THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

An analytical framework for studying transcriptional regulation

CHRISTOPH SEBASTIAN BÖRLIN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Biology and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020

An analytical framework for studying transcriptional regulation
CHRISTOPH SEBASTIAN BÖRLIN

ISBN 978-91-7905-332-1

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie 4799
ISSN 0346-718X

Division of Systems and Synthetic Biology
Department of Biology and Biological Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Cover: Schematic representation of this thesis

Printed by Chalmers Reproservice
Gothenburg, Sweden 2020

An analytical framework for studying transcriptional regulation

CHRISTOPH SEBASTIAN BÖRLIN
Department of Biology and Biological Engineering
Chalmers University of Technology

ABSTRACT

The state and behavior of any living cell is controlled by a complex interplay of different regulatory processes, with the regulation of transcription playing a major role. When a cell adapts to a new environment it often does that by modulating gene transcript levels, mainly through changes in transcription factor binding events. Therefore, understanding the transcriptional regulation is vital for many biological research fields ranging from understanding cancer metabolism to metabolic engineering.

In this thesis, I present and apply an analytical framework for studying transcriptional regulation in a well-characterized eukaryotic model organism, the yeast *S. cerevisiae*. The framework is a combination of advanced sequencing methods like Chromatin Immunoprecipitation followed by DNA sequencing (ChIP-seq / ChIP-exo) and Cap Analysis of Gene Expression (CAGE) with bioinformatic approaches.

The relative binding location of transcription factors in relation to the transcription start site is important for interpretation, therefore the transcription start sites of all genes active in multiple controlled growth environments were determined using CAGE. To use and analyze the gathered data in a reliable and efficient way a high-quality bioinformatics pipeline was established.

After establishing the required analytical framework, I employed it in various projects, all aimed to gain a better understanding of yeast transcriptional regulation. In a detailed study of a single transcription factor, I investigated Leu3, the main regulator of leucine biosynthesis. Here, I was able to show that its binding behavior is affected by the availability of leucine in the media, an adaptive behavior that has not been reported before.

Metabolic engineering will be increasingly important to support the needs of our society and in order to help with this, I developed a tool for fine tuning conditional gene expression levels using hybrid promoters. This tool is based on a machine learning approach and can be used to improve productivity in large scale fermentations.

In conclusion, this thesis lays the foundation for future large-scale studies of transcriptional regulation in *S. cerevisiae* and can also serve as a blueprint on how to study it in different organisms.

Keywords: *S. cerevisiae,* transcription factor, transcriptional regulation, ChIP-exo

LIST OF PUBLICATIONS

This thesis is based on the work contained in the following papers and manuscripts.

I. **Börlin CS**, Cvetesic N, Holland P, Bergenholm D, Siewers V, Lenhard B & Nielsen J. *Saccharomyces cerevisiae* displays a stable transcription start site landscape in multiple conditions. *FEMS Yeast Research* 2019;**19**.

II. **Börlin CS**, Bergenholm D, Holland P, & Nielsen J. A bioinformatic pipeline to analyze ChIP-exo datasets. *Biology Methods and Protocols* 2019;**4**:1–9.

III. Holland P, Bergenholm D, **Börlin CS**, Liu G, & Nielsen J. Predictive models of eukaryotic transcriptional regulation reveals changes in transcription factor roles and promoter usage between metabolic conditions. *Nucleic Acids Research* 2019;**47**:4986–5000.

IV. **Börlin CS**, Bergenholm D, Kerkhoven EJ, Siewers V & Nielsen J. Analyzing and predicting conditional gene expression changes using transcription factor binding data. *Manuscript*.

V. **Börlin CS**, Nielsen J & Siewers V. The transcription factor Leu3 shows differential binding behavior in response to changing leucine availability. *FEMS Microbiology Letters* 2020;**367**.

VI. Bergenholm D, **Börlin CS**, Holland P & Nielsen J. T-rEx: A *Saccharomyces cerevisiae* transcription factor explorer. *Manuscript*.

CONTRIBUTION SUMMARY

I. I co-designed the study, carried out the experiments, analyzed the data and wrote the manuscript

II. I designed the study, wrote the code, performed the data analysis and wrote the manuscript.

III. I assisted with performing the experiments and the analysis, wrote part of the manuscript.

IV. I designed the study, wrote the code, performed the analysis and wrote the manuscript.

V. I designed the study, performed the experiments and data analysis, wrote the manuscript.

VI. I assisted with writing the scripts and performing the analysis, wrote part of the manuscript.

PREFACE

This dissertation serves as partial fulfillment of the requirements to obtain the degree of Doctor of Philosophy at the Department of Biology and Biological Engineering at the Chalmers University of Technology. The PhD studies were carried out between October 2016 and September 2020 at the Division of Systems and Synthetic Biology (SysBio) under the supervision of Jens Nielsen and co-supervised by Verena Siewers and Eduard Kerkhoven. This thesis was examined by Dina Petranovic.

This thesis was funded by the European Union's Horizon 2020 research and innovation programme [Marie Skłodowska-Curie grant agreement No 722287 (PAcMEN); grant agreement No 720824 (CHASSY)], the Knut and Alice Wallenberg Foundation and the Novo Nordisk Foundation [grant number NNF10CC1016517].

Christoph Sebastian Börlin

September 2020

TABLE OF CONTENTS

1 INTRODUCTION

This thesis is about creating a framework to efficiently study transcriptional regulation. Before diving into how I did this and why it is an interesting field of study with many applications, we must start at the beginning and understand what transcription is.

1.1 WHAT IS TRANSCRIPTION?

In order to explain what transcription is, we have to start at a molecular understanding of what life is and how cellular organism behave and interact with their environments. As this thesis will focus on a yeast species, *Saccharomyces cerevisiae* (also known as baker's yeast), let us take a look at the (simplified) life of a yeast cell used to brew beer. During the brewing process the cell consumes sugar molecules, mainly in the form of glucose from the barley, and converts them into alcohol molecules, more specifically into ethanol. So how is that done? First the glucose has to be imported into the cell by transport proteins (long-chains of amino-acids strung together and folded into three dimensional structures), then the glucose will be converted through many intermediate steps performed by enzymes (a specific class of proteins) into ethanol and finally it will be transported out of the cell.

One can compare this to a miniature factory. First the raw goods have to be delivered to the factory where they are processed by an assembly line of machines (the enzymes) towards the final product that will be shipped to the customer. Besides the enzymes and transport proteins there are many more proteins that play a role in the cell, many involved in regulating the processed and responding to the outside environments. One could think of them as a mix of support utilities like conveyer belts and power lines in the factory, as well as managers talking to customers and suppliers. So now that we have established that proteins are the machines inside the cell that get the work done, the question is how are they produced?

Proteins are produced by a process called translation, where a ribosome (a complex of multiple proteins) reads a messenger RNA (mRNA) molecule and translates the instructions encoded in the mRNA into a chain of amino acids; the protein. One can look at the mRNA as an order for a new machine for our factory that conveniently already includes all necessary instructions in how to build that machine. But where does the mRNA come from and how can the assembly instructions be encoded in there?

Now we are finally getting to the process called transcription. The assembly instruction for every protein is encoded in its gene, which is part of the genome (consisting of DNA) of the organism. In the process called transcription, a selected part of the DNA is transcribed into mRNA. One can interpret the whole DNA as a product catalog of a

machine manufacturer where the customer can pick the page of a machine he or she would like to order. As the mRNA is basically just a one-page copy of that catalog, the process is called transcription (from the verb “to transcribe” meaning to make a written copy of something). Later, the instructions encoded in the mRNA are then translated to obtain the needed amino acid sequence to build that specific protein, hence that step is called translation.

This three-step process from DNA being transcribed to mRNA and then being translated to proteins is also known as the central dogma of molecular biology and a graphical representation is shown in Figure 1, including a comparison to the product ordering analogy I used here to explain it.

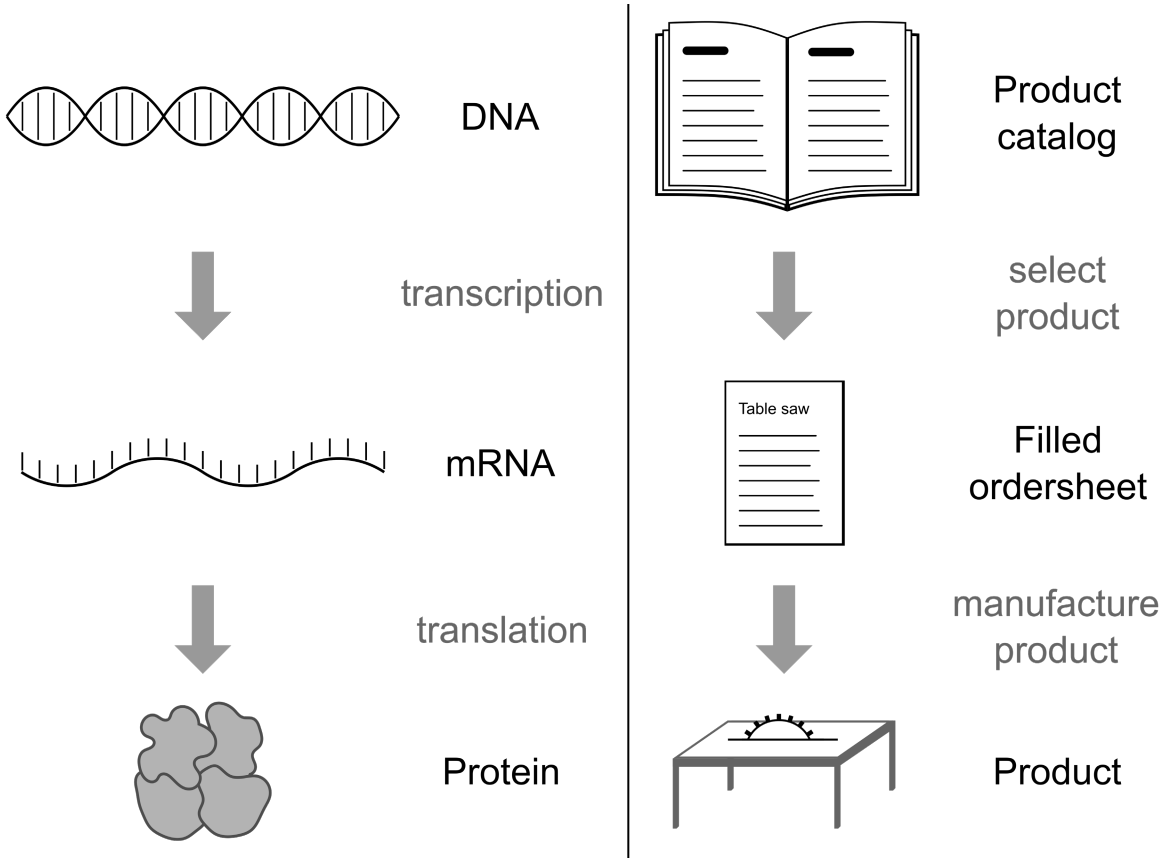

Figure 1: Graphical summary of the transcription and translation process to make proteins starting from DNA.

1.2 HOW IS TRANSCRIPTION REGULATED?

To better understand the whole process of transcription and how it is regulated we first have to dive deeper into the exact molecular mechanism at play. The production of mRNA or any other kind of RNA is done by large enzymes called RNA polymerases, which consist of many subunits (Hahn 2004). In yeast and other eukaryotes all protein-coding genes are transcribed by the RNA polymerase II (Alberts *et al.* 2015).

Transcription is initiated in the core promoter, a stretch of 60-70 base-pairs defined as the minimal stretch of DNA necessary to start transcription from, at least *in vitro* (Danino *et al.* 2015; Haberle and Lenhard 2016). In the process, the RNA polymerase II is recruited to the core promoter by a group of general transcription factors. Together they form the transcription preinitiation complex (PIC) which can then start the transcription at the transcription start site (TSS) (Smale and Kadonaga 2003; Sandelin *et al.* 2007; Hahn and Young 2011).

Unfortunately, this process is not as straightforward *in vivo* as it is *in vitro*. *In vivo* the basal level of transcription from a core promoter is basically zero (Struhl 1999). This means that for transcription to take place, the polymerase recruitment and transcription initiation process needs to be activated, which can be done by various mechanisms. This is often done by members of a class of proteins called transcription factors (TFs) binding to specific activating sequences upstream of the TSS, called upstream activation sequences (UAS) (Struhl 1999).

Before going into more details about what TFs exactly are and how they regulate transcription, we first have to examine why the ground state of a yeast core promoter is inactive *in vivo* while it is active *in vitro*.

DNA is not floating around unstructured in the eukaryotic cell, but highly condensed in a structure called chromatin. Chromatin is made out of nucleosomes, which is a stretch of DNA tightly wrapped around a protein complex consisting of histones (Alberts *et al.* 2015). Because the DNA is wrapped around the histones it is not openly accessible for the PIC to assemble there and initiate transcription. First, the chromatin has to be opened, which is often done through proteins called chromatin remodelers, recruited by the TFs, or by direct binding of pioneer TFs that can open the chromatin structure on their own (Zaret and Carroll 2011). The strong involvement of TFs in chromatin remodeling and opening explains why TFs are often necessary for transcription to occur. TFs are however not the only mechanism by which the chromatin can be remodeled and thereby made accessible. Longer DNA stretches with high ratios of adenine and thymine bases can for example also improve chromatin accessibility by decreasing nucleosome stability (Iyer and Struhl 1995).

After opening the chromatin and thereby making the DNA more accessible, the general transcription factors can be recruited which in turn recruit the RNA polymerase II, thereby assembling the PIC. With a functionally assembled PIC, transcription initiation can finally occur and an overview of the three main stages of transcription initiation is shown in Figure 2.

Now we can go back to the topic of transcription factors and understand what exactly they are and what roles they play in more detail.

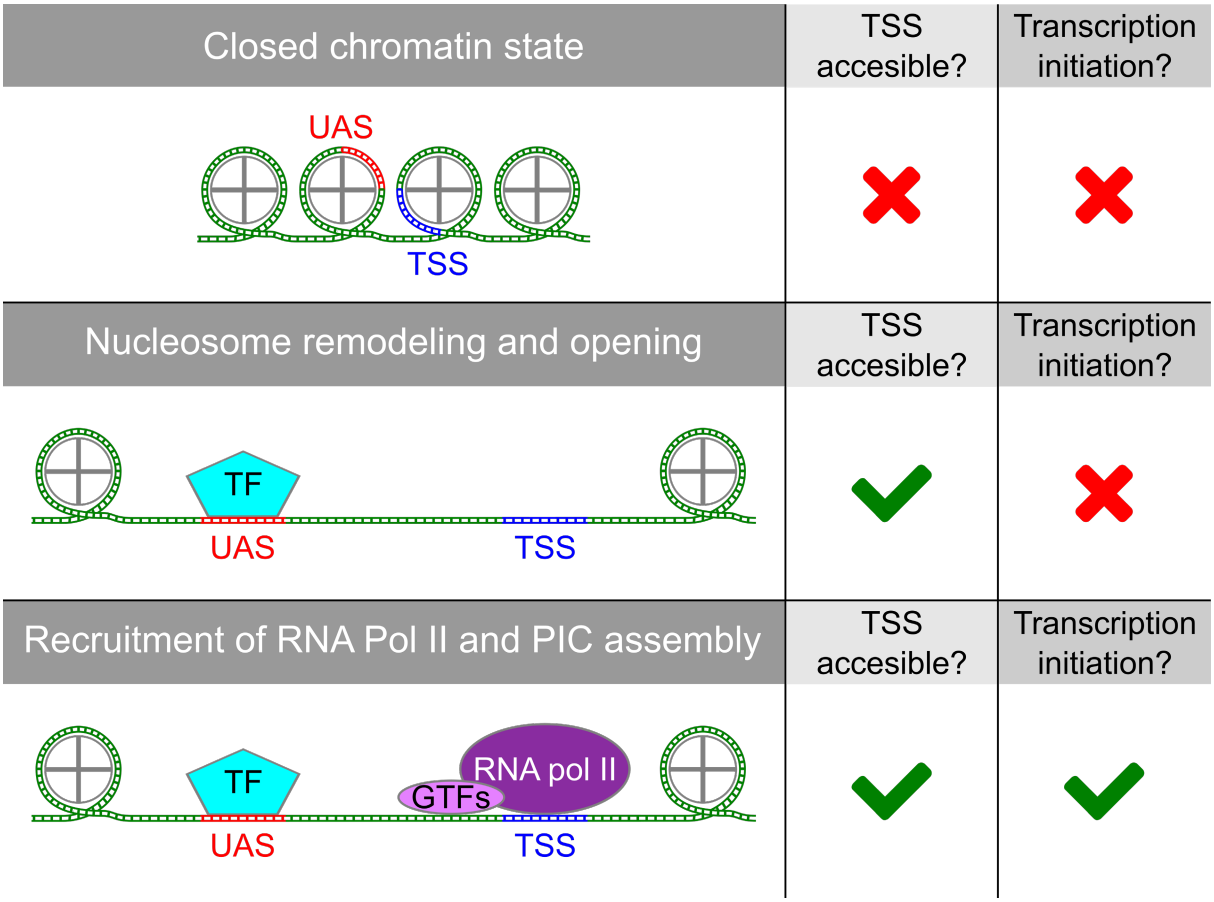

Figure 2: Overview of necessary steps to initiate transcription. UAS = upstream activation sequence, TSS = transcription start site, GTFs = general transcription factors.

1.3 WHAT ARE TRANSCRIPTION FACTORS?

The most common definition says that a TF is a protein that has two distinct characteristics, it is (i) able to bind DNA in a sequence-specific matter and (ii) able to influence the transcription of a gene (either inhibiting or enhancing) (Hughes and de Boer 2013). This definition is however not without issues as there are several edge cases to consider. For example, in yeast there is the case of the transcriptional activator Met4, which is sometimes considered a TF (Hahn and Young 2011), despite the fact that it does not have its own DNA binding domain and relies on Met31 / Met32 or Cbf1 for binding (Carrillo *et al.* 2012). Therefore, there are other definitions out there, leading to the situation that different publications will mention different total number of yeast TFs (mainly in the range 141 to 251 TFs) (Hughes and de Boer 2013).

In this thesis, I will follow the most common definition as also described by Hughes and de Boer in their review, where they state that there are around 209 TFs in *S. cerevisiae* (Hughes and de Boer 2013).

TFs can be classified into different structural groups based on their DNA binding domain (DBD), and an overview of the groups in *S. cerevisiae* and their relative sizes is shown in Figure 3. This is based on data about the DBD of 202 TFs obtained from the YeTFaSco database (De Boer and Hughes 2012). The largest group of TFs is the zinc binding type, which has a DBD that is stabilized by one or more zinc ions (Hahn and Young 2011). Adr1 and Leu3 are both member of this large group, which can be further subdivided into zinc fingers (also called C2H2 zinc fingers), zinc clusters (also called Zn2Cys6 clusters) and GATA fingers (also called C4 fingers). Zinc fingers are common among most eukaryotes while zinc clusters are unique to fungi (Hahn and Young 2011; Hughes and de Boer 2013). The second largest group are the zipper type DBDs, which are characterized having a basic region and a dimerization motif. Examples of this class are Gcn4, Ino2 and Ino4. The zipper type DBDs can be subdivided into basic zippers (bZIP) and basic helix-loop-helix (bHLH) types (Hahn and Young 2011).

The third largest group is the helix-turn-helix (HTH) type that is defined by two alpha helices in the protein structure connected by a short loop region. Members of this group are for example Mcm1 and Hsf1. The main subclass of the HTH group are the Homeodomain TFs, while Forkhead TFs and MADS box TFs are associated with this class due to their high similarity to the Homeodomain DBD (Hahn and Young 2011). Besides these three large groups there are a number of TFs that have their quite unique DBDs and for some TFs the exact DBD is still unknown or not classified.

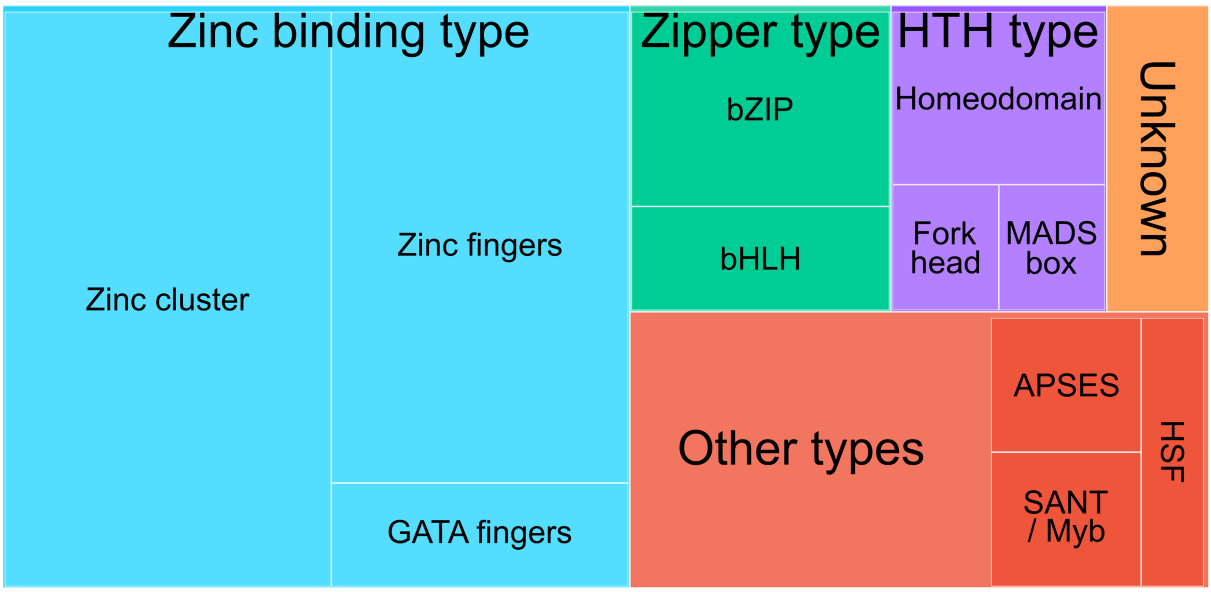

Figure 3: Distribution of TF types and subtypes. 202 non-dubious TFs and their DBD extracted from YeTFaSCo are shown, subtypes with less than 5 members are not shown. Number of TFs in each main group are: Zinc binding type 105; Other types 46; Zipper type 23; HTH type 19; Unknown 9.

Besides their differences in structure and how they bind to DNA, TFs also show a wide range of behaviors when it comes to conditional responses. In their 2004 paper, Harbison *et al*. classified TFs into four groups, as shown in Figure 4. There are TFs that are (i) condition invariant, meaning that they will always bind to the same target genes independent of the environmental condition. One example of this behavior is the TF Put3 (Axelrod, Majors and Brandriss 1991), a TF involved in regulating proline utilization processes. Here I would like to note that Leu3, the example given as a condition invariant TF by Harbison *et al*., is responding to changes in the environment, as I will later show in this thesis. Another group of TFs is (ii) condition responsive, meaning that they are inactive until a trigger activates them. This is the case for many stress responsive TFs, like Msn2 (Schmitt and Mcentee 1996; Estruch 2000). Next, are TFs that are (iii) condition expanded, which means that they have a subset of their possible targets that they always or most of the time bind to and then - once activated - they expand their target range. This behavior can for example be seen in Gcn4, the main regulator of amino acid metabolism, which under amino acid starvation binds to many more targets than in the presence of sufficient amino acid levels (Albrecht *et al.* 1998). The last group are (iv) condition altered TFs, which bind to different target genes in response to the environment. The targets do not have to be exclusive for that condition, there can be a conserved core-response of that TF. An example would be Ste12, which can either promote a mating phenotype or filamentous growth, partially depending on its binding partners (Zeitlinger *et al.* 2003).

