NET-seq signal (at least one read) was detected in at least one library (2129 annotated miRNA genes).

Another nascent RNA processing intermediate that is visible in NET-seq originate during the process of splicing. Figure 5.6 shows NET-seq signal at the internal exons that are retained during splicing. An accumulation of signal is visible at the 3′ ends of the exons and the 3′ ends of the upstream introns.

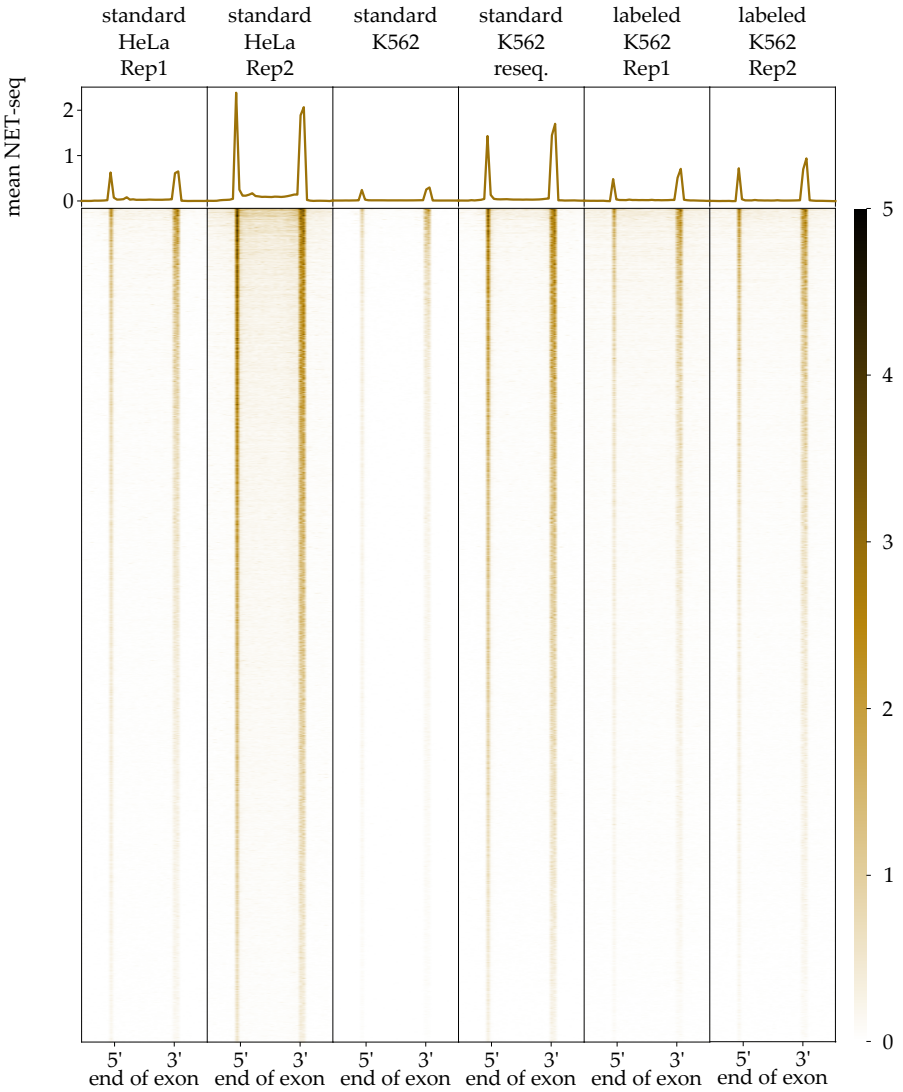

Figure 5.6: **NET-seq signal at exons visualized with deepTools.** Shown is the NET-seq signal at the region between 50 base pairs upstream of the 5′ end and 50 base pairs downstream of the 3′ end of the internal exons. DeepTools heatmaps are visualized for six NET-seq experiments conducted in two different cell lines (HeLa S3 and K562) and using two different NET-seq variants (standard and with labeling). Color-coded is the average NET-seq signal in 5 base pair long bins. Profiles presented above the heatmaps show the average NET-seq signal. The plots include only internal exons of the transcripts, which are expressed either in HeLa S3 or K562 cell line.

Non-nascent chromatin associated RNAs

As NET-seq purification selects all chromatin associated RNAs, in NET-seq genomic tracks show signal originating not only from nascent RNAs but also from all RNA species that form transiently stable complexes with the chromatin or nascent RNA themselves. The chromatin associated RNA species include RNA molecules executing their functions in the proximity of chromatin and nascent RNAs, such as snRNAs that are involved in splicing or snoRNAs. To quantify the contribution of the non-nascent, chromatin-associated RNAs we used deepTools package [61] to create heatmap summary plots.

Figure 5.7 shows the average NET-seq signal at snoRNA genes. In all libraries, we can see an accumulation of NET-seq signal at the 3' end of snoRNA genes that correspond to the full length transcripts. There is a subpopulation of loci, for which the signal is extremely high, reaching an average of over 80000 reads per position in the standard NET-seq library prepared for K562 cell line and sequenced with a high sequencing depth. All 40 genes in the first cluster belong to the C/D box snoRNAs.

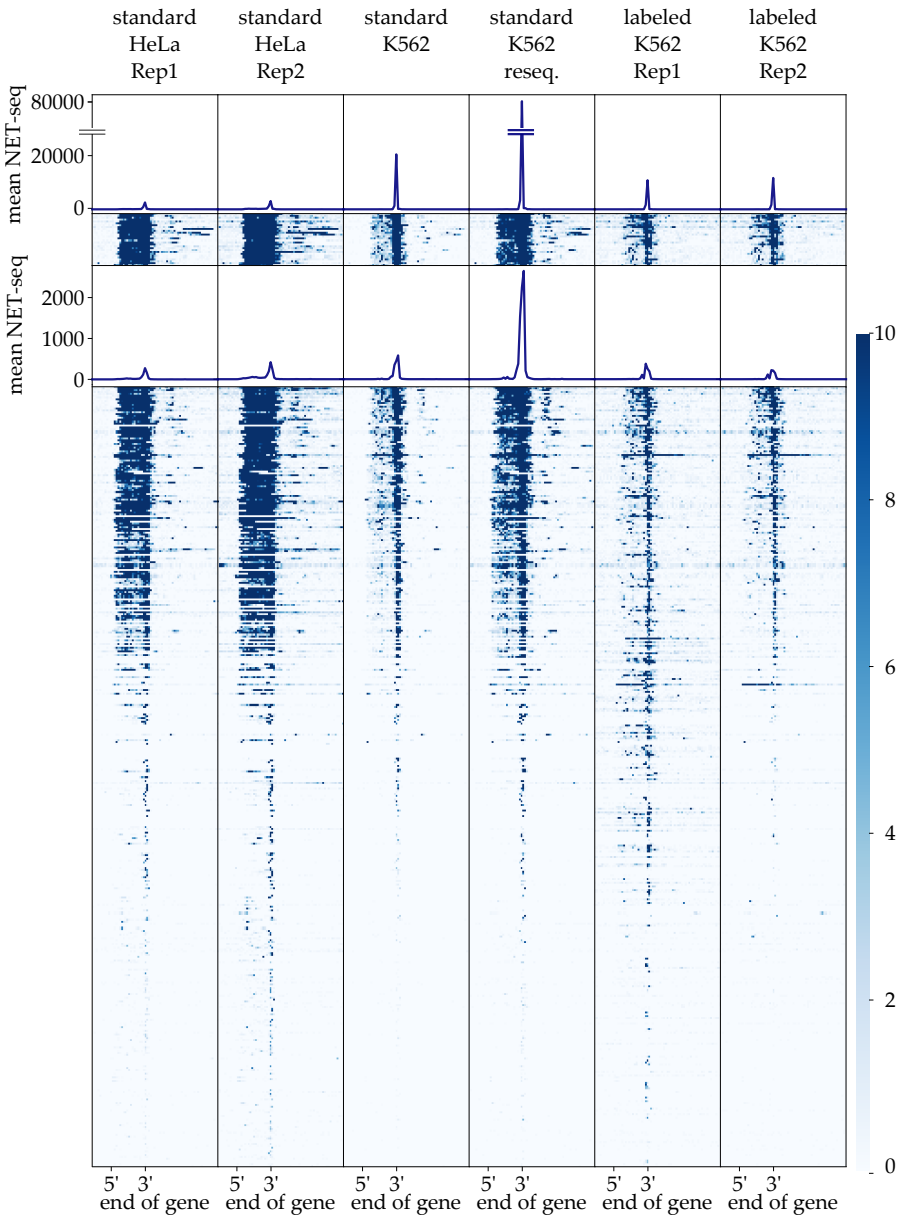

Figure 5.7: **NET-seq signal at snoRNA genes visualized with deepTools.** Shown is the NET-seq signal at the region between 100 base pairs upstream of the 5′ end and 200 base pairs downstream of the 3′ end of the snoRNA genes. DeepTools heatmaps are visualized for six NET-seq experiments conducted in two different cell lines (HeLa S3 and K562) and using two different NET-seq variants (standard and with labeling). Color-coded is the average NET-seq signal in 5 base pair long bins. Profiles presented above the heatmaps show the average NET-seq signal. The plots include only those genomic region, for which NET-seq signal (at least one read) was detected in at least one library (558 snoRNA genes).

Another category of RNA species that form transiently stable complexes with the nascent RNAs consists of snRNA genes. Figure 5.8 shows the average NET-seq signal for snRNA genes. Here, we can see an accumulation

of NET-seq signal at the 3′ end of snRNA genes that correspond to the full length transcripts. The signal is especially pronounced in standard K562 libraries and it has higher average than the snRNA signal in the data set with the highest sequencing depth (standard HeLa Rep2).

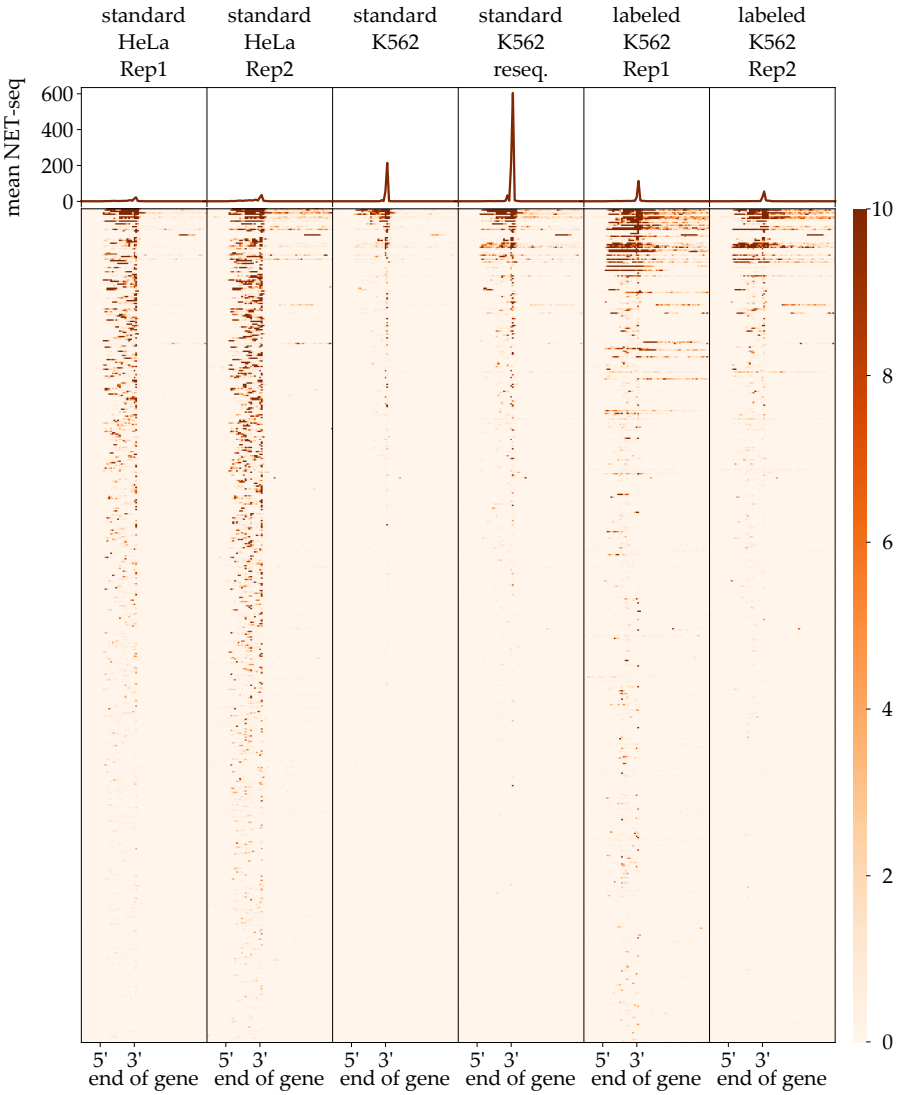

Figure 5.8: **NET-seq signal at snRNA genes visualized with deepTools.** Shown is the NET-seq signal at the region between 100 base pairs upstream of the 5′ end and 200 base pairs downstream of the 3′ end of the snRNA genes. DeepTools heatmaps are visualized for six NET-seq experiments conducted in two different cell lines (HeLa S3 and K562) and using two different NET-seq variants (standard and with labeling). Color-coded is the average NET-seq signal in 5 base pair long bins. Profiles presented above the heatmaps show the average NET-seq signal. The plots include only those genomic region, for which NET-seq signal (at least one read) was detected in at least one library (608 snRNA genes).

This analysis provided a justification for black-listing the genomic regions generating NET-seq signal that does not reflect Pol II occupancy. Before

conducting any further analyses, we masked the above mentioned regions to avoid misinterpreting local enrichments found there as Pol II pausing sites.

5.4.2 *Parameters affecting the number of detected pausing sites*

The next step after examining regions where NET-seq signal does not reflect Pol II occupancy was defining the sites that should be detected and designing a peak calling algorithm that is suitable for sparse Pol II occupancy profiles. We examined two approaches (see Section 5.2.3) for calculating the p-value of each candidate peak, which were derived for different definitions of a pausing site based on the NET-seq signal. The first definition characterizes a pausing site as a local maximum with a significantly higher value than the expected value given the local NET-seq signal. To calculate the p-values of candidate peaks, we used the Poisson distribution with the mean λ equal to the average number of reads per position within the window centred at the potential pausing site position. The second definition characterizes a pausing site as a local maximum with a significantly higher value than the expected value of the maximum given the local NET-seq signal. For this definition, we determine the statistical significance of candidate peaks based on the distribution of expected values of the maximum given the number of reads within the window centred at the potential pausing site position.

To better understand the differences between the proposed approaches, we first visualized the distributions used for statistical testing. For a selected total number of reads in the window M and for a selected number of positions l with non-zero signal intensity, we compared the empirical distribution of the maxima for pairs (M,l) and Poisson distribution with the mean λ equal $\frac{M}{l}$. Figure 5.9 shows a comparison of the distributions for the total number of reads M equal 3376 and number of positions l equal 115. As the empirically obtained distribution models the expected maximum number of reads, it is centered in the upper tail of the corresponding Poisson distribution.

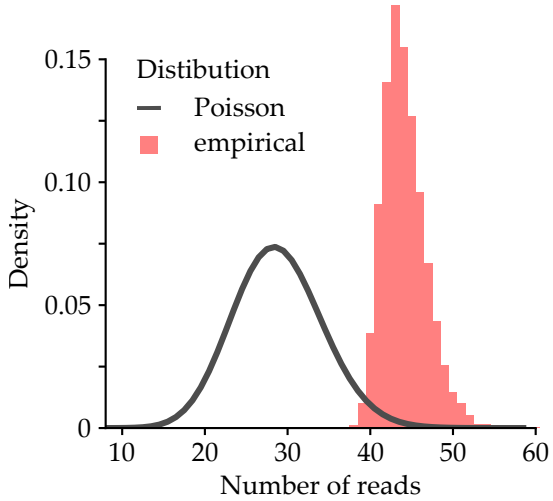

Figure 5.9: **Distributions of the expected number of reads (modeled with Poisson distribution; in grey) and the expected maximum number of reads (empirically derived; in red).** The distributions were calculated for the total number of reads $M = 3376$ and number of non-zero positions $l = 115$.

To further compare the proposed parametrical and non-parametrical approach, we examined the thresholds to call a peak significant using a window of 200 nucleotide width. Both of the approaches determine the statistical significance of a peak based on three values: the peak intensity x_i, total number of reads in the window M and number of positions l with non-zero signal intensity. We calculated the 95th percentile of the Poisson distribution of the mean λ equal $\frac{M}{l}$ and the empirical distribution of the maxima for pairs (M,l), with the total number of reads M ranging from 6 to 1500 and the number of positions l with non-zero signal intensity between 2 and 200. Figure 5.10 shows the difference between the 95th percentile of the empirically derived distribution of the maxima and the Poisson distribution. The thresholds are slightly higher for the empirical distribution of the maxima with the exception for regions with extremely small number of non-zero positions (less than 5 out of 200).

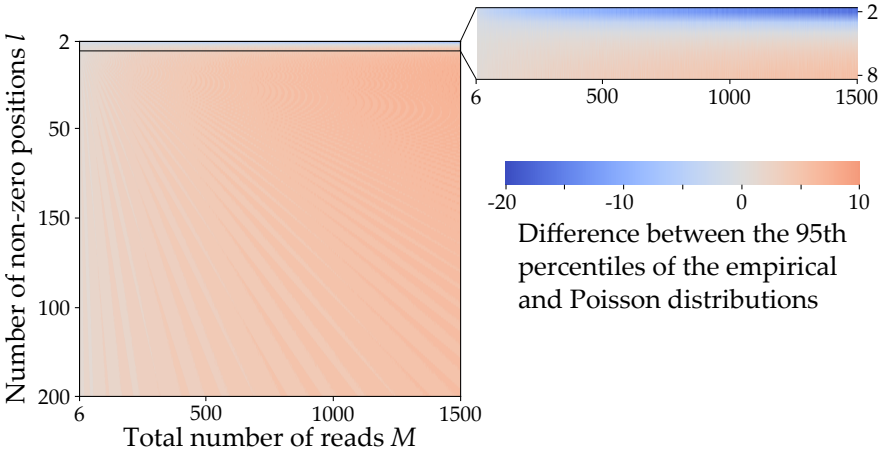

Figure 5.10: **Heatmap showing the difference between the 95th percentile of the empirical distribution of the maxima and Poisson distribution.** The difference depends on the total number of reads in the window M (x-axis) and number of positions l with non-zero signal intensity (y-axis).

Determining peak significance using empirical distribution is computationally more costly than using Poisson distribution due to the resampling performed to generate the empirical distribution. For data sets with high signal intensity, e.g. where the PCR duplicates were not removed, using the resampling approach might lead to long computing time. Therefore, we wanted to examined whether setting a higher p-value cut-off for the parametrical approach yields similar results to the non-parametrical approach with a lower cut-off, especially for the high total number of reads. We calculated the difference between the 95th percentile of the empirical distribution of the maxima and the 99.5th percentile of the Poisson distribution with the mean λ equal $\frac{M}{l}$. Figure 5.11 shows that the difference between the significance threshold is close to 0, except for regions with very small number of non-zero positions (less than 10 out of 200). This finding indicates that for data sets with high signal intensity and low sparsity, the parametrical approach is a solution that yields results similar to the non-parametrical approach.

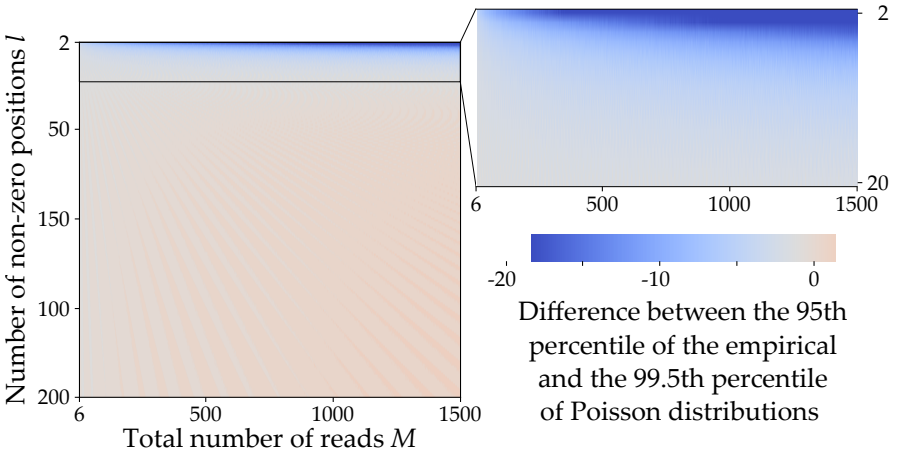

Figure 5.11: **Heatmap showing the difference between the 95th percentile of the empirical distribution of the maxima and 99.5th percentile of the Poisson distribution.** The difference depends on the total number of reads in the window M (x-axis) and number of positions l with non-zero signal intensity (y-axis).

Next, we examined how the parameter of the testing method, namely the window width, affects the number of pausing sites detected. We detected the peaks in NET-seq library available for HeLa S3 cells (standard HeLa Rep2). Then, we performed the evaluation of the statistical significance of the peaks with the non-parametrical testing using five different widths of the window equal to: 20, 50, 100, 200, and 500 nucleotides. Figure 5.12 shows that the number of called peaks depends on the size of the window, with the larger window yielding more peaks. We compared the peaks detected using 50- and 200- nucleotide window, focusing on the differences between peaks detected with only one or both of the windows. Peaks detected with only one window size have significantly lower median then the peaks called significant with both window widths.

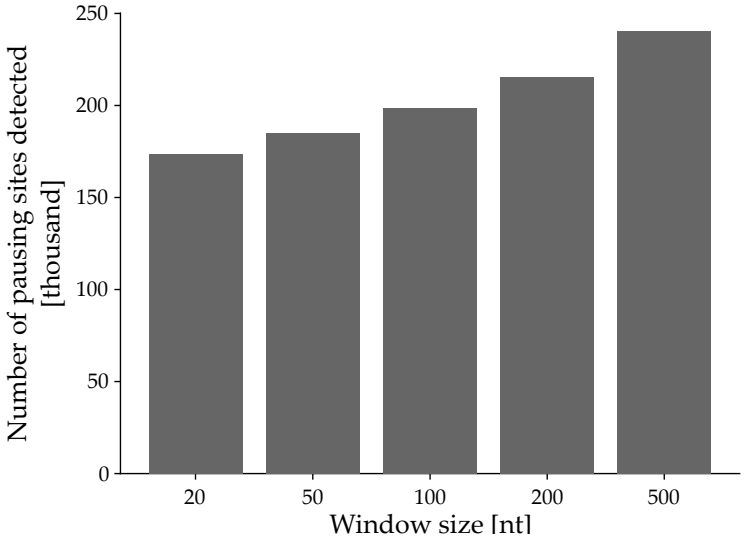

Figure 5.12: **Relationship between the window width and number of called peaks.**

To learn about the sensitivity of peak calling, we randomly downsampled a high-coverage NET-seq data set available for HeLa S3 cells (standard HeLa Rep2) to simulate lower sequencing depths of NET-seq libraries. For each of the simulated sequencing depths, we performed a peak detection and evaluation with the non-parametrical testing, setting the window width to 200 nucleotides. As Figure 5.13 shows, the number of called peaks correlated with the sequencing depth of NET-seq data and dropped proportionally to the decreasing number of sequencing reads. Since lowering the sequencing coverage unavoidably also reduces the library complexity, defined by the number of unique DNA fragments present in a given library, we cannot rule out that the drop in peak identification was partially caused by the decrease in library complexity.

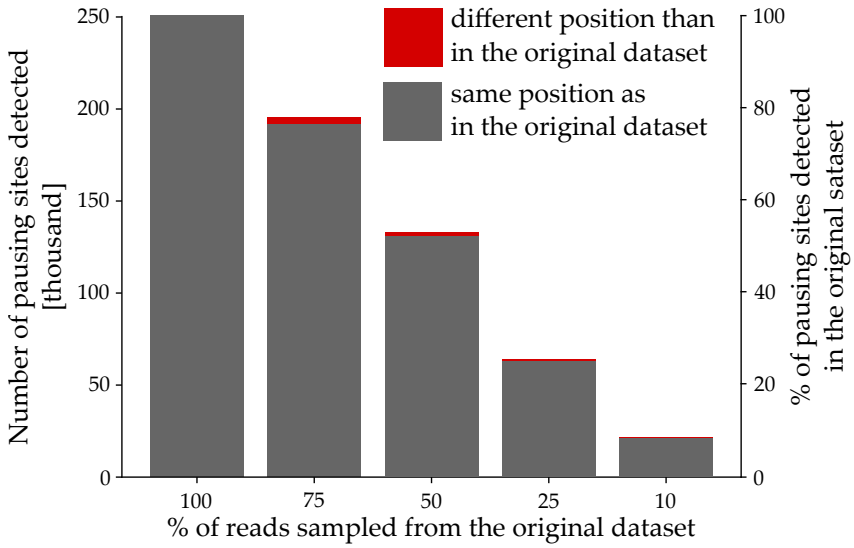

Figure 5.13: **Relationship between the sequencing depth of NET-seq data and number of called peaks.** A reduction in the sequencing depth was obtained by random subsampling of raw reads.

As a final part of the assessment, we examined the robustness of the peak calling by applying the algorithm to two technical NET-seq replicate data sets prepared using HEK293T cells. As the two data sets had an almost identical sequencing coverage, it allowed us to determine the reliability of the peak calling irrespectively of the sequencing depth and the biological variation. Figure 5.14 shows that vast majority of detected peaks was called at the same nucleotide position in both replicates. To assess the significance of the overlap between the sets of detected peaks, we performed Fisher exact test obtaining p-value smaller than 2.2^{-308}. This finding indicates that pause detection algorithm calls high-resolution peaks reliably and that the effect of technical variation is minimal.

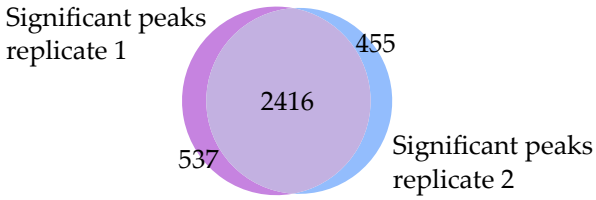

Figure 5.14: **Venn diagram showing the overlap of significant peaks detected for technical NET-seq replicates obtained for human HEK293T cells.**

5.4.3 *Characterization of Pol II pausing sites in human cell lines*

After examining the parameters affecting the detection of the peaks, we applied the peak detection algorithm to two biological replicates of NET-seq obtained for the HeLa S3 cell line (standard HeLa Rep1 and Rep2). Downstream analyses were perform using only high-confidence pausing sites that were detected in both biological replicates at the same nucleotide position. Based on their location, pausing sites were classified into one of four major categories: promoter-proximal, gene-body, antisense or intergenic. Figure 5.15 shows the distribution of pausing sites over those genomic regions. About 75% and 20.6% of occupancy peaks were intragenic (promoter-proximal, gene-body and convergent antisense) or intergenic (including divergent antisense), respectively. We found that 17.3% of peaks were located in the promoter-proximal region of genes corresponding to promoter-proximal pauses. Notably, we found that the majority of Pol II pausing sites occurred outside of promoter-proximal regions, mainly throughout the gene-body. About 31% and 69% of the gene-body pauses occurred over exons and introns, respectively.

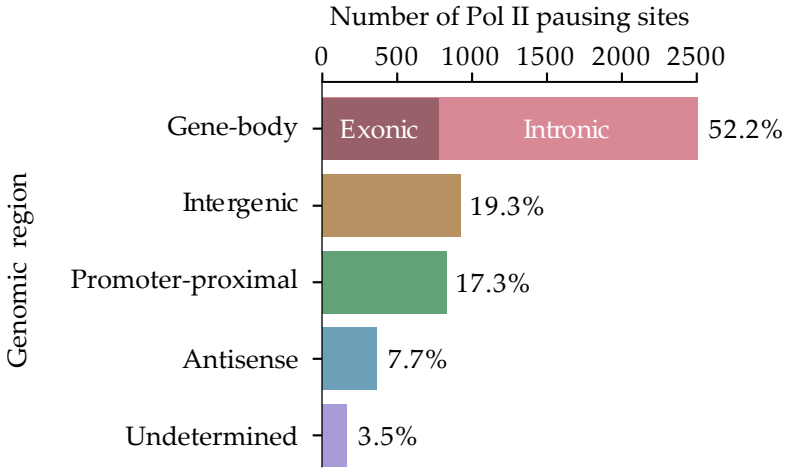

Figure 5.15: **Pausing site distribution over different genomic regions in the HeLa S3 cell line.**

We first investigated whether distinct DNA sequences were enriched in a close proximity to pausing sites. A main advantage of NET-seq over ChIP-seq data is the high spatial resolution, which allowed us to precisely extract the DNA sequences underlying Pol II pause positions and to calculate the nucleotide frequencies in relation to Pol II position. To derive the enrich-

ment logos, we first calculated the background nucleotide distribution for promoter-proximal and gene-body regions. Then, the extracted pausing sequences and the background frequencies were used to create enrichment logos with Logolas [18]. The enrichment logos with the positions aligned to the RNA-DNA hybrid are shown in Figures 5.16B and 5.16C. The motif obtained for promoter-proximal pausing sites consists of two parts: the G_{-10} at the upstream fork junction of the RNA-DNA hybrid and the $Y_{-2}G_{-1}Y_{+1}$, where Y is thymine or cytosine, at the region spanning the active site of Pol II and the downstream fork junction of the RNA-DNA hybrid. We did not recover a clear motif linked to Pol II pausing at gene-body.

Next, we looked into the DNA sequence signatures enriched at pausing sites detected in the HiS-NET-seq. We were interested if the additional enrichment procedure results in a different set of pausing sites detected. The obtained enrichment logos with the positions aligned to the RNA-DNA hybrid are shown in Figures 5.16D and 5.16E.

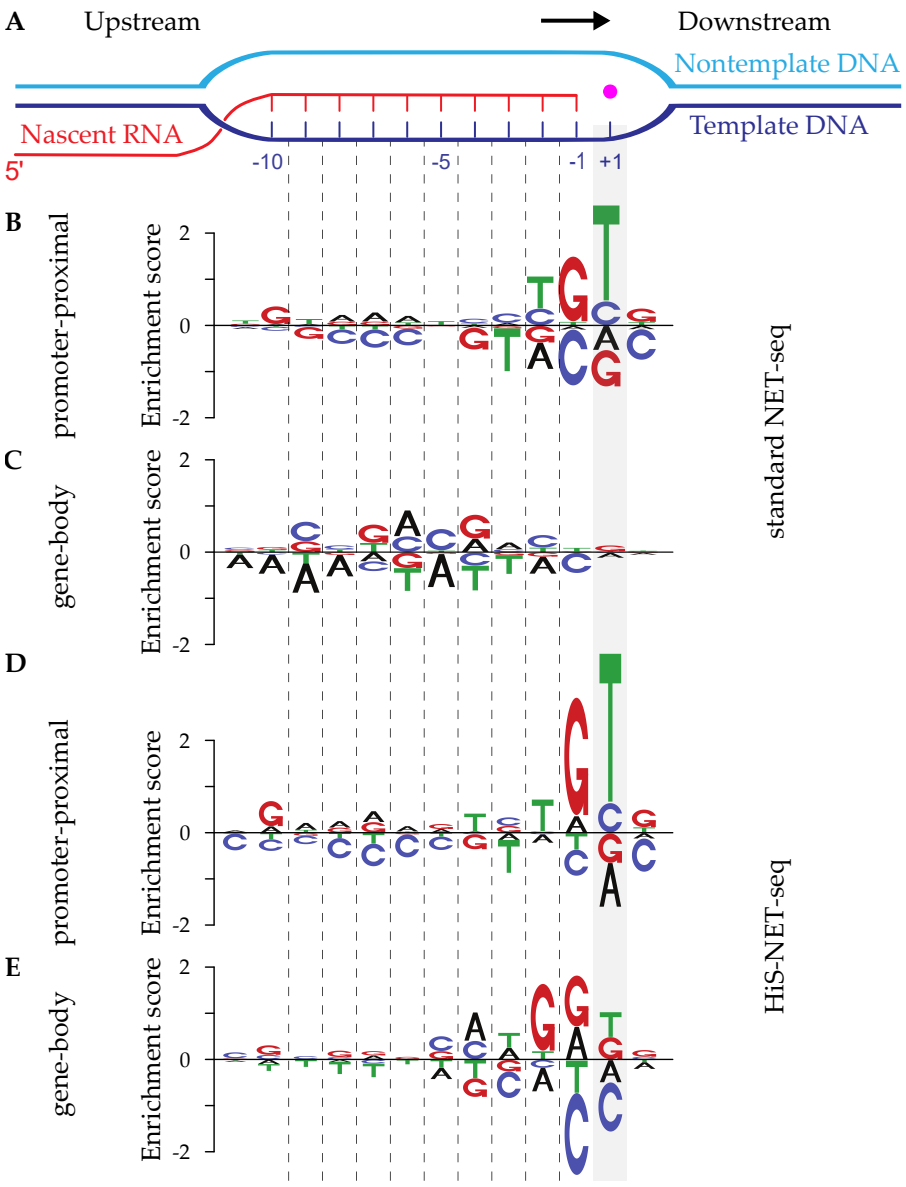

Figure 5.16: **Pausing motif discovery and sequence analysis in human cell lines. (A)** Schematic view of the transcription bubble. The pink dot corresponds to a Mg^{2+} ion marking the active site of Pol II. -1 refers to the last nucleotide of the nascent RNA. +1 indicates the position in the DNA template where the next incoming NTP binds. The direction of transcription is indicated by a black arrow. This model is based on recent evidence from structural studies indicating that the RNA-DNA hybrid that spans the active site of the mammalian Pol II elongation complex is 9–10 bp long. **(BC)** Enrichment logos for promoter-proximal pause **(B)** and gene-body pause sites **(C)** retrieved using standard NET-seq protocol. **(DE)** Enrichment logos for promoter-proximal pause **(D)** and gene-body pause sites **(E)** retrieved using HiS-NET-seq protocol.

The motifs retrieved for promoter-proximal pausing sites showed the same enrichment profiles for both the standard and the high sensitivity variants of NET-seq. In difference to standard NET-seq, where no strong enrichments or depletions were found, we did recover a sequence linked to Pol II pausing at gene-body using HiS-NET-seq. The pause signature obtained for the gene-body differed from the motif retrieved for promoter-proximal pausing but pausing in both of the regions showed an enrichment of guanines and thymines at the positions spanning the active site of Pol II and the downstream fork junction of the RNA-DNA hybrid. Additionally, the pausing signatures for gene-body pausing different between the two variants of NET-seq protocols.

5.4.4 *Pol II pausing detection in NET-seq data from various organisms*

We next asked whether the sequence underlying RNA polymerase pausing sites shared similarities between species. To address this question, we reanalyzed available NET-seq data for bacteria (*Escherichia coli*), budding yeast (*Saccharomyces cerevisiae*) and plants (*Arabidopsis thaliana*). First, we applied the pausing detection algorithm to find the pausing sites for those organisms. In difference to the analyses performed for human cell lines, we created the enrichment logos for all sites in the intragenic region without further categorizing them based on their genomic position, since clear evidence for promoter-proximal pausing has not yet emerged in these organisms. The retrieved sequence motifs underlying pausing in *E. coli*, *S. cerevisiae* and *A. thaliana* are presented in Figure 5.17, together with the motif obtained for the promoter-proximal pausing in human cell lines. All highly enriched nucleotides overlapped with the region of the downstream fork junction of the RNA-DNA hybrid. The sequence motif obtained for bacteria was highly similar to the motif underlying promoter-proximal pauses in human cells, with the bacterial pausing motif being shifted upstream by a single nucleotide as compared to the human consensus pausing sequence. Additionally, for all species investigated here the consensus pausing motif contained a TG dinucleotide at the active site region of RNA polymerase.

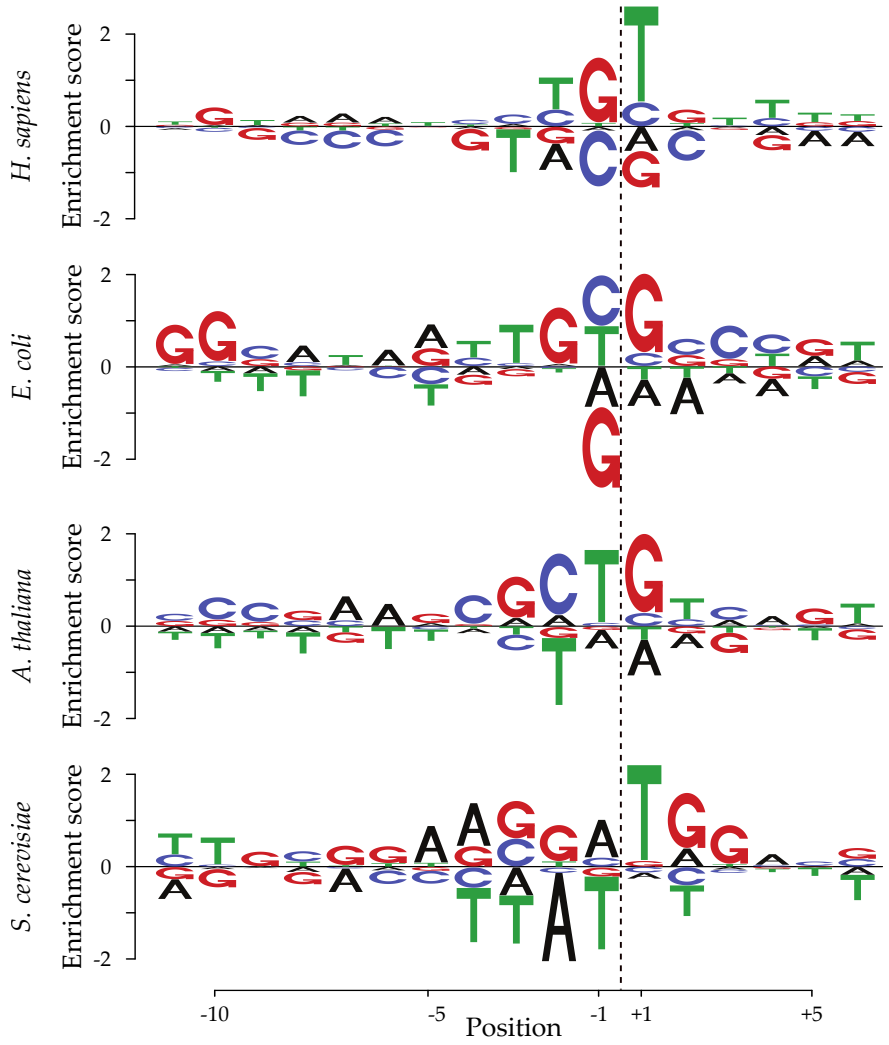

Figure 5.17: **Comparison of DNA sequences at pausing sites in model organisms.** Enrichment logos for *H. sapiens* (pauses in promoter-proximal region), *E. coli* (pauses within gene), *S. cerevisiae* (pauses within gene) and *A. thaliana* (pauses within gene).

5.5 DISCUSSION

Here, we proposed refinements to the NET-seq data processing and developed an algorithm for robust peak detection from single-nucleotide resolution profiling data to investigate transcriptional pausing. We examined genomic regions generating NET-seq signal that does not reflect Pol II occupancy and originates from processes other than transcription conducted by Pol II. To show that artifacts originating during library preparation can have a strong impact on data interpretation, we visualized the average NET-seq signal at genes transcribed by other polymerases, at genes encoding chromatin associated RNAs and at splice sites. We observed an accumulation of NET-seq signal at 3′ ends of genes transcribed by other polymerases and encoding chromatin associated RNAs. The NET-seq signal accumulation there can easily be mistaken for a Pol II pausing site in the intronic region of the host gene, as the rRNA, tRNA, snRNA and snoRNA are relatively short and

some of them are located in the intronic regions of protein-coding and long non-coding genes. Additionally, we found a group of C/D box snoRNA genes with extremely high signal intensity at their 3′ ends.

Moreover, we identified a NET-seq signal accumulation in the miRNA genes. The metagene plots for miRNA genes showed strong NET-seq signal at the 3′ end of the mature miRNA and a weaker signal accumulation at three additional locations: the 5′ end of the mature miRNA, the 5′ end of pre-miRNA, and the 3′ end of the pre-miRNA. Similarily to the peaks obtained at splicing intermediates, the NET-seq peak at the ends of the pre-miRNA might correspond to the first step of the miRNA maturation, as it happens co-transcriptionally in the nucleus. The NET-seq peak at the 3′ ends of the mature miRNA is likely to originate from the mature miRNAs serving their function in the proximity of the chromatin. Likewise with rRNA, tRNA, snRNA, and snoRNA genes, miRNA genes are frequently located in the intronic regions. This analysis provides a justification for black-listing and masking the genomic regions generating NET-seq signal that does not reflect Pol II occupancy.