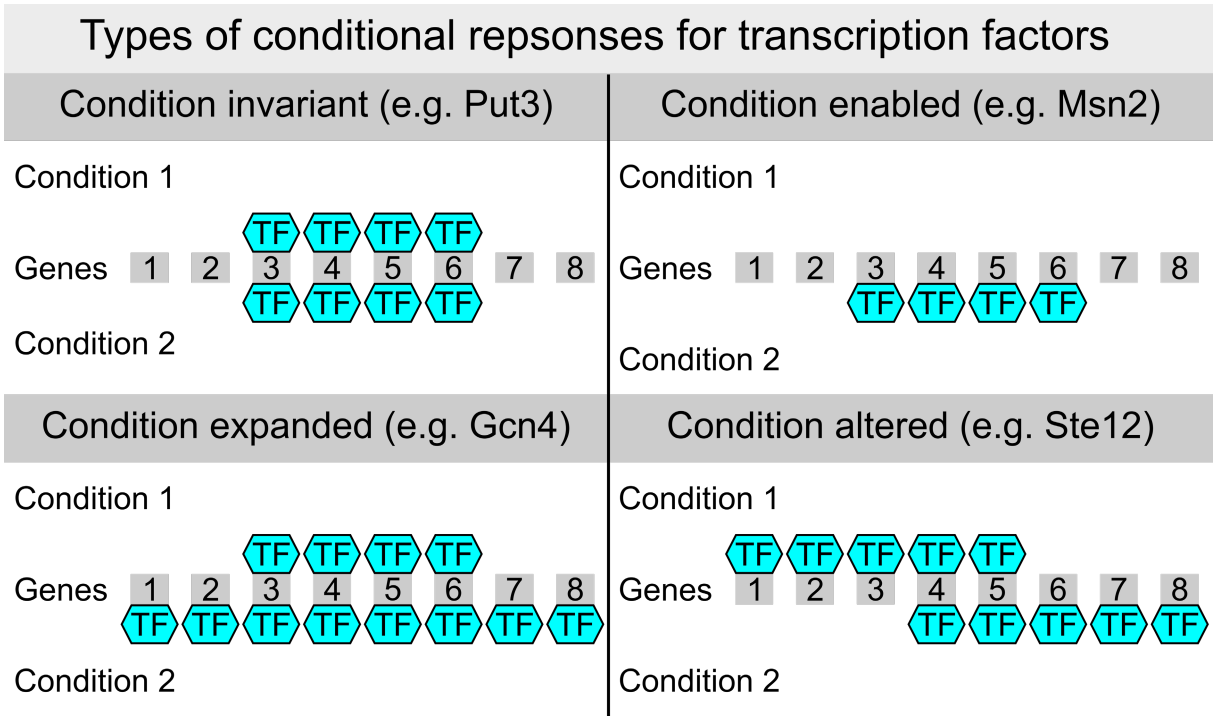

Figure 4: Overview of different types of conditional response from transcription factors. For each type of conditional response, the binding behavior is shown for two conditions on a set of example genes. Condition 1 is shown as binding above the gene boxes, condition 2 is shown as binding below the gene boxes.

As shown above, TFs can exhibit a large range of gene targeting behavior and how the targets change depends on the cellular environment. But how are TFs regulated themselves so that they can respond to changes in the environment?

There are many different molecular mechanisms by which TFs are regulated and the major ones are shown in Figure 5. The probably most intuitive regulation is through changing the protein level of the TF, which can be achieved by either increasing the transcription rate or the translation efficiency. Even though this seems like a very straight forward mechanism, how this is exactly done can nevertheless be quite convoluted. An excellent example of this is the TF Gcn4, whose translational efficiency depends on the overall availability of amino acids. This is achieved through the presence of four short open reading frames upstream of the Gcn4 reading frame that are transcribed together (Mueller and Hinnebusch 1986). If the amino acid levels are not sufficient the translation efficiency is increased by enabling more ribosomes to either read through these short open reading frames or reinitiate at the Gcn4 coding start site. This results in an increase in Gcn4 levels and thereby enabling the TF to induce the expression of amino-acid synthesis genes (Hinnebusch 1988; Hinnebusch and Natarajan 2002).

Gcn4 is also interesting because it does not only respond by an increase in translational efficiency but also by an subsequent increase in transcription rate to further increase its protein levels (Albrecht *et al.* 1998).

Another way to regulate TFs is through protein-protein interactions, for example in the case of the TF Gal4 that is inhibited when it is bound by Gal80 (Egriboz *et al.* 2013), or by phosphorylation of the TF, for example Adr1, which is inactive when phosphorylated (Cherry *et al.* 1989). Binding of the TF to metabolites is an alternative way to control TF activity levels; for example Leu3, the main regulator of leucine metabolism is activated once it is bound to alpha-isopropylmalate, an intermediate in the leucine synthesis pathway (Kohlhaw 2003). The activity of TFs can also be regulated by influencing their concentration in the nucleus, either by targeted import or export mechanisms. Examples for this are the TFs Rtg1 and Rtg3 that are imported into the nucleus when the cells are grown in media containing urea or ammonia (Komeili *et al.* 2000).

I would also like to note that these different mechanisms are not exclusive, and many TFs are regulated by a combination of them. This is especially the case for TFs that are regulated by nuclear exclusion and localization, as this is often achieved through phosphorylation events, that either target them for transport into or out of the nucleus, for example the aforementioned Rtg3 is regulated in such a way (Komeili *et al.* 2000). Another example is Pho4 which is exported out of the nucleus after phosphorylation (Komeili and O'Shea 1999).

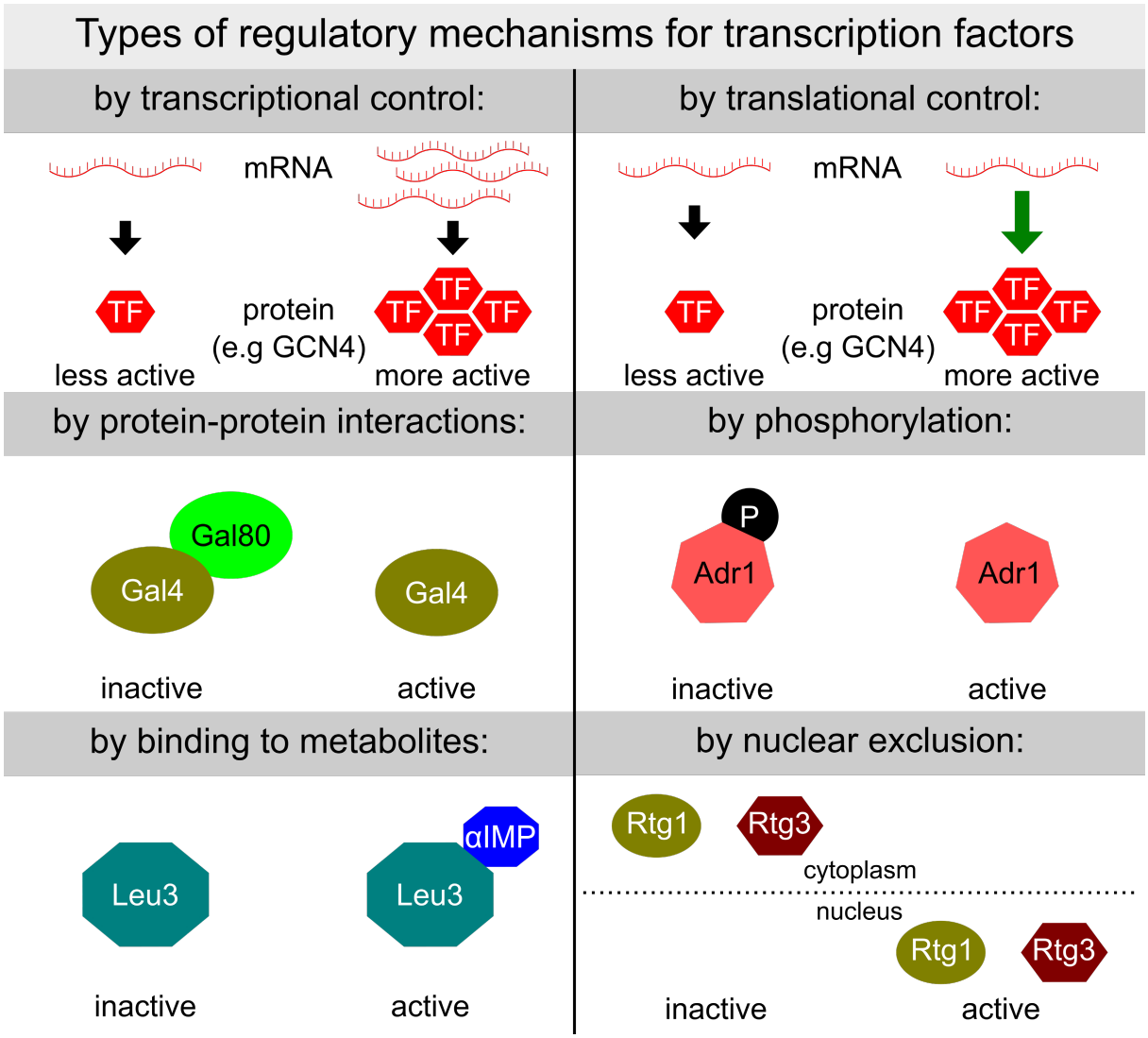

Figure 5: Overview of common regulatory mechanisms influencing the activity of transcription factors.

1.4 WHY STUDY TRANSCRIPTIONAL REGULATION?

Now that we have a good understanding of what transcription is and how it is regulated, we should take a step back and talk about why studying this topic is important and interesting, before moving on to how to study it.

First of all, I would like to note how limited our current knowledge of the molecular processes and regulatory steps inside the cells still is. Therefore, gaining a better understanding of any of those processes is a valid research goal on its own, because sometimes we do not even know what we do not know. But besides a general need for basic research what are more tangible fields where we could apply the gained knowledge about transcriptional regulation?

When it comes to using improved knowledge about molecular biology, one important area is always medicine, often related to cancer research and treatment. It has been shown that TFs and therefore transcriptional regulatory processes play an important role in cancer development and progression (Nebert 2002). But being linked to cancer development is not unique to TFs and it would actually be quite difficult to find an important cellular mechanism or process that would not cause cancer when it goes horribly wrong.

So why am I interested in studying TFs? My motivation comes from a mix of fascination for basic research combined with the potential application of this knowledge to the area of metabolic engineering. Before I start and describe how knowledge about TFs can be used in metabolic engineering, let me first explain what that is.

As explained before in the introduction, cells take in nutrients from the environment and convert these into products, for example into ethanol in the case of beer and wine fermentation processes. In metabolic engineering one genetically modifies the organism with the purpose to increase the rate or efficiency of producing a desired chemical, for example to increase the amount of ethanol produced in a microbial bioethanol production plant. Metabolic engineering can also enable the organism to synthesize products it was not able to produce before, for example the production of terpenoids by yeast for the use in fragrances or pharmaceuticals (Zhang, Nielsen and Liu 2017).

Metabolic engineering is an important tool to achieve the transformation towards a bio-based economy, where many chemicals will be produced by sustainable fermentation processes, thereby decreasing our dependency on fossil derived resources. Therefore, metabolic engineering will be increasingly important to support the needs of our society in the future.

So how can TFs and knowledge about their roles be used in metabolic engineering? As TFs are versatile and play an important role in many cellular processes there are plenty of ways to use them in metabolic engineering.

One could use them to upregulate their target pathways to achieve a desired phenotype. This has been shown by overexpressing the stress-responsive TF Msn2, which led to an increase in furfural resistance and improved ethanol production rates (Sasano *et al.* 2012). The problem with this approach is that we unfortunately lack sufficient knowledge to identify the best TF or group of TFs that we should overexpress to achieve our goal.

But there are still ways how one can use the power of TFs in metabolic engineering already today. One of these options is through the creation of a library of artificial TFs, where a known activator domain is fused with different artificial DNA binding domains, and then select those artificial TFs that show the desired improvement. The applicability of this idea has already been shown for improving the phenotypes of yeast cells, for example to achieve an increased thermotolerance (Park *et al.* 2003). The drawback, however, is that one has to create an extensive library and invest resources to screen the library, which takes time and effort. This bottleneck can however be alleviated by more knowledge about TFs as stated before.

Modulation of TFs as metabolic engineering strategy has not only been successfully demonstrated in yeast but also in other organisms, for example in bacteria and plant cells (Broun 2004; Grove 2017), making it a promising research area for many applications.

Besides working with native or fully artificial TFs one can also take a route in-between and modify native TFs for increasing productivity, for example using a constitutively active form of Leu3 for increased isobutanol production (Park, Kim and Hahn 2014). TFs can also be used as part of a biosensor where a TF induces the expression of specific target genes based on the input from the sensor. Increased productivity was achieved by linking the intra-cellular levels of malonyl-CoA to the synthesis of the product (David, Nielsen and Siewers 2016).

Apart from these few examples, there are many more possibilities to use TFs in the field of metabolic engineering. Therefore, improved knowledge about TFs would allow us to speed up the metabolic engineering efforts and thereby help with the transition towards a bio-based economy.

1.5 HOW CAN WE STUDY IT?

Now that we have established that the study of transcriptional regulation and transcription factors is interesting and important, how can we do this? There are two common ways to study TFs, either by changing the expression level of the TF and measure the effects or by identifying the genomic location where the TF is binding.

An example for modulating the TF abundance in order to infer their function is the large-scale knock-out study by Hu *et al*., showing that one can use such a dataset to reconstruct a transcriptional regulatory network (Hu, Killion and Iyer 2007). The computationally identified gene targets for their TF showed a good enrichment for known motifs. In addition, using their computational model they could assign TFs to be either activating or repressing. Another example of this approach is the study by Hackett *et al.*, where they used an inducible promoter to selectively overexpress a single TF and measure the short-term changes in genome-wide gene expression levels using microarray measurements at different time points 5 to 90 minutes after induction (Hackett *et al.* 2020).

The issues with these approaches are that other TFs can buffer the effect of a changed TF abundance, and that there is no proof that the observed changes are directly caused by the changed TF and not by the adaption of the cell to the new state. Both these issues have been reduced in the approach used by Hackett *et al.*, by measuring the short-term changes in expression directly after the increased expression of the TF. Regardless, the overexpression has its own challenges, because many TFs are activated by a certain metabolic state and just producing more TF molecules that will be inactive in the studied condition will not reveal their metabolic function.

The other common approach to identify regulatory targets of TFs are based on identifying the binding sites of a specific TF. Most approaches to achieve this are based on a method called Chromatin Immunoprecipitation (ChIP) (Carey, Peterson and Smale 2009). The principle behind ChIP is that one first fixates all DNA bound proteins for example by using formaldehyde and then employ an antibody against a specific TF of interest (or against a tag attached to that TF) to selectively enrich for this TF and simultaneously filter out all other TFs that were also bound to the DNA. How the TF-bound DNA is then identified has changed dramatically over the years. The first large-scale method used DNA microarrays, resulting in the ChIP-chip method (Ren *et al.* 2000). After high-throughput sequencing technologies were becoming more accessible, an improved protocol for Illumina sequencers was developed, called ChIP-seq (Johnson *et al.* 2007). The next and currently last step of protocol development introduced a lambda exonuclease treatment, resulting in ChIP-exo (Rhee and Pugh 2011; Rossi, Lai and Pugh 2018). The exonuclease treatment will digest

DNA that is not directly bound by any TF, thereby improving the resolution down to the single nucleotide level. The exonuclease treatment also improved the signal-to-noise ratio by degrading unbound DNA strands. The individual steps involved in the ChIP-exo protocol are covered in Section 2.2.

To date many large scale studies of *S. cerevisiae* TFs have been performed using different ChIP methods, mainly by the now outdated ChIP-chip, like the study by Lee and the extension by Harbison (Lee *et al.* 2002; Harbison *et al.* 2004), which were performed for a total of 158 TFs during growth in rich culture medium and occasionally other conditions, or a study of 30 TFs involved in DNA damage response (Workman *et al.* 2006). The main conclusion gained from these large-scale studies is that the binding behavior of many TFs is highly dependent on the exact growth condition. Therefore, one can unfortunately not run a single experiment using rich media and get all insights into the TF. However, if one compares two different conditions, one can gain meaningful insights into specific regulatory programs, for example into the response to DNA damage (Workman *et al.* 2006). These studies also showed that the number of gene targets per TF varies significantly, with some TFs being quite specialized with very few targets, while other TFs are acting more as general regulators of transcription have several hundred targets.

Besides ChIP based methods to identify TF binding sites there are methods that identify all binding sites of all TFs simultaneously, like ATAC-seq (Li *et al.* 2019) or DNase-seq (He *et al.* 2014). The main different is that these methods identify TF binding peaks but cannot identify which specific TF is binding. So, to identify the binding sites of a specific TF one still needs a ChIP based method.

1.6 AIMS OF THIS THESIS

The currently best way to study TFs and reliably identify their targets in my opinion is by using large-scale ChIP-exo experiments involving many TFs in different environmental conditions. Therefore, the aim of this thesis was to establish a robust and efficient framework to streamline the process and serve as a base for future ChIP-exo studies.

In addition, I wanted to use that framework to study TF binding events using machine learning approaches to see how they can be used in this context and if it would be possible to create tools for targeted metabolic engineering approach. Given the vast amount of TF binding data already collected and the simplified processes of gathering more data, because of my framework, I believe that machine learning will become more and more important in the study of transcriptional regulation.

Besides large-scale studies, I also wanted to show how this framework can be used to gather detailed insights about single TFs with selected small-scale experiments. This shows how to efficiently augment the knowledge gained through large scale experiments, possibly using machine learning approaches, by detailed studies to further improve the knowledge base we have.

2 EXPERIMENTAL SETUP AND METHODS

In this section I will cover the employed experimental setup for the majority of experiments in this thesis as well as explaining the main methods I used.

2.1 CHEMOSTATS AND GROWTH CONDITIONS

All experiments performed in this thesis were using the *S. cerevisiae* strain CEN.PK113-7D (van Dijken *et al.* 2000) and the majority of experiments were done using a continuous cultivation method called chemostats. A schematic overview of such a reaction vessel setup is shown in Figure 6 A, the chemostat system used here was built by D2Biotech (D2biotech.com). Fresh medium is constantly pumped into the vessel at a fixed rate while spent medium containing yeast cells is drained to achieve a constant volume. In addition, gas is pumped into the vessel for aeration, and the medium is continuously stirred. Together, this system forces the cells to grow at a growth rate corresponding to the pump rate set for the fresh medium inflow (this is also called the dilution rate). The growth rate equals the dilution rate because of the limited supply of fresh nutrients. If there are too many cells in the vessel, they can only grow slower than the dilution rate due to the competition for nutrients and are therefore slowly washed out (the number of cells decreases). If there are too few cells in the vessel, they have an overabundance of nutrients and can grow faster than the dilution rate until the cells reach a steady state of cell count and growth rate. Therefore, this system provides a robust base for biological experiments as the cells reside in a reproducible steady state. In addition, the fixed growth rate enables the straightforward comparison of cells grown at different growth conditions without the influence of changes in the growth rate. This is important as various media compositions were used in this thesis, forcing the cells to employ different metabolic pathways for growth and survival. If the cells had been grown in shake flasks, where they had reached their condition-specific maximum growth rate, many of the observed differences between the conditions would have originated from the different growth rate and not the different metabolic program employed. The four main media compositions used in this thesis were: (i) fermentative glucose metabolism using glucose limitation in an anaerobic environment; (ii) gluconeogenic respiration using ethanol limitation; (iii) respiratory glucose metabolism using glucose limitation; and (iv) aerobic fermentation using nitrogen limitation. An overview of the conditions is shown in Figure 6 B.

These conditions were chosen to cover a wide range of metabolic states and programs of the cells. In addition, these conditions represent industrially relevant conditions and are therefore interesting to explore. In industrial fermentation processes, such as second generation bioethanol production, the expenses for the carbon source can make up 45–58% of the overall product cost (Hamelinck and Faaij 2006). Therefore, industrial fermentations are often nutrient-limited conditions to minimize wasted resources, similar to the glucose-limited condition studied here. Many industrial cultivations are also performed in anaerobic fermentations, as it is surprisingly expensive to aerate the sizeable tanks that can typically contain up to 200,000 liter (Humbird, Davis and McMillan 2017). A better understanding of anaerobic conditions therefore also has industrial relevance.

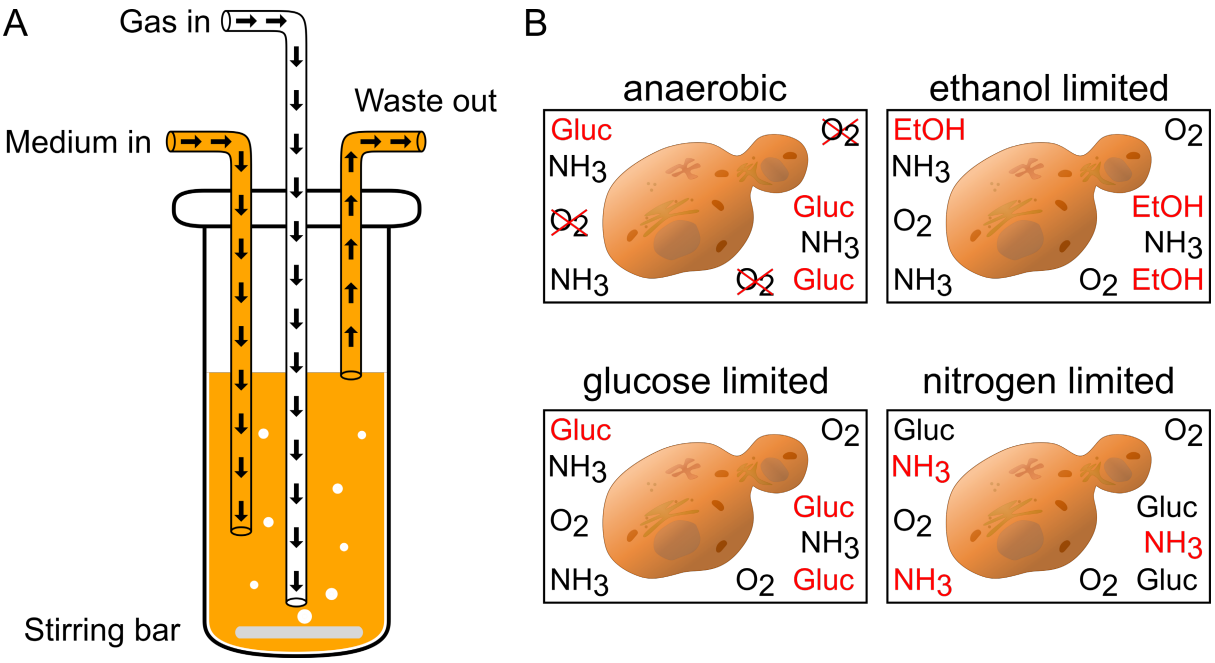

Figure 6: Overview of chemostat and cultivation conditions. A: Overview of a chemostat reaction vessel. **B:** Graphical representation of the four used media compositions.