We developed a resampling-based peak caller suitable for sparse occupancy tracks. Using technical replicates of NET-seq data, we showed that the majority of detected peaks was called at the same nucleotide position in both replicates, indicating that the algorithm calls high-resolution peaks reliably and that the effect of technical variation is minimal. We examined the impact of the window size, which is a parameter of the peak caller, on peak detection. The number of peaks detected depends on the window size, with the larger windows yielding more peaks. However, peaks detected with only one window size have significantly lower median then the peaks robustly called with different sizes of windows. Therefore, the peak detection can be performed multiple times with different window sizes and the intersection of the results from different runs provides a set of pausing sites obtained with a more stringent approach.

Additionally, we examined the relationship between the sequencing depth and the number of detected peaks. The number of called peaks correlates with the sequencing depth of NET-seq data. Since lowering the sequencing coverage unavoidably also reduces the library complexity, we cannot rule out that the drop in peak identification is partially caused by the decrease in library complexity. Finally, we compared the proposed non-parametrical, resampling-based peak caller to the parametrical approach using Poisson distribution for testing. For the same significance threshold, the non-parametrical approach requires higher peak intensity to call the peak significant for most of the local read densities, where local read density is defined by the total number of reads in the window M and the number of positions l with non-zero signal intensity. However, for regions with extremely small number of non-zero positions, testing using Poisson distribution provides more stringent results. Additionally, as the resampling can be computationally costly, we examined whether it is possible to find a pair of p-value cut-offs for which the parametrical and non-parametrical approach yield similar results, especially for the high local density of reads. We showed that setting the p-values to 0.05 and 0.005 for the non-parametrical and the parametrical approach respectively allows calling peaks of the same intensities for low signal sparsity. Therefore, we recommend the usage of the parametrical approach as an alternative to the resampling approach, especially in case of the data sets with high signal intensity, e.g. where the PCR duplicates were not removed.

We applied the proposed peak detection algorithm to investigate the Pol II pausing sites detected in NET-seq. The transcriptional pausing landscape of Pol II in human cells proved to be more diverse than originally anticipated. The majority of detected Pol II pausing sites are located outside of promoter-proximal gene regions, with a large fraction of these pauses distributed along the gene-body. A large set of promoter-proximal and gene-body pauses occur non-randomly at the same genomic nucleotide position in biological replicate measurements. Moreover, pausing is not restricted to the sense direction, but is also prevalent throughout antisense transcription, raising a question whether antisense pausing interferes with the sense transcription.

We investigated whether distinct DNA motifs are enriched at or in close proximity to pausing sites. The high spatial resolution of NET-seq data allowed us to precisely extract the DNA sequences underlying Pol II pause positions. We derived sequence signatures in relation to Pol II position for promoter-proximal and gene body pausing, taking into account the differences in the nucleotide compositions in those regions. The uncovered sequence motif that underlies promoter-proximal pausing in human cells differs from the motifs observed in previous studies. Although the motif position overlapping with the downstream fork junction of the RNA-DNA hybrid is consistent with recent studies, the nucleotide sequence differs strongly from sequence elements that have been implicated in promoter-proximal pausing [26, 73]. The possible explanations for this difference are the high spatial resolution of pause site detection that allowed us to precisely extract the underlying DNA sequence and the sequence context based normalization. The latter was critical to minimize sequence biases originating from the high GC content in promoter-proximal gene regions [36]. Interestingly, the promoter-proximal pausing motif shows similarities to the consensus sequence of the following core promoter elements located downstream of the TSS: the downstream core promoter element (DPE) [8], the downstream core element (DCE) [45, 46] and the recently uncovered human DPR core promoter element [70] (see Figure S1). The region where these core promoter elements are located strongly overlap with the region where promoter-proximal pausing of Pol II usually occurs. These findings extend previous observations that have linked core promoter elements with transcriptional pausing in *Drosophila* [31, 42, 63].

Whereas a clear sequence motif was uncovered for pauses in the promoter-proximal region, no motif was retrieved for gene-body pauses detected using the standard NET-seq library preparation. Conversely, an enrichment of guanines and thymines at the downstream fork junction of the RNA-DNA hybrid was uncovered for the gene-body pausing sites detected in HiS-NET-seq. This observation indicates that the sets of pausing sites called in the different NET-seq variants exhibit different characteristics and they might be evoked by different mechanisms. The set of the sites detected in the standard NET-seq libraries might be more homogeneous and therefore no clear motif was observed. Another possible explanation is that not all of the peaks detected in the gene-body region in the standard NET-seq libraries corresponds to the Pol II pausing sites. The DNA signature might be than masked by the unspecific peaks and could be retrieved only thanks to the additional enrichment step performed in the HiS-NET-seq library. Last explanation is that some of the pausing sites detected in the HiS-NET-seq library are caused by the addition of the 4-thiouridine. The observed relative enrichment of the adenine at the -1 position of the template DNA might

arise due to the position where polymerase was unable to continue the transcription upon 4-thiouridine incorporation.

Widespread transcriptional pausing of Pol II in human cells has apparent similarities to RNA polymerase pausing in bacteria. Similarly to bacteria [16], pausing in human cells occurs throughout the transcribed region and is not restricted to promoter-proximal gene regions. Moreover, the sequence motif $G_{-10}Y_{-2}G_{-1}Y_{+1}$, where Y denotes cytosine or thymine, uncovered for Pol II promoter-proximal pausing bears a striking resemblance to the bacterial pausing motif $G_{-10}Y_{-3}G_{-2}Y_{-1}G_{+1}$ with one main difference. The $Y_{-2}G_{-1}Y_{+1}$ portion of the Pol II promoter-proximal pausing motif spanning the active site is shifted downstream by a single nucleotide as compared to the bacterial consensus pause sequence. A potential explanation for this difference can be found in the oscillating behavior of the elongation complex oscillating in a thermal equilibrium by one nucleotide position between pre- and post-translocated states [5]. It can be that human Pol II and bacterial RNA polymerase were preferentially captured in the post- or pre-translocated state, respectively. The similarities of transcriptional pausing in human and bacterial cells are also consistent with the similar 3D architecture of the active site between Pol II and bacterial RNA polymerase, and with the conserved catalytic mechanism [17, 67]. These similarities point toward a possible conservation of the sequence-dependent transcriptional pausing mechanism.

6

INVESTIGATING THE CAUSES OF POL II PAUSING USING INTERPRETABLE MACHINE LEARNING

In this chapter, we further investigate pausing sites detected in NET-seq, focusing on potential causes of pausing in different genomic regions. We go beyond simple nucleotide enrichments visualized by logos, more specifically examining sequence characteristics of the pausing sites. We build classifiers that discriminate between pausing- and non- pausing sites using the sequence-dependent features of their locations. We examine the features that are important for the classification, gaining insight into potential mechanisms of Pol II pausing. Finally, we compare the results obtained for different model organisms of various biological complexity to find if the pausing locations share similarities between species. The code of the modeling pipeline is available on GitHub: https://github.molgen.mpg.de/gajos/ClassifierPausing.

6.1 MOTIVATION

Many transcription factors, DNA and chromatin characteristics have been reported to be implicated in transcriptional pausing (see Section 2.3). However, it is unclear, what are the relative contributions of these factors on the observed Pol II pausing. The elements involved in pausing are usually described and analysed independently, and the effects evoked by other regulators are not taken into consideration and discussed. A comprehensive assessment of the relative contribution of genetic factors to transcriptional pausing could further our understanding of the pausing mechanism, especially if performed in an unbiased and quantitative manner.

Moreover, most of the pausing factors and elements were identified by either describing genes with a high fraction of polymerases stuck in the promoter-proximal region or comparing genes exhibiting strong and weak pausing phenotypes. Such analyses inform us about global attributes of the paused genes such as promoter elements but often fail to capture characteristics of the position at which Pol II pauses. Our knowledge about the features typical for the single-nucleotide pausing is limited to the average distance from the transcription start site and sequence logos obtained for pausing sites. A better understanding of the properties differentiating positions where polymerase pauses from those where it can progress without interruption could help us decipher the role of the local DNA elements in regulating or invoking pausing.

An interpretable machine learning classifier able to discriminate the pausing and non-pausing sites based on only DNA sequence characteristics can be the first step toward closing those knowledge gaps. The features identified by the model as important for distinguishing pausing sites are promising candidates for perturbation experiments that might show the causal relationship between the selected features and transcriptional pausing. If the classification is performed separately in different genomic regions (e.g. such as a promoter-proximal region and gene-body region), the result can help us answer the question of whether pauses in distinct genomic regions are driven by the same factors and mechanisms. Finally, we can compare the

features identified as important for distinguishing pausing sites to find out whether the sites prone to pausing share similarities between species.

6.2 METHODS

6.2.1 *Creating Training, Test and Validation Sets*

We defined high-confidence pausing sites as pausing sites that were detected in all biological replicates at the same nucleotide position. Only high-confidence pausing sites were used to create a set of pausing sites (positive set). A set of non-pausing sites (negative set) was generated by sampling random positions at which pausing does not occur. To avoid creating artificial differences between both sets, a non-pausing site was sampled from the region [x, x+20] nucleotides downstream or upstream of a pausing location, where distance x depends on the region of the pausing site (50 for promoter-proximal pauses, 300 for pauses in other regions). Pausing and non-pausing sites are subsequently referred to as genomic sites.

6.2.2 *Building Classifiers*

Machine learning models were developed to distinguish pausing from non-pausing sites based on the genomic features. Each model was tuned, trained and tested on n sites, with two equally sized sets of pausing and non-pausing sites. The classification models were implemented with the scikit-learn Python package and use p predictor variables (genomic features), which showed a variable degree of correlation between each other. 20% of the n sites were used to optimize the hyperparameters of the models using 5-fold cross-validation. The models were trained and tested on the remaining 80% of observations using 70% of them for the training and the remaining 30% for testing. The model performance was assessed using the area under the curve (AUC) values of the ROC curve (ROC-AUC) and precision-recall curve (PR-AUC). The importance of each feature was computed using the Permutation Importance, permuting each of the features 10 times. The number of sites n, the number of predictor variables p are listed in Supplemental Table S2. .

6.3 DATA

As our primary source of training and testing examples, we used the set of pausing sites detected in two biological replicates of the standard NET-seq obtained for the HeLa S3 cell line. Additionally, for testing and refining training examples, we used pausing sites detected in two biological replicates of the HiS-NET-seq obtained for the K562 cell line. The pausing site locations in other model organisms were derived using the following NET-seq data: *Escherichia coli* [43] (GEO accession: GSM1367304), *Saccharomyces cerevisiae* [28] (GEO accessions: GSM1673641 and GSM1673642), and *Arabidopsis thaliana* [38] (GEO accessions: GSM3814845 and GSM3814846).

6.4 RESULTS

6.4.1 *Feature engineering*

To analyse the impact of the DNA sequence and sequence-dependent factors on the transcriptional pausing, we engineered a set of sequence-dependant features that were previously implicated in pausing. For all species and genomic regions, we calculated the differences in nucleotide skewness and identity at positions of interest, thermodynamic features of the RNA-DNA hybrid, and free energy of the nascent RNA.

Since some of the tools and databases were only available for human cell lines, additional features were calculated for human samples. The human-specific features include the DNA shape and form, binding motifs of transcription factors (TFs) and RNA binding proteins (RBPs). A comprehensive list of features is given in Supplemental Table S3.

DNA skewness

XY skewness was defined as $\frac{|X|-|Y|}{|X|+|Y|}$, where $|X|$ denotes the number of nucleotides X. X and Y represent standard DNA nucleotides. A difference of XY skewness for all pairs of DNA nucleotides was calculated between the positions located 10 nucleotides upstream and downstream of the genomic site in a 20 nucleotide window.

Nucleotide identity

The nucleotide identity was extracted from the +1, -1, -2, -3, -10, -11 position from the genomic site, corresponding to the active centre of the RNA polymerase and both ends of the RNA-DNA hybrid.

Thermodynamic features of RNA-DNA hybrid

Thermodynamic features such as the entropy, enthalpy, Gibbs free energy, and the melting temperature of the 10 nucleotide RNA-DNA hybrid was calculated using MELTING [44] with the RNA-DNA model parameters.

Potential for nascent RNA hairpin formation

The minimum free energy of the stretch of a nascent RNA between positions 11 and 29, where position 1 corresponds to the last nucleotide added to the nascent RNA, was calculated using the ViennaRNA package [48]. The region between nucleotides 11 and 29 of the nascent RNA corresponds to the region of RNA hairpin structure formation.

DNA shape

DNAshapeR [12] was used to calculate DNA shape descriptors including the minor groove width (MGW), roll, propeller twist (ProT), helix twist (HelT), and potential energy (EP). The features consist of a minimum, maximum, mean, span, and mean value of the first derivative of the descriptors calculated for the region 10 nucleotides upstream and 5 nucleotides downstream of the genomic site. The region encompasses the RNA-DNA hybrid and a 5 nucleotide long DNA fragment downstream of the polymerase active centre.

DNA form

The presence of non-B DNA forms in the region 100 nucleotides upstream or downstream of the genomic site was also taken into consideration. Regions of the human genome that tend to form non-B DNA structures such as Z-DNA, A-phased repeats, inverted repeats, mirror repeats, and direct repeats, were extracted from non-B database v2.0 [10]. Additionally, pqsfinder [33] was used to predict the locations of the G-quadruplexes.

Transcription Factor binding motif

The presence of a transcription factor binding motif (TFBM) was determined for three regions referred to as 'polymerase footprint' (25 nucleotides upstream and downstream of the genomic site), 'upstream' (100 nucleotides upstream of the polymerase footprint), and 'downstream' (100 nucleotides downstream of the polymerase footprint). 810 position weight matrices (PWMs) describing 639 human transcription factors were downloaded from the JASPAR database [37]. To limit the number of features added to the model, the PWMs were clustered into 111 consensus matrices using RSAT matrix-clustering [9]. The consensus matrices were later used to scan for motif occurrences using the MOODS package [39].

RNA Binding Protein binding motif

1194 PWMs describing human RNA Binding Protein (RBP) motifs were downloaded from ATtRACT database [23]. To limit the number of features added to the model, the PWMs were clustered into 240 consensus matrices using RSAT matrix-clustering [9]. The upstream region of a pausing site was scanned for motifs using the MOODS package [39]. The RBP motif search was restricted to that region because the complementary sequence corresponds to the nascent RNA that RBPs can bind.

6.4.2 *Modeling promoter-proximal pausing in human cells*

Feature examination and selection

The first step of the analysis included a visual examination of all the engineered features (not shown) and an inspection of dependencies between the features. We excluded the features that took the same value for all of the examples. As the presence of correlated features might influence the estimates of feature importance, we calculated the Spearman correlation coefficient between all pairs of the features. We detected pairs of the features, for which the absolute value of the coefficient was larger than 0.7.Figure 6.1 presents the highly correlated feature pairs. We can see that the features describing the shape of the DNA showed a high absolute correlation, and some of them (e.g. mean potential energy) are correlated with the GC content in the region. Additionally, high correlation coefficients were observed for pairs of features describing thermodynamic parameters of the RNA-DNA hybrid, including the GC content of the hybrid. Therefore, to avoid the misinterpretation of feature importances, we present them later in this Chapter aggregated into the feature type categories.

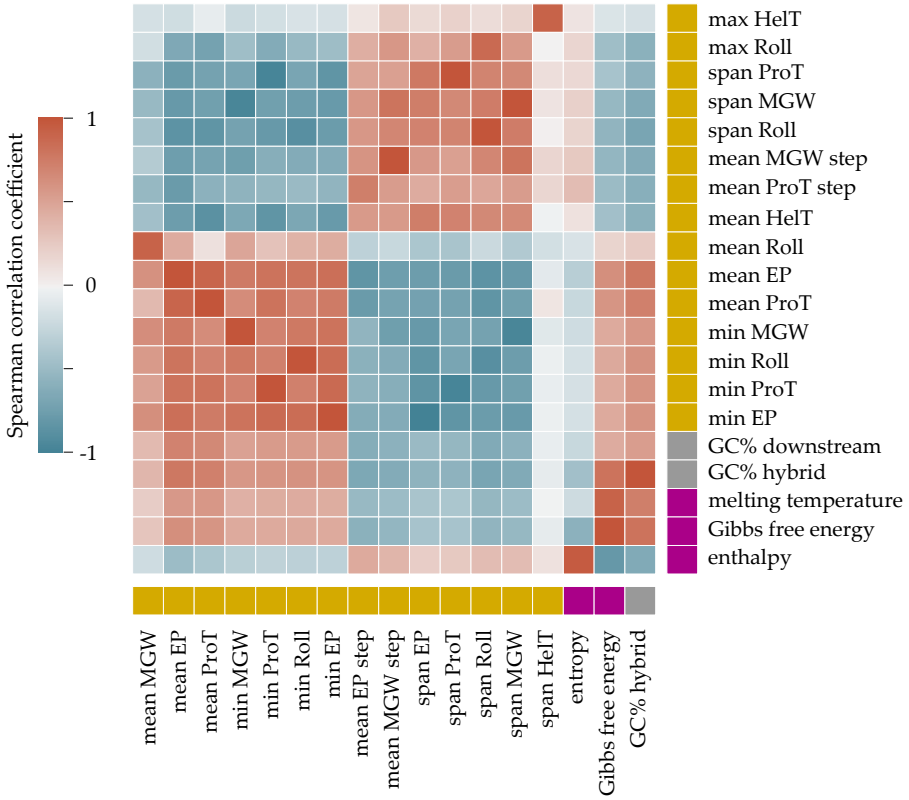

Figure 6.1: **Highly correlated feature pairs as determined by the Spearman correlation coefficient.** The heatmap shows the pairs with high negative (blue) and positive (red) correlation coefficients. Coloured squares on the right and below the heatmap indicate the type of the feature as follow: DNA shape (yellow), GC content (grey), and thermodynamic features of RNA-DNA hybrid (purple).

Finding a model suitable for classifying pausing and non-pausing sites

After we familiarized ourselves with the designed features and recognized the dependencies between them, we used the validation set to find the most suitable model for our classification problem. We tested three different learning algorithms: a logistic regression, a random forest, and a gradient boosting tree-based classifier. For each of the learning algorithms, the model can be further defined by a set of hyperparameters characteristic for the learning algorithm. In case of the logistic regression, we performed a grid search for the inverse of the regularization strength C and the parameter α of the elastic-net, using the following values: $C = \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ and $\alpha = \{0, 0.25, 0.5, 0.75, 1\}$. A random forest and a gradient boosting classifier use the same types of hyperparameters as they are both ensemble, tree-based approaches. Their hyperparameters include the number of trees N, the maximum number of features ω considered at each split, the maximum number of levels h in a tree, and the minimum number of samples s required to be at a leaf node. Additionally, a gradient boosting classifier is characterized by the learning rate r. We used the following values to perform a grid search: $N = \{400, 800, 1500, 3000, 5000\}$, $h = \{2, 4, 8, 16, 32, 64, 128\}$,

$s = \{2, 5, 10, 15\}$, and $r = \{0.001, 0.005, 0.01\}$. The maximal number of features ω considered at each split equalled the square root of all of the features. Effectively, we tested 585 models characterized by different hyperparameter combinations: 25 logistic regressions, 140 random forests and 420 gradient boosting classifiers.

Figure 6.2 shows the results obtained for the validation set. In the lower-left, we can see a group of underfitted models obtaining low performance on both test and train subsets of the validation set. This group consist of logistic regression models with high regularization ($C = 0.001$), which have a too small capacity to learn important aspects of the problem. In the upper right corner, there is a group of overfitted models showing a large gap between the train and test subset performance. The large number of overfitted models might be the result of both: the small size of the validation data set and the random hyperparametrization leading to too large model capacity. Among overfitted models, we found no random forests, logistic regression with low regularization ($C > 1$) and gradient boosting classifiers with $rN < 4$.

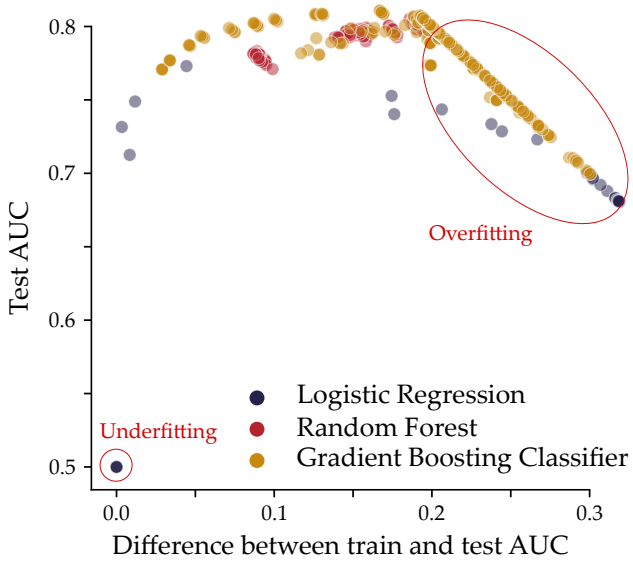

Figure 6.2: **Grid search results: performance of the validated models measured using mean ROC-AUC of 5-fold cross-validation.** The x-axis corresponds to the difference between performance using the training and test sets and the y-axis shows the performance of the test set.

We examined how the performance depends on the choice of the values of the hyperparameters. For logistic regression models, the performance depends on the regularization strength C, as shown in Figure 6.3. Most of the highly regularized logistic regression models (with the inverse of the regularization strength C equal 0.001) achieve classification that is not better than a random guess. Models with low regularization strength (C greater than 1), perform better than the highly regularized models, however, they do not reach the performance obtained with the moderate regularization strength.

In difference to the logistic regression, the performance of random forest models proved not to be sensitive to the selection of the hyperparameter values within the chosen range. Figure 6.4 shows that all tested models

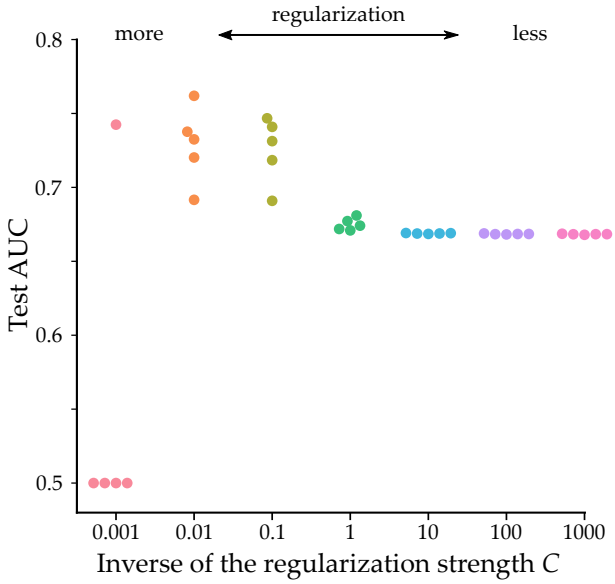

Figure 6.3: **The dependence between the performance of the validated logistic regression models and their hyperparameters.** The x-axis corresponds to the inverse of the regularization strength C. The y-axis shows performance measured using mean ROC-AUC of 5-fold cross-validation.

obtained high classification performance (mean ROC-AUC above 0.75). However, models with low maximum tree depth h (h smaller than 5) tend to obtain lower performance than the models with deeper trees.

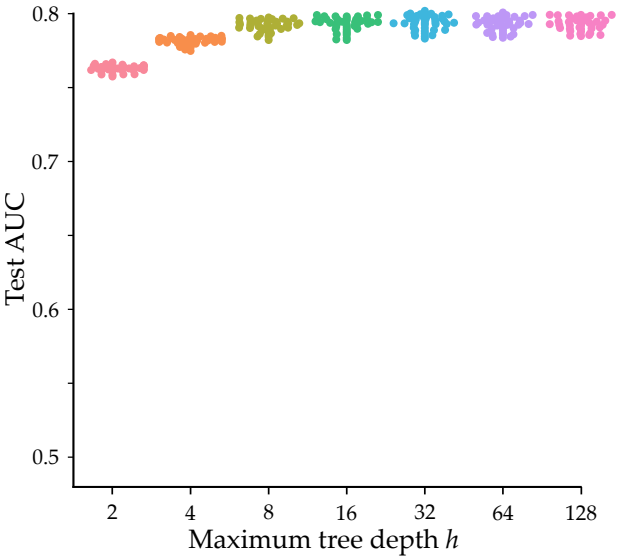

Figure 6.4: **The dependence between the performance of the validated random forest classifiers and their hyperparameters.** The x-axis corresponds to the maximum tree depth h. The y-axis shows performance measured using mean ROC-AUC of 5-fold cross-validation.

Figure 6.5 shows that the performance of the tree-based gradient boosting classifier depends on the product of the learning rate r and the number of trees N. Models with both a low learning rate and a small number of trees perform worse on average than models with high values of those hyperparameters. Additionally, the model performance is influenced by the number of samples necessary to form a leaf, with classifiers requiring a small number of samples (the minimum number of samples s at the leaf node equals 2) obtaining lower performance.

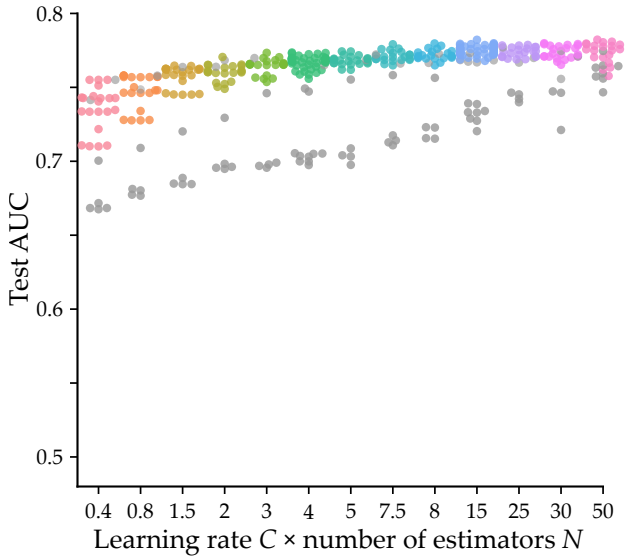

Figure 6.5: **The dependence between the performance of the validated gradient boosting classifiers and their hyperparameters.** The x-axis corresponds to the product of the learning rate r and the number of trees N. The y-axis shows performance measured using mean ROC-AUC of 5-fold cross-validation. Models with less than three samples required at the leaf node are marked in grey.

Having examined the influence of the hyperparameters on models performance, we were able to find models that can be applied to our classification problem. For training, we chose the best performing model for each of the algorithm types:

- a logistic regression with the inverse of the regularization strength C =0.01 and the parameter α=0.25,
- a random forest consisting of N=1500 trees of a maximum depth $h = 32$, and with at least $s = 2$ samples at each leaf,
- a gradient boosting classifier with a learning rate $r = 0.01$ consisting of N=5000 subsequent trees of a maximum depth $h = 32$, and with at least $s = 10$ samples at each leaf.

We trained the selected models and evaluated the classification using the test set. Figure 6.6 presents the performance of the chosen models on the test set. All models obtained high classification scores measured using the AUC-ROC: 0.85, 0.84, and 0.8 for the random forest, the gradient boost classifier, and the logistic regression respectively.

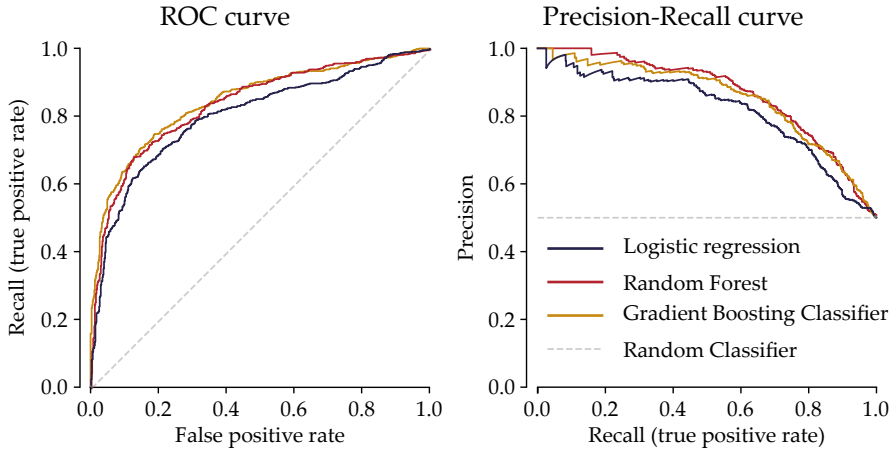

Figure 6.6: **ROC and precision-recall curves of the chosen models on the test set.** Grey lines indicate the expected performance result for a random classifier.

We performed a further assessment of the model performance using pausing sites detected in the HiS-NET-seq library prepared using the K562 cell line. As all of the features depend only on the DNA sequence, the models should be applicable to different human cell types. Additionally, if the misclassification rate for an independent set of sites is low it shows that the model does not tend to overfit. To avoid artificially high classification scores, we removed all of the pausing sites that were also present in the training data set. As Figure 6.7 shows, all the chosen models achieved even higher classification scores than for the test set derived from the same cell line as the training set. The AUC-ROC equalled 0.88, 0.87, and 0.86 for the random forest, the gradient boost classifier, and the logistic regression respectively. The higher scores might result from the fact that HiS-NET-seq variant includes additional enrichment step and therefore the resulting Pol II occupancy tracks are less noisy.

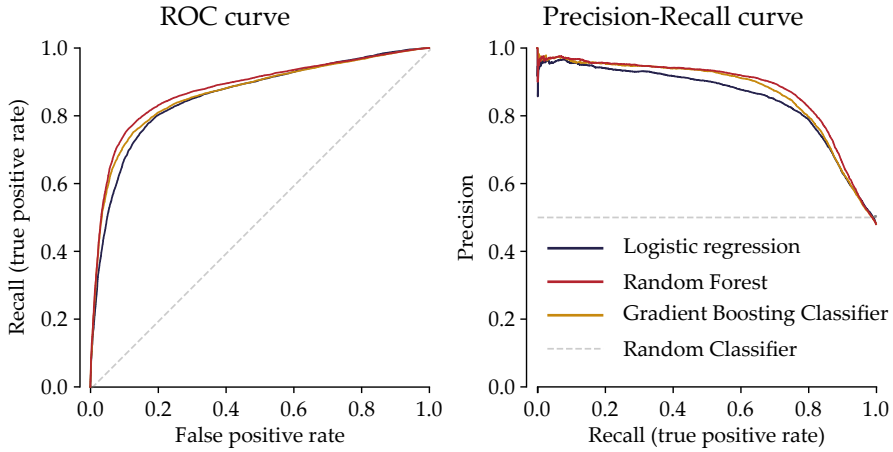

Figure 6.7: **ROC and precision-recall curves of the best performing models on the test 4sU.** Grey lines indicate the expected performance result for a random classifier.

Examining features predictive for pausing in the promoter-proximal region

To gain an insight into the mechanics of promoter-proximal pausing, we examined the features that were important for making correct discrimination between pausing and non-pausing sites. Those features are of interest, as they might contribute to evoking pausing at specific sites. We calculated feature importance measured as the decrease in AUC-ROC upon applying permuting values of each of the features 10 times. Figure 6.8 reports the results for the random forest model, as it obtained the highest classification scores on the test sets. The features for which the decrease in AUC-ROC upon the permutations was significantly larger than zero include: the nucleotide identities at the positions +1, -1, -2, -3 and the stacking energy between the nucleotides in the active centre of the polymerase (+1 and -1). All of the listed features achieved positive feature importance for the logistic regression and the gradient boosting classifier as well (shown in Supplemental Figures S2 and S3).

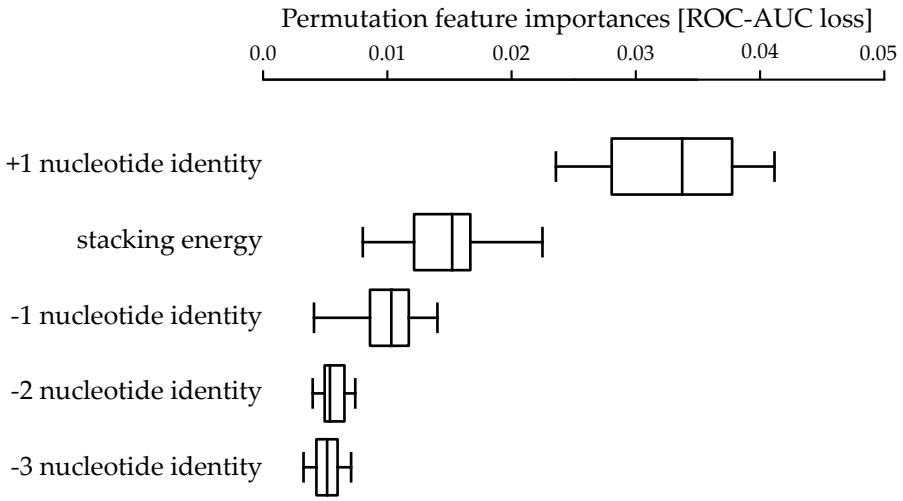

Figure 6.8: **Permutation feature importance calculated for the random forest using promoter-proximal pausing sites.** The distribution of the Permutation Importance was computed for each feature by permuting the feature 10 times. Only features with non-zero importance, as indicated by a *t*-test are plotted.

The permutation feature importance allows us to identify features predictive for classification, but it does not indicate which value increases the chance of belonging to the positive class. For example, Figure 6.8 shows that the stacking energy between the nucleotides at positions +1 and -1 is relevant for pausing, but we do not know from the permutation importance boxplot alone which ranges of energies characterize pausing sites. Therefore, we visualized the distribution of the important features in Figure 6.9. In this way, we can recognize the values that distinguish pausing and non-pausing sites in the promoter-proximal region. For pausing sites, we typically find guanine at the -1 position and thymine at the +1 position. We also observe higher stacking energy between those two positions comparing to the population of the non-pausing sites. Additionally, the distributions of nucleotide frequencies at positions -2 and -3 differ between the pausing and non-pausing sites.

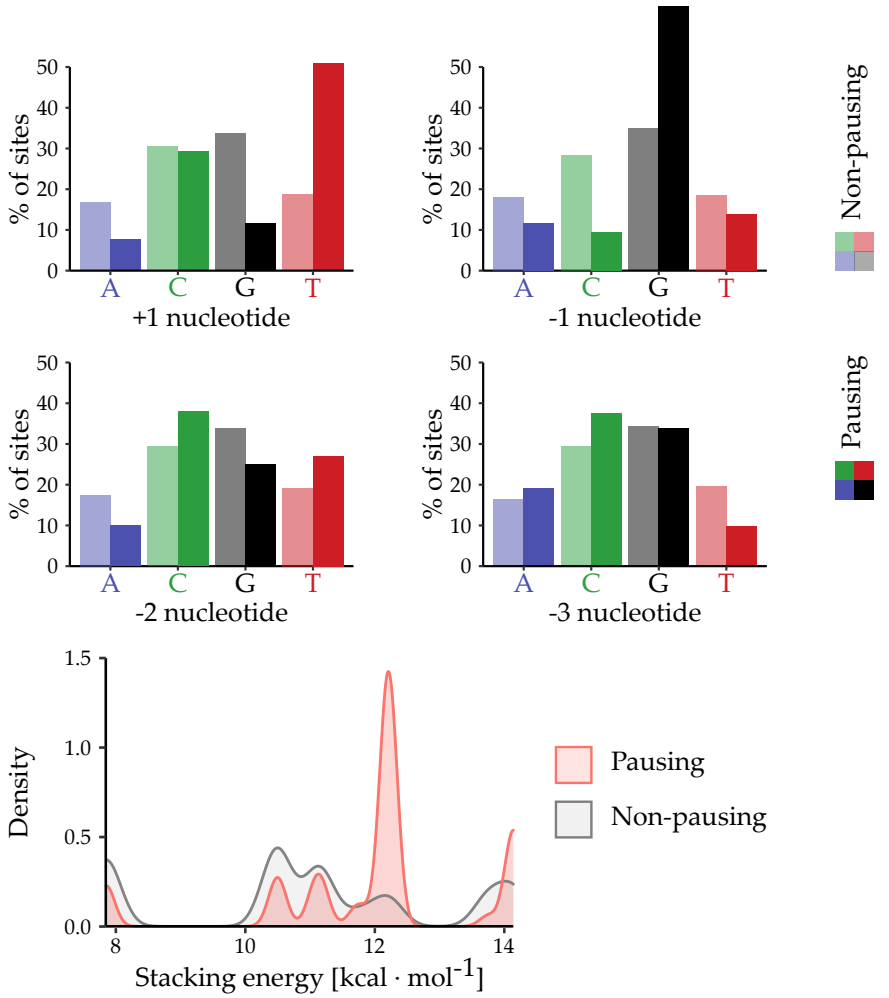

Figure 6.9: **Distributions of features important for classification performed by the random forest using promoter-proximal pausing sites.** The top two rows show frequencies of nucleotide identities at the positions of interest, with the saturation-coded class affiliation: pausing (saturated) and non-pausing (pale). The bottom row illustrates the stacking energy between the nucleotides at positions +1 and -1, with the colour-coded class affiliation: pausing (red) and non-pausing (grey).

6.4.3 *Modeling gene-body pausing in human cells*

Before training a new model for classifying pausing and non-pausing sites in the gene-body region, we applied the chosen classifiers trained to distinguish pausing and non-pausing sites in the promoter-proximal region. If these classifiers obtain high classification rates, we could assume that the pausing principles in the gene-body region do not differ from the pausing mechanism observed in the promoter-proximity and creating a model specialized for classifying gene-body pausing sites would not be necessary. The achieved results are presented in Figure 6.10. All models obtained low classification scores measured using the AUC-ROC: 0.56, 0.55, and 0.53 for the random forest, the gradient boost classifier, and the logistic regression respectively. As the predictive power of the models trained using promoter-proximal pausing sites is not much better than a random guess for predicting the class label of gene-body sites, we concluded that the pausing sites in the promoter-proximal and gene-body region are not generated by the same underlying distribution.

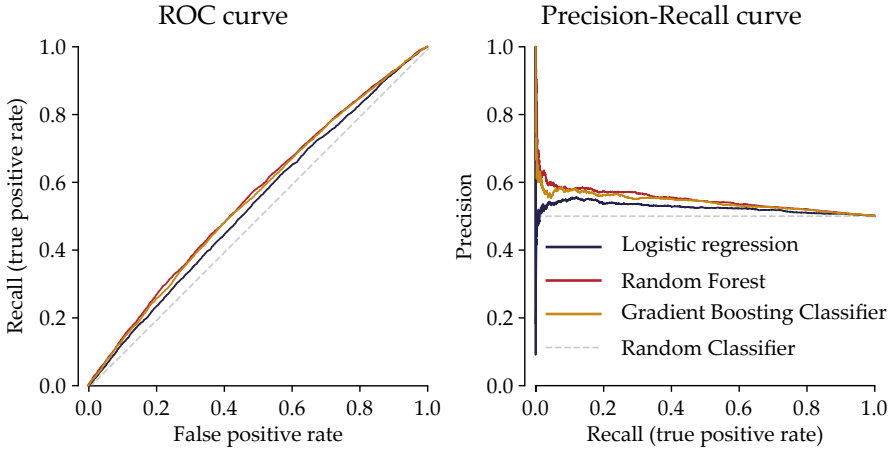

Figure 6.10: **ROC and precision-recall curves of the models trained using promoter-proximal pausing sites applied to the test set of gene-body sites.** Grey lines indicate the expected performance result for a random classifier.

Creating a model predicting gene-body pausing in human cells using standard NET-seq

As the models trained using promoter-proximal sites do not reach satisfactory accuracy for the gene-body site classification, we created validation training and test sets using gene-body sites examples, which we later used to develop new classifiers. We evaluated tested three different learning algorithms: a logistic regression, a random forest, and a gradient boosting tree-based classifier. To find the best-suited models, we performed a grid search using the same hyperparameter values as in Section 6.4.2. Figure 6.11 shows the results obtained for the validation set. In the lower-left, we can see a group of underfitted models obtaining low performance on both test and train subsets of the validation set. In the upper right corner, there is a group of overfitted models showing a large gap between the train and test subset performance.

All models, regardless of their capacity, obtained a moderate classification performance, with the AUC-ROC smaller than 0.7.

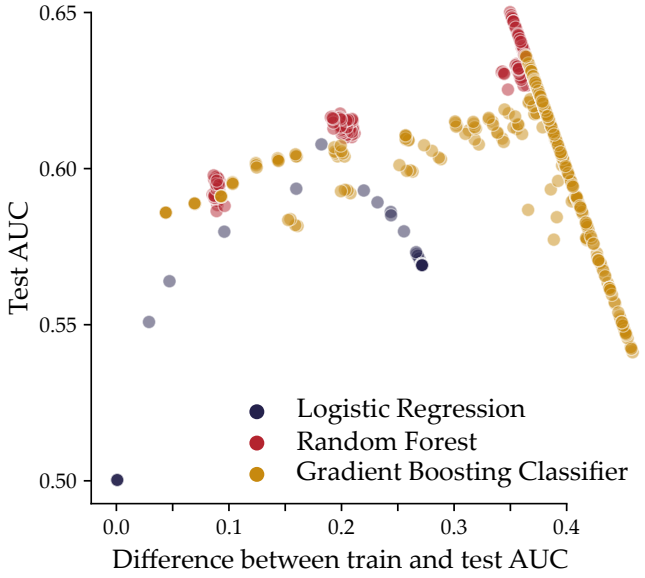

Figure 6.11: **Grid search results: performance of the validated models measured using mean ROC-AUC of 5-fold cross-validation.** The x-axis corresponds to the difference between performance using the training and test sets and the y-axis shows the performance of the test set.