2.2 CHIP-EXO METHODOLOGY

Inside the cells the TF of interest as well as other TFs and DNA binding proteins are constantly binding to and dissociating from the DNA as well as moving along the DNA during all stages of the cells life cycle (Marklund *et al.* 2020). Therefore, the first step of measuring the TF binding is to create a snapshot of the currently bound proteins. This is achieved by treating the cells with formaldehyde, which covalently cross-links all DNA–protein complexes, stopping the constant dissociation (Solomon and Varshavsky 1985). Due to the small molecular size of formaldehyde it also only links macromolecules together that are maximum 2 Å apart, thereby avoiding to cross-link every protein present in the nucleus with the DNA (Hoffman *et al.* 2015). Next, the DNA is sheared using sonication, to achieve an average length of 200 to 500 bp. Then, antibodies against the TF or its attached tag are used to enrich the mixture of sheared DNA-protein complexes for the TF of interest. This enrichment process using antibodies is why the method is called chromatin immunoprecipitation (Carey, Peterson and Smale 2009). Up until this point the process is exactly the same for the ChIP-chip, ChIP-seq and ChIP-exo method.

For ChIP-chip and ChIP-seq one would now reverse the DNA-protein crosslinking, degrade all proteins using proteinase K and then identify the bound DNA either by DNA microarrays (ChIP-chip) or by sequencing (ChIP-seq) (Ren *et al.* 2000; Johnson *et al.* 2007). The obtained resolution of these methods depends then on the exact specification of the DNA microarray and on the size of the DNA fragments after sonication, but in general only a resolution of up to 100 bp can be achieved (Rhee and Pugh 2011).

To increase the resolution to the single nucleotide level, a treatment step with the lambda exonuclease was added, giving ChIP-exo its name (Rhee and Pugh 2011, 2012). The lambda exonuclease will selectively degrade DNA from the 5′ to the 3′ end in a processive manner. The DNA that is covered by the TF is protected from the exonuclease and therefore the 5′ end of both the DNA strands will end at the border of the TF. This means that the achievable resolution of the footprint of the TF is defined by its actual binding size and not by the size of the DNA fragments after sonication. Next, the cross-linking is reversed, all proteins are degraded, and the DNA fragments are prepared for sequencing by adding the necessary Illumina sequencing linkers.

After high-throughput sequencing, the obtained reads are mapped to the genome of the organism and the TF binding peaks can be detected and analyzed with specialized software, for example with GEM to call the TF binding peaks (Guo, Mahony and Gifford 2012). The process of steps of the ChIP-exo method are summarized in Figure 7.

The original ChIP-exo protocol published in 2012 involved thirteen enzymatic steps, making the process long and cumbersome. To address this issue and to increase the adoption of ChIP-exo in the research community, in 2018 an improved protocol called ChIP-exo 5.0 with only five enzymatic step was published (Rossi, Lai and Pugh 2018).

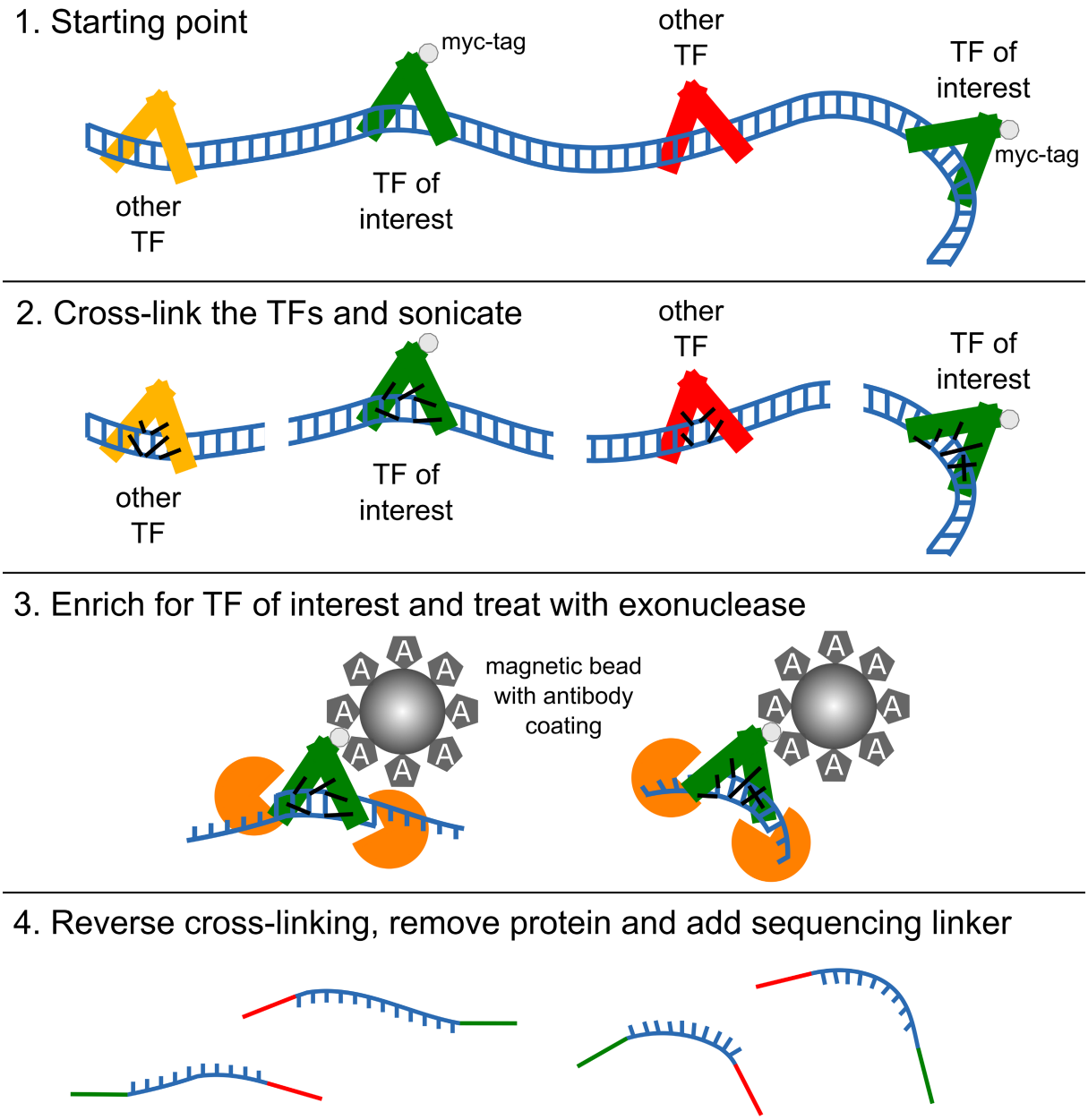

Figure 7: Steps of the ChIP-exo method. After sampling the TFs are crosslinked to the DNA using formaldehyde and the DNA is sheared into 200 to 500 bp long stretches using sonication. Next magnetic antibodies are used for enrichment of the TF of interest and 5′ ends of the unbound DNA are removed using the lambda exonuclease. Finally, the cross-linking is reverse, the proteins are degraded, and the sequencing linkers are added to finish the library preparation.

2.3 MACHINE LEARNING

As two of the papers included in this thesis rely heavily on machine learning as part of the specialized software to analyze and repurpose ChIP-exo data, I will give a brief introduction covering the different types of machine learning and the main principles on how to use them. The basic idea of most machine learning approaches is that we have data available and would like to train a computational model on it, in order to predict one or multiple characteristics. For the sake of completeness, I have to mention one type of machine learning called reinforcement learning, that does not start with data, but I will not go into more details about it and focus on the other two approaches.

What kind of data we have available and what kind of features we would like to learn and predict determines whether we can employ supervised or non-supervised methods. If the data we have also contain labels, for example a set of pictures of cats and dogs that are individually annotated as either depicting a cat or a dog, we can use supervised methods. This is what most people would think of when one talks about machine learning. If we do not have these labels available, we can only use unsupervised methods, like a clustering algorithm to try and identify interesting subgroups in our data set. In the cat and dog example, that could result in one cluster of images only depicting dogs and another one only containing images of cats, while the model would still not be able to say what each cluster contains as it has no labels available.

In this thesis I will only be using supervised learning and this approach can be further divided into two classes: classification and regression. Classification is when the model learns to which class an observation belongs. A widely used example data set for this type of machine learning is on the iris flower, where the flower is either classified as being *Iris setosa*, *Iris virginica*, or *Iris versicolor* based on the length and width of the petal and the sepal (Fisher 1936). The other class of supervised learning is regression, where the model output is not the class but any numerical value. This means, that while a classification algorithm trained on the iris flower set will never be able to classify something for example as a cactus, a regression algorithm can output a value that is not present in the training data. The typical example for learning about regression is predicting the price of a house based on the size, the number of rooms, a rating of the neighborhood, etc., for example by using the Boston Housing dataset from 1976 (Rubinfeld and Harrison 1978). A graphical overview of these two classes and their differences is shown in Figure 8.

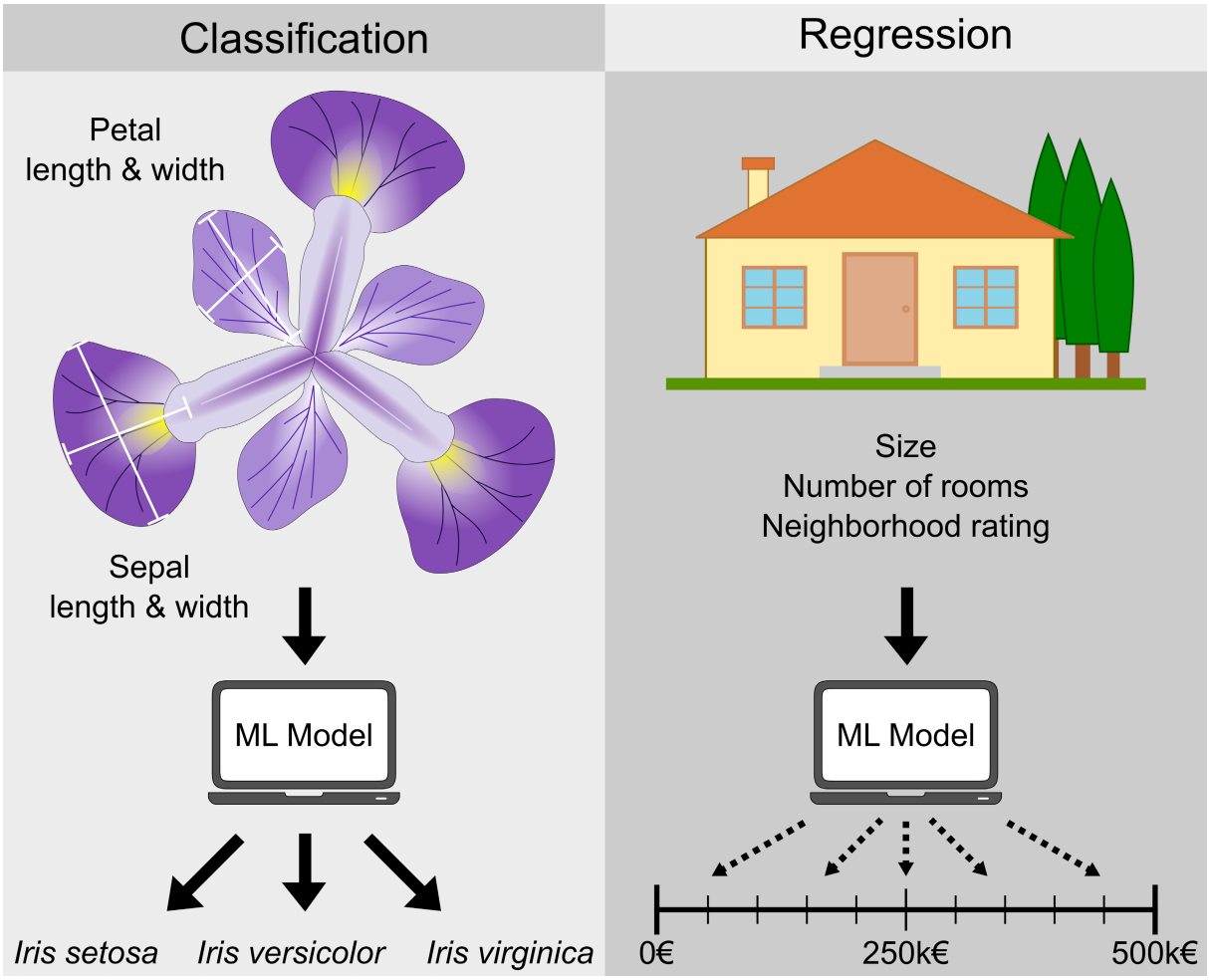

Figure 8: Overview of the two supervised machine learning classes. While classification assigns one of the three known classes to the data, regression can output any number for the house, even if that number was not present in the training data.

How can we evaluate what the model learned and how good the model is? This evaluation is done by predicting the labels of our data and compare them to the actual values, to see how far off the predictions are. In regression problems, one could then take the mean absolute error as an evaluation metric. When choosing and evaluating a machine learning model one commonly observes two types of issues called under- and over-fitting. To illustrate what they exactly are, let us have a look at the example in the first row of Figure 9, where we have data with one input variable (X) and one label / output variable (Y). The relationship between X and Y is a noisy quadratic function. We now fit three different regression models on the data, a linear model, a quadratic model and a 6^{th} order polynomial model, as shown in the second row of Figure 9.

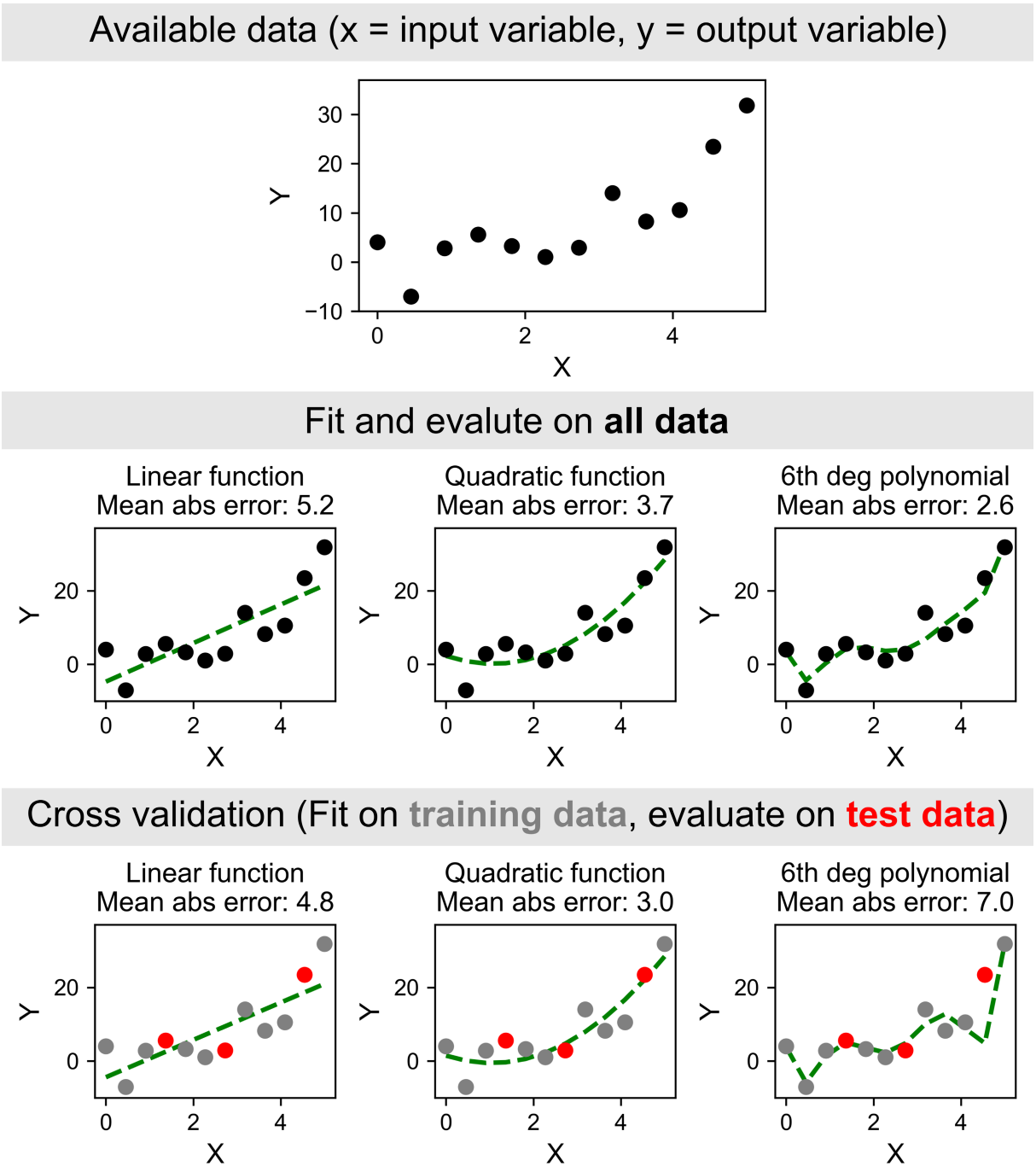

Figure 9: Under- and over-fitting in machine learning. Example of fitting data that was generated with a quadratic function. The linear fit does not fit the data well (it is underfitting), while the 6th degree polynomial is overfitting. The overfitting can only be detected when using cross validation. If wrongly fitted and evaluated on all data, the 6th degree polynomial fit produces the smallest error of the three.

The goodness of fit for these three models is then evaluated by comparing the predicted values of Y to the true data for Y and the mean absolute error is calculated and displayed. One can observe that the linear model does not capture the full dynamics of the underlying data, a behavior called underfitting. It therefore has the highest error of the three models. The quadratic model (that is also underlying the data) fits very well to the data as expected. Interestingly, it does however not have the lowest error, which is achieved by the 6^{th} degree polynomial model. The drawback with that result is that this model captures a lot of the noise in the data (see the initial dip in the prediction curve) and will therefore probably not generalize very well for other points, a phenomenon called overfitting. For a simple data set like the present one here with just one variable it is easy to plot the results and examine which model fits the best visually, but how can this be achieved for larger data sets with many variables if one cannot use the amount of error as a guide?

To solve this problem, one commonly uses a method called cross-validation. The idea is that one splits the data into two sets, one training data set and one test set. In the third row of Figure 9, this is displayed by assigning 75% of the data to the training data in grey and 25% to the test data in red. The models are then trained on the training data and are used to predict the labels for the test data. Now the model with the lowest error is indeed the quadratic model, outperforming both other models by a large margin. By using cross-validation we can now be more confident that the model really learned the characteristics of the data and will generalize well to new data when it produces a low error measurement. The most used cross-validation technique is called k-fold cross-validation, with the k standing for an integer typically in the range of 5 to 10. This means that the data is split into k-folds, and then the model is trained on all, except one, folds and evaluated on the left-out fold. This is repeated so that every fold is used once for evaluation. This method has the advantage that all the data can be used for training and that the models will be evaluated on all data points. This is important, because the random assignment of data points into training and test sets could otherwise have unwanted effects. It could have happened that in our example the three points with the lowest x values had been assigned to the test data, which would not have given us a fair evaluation of the model across the whole range of values.

After successfully training a machine learning model, one can then interrogate it to try and understand which characteristics the model learned and are therefore of interest.

Supervised machine learning techniques can therefore be a powerful tool to improve our understanding of biology and have been applied in a variety of research projects like the prediction of essential genes (Hwang *et al.* 2009), classifying breast cancer types (Amrane *et al.* 2018) or predicting protein structures (Senior *et al.* 2020).

3 ESTABLISHING THE FRAMEWORK

The aim of this thesis was to establish a framework to efficiently analyze large scale data sets of TF binding data and apply it to gain insights into TF biding events. This framework, which is presented here in this section consists of two parts and covers **Paper I** and **II**. First, the transcription start sites for each expressed gene had to be determined (**Paper I**) and then an efficient bioinformatics pipeline had to be constructed (**Paper II**).

3.1 IDENTIFICATION OF TRANSCRIPTION START SITES

When analyzing TF binding data, the exact location of the binding site is important, which can be obtained by mapping the sequencing reads against the genome of the organism. This, however, only provides the absolute position on the chromosome, for example that the TF Gcn4 is binding to chromosome 10 around the base pair 161962, which on its own is not informative. One can use the genome annotations to search if any genes are located close to the detected binding site, and here, this would tell us that the binding occurs 843 bp upstream of the start of the *URA2* gene. This is already much more informative, but it is known that the actual site where transcription starts is not identical to the start of the coding sequence. This means that the distance of the TF binding site to the ATG is difficult to compare between different genes, because the distance from the ATG to the TSS is different for each gene. To overcome this hurdle. one would need to know the exact location of the Transcription Start Site (TSS), and in this example it would reveal that the Gcn4 binding site is 317 bp upstream of the *URA2* TSS, a measurement that is now comparable across different genes. The only issue left here is that the exact locations of the TSSs in *S. cerevisiae* have previously only been reported using rich medium in shake flasks (Parky *et al.* 2014; Wery *et al.* 2016), but not in the growth conditions we were interested in. It is known that the growth conditions of the cells influences the TSS in other eukaryotic organisms (Reyes and Huber 2018), therefore we could not use the already published data because it was not known if the growth condition also influences the position of the TSS in yeast. To address this issue, the TSS were experimentally determined as described in the first paper of this thesis, **Paper I**.

To identify and map the TSS at a high resolution I used Cap Analysis of Gene Expression (CAGE), an RNA sequencing method to specifically sequence the 5′ ends of intact mRNA molecules (Kodzius *et al.* 2006). More specifically I used the non-amplification non-tagging CAGE method which has the best resolution and signal to noise ratio of all CAGE protocols (Murata *et al.* 2014).To be able to identify the TSS in the four conditions used in the various experiments across this thesis and to asses if the TSS landscape is changing in different growth conditions, the cells were grown in

these four different conditions using chemostats with a fixed dilution rate of 0.1/h. The four conditions were: (i) fermentative glucose metabolism using glucose limitation in an anaerobic environment, (ii) gluconeogenic respiration using ethanol limitation, (iii) respiratory glucose metabolism using glucose limitation and (iv) aerobic fermentation using nitrogen limitation. After the cells reached steady state for at least 24 hours, the cells were collected, and CAGE was performed. The workflow of the method is shown in Figure 10. The first step was to extract the mRNA from the yeast cells, which results in a mix of intact mRNA molecules that were still capped at their 5′ end and partially degraded ones that already lost the cap. Next, a reverse transcriptase was used to create a hybrid RNA – cDNA double strand using random hexamer primers, and the cap was biotinylated. As the starting point for the reverse transcriptase was random it was not guaranteed that it will continue long enough to reach the 5′ end of the mRNA molecule. Therefore, the hybrid strands were treated with RNAse I that selectively degraded single stranded RNA, resulting in the removal of the 5′ cap if the reverse transcriptase did not transcribe until the end. Now, all molecules with a biotinylated cap were extracted using streptavidin beads. In the last steps before sequencing, the cap was removed, the mRNA was replaced with a second DNA strand and the necessary linkers were added at both ends for subsequent Illumina sequencing. After sequencing, the reads were mapped against the reference genome for the used CEN.PK113-7D strain (Salazar *et al.* 2017). The 5′ end of each read now denoted the starting position for an individual transcription event. The data were analyzed using CAGEr (Haberle *et al.* 2015), where first the individual 5′ end read positions were clustered together into TSS clusters, then the clusters were merged across the conditions and the cluster width as well as the main TSS position was determined. Finally, the TSS clusters were assigned to genes if their main TSS position was within 1 kb upstream of the start of the gene coding region. An example of how the data looked for two different promoter regions is shown in Figure 11 A and B. For each of the four condition the individual read distribution is shown, the height of each blue bar denotes how many transcription start events originated from that exact nucleotide position. One can also see how the individual transcription start events were clustered together (the row labeled Cluster annotation), and where the dominant TSS is (thin vertical blue line in the row labeled Cluster annotation Dom TSS).