For training using the gene-body site examples, we chose the best performing model for each of the algorithm types:

- a logistic regression with the ridge regularization (the parameter $\alpha = 0.25$) the inverse of the regularization strength $C = 0.01$,
- a random forest consisting of N=3000 trees of a maximum depth $h = 4$, and with at least $s = 10$ samples at each leaf,
- a gradient boosting classifier with a learning rate $r = 0.005$ consisting of N=5000 subsequent trees of a maximum depth $h = 4$, and with at least $s = 2$ samples at each leaf.

The selected models were trained using the gene-body training set and evaluated with the test examples. Figure 6.12 presents the performance of the chosen models on the gene-body test set. All models obtained moderate classification scores measured using the AUC-ROC: 0.68, 0.69, and 0.6 for the random forest, the gradient boost classifier, and the logistic regression respectively.

The modest classification performance might be a result of a combination of the following factors: the inappropriate choice of the model type, the overfitting or too many erroneous examples in the train set. Other explanations include a possibility that the gene-body pausing cannot be predicted from the sequence-dependent features alone, either because pausing in the gene-body region is stochastic or depends on other factors like chromatin modifications. We reasoned that the model types are not causing the problem, as the chosen algorithms were able to provide a good classification for the promoter-proximal pausing and the problem complexity is not expected to

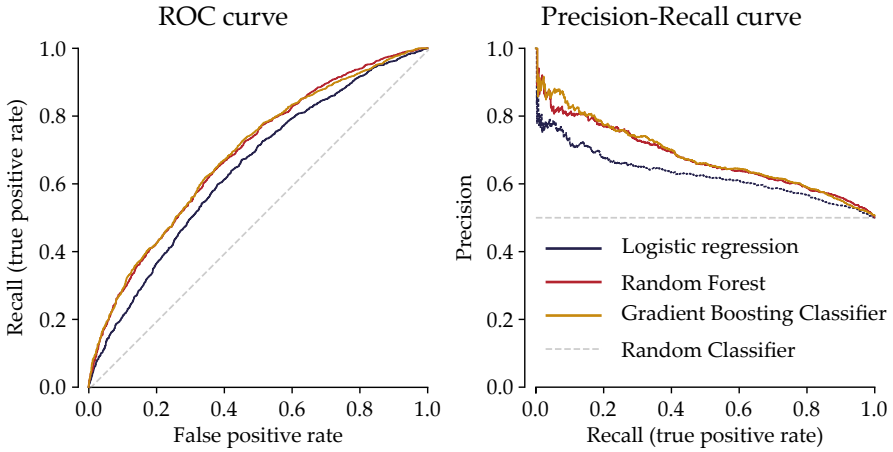

Figure 6.12: **ROC and precision-recall curves of the best performing models on the gene-body test sets.** Grey lines indicate the expected performance result for a random classifier.

be significantly different for a different genomic region. As Figure 6.11 shows, all of the validated models had only moderate predictive power, regardless of their capacity. Therefore, we assumed that the performance cannot be further improved by choosing a model with lower capacity. We attempted to refine the training set, by using the pausing sites detected in the HiS-NET-seq. This NET-seq variant includes an additional selection that ensures clearer enrichment of the nascent RNA species.

Creating a model predicting gene-body pausing in human cells using HiS-NET-seq

We repeated the tuning and training procedure with the learning set derived from the gene-body pausing sites detected in HiS-NET-seq. For training, we chose the best performing model for each of the tested algorithm types:

- a logistic regression with the inverse of the regularization strength $C = 0.01$ and the parameter $\alpha = 0.25$,
- a random forest consisting of N=800 trees of a maximum depth $h = 32$, and with at least $s = 2$ samples at each leaf,
- a gradient boosting classifier with a learning rate $r = 0.01$ consisting of N=800 subsequent trees of a maximum depth $h = 4$, and with at least $s = 15$ samples at each leaf.

We trained the selected models and evaluated the classification using the test set. Figure 6.13 presents the performance of the chosen models on the test set. All models obtained high classification scores measured using the AUC-ROC: 0.88, 0.87, and 0.8 for the random forest, the gradient boost classifier, and the logistic regression respectively.

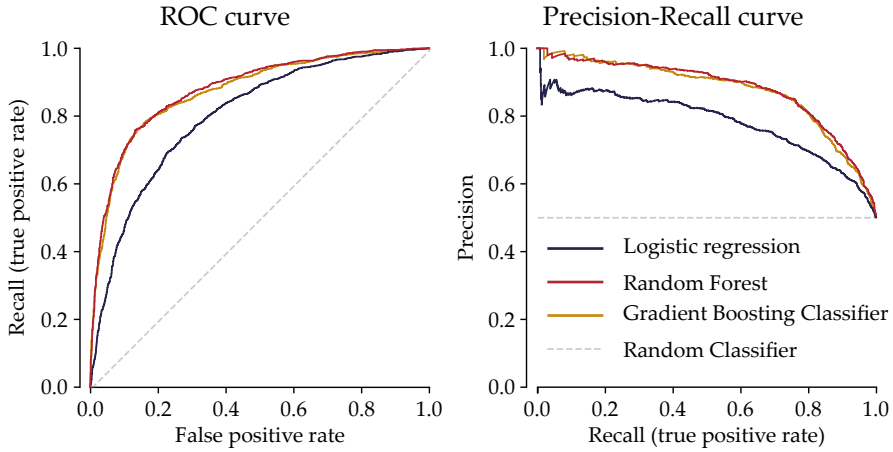

Figure 6.13: **ROC and precision-recall curves of the best performing models on the test sets.** Grey lines indicate the expected performance result for a random classifier.

Examining features predictive for pausing in the gene-body region

To gain an insight into the mechanics of gene-body pausing, we examined the features that were important for making correct discrimination between pausing and non-pausing sites. Those features are of interest, as they might contribute to evoking pausing at specific sites. We calculated feature importance measured as the decrease in AUC-ROC upon applying permuting values of each of the features 10 times. Figure 6.14 reports the results for the random forest model, as it obtained the highest classification scores on the test sets. The features for which the decrease in AUC-ROC upon the permutations was significantly larger than zero include: the nucleotide identities at the positions -2, -1, +1, -3 and the stacking energy between the nucleotides in the active centre of the polymerase (+1 and -1). All of the listed features achieved positive feature importance for the logistic regression and the gradient boosting classifier as well (shown in Supplemental Figures S4 and S5).

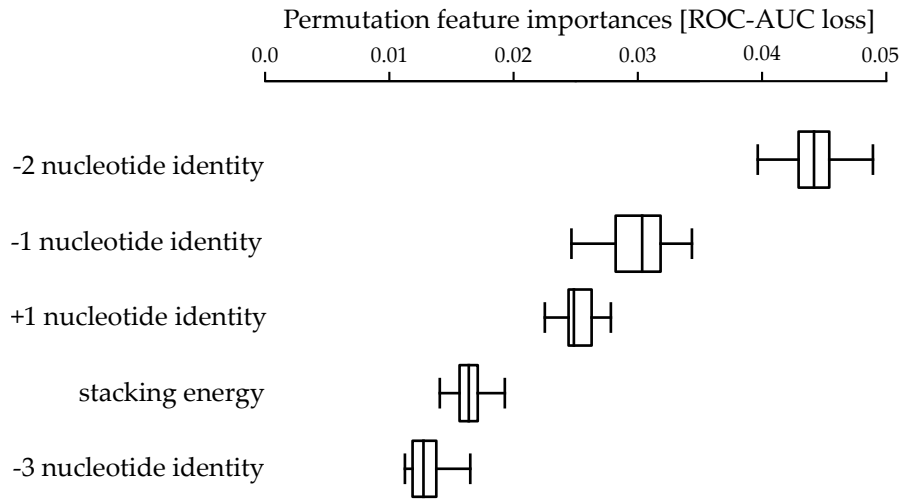

Figure 6.14: **Permutation feature importance calculated for the random forest using gene-body pausing sites detected in HiS NET-seq.** The distribution of the Permutation Importance was computed for each feature by permuting the feature 10 times. Only features with non-zero importance, as indicated by a *t*-test are plotted.

As previously mentioned, the permutation feature importance allows us to identify features predictive for classification, but it does not indicate which value increases the chance of belonging to the positive class. Therefore, we visualized the distribution of the important features in Figure 6.15. In this way, we can recognize the values that distinguish pausing and non-pausing sites in the gene-body region. For pausing sites, we typically find guanine or an adenine at the -1 position and thymine or guanine at the +1 position. We also observe higher stacking energy between those two positions comparing to the population of the non-pausing sites. Another difference can be noticed at the -2 position, where guanine is observed for the majority of the pausing sites. Additionally, the distributions of nucleotide frequencies at the -3 position differ between the pausing and non-pausing sites.

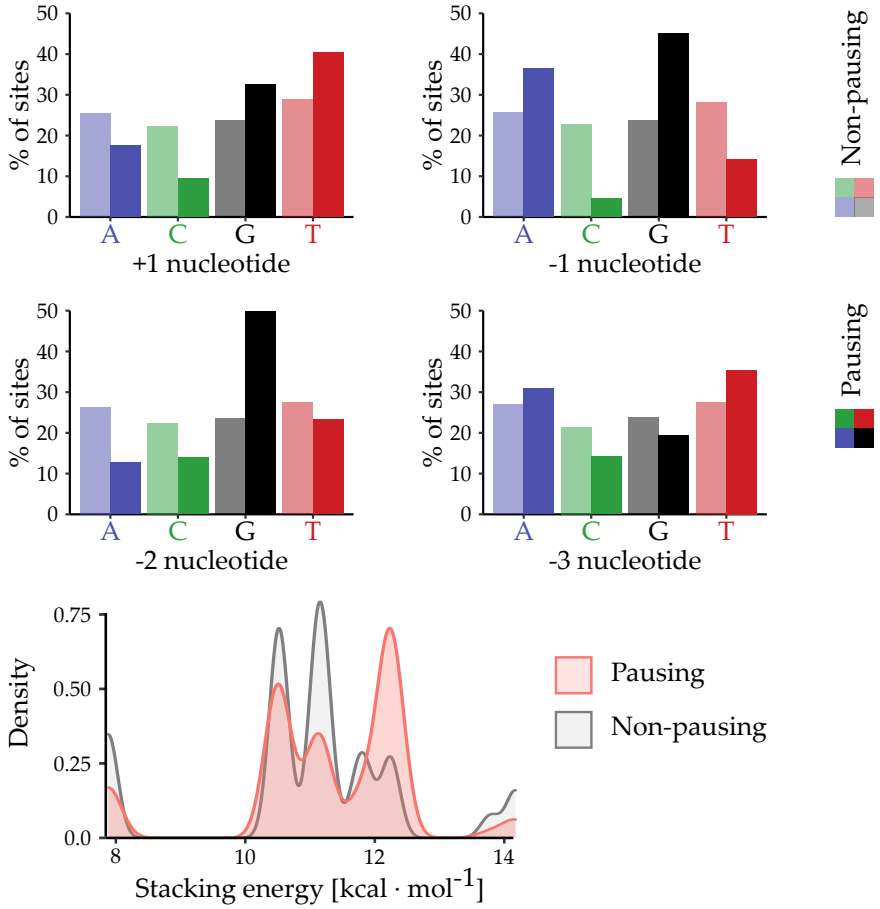

Figure 6.15: **Distributions of the features important for classification performed by the random forest using gene-body pausing sites detected in HiS NET-seq.** The top two rows show frequencies of nucleotide identities at the positions of interest, with the saturation-coded class affiliation: pausing (saturated) and non-pausing (pale). The bottom row illustrates the stacking energy between the nucleotides at positions +1 and -1, with the colour-coded class affiliation: pausing (red) and non-pausing (grey).

6.4.4 *Modeling transcriptional pausing in non-human organisms*

Since different types of transcriptional pausing have been reported in various organisms, we wanted to address the question of whether the location of pausing sites share similarities between species. We used NET-seq data available for 3 different organisms: *Escherichia coli* (bacteria), *Saccharomyces cerevisiae* (budding yeast, unicellular eukaryote) and *Arabidopsis thaliana* (plant, multicellular eukaryote). In difference to the modeling approach taken for the human pausing sites, where we trained different models for pausing sites situated in distinct genomic regions, here our positive set included all pauses within genes since clear evidence for promoter-proximal pausing has not yet emerged in these organisms. Additionally, since some of the tools and databases were only available for human cell lines, some of the features were not calculated for non-human organisms. An overview of features used can be found in Supplemental Table S4.

Table 6.1: AUC-ROC scores obtained for the chosen model organisms.

Model organism	Classification model		
	Logistic Regression	Random Forest	Gradient Boosting
E. Coli	0.83	0.86	0.85
S. Cerevisiae	0.88	0.90	0.90
A. Thaliana	0.85	0.88	0.87

We repeated the tuning and training procedure for all 3 organisms. For training, we chose the best performing model for each of the tested algorithm types:

- a logistic regression with the inverse of the regularization strength $C = 0.01$ and the parameter $\alpha = 0.25$,
- a random forest consisting of N=800 trees of a maximum depth $h = 32$, and with at least $s = 2$ samples at each leaf,
- a gradient boosting classifier with a learning rate $r = 0.01$ consisting of $N = 800$ subsequent trees of a maximum depth $h = 4$, and with at least $s = 15$ samples at each leaf.

The selected models were trained using the training set and evaluated with the test examples derived for the chosen organisms. Figures 6.16, 6.17, and 6.18 present the performances of the chosen models on the test sets derived for *E. Coli*, *S. Cerevisiae*, and *A. Thaliana* respectively. All models showed high classification scores measured using the AUC-ROC, with the random forest showing the best performance for the chosen model organisms. Table 6.1 summarizes the AUC-ROC values obtained by the selected models.

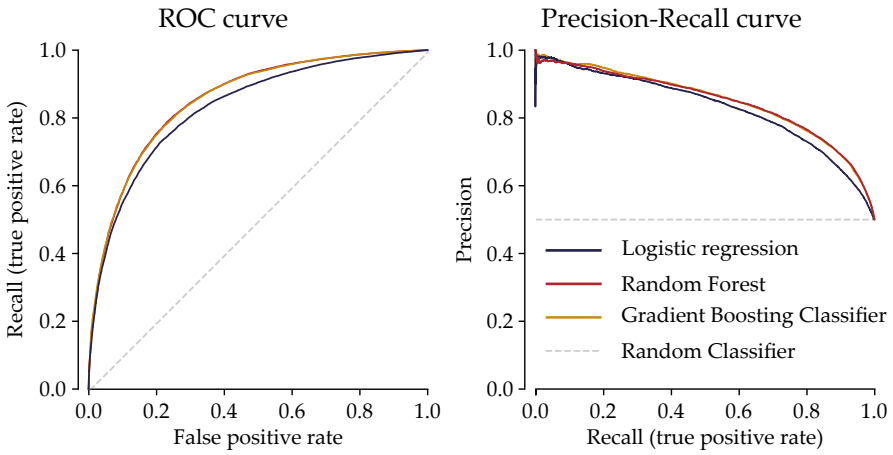

Figure 6.16: **ROC and precision-recall curves of the best performing models on the test sets obtained for *E. Coli*.** Grey lines indicate the expected performance result for a random classifier.

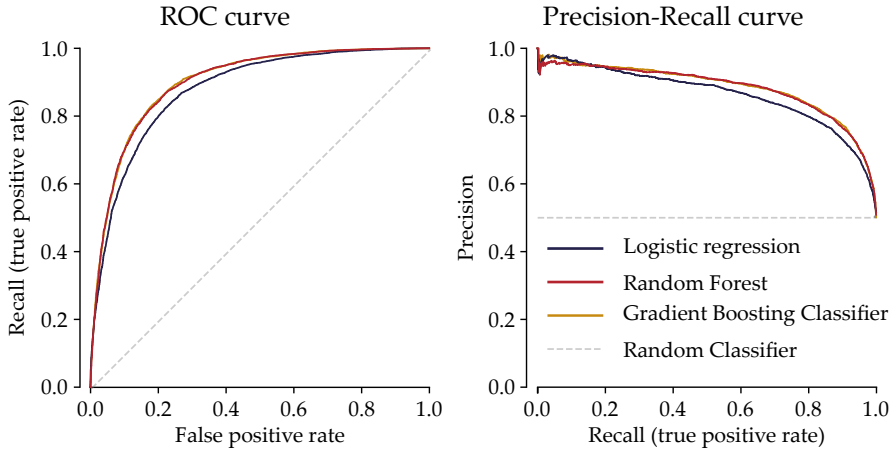

Figure 6.17: **ROC and precision-recall curves of the best performing models on the test sets obtained for *S. cerevisiae*.** Grey lines indicate the expected performance result for a random classifier.

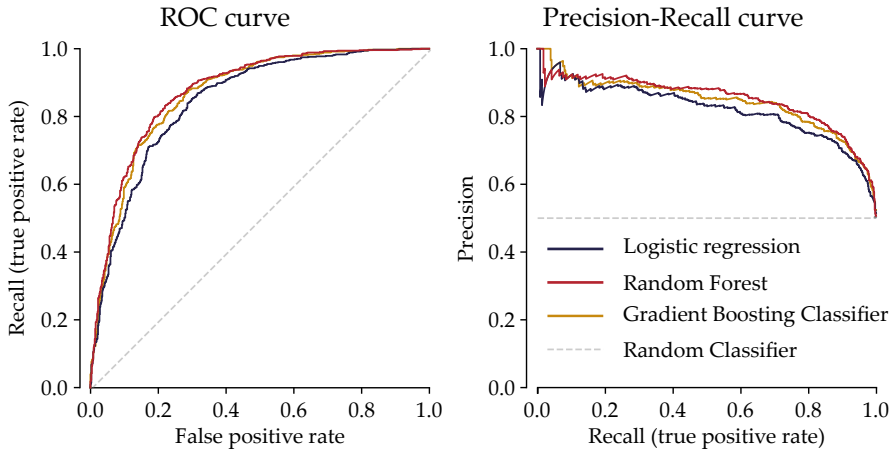

Figure 6.18: **ROC and precision-recall curves of the best performing models on the test sets obtained for *A. Thaliana*.** Grey lines indicate the expected performance result for a random classifier.

Examining features predictive for pausing in non-human organisms

To gain an insight into the mechanics of pausing in the selected organisms, we examined the features that were important for making correct discrimination between pausing and non-pausing sites. We calculated feature importance measured as the decrease in AUC-ROC upon applying permuting values of each of the features 10 times. In the subsections below, we discuss the features important for the model organisms.

Features predictive for transcriptional pausing in E. Coli

Figure 6.19 reports the results for the random forest model, as it obtained the highest classification scores on the test sets. The features for which the decrease in AUC-ROC upon the permutations was significantly larger than zero include: the nucleotide identities at the positions +1, -1, -2, -3, -10, -11 and the stacking energy between the nucleotides in the active centre of the polymerase (+1 and -1).

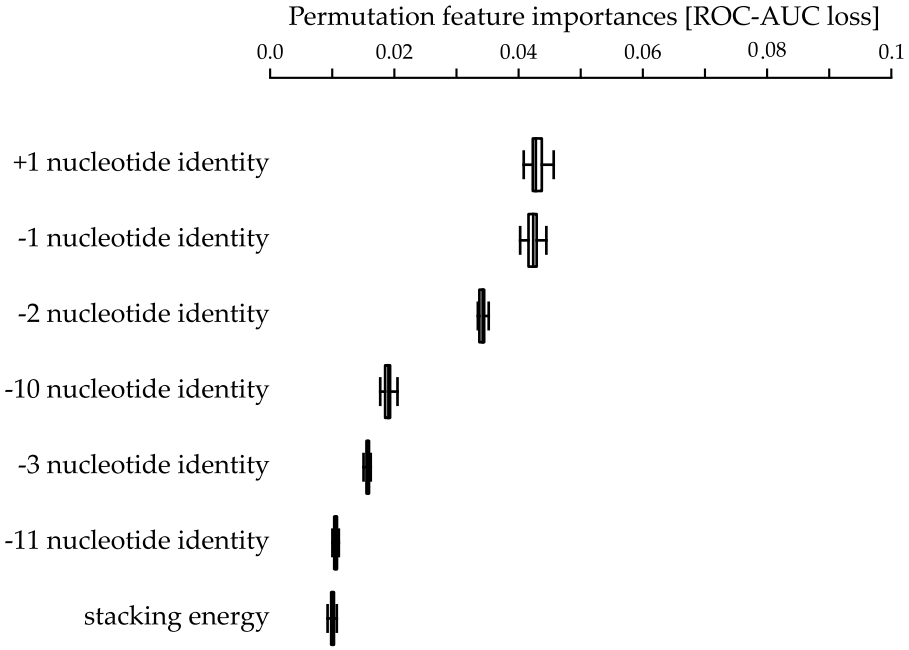

Figure 6.19: **Permutation feature importance calculated for the random forest using pausing sites detected in bacterial NET-seq.** The distribution of the Permutation Importance was computed for each feature by permuting the feature 10 times. Only features with non-zero importance, as indicated by a *t*-test are plotted.

As previously mentioned, the permutation feature importance allows us to identify features predictive for classification, but it does not indicate which value increases the chance of belonging to the positive class. Therefore, we visualized the distributions of the features predictive for classification of pausing and non-pausing sites in *E. Coli* in Figure 6.20. For pausing sites, we typically find guanine at the positions +1, -2, -10, and -11. We also observe thymine or a cytosine at position -1 with a higher frequency than for the non-pausing sites. Another difference can be noticed at the -3 position, where thymine is more frequently present at the pausing sites.

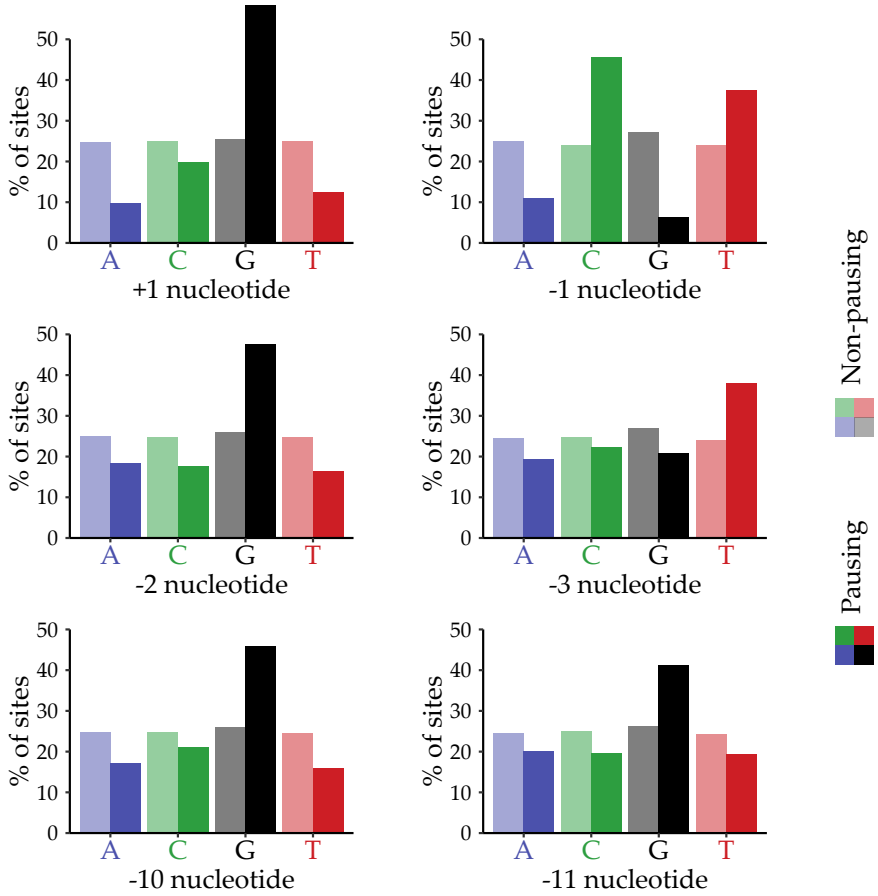

Figure 6.20: **Distributions of the features important for classification performed by the random forest using pausing detected in bacterial NET-seq.** Frequencies of nucleotide identities at the positions of interest are shown, with the saturation-coded class affiliation: pausing (saturated) and non-pausing (pale).

Features predictive for transcriptional pausing in S. Cerevisiae

Figure 6.21 reports the results for the random forest model, as it obtained the highest classification scores on the test sets. The features for which the decrease in AUC-ROC upon the permutations was significantly larger than zero include: the nucleotide identities at the positions +1, -1, -2, -3, the features describing the thermodynamical state of the hybrid and the GC content downstream of the analysed sites.

The distributions of the features that are most predictive for the classification of pausing and non-pausing sites in *S. Cerevisiae* are visualized in Figure 6.22. For pausing sites, we find at the majority of pausing sites thymine at the +1 position and adenine at the -1 position. We typically observe with a higher frequency than for the non-pausing sites guanine at position -2 and guanine or a cytosine at position -3.

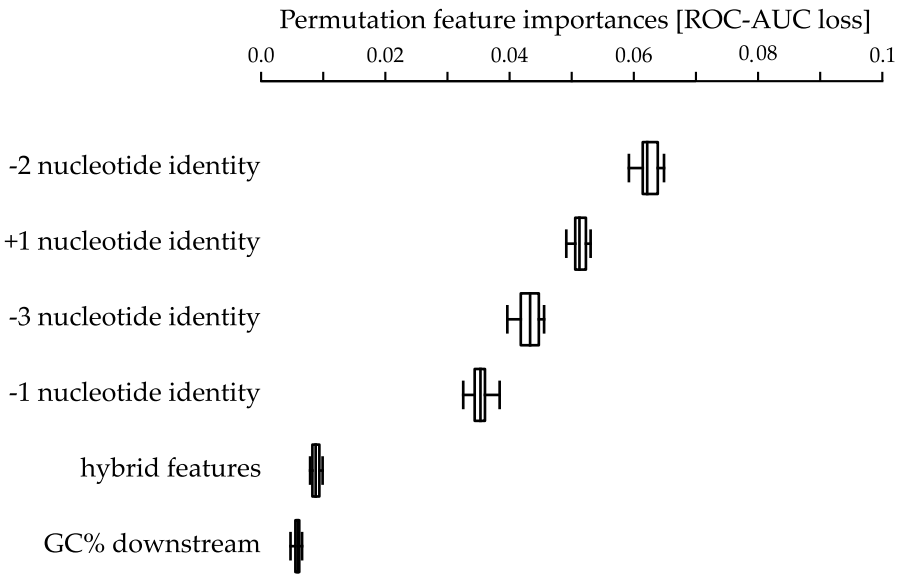

Figure 6.21: **Permutation feature importance calculated for the random forest using pausing sites detected in yeast NET-seq.** The distribution of the Permutation Importance was computed for each feature by permuting the feature 10 times. Only features with non-zero importance, as indicated by a *t*-test are plotted.

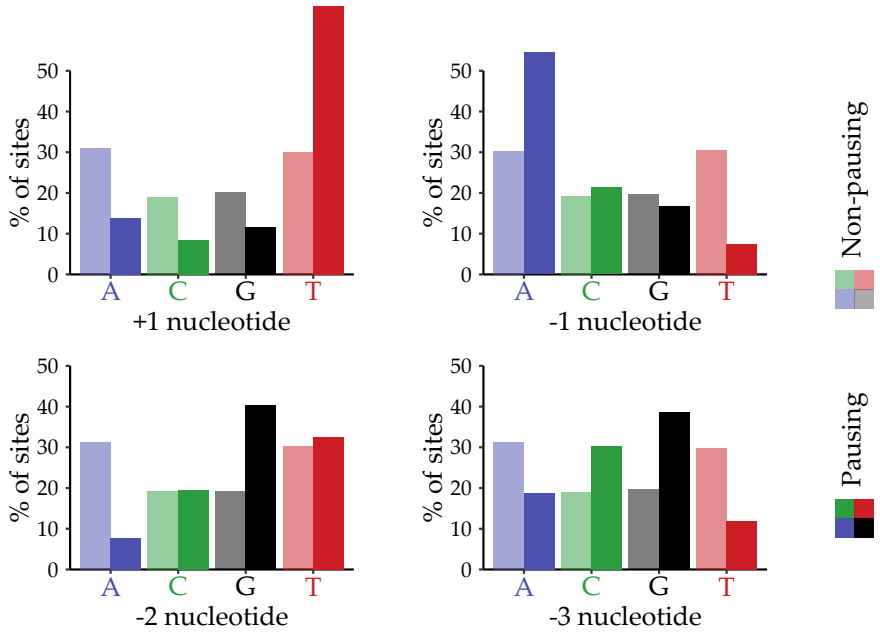

Figure 6.22: **Distributions of the features important for classification performed by the random forest using pausing sites detected in yeast NET-seq.** Frequencies of nucleotide identities at the positions of interest are shown, with the saturation-coded class affiliation: pausing (saturated) and non-pausing (pale).

Features predictive for transcriptional pausing in A. Thaliana

Figure 6.23 reports the results for the random forest model, as it obtained the highest classification scores on the test sets. The features for which the

decrease in AUC-ROC upon the permutations was significantly larger than zero include: the nucleotide identities at the positions +1, -1, -2, the GC content (especially in the region upstream of the analysed sites), and the potential of the nascent RNA to form a hairpin within the RNA exit channel.

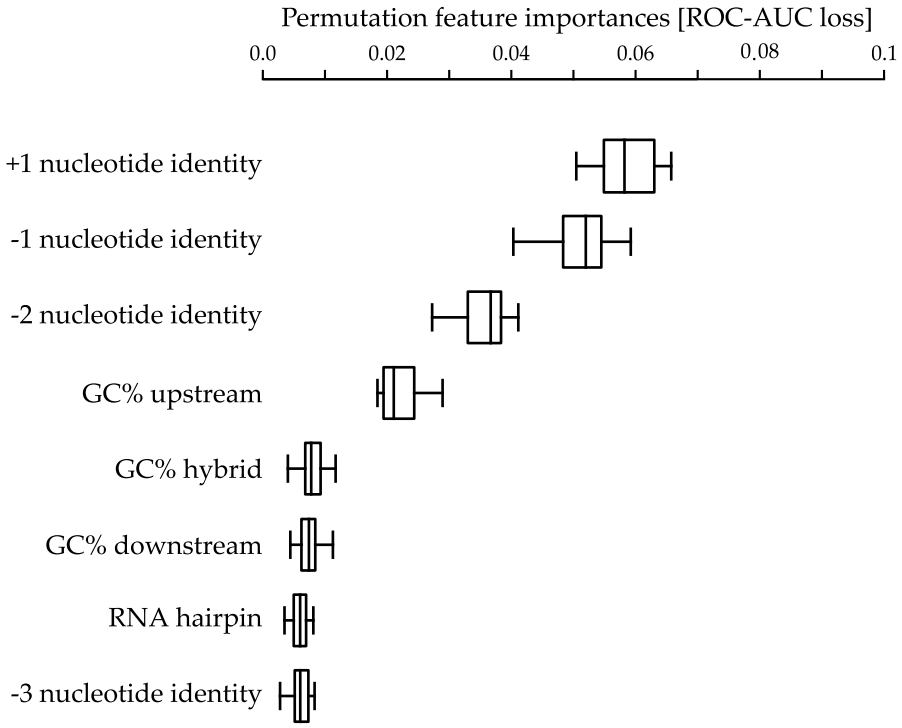

Figure 6.23: **Permutation feature importance calculated for the random forest pausing sites detected in plant NET-seq.** The distribution of the Permutation Importance was computed for each feature by permuting the feature 10 times. Only features with non-zero importance, as indicated by a *t*-test are plotted.

The distributions of the features that are most predictive for the classification of pausing and non-pausing sites in *A. Thaliana* are visualized in Figure 6.24. For pausing sites, we typically find guanine at position +1, thymine at position -1, and cytosine at position -2. We also observe a higher GC content upstream of the pausing than non-pausing sites.

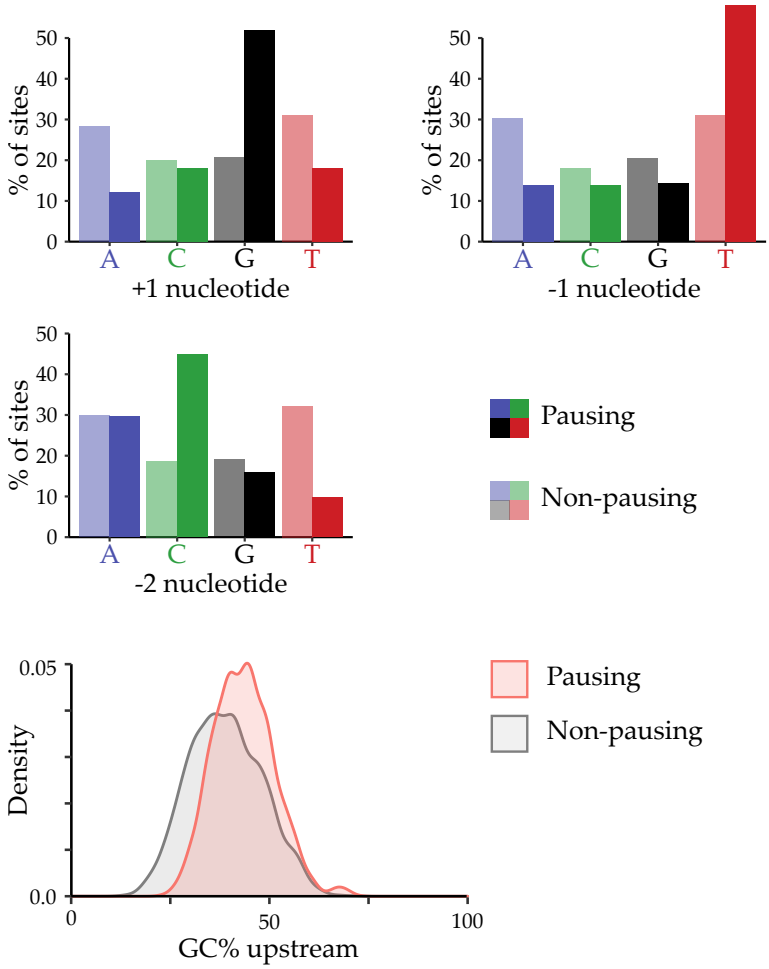

Figure 6.24: **Distributions of features important for classification performed by the random forest using pausing sites detected in plant NET-seq.** The top two rows show frequencies of nucleotide identities at the positions of interest, with the saturation-coded class affiliation: pausing (saturated) and non-pausing (pale). The bottom row illustrates the GC content upstream of the site, with the colour-coded class affiliation: pausing (red) and non-pausing (grey).

6.5 DISCUSSION

Here, we attempted to identify the determinants of the Pol II pausing positions within a gene through the integration of a comprehensive list of features derived from the DNA sequences underlying the pausing sites. We used the DNA sequence features to train separate classification models for two regions of a gene, namely promoter-proximal and gene-body region. The task of those machine learning models was to discriminate between positions where Pol II transcribes in a processive manner (a negative set) and pauses (a positive set). In this way, we were able to systematically analyse the contribution of features that were previously linked to Pol II pausing or likely to interrupt processive transcription.

For the classification of the sites in the promoter-proximal region, we examined three machine learning algorithms: a logistic regression, a random forest classifier and a gradient boosting classifier based on regression trees. For each of the machine learning algorithms, we used the validation set to 1) find the hyperparameters leading to the highest performance and 2) explain why a given set of parameters leads to high or low performance. We trained the best performing model for each of the algorithm types using promoter-proximal pausing sites detected in the standard NET-seq libraries created for the HeLa S3 cell line. The lowest classification performance was obtained for the logistic regression model (AUC-ROC 0.80); however, it was not much worse than the performance of the tree based models (AUC-ROC of the random forest: 0.85, AUC-ROC of the gradient boosting classifier: 0.84), which suggests that including the interplay between the features is not necessary to correctly classify the site and that promoter-proximal pausing sites form quite a homogeneous group. We confirmed the generalizability of our model through validation of model predictions with an independent set of pausing sites detected in HiS-NET-seq libraries created using K562 cell line. The classification performance was higher for the test set derived for the HiS-NET-seq than for the standard NET-seq, even though the training was performed using standard NET-seq data.

To uncover the rules that control Pol II pausing in the promoter-proximal region, we analysed the importance of the sequence derived features. The features important for promoter-proximal pausing include the nucleotide identities at the positions +1 and -1 and the stacking energy between the them. Therefore, our study links YGT sequence in the active center of the polymerase and promoter-proximal pausing, where Y denotes cytosine or thymine. This finding is consistent with a previous *in vitro* study showing that a TG dinucleotide motif can provoke a slowdown of transcribing bacterial RNA polymerase and that a repeat of this motif (TGTG) had an even stronger effect on pause induction [32]. This motif also has interesting similarities with a pause stabilizing element (Inr-G) in *Drosophila* that contains a GT dinucleotide [63]. Further important features include the nucleotide identity at the position -3, with cytosine being more frequently observed there for pausing sites. Interestingly, we did not identify other factors previously implicated in the transcriptional pausing such as GC-content [65, 69, 77]. The discrepancy is likely to result from the difference in formulating the research question. Here, we set out to find the determinant of the exact positions where polymerase pauses, whereas previous studies were examining the characteristics of genes prone to promoter-proximal pausing. Therefore, the features that were identified in previous previous studies might play a role

in promoter-proximal pausing, but they are not defining the exact pausing position within the gene.

To examine if the positions of gene-body pausing are determined by the same elements as promoter-proximal pauses, we applied the models trained with the promoter-proximal pausing sites to predict pausing in the gene-body region. All models achieved the correct classification rate close to the random guess, which suggests that the pausing sites in the promoter-proximal and gene-body region are not generated by the same underlying distribution. Therefore, we again examined three machine learning algorithms (a logistic regression, a random forest and a gradient boosting classifier based on regression trees) and trained the best performing model for each of the algorithm types using gene-body pausing sites detected in the standard NET-seq libraries created for the HeLa S3 cell line. All models achieved modest classification performance (AUC-ROC below 0.7). The high misclassification rate might be a result of a combination of the following factors: the inappropriate choice of the model type, overfitting or too many erroneous examples in the training set. Other explanations include a possibility that the gene-body pausing cannot be predicted from the sequence-dependent features alone, either because pausing in the gene-body region is stochastic or it depends on other factors like chromatin modifications. We reasoned that the model types are not the cause of the modest classification performance, as the chosen algorithms were able to provide a good classification for the promoter-proximal pausing and the problem complexity is not expected to be significantly different for different genomic regions. Additionally, all of the validated models had only moderate predictive power, regardless of their capacity. Therefore, we assumed that the modest performance is not caused by overfitting and cannot be further improved by choosing a model with lower capacity. Hence, we set out to refine the training set by using the pausing sites detected in the HiS-NET-seq libraries created for K562 cell line. This NET-seq variant includes an additional selection step that ensures better enrichment of the nascent RNA species. All models achieved high classification (AUC-ROC above 0.87 for the tree-based models and AUC-ROC of 0.80 for the logistic regression).

The most predictive feature for the gene-body pausing is the nucleotide identity at the position -2, with guanine observed at this position for the majority of pausing sites. Other important features include nucleotide identity at the positions +1, -1, and -3 and the stacking energy between the nucleotides in the active center of the Pol II. For pausing sites in gene-body regions, we typically find guanine or adenine at the -1 position and thymine or guanine at the +1 position. We also observe higher stacking energy between those two positions comparing to the population of the non-pausing sites. Another difference can be noticed at the -3 position, where adenine or thymine are observed for the majority of the pausing sites. Similarly to the promoter-proximal pausing, the nucleotides predictive for pausing are located close to the active center of the polymerase and the pausing sites exhibit increased stacking energy between the -1 and +1 nucleotides, with guanine and thymine being most frequently observed at those positions. This finding might indicate that pausing within the gene-body may be evoked by the same sequence elements as promoter-proximal pausing, at least for a subset of pausing sites.

We next asked whether the sequence determinants of pervasive RNA polymerase pausing were evolutionary conserved. To address this question,

we applied our modeling approach to reveal determinants for pauses detected in NET-seq data available for bacteria (*Escherichia coli*), budding yeast (*Saccharomyces cerevisiae*) and plants (*Arabidopsis thaliana*). All models showed high classification scores (AUC-ROC above 0.83). For all of the investigated species, identities of the nucleotides at the active center of the polymerase (+1 and -1 nucleotide) and in the upstream part of the RNA-DNA hybrid (-2 nucleotide) were predictive for the polymerase pausing. However, the distributions of nucleotide frequencies at the important positions varied for different species. Additionally, *E. coli* predictive features include nucleotide identities at positions -3, -10, and -11, which is consistent with previous studies [43]. In *A. thaliana* increased GC content is observed upstream of the pausing sites.