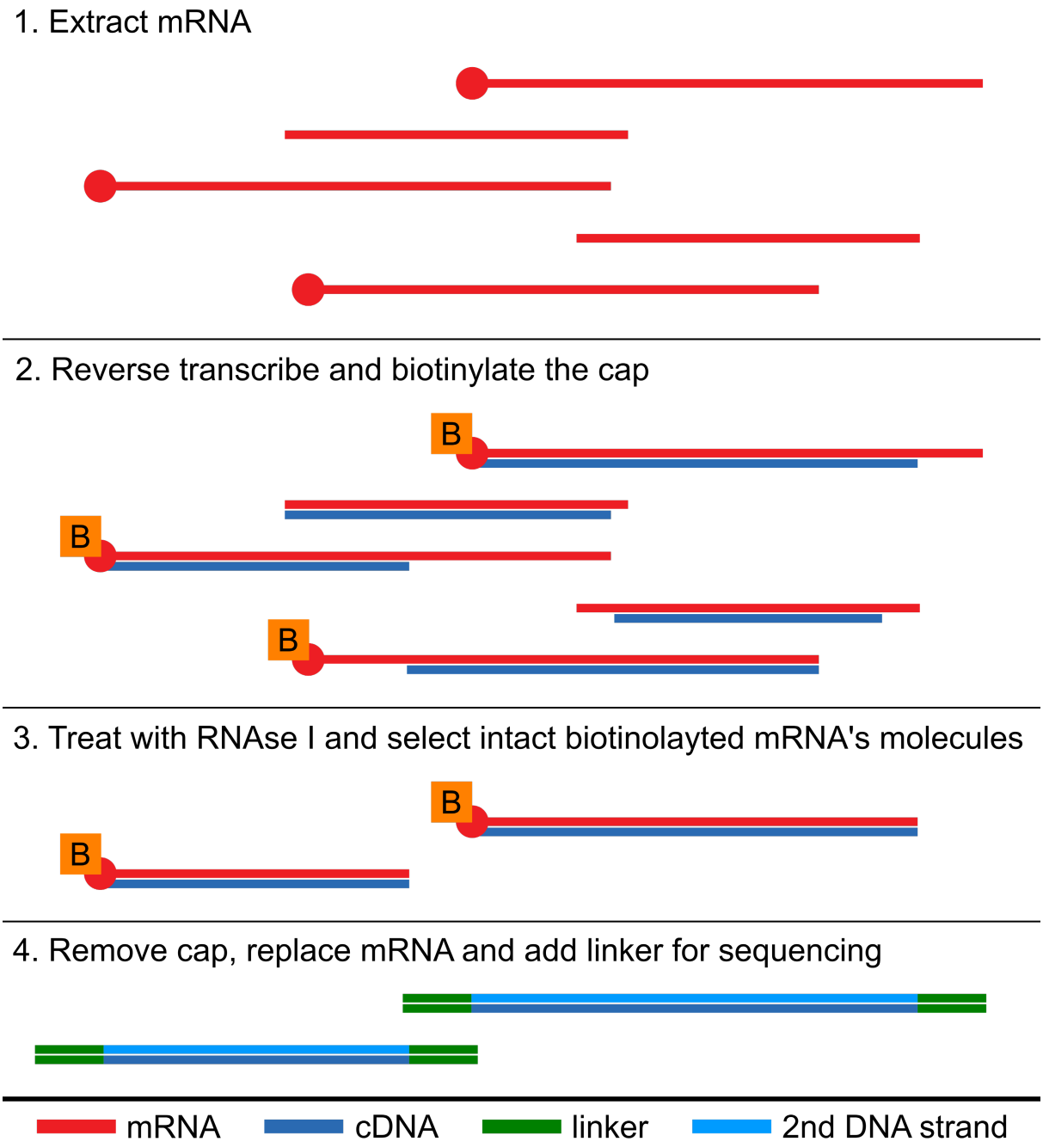

Figure 10: CAGE workflow. After the mRNA is extracted from the cells, it was reverse transcribed into a cDNA-mRNA hybrid molecule and the mRNA cap structure was biotinylated. Next an RNAse I treatment was used to selectively degrade RNA molecules without a cDNA binding partner and magnetic beads coated in streptavidin were used to select for mRNA molecule with an intact cap. Finally, the cap is removed, the mRNA is replaced by DNA and the necessary linkers are added for sequencing.

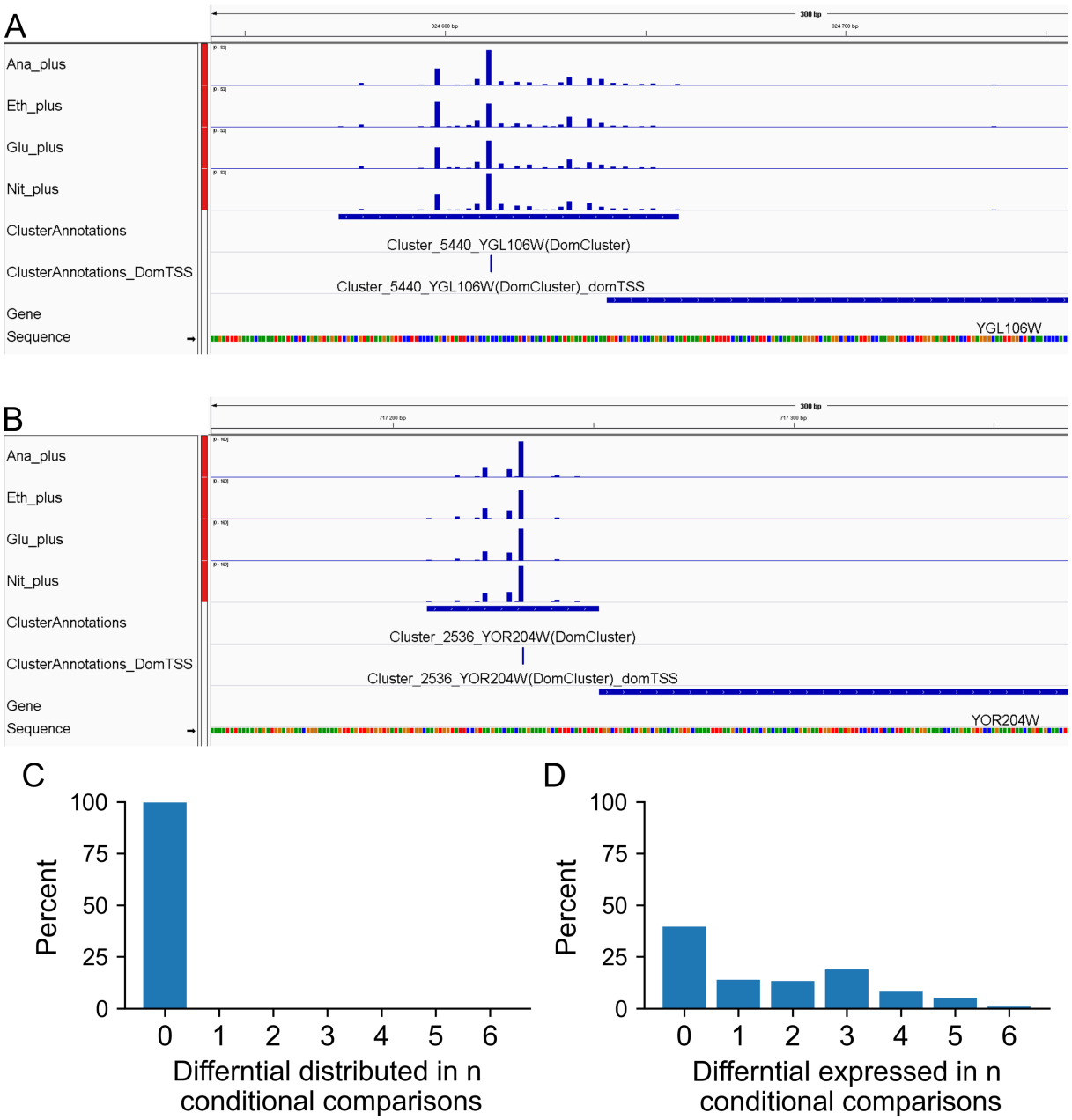

Figure 11: Overview of TSS cluster. A: Screenshot from IGV showing the broad CAGE read distribution for the constitutively expressed gene YGL106W (*MLC1*). **B:** Screenshot from IGV showing the peaked CAGE read distribution for the constitutively expressed gene YOR204W (*DED1*). **C:** Results for detecting shifted TSS, showing the proportions of clusters associated to genes that were detected as differentially distributed (using getShiftingPromoters from CAGEr, shifting score > 0.6 and adj. p-value of Kolmogorov-Smirnov test < 0.01) in n pairwise comparisons of the different conditions (four conditions = six possible comparisons). **D:** Results for differential gene expression analysis of RNA-seq data, showing the proportions of genes that were detected as differentially expressed (using DEseq2, adj. p-value < 0.001) in n pairwise comparisons of the four different conditions.

3.1.1 Stability of the TSS landscape

One of the reasons the CAGE experiment was performed was to determine whether the TSS landscape is different in the four different conditions, and in Figure 11 A and B one can already observe that even though the two TSS cluster have quite different width and shapes there are no obvious differences between the four conditions. To quantify the stability, I tested whether the read distribution for each cluster was significantly different in any of the four conditions, using six pair-wise comparisons. The assessment was performed using the Kolmogorov–Smirnov test with an adjusted p-value threshold of 0.01. The results for this are shown in Figure 11 C, and 99.7% of all cluster show no differential distribution. This is especially striking because more than 50% of all cluster show a significant change in expression level in at least one pairwise comparison as shown in Figure 11 D. This means that, even though the overall expression levels are changing quite dramatically, the distribution of transcription start events stays the same. To further validate the stability of the TSS landscape across conditions, we compared the length of the 5′ untranslated region (5′ UTR, the region between the TSS and the start of the coding sequence) we obtained from our data with the previously published data for another *S. cerevisiae* strain (Parky *et al.* 2014) and found that the average difference is less than 9 bp. Therefore, for all future experiments where one maps TF binding events in relation to the TSS, the same TSS annotation can be used independent of the actual growth condition.

3.1.2 Cluster shape index

In Figure 11 one could already observe that different TSS clusters have distinct widths and shapes and Hoskins *et al.*, developed a metric called Shape Index (SI) to measure how broad or peaked a cluster is (Hoskins *et al.* 2011). The SI for a cluster can be calculated using the following formula:

$$SI = 2 + \sum_{i}^{L} p_i * log2\,(p_i)$$

$$p_i = \text{proportion of counts at position i in the cluster}$$

$$L = \text{position with at least 1 tag}$$

Clusters with an SI of less than −1 are classified as broad (e.g. the cluster in Figure 11 A) and clusters with an SI of larger than −1 are classified as peaked (e.g. the cluster in Figure 11 B). The distribution of all SI values is shown in Figure 12 A. One can observe that the majority of all cluster are classified as peaked. The SI, which is calculated for each individual condition, is also very stable across the different conditions, with a minimal Pearson correlation coefficient of 0.92 between any two conditions, as shown in Figure 12 B.

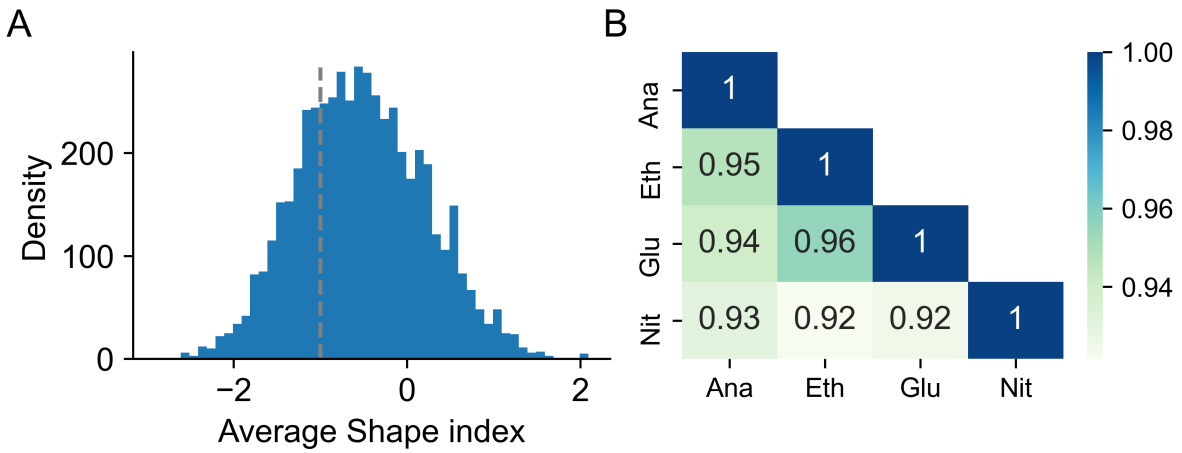

Figure 12: Overview of shape index characteristics. A: Distribution of average shape index of each cluster, bin size = 0.1. Grey dashed line at -1 denotes the border which separates clusters classified as peaked (shape index > −1) and clusters classified as broad (shape index ≤ −1). **B:** Pearson correlation coefficient for each pairwise comparison of the condition specific shape index.

The shape index also has another very interesting characteristics, it shows a quite strong negative correlation with the expression levels as displayed in Figure 13 A. This means that a gene which TSS cluster has a higher shape index (more peaked) tends to have a lower expression value than a gene with a broad TSS cluster distribution. This correlation also seems to be exclusively linked to the actual distribution of transcription start events and not just the width of the cluster span (Figure 13 B). The link between the shape index and the gene expression can also be observed in higher organisms like *Drosophila melanogaster*, where there is a remarkable connection between the shape index and the gene expression level during different developmental phases (Hoskins *et al.* 2011).

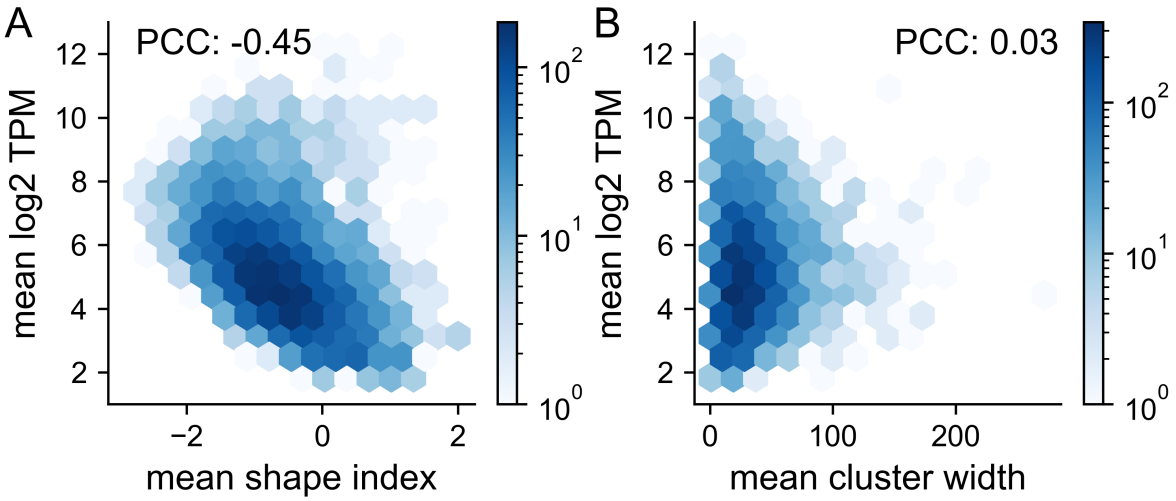

Figure 13: Correlation of TSS cluster characteristics with expression levels. A: Comparison of the mean shape index and the mean CAGE expression levels showing an anti-correlation with a Pearson correlation coefficient −0.45. **D:** Comparison of the mean promoter width with mean CAGE expression levels showing no correlation. TPM = transcript per million. Color denotes the number of observations per hex tile.

3.2 CREATION OF THE BIOINFORMATICS PIPELINE

After obtaining the high resolution TSS annotations, the next step was to analyze the already gathered ChIP-exo data. Even though the original ChIP-exo protocol has been published several years ago (Rhee and Pugh 2012) there had not been a complete bioinformatics pipeline available to process the generated sequencing reads in a straight forward matter. Therefore, we developed such a pipeline, which will be presented here (**Paper II**).

Analyzing ChIP-exo data is more complicated than analyzing ChIP-seq data, as the borders of the TF binding area are captured in different reads, one side by reads from the positive strand and the other side by reads from the negative strand. This means that one has to estimate the distance between the borders and trim the reads to the correct length so that reads coming from both sides will overlap in the middle to mark the TF binding site. This procedure is visualized in Figure 14. If the reads are not extended long enough, they will not overlap in the middle and will therefore not be recognized as a peak. If the reads are on the other hand extended too long, they will extend over the border of the TF and increase the measured footprint, thereby reducing the overall resolution which is the main reason for using ChIP-exo in the first place. To overcome this issue, we developed a formula to calculate the optimal trim length of a TF based on its size:

$$TF\ weight = sequence\ length\ [nuc] * \frac{1\ AA}{3\ nuc} * 110 \frac{Da}{AA}$$

$$TF\ radius = 0.066\ \frac{nm}{daltons} *\ TF\ weight\ [Da]^{\frac{1}{3}}$$

$$TF\ footprint = 3 * TF\ radius * 3.03 \frac{bp}{nm}$$

The first step is to calculate the weight of the TF based on the amount of amino acids and the average amino acid weight of 110 Da. Next, the size of the TF is estimated based on a spherical shape using a previous published formula (Erickson 2009). Based on the assumption that most TFs bind as dimers that overlap by half the diameter, we multiply the radius by 3 to obtain the footprint in nm and convert this to length in bp units. This footprint then corresponds to the optimal trim length.

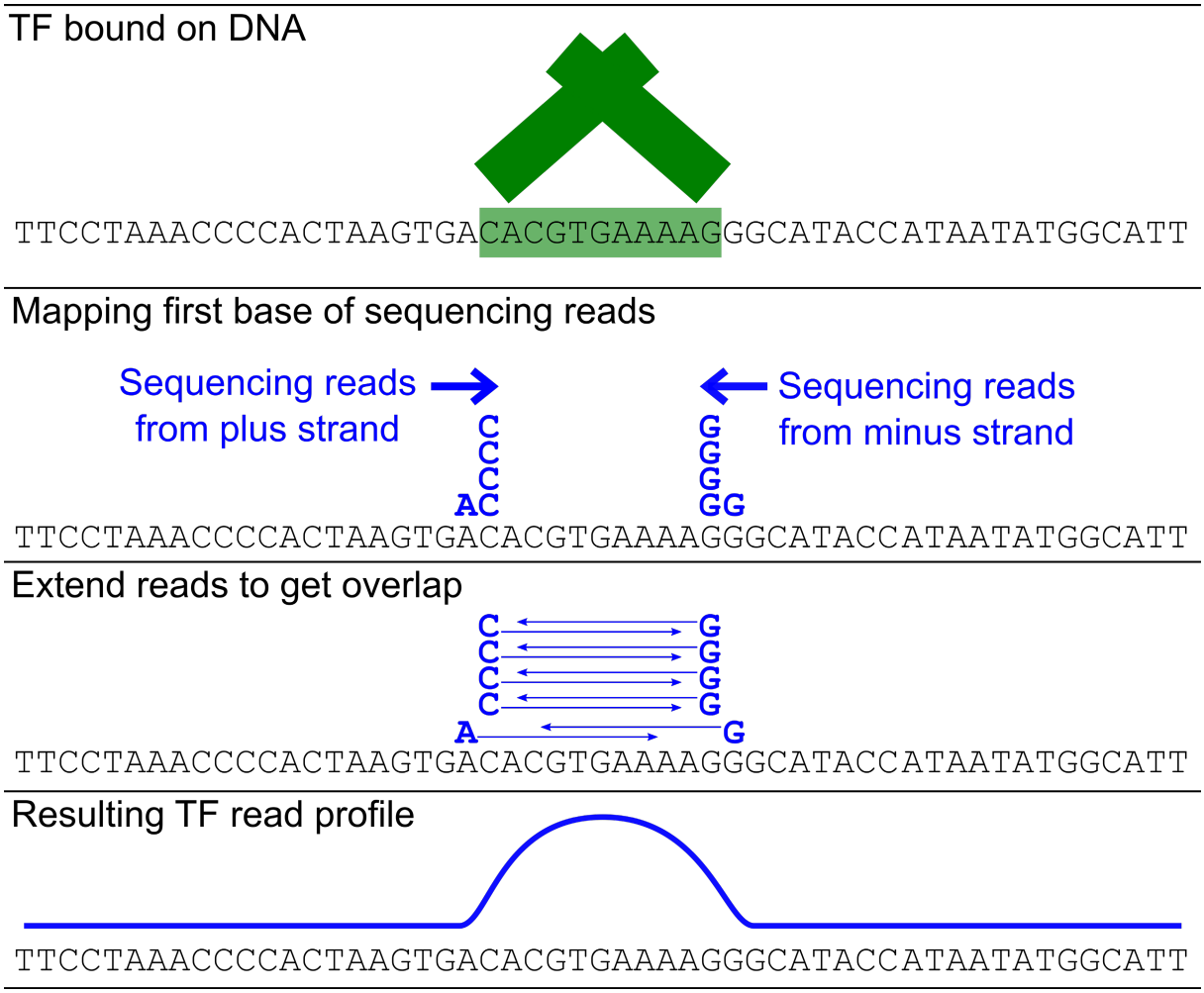

Figure 14: Overview of ChIP-exo read trimming procedure. To reconstruct the binding footprint of the TF, as shown in the upper row, from the sequencing data, first the reads were mapped to the genome and the position of the first base of each read was marked. Then each read was extended to the previously determined trim length so that reads from both sides overlap. This overlap corresponds to the real binding behavior in the cell, as shown in the last row now.

3.2.1 Pipeline workflow

The pipeline consists of several steps that are executed in a stepwise fashion and an overview is shown in Figure 15. The pipeline starts with mapping the raw sequence reads to the genome using the well-known tool Bowtie2 (Langmead and Salzberg 2012). Next, PCR duplicates that are created during the library construction are removed. The resulting SAM / BAM files are handled using Samtools and Bamtools (Li *et al.* 2009). The main path going forward is to trim the reads to the previously determined optimal trim length using Bamtools. The overlap of reads from both strands are subsequentially counted to produce the genome-wide read profile files. Using the same trimmed reads, GEM (Guo, Mahony and Gifford 2012) is used to identify TF binding peaks. These binding peaks can then be assigned to genes based on their distance to the TSS and result in the gene target list. With looking at only the first nucleotide of each mapped read, one can also create the peak centered read distribution plots as shown in the next section.

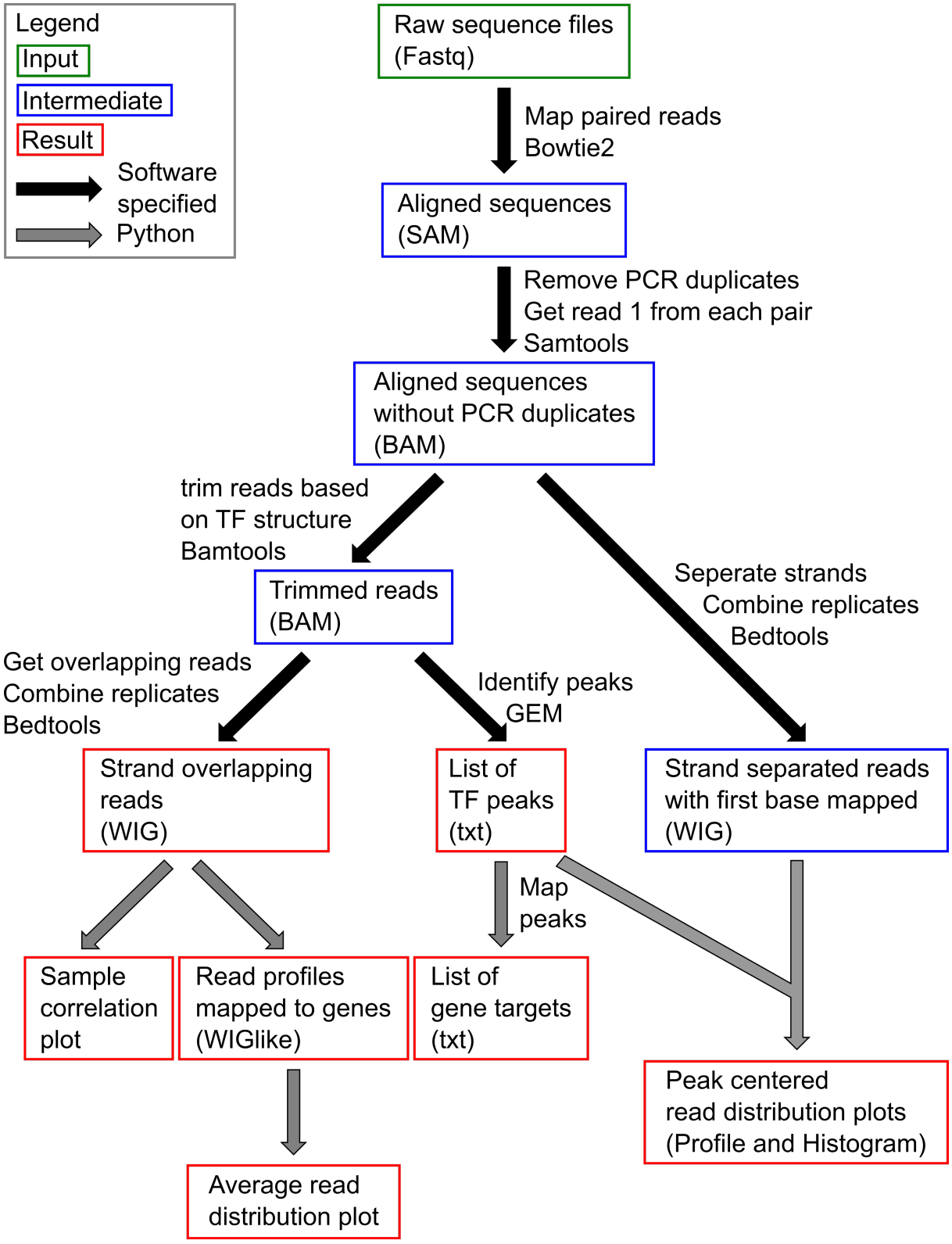

Figure 15: Overview of bioinformatics pipeline for ChIP-exo data. Here, all steps are shown with computational steps represented as arrows and boxes for files. The pipeline starts with a number of computational steps using existing software tools (marked with black arrows), producing intermediate result files (marked with blue boxes). The final analysis of the output is done using custom made python scripts (marked with gray arrows) producing the final output files (marked with red boxes).

3.2.2 Pipeline output

In Figure 16, three of the graphical outputs of the pipeline are shown using the TF Gcn4 in the glucose-limited condition as an example. In panel A, one can clearly observe that there is a strong enrichment of reads upstream of the TSS, matching the region where most TF binding is to be expected. Panel B and C show a detailed view of how reads are distributed on both strands around the detected peaks. This shows that there is a clearly defined border of the TF binding, which one would expect based on how the ChIP-exo reads are generated. From Figure 16 C one can also estimate the overall footprint of the Gcn4 binding to roughly 25 bp, which highlights the high resolution of the ChIP-exo method down into the single nucleotide space.

These output figures can be easily employed for quality control purposes. If there was no enrichment of reads upstream of the TSS as observed in Figure 16 A, or no clearly visible peak borders as observed Figure 16 B and C, this would indicate that the sequencing quality is sub-par. In addition, the pipeline also produces a correlation plot for the individual replicates, which also serves as a quality control step.

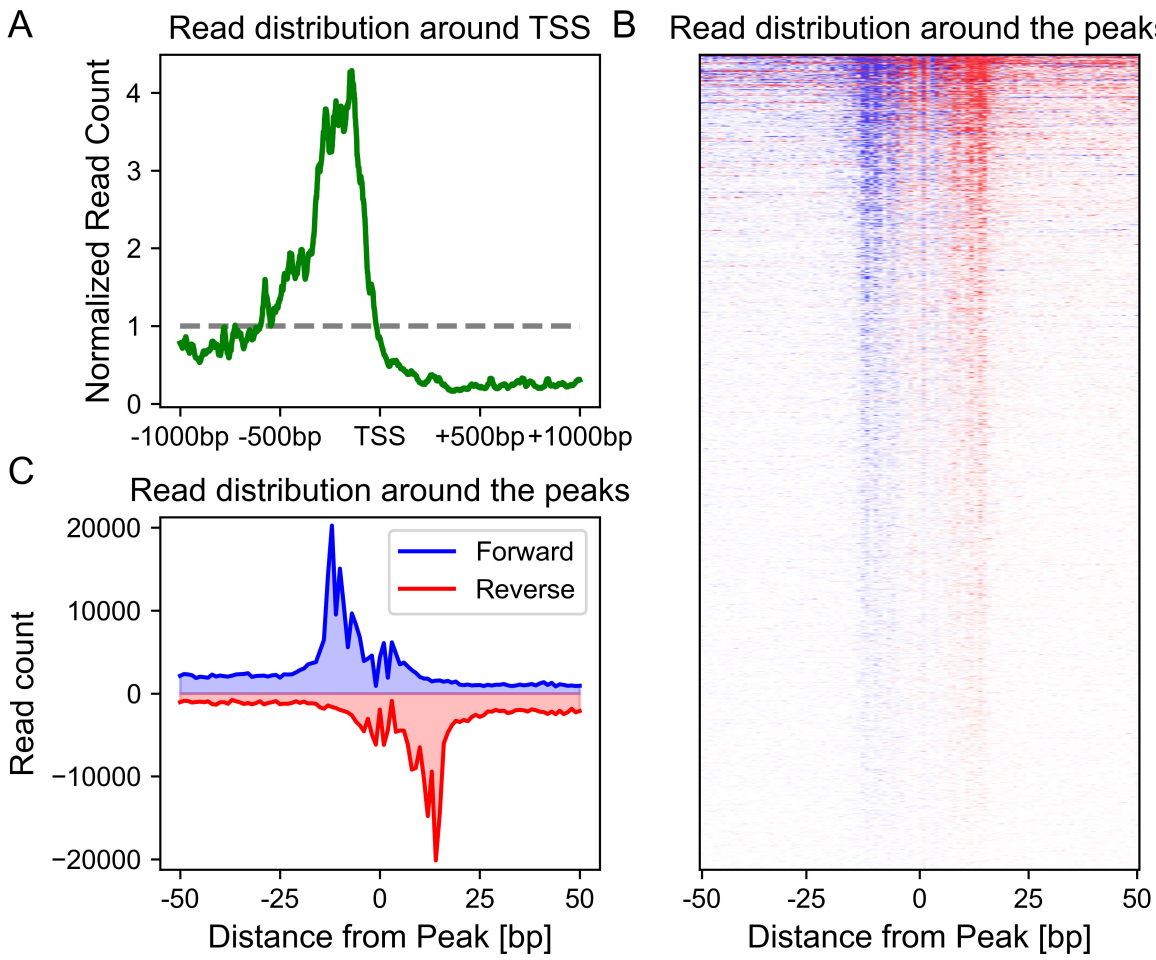

Figure 16: Overview of the pipeline output figures. A: The overall read count distribution (coming from overlapping reads on both strands) is shown across the promoter region. One can observe an enrichment of reads upstream of the TSS. **B:** This histogram shows the read distribution (with only the first base of each read mapped to the genome) around all peaks for Gcn4 in glucose-limited conditions on an individual peak level, where every line corresponds to a single peak. **C:** This plot shows the average read distribution from all the Peaks shown individually in B.

4 APPLICATION OF THE FRAMEWORK

After the successful establishment of the complete framework, I was able to use it to investigate different aspects of TF binding (**Papers III** to **XI**). The applications in this section will range from detailed studies of single TFs to machine learning based analysis of many TFs simultaneously and will also include a web tool for increased accessibility and usability of the gathered TF binding data.

4.1 LINKING TF BINDING TO GENE EXPRESSION

The first project I want to talk about is a large-scale study of many TFs involved in central carbon metabolism (**Paper III)**. In the project the underlying ideas of the analytical framework were already employed, even though the framework was still under development and was therefore not yet published. We collected data for 16 TFs and combined them with previously published data for seven TFs (Bergenholm *et al.* 2018; Ouyang *et al.* 2018) in four different metabolic conditions: (i) fermentative glucose metabolism using glucose limitation in an anaerobic environment, (ii) gluconeogenic respiration using ethanol limitation, (iii) respiratory glucose metabolism using glucose limitation and (iv) aerobic fermentation using nitrogen limitation. Throughout the thesis I will mainly focus on the ethanol and the glucose-limited growth condition. The 21 TFs analyzed here were: Cat8, Cbf1, Ert1, Gcn4, Gcr1, Gcr2, Hap1, Hap4, Ino2, Ino4, Leu3, Oaf1, Pip2, Rds2, Rgt1, Rtg1, Rtg3, Sip4, Stb5, Sut1 and Tye7. They were selected due to their previously published involvement in processes related to the central carbon metabolism, mainly based on ChIP-chip datasets (Harbison *et al.* 2004).

We were interested to see if we can connect the TF binding events, that we measured using ChIP-exo with the gene expression levels that were obtained from RNA sequencing (RNA-seq) in the exact same conditions. To model this we employed a regression method called multivariate adaptive regression splines (MARS) (Friedman 1991). The difference to regular multiple linear regression methods is that MARS can introduce a hinge in the model to account for non-linear behavior if beneficial for the fit. The results for predicting the gene expression levels in the glucose and ethanol-limited condition using the TF binding events are shown in Figure 17. Being able to explain up to 42% of observed gene expression variability with a rather simple model using only 21 TFs, out of more than about 209 TFs in yeast (Hughes and de Boer 2013), is a great start.

The build-in feature selection of MARS was used to only select the TFs with the highest predictive power and the best splines for each selected TF are also shown in Figure 17 (lower panels). One can see that the hinges are used in all cases, mainly to add an unresponsive part to the model. This can be seen as either an activation threshold that

has to be crossed before more binding translates to more expression (e.g. Cat8 in ethanol limited growth) or a saturation effect after which more binding does not correlate with more expression anymore (e.g. Gcn4 in ethanol limitation). The introduction of hinges in all cases also explains why MARS outperformed simple multiple linear regression models, indicating that at least partial non-linear relationships between TF binding and gene expression are occurring.

As we have only mapped a small subset of all TFs in yeast, it would be interesting to see if we can identify subsets of genes for which we have a higher coverage of TF binding events and that we can therefore better predict their expression level. To do this, we clustered metabolic genes based on their relative change in expression levels across the conditions. The resulting 16 clusters had strong enrichments for specific metabolic processes as determined by GO term analysis and are shown in Figure 18.

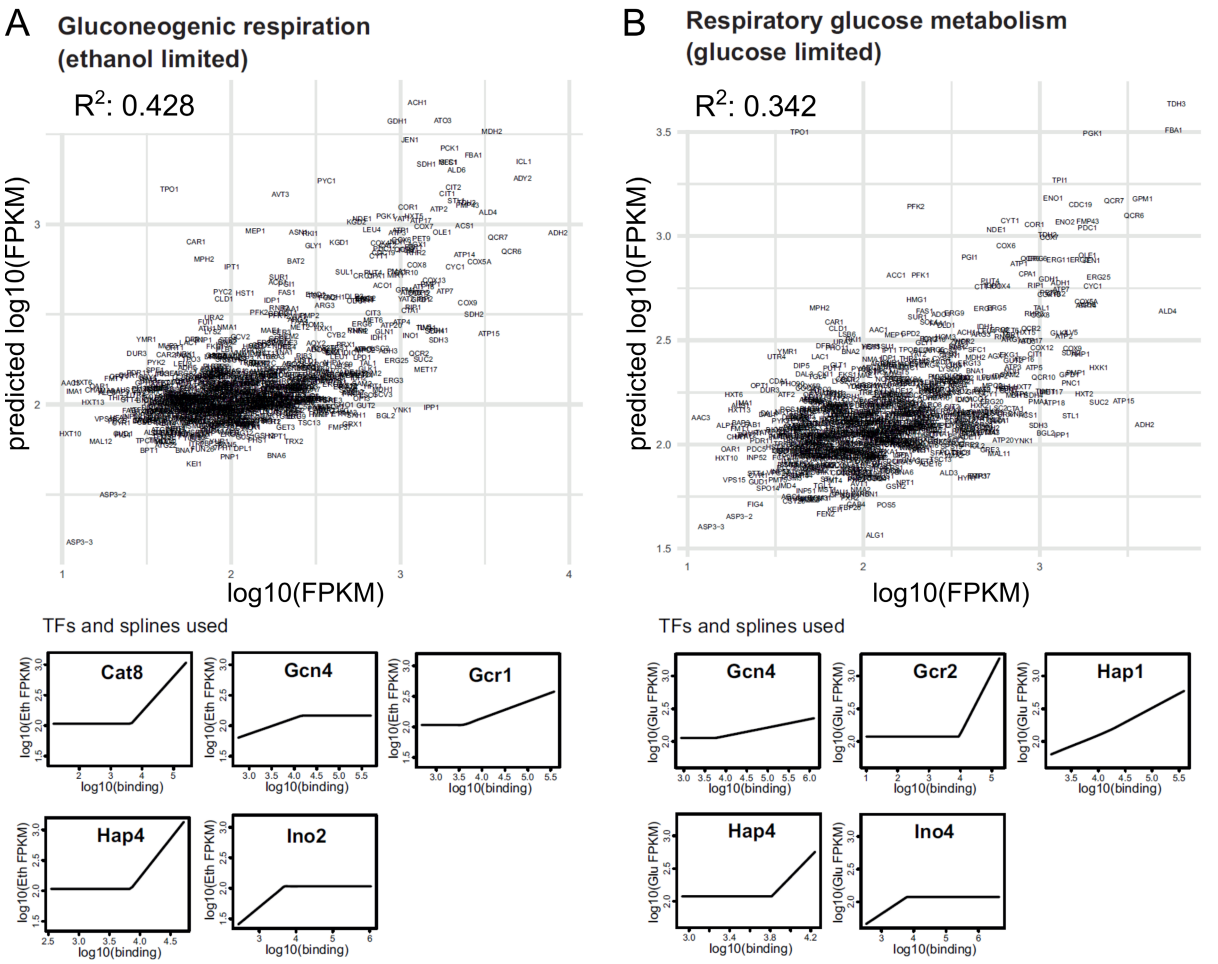

Figure 17: Results for multivariate adaptive regression splines models used to predict gene transcript levels. A: Cross-validated predictions for gene expression levels in ethanol-limited condition. Selected TFs and splines are shown below. **B:** Cross-validated predictions for gene expression levels in glucose-limited condition. Selected TFs and splines are shown below. FPKM: fragments per kilobase per million reads mapped.

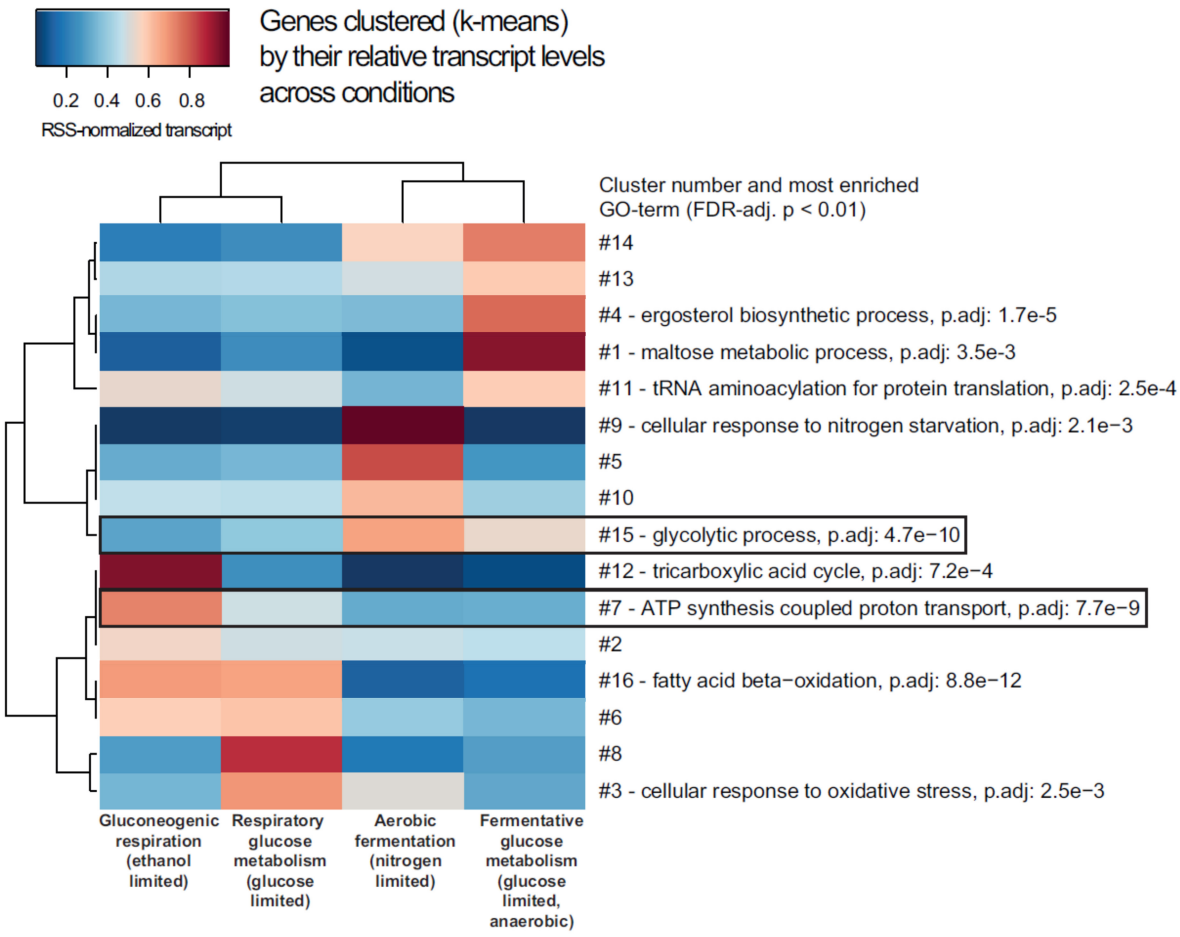

Figure 18: Results of expression-based gene clustering. For clusters which have at least one significantly enriched GO term, the top GO term is shown. The two marked clusters were selected for further analysis.

Next, we used multiple linear regression to predict the gene expression levels of the genes in each cluster using the TF binding events and the results for two of the clusters are shown in Figure 19. The predictive power of the models clearly increased, showing that subsets of genes are co-regulated by similar TFs and mechanisms. The increased ability to accurately predict gene expression levels when focusing on gene subsets (where relatively more TF data are available) also indicates that the amount of available TF binding data is the limiting factor in our current models. Therefore, the approach can be further improved by gathering more data, which would enable us to build stronger predictive models of gene expression based on TF binding data.

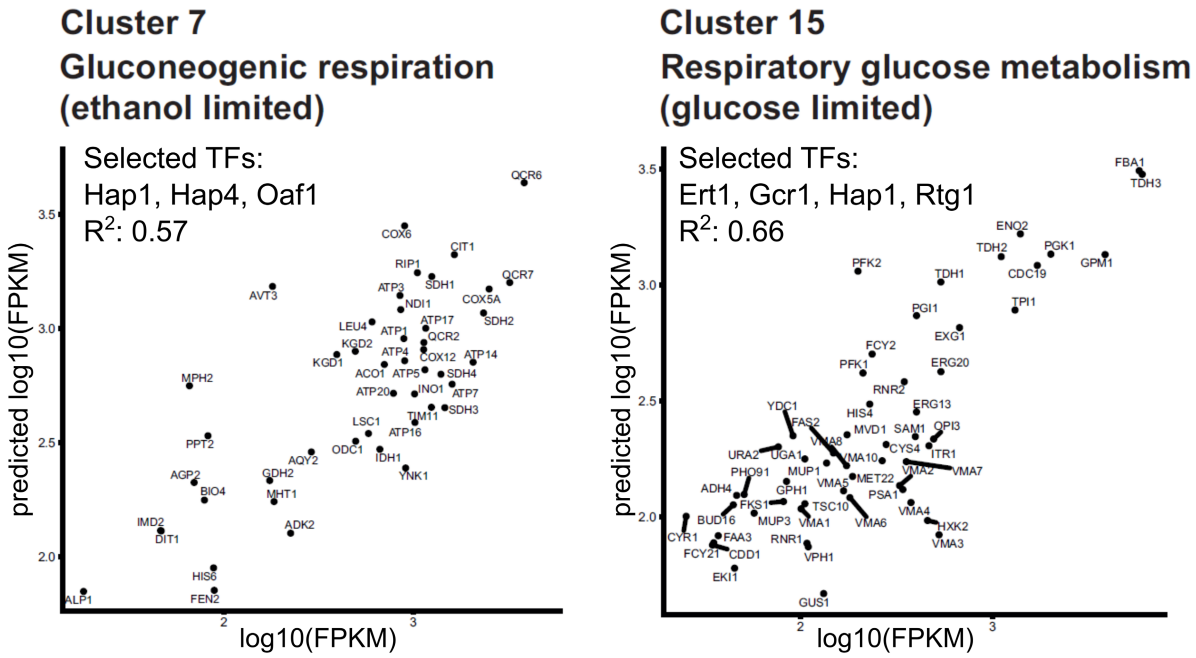

Figure 19: Results for training MARS models on gene clusters. Cross-validated gene expression predictions for two different clusters are shown. The selected TFs as well as the resulting R^2 score is shown. FPKM: fragments per kilobase per million reads mapped.

Using MARS, we could show that employing machine learning approaches can generate insights into the relationships between TF binding events and gene expression levels. It was especially interesting to see that most linear splines got a hinge, indicating that non-linear effects play a role in TF binding. This could mean that by using more complex machine learning models the predictive power could be further increased using our large-scale datasets.