In this study, we identified the most predictive features for transcriptional pausing in different genomic regions and various species. Our approach can be used to generate hypotheses that are supported by existing data and that can be further validated experimentally. It is important to note that although machine learning approaches help us systematically evaluate correlative relationship between the input features and the output measurements, a high predictive power of a feature does not necessarily imply that the feature is actually causative for the phenomenon being predicted. Additionally, a machine learning model is only as good as the data used for training it is. Therefore, machine learning approaches might fail if the measured data is of low quality. Lastly, the performance of classical machine learning approaches depends heavily on the quality and the relevance of the features included in the model. A possible extension of the presented study overcoming the last limitation is training a deep learning model, which embeds the computation of features into the machine learning model itself to yield an end-to-end model.

7

SUMMARY AND CONCLUSIONS

In this thesis, we investigated what are the causes of Pol II pausing. We addressed this question in two steps. First, we set out to find Pol II pausing sites in the NET-seq data. Second, we attempted to identify the determinants of Pol II in an unbiased manner based on the underlying DNA sequence.

Pausing polymerases result in local enrichments of NET-seq signal because the longer the polymerase stays at the given position, the more frequently it is encountered at this position. However, not all of the occupancy peaks correspond to pausing polymerases. Our careful examination of the NET-seq characteristics revealed genetic locations exhibiting a potential to generate peaks that do not reflect Pol II pausing. Those blacklisted regions include genes transcribed by other human polymerases, genes encoding RNA species serving their function in the close proximity of the nascent RNA and chromatin, and single-nucleotide positions corresponding to the 3′ ends of transient RNA products of the nascent RNA processing. We proposed an improved NET-seq processing pipeline that limits the presence of artificial peaks thanks to masking the blacklisted regions and recognizing positions prone to generate high signal due to the reverse-transcriptase mispriming. The improved NET-seq pipeline includes collapsing the PCR duplicated reads before the read mapping, which decreases the processing time.

We designed and implemented a tool to detect pausing sites in the high-resolution Pol II occupancy tracks. We proposed two approaches: a resampling-based and a parametrical approach. We compared the two approaches and showed the parameters for which they yield a similar sets of peaks. We examined how the choice of parameters influences the significance threshold. We performed random downsampling of a NET-seq library to investigate how the sequencing depth and the library complexity influence the number of peaks detected, showing that the number of called peaks correlates with the sequencing depth and drops proportionally to the decreasing number of reads. Finally, we applied the peak calling algorithm to two technical replicates of a NET-seq library and showed the robustness of the peak calling approach.

Following the technical examinations and quality checks of the NET-seq data, we examined the distribution of the pausing sites over the human genome. We showed that Pol II pausing is not limited to the promoter-proximal region, but it occurs in the gene-body region in both sense and antisense directions as well as in the intergenic regions. Then, we investigated the sequences underlying pausing sites detected using two NET-seq variants in the promoter-proximal and gene-body regions. For the promoter-proximal region, we found a motif consisting of two parts: the G_{-10} at the upstream fork junction of the RNA-DNA hybrid and the $Y_{-2}G_{-1}Y_{+1}$, where Y is thymine or cytosine, at the region spanning the active site of Pol II and the downstream fork junction of the RNA-DNA hybrid. The comparison of transcriptional pausing sites in different model organisms revealed similarities between the motifs underlying human promoter-proximal pausing sites and pervasive pausing sites in bacteria. This similarity points toward a possible conservation of the sequence-dependent transcriptional pausing

mechanism. For the gene-body region, we did not find an underlying motif using the standard NET-seq variant. However, we did find a guanine-rich motif encompassing the downstream fork junction characteristic for pausing sites in gene-body detected using HiS-NET-seq.

We then set out to identify the determinants of Pol II in an unbiased manner based on the underlying DNA sequence. We created a large number of features, including factors that were previously linked to transcriptional pausing, but also factors that were not yet connected to Pol II pausing. To predict the predisposition of a genomic site to evoke Pol II pausing, we tested machine learning models such as logistic regression and two tree-based ensemble models. We showed that these models are able to discriminate with a high accuracy the pausing and non-pausing sites based only on features derived from the underlying DNA sequence. However, we were able to train the model for the gene-body sites only using HiS-NET-seq data and not with standard NET-seq data.

Examining feature importance helped us to identify the most important features in the model, namely the main potential determinants of Pol II pausing in different genomic regions. For all investigated species and genomic regions, identities of the nucleotides at the active center of the polymerase and in the upstream part of the RNA-DNA hybrid were predictive for the polymerase pausing. However, it is important to note that the distributions of nucleotide frequencies at the important position varied between the species and genomic regions. Additionally, our study linked the YGT sequence in the active center of the polymerase and promoter-proximal pausing, where Y denotes cytosine or thymine. Surprisingly, we did not identify other factors previously implicated in the transcriptional pausing. The difference is likely to result from the difference in formulating the research question. Here, we set out to find the determinant of the exact positions where polymerase pauses, whereas previous studies were examining the characteristics of genes prone to promoter-proximal pausing. Therefore, the features that were identified in previous previous studies might play a role in promoter-proximal pausing, but they are not defining the exact pausing position within the gene.

Our approach of creating a large number of features, including factors that were not yet previously linked to a given phenomenon, can be used to generate hypotheses that are supported by the existing data and that can be further validated experimentally. Although machine learning approaches help us systematically evaluate correlative relationships between the input features and the phenomenon, a high predictive power of a feature does not necessarily imply that the feature is actually causative for the phenomenon being predicted. Nevertheless, the features identified by the model as important for distinguishing pausing sites are promising candidates for perturbation experiments that might show the causal relationship between the selected features and transcriptional pausing.

8 CONTRIBUTIONS TO OTHER PROJECTS

In this Chapter, we list the additional projects, which are not covered in previous chapters, that the author contributed to during her work as a PhD candidate. We briefly describe these projects and the author's contributions.

8.1 ANALYSING CHANGES IN ISOFORM COMPOSITION AND CODING POTENTIAL DURING NEURONAL DIFFERENTIATION

Alternative splicing is a mechanism that increases transcriptomic diversity and can be regulated co-transcriptionally (see Section 2.2.4). We were interested in how high is the impact of the Pol II transcriptional speed on the alternative splicing regulation. As a model for our study, we chose the neuronal differentiation system, because alternative splicing occurs at high frequency in brain tissues and contributes to every step of nervous system development [68]. Olga Jasnovidova conducted short- and long-read RNA-seq time-course experiments during the first five days of the differentiation, providing the transcriptome profiles of the pluripotent, progenitor and neuronal cells. This model system allowed us to address unanswered questions regarding transcription, splicing and neuronal differentiation.

First, we described the landscape of the transcriptional changes, including the quantification of different alternative splicing events at subsequent developmental stages. Next, we addressed the question, whether the changes observed on the transcriptome level have the potential to affect the sequence and properties of the encoded protein. This part resulted in the tool IsoTV (Isoform Transcript Visualizer) [2], which is described in the following section. Additionally, we investigated the causes of the alternative splicing events at different developmental stages, including the potential contribution of Transcription Factors, RNA Binding Proteins, μRNAs and Pol II transcriptional speed. Similarly to the approach presented in Chapter 6, we intend to train a machine learning model predicting the occurrence of the alternative splicing event based on the presence of binding motifs and NET-seq signal. This approach will allow us to quantify the contributions of various factors in regulating alternative splicing. The computational part of the project was conducted by the author, supported by Siddharth Annaldasula.

8.2 QUANTIFICATION AND VISUALIZATION OF THE CODING POTENTIAL OF MRNA ISOFORMS DETECTED BY ONT LONG-READ SEQUENCING

During our work on transcript isoform changes in neuronal differentiation, we were interested in the functional differences between isoforms originating from the same gene. In that project, we used long-read RNA sequencing, which allowed us to detect non-canonical transcript isoforms. However, the computational tools for protein isoform analysis and visualization, especially regarding novel isoforms, were missing. This observed need resulted in the creation of the tool IsoTV (Isoform Transcript Visualizer) that was published as an application note in Bioinformatics [2].

IsoTV was implemented in the form of a versatile Snakemake pipeline. The pipeline consists of two main parts: (1) processing the raw ONT long-reads to *de novo* assemble a transcriptome and quantify isoform expression, and (2) translating the obtained transcript isoforms into predicted protein isoforms and visualizing their functional features. The first part was developed by the author to process raw or basecalled ONT reads in order to *de novo* assemble the transcriptome. This sub-workflow was inspired by ONT's long-read processing pipeline and includes basecalling the raw reads, reconstructing the transcriptome comprehensively for all samples and quantification of the transcript expression. The second part was conceptualized together with Siddarth Annaldasula and implemented by him. It consists of the transcript isoform translation accounting for the presence of upstream open reading frames (uORFs) and the protein isoform visualization. The visualization module incorporates various tools to predict protein domains, secondary structure, disordered regions and post-translational modification sites. The final product of the pipeline is a consolidated report generated for chosen input genes. It consists of intuitive visualizations depicting gene and isoform expression, transcript composition and functional features of the translated transcript isoforms. Finally, we demonstrated the functionality of IsoTV on cancer cell lines sequenced using ONT long-reads.

8.3 COMPARING FUNCTIONS OF HUMAN TFIIS PARALOGS USING MULTIOMICS DATA

In Chapter 5 we described the pause detection algorithm developed for sparse data. This peak caller is intended to be used for a currently ongoing project led by Yelizaveta Mochalova. The project focuses on the role of TFIIS paralogs in transcription regulation. TFIIS, which stands for Transcription elongation Factor IIS, is a transcription factor rescuing polymerases stalled as a consequence of a reverse-translocation movement called backtracking. As a result of backtracking, RNA 3'end is mislocalized to a pore in Pol II, effectively preventing elongation. TFIIS stimulates the cleavage of the backtracked Pol II, realigning the nascent RNA with the DNA template and enabling the polymerase to resume transcription [59]. There are four paralogs of TFIIS showing various expression levels in different cell lines and tissues [14]. Our main aim is to explore functional dissimilarities of the human TFIIS paralogs.

To find the differences between the human TFIIS paralogs, Yelizaveta created mutant cell lines exhibiting abnormal expression of TFIIS-encoding genes. The modified cell lines were derived from immortalized human embryonic kidney cells (HEK293) and included cell lines overexpressing one of the main TFIIS paralogs, namely genes TCEA1 or TCEA2, and cell lines, in which one or both of these genes were deleted. These modified cell lines were used to provide an extensive characterization of the resulting phenotypes. The characterization included profiling transcriptome using RNA-seq, describing Pol II occupancy using NET-seq and finding TFIIS binding sites and interactome of the paralogs using ChIP-seq and IP-MS respectively. The author's contribution comprised analyses of the data and integration of these multiomics approaches. Unexpectedly, we found that deletion of the TCEA2 gene, whose expression was described as testis-specific [74], caused disregulation of a larger number of genes than the deletion of ubiquitously expressed TCEA1. Additionally, the pausing detection algorithm will be

applied to better understand, how the unresolved backtracking affects gene expression.

[1] Bruce Alberts *et al. Podstawy Biologii komórki*. Wydawnictwo Naukowe PWN, 2015.

[2] Siddharth Annaldasula, Martyna Gajos, and Andreas Mayer. "IsoTV: processing and visualizing functional features of translated transcript isoforms." In: *Bioinformatics* (2021). ISSN: 1367-4803.

[3] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: A practical and powerful approach to multiple testing." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 289–300.

[4] Carrie Bernecky, Jürgen M Plitzko, and Patrick Cramer. "Structure of a transcribing RNA polymerase II-DSIF complex reveals a multidentate DNA-RNA clamp." In: *Nat. Struct. Mol. Biol.* 24.10 (2017), pp. 809–815.

[5] Aleksandra Bochkareva, Yulia Yuzenkova, Vasisht R Tadigotla, and Nikolay Zenkin. "Factor-independent transcription pausing caused by recognition of the RNA-DNA hybrid sequence." In: *EMBO J.* 31.3 (2012), pp. 630–639.

[6] Morgane Boone, Andries De Koker, and Nico Callewaert. "Capturing the 'ome': the expanding molecular toolbox for RNA and DNA library construction." In: *Nucleic Acids Research* 46.6 (2018), pp. 2701–2721.

[7] Leo Breiman. "Random Forests." In: *Machine Learning* 45.1 (2001), pp. 5–32.

[8] T W Burke and J T Kadonaga. "The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila." In: *Genes Dev.* 11.22 (1997), pp. 3020–3031.

[9] Jaime Abraham Castro-Mondragon, Sébastien Jaeger, Denis Thieffry, Morgane Thomas-Chollier, and Jacques van Helden. "RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections." In: *Nucleic Acids Res.* 45.13 (2017), e119.

[10] Regina Z. Cer *et al.* "Non-b db v2.0: A database of predicted non-b dna-forming motifs and its associated tools." In: *Nucleic Acids Research* 41.D1 (2012).

[11] Alan Cheung and Patrick Cramer. "A movie of RNA Polymerase II transcription." In: *Cell* 149.7 (2012), pp. 1431–1437.

[12] Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. "DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding." In: *Bioinformatics* 32.8 (2016), pp. 1211–1213.

[13] L Stirling Churchman and Jonathan S Weissman. "Nascent transcript sequencing visualizes transcription at nucleotide resolution." In: *Nature* 469.7330 (2011), pp. 368–373.

[14] GTEx Consortium. "Genetic effects on gene expression across human tissues." In: *Nature* 550.7675 (2017), pp. 204–213.

[15] Leighton J. Core, Joshua J. Waterfall, and John T. Lis. "Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters." In: *Science* 322.5909 (2008), pp. 1845–1848.

[16] Leighton Core and Karen Adelman. "Promoter-proximal pausing of RNA polymerase II: A NEXUS of gene regulation." In: *Genes & Development* 33.15-16 (2019), pp. 960–982.

[17] P Cramer, D A Bushnell, and R D Kornberg. "Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution." In: *Science* 292.5523 (2001), pp. 1863–1876.

[18] Kushal K. Dey, Dongyue Xie, and Matthew Stephens. "A new sequence logo plot to highlight enrichment and depletion." In: *BMC Bioinformatics* 19.473 (2018).

[19] P. D'haeseleer. "What are DNA sequence motifs?" In: *Nature Biotechnology* 24 (2006), pp. 423–425.

[20] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. *All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously*. 2019. arXiv: 1801.01489 [stat.ME].

[21] Martyna Gajos, Olga Jasnovidova, Alena van Bömmel, Susanne Freier, Martin Vingron, and Andreas Mayer. "Conserved DNA sequence features underlie pervasive RNA polymerase pausing." In: *Nucleic Acids Research* 49.8 (2021), pp. 4402–4420.

[22] D S Gilmour and J T Lis. "RNA polymerase II interacts with the promoter region of the noninduced HSP70 gene in Drosophila Melanogaster cells." In: *Molecular and Cellular Biology* 6.11 (1986), pp. 3984–3989.

[23] Girolamo Giudice, Fátima Sánchez-Cabo, Carlos Torroja, and Enrique Lara-Pezzi. "ATtRACT-a database of RNA-binding proteins and associated motifs." In: *Database* 2016 (2016).

[24] Mark E. Glickman, Sowmya R. Rao, and Mark R. Schultz. "False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies." In: *Journal of Clinical Epidemiology* 8 (2014), pp. 850–857.

[25] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[26] Saskia Gressel, Björn Schwalb, Tim Michael Decker, Weihua Qin, Heinrich Leonhardt, Dirk Eick, and Patrick Cramer. "CDK9-dependent RNA polymerase II pausing controls transcription initiation." In: *Elife* 6 (2017).

[27] Sveinung Gundersen, Matúš Kalaš, Osman Abul, Arnoldo Frigessi, Eivind Hovig, and Geir Sandve. "Identifying elemental genomic track types and representing them uniformly." In: *BMC Bioinformatics* 12.1 (2011), p. 494.

[28] Kevin M Harlen, Kristine L Trotta, Erin E Smith, Mohammad M Mosaheb, Stephen M Fuchs, and L Stirling Churchman. "Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue." In: *Cell Rep.* 15.10 (2016), pp. 2147–2158.

[29] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer, 2009, pp. I–XXII, 1–745. ISBN: 9780387848570.

[30] Anne Helmrich, Monica Ballarino, and Laszlo Tora. "Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes." In: *Molecular Cell* 44.6 (2011), pp. 966–977.

[31] David A Hendrix, Joung-Woo Hong, Julia Zeitlinger, Daniel S Rokhsar, and Michael S Levine. "Promoter elements associated with RNA Pol II stalling in the Drosophila embryo." In: *Proc. Natl. Acad. Sci. U. S. A.* 105.22 (2008), pp. 7762–7767.

[32] Ewa Heyduk and Tomasz Heyduk. "DNA template sequence control of bacterial RNA polymerase escape from the promoter." In: *Nucleic Acids Research* 46.9 (2018), pp. 4469–4486.

[33] Jirí Hon, Tomáš Martínek, Jaroslav Zendulka, and Matej Lexa. "pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R." In: *Bioinformatics* 33.21 (2017), pp. 3373–3379.

[34] Masahiko Imashimizu, Hiroki Takahashi, Taku Oshima, Carl McIntosh, Mikhail Bubunenko, Donald L Court, and Mikhail Kashlev. "Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo." In: *Genome Biol.* 16 (2015), p. 98.

[35] Alan Julian Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. 1st ed. Springer Publishing Company, Incorporated, 2008.

[36] Olga V Kel-Margoulis, Dmitri Tchekmenev, Alexander E Kel, Ellen Goessling, Klaus Hornischer, Birgit Lewicki-Potapov, and Edgar Wingender. "Composition-sensitive analysis of the human genome for regulatory signals." In: *In Silico Biol.* 3.1-2 (2003), pp. 145–171.

[37] Aziz Khan *et al.* "JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework." In: *Nucleic Acids Res.* 46.D1 (2018), pp. D260–D266.

[38] Peter Kindgren, Maxim Ivanov, and Sebastian Marquardt. "Native elongation transcript sequencing reveals temperature dependent dynamics of nascent RNAPII transcription in Arabidopsis." In: *Nucleic Acids Res.* 48.5 (2020), pp. 2332–2347.

[39] Janne Korhonen, Petri Martinmäki, Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen. "MOODS: fast search for position weight matrix matches in DNA sequences." In: *Bioinformatics* 25.23 (2009), pp. 3181–3182.

[40] A. R. Kornblihtt. "Multiple links between transcription and splicing." In: *RNA* 10.10 (2004), pp. 1489–1498.

[41] Tomoya Kujirai, Haruhiko Ehara, Yuka Fujino, Mikako Shirouzu, Shun-ichi Sekine, and Hitoshi Kurumizaka. "Structural basis of the nucleosome transition during RNA polymerase II passage." In: *Science* 362.6414 (2018), pp. 595–598.

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

[43] Matthew H Larson, Rachel A Mooney, Jason M Peters, Tricia Windgassen, Dhananjaya Nayak, Carol A Gross, Steven M Block, William J Greenleaf, Robert Landick, and Jonathan S Weissman. "A pause sequence enriched at translation start sites drives transcription dynamics in vivo." In: *Science* 344.6187 (2014), pp. 1042–1047.

[44] N Le Novère. "MELTING, computing the melting temperature of nucleic acid duplex." In: *Bioinformatics* 17.12 (2001), pp. 1226–1227.

[45] Dong-Hoon Lee, Naum Gershenzon, Malavika Gupta, Ilya P Ioshikhes, Danny Reinberg, and Brian A Lewis. "Functional characterization of core promoter elements: the downstream core element is recognized by TAF1." In: *Mol. Cell. Biol.* 25.21 (2005), pp. 9674–9686.

[46] B A Lewis, T K Kim, and S H Orkin. "A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts." In: *Proc. Natl. Acad. Sci. U. S. A.* 97.13 (2000), pp. 7172–7177.

[47] Xiao Li and Xiang-Dong Fu. "Chromatin-associated RNAS as facilitators of functional genomic interactions." In: *Nature Reviews Genetics* 20.9 (2019), pp. 503–519.

[48] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. "ViennaRNA Package 2.0." In: *Algorithms Mol. Biol.* 6 (2011), p. 26.

[49] E. Louie, J. Ott, and J. Majewski. "Nucleotide frequency variation across human genes." In: *Genome Research* 13.12 (2003), pp. 2594–2601.

[50] Andreas Mayer, Julia di Iulio, Seth Maleri, Umut Eser, Jeff Vierstra, Alex Reynolds, Richard Sandstrom, John A Stamatoyannopoulos, and L Stirling Churchman. "Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution." In: *Cell* 161.3 (2015), pp. 541–554.

[51] Andreas Mayer and L Stirling Churchman. "Genome-wide profiling of RNA polymerase transcription at nucleotide resolution in human cells with native elongating transcript sequencing." In: *Nature Protocols* 11.4 (2016), pp. 813–833.