We were not the first to employ machine learning models to link TF binding events with gene expression patterns. There have been studies in other organisms, for example in mouse embryonic stem cells where the authors were able to explain up to 65% of gene expression variability with ChIP-seq data from 12 key TFs (Ouyang, Zhou and Wong 2009). In related studies histone modification marks, instead of TF binding, have been used as input data employing different machine learning models, such as deep learning models (Singh *et al.* 2016).

4.2 ANALYZING CONDITIONAL GENE EXPRESSION CHANGES

After this successful application of machine learning approaches to understand TF binding patterns and their effect on gene expression, I wanted to see how far I can push this and if I could create a machine learning model that would be useful in metabolic engineering applications, which lead to **Paper IV**.

It has been shown before that gene expression levels are not only depended on TF binding events, but are also influenced by a variety of other processes and characteristics, like the shape of the TSS (see **Paper I**), or the codon bias (dos Reis, Wernisch and Savva 2003; Zhang *et al.* 2012). Therefore, a machine learning model trained on TF binding data alone will never be able to capture the full dynamic of gene expression. To overcome this, I decided to focus on the change of gene expression levels between two conditions. This eliminates the effect of many sequence-related features, as the promoter and coding sequence of a gene does not change between conditions. Therefore, one can assume that changes in TF binding patterns are the main driver of differential gene expression. The two conditions chosen are the growth during glucose-limitation and during ethanol-limitation. This decision was based on three reasons: (i) these are the two conditions where we observe the most TF binding events, (ii), a Crabtree positive yeast cell like *S. cerevisiae* grown in an aerobic batch cultivation on glucose will go through both of them during the growth process; and (iii) these two conditions are also of industrial relevance for large-scale fermentation processes. In addition, these two conditions are well studied with many genome wide comparative studies about gene expression levels in both batch and chemostat cultures, showing an extensive reorganization of the central carbon metabolism (DeRisi, Iyer and Brown 1997; Daran-Lapujade *et al.* 2004; Wu *et al.* 2004). This means that there is already extensive knowledge we can use to spot-check predictions from our model later on.

Using RNA-seq data, the gene expression changes between the two conditions were calculated and the distribution is shown in Figure 20 A. One can see that even though there is no global increase or decrease of gene expression, many genes are affected, and the average change is 0.2 log2 fold. Looking at the change of TF binding patterns, there are a total of 7581 TF peaks present in either or both conditions. Of these, 1463 are only present in glucose and 887 are only present in ethanol. For the remaining 5231 peaks the distribution of log2 binding strengths ratios is shown in Figure 20 B. One can observe that there is again no global decrease or increase of TF binding strength, but there are many individual changes with an absolute log2 ratio average of 0.54. This large change, together with the over 2000 peaks that are present in only one condition, shows that TF binding patterns change significantly between the two conditions, which could explain the observed change in gene expression levels.

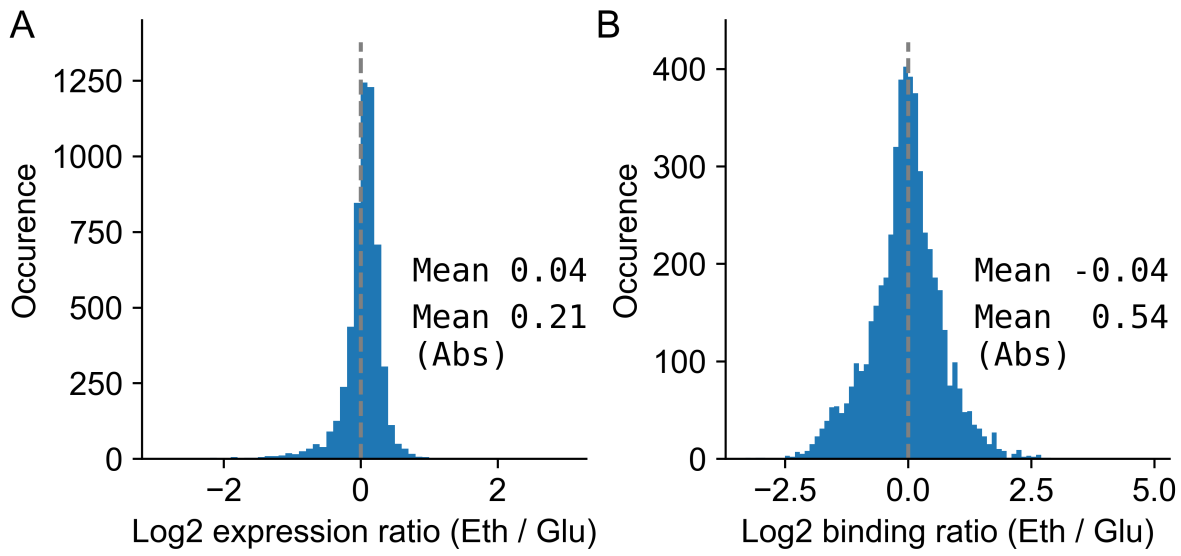

Figure 20: Overview of conditional expression changes between the ethanol and the glucose phase. A: Distribution of log2 expression ratios. **B:** Distribution of log2 binding ratios for all TF peaks present in both conditions. The grey dashed line is at 0 for orientation and the mean of all ratios as well as the mean of the absolute of all ratios is displayed. Bin width is 0.1.

4.2.1 Feature engineering and selection

Improving the feature engineering approach from **Paper III**, I created seven types of features using the data of 20 TFs from **Paper III** (excluding Hap4 due to low quality) and previously published data for Cst6 (Liu, Bergenholm and Nielsen 2016). In order to obtain the seven feature types, first the whole promoter (−1000 bp to +500 bp, relative to TSS) was binned into 50 bp long intervals and the sum of normalized reads for each TF was calculated for binding in glucose, binding in ethanol and the difference between them (ethanol – glucose). In addition, the number of positions with reads were counted for these three conditions; for the difference we distinguish between positions with read counts above zero (where binding in ethanol was stronger) and positions with read counts below zero (where binding in glucose was stronger).

As it is known that TFs often act together, groups of TFs were created to combine them as additional features. These groups were: (i) all TFs together; (ii) all zipper type TFs; (iii) all zinc cluster type TFs; and (iv) the 6 known TF pairs in the data (Ino2 - Ino4, Oaf1 - Pip2, Cat8 - Sip4, Gcr1 - Gcr2, Ert1 - Rds2 and Rtg1 - Rtg3).

These features that were initially based on 50 bp promoter stretches were subsequently combined into four larger intervals to reduce the number of features. These intervals were: (i) full length promoter [from −1000 bp to +500 bp relative to the TSS]; (ii) the first 500 bp of the promoter [−1000 bp to −500 bp]; (iii) the second 500 bp of the promoter [−500 bp to 0]; and (iv) the first 500 bp after the TSS [0 to +500 bp].

This resulted in a total of 840 features, which were the starting point for training a tree-based gradient boosting regression model on the data and score its performance using a five-fold cross-validation scheme. To further reduce the number of features and select the most important ones, a very stringent feature selection was performed using sequential forward-selection, a method that starts with finding the most important single feature and then adds more features step-wise to find the best performance, which was achieved here with only 26 features. The stringent feature selection process led to an impressive gain in performance to a final cross validated R^2 score of 0.519 (starting from 0.362 for all features). Analyzing the importance scores of the features can provide valuable insights into the used data. Here, we split each feature into its four parts: (i) the TF (or TF group); (ii) the interval; (iii) the data processing type; and (iv) the condition (glucose, ethanol, or the difference between them). For each part all scores were aggregated, and the averages are shown in Figure 21.

The two most important TFs were both pairs, Cat8 - Sip4 and Gcr1 - Gcr2. Cat8 and Sip4 are both involved in the activation of gluconeogenesis (Hedges, Proft and Entian 1995; Hiesinger *et al.* 2001), while Gcr1 and Gcr2 have been shown to play an important role in the activation of glycolysis genes (Chambers, Packham and Graham 1995). This encouragingly shows that our ML model learned true biological knowledge, giving us greater confidence in the model and our approach.

Overall, Cat8 was the most important TF in our model, as it is also part of the zipper type TF group (rank 3) and was on rank 4 on its own. This indicates that Cat8 plays an even bigger role in regulating conditional gene expression changes in the glucose and ethanol condition than was previously known and should warrant future studies about it. Besides the feature importance of TFs, we also have data about the other feature types and looking at the intervals, we can observe that the most important interval was the 500 bp directly upstream of the TSS, where also most TF binding events were taking place. This again is an indicator that the model learns biological relevant patterns and not noise. Interestingly the 500 bp interval downstream of the TSS was more important than the 500 bp stretch that was located 500 bp upstream of the TSS. This would indicate that binding of TFs downstream of the TSS play a role in regulating expression in addition to any binding events occurring upstream of the TSS.

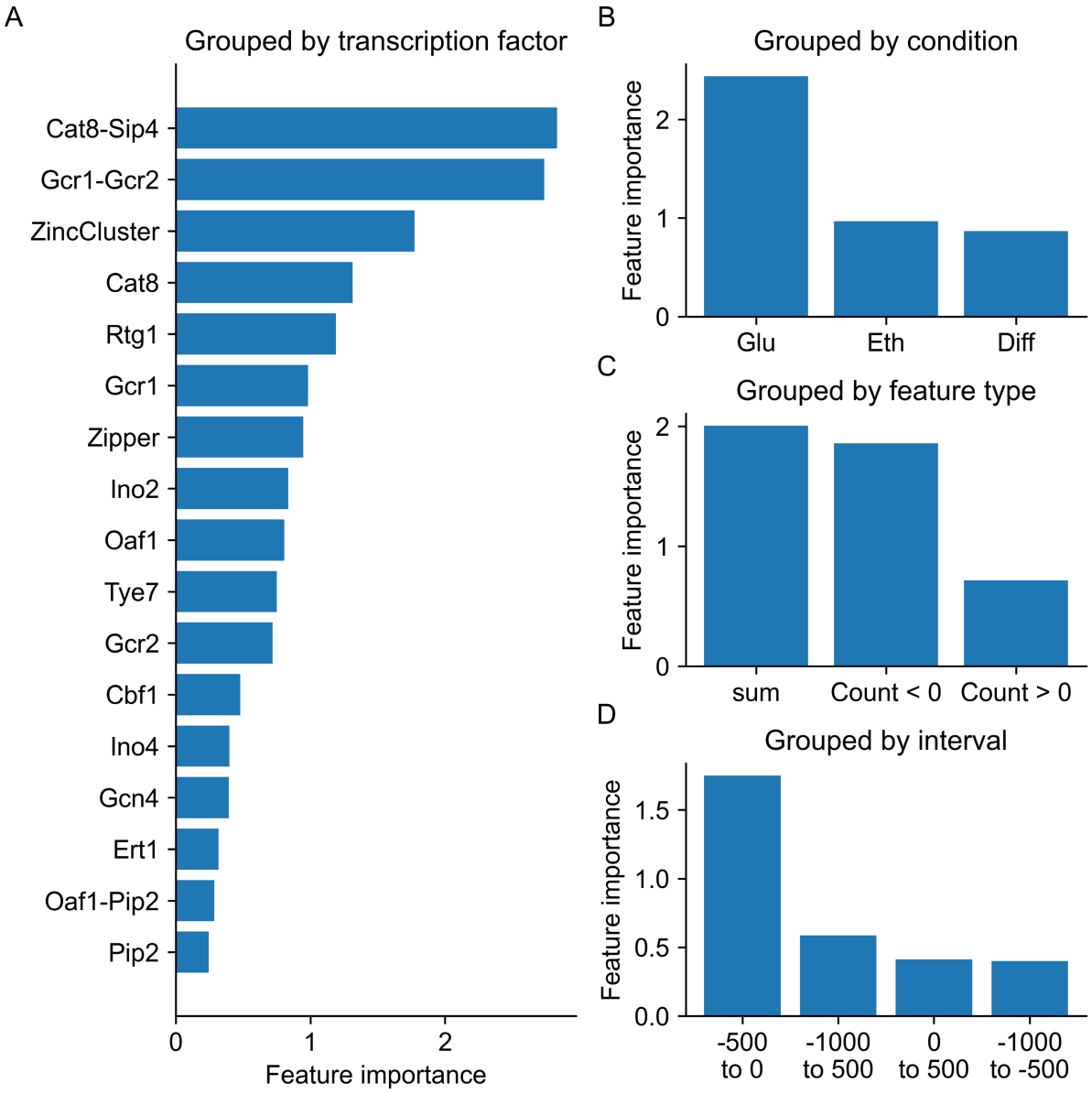

Figure 21: Feature importance plots. The 26 selected features were aggregated by either **(A)** the transcription factor, **(B)** the condition, **(C)** the type of feature or **(D)** the interval. For all groups the feature importance values were averaged and displayed.

4.2.2 Promoter engineering

Fine tuning the conditional gene expression behavior of promoters is of interest in metabolic engineering projects, because it can be used to increase the productivity of fermentation processes. Examples for this are the usage of promoters induced by high glucose levels like *HXT1* (Scalcinati *et al.* 2012; Teixeira *et al.* 2018) or by ethanol, like *ADH2* (Kealey *et al.* 1998) and *ICL1* (Maury *et al.* 2018). Besides direct application of conditional promoters in metabolic engineering projects, other studies aimed to support these approaches by providing datasets of conditional promoters (Peng *et al.* 2015). Because of my strong interest in metabolic engineering and the transition towards a bio-based economy I developed an online promoter engineering tool, called HYENA (**Hy**brid promoter d**e**sig**n** using **a**dvanced transcription factor binding predictions) to further aid in this endeavor.

The workflow of HYENA is displayed in Figure 22. After the user selects a promoter of a metabolic gene to engineer and the desired gene expression ratio between the ethanol and glucose phase, the tool creates hybrid promoters by exchanging a 50 bp long stretch of the upstream promoter region (−750 to −250 bp relative to TSS) with the corresponding 50 bp stretch from another promoter. This is done for each of the 10 short regions and 1037 other metabolic genes. In the next step, the expression ratios for these 10370 hybrid promoters are predicted using the machine learning model and then ranked according to their difference to the user-designated expression ratio. The top candidates for feature engineering are displayed, together with their hybrid promoter sequence. HYENA is available online at https://hyena-toolbox.herokuapp.com/ and the source code is available through GitHub https://github.com/SysBioChalmers/Hyena.

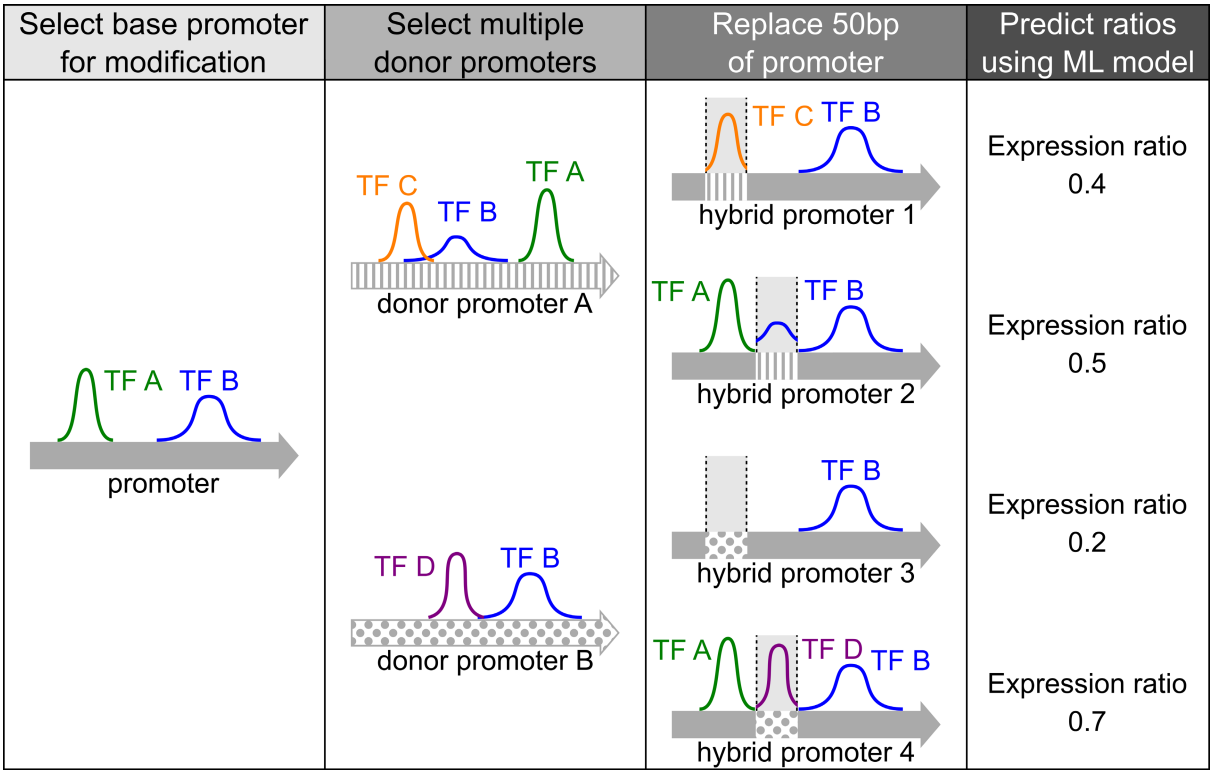

Figure 22: Promoter engineering strategy in Hyena. A 50 bp stretch of the promoter is replaced by the corresponding 50 bp stretch of different donor promoters creating hybrid promoters. These are then evaluated using the machine learning model and ranked according to the difference in predicted expression ratio to the desired expression ratio.

After creating the model, the next step is to validate it. In absence of actual measurements of these hybrid promoters and their expression patterns, can we instead use previous data and knowledge for an initial validation? In a first step I ran HYENA to independently increase or decrease the expression ratio of each gene by 0.5, thereby making its expression pattern more pronounced in either the ethanol phase (increased ratio) or in the glucose phase (decreased ratio). I then analyzed which three donor promoters were chosen most frequently, as shown in Table 1.

Table 1: Overview of common donor promoters chosen by HYENA.

# top donor promoter	Stronger in glucose	Stronger in ethanol
1	*TPO1* (483 times)	*ATO3* (305 times)
2	*CDC19* (239 times)	*HXT3* (219 times)
3	*ENO1* (180 times)	*ADY2* (148 times)

The first observation is that half of the top candidates are involved or directly linked to glycolysis (*CDC19*, *ENO1* and *HXT3*). As the flux through glycolysis dramatically changes between the two conditions, it is not surprising that this pathway harbors great donor genes and beneficial properties have already been shown for promoters of other glucose transporter genes, like *HXT1* (Scalcinati *et al.* 2012; Teixeira *et al.* 2018). This indicates that these top donor genes could indeed cause the desired change in gene expression ratio, providing a first validation step of HYENA. Interestingly, four out of the six most common donors encode transport proteins involved in quite different metabolic areas, the transport of acetate (*ADY2*), ammonium (*ATO3*) glucose (*HXT3*) or polyamines (*TPO1*). It makes sense that transporter genes show a very condition specific gene expression pattern as they are not always required, while this further indicates that transport proteins are potentially heavily regulated by TF binding events, rendering other transport proteins also interesting donors for future modifications.

Besides the commonly selected promoters in Table 1, with for example *TPO1* being selected for nearly 50% of all 1038 metabolic genes, I would like to note that in this initial test of HYENA, 390 different donor promoters were proposed by HYENA at least once to increase the expression in ethanol, 82 of these at least in five occasions. In the opposite direction, the numbers are very similar with 338 and 64 respectively. This demonstrates that there is a vast variety of potential donor promoters that can be used for fine tuning of conditional gene expression levels, in addition to the few that are currently already used in laboratories, and that this can only be efficiently unlocked using a machine-learning approach.

In **Paper III** and **IV** I demonstrated how to use machine learning models to gain insight into the effects of TF binding patterns and how one can use these models to guide promoter engineering. Both showed promising progress and highlighted what can be done using ChIP-exo data. Both of these two big data approaches have only been made possible by having a large data set of ChIP-exo data available, which was processed in a fast and reliable way using our analytical framework. Therefore, the framework I developed can serve as the basis for future large-scale studies involving more TFs and more conditions to further improve the knowledge we have and allow us to even better regulate conditional gene expression. Taken together this will hopefully help and contribute to improved metabolic engineering strategies facilitating the transition towards a bio-based economy.

4.3 INVESTIGATING CONDITIONAL BINDING OF LEU3

After these big data inspired studies, I would like to switch gears and talk about a detailed investigation of a single TF, namely Leu3, the main regulator of leucine biosynthesis (**Paper V**). Leu3 is a zinc knuckle type transcription factor that is activated by binding to alpha-isopropylmalate, an intermediate product of leucine biosynthesis (Kohlhaw 2003). It has been shown previously that Leu3 binds to the promoters of the majority of genes involved in leucine biosynthesis like *LEU1*, *LEU2*, *LEU4*, *ILV2*, *ILV3*, *ILV5* and *BAT1* (Kohlhaw 2003; MacIsaac *et al.* 2006). Interestingly Leu3 has been reported as being one of the few TFs that are always bound to their target gene, independent if it is activated or not (Kirkpatrick and Schimmel 1995; Harbison *et al.* 2004). This contrasts with what we observed in a large-scale study of multiple TFs using ChIP-exo in the four chemostat conditions (see **Paper III** for more details).

Because our large-scale study was based on four vastly different environmental growth conditions it was impossible to connect this differential binding behavior of Leu3 directly to leucine availability. In order to elucidate this, I designed a simplified growth experiment were the only changing environmental variable was the availability of leucine or the two other branched-chain amino acids (valine and isoleucine) in the media. An overview of the setup is shown in Figure 23.

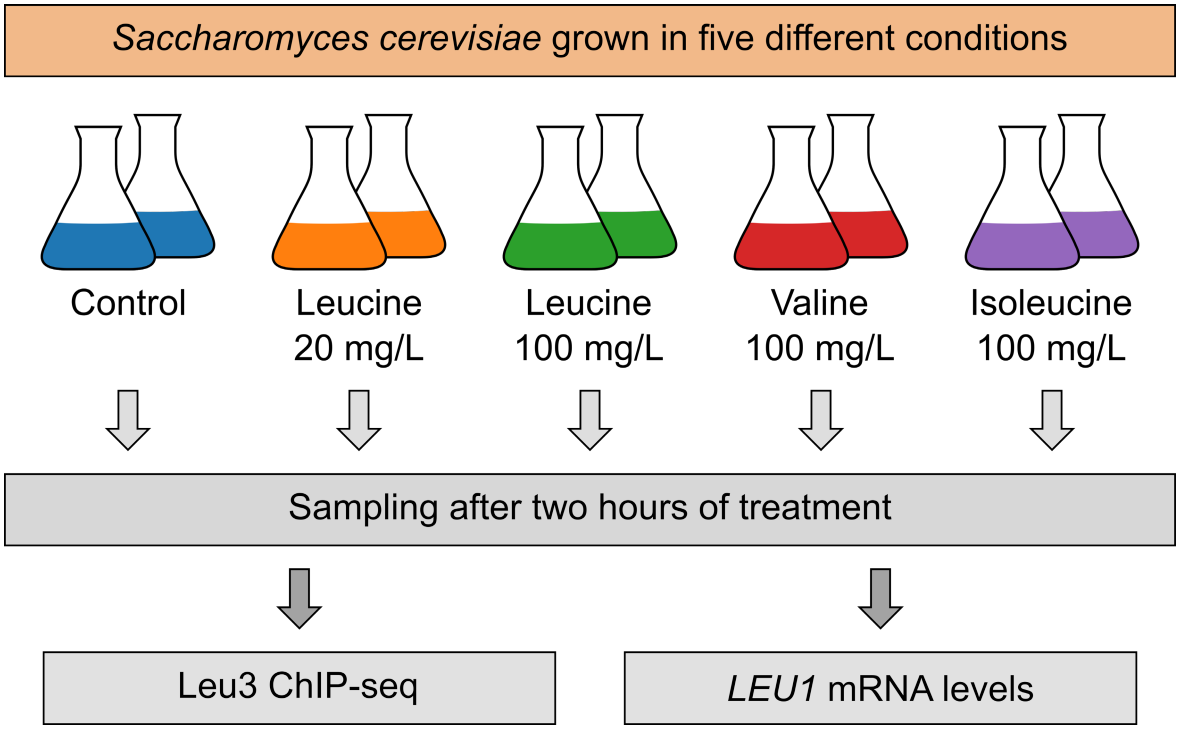

Figure 23: Overview of experimental setup for measuring Leu3 activity. The yeast cells were grown in different media compositions in shake flasks and after the treatment the cells were collected for ChIP-seq and qPCR experiments.