[52] Andreas Mayer, Heather M Landry, and L Stirling Churchman. "Pause & go: from the discovery of RNA polymerase pausing to its functional implications." In: *Curr. Opin. Cell Biol.* 46 (2017), pp. 72–80.

[53] Matthieu Meryet-Figuiere, Babak Alaei-Mahabadi, Mohamad Moustafa Ali, Sanhita Mitra, Santhilal Subhash, Gaurav Kumar Pandey, Erik Larsson, and Chandrasekhar Kanduri. "Temporal separation of replication and transcription during S-phase progression." In: *Cell Cycle* 13.20 (2014), pp. 3241–3248.

[54] Gracjan Michlewski and Javier F. Cáceres. "Post-transcriptional control of mirna biogenesis." In: *RNA* 25.1 (2018), pp. 1–16.

[55] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/. 2019.

[56] Mariangela Morlando, Monica Ballarino, Natalia Gromak, Francesca Pagano, Irene Bozzoni, and Nick J Proudfoot. "Primary microrna transcripts are processed co-transcriptionally." In: *Nature Structural & Molecular Biology* 15.9 (2008), pp. 902–909.

[57] Harvey Motulsky. *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. Third. Oxford University Press, 2013.

[58] S. Nechaev, D. C. Fargo, G. dos Santos, L. Liu, Y. Gao, and K. Adelman. "Global analysis of short RNAS reveals widespread promoter-proximal stalling and arrest of Pol II in drosophila." In: *Science* 327.5963 (2009), pp. 335–338.

[59] Melvin Noe Gonzalez, Daniel Blears, and Jesper Q. Svejstrup. "Causes and consequences of RNA polymerase II stalling during transcript elongation." In: *Nature Reviews Molecular Cell Biology* 22.1 (2020), pp. 3–21.

[60] Allison Piovesan, Maria Pelleri, Francesca Antonaros, Pierluigi Strippoli, Maria Caracausi, and Lorenza Vitale. "On the length, weight and GC content of the human genome." In: *BMC Research Notes* 12 (2019).

[61] Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. "Deeptools2: A next generation web server for deep-sequencing data analysis." In: *Nucleic Acids Research* 44.W1 (2016).

[62] Arthur L. Samuel. "Some studies in machine learning using the game of Checkers." In: *IBM Journal of Research and Development* (1959), pp. 71–105.

[63] Wanqing Shao, Sergio G-M Alcantara, and Julia Zeitlinger. "Reporter-ChIP-nexus reveals strong contribution of the initiator sequence to RNA polymerase pausing." In: *Elife* 8 (2019).

[64] Haridha Shivram and Vishwanath R Iyer. "Identification and removal of sequencing artifacts produced by mispriming during reverse transcription in multiple RNA-seq technologies." In: *RNA* 24.9 (2018), pp. 1266–1274.

[65] Stefan Sigurdsson, A. Barbara Dirac-Svejstrup, and Jesper Q. Svejstrup. "Evidence that transcript cleavage is essential for RNA polymerase II transcription and cell viability." In: *Molecular Cell* 38.2 (2010), pp. 202–210.

[66] Vikram Singh and Pawan K. Dhar. *Systems and synthetic biology*. Springer, 2015.

[67] T A Steitz. "A mechanism for all polymerases." In: *Nature* 391.6664 (1998), pp. 231–232.

[68] Chun-Hao Su, Dhananjaya D, and Woan-Yuh Tarn. "Alternative splicing in neurogenesis and brain development." In: *Frontiers in Molecular Biosciences* 5 (2018).

[69] Karol Szlachta, Ryan G Thys, Naomi D Atkin, Levi C T Pierce, Stefan Bekiranov, and Yuh-Hwa Wang. "Alternative DNA secondary structure formation affects RNA polymerase II promoter-proximal pausing in human." In: *Genome Biol.* 19.1 (2018), p. 89.

[70] Long Vo Ngoc, Cassidy Yunjing Huang, California Jack Cassidy, Claudia Medrano, and James T Kadonaga. "Identification of the human DPR core promoter element using machine learning." In: *Nature* 585.7825 (2020), pp. 459–463.

[71] Seychelle M Vos, Lucas Farnung, Henning Urlaub, and Patrick Cramer. "Structure of paused transcription complex Pol II-DSIF-NELF." In: *Nature* 560.7720 (2018), pp. 601–606.

[72] Irina O Vvedenskaya, Hanif Vahedian-Movahed, Jeremy G Bird, Jared G Knoblauch, Seth R Goldman, Yu Zhang, Richard H Ebright, and Bryce E Nickels. "Interactions between RNA polymerase and the "core recognition element" counteract pausing." In: *Science* 344.6189 (2014), pp. 1285–1289.

[73] Jason A Watts, Joshua Burdick, Jillian Daigneault, Zhengwei Zhu, Christopher Grunseich, Alan Bruzel, and Vivian G Cheung. "cis Elements that Mediate RNA Polymerase II Pausing Regulate Human Gene Expression." In: *Am. J. Hum. Genet.* 105.4 (2019), pp. 677–688.

[74] Zoe A. Weaver and Caroline M. Kane. "Genomic characterization of a testis-specific TFIIS (TCEA2) gene." In: *Genomics* 46.3 (1997), pp. 516–519.

[75] CH Wu, Y Yamaguchi, LR Benjamin, M Horvat-Gordon, J Washinsky, E Enerly, J Larsson, A Lambertsson, H Handa, and D. Gilmour. "NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in drosophila." In: *Genes & Development* 17.11 (2003), pp. 1402–1414.

[76] Zhang Y *et al.* "Model-based Analysis of ChIP-Seq (MACS)." In: *Genome Biology* 9.9 (2008), R137.

[77] B. Zamft, L. Bintu, T. Ishibashi, and C. Bustamante. "Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases." In: *Proceedings of the National Academy of Sciences* 109.23 (2012), pp. 8948–8953.

LIST OF TABLES

Table S1: Tools and parameters used for preprocessing NET-seq data.

Step	**Tool version**	**Parameters**
adapter removal	cutadapt v3.4	-a ATCTCGTATGCCGTCTTCTGCTTG -a AAAAAAAAAAGGGGGGGGGGGGG -a GGGGGGGGGGGGGGGGGGGGGG -e 0.2 -q 5 –max-n 0.9
collapse of PCR duplicates	starcode v1.3	-d 0
read mapping	STAR 2.7.3a	outFilterMultimapNmax 1 clip3pAdapterSeq ATCTCGTATGCCGTCTTCTGCTTG clip3pAdapterMMp 0.21 clip3pAfterAdapterNbases 1 outSJfilterOverhangMin 3 1 1 1 outSJfilterDistToOtherSJmin 0 0 0 0 alignIntronMin 11 alignEndsType EndToEnd
others	custom Python & bash scripts	available at Mayer lab's github: https://github.molgen.mpg.de/MayerGroup

Table S2: The number of sites and predictor variables used for machine learning modeling.

Species	**NET-seq adaptation**	**Genomic region**	**Number of sites n**	**Number of genomic features p**
H. sapiens		promoter-proximal	5244	670
H. sapiens		gene-body	15524	670
H. sapiens	HiS-NET-seq	promoter-proximal	23188	670
H. sapiens	HiS-NET-seq	gene-body	12838	670
E. coli		gene	157192	39
S. cervisea		gene	52500	39
A. thaliana		gene	5912	39

Table S3: List of features included in machine learning models predicting pausing sites in human genome.

Category	**Features**
Skewness	AT-, AC-, AG-, CT-, GC-, GT- skewness
Nucleotide identity	At positions +1, -1, -2, -3, -10, -11
Thermodynamics	Entropy, enthalpy, Gibbs free energy, and melting temperature
RNA hairpin	Minimum free energy of nascent RNA fragment
DNA shape	Minor Groove Width, Roll, Propeller Twist, Helix Twist, Potential Energy
DNA structures	Z-DNA, A-phased repeats, inverted repeats, mirror repeats, direct repeats, G-quadruplexes
Transcription factor binding motifs	111 consensus motifs of 639 human transcription factors
RNA binding protein motifs	240 consensus motifs of 160 human RNA binding proteins

Table S4: List of features included in machine learning models predicting pausing sites in non-human model organism genomes.

Category	**Features**
Skewness	AT-, AC-, AG-, CT-, GC-, GT- skewness
Nucleotide identity	At positions +1, -1, -2, -3, -10, -11
Thermodynamics	Entropy, enthalpy, Gibbs free energy, and melting temperature
RNA hairpin	Minimum free energy of nascent RNA fragment

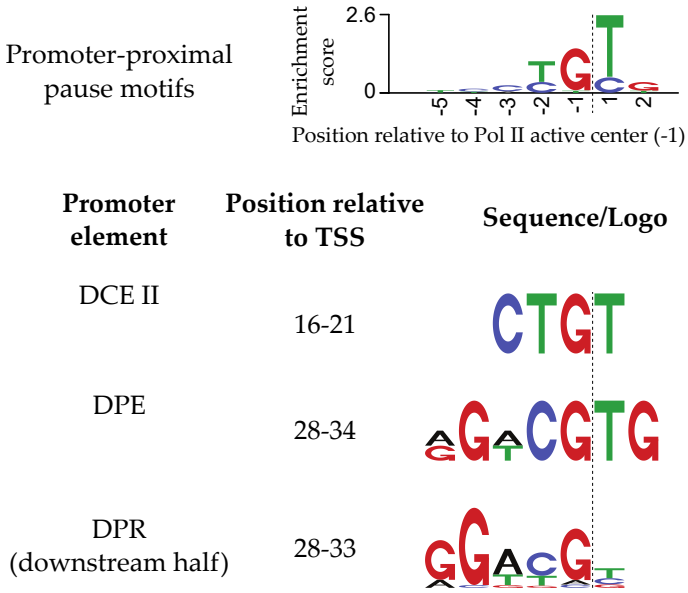

Figure S1: **Comparison of the promoter-proximal Pol II pausing motif (top row) with core promoter elements (following rows)**. For the downstream core promoter element (DPE) and the DPR core promoter element consensus DNA sequences are shown, whereas for the downstream core element (DCE) II only the most frequently appearing nucleotides are shown.

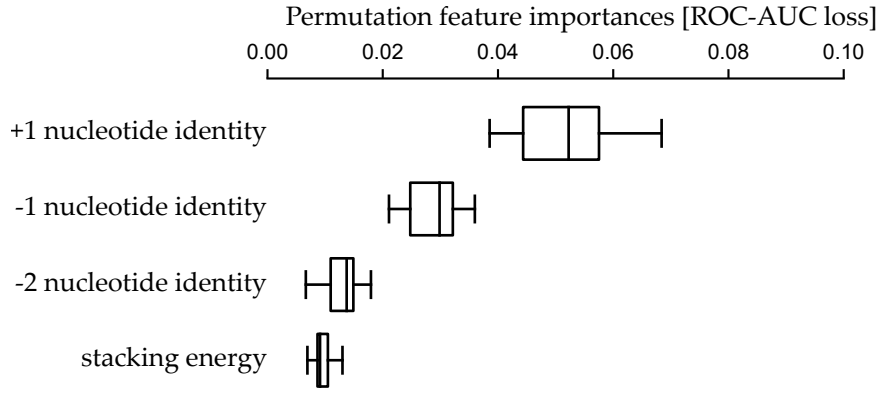

Figure S2: **Permutation feature importance calculated for the logistic regression using promoter-proximal pausing sites.** The distribution of the Permutation Importance was computed for each feature by permuting the feature 10 times. Only features with non-zero importance, as indicated by a *t*-test are plotted.

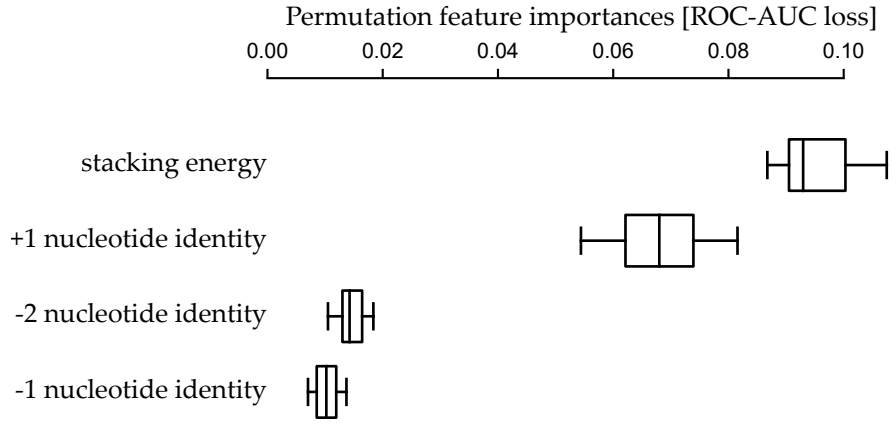

Figure S3: **Permutation feature importance calculated for the gradient boosting classifier using promoter-proximal pausing sites.** The distribution of the Permutation Importance was computed for each feature by permuting the feature 10 times. Only features with non-zero importance, as indicated by a *t*-test are plotted.

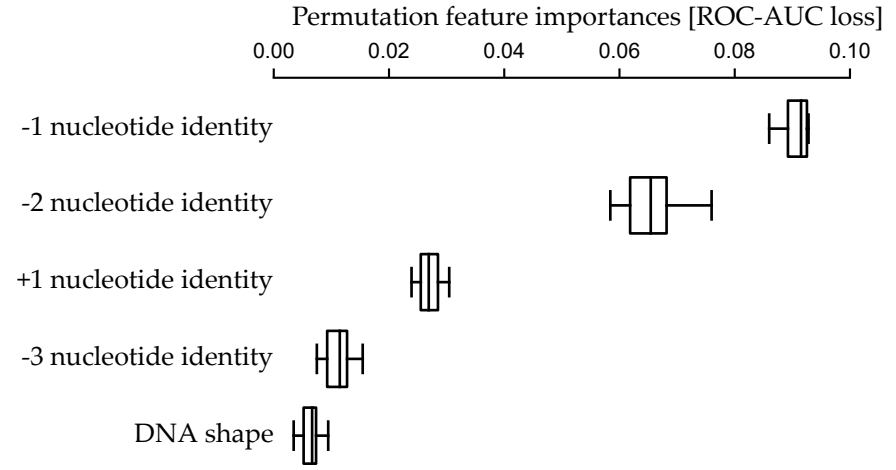

Figure S4: **Permutation feature importance calculated for the logistic regression using gene-body pausing sites detected in HiS-NET-seq.** The distribution of the Permutation Importance was computed for each feature by permuting the feature 10 times. Only features with non-zero importance, as indicated by a *t*-test are plotted.

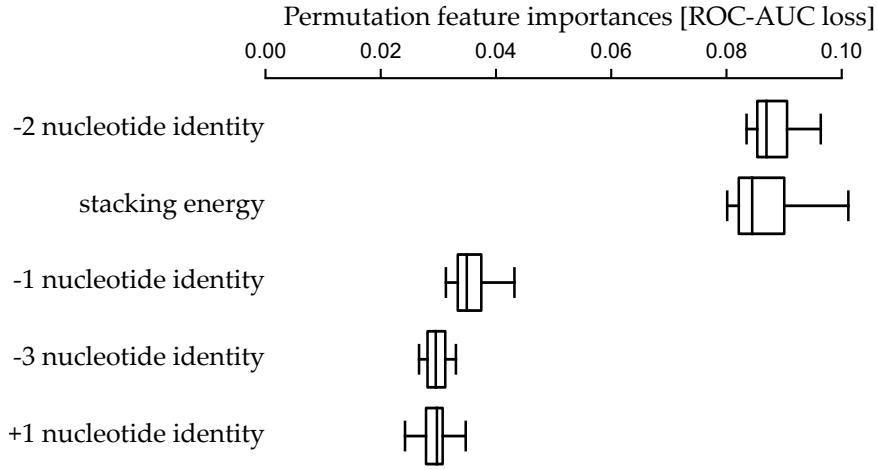

Figure S5: **Permutation feature importance calculated for the gradient boosting classifier using gene-body pausing sites detected in HiS-NET-seq.** The distribution of the Permutation Importance was computed for each feature by permuting the feature 10 times. Only features with non-zero importance, as indicated by a *t*-test are plotted.

ZUSAMMENFASSUNG

Ein allgemeines Merkmal der Genexpression in menschlichen Zellen ist das Pausieren der RNA Polymerase II (Pol II). Verschiedene Aspekte wie Transkriptionsfaktoren, DNA Sequenzen und Eigenschaften des Chromatins werden mit dem Prozess in Verbindung gebracht. Der relative Beitrag dieser Faktoren zur Entstehung der beobachteten Pausen ist unbekannt. Darüber hinaus hat sich die bisherige Forschung bei Metazoen hauptsächlich auf Pol II Pausen während der frühen Elongationsphase, im promoter-proximalen Bereich, konzentriert. Die Ursachen für das Pausieren außerhalb dieser Regionen sind unbekannt.

Um das Verständnis der Ursachen von Transkriptionspausen zu verbessern, haben wir einen Algorithmus entwickelt, der Pol II Signale verarbeitet und Pausen präzise bis auf ein einzelnes Nukleotid lokalisiert. Die Pol II Signalmessungen werden mithilfe von NET-seq (Native Elongating Transcript Sequencing), einer hochauflösenden Methode, erstellt. Bei der Untersuchung der Methode identifizierten wir systematische Fehler in den Messdaten, welche zur Anpassung bei der Datenverarbeitung führte. Diese algorithmischen Verbesserungen zeigten, dass Pol II Pausen in menschlichen Zellen weit verbreitet sind und verteilt über das gesamte Genom, an einzelnen Nukleotiden, beobachtet werden können.

Für eine unvoreingenommene Identifizierung der Sequenzspezifischen Faktoren, die zum Pausieren der Pol II beitragen, wurden eine Reihe von Methoden des maschinellen Lernens angewandt. Mit hoher Sicherheit detektierte Transkriptionspausen wurden genutzt, um Prädispositionen in DNA-Abschnitten zu lernen und vorherzusagen. Für jedes dieser Beispiel Regionen werden beschreibende Merkmale erstellt. Darunter befinden sich Faktoren, die zuvor mit Transkriptionspausen in Verbindung gebracht wurden, sowie Merkmale ohne bekannte Assoziation. Unsere Analyse identifiziert ein neues DNA Sequenzmotiv und andere relevante Sequenzeigenschaften, welche dem pausieren der Pol II zugrunde liegen. Interessanterweise sind die identifizierten Sequenzeigenschaften sowohl in menschlichen Zellen als auch in Bakterien zu finden. Unsere Studie deutet darauf hin, dass Transkriptionspausen in menschlichen Zellen sequenzabhängig und evolutionär konserviert sind.

ABSTRACT

Pausing of transcribing RNA polymerase II (Pol II) has emerged as a general feature of gene expression in human cells. Many transcription factors, DNA sequences and chromatin characteristics have been implicated in inducing transcriptional pausing. However, it is unclear what are the relative contributions of these factors on the observed Pol II pausing. Furthermore, research in metazoans has mainly focused on Pol II promoter-proximal pausing, leaving the causes of pausing outside of this region unknown.

To reliably detect real transcriptional pausing sites and advance the understanding of the causes of this phenomenon, we developed a pausing detection algorithm for nucleotide-resolution Pol II occupancy data. We scrutinized the characteristics and potential shortcomings of Native Elongating Transcript sequencing (NET-seq), which is one of the high-resolution methods of Pol II profiling, and we used our observations to improve the NET-seq processing pipeline. Leveraging the improved processing pipeline and the developed pausing detection algorithm revealed widespread genome-wide Pol II pausing at a nucleotide resolution in human cells.

Next, we set out to identify the determinants of Pol II pausing in an unbiased manner based on the underlying DNA sequence. To predict the predisposition of a genomic site to evoke Pol II pausing, we applied a range of machine learning approaches using previously identified high-confidence pausing sites. For each of the sites, we created a large number of features, including both factors that were previously linked to transcriptional pausing and factors that were not yet implicated in invoking pausing. Our analysis revealed DNA sequence properties underlying widespread Pol II pausing including a new pausing motif. Interestingly, key sequence determinants of RNA polymerase pausing are shared by human cells and bacteria. Our study indicates that transcriptional pausing in human cells is sequence-induced and that the determinants of Pol II pausing might be evolutionary conserved.

DECLARATIONS

Selbstständigkeitserklärung

Name: Gajos
Vorname: Martyna
geb.am:
Matr.Nr.:

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht.
Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Date: Signature:

Martyna Gajos

Declaration of authorship

Name: Gajos
First name: Martyna
Date of birth:
Student-ID:

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

Date: Signature:

Martyna Gajos

SHORT RESUME

For reasons of data protection, the curriculum vitae is not published in the electronic version.