In short, I grew the cells in shake flasks using slow-release glucose feed beads to achieve a glucose limited condition, similar to the chemostat setup. After addition of the branched-chain amino acids I harvested the cells for ChIP-Seq and qPCR analysis. The qPCR analysis of *LEU1* mRNA levels was performed to validate the experimental setup and a strong reduction in *LEU1* expression was observed as expected.

Next, I mapped the TF binding sites, which was done by ChIP-seq. As this was a smaller-scale experiment, I decided to use ChIP-seq instead of ChIP-exo and trade the high resolution for a higher throughput. In addition, the sequencing was done using the novel Oxford Nanopore Sequencing Technology, where full length long reads are sequenced by translocation through a pore (see (Jain *et al.* 2016) for a comprehensive review of the technology). Using this setup, I achieved a sufficiently high quality for performing robust ChIP-seq analysis. The bioinformatics pipeline as introduced in **Paper II** was designed for ChIP-exo, therefore some small adjustments had to be made so that it was usable in this project, but the overall workflow stayed the same. The reads were first mapped to the genome using Bowtie2 and the peaks were detected using GEM. The numbers of detected peaks per condition are shown in Figure 24 A.

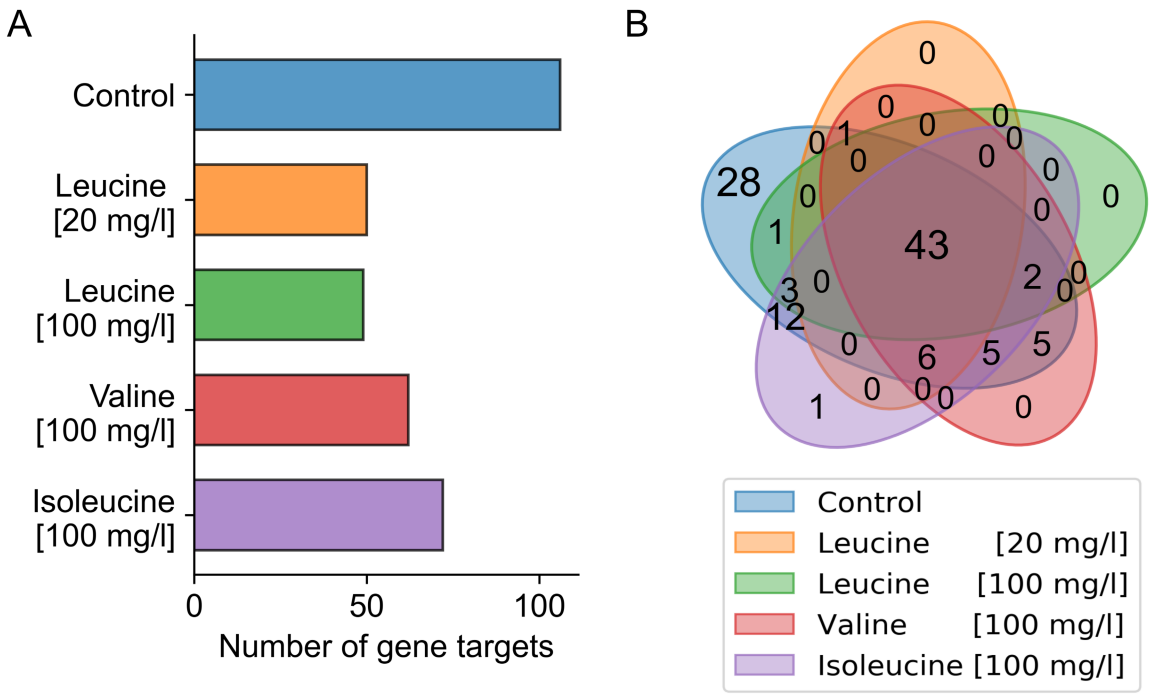

Figure 24: Detected gene targets for Leu3. A: Number of detected Leu3 binding targets for each of the for each of the five conditions. Only peaks detected by GEM with a signal to noise ratio of ≥ 2 that are within 1000 bp of a TSS were taken into account. **B:** Distribution and overlap of the Leu3 gene targets between the five conditions.

The number of gene targets per condition responded to the availability of the branched-chain amino acids in the media, with the most targets bound in the control condition and less than half of these in either leucine condition. In Figure 24 B, the overlap of gene targets between the different conditions is shown, and less than 50% of all detected 107 gene targets are bound in all five conditions. This is in contrast to earlier studies which concluded that Leu3 always binds to its gene targets (Kirkpatrick and Schimmel 1995), but matches with what we have observed in the large-scale ChIP-exo study (**Paper III**).

Investigating the binding behavior of Leu3 on genes involved in the leucine biosynthetic pathway as shown in Figure 25, one can observe that even though these gene targets are always bound, the binding strength is significantly decreased. The addition of leucine to the media reduced the binding strength by more than 50%. This reduction without abolishing binding, could explain the conclusion from the earlier studies concerning the quite unique binding behavior of Leu3.

This finding nicely demonstrates the value of large-scale studies that are then coupled with detailed investigations of specific aspects to redefine our knowledge about biological processes.

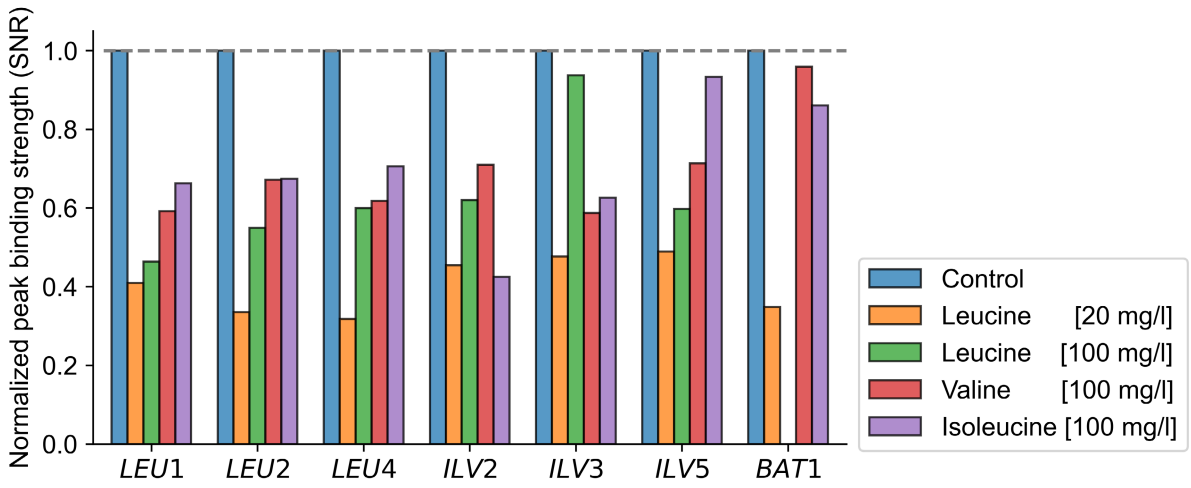

Figure 25: Leu3 binding strength on leucine metabolic genes. Peak binding strength (signal to noise ratio, SNR) normalized to control levels of Leu3 on genes involved in leucine biosynthesis in the five different conditions used.

4.4 IMPROVING DATA ACCESSIBILITY

I strongly believe that making research data and results as open and accessible as possible is important and should always be considered. Therefore, I am very proud to have contributed to an R Shiny online app that makes all the gathered ChIP-exo data and analysis easily accessible. Here, I will present **Paper VI**, the *S. cerevisiae* Transcription factor Explorer (T-rEx), available at https://www.sysbio.se/tools/trex/.

The main goal of T-rEx was to make all the gathered ChIP-exo data, that was processed using our published framework (**Paper I** and **II),** available. On the TF overview page, one can see the gene targets for each TF and obtain the consensus motif, the sequence map as well as the peak and read distribution plots, all produced by our analytical pipeline. An example overview page for Gcn4 is shown in Figure 26.

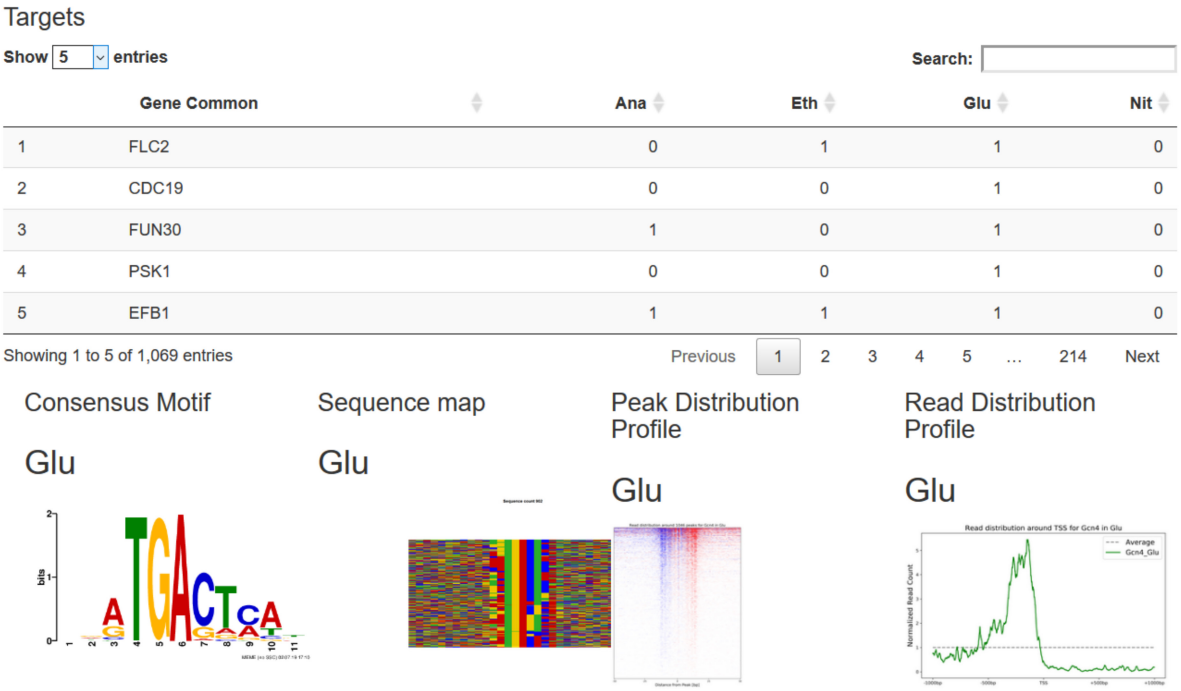

Figure 26: T-rEx overview page; showing pipeline outputs for Gcn4. For the selected TF the gene targets in each condition are shown. In addition, other characteristics like the consensus motif or the peak distribution profile are shown.

Besides these summary plots one can also obtain the complete binding profiles for all TFs on each individual gene. As an example, the binding pattern of Gcn4 and Leu3 on *ILV2*, a gene involved in branched-chain amino acid metabolism, is shown in Figure 27 for two different conditions, glucose-limited and nitrogen-limited. One can clearly see that both TFs show a single strong peak upstream of the TSS in the glucose-limited condition, with Gcn4 being located closer to the TSS. In comparison, the nitrogen-limited condition only exhibits a weaker Gcn4 binding peak and no peak for Leu3. This again shows the conditional binding behavior of Leu3 (see **Paper V**).

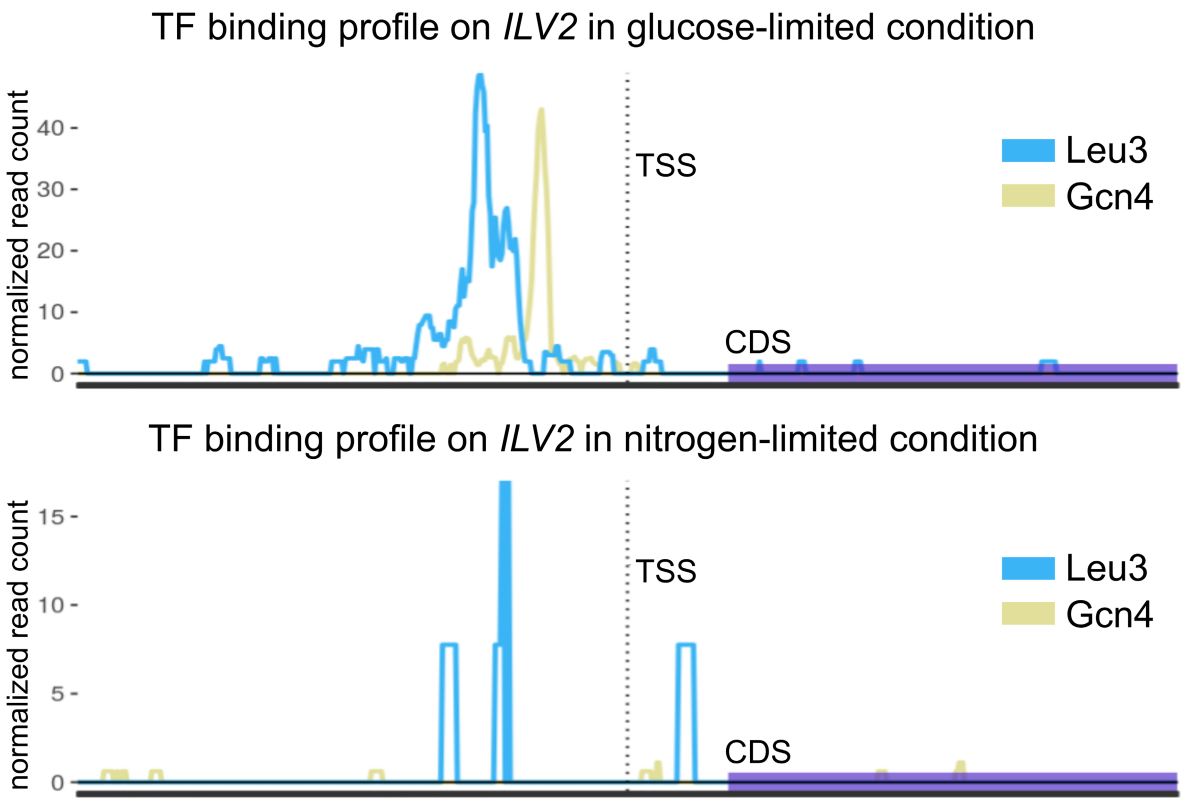

Figure 27: T-rEx binding profile output; showing Gcn4 and Leu3 on *ILV2* in two different conditions. TSS = Transcription start site, CDS = Coding sequence start.

The tool also allows the user to zoom into the graph and inspect the genome sequence at this region. This could for example be used to identify the underlying motifs or to design guide RNAs for promoter engineering using CRISPR/Cas9.

T-rEx is not only a powerful tool for visualizing ChIP-exo data and TF characteristics, it also includes a number of analytical tools. The user can select genes based on GO terms and for example plot which TFs bind to these genes. This allows the user to obtain a quick overview of which TFs are relevant in that subset of genes and how the TFs and gene targets cluster together. Such a heatmap is shown in Figure 28 for 16 genes related to branched-chain amino acids. Gcn4, the master regulator of amino acid metabolism and Leu3, the main regulator of leucine biosynthesis, are the two TFs with the most binding targets. Ino2 and Ino4 also bind to a number of gene targets, and this two TF, that often bind together as heterodimers, also binds to the exact same targets. From the seven Ino2 / Ino4 targets, five of them are also bound by Gcn4, while two (*AVT4* and *PDC1*) are not. Therefore, such a heatmap can be useful to identify the gene targets and to identify subsets of genes that are bound by specific TF or TF pairs.

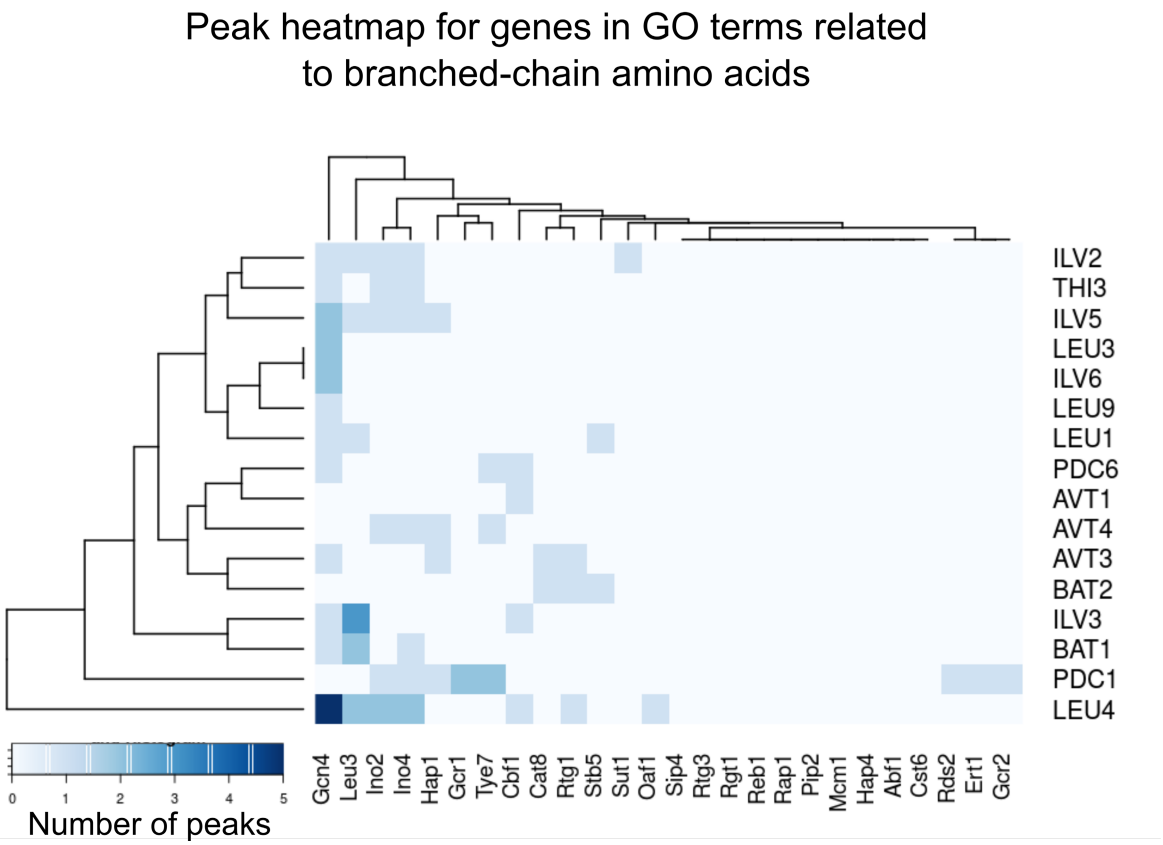

Figure 28: T-rEx analytical options; heatmap for number of peaks on genes related to branched-chain amino acids. This heatmap can be used to identify relevant TFs and their targets in addition to showing possible subgroups with distinct TF binding patterns.

To summarize, the developed open-source online app T-rEx is a powerful tool to investigate high-resolution TF binding data in an accessible and user-friendly way. It can for example be used to find TF targets as well as perform various analytical steps, like cluster genes together based on their TF binding pattern or to find TFs that bind together to the same subset of genes.

5 CONCLUSION AND OUTLOOK

Here in this thesis I presented the development of an analytical framework to study transcriptional regulation in the yeast species *S. cerevisiae* and give examples for its many possible applications.

The framework that I built was based on two projects. The first project aimed to establish the exact transcription start sites for all expressed genes in a set of industrially relevant growth conditions using the state-of-the-art method called Cap Analysis of Gene expression (CAGE) (**Paper I**). With that data we were also able to show that the transcription start sites in yeast are stable across the chosen conditions, which was an important piece of knowledge needed for the larger framework. The second project aimed to create a unified data processing pipeline for analyzing transcription factor binding sites from ChIP-exo data using a combination of published software tools and custom written scripts (**Paper II**). The focus of this pipeline was to include extensive quality control measures to ensure that the obtained results are reliable.

Together, this framework enables us to process large quantities of ChIP-exo data in a fast, reliable and unified way to increase the usability of the gathered data. The usefulness of this framework is demonstrated with the four different projects described in this thesis, covering a wide range of possible applications, ranging from data visualization and accessibility (**Paper VI**), to single TF studies (**Paper V**), to large scale studies using advanced machine learning methods (**Paper III and IV**).

In believe that the use of machine learning models to study transcriptional regulation in yeast using TF binding sites as model inputs is a promising direction for further research and development. This view is based on several aspects: (i) the data we have already gathered show how complex the individual regulation is with a plethora of interactions, which will be otherwise difficult to decipher; (ii) the increased data availability makes machine learning approaches more viable; (iii) we are getting better at understanding how the machine learning models are working, and how to make them more explainable, which will help in getting biological insights through their application. In addition, I think it is a promising direction because of the great potential this has for applications in the area of metabolic engineering, for example through the design of custom hybrid promoters with clearly defined conditional response patterns.

All in all, this thesis provides a reliable framework to study transcriptional regulation in an easier and more straightforward fashion than previously possible with an increased range of potential applications, especially using machine learning methods to guide metabolic engineering. The establishment of this framework was made possible by the continuous technological and methodological progress in regard to sequencings

technology as well as machine learning and I therefore anticipate that this framework will continued to be used to investigate TF binding patterns and their connection to transcriptional regulatory events.

So, talking about the future, what would be the next step in continuing this project? To date, we have mapped only around 10% of all the TFs in *S. cerevisiae*, therefore one should focus on performing more ChIP-exo experiments to gain a better idea of the whole regulatory network and not only focusing on central carbon metabolism. Personally, I would first extend the list of studied TFs by adding more TFs related to amino acid metabolism (besides Leu3 and Gcn4 that are already mapped), as that is a vital part of metabolism and closely connected to the central carbon metabolism we already cover reasonably well. After that I would step aside from the core metabolism and instead start investigating the many TFs involved in stress response mechanisms. After achieving a good coverage of all TFs (> 70%) one should start focusing on mapping them in more conditions, because I think at that point the added value of the next TF is quite small in comparison with gaining insights into a completely new condition. Besides the four condition already presented in this thesis, I think one should focus on industrially relevant conditions, including high temperature stress and exposure to different acids and growth inhibitors. It would also be interesting to study the influence of different growth rates on TF binding patterns. With all of this additional data, our knowledge of the whole regulatory network would increase dramatically, and the different machine learning algorithms that are employed would become so much more powerful.

But is this goal actually achievable with the methods we have available to date? In order to perform such a large scale experiment in a reasonable timeframe one has to invest in automation because otherwise just performing the necessary ChIP-exo experiments to map all TFs for a single condition will take roughly a year of full time work and then one still has not created a single tagged strain or run the cultivations. I would also love to see more than ten conditions mapped in different growth rates, further increasing the number of experiments that have to be done.

So, can one automate the ChIP-exo experiment? Yes, one can. The new ChIP-exo 5.0 protocol (Rossi, Lai and Pugh 2018) can be performed using magnetic beads for all of the individual washing and purification steps, meaning that it can be performed by most pipetting robots in 96 well plates using a magnetic stand as is already in use for many other sequencing libraries preparations. Besides the ChIP-exo experiments the other aspect of this project with much hands-on time is the bioreactor cultivation. I do not think that we will be able to completely automate that in the coming years, but with different companies working on smaller reactors for increased throughput, the hands-on time per single TF will at least be significantly reduced.

The only part of this project that will be rather difficult to automate is the creation of the strains with a tagged TF and the necessary validation, but as this has only to be done once for every of the ~209 TFs, it is at least not a recurring work.

As exciting as getting access to all this data would be, I have to admit this plan is largely just more of the same, so how could one move this project forwards even further? The framework I have created should be easily adaptable to single cell ChIP-exo as soon as such data becomes available. I guess that we will see a single-cell ChIP-exo protocol rather soon, as single cell ChIP-seq was already published (Rotem *et al.* 2015) and adapting it to ChIP-exo should pose no unmanageable obstacles. This would provide valuable insights into the cell-to-cell variability of TF binding patterns, something that we are currently lacking completely, without sacrificing the great resolution of ChIP-exo. It would also enable measuring the percentage of cells having the TF of interest bound at a given location in the genome, making the comparison of the binding strength of different TFs easier and more reliable.

To further improve the insights into cell-to-cell variability and the presence or absence of different subpopulations, one could try to combine single cell ChIP-exo with simultaneous RNA-seq of the same cell. This would allow us to directly connect the TF binding event observed in a subpopulation with the transcriptional activity levels in that subpopulation. This sounds quite farfetched, but something very similar has already be done by combining single cell ATAC-seq and RNA-seq (Reyes *et al.* 2019), so there should be no fundamental obstacles. As single cell ChIP-exo is however not yet available I unfortunately do not expect that technology to become widely available in the next couple of years.

Assuming that in maybe ten years' time we have finally mapped all TFs, using single cell ChIP-exo combined with simultaneous single-cell RNA-seq, would we then be able to understand everything? Unfortunately, no. There would still be some specific binding patterns that we would not be able to resolve completely. To better explain this let me give an example: assume that we have identified a subpopulation of cells based on a gene expression pattern from RNA-seq and we measured that at a single location in the promoter of a gene Ino2 as well as Ino4 is bound in 50% of all cells. The interesting thing about this pair is that they are known to bind either as homodimer or heterodimer. This could mean that in 50% of the cells the heterodimer Ino2-Ino4 is binding and in the other half there is no binding of Ino2 or Ino4 at all. It could however also mean that we have up to four distinct sub-subpopulations of equal size: 25% of the cells are bound by a Ino2 homodimer, 25% by a Ino4 homodimer, 25% by a Ino2-Ino4 heterodimer and the last 25% are not bound by either of them (many other percentage distributions would also be possible). This means that even with single-cell ChIP-exo one would not be able to identify all occurring sub-subpopulations.

So, is there a possible way to overcome this and push the mapping of TF binding events into a completely new level? Yes, I believe there is. This will however not be achieved by using any sort of ChIP method as this inherently implies that we can only measure one TF at the time. Therefore, one would need to swap out the antibody-based enrichment step with a procedure that can identify all the proteins bound to the DNA at the same time. This could be done by a mass-spectrometry based methods, but I believe that the way to go in protein identification is through a protein sequencing device based on a nanopore. This concept is already working for DNA (and was also employed in this thesis), and recently a very important milestone for protein sequencing was reached, as it is now possible to distinguish all 20 amino acids from each other using an aerolysin nanopore (Ouldali *et al.* 2020). The added advantage of such a method is that it would dramatically cut the number of experiments necessary to map all TFs in a single condition from ~209 to 1 (ignoring necessary replicates), meaning that it would be able to map so many more conditions in a reasonable timeframe, one could even start to perform it on gene-knockout strains to measure the effect of that knockout.

But this simultaneous TF mapping method is so far down the road that for now we should focus on gathering as much ChIP-exo / single cell ChIP-exo data as possible and with the framework developed in this thesis this will be easier than before.

All in all, I think that the field of studying transcriptional regulation is currently at a very exiting stage, because we are just gathering momentum in collecting high quality data in high resolution and there are many possibilities for future methodological advances. In addition, the improved applications of machine learning to the vast data that is already collected or that will be collected in the future means that we will be getting better at understanding transcriptional regulation in yeast and that the possible applications for modulating transcriptional regulation will increase. Another interesting venue for analyzing TF binding patterns in the future will be through combining it with genome scale metabolic models that are continuously getting better and include more features. This could pave the way to a complete *in-silico* model of a living eukaryotic cell, including transcriptional regulation and continues adaptation to an ever-changing environment.

It will be therefore very exciting to see where the field is going to be in the next five to ten years and I am proud to have contributed to its advancement.

6 ACKNOWLEDGEMENTS

First of all, I would like to thank Jens Nielsen, my main supervisor, for giving me the opportunity to write my PhD thesis in his lab and giving me all the support and academic freedom I could have asked for. I also want to say thank you to my two co-supervisors, Verena Siewers and Eduard Kerkhoven for doing a great job in supporting me throughout this journey. Thanks to Dina Petranovic, for being my examiner and thanks to Chris Workman for being my opponent.

Before I start thanking different groups of people at SysBio, I would like to thank all of you for making SysBio such an amazing work environment, with a great culture. Essential to this are all the current and former members of the Core Value Team. I am proud to have also been part of the CVT and I want to say a special thanks to Kate, Yassi and Francesca, my co-authors on the Nature correspondence piece about working culture, it was a lot of fun to write it with you. Very important for our well-functioning lab are all the wet and dry lab research engineers that enable us to focus more on our work and less on lab organization and maintenance, and everyone involved in the administration who are always there to help. Thank you.

A big thank you to David, Petter and Liming, my teammates working together on the different transcription factor projects, without you I would not have been able to make that much progress and develop all the cool tools. I also want to say thank you to Michi and Oliver for your help in the wet-lab and for making working in the lab so much more fun. Thanks to our daily lunch group at 11:30 sharp, mainly consisting of Michi, Louis, David, Carl, John and Olena. Thanks also to the very active fika group around Verena, Christer, Joakim, Ed and Michi. I will definitely miss working with all of you.

Thanks also to Nevena Cvetesic and Boris Lenhard from the Imperial College in London, for the great collaboration on performing the CAGE experiment and welcoming me in your lab.

Besides working I was also engaged in quite a number of leisure activities at SysBio and I want to thank the different groups of people who were involved. Thanks to everyone involved in the amazing SysBio Ski-trip (especially Ed, Verena, Ana, Michi and Oliver) and I hope that I will now also be invited for joining next year again. Thanks to the SysBio lifting group (Michi, JC and Oliver), the climbing group (Oliver, Ben, Gatto and David) as well as the squash group (mainly Ed, Kate and Raghu). Apart from doing sports I also always had fun joining the pub-quiz sessions with Verena, Ana, Michi, Ben, Jon, Dan, and others. Friday evenings at Foxes were also a very important event and I hope to see Michi, Verena, Ed, Louis, Lauren, Jon, Suvi, Tyler, Lucy, Martin, Ben, Avlant, Gatto and many others there again soon (or at least soonish when its safe again). Board games are also always fun to play and for this thank you to Oliver, Marta,

Dany, Max, Veronica, John and Paulo. Special thanks also to Lucy and Tyler for often hosting us, especially Michi and me for our Sunday football sessions with tons of amazing food.

I also want to thank everyone involved in my ITN PAcMEN (especially Gang, Paul, Helén, Anna and Paul) for all the great meetings and discussions we had, and of course also to the European Union for funding my PhD.

I would also like to thank my parents and my sister & her family (especially Maya who I hope will read this thesis one day) for their on-going support. I also want to thank my amazing MoBi buddies from Heidelberg, without whom I would have probably not even finished my bachelor's degree and definitely would had a lot less fun doing so. Finally thank you Rosemary for making sure I am not going mad while being stuck in home office writing my thesis, it would have been so much more difficult without you.

Again, thanks to all of you, it was an amazing four years for me.

Christoph

7 REFERENCES

Alberts B, Johnson A, Lewis J *et al. Molecular Biology of the Cell.* Garland Science, 2015.

Albrecht G, Mösch HU, Hoffmann B *et al.* Monitoring the Gcn4 protein-mediated response in the yeast Saccharomyces cerevisiae. *J Biol Chem* 1998;**273**:12696–702.

Amrane M, Oukid S, Gagaoua I *et al.* Breast cancer classification using machine learning. *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2018*. Institute of Electrical and Electronics Engineers Inc., 2018, 1–4.

Axelrod JD, Majors J, Brandriss MC. Proline-independent binding of PUT3 transcriptional activator protein detected by footprinting in vivo. *Mol Cell Biol* 1991;**11**:564–7.

Bergenholm D, Liu G, Holland P *et al.* Reconstruction of a Global Transcriptional Regulatory Network for Control of Lipid Metabolism in Yeast by Using Chromatin Immunoprecipitation with Lambda Exonuclease Digestion. Wilmes P (ed.). *mSystems* 2018;**3**:e00215-17.

De Boer CG, Hughes TR. YeTFaSCo: A database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res* 2012;**40**, DOI: 10.1093/nar/gkr993.

Broun P. Transcription factors as tools for metabolic engineering in plants. *Curr Opin Plant Biol* 2004;**7**:202–9.

Carey MF, Peterson CL, Smale ST. Chromatin Immunoprecipitation (ChIP). *Cold Spring Harb Protoc* 2009;**2009**:pdb.prot5279.

Carrillo E, Ben-Ari G, Wildenhain J *et al.* Characterizing the roles of Met31 and Met32 in coordinating Met4-activated transcription in the absence of Met30. *Mol Biol Cell* 2012;**23**:1928–42.

Chambers A, Packham EA, Graham IR. Control of glycolytic gene expression in the budding yeast (Saccharomyces cerevisiae). *Curr Genet* 1995;**29**:1–9.

Cherry JR, Johnson TR, Dollard C *et al.* Cyclic AMP-dependent protein kinase phosphorylates and inactivates the yeast transcriptional activator ADR1. *Cell* 1989;**56**:409–19.

Danino YM, Even D, Ideses D *et al.* The core promoter: At the heart of gene expression. *Biochim Biophys Acta - Gene Regul Mech* 2015;**1849**:1116–31.

Daran-Lapujade P, Jansen MLA, Daran JM *et al.* Role of Transcriptional Regulation in Controlling Fluxes in Central Carbon Metabolism of Saccharomyces cerevisiae: A chemostat culture study. *J Biol Chem* 2004;**279**:9125–38.

David F, Nielsen J, Siewers V. Flux Control at the Malonyl-CoA Node through Hierarchical Dynamic Pathway Regulation in Saccharomyces cerevisiae. *ACS Synth Biol* 2016;**5**:224–33.

DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science (80-)* 1997;**278**:680–6.

van Dijken J., Bauer J, Brambilla L *et al.* An interlaboratory comparison of physiological and genetic properties of four *Saccharomyces cerevisiae* strains. *Enzyme Microb Technol* 2000;**26**:706–14.

Egriboz O, Goswami S, Tao X *et al.* Self-Association of the Gal4 Inhibitor Protein Gal80 Is Impaired by Gal3: Evidence for a New Mechanism in the GAL Gene Switch. *Mol Cell Biol* 2013;**33**:3667–74.

Erickson HP. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. *Biol Proced Online* 2009;**11**:32–51.

Estruch F. Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. *FEMS Microbiol Rev* 2000;**24**:469–86.

Fisher RA. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Ann Eugen* 1936;**7**:179–88.

Friedman JH. Multivariate Adaptive Regression Splines. *Ann Stat* 1991;**19**:115–23.

Grove A. Regulation of Metabolic Pathways by MarR Family Transcription Factors. *Comput Struct Biotechnol J* 2017;**15**:366–71.

Guo Y, Mahony S, Gifford DK. High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. Aerts S (ed.). *PLoS Comput Biol* 2012;**8**:e1002638.

Haberle V, Forrest ARR, Hayashizaki Y *et al.* CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res* 2015;**43**:e51–e51.

Haberle V, Lenhard B. Promoter architectures and developmental gene regulation. *Semin Cell Dev Biol* 2016;**57**:11–23.

Hackett SR, Baltz EA, Coram M *et al.* Learning causal networks using inducible transcription factors and transcriptome-wide time series. *Mol Syst Biol* 2020;**16**, DOI: 10.15252/msb.20199174.

Hahn S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* 2004;**11**:394–403.

Hahn S, Young ET. Transcriptional Regulation in Saccharomyces cerevisiae : Transcription Factor Regulation and. *Genetics* 2011;**189**:705–36.

Hamelinck CN, Faaij APC. Outlook for advanced biofuels. *Energy Policy* 2006;**34**:3268–83.

Harbison CT, Gordon DB, Lee TI *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;**431**:99–104.

He HH, Meyer CA, Hu SS *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods* 2014;**11**:73–8.

Hedges D, Proft M, Entian KD. CAT8, a new zinc cluster-encoding gene necessary for derepression of gluconeogenic enzymes in the yeast Saccharomyces cerevisiae. *Mol Cell Biol* 1995;**15**:1915–22.

Hiesinger M, Roth S, Meissner E *et al.* Contribution of Cat8 and Sip4 to the transcriptional activation of yeast gluconeogenic genes by carbon source-responsive elements. *Curr Genet* 2001;**39**:68–76.

Hinnebusch AG. *Mechanisms of Gene Regulation in the General Control of Amino Acid Biosynthesis in Saccharomyces Cerevisiae*., 1988.

Hinnebusch AG, Natarajan K. Gcn4p, a Master Regulator of Gene Expression, Is Controlled at Multiple Levels by Diverse Signals of Starvation and Stress. *Eukaryot Cell* 2002;**1**:22–32.

Hoffman EA, Frey BL, Smith LM *et al.* Formaldehyde crosslinking: A tool for the study of chromatin complexes. *J Biol Chem* 2015;**290**:26404–11.

Hoskins RA, Landolin JM, Brown JB *et al.* Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res* 2011;**21**:182–92.

Hu Z, Killion PJ, Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet* 2007;**39**:683–7.

Hughes TR, de Boer CG. Mapping yeast transcriptional networks. *Genetics* 2013;**195**:9–36.

Humbird D, Davis R, McMillan JD. Aeration costs in stirred-tank and bubble column bioreactors. *Biochem Eng J* 2017;**127**:161–6.

Hwang YC, Lin CC, Chang JY *et al.* Predicting essential genes based on network and sequence analysis. *Mol Biosyst* 2009;**5**:1672–8.

Iyer V, Struhl K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J* 1995;**14**:2570–9.

Jain M, Olsen HE, Paten B *et al.* The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* 2016;**17**, DOI: 10.1186/s13059-016-1103-0.

Johnson DS, Mortazavi A, Myers RM *et al.* Genome-wide mapping of in vivo protein-DNA interactions. *Science (80-)* 2007;**316**:1497–502.

Kealey JT, Liu L, Santi D V *et al.* Production of a polyketide natural product in nonpolyketide-producing prokaryotic and eukaryotic hosts. *Proc Natl Acad Sci U S A* 1998;**95**:505–9.

Kirkpatrick CR, Schimmel P. Detection of leucine-independent DNA site occupancy of the yeast Leu3p transcriptional activator in vivo. *Mol Cell Biol* 1995;**15**:4021–30.

Kodzius R, Kojima M, Nishiyori H *et al.* CAGE: cap analysis of gene expression. *Nat Methods* 2006, DOI: 10.1038/nmeth0306-211.

Kohlhaw GB. Leucine Biosynthesis in Fungi: Entering Metabolism through the Back Door. *Microbiol Mol Biol Rev* 2003;**67**:1–15.

Komeili A, O'Shea EK. Roles of phosphorylation sites in regulating activity of the transcription factor pho4. *Science (80-)* 1999;**284**:977–80.

Komeili A, Wedaman KP, O'shea EK *et al.* Mechanism of Metabolic Control: Target of Rapamycin Signaling Links Nitrogen Quality to the Activity of the Rtg1 and Rtg3 Transcription Factors. *J Cell Biol* 2000;**151**:863–78.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.

Lee TI, Rinaldi NJ, Robert F *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Sci Signal* 2002;**298**:799.

Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.

Li Z, Schulz MH, Look T *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* 2019;**20**, DOI: 10.1186/s13059-019-1642-2.

Liu G, Bergenholm D, Nielsen J. Genome-wide mapping of binding sites reveals multiple biological functions of the transcription factor Cst6p in *Saccharomyces cerevisiae*. *MBio* 2016;**7**:1–10.

MacIsaac KD, Wang T, Gordon DB *et al.* An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* 2006;**7**, DOI: 10.1186/1471-2105-7-113.

Marklund E, van Oosten B, Mao G *et al.* DNA surface exploration and operator bypassing during target search. *Nature* 2020, DOI: 10.1038/s41586-020-2413-7.

Maury J, Kannan S, Jensen NB *et al.* Glucose-dependent promoters for dynamic regulation of metabolic pathways. *Front Bioeng Biotechnol* 2018;**6**, DOI: 10.3389/fbioe.2018.00063.

Mueller PP, Hinnebusch AG. Multiple upstream AUG codons mediate translational control of GCN4. *Cell* 1986;**45**:201–7.

Murata M, Nishiyori-Sueki H, Kojima-Ishiyama M *et al.* Detecting expressed genes using CAGE. *Methods Mol Biol* 2014;**1164**:67–85.

Nebert DW. Transcription factors and cancer: an overview. *Toxicology* 2002:213–35.

Ouldali H, Sarthak K, Ensslen T *et al.* Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat Biotechnol* 2020;**38**:176–81.

Ouyang L, Holland P, Lu H *et al.* Integrated analysis of the yeast NADPH-regulator Stb5 reveals distinct differences in NADPH requirements and regulation in different states of yeast metabolism. *FEMS Yeast Res* 2018;**18**, DOI: 10.1093/femsyr/foy091.

Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A* 2009;**106**:21521–6.

Park K-S, Lee D, Lee H *et al.* Phenotypic alteration of eukaryotic cells using randomized libraries of artificial transcription factors. *Nat Biotechnol* 2003;**21**:1208–14.

Park SH, Kim S, Hahn JS. Metabolic engineering of Saccharomyces cerevisiae for the production of isobutanol and 3-methyl-1-butanol. *Appl Microbiol Biotechnol* 2014;**98**:9139–47.

Parky D, Morrisy AR, Battenhouse A *et al.* Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic Acids Res* 2014;**42**:3736–49.

Peng B, Williams TC, Henry M *et al.* Controlling heterologous gene expression in yeast cell factories on different carbon substrates and across the diauxic shift: A comparison of yeast promoter activities. *Microb Cell Fact* 2015;**14**, DOI: 10.1186/s12934-015-0278-5.

dos Reis M, Wernisch L, Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res* 2003;**31**:6976–85.

Ren B, Robert F, Wyrick JJ *et al.* Genome-wide location and function of DNA binding proteins. *Science (80-)* 2000;**290**:2306–9.

Reyes A, Huber W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res* 2018;**46**:582–92.

Reyes M, Billman K, Hacohen N *et al.* Simultaneous Profiling of Gene Expression and Chromatin Accessibility in Single Cells. *Adv Biosyst* 2019;**3**:1900065.

Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 2011;**147**:1408–19.

Rhee HS, Pugh BF. ChiP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol* 2012:1–14.

Rossi MJ, Lai WKM, Pugh BF. Simplified ChIP-exo assays. *Nat Commun* 2018;**9**:2842.

Rotem A, Ram O, Shoresh N *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* 2015;**33**:1165–72.

Rubinfeld DJ, Harrison DJ. Hedonic housing prices and the demand for clean air. *J Enviromental Econ Manag* 1978;**5**:81–102.

Salazar AN, de Vries ARG, van den Broek M *et al.* Nanopore sequencing enables near-complete de novo assembly of Saccharomyces cerevisiae reference strain CEN.PK113-7D. *FEMS Yeast Res* 2017;**17**, DOI: 10.1093/femsyr/fox074.

Sandelin A, Carninci P, Lenhard B *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* 2007;**8**:424–36.

Sasano Y, Watanabe D, Ukibe K *et al.* Overexpression of the yeast transcription activator Msn2 confers furfural resistance and increases the initial fermentation rate in ethanol production. *J Biosci Bioeng* 2012;**113**:451–5.

Scalcinati G, Knuf C, Partow S *et al.* Dynamic control of gene expression in Saccharomyces cerevisiae engineered for the production of plant sesquitepene α-santalene in a fed-batch mode. *Metab Eng* 2012;**14**:91–103.

Schmitt AP, Mcentee K. Msn2p, a zinc finger DNA-binding protein, is the transcriptional activator of the multistress response in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 1996;**93**:5777–82.

Senior AW, Evans R, Jumper J *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* 2020;**577**:706–10.

Singh R, Lanchantin J, Robins G *et al.* DeepChrome: Deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 2016;**32**:i639–48.

Smale ST, Kadonaga JT. The RNA Polymerase II Core Promoter. *Annu Rev Biochem* 2003;**72**:449–79.

Solomon MJ, Varshavsky A. Formaldehyde-mediated DNA-protein crosslinking: A probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A* 1985;**82**:6470–4.

Struhl K. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* 1999;**98**:1–4.

Teixeira MC, Monteiro PT, Palma M *et al.* YEASTRACT: An upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. *Nucleic Acids Res* 2018;**46**:D348–53.

Wery M, Descrimes M, Vogt N *et al.* Nonsense-mediated decay restricts LncRNA Levels in Yeast Unless Blocked by Double-Stranded RNA Structure. *Mol Cell* 2016;**61**:379–92.

Workman CT, Mak HC, McCuine S *et al.* A systems approach to mapping DNA damage response pathways. *Science* 2006;**312**:1054–9.

Wu J, Zhang N, Hayes A *et al.* Global analysis of nutrient control of gene expression in Saccharomyces cerevisiae during growth and starvation. *Proc Natl Acad Sci U S A* 2004;**101**:3148–53.

Zaret KS, Carroll JS. Pioneer transcription factors: Establishing competence for gene expression. *Genes Dev* 2011;**25**:2227–41.

Zeitlinger J, Simon I, Harbison CT *et al. Program-Specific Distribution of a Transcription Factor Dependent on Partner Transcription Factor and MAPK Signaling Genetic Assays Indicate That Both MAPKs Activate Ste12.*, 2003.

Zhang Y, Nielsen J, Liu Z. Engineering yeast metabolism for production of terpenoids for use as perfume ingredients, pharmaceuticals and biofuels. *FEMS Yeast Res* 2017;**17**, DOI: 10.1093/femsyr/fox080.

Zhang Z, Li J, Cui P *et al.* Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. *BMC Bioinformatics* 2012;**13**:43